

# ANALYSIS AND MODELING SPATIOTEMPORAL EVENTS ON COMPLEX SPATIAL REGIONS

**Somnath Chaudhuri**

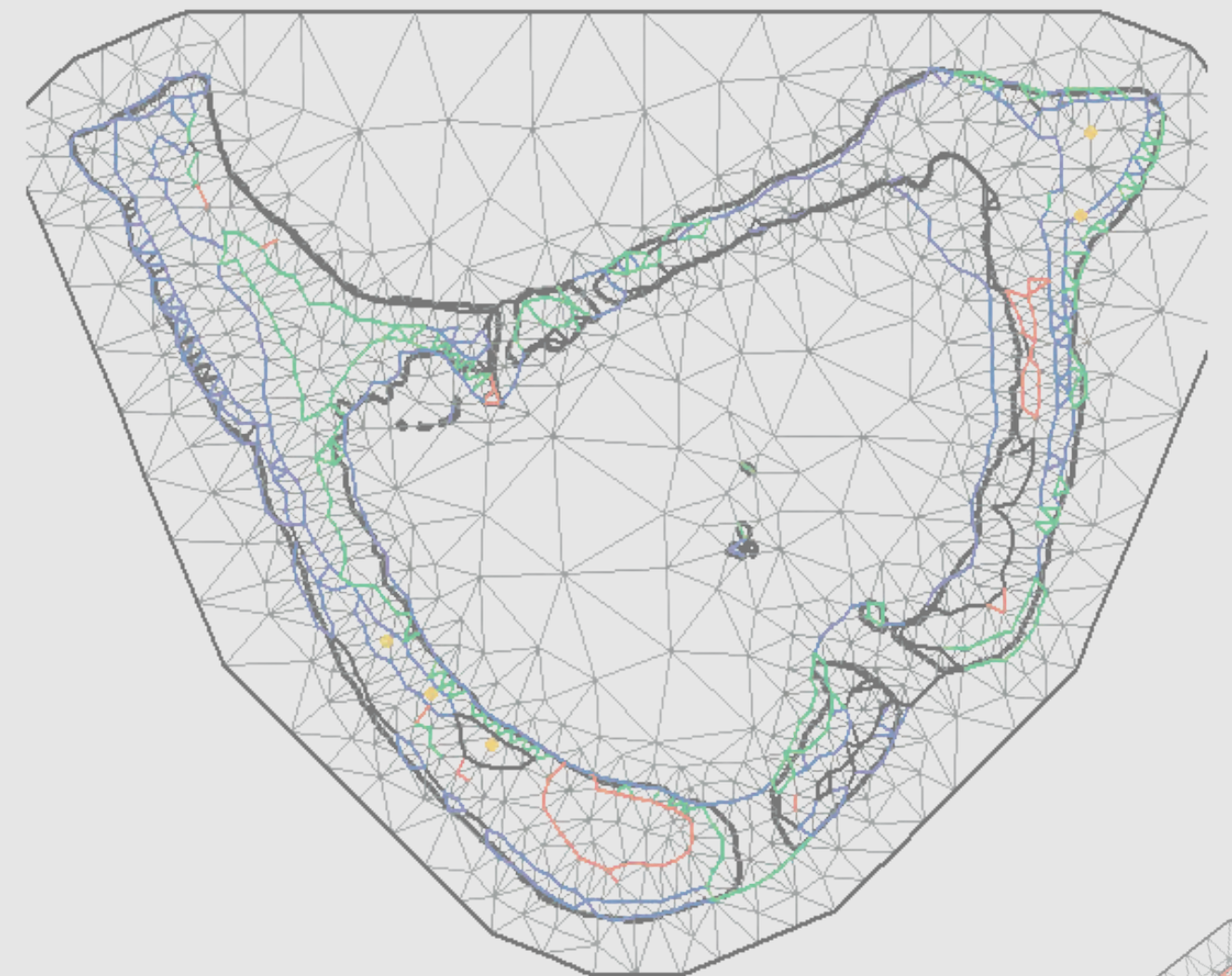
**ADVERTIMENT.** L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

**ADVERTENCIA.** El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

**WARNING.** Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

# DOCTORAL THESIS

## ANALYSIS AND MODELING SPATIOTEMPORAL EVENTS ON COMPLEX SPATIAL REGIONS



2023  
Somnath Chaudhuri

2023 | DOCTORAL THESIS SOMNATH CHAUDHURI





**DOCTORAL THESIS**

**Analysis and Modeling Spatiotemporal Events  
on Complex Spatial Regions**

**Somnath Chaudhuri**

**2023**

PhD program in Molecular Biology, Biomedicine and Health

Under the supervision of:  
Prof. Marc Saez Zafra and Dr. Pablo Juan Verdoy

Thesis delivered to obtain the doctoral degree by the Universitat de Girona





El **Prof. Marc Saez Zafra**, de la Universitat de Girona i membre del Grup de Recerca en Estadística, Econometria i Salut (GRECS) i el **Dr. Pablo Juan Verdoy**, de la Departamento de Matemáticas, Universidade Jaume I, Castellón,

### CERTIFIQUEM:

Que el treball titulat Analysis and Modeling Spatiotemporal Events on Complex Spatial Regions, que presenta Somnath Chaudhuri per a l'obtenció del títol de doctor, ha estat realitzat sota la seva direcció i que compleix els requisits per poder optar a Menció Internacional. La tesi es presenta com un compendi d'articles, indicant la idoneïtat d'aquest format i demostrant la rellevància de la contribució específica del doctorand a les publicacions presentades. Així mateix, la tesi reflecteix fidelment el treball realitzat pel doctorand, que ha estat elaborada d'acord amb el codi de bones pràctiques de l'Escola de Doctorat i que no conté cap element plagiat.

I, perquè així consti i tingui els efectes oportuns, signem aquest document.

Prof. Marc Saez Zafra

*Universitat de Girona, Girona*

Dr. Pablo Juan Verdoy

*Universidade Jaume I, Castellón*

Girona, de de 2023



# ACKNOWLEDGEMENTS

---

I would like to take this opportunity to express my sincere gratitude to all those who have supported and guided me throughout my PhD journey. First and foremost, I would like to thank my supervisor, Dr. Marc Saez for his constant encouragement, invaluable guidance, and unwavering support. I would like to express my sincere gratitude to my co-supervisor Dr. Pablo Juan for his continuous support, motivation, and thoughtful comments. I am extremely lucky to have supervisors like them who respond to my every query so promptly and with patience. Their inimitable energetic style of inspiration makes me enjoy the work and complete it on scheduled time. Their expertise, patience, and kindness have been instrumental in shaping my research and helping me to reach this important milestone in my academic career.

My sincere thanks go to my research guide in the King Abdullah University of Science and Technology (KAUST) Dr. Håvard Rue for his thoughtful comments, guidance and immense knowledge for the dedication of research.

Several faculty members and colleagues from various departments of different universities have played a crucial role in my research. I would like to express my sincere gratitude for their valuable contributions of Dr. Elias Teixeira Krainski and Dr. David Bolin (KAUST, Saudi Arabia), Dr. Diego Varga, Dr. Maria Antonia Barceló Rado and Dr. Laura Serra (University of Girona, Spain) and Dr. Jorge Mateu and Dr. Sergio Trilles Oliver (University Jaume I, Spain).

I am also thankful to all the respected teachers from the University of Girona for sharing their knowledge during the course, who through their teachings made me equipped with the necessary background needed to undertake such a work.

I must express my gratitude to my family, especially to my parents, their unconditional love and support have been my source of strength throughout my life and academic pursuits. Their continuous encouragement and belief in me have been a constant source of motivation and inspiration.

I owe special thanks to a very special person, my life-partner, Pam, for her motivation and for always being there for me. I cannot thank enough for her patience, understanding, and continuous support throughout this journey. Her love and support have kept me grounded and focused on achieving my goals, and I am forever grateful to her.

Special thanks go to my friends Marcello, Erin, Lluna, Erik, Ignacio, Mutaz, Sadoon, Saurya, Eman and Samrat for their support and encouragement during the times of difficulty.



I would like to express my gratitude to the journal editors and anonymous reviewers for their valuable comments and suggestions on my scientific publications.

I acknowledge the University of Girona for funding my PhD scholarship through IFUdG Scholarship program in collaboration of Banco Santander, Spain. I intend to extend my heartfelt thanks to the Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain and CIBER of Epidemiology and Public Health (CIBERESP), Spain for their support and sponsorship throughout my thesis work.

I would like to thank the R-INLA community and Informació, tràmits i serveis de la Generalitat de Catalunya (gencat) for providing open-source tools, data, and support.

Moltíssimes gràcies a tots!

Somnath Chaudhuri





# FUNDING

---

This doctoral thesis was supported by:

**Universitat de Girona (UdG)** through:

- a predoctoral contract through IFUdG Scholarship program in collaboration of Banco Santander, Spain.

**King Abdullah University of Science & Technology (KAUST)** through:

- supported a 4-month stay (from 5<sup>th</sup> February to 28<sup>th</sup> May 2022) at the Division for Computer, Electrical and Mathematical Science and Engineering, KAUST, Saudi Arabia under the supervision of Prof. Håvard Rue.



# LIST OF PUBLICATIONS

---

This thesis is presented as a compendium of three publications:

Publication	Quality criteria	Status
Chaudhuri, S., Giménez-Adsuar, G., Saez, M., & Barceló, M. A. (2022). PandemonCAT: Monitoring the COVID-19 Pandemic in Catalonia, Spain. <i>International Journal of Environmental Research and Public Health</i> , 19(8), 4783. <a href="https://doi.org/10.3390/ijerph19084783">https://doi.org/10.3390/ijerph19084783</a>	Impact Factor (2021): 4.614 Public, Environmental & Occupational Health, position 45 out of 182 (Q1).	Published
Chaudhuri, S., Juan, P., & Mateu, J. (2022). Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation. <i>Journal of Applied Statistics</i> , 1-22. <a href="https://doi.org/10.1080/02664763.2022.2104822">https://doi.org/10.1080/02664763.2022.2104822</a>	Impact Factor (2021): 1.416 Statistics & Probability, position 73 out of 125 (Q3).	Published
Chaudhuri, S., Juan, P., Varga, D., & Saez, M. (2023). Spatiotemporal modeling of traffic risk mapping: A study of urban road networks in Barcelona, Spain. <i>Spatial Statistics</i> . <a href="https://doi.org/10.1016/j.spasta.2022.100722">https://doi.org/10.1016/j.spasta.2022.100722</a>	Impact Factor (2021): 2.125 Statistics & Probability, position 41 out of 125 (Q2).	Published

Other related publications during the thesis period are listed in **Annex 1**.



# LIST OF ABBREVIATIONS

---

**BYM**, Besag-York-Mollié  
**CAR**, Conditional Autoregressive  
**CBD**, Central Business District  
**CPO**, Conditional Predictive Ordinates  
**CV**, Cross Validation  
**DIC**, Deviance Information Criterion  
**EuroRAP**, European Road Assessment Program  
**GF**, Gaussian Field  
**GIS**, Geographic Information Systems  
**GLM**, Generalized Linear Model  
**GLMM**, Generalized Linear Mixed Models  
**GMRF**, Gaussian Markov Random Field  
**HDX**, Humanitarian Data Exchange  
**INLA**, Integrated Nested Laplace Approximations  
**LGCP**, Log Gaussian Cox Process  
**LSOA**, Layer Super Output Area  
**MAUP**, Modifiable Areal Unit Problem  
**MCMC**, Markov Chain Monte Carlo  
**MIDP**, Misaligned Data Problem  
**NPI**, Nonpharmaceutical Interventions  
**ODbL**, Open Database License  
**OSM**, Open Street Map  
**PC**, Penalized Complexity  
**PCA**, Principal Component Analysis  
**POI**, Points of Interest  
**RH**, Relative Humidity  
**SAR**, Simultaneous Autoregressive  
**SDM**, Species Distribution Model  
**SPDE**, Stochastic Partial Differential Equation  
**WAIC**, Watanabe-Akaike Information Criterion  
**WHO**, World Health Organization





# LIST OF FIGURES

---

Figure 1: Realizations of areal data .....	18
Figure 2: Neighborhood structure in areal data .....	19
Figure 3: Example of geostatistical data .....	20
Figure 4: Regular and irregular lattice.....	24
Figure 5: Three triangulations for the SPDEtoy dataset.....	29
Figure 6: Example of complex spatial region .....	33
Figure 7: Traffic accident locations on buffered road segments .....	36
Figure 8: Clipped polygon of buffered road segments .....	37
Figure 9: SPDE triangulation for entire study area and network mesh (London, UK) .....	37
Figure 10: SPDE triangulation for entire study area and network mesh (Barcelona, Spain).....	38
Figure 11: Barrier objects.....	42
Figure 12: Mesh with barrier object .....	42
Figure 13: Graph data structure of traffic accident (Barcelona, Spain).....	44
Figure 14: Republic of Maldives geographical location and island structure .....	114
Figure 15: Locations of tsunami affected islands for individual atolls of Maldives .....	117
Figure 16: Tsunami affected regions of Baa and Raa atolls.....	118
Figure 17: SPDE triangulation with tsunami affected regions .....	121
Figure 18: Barrier object and barrier object with SPDE triangulation .....	122
Figure 19: Combined spatial region for Haa Alifu, Haa Dhaalu, Shaviyani and Noonu atolls .....	127
Figure 20: Combined spatial region for Raa and Baa atolls.....	128
Figure 21: Combined spatial region for Gaafu Alifu and Gaafu Dhaal atolls .....	128
Figure 22: Mesh for Shaviyani group atolls .....	129
Figure 23: Mesh for Baa group atolls .....	129
Figure 24: Mesh for Kaafu atoll .....	130
Figure 25: Mesh for Meemu atoll.....	130
Figure 26: Mesh for Laamu atoll .....	131
Figure 27: Mesh for GDhaalu group atolls.....	131
Figure 28: Mesh for Seenu atoll .....	131
Figure 29: Geographical location and road network with traffic accidents in Barcelona .....	138
Figure 30: Region mesh for Barcelona road network .....	139
Figure 31: Buffered road polygon and network mesh.....	139
Figure 32: Barrier object and mesh with barrier object.....	142
Figure 33: Graph data structure of the traffic accident locations as nodes and road networks as edges .....	146
Figure 34: Road network conversion to graph data structure .....	147
Figure 35: Marginal posterior mean of the spatial random effect .....	150
Figure 36: Marginal posterior distributions of network mesh model hyperparameters.....	150
Figure 37: Marginal posterior distributions of barrier model hyperparameters .....	151
Figure 38: Marginal posterior distributions of graph model hyperparameters .....	151



# Resum

---

L'estadística espacial es basa tradicionalment en models estacionaris com els camps de Matérn. Tot i això, l'aplicació de models estacionaris a regions espacials complexes que tenen barreres físiques com illes o àrees costaneres pot resultar en un suavitzat inadequat d'aquestes regions. A més, en moltes aplicacions ambientals, com a sistemes de rierols o xarxes de carreteres urbanes, és essencial definir models estadístics en xarxes lineals.

La tesi de recerca actual explora els beneficis i les limitacions de les aproximacions de Laplace imbricades integrades (INLA) juntament amb l'equació diferencial parcial estocàstica tradicional (SPDE) per al modelatge espaciotemporal bayesià. L'estudi se centra en regions espacials distribuïdes complexes que tenen barreres físiques, així com en xarxes lineals com les xarxes de carreteres urbanes.

La motivació darrere de l'article de recerca inicial és dissenyar una aplicació per monitoritzar la dinàmica de la pandèmia de COVID-19 en un context espai-temporal a la regió de Catalunya, Espanya. En aquest cas, hem utilitzat INLA-SPDE, però en regió espacial contínua. Els dos articles següents van involucrar l'ús de triangulació de xarxa explícita per explorar i analitzar l'ocurrència d'accidents de trànsit a les xarxes vials urbanes al Regne Unit i Espanya. Vam proposar el nou concepte de triangulació espacial restringida a xarxes lineals. Però les regions frontereres complexes creen estructures espacials fictícies que donen com a resultat dependències espacials artificials. Als següents articles proposats, hem explorat estratègies computacionals alternatives per dissenyar models de barrera no estacionaris. Inicialment, hem utilitzat el model de barrera per analitzar la variació espacial del risc de tsunami a la República de Maldives. Després implementem models de barrera a xarxes lineals. Però en tots dos casos, els límits es troben dins del domini espacial d'interès, cosa que impedeix que es redueixin els efectes dels límits alts. L'article final proposat presenta una estratègia nova per utilitzar mètriques no euclidianes en estructures gràfiques, com a alternativa a la metodologia de distància euclidiana convencional. En aquest cas, és un desafiament trobar classes flexibles de funcions que siguin definides positives per formular camps gaussians en gràfics mètrics. Utilitzant el concepte esmentat, s'ha desenvolupat una nova categoria de processos gaussians en gràfics mètrics compactes. Els camps de Whittle-Matérn emprats en aquest enfocament es defineixen mitjançant un SPDE fraccionari en un gràfic mètric. Els camps proposats són una extensió natural dels camps gaussians amb funcions de covariància de Matérn en dominis euclidians a configuracions gràfiques mètriques no euclidianes.

S'ha fet servir un període de deu anys (2010-2019) de registres diaris d'accidents de trànsit de Barcelona, Espanya, per avaluar els tres models esmentats anteriorment. En comparar el rendiment del model, utilitzant mètriques d'avaluació, observem que l'SPDE fraccionari proposat al model de gràfic de mètriques supera la triangulació de xarxa i els models de barrera. A causa d'aquesta flexibilitat, es pot aplicar a una àmplia gamma de problemes ambientals, especialment aquells que involucren regions espacials complexes o distribuïdes, com ara illes, xarxes de carreteres o àrees delimitades per límits.



# Resumen

---

La estadística espacial se basa tradicionalmente en modelos estacionarios como los campos de Matérn. Sin embargo, la aplicación de modelos estacionarios a regiones espaciales complejas que tienen barreras físicas como islas o áreas costeras puede resultar en un suavizado inadecuado de tales regiones. Además, en muchas aplicaciones ambientales, como sistemas de arroyos o redes de carreteras urbanas, es esencial definir modelos estadísticos en redes lineales.

La tesis de investigación actual explora los beneficios y las limitaciones de las aproximaciones de Laplace anidadas integradas (INLA) junto con la ecuación diferencial parcial estocástica tradicional (SPDE) para el modelado espaciotemporal bayesiano. El estudio se centra en regiones espaciales distribuidas complejas que tienen barreras físicas, así como en redes lineales como las redes de carreteras urbanas.

La motivación detrás del artículo de investigación inicial es diseñar una aplicación para monitorear la dinámica de la pandemia de COVID-19 en un contexto espaciotemporal en la región de Cataluña, España. En este caso, hemos utilizado INLA-SPDE pero en región espacial continua. Los siguientes dos artículos involucraron el uso de triangulación de red explícita para explorar y analizar la ocurrencia de accidentes de tráfico en las redes viales urbanas en el Reino Unido y España. Propusimos el novedoso concepto de triangulación espacial restringida a redes lineales. Pero las regiones fronterizas complejas crean estructuras espaciales ficticias que dan como resultado dependencias espaciales artificiales. En los siguientes artículos propuestos, hemos explorado estrategias computacionales alternativas para diseñar modelos de barrera no estacionarios. Inicialmente, hemos utilizado el modelo de barrera para analizar la variación espacial del riesgo de tsunami en la República de Maldivas. Luego implementamos modelos de barrera en redes lineales. Pero en ambos casos, los límites se encuentran dentro del dominio espacial de interés, lo que impide que se reduzcan los efectos de los límites altos. El artículo final propuesto presenta una estrategia novedosa para utilizar métricas no euclidianas en estructuras gráficas, como alternativa a la metodología de distancia euclidiana convencional. En este caso, es un desafío encontrar clases flexibles de funciones que sean definidas positivas para formular campos gaussianos en gráficos métricos. Utilizando el concepto mencionado, se ha desarrollado una nueva categoría de procesos gaussianos en gráficos métricos compactos. Los campos de Whittle-Matérn empleados en este enfoque se definen a través de un SPDE fraccionario en un gráfico métrico. Los campos propuestos son una extensión natural de los campos gaussianos con funciones de covarianza de Matérn en dominios euclidianos a configuraciones gráficas métricas no euclidianas.

Se ha utilizado un período de diez años (2010-2019) de registros diarios de accidentes de tráfico de Barcelona, España, para evaluar los tres modelos mencionados anteriormente. Al comparar el rendimiento del modelo, utilizando métricas de evaluación, observamos que el SPDE fraccional propuesto en el modelo de gráfico de métricas supera la triangulación de red y los modelos de barrera. Debido a esta flexibilidad, se puede aplicar a una amplia gama de problemas ambientales, especialmente aquellos que involucran regiones espaciales complejas o distribuidas, como islas, redes de carreteras o áreas delimitadas por límites.



# Abstract

---

Spatial statistics is traditionally based on stationary models like Matérn fields. However, applying stationary models to complex spatial regions having physical barriers like islands or coastal areas can result in inappropriate smoothing of such regions. Additionally, in many environmental applications such as stream systems or urban road networks, it is essential to define statistical models on linear networks.

The current research thesis explores the benefits and limitations of integrated nested Laplace approximations (INLA) along with traditional stochastic partial differential equation (SPDE) for Bayesian spatiotemporal modeling. The study focuses on complex distributed spatial regions having physical barriers, as well as linear networks like urban road networks.

The motivation behind the initial research article is to design an application to monitor the dynamics of COVID-19 pandemic in a spatiotemporal context in the region of Catalonia, Spain. In this case, we have used INLA-SPDE but in continuous spatial region. The following two articles involved utilizing explicit network triangulation to explore and analyse the occurrences of traffic accidents on urban road networks in UK and Spain. We proposed the novel concept of spatial triangulation restricted to linear networks. But complex boundary regions create fictitious spatial structures resulting in artificial spatial dependencies. In the following proposed articles, we have explored alternative computational strategies to design nonstationary barrier models. Initially, we have used barrier model to analyse spatial variation of tsunami risk in the Republic of Maldives. Then we implemented barrier models on linear networks. But in both cases, boundaries lie within the spatial domain of interest, preventing the high boundary effects from being reduced. The final proposed article presents a novel strategy for utilizing non-Euclidean metric on graph structures, as an alternative to the conventional Euclidean distance methodology. In this case, it is challenging to find flexible classes of functions that are positive definite to formulate Gaussian fields on metric graphs. Utilizing the mentioned concept, a novel category of Gaussian processes has been developed on compact metric graphs. The Whittle-Matérn fields employed in this approach are defined through a fractional SPDE on a metric graph. The proposed fields are a natural extension of Gaussian fields with Matérn covariance functions on Euclidean domains to non-Euclidean metric graph settings.

A ten-year period (2010-2019) of daily traffic-accident records from Barcelona, Spain have been used to evaluate the three models referred above. While comparing model performance using evaluation metrics, we observed that the proposed fractional SPDE on metric graph model outperform network triangulation and barrier models. Due to this flexibility, it can be applied to a wide range of environmental issues, especially those involving complex or distributed spatial regions, such as islands, road networks, or areas demarcated by boundaries.





# Table of Contents

---

<b>ACKNOWLEDGEMENTS</b> .....	<b>III</b>
<b>FUNDING</b> .....	<b>VII</b>
<b>LIST OF PUBLICATIONS</b> .....	<b>IX</b>
<b>LIST OF ABBREVIATIONS</b> .....	<b>XI</b>
<b>LIST OF FIGURES</b> .....	<b>XIII</b>
<b>RESUM</b> .....	<b>XV</b>
<b>RESUMEN</b> .....	<b>XVII</b>
<b>ABSTRACT</b> .....	<b>XIX</b>
<b>TABLE OF CONTENTS</b> .....	<b>XXI</b>
<b>1. INTRODUCTION</b> .....	<b>1</b>
1.1 Background .....	1
1.2 Application of Spatiotemporal Analysis .....	2
1.2.1 Climate and Meteorology .....	2
1.2.2 Ecology and Environmental Management .....	3
1.2.3 Epidemiology and Health .....	3
1.2.4 Urban Issues .....	4
1.2.5 Crime and Antisocial Activities.....	4
1.2.6 Traffic Accidents and Transport Management .....	5
1.2.7 Air Pollution and Health.....	6
1.2.8 Disaster Prevention and Management .....	6
<b>2. RATIONALE</b> .....	<b>7</b>
<b>3. OBJECTIVES</b> .....	<b>8</b>
<b>4. METHODOLOGY</b> .....	<b>9</b>
4.1 Statistical Methods for Spatiotemporal Analysis.....	9
4.2 Overview of Bayesian Inference.....	9
4.3 Markov Chain Monte Carlo (MCMC) .....	11
4.3.1 Challenges in MCMC .....	13
4.4 Integrated Nested Laplace Approximations (INLA).....	13
4.4.1 Overview of R-INLA Package.....	17
4.5 Spatial Data .....	17
4.5.1 Areal Data.....	18
4.5.2 Geostatistical data.....	19
4.5.3 Spatial Point Patterns.....	20
4.6 Extended Geostatistical Paradigm .....	21
4.7 Spatial and Spatiotemporal Modeling .....	22
4.7.1 Modeling for Areal Data.....	23
4.7.2 Spatiotemporal Modeling .....	25
4.8 Stochastic Partial Differential Equation (SPDE) Approach for Geostatistical Data.....	27
4.8.1 Application of INLA-SPDE in Spatial and Spatiotemporal Modeling .....	32
4.9 Challenges in Traditional SPDE Triangulation Approach.....	33

4.9.1	Complex Distributed Spatial Regions.....	33
4.9.2	Modeling on Linear Networks.....	34
4.10	Proposed Methodologies.....	36
4.10.1	Network Triangulation.....	36
4.10.2	Barrier Model.....	39
4.10.3	Application of Barrier Model in Disjoint Spatial Regions.....	40
4.10.4	Application of Barrier Model on Linear Network.....	41
4.10.5	Exponential Graph Model for Linear Network Problems.....	43
<b>5.</b>	<b>RESULTS.....</b>	<b>46</b>
5.1	Article 1: PandemonCAT.....	47
5.2	Article 2: Modeling Traffic Accidents in London, UK.....	70
5.3	Article 3: Risk Mapping Road Networks in Barcelona, Spain.....	93
5.4	Natural Hazards in Islands. Nonstationary Approach with Barriers.....	113
5.4.1	Introduction.....	113
5.4.2	Data Settings.....	116
5.4.3	Methodology.....	119
5.4.4	Results and discussions.....	124
5.5	Enhanced spatial modeling on linear networks using Gaussian Whittle-Matérn family.....	133
5.5.1	Introduction.....	133
5.5.2	Data Settings.....	137
5.5.3	Methodology.....	138
5.5.4	Results and Discussion.....	148
<b>6.</b>	<b>DISCUSSION.....</b>	<b>154</b>
<b>7.</b>	<b>CONCLUSIONS.....</b>	<b>158</b>
	<b>REFERENCES.....</b>	<b>159</b>
<b>8.</b>	<b>ANNEX.....</b>	<b>184</b>
8.1	List of Additional Publications.....	184
8.2	Article 4: Climate Pattern in Complex Islands.....	185
8.3	Article 5: Trend Detection on Linear Networks.....	201





# 1. INTRODUCTION

---

## 1.1 Background

Spatiotemporal events refer to phenomena that occur in both space and time, meaning they have a location and a specific time of occurrence. These events can range from natural phenomena, such as the movement of weather patterns or the migration of animals, to human-made events, such as the movement of vehicles in a city or the spread of a disease. Understanding spatiotemporal events is important in various fields, including geography, environmental science, epidemiology, urban planning, and others for predicting and controlling their outcomes, identifying patterns, and informing decision-making. One of the most significant advantages of studying spatiotemporal events is that it enables the identification of complex patterns that are not evident when analyzing data in space or time alone. For example, in epidemiology, analyzing the spatiotemporal spread of a disease can help to identify areas at high risk and inform targeted interventions to control the outbreak. In urban planning, studying traffic patterns can inform the design of transportation systems and improve traffic flow.

The analysis of spatiotemporal events involves the use of various techniques to examine the patterns and relationships between spatial and temporal components of the event. These techniques include visualization, clustering, and regression analysis, among others. It is difficult to determine the first research work on spatiotemporal analysis, as the concept has been explored by various scholars over time. However, the formal development of spatiotemporal analysis as a distinct field of study began in the mid-20th century, with the development of new technologies such as geographic information systems (GIS) and remote sensing. One of the pioneers in spatiotemporal analysis is Waldo Tobler, who proposed the concept of "first law of geography" in 1970, which states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Tobler also developed the idea of GIS, which are computer-based tools for storing, manipulating, and analyzing spatial data. The research work by Tobler (1970) to model urban growth using computer simulations, with a focus on spatiotemporal patterns is considered as one of the pioneering efforts in this domain. One of the earliest applications of spatiotemporal analysis was in the field of meteorology, where researchers used spatial and temporal data to study weather patterns and make forecasts (Glahn and Lowry, 1972). This approach represented a significant advance over earlier methods, which relied primarily on manual analysis of meteorological data. It also provides a useful historical perspective on the early applications of spatiotemporal analysis in meteorology. In the 1970s and 1980s, the development of GIS technology and the availability of satellite imagery led to the rapid growth of spatiotemporal analysis in fields such as environmental science and urban planning. GIS allowed researchers to integrate spatial and temporal data from a variety of sources and analyze the relationships between environmental factors and land use patterns. Early in 1981, Burrough in his study introduced a methodology for identifying and modeling the spatial dependence between observations. The author proposed using a spatial autocorrelation function to identify the spatial interaction models. Zirschky (1985) demonstrated the use of geostatistical methods like kriging for spatial interpolation, to estimate the yield at unsampled

locations in agricultural field trials. An interesting research work by [Opensaw \(1984\)](#) highlighted the issue of spatial scale in analyses and the potential effect of diverse geographic units on outcomes. This problem is particularly relevant to spatiotemporal analysis since patterns can differ depending on the temporal and spatial resolution of the data. Another key figure in the history of spatiotemporal analysis is Michael Goodchild, who is often credited with coining the term geographic information science (GIScience) in the 1990s ([Goodchild, 1991](#)). His work focused on developing new methods for analyzing spatial data, including spatial statistics and spatial modeling. Following the trend, Later in [1995](#), [Bailey and Gatrell](#) provided a comprehensive overview of spatiotemporal analysis methods, including exploratory data analysis, visualization, and spatial statistics. Remote sensing, GIS, and availability of open data have a significant impact on spatiotemporal modeling, enabling researchers to integrate and analyze spatial and temporal data at a high resolution ([Gitelson et al., 2002](#); [An et al., 2018](#); [Singh 2019](#); [Comber and Wulder, 2019](#); [Song et al., 2019](#); [Muhammad et al., 2022](#); [Apostolopoulos et al., 2022](#)). It was stated early in [2009](#) by [Elith and Leathwick](#), that "analysis of spatiotemporal data is a rapidly developing field, with new statistical models and techniques appearing regularly, and it presents some of the most challenging problems in modern statistics." Overall, literature shows that, spatiotemporal analysis is a rapidly evolving field that seeks to integrate spatial and temporal information to better understand the dynamics of natural and human systems. One of the most notable recent advances in spatiotemporal analysis is the development of big data technologies, which have made it possible to analyze massive amounts of spatial and temporal data ([Li et al., 2017](#); [Wang et al., 2019](#); [Zhou et al., 2021](#)). This has led to new opportunities for understanding and modeling complex spatiotemporal processes, such as climate change, disease outbreaks, and urban growth.

## **1.2 Application of Spatiotemporal Analysis**

Spatiotemporal analysis is an essential tool for understanding complex systems and processes that involve both space and time. Its applications are vast and wide-ranging, and it is becoming increasingly important in fields as diverse as healthcare, environmental science, urban planning, and transportation.

### **1.2.1 Climate and Meteorology**

The application of spatiotemporal analysis has displayed considerable potential in advancing our understanding of weather patterns and climate change within the field of meteorology and climate science. Such research works have led to a better understanding of the relationships between climate variables over space and time, and the identification of the potential impacts of climate change. A crucial use of spatiotemporal analysis in this field is in the enhancement of weather and climate models. [Kumari et al. \(2021\)](#) conducted a study that applied spatiotemporal analysis to enhance the precision of rainfall forecasts in a regional climate model. Correspondingly, in the United States of America (USA), the [National Climate Assessment Report \(2014\)](#) utilized spatiotemporal analysis to assess the impacts of climate change on various sectors in the country, including agriculture, water resources, and human health. Another important area of research in this domain has been the development of high-resolution climate

models that can simulate the complex interactions between the atmosphere, oceans, and land surface. These models can be used to make more accurate climate projections and to investigate the potential impacts of climate change on the region. For example, a study published in 2018 used a high-resolution climate model to project changes in precipitation patterns in Europe under different climate scenarios (Mizuta et al., 2018). Other studies that have investigated the application of spatiotemporal modeling in meteorology and climate science include (Handcock and Wallis, 1994; Compo et al., 2011; Guo et al., 2019; Lin et al., 2020; Chaudhuri et al., 2021; Wang et al., 2022). One area where spatiotemporal analysis has been extensively utilized is in assessing the impacts of climate change on agriculture. For example, a study by Lobell et al. (2011) used spatiotemporal analysis to investigate the relationship between temperature and maize yield in the United States. Similar study to assess the vulnerability of wheat production to climate change in Europe have been conducted by Senapati et al. (2021). In addition, spatiotemporal analysis has been employed to evaluate the effectiveness of adaptation measures in the agricultural sector. For example, a study by Carr et al. (2022) used spatiotemporal analysis to assess the impact of different adaptation strategies on crop production in west African nations.

### **1.2.2 Ecology and Environmental Management**

Similar applications are observable in the domain of environmental management to study the dynamics and patterns of environmental variables and their interactions with ecological systems. It covers a wide range of research works such as monitoring land-use change (Wrenn et al., 2014; Wang et al., 2018), natural resource management (Gruijter et al., 2006; Meseguer Costa et al., 2016; Paradinas et al., 2017) and biodiversity conservation (Adler and Lauenroth, 2003; Yi et al., 2018; Varga et al., 2019). Other relevant research works include urban and regional planning (Bird et al., 2014; Zeng et al., 2015; Wang et al., 2021), marine ecosystem-based management (Dunn et al., 2011; Grüss et al., 2018) and others related to forest fire management (Juan et al., 2012; Serra et al., 2012; Serra et al., 2014a; Serra et al., 2014b; Díaz-Avalos and Juan, 2022). Literature shows that spatiotemporal analysis has a significant impact on environmental management due to its ability to provide a comprehensive understanding of environmental systems and interactions between environmental factors and ecosystems. It allows for the identification of critical areas for conservation and restoration. This can aid in the identification and monitoring of environmental changes over time. This information can be used to evaluate the effectiveness of management strategies, assess the impacts of climate change, and inform policy decisions related to environmental management.

### **1.2.3 Epidemiology and Health**

The application of spatiotemporal analysis in epidemiology and health has been gaining increasing attention over the past decade. It helps the researchers to examine the spatial and temporal patterns of diseases and health outcomes (Cromley and McLafferty, 2012; Juan et al., 2017). By analyzing spatiotemporal patterns of disease outbreaks and related determinants researchers can identify areas of high risk and track the spread of the disease. One area where spatiotemporal analysis has been applied in epidemiology and health is in the study of infectious diseases. For example, spatiotemporal modeling has been used to examine the transmission dynamics of infectious diseases such as malaria, dengue fever, Ebola and Severe Acute



Respiratory Syndrome (SARS) (Yu et al., 2004; Hsieh and Ma, 2009; Bhatt et al., 2013; Faye et al., 2015). Spatiotemporal analysis has also been used to identify hotspots of disease transmission and to understand the impact of socioeconomic factors on public health in different spatial regions (Borrell et al., 2010; Gotsens et al., 2013; Borrell et al., 2014; Hoffmann et al., 2014; Mari-Dell'olmo et al., 2015; Maynou et al., 2015; Maynou-Pujolràs et al., 2016a; Maynou-Pujolràs et al., 2016b; Maynou and Saez, 2016; Povedano et al., 2018; Saez et al., 2018; ). In addition, spatiotemporal analysis has been applied in infectious disease control for animal health (Allepuz et al., 2010; Allepuz et al., 2010; Allepuz et al., 2011). For example, spatiotemporal analysis has been used to study the distribution of cancer and other chronic diseases (Saurina et al., 2010; Puigpinós-Riera et al., 2011; Saez et al., 2013; Renart-Vicens et al., 2014; Aguilar-Palacio et al., 2017; Barceló et al., 2021). Spatiotemporal analysis has also been used to examine the relationship between environmental factors and health outcomes, such as air pollution and asthma (Barceló et al., 2009; Blangiardo et al., 2016; Bennett et al., 2019; Saez and López-Casasnovas, 2019). Studies used spatiotemporal models to investigate the inequalities in suicide mortality rates and the economic recession in England (Saurina et al., 2013) and Catalonia, Spain (Saurina et al., 2015).

During the recent COVID-19 pandemic, spatiotemporal analysis has been extensively utilized to detect and examine patterns and trends in the transmission of the disease (Al-Kindi et al., 2020; Gross et al., 2020; Shariati et al., 2020; Niraula et al., 2022), investigate the impact of covariates (Briz-Redón and Serrano-Aroca, 2020; Díaz-Avalos et al., 2020; Chaudhuri et al., 2022a), and analyze the dynamics of COVID-19 both prior to and subsequent to the development of vaccinations (Grauer et al., 2020; Kraemer et al., 2021; Franch-Pardo et al., 2021; Zapata-Cachafeiro et al., 2022).

Thus, by combining spatial and temporal information, researchers can identify hotspots and high-risk areas for various health outcomes, including infectious diseases, chronic diseases, and environmental exposures and can develop targeted interventions and policies to improve public health.

#### **1.2.4 Urban Issues**

Spatiotemporal analysis has become increasingly important in analyzing urban issues and informing urban planning and policy decisions. It allows researchers to study how urban areas change over time and space and can provide valuable insights into patterns of urban growth and development, as well as the social, economic, and environmental factors that shape them (Braulio-Gonzalo et al., 2016; Bovea et al., 2018a; Bovea et al., 2018b; Braulio-Gonzalo et al., 2021a; Braulio-Gonzalo et al., 2021b; Juan et al., 2022). It also allows researchers to combine and analyze different types of data related to specific locations and time periods, such as population density, land use, transportation, and environmental factors.

#### **1.2.5 Crime and Antisocial Activities**

Spatiotemporal analysis has become increasingly important in analyzing urban issues, such as crime. Crime patterns are not random in space and time but are influenced by various factors such as social, economic, and environmental conditions. Spatiotemporal analysis techniques can

help to identify spatial and temporal patterns in crime that are useful for developing crime prevention strategies, allocation of police resources, and urban planning. Numerous studies have applied spatiotemporal analysis to crime in urban areas (Cusimano et al., 2010; Irvin-Erickson et al., 2015; Rummens et al., 2017; Quick et al., 2019; Zhuang and Mateu, 2019; Boqué et al., 2022; Serra et al., 2022; Vlad et al., 2023). For example, Schutte and Breetzke (2018) found that the relationship between weather conditions and crime varies depending on the type of crime and geographic location. Similarly, Valente (2019) examined the spatial and temporal distribution of violent crimes in a state capital of Brazil. Their study found that homicide rates exhibit significant spatiotemporal clustering, and that high homicide rates are associated with social and economic deprivation, racial segregation, and drug markets. In the same year, Hu et al. (2018) proposed spatial scan statistic for crime, which involves the detection of high-risk areas for crime through the identification of clusters or hotspots of crime incidents. The method takes into account both the spatial and temporal dimensions of the data, which allows for the identification of hotspots that are significant over both space and time. Study by Harries (2006) identify crime concentrations in Baltimore City and explored the motivation behind crime incidents by analyzing the spatiotemporal patterns of specific crime types. The paper by Mata et al. (2016) proposes an approach to generate safe routes in mobile devices by integrating crowd-sensed and official crime data using a semantic processing technique and a Bayes algorithm. The approach is aimed at providing estimations defined by crime rates and uses a geospatial repository to store crime event data in Mexico City. Thus, the application of spatiotemporal analysis to crime in urban areas has provided valuable insights into the spatial and temporal patterns of crime, and the factors that influence crime rates. By identifying high-risk areas and times, policymakers and law enforcement officials can develop targeted interventions to prevent crime and improve the safety and well-being of urban residents (Davis et al., 2005; Frazier et al., 2013; Rummens et al., 2017; Morris et al., 2019; Contreras and Hipp, 2020; Ceccato et al., 2022).

## 1.2.6 Traffic Accidents and Transport Management

Similar to crime, the distribution of traffic accidents in urban areas is also spatiotemporal. The spatiotemporal analysis of traffic accidents involves the use of spatial and temporal data to identify patterns and relationships in traffic accidents, with the goal of predicting and avoiding future accidents. A number of research studies in this domain have shown promising results in improving road safety and reducing traffic accident costs (Plug et al., 2011; Prasannakumar et al., 2011; Wang et al., 2013; Kaygisiz et al., 2015; Liu and Sharma, 2017; Liu and Sharma, 2018; Mahata et al., 2019; Chaudhuri et al., 2022b, Chaudhuri et al., 2023). By understanding the spatiotemporal patterns of traffic accidents, transportation planners and policymakers can implement targeted interventions such as road design improvements, traffic calming measures, and enhanced enforcement strategies to reduce the frequency and severity of accidents.

Another application of spatiotemporal analysis in handling urban challenges is in transportation planning. By analyzing the spatiotemporal patterns of traffic congestion, researchers can identify bottlenecks and areas of high demand. This information can be used to optimize the design of transportation networks and improve the flow of traffic. Literature shows the effectiveness of spatiotemporal modeling in managing traffic congestion in urban environments (Wang et al.,

2013; Duan et al., 2018; Tascikaraoglu, 2018; Zhang et al., 2019; Niu et al., 2019; Afrin and Yodo, 2021; Zeng et al., 2022).

### **1.2.7 Air Pollution and Health**

Air pollution is a major concern in urban areas due to its adverse effects on human health and the environment. In recent years, spatiotemporal analysis has been widely used in urban pollution modeling to assess the distribution and concentration of air pollutants over space and time. The spatiotemporal distribution of air pollution has been studied in many urban areas. For example, spatio-temporal modelling of air pollution by Lindström et al. (2014) is a popular research work in this domain. In another study Lertxundi-Manterola and Saez (2009) modeled nitrogen dioxide (NO<sub>2</sub>) and fine particulate matter (PM<sub>10</sub>) air pollution in the metropolitan areas of Barcelona and Bilbao, Spain. A study by Paoletti et al. (2014) shows that ozone levels in European and USA cities are increasing more than at rural sites but they pointed out that the peak values are having a decreasing trend. Literature shows several research works to investigate the distribution and concentration of air pollutants in urban areas, identify major sources of pollution, and model the impact of mitigation strategies on air quality (Barceló et al., 2009; Yanosky et al., 2014; Barceló et al., 2016; Vicente et al., 2018; Trilles et al., 2019; Vicente et al., 2019; Saez et al., 2020; Mota-Bertran et al., 2021; Trilles et al., 2021; Dimakopoulou et al., 2022; Saez and Barceló, 2022).

### **1.2.8 Disaster Prevention and Management**

Analyzing and understanding the spatial and temporal aspects of natural hazards and disasters has become increasingly important with the advent of spatiotemporal analysis. Some common applications in disaster management include, hazard mapping and risk assessment, damage assessment and emergency response efforts. Studies demonstrate varied use of spatiotemporal analysis in hazard mapping and risk assessment, from landslides (Lateltin et al., 2005; Bednarik et al., 2012; Yang et al., 2015; Nahayo et al., 2019) to floods (Di Baldassarre et al., 2009; Hagemeyer-Klose et al., 2009; Dottori et al., 2016; Franci et al., 2016; Popa et al., 2019; Ha et al., 2023), across different geographic regions. Different research studies show that spatiotemporal analysis can be helpful in assessing damage caused by natural hazards such as cyclones, landslides, and earthquakes (Kiremidjian and Shah, 1997; Stein et al., 2012; Poompavai and Ramalingam, 2013; Giardini et al., 2018; Sahoo et al., 2018; Quesada-Román et al., 2022; Sreejaya et al., 2022). Other studies that have investigated the impacts of spatiotemporal analysis in emergency response management after natural hazards include (Huang and Xiao, 2015; Wang et al., 2016; Han et al., 2019; Karimiziarani et al., 2022).

After analyzing the discussions, it can be scientifically affirmed that the application of spatiotemporal analysis has proven to be a valuable tool across diverse fields such as meteorology, ecology and environmental management, epidemiology and public health, urban issues encompassing crime, pollution, and traffic, as well as disaster prevention and management. This approach aids in the identification and analysis of spatial and temporal patterns of phenomena, providing insights for informed decision-making processes.

## 2. RATIONALE

---

The integrated nested Laplace approximations (INLA) along with stochastic partial differential equation (SPDE) methodology is a powerful statistical modeling tool that combines two effective techniques: INLA, a rapid and precise Bayesian inference method, and SPDE, a flexible and scalable method for spatial modeling, providing significant advantages over other methods. To utilize INLA-SPDE for spatiotemporal modeling, the first essential step is to construct an SPDE triangulation or mesh over the spatial study area.

Literature shows, for continuous spatial regions, traditional SPDE triangulation process is typically suitable and effective. However, in case of complex land structures like coastal areas or distributed islands, or in case of study areas separated by physical barriers like roads or, water bodies, conventional SPDE triangulation processes can be inadequate. This is because these methods inappropriately smooth over physical barriers and can lead to unrealistic assumptions.

Moreover, it has been observed that conventional INLA-SPDE techniques are frequently used to model spatiotemporal events that are strictly confined to linear networks like traffic accidents or street crimes. Although creating a mesh for the entire region facilitates fitting the INLA model, predicting events using this approach can be problematic. This is because the observed events are discrete spatial points that are precisely located on the road network. Models that are fitted using a region mesh cover the entire study area, which can lead to predicted events appearing in areas without road networks, resulting in unrealistic predictions.

As a result, traditional SPDE triangulation methods for spatiotemporal modeling may not be appropriate for all types of spatial regions. Therefore, there is a need for more precise and scientifically sound approaches that can account for physical barriers such as coastlines and accurately forecast events on linear networks while minimizing or eliminating boundary effects in complex spatial regions. Unfortunately, to date, there are limited scientific publications on this topic, emphasizing the need for further research to develop and evaluate innovative techniques. Additional research is necessary to devise and assess advanced techniques that can enhance the accuracy and precision of the INLA-SPDE methodology in modeling spatiotemporal events in complex land structures having physical barriers and on linear networks.

### 3. OBJECTIVES

---

Traditionally, spatial statistics has relied on stationary models such as Matérn fields. However, these models may not be appropriate for analyzing complex spatial regions that have physical barriers like islands or coastal areas, as they can lead to inappropriate smoothing of these regions. Furthermore, in many environmental applications such as stream systems or urban road networks, it is necessary to develop statistical models specifically designed for linear networks. Concerning the challenges associated with modeling complex distributed spatial regions, particularly coastal areas and islands, as well as linear networks, it is evident that a novel and comprehensive modeling approach is required to address these issues. This may entail enhancing and refining the SPDE triangulation approach, particularly with regards to linear networks, or developing a generalized approach to model spatial and spatiotemporal events within complex land structures.

Thus, the principal objective of the current thesis is two-fold:

1. On one side we seek to establish a modeling framework for investigating spatiotemporal phenomena in complex spatial regions with physical barriers.
2. Secondly, we aim to develop an innovative and realistic computational strategy for constructing spatial triangulations that are constrained to linear network topologies.

This framework will allow us to explore options such as nonstationary barrier models, as well as to investigate the potential benefits of alternative models, such as graph models, for enhancing the efficiency of our analyses.

## 4. METHODOLOGY

---

### 4.1 Statistical Methods for Spatiotemporal Analysis

The diverse studies reported in the preceding section establishes the fact that, understanding and analyzing spatiotemporal events is important because it allows us to gain insight into the dynamic behavior of complex systems and develop predictive models that can inform decision-making in many scientific fields. Overall, these works (and many others) contributed to the development of spatiotemporal analysis as a distinct area of research. One of the key benefits of spatiotemporal analysis is the ability to identify patterns and trends in data that may be difficult or impossible to detect using other methods. For example, in ecology, spatiotemporal analysis can be used to track the movements of animals, monitor changes in biodiversity, and identify the impact of climate change on ecosystems. In epidemiology, spatiotemporal analysis can be used to identify clusters of disease outbreaks and track the spread of infectious diseases. Another benefit of spatiotemporal analysis is the ability to develop predictive models that can inform decision-making. Moreover, in urban planning, spatiotemporal analysis can be used to understand population dynamics, track the development of urban sprawl. Similarly, can be implemented to develop models that predict the impact of different development scenarios on traffic congestion, air quality, and other environmental factors. The development of new methods and the availability of large datasets will continue to advance spatiotemporal analysis and its applications in the future.

Thus, exploring spatiotemporal events and their analysis are important in various fields and can provide valuable insights into the underlying processes that drive these events. Some of the key tools used in spatiotemporal analysis include GIS, remote sensing, spatial statistics, and spatial modeling. Statistical methods are commonly used to analyze and understand spatiotemporal events and phenomena. The analysis requires the use of a range of statistical methods that account for both spatial and temporal dimensions of the data. These methods can help identify patterns and relationships in the data and can be used to make predictions about future events. Statistical methods frequently employed in spatiotemporal analysis can be classified into several categories based on their purpose and application such as, exploratory analysis, interpolation methods, regression methods, cluster analysis, geostatistics, time series analysis, principal component analysis (PCA), spatial point pattern analysis, generalized linear model (GLM). These classifications are not mutually exclusive, and some methods can belong to multiple categories. The choice of method will depend on the research question, data type, spatial and temporal scales, and available computational resources.

### 4.2 Overview of Bayesian Inference

With Bayesian analysis, it is possible to use models that are flexible enough to accommodate non-linear associations and data that is not normally distributed. The statistical framework provides a way to estimate model parameters and predict values at unsampled locations. The methodology involves specifying prior distributions for the model parameters, fitting the model using Bayesian inference, and making posterior inferences, which can provide more intuitive and

interpretable results than frequentist  $p$ -values. It is particularly useful for spatiotemporal modeling because it can handle the complex interactions between space and time which can be difficult to model using traditional statistical methods. Moreover, it is not restricted to normally distributed data, making it applicable in a wide range of fields. It can handle a large number of covariates and allows for the inclusion of new covariates at a later stage. Furthermore, it is possible to analyze the significance level of each covariate, which enhances its usefulness in statistical modeling. Additionally, Bayesian inference allows us to quantify uncertainty and make probabilistic statements about the parameters of interest in a statistical model (Moraga, 2019). Bayesian models also allow incorporation of prior knowledge and uncertainty into the analysis. Bayesian inference is based on Bayes' theorem, which relates the conditional probabilities of the data given the parameters (the likelihood function) and the prior probabilities of the parameters. The posterior probability distribution of the parameters is then computed using Bayes' theorem. Mathematically, Bayes' theorem can be expressed as follows:

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{p(y)}$$

where  $p(\theta|y)$  is the posterior distribution of the parameters  $\theta$  (that take values in a parametric space  $(\Theta)$  ) given the observed data  $y$ ,  $p(y|\theta)$  is the likelihood function, which represents the probability of the observed data given the parameters,  $p(\theta)$  is the prior distribution of the parameters, which represents our beliefs about the parameters before observing the data, and  $p(y)$  is the marginal likelihood, which is the probability of the observed data averaged over all possible values of the parameters. The marginal likelihood is equal to

$$\int_{\Theta} p(y | \theta) p(\theta) d\theta$$

which make it difficult to calculate. The posterior distribution  $p(\theta|y)$  is usually a complex, multi-dimensional distribution that can be difficult to calculate for many models since the marginal likelihood  $p(y)$  is often hard to estimate. As a result, the posterior distribution is often estimated without computing the marginal likelihood. This is why Bayes' theorem is often written in the proportional form:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

where the posterior distribution is proportional to the product of the likelihood function and prior distribution (Gomez-Rubio, 2020). If the posterior distribution cannot be obtained in a closed form, alternative methods must be used to estimate or sample from it. In such cases, the Ergodic theorem can be applied to estimate moments and other relevant quantities using a sample from the posterior. (Brooks et al., 2011). Computational techniques generally focus on estimating the integrals that arise in Bayesian inference. For example, the posterior mean of parameter  $\theta_i$  (with values in the parameter space  $(\Theta)$  ) is computed as:

$$\int_{\Theta_i} \theta_i p(\theta_i | y) d\theta_i$$

Distribution  $p(\theta_i|y)$  is the marginal posterior distribution of univariate parameter  $\theta_i$ . Similar integrals can be used to compute various other moments, including the posterior variance. To approximate these integrals, numerical integration methods and the Laplace approximation (Tierney and Kadane, 1986) are considered appropriate techniques. Thus, the posterior distribution provides a complete summary of the uncertainty about the parameters, given the observed data. It can be used to estimate the values of the parameters, make predictions about new data, and quantify the uncertainty in these estimates and predictions. In Bayesian inference, we update our beliefs about the parameters based on the observed data, by computing the posterior distribution. This requires specifying a prior distribution for the parameters, which reflects our beliefs about the parameters before observing the data. It is worthy to mention that the choice of prior distribution can have a significant impact on the posterior distribution, particularly for small or sparse data sets (Coles and Powell, 1996).

In spatiotemporal modeling, Bayesian inference is particularly useful because it allows to estimate the parameters of a model while taking into account uncertainty in the data. By incorporating prior knowledge or beliefs, we can also make better use of limited data and improve the accuracy of our estimates. In addition, Bayesian approaches are capable of handling complex models that are difficult to fit using classical methods such as repeated measures, missing data, and multivariate data (Hanson and Branscum, 2006).

One significant drawback of Bayesian inference is it requires the specification of prior distributions for all model parameters. Choosing appropriate prior distributions can be challenging, and the results can be sensitive to the choice of priors. Moreover, though Bayesian analysis offers the advantage of flexible models for non-linear relationships and non-normal data, one of its challenges is the computation of the posterior distribution. This distribution reflects our level of uncertainty regarding the parameters based on the available data and can be either intractable or difficult to calculate. This has led to the development of several computational methods for approximating the posterior distribution. Two popular methods for approximating the posterior distribution are Markov chain Monte Carlo (MCMC) and Laplace approximation (Moraga, 2019). MCMC is a simulation-based approach that uses a Markov chain to sample from the posterior distribution. Laplace approximation, on the other hand, is a deterministic method that approximates the posterior distribution with a Gaussian distribution centered around the mode of the posterior.

### 4.3 Markov Chain Monte Carlo (MCMC)

MCMC methods are often used in Bayesian analysis to estimate the posterior distribution of model parameters. It is a powerful statistical method used for Bayesian inference in complex models that are difficult or impossible to analyze analytically. Bayesian analysis involves the use of prior knowledge to update beliefs about the probability of a hypothesis given the data. MCMC methods allow for the simulation of samples from the posterior distribution of the model parameters using Markov chains and are widely used in spatiotemporal analysis for modeling complex processes that vary in space and time (Geweke, 1992; Gilks et al. 1996; Brooks et al. 2011). MCMC methods rely on the use of Markov chains to simulate samples from the posterior distribution of the model parameters. A Markov chain is a sequence of random variables, where each variable depends only on the previous variable in the sequence (Gomez-Rubio, 2020). The



Markov chain is constructed so that its equilibrium distribution is the target posterior distribution of interest. The chain is started at some initial value, and then moves from one state to the next according to a transition probability distribution. The transition probability distribution is designed so that the Markov chain is reversible and satisfies the detailed balance condition (Blangiardo et al., 2013). A Markov chain is a sequence of random variables that follows the Markov property, which states that the probability of moving to a new state depends only on the current state, and not on any previous states. Mathematically, it can be expressed as:

$$P(X_{t+1} | X_t, X_{t-1}, \dots, X_0) = P(X_{t+1} | X_t)$$

where  $X_t$  is the state at time  $t$ , and  $X_{t+1}$  is the state at time  $t + 1$ . The MCMC algorithm involves generating a Markov chain of parameter values that converge to the target posterior distribution. The algorithm starts with an initial value of the parameters, and then proposes a new value from a proposal distribution. The proposed value is then accepted or rejected based on its probability, which is calculated using the acceptance ratio. Mathematically, the acceptance ratio is given by:

$$\min\left(1, \frac{P(D|\theta') \cdot P(\theta')}{P(D|\theta) \cdot P(\theta)}\right)$$

where  $\theta$  is the current parameter value,  $\theta'$  is the proposed parameter value,  $P(D|\theta)$  is the likelihood of the data given the current parameter value,  $P(\theta)$  is the prior distribution of the parameter,  $P(D|\theta')$  is the likelihood of the data given the proposed parameter value, and  $P(\theta')$  is the prior distribution of the proposed parameter value. The MCMC algorithm is designed to generate a Markov chain that converges to the target posterior distribution. Convergence is typically assessed by monitoring the autocorrelation of the chain, which is a measure of how closely the values of the chain are related to each other. A well-converged Markov chain will have low autocorrelation and will produce accurate estimates of the posterior distribution (Cowles and Carlin, 1996).

In summary, the MCMC methodology for Bayesian inference involves generating a Markov chain of parameter values that converge to the target posterior distribution, by proposing new parameter values and accepting or rejecting them based on their probability. MCMC methods have become a standard tool in Bayesian analysis for spatiotemporal analysis because they allow for the incorporation of spatial and temporal dependence in models, while providing a flexible framework for estimating parameters and making predictions. Several studies have implemented MCMC-based spatiotemporal models to understand various environmental and ecological phenomena. For example, Wikle et al. (1998) discuss the use of hierarchical Bayesian models for analyzing spatiotemporal data. The paper presents an overview of the theory and implementation of hierarchical Bayesian models, with a particular emphasis on using MCMC algorithms in a Bayesian framework for parameter estimation. A second study conducted by Wikle in 2003 explores the application of hierarchical Bayesian models in predicting the spread of ecological processes with a focus on using MCMC. Literature shows similar applications in ecological modeling, climate modeling, environmental risk assessment, biodiversity and conservation and in other socio-economic issues (Malve et al., 2007; Gallagher et al., 2009; de Figueiredo et al., 2019; Wang et al., 2022). Interestingly, urban planning and management require informed decision-making under uncertainty, making Bayesian inference a natural fit for urban

applications. In that context, MCMC has been applied in various areas such as transportation modeling, urban planning and urban growth modeling (Brooks et al., 2011; Gielen et al., 2018; Mustafa et al., 2021), crime analysis (Hubin and Storvik, 2018; Bresson et al., 2021) and modeling urban air pollution and effects on public health (Brooks et al., 2011; Zakaria and Noor, 2018; Zhu et al., 2021). MCMC provides a flexible framework for Bayesian inference, allowing the incorporation of complex models and the propagation of uncertainties through these models. Moreover, MCMC is a powerful tool that can make probabilistic predictions and handle uncertainty, which is particularly valuable in epidemiology where data are often incomplete or uncertain (Cauchemez et al., 2004 ; Hamra et al., 2013; Shim et al., 2019; Safford et al., 2021).

### 4.3.1 Challenges in MCMC

Based on the preceding discussions, it can be asserted that MCMC using Bayesian inference is beneficial in spatiotemporal modeling, but it also has certain drawbacks and challenges. One of the primary issues with MCMC in spatiotemporal modeling is that it can be computationally intensive and time-consuming, especially when dealing with large datasets or complex models (Rue et al., 2009; Blangiardo and Cameletti, 2015). Another challenge is that selecting appropriate priors and tuning the MCMC algorithm can be difficult, which can lead to biased or inefficient estimates. Additionally, MCMC can struggle to handle non-linear and non-Gaussian models, which are common in spatiotemporal modeling (Gomez-Rubio, 2020). These issues need to be carefully considered when using MCMC in spatiotemporal modeling.

Recent advances in computing technology and real-time data collection have led to an increase in the complexity and amount of spatiotemporal data available for analysis. MCMC methods require repeated sampling from a complex distribution, which can be computationally expensive, particularly for large datasets. The increase in the amount and complexity of spatiotemporal data has created a big challenge for researchers and analysts who wish to use MCMC methods for modeling and inference. The large number of parameters and the high dimensionality of the data can cause the Markov chains to mix slowly, leading to a long computational time to generate a sufficient number of samples for accurate inference. Additionally, the high correlation between neighboring data points in spatiotemporal data can also slow down the convergence of the Markov chains (Rue et al., 2009; Gómez-Rubio et al., 2014; Zhang et al., 2016). To address these challenges, researchers have developed a range of techniques to make MCMC methods more computationally efficient for spatiotemporal data. Some of these techniques include parallel computing, which involves distributing the computation across multiple processors to speed up the sampling process (Wikle, 2003). Other approaches include using more efficient algorithms specifically designed for spatiotemporal data. In the following section we will discuss in detail about a novel algorithm that uses integrated nested Laplace approximations (INLA) for Bayesian spatiotemporal modeling (Rue et al., 2009).

## 4.4 Integrated Nested Laplace Approximations (INLA)

INLA is a statistical methodology that is specifically designed for modeling latent Gaussian models, which are an extensive and versatile class of models that include linear mixed, spatial, and spatio-temporal models. Due to this versatility, INLA has been successfully applied in a

diverse range of fields, such as [Paul et al. \(2010\)](#), [Martino et al. \(2011\)](#), [Roos and Held \(2011\)](#), [Schrödle and Held \(2011\)](#), [Li et al. \(2012\)](#), [Riebler et al. \(2012\)](#), [Ruiz-Cárdenas et al. \(2012\)](#). This popular computational method has been developed as a computationally efficient alternative to MCMC ([Rue et al., 2009](#)). The approach combines the advantages of two existing Bayesian computation methods, MCMC and Laplace approximation and provide fast and accurate estimates of posterior distributions in complex hierarchical models. In particular, they focus on estimating the posterior marginals of the model parameters. Hence, instead of estimating a complex multivariate joint posterior distribution they focus on obtaining approximations to simple univariate posterior distributions ([Gomez-Rubio, 2020](#)).

The integrated nested Laplace approximation method enables the use of approximate Bayesian inference in latent Gaussian models, including generalized linear mixed models, as well as spatial and spatio-temporal models. Based on the computational properties, INLA focus on latent Gaussian Markov random field (GMRF) models ([Rue et al., 2009](#); [Krainski et al., 2018](#)). This covers a wide range of models as reported by [Rue et al. \(2017\)](#) and [Bakka et al. \(2018\)](#).

Prior to delving into the details of the INLA methodology, it is necessary to first provide an overview of the concept of GMRF. In practice, it is crucial for the latent field to be both Gaussian and a sparse GMRF ([Rue and Held, 2005](#); [Held and Rue, 2010](#)). A GMRF is a Gaussian model with additional conditional independence properties, such that  $x_i$  and  $x_j$  are conditionally independent given the remaining elements  $x_{ij}$  for a subset of pairs  $\{i, j\}$ . A common example is the first-order auto-regressive model,

$$x_t = \phi x_{t-1} + \epsilon_t, \quad t = 1, 2, \dots, m,$$

with  $\phi$  is a constant that determines the relationship between  $x_t$  and  $x_{t-1}$  and Gaussian innovations  $\epsilon$ . Although the resulting covariance matrix is dense, the precision matrix is tridiagonal and can be factorized in  $O(m)$  time, which provides a significant computational advantage ([Rue and Held, 2005](#)).

In general, the computational cost of using GMRFs depends on the sparsity pattern in the precision matrix. For models with a spatial structure, the cost is  $O(m^{3/2})$  paired with a  $O(m \log(m))$  memory requirement ([Rue and Held, 2005](#)). This reduction in both computational and memory requirements make it feasible to run larger models using GMRFs.

In particular, the models are defined by the equations:

$$y_i | \mathbf{x}, \boldsymbol{\theta} \sim \pi(y_i | x_i, \boldsymbol{\theta}), i = 1, \dots, n,$$

$$\mathbf{x} | \boldsymbol{\theta} \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q}(\boldsymbol{\theta})^{-1}),$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$$

where  $\mathbf{y} = (y_1, \dots, y_n)$  is the observed data,  $\mathbf{x}$  is a Gaussian field, and  $\boldsymbol{\theta}$  are hyperparameters. The mean of the latent Gaussian field  $\mathbf{x}$  is given by  $\boldsymbol{\mu}(\boldsymbol{\theta})$  and the precision matrix (i.e., the inverse of the covariance matrix) is given by  $\mathbf{Q}(\boldsymbol{\theta})$ . The observed data  $\mathbf{y}$  and the Gaussian field  $\mathbf{x}$  can both be high-dimensional. As approximations are computed using numerical integration over the hyperparameter space, it is essential to limit the size of the hyperparameter vector  $\boldsymbol{\theta}$  to generate fast inferences ([Rue et al., 2009](#)). In several cases, the observations  $y_i$ , are part of an exponential family whose mean is given by  $\mu_i = g^{-1}(\eta_i)$ .

The linear predictor  $\eta_i$  accounts for the effects of different covariates in an additive manner

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) \quad \dots\dots\dots (1)$$

Here,  $\alpha$  is a scalar representing the intercept, the linear impact of the covariates  $z_{ki}$  on the response is captured by the coefficients  $\beta_k$ , and  $f^{(j)}(\cdot)$  are a set of random effects defined in terms of some covariates  $u_{ji}$ . The term  $n_\beta$  is usually specified prior to model fitting and reflects the number of predictor variables included in the model and  $n_f$  represents the number of covariates. This approach allows for the inclusion of various types of models due to the flexibility of the  $f^{(j)}$  functions, which can take different forms including spatial and spatio-temporal models (Moraga, 2019).

INLA employs a hybrid approach that involves analytical approximations and numerical algorithms for sparse matrices in order to approximate the posterior distributions using closed-form expressions. These approximated posteriors can subsequently undergo post-processing to compute relevant quantities such as posterior expectations and quantiles. Specifically, consider the vector of latent Gaussian variables, where  $\boldsymbol{\theta}$  denotes the vector of hyperparameters that may not follow a Gaussian distribution. This strategy enables quicker inference and sidesteps issues related to sample convergence and mixing, making it feasible to fit large datasets and investigate different models (Rue et al., 2009). It produces fast and precise approximations to the posterior marginals of the components of the latent Gaussian variables represented as:

$$\pi(x_i | \mathbf{y}), i = 1, \dots, n,$$

also, for the posterior marginals for the hyperparameters of the Gaussian latent model

$$\pi(\theta_j | \mathbf{y}), j = 1, \dots, \dim(\boldsymbol{\theta})$$

The posterior marginals of each element  $x_i$  of the latent field  $x$  are

$$\pi(x_i | \mathbf{y}) = \int \pi(x_i | \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$$

and the posterior marginals for the hyperparameters can be represented as

$$\pi(\theta_j | \mathbf{y}) = \int \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}_{-j}$$

The nested formulation is employed to estimate the posterior distribution  $\pi(x_i | \mathbf{y})$  by combining analytical approximations of the full conditionals  $\pi(x_i | \boldsymbol{\theta}, \mathbf{y})$  and  $\pi(\boldsymbol{\theta} | \mathbf{y})$  with numerical integration routines for integrating out  $\boldsymbol{\theta}$ .

Similarly, the approximation of  $\pi(\theta_j | \mathbf{y})$  is achieved by approximating  $\pi(\boldsymbol{\theta} | \mathbf{y})$  and integrating out  $\boldsymbol{\theta}_{-j}$ . Specifically, the Gaussian approximation for the posterior distribution of the latent field,  $\widetilde{\pi}_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ , evaluated at the posterior mode,  $\mathbf{x}^*(\boldsymbol{\theta}) = \arg \max_{\mathbf{x}} \pi_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})$ , is used to estimate the posterior density of the hyperparameters (Rue et al., 2009; Lindgren and Rue, 2015; Rue et al., 2017)

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}$$

Then, INLA constructs the following nested approximations:

$$\tilde{\pi}(x_i|\mathbf{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}$$

Finally, these approximations can be integrated numerically with respect to  $\boldsymbol{\theta}$

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k, \mathbf{y})\tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y}) \times \Delta_k,$$

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \sum_l \tilde{\pi}(\boldsymbol{\theta}_l^*|\mathbf{y}) \times \Delta_l^*$$

where,  $\Delta_k(\Delta_l^*)$  denotes the area weight corresponding to  $\boldsymbol{\theta}_k(\boldsymbol{\theta}_l^*)$ . The approximations for the posterior marginals for the  $x_i$ 's conditioned on selected values of  $\boldsymbol{\theta}_k$ ,  $\tilde{\pi}(x_i|\boldsymbol{\theta}_k, \mathbf{y})$ , can be computed in different ways such as, using a Gaussian, a Laplace, or a simplified Laplace approximation. The simplest and fastest solution is to use a Gaussian approximation derived from  $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  (Rue et al., 2009).

From the above discussions, we can summarize that, INLA combines analytical approximations with numerical integration to obtain approximations of the posterior distributions of the latent Gaussian variables  $\mathbf{x}$  and hyperparameters  $\boldsymbol{\theta}$ , which are computed by integrating out the hyperparameters from the full conditional posterior distributions of the latent Gaussian variables. The posterior marginals are then integrated numerically with respect to the hyperparameters to compute posterior expectations and quantiles. The nested formulation of the approximations allows for accurate and efficient estimation of the posterior distributions of the latent Gaussian variables and hyperparameters (Rue et al., 2009; Rue et al., 2017). Specific advantages of INLA can be reported as, INLA can handle complex hierarchical models with many levels of nesting, which are common in spatiotemporal modeling. This is because INLA approximates the posterior distribution of the model parameters using a Laplace approximation, which provides a computationally efficient way to integrate out the latent variables in the model. Secondly, it allows inclusion of spatial and temporal dependencies in the model using random effects. This allows for the modeling of non-stationary spatial and temporal processes, which is important in many applications such as disease mapping, environmental monitoring, and climate modeling. Finally, INLA can handle missing data and irregularly spaced data, which are common in spatiotemporal data sets. This is because INLA uses a GMRF representation of the spatial and temporal random effects, which allows for the efficient computation of the likelihood and posterior distribution even when the data is incomplete or irregularly spaced.

### 4.4.1 Overview of R-INLA Package

The R-INLA is an R package (Lindgren and Rue, 2015) used to implement approximate Bayesian inference using the INLA approach. INLA website (<http://www.r-inla.org>) provides instructions on extensive documentation, examples, a discussion forum, and other resources that cover the theory and applications of INLA.

In order to utilize INLA for model fitting, there are two necessary steps. First, we must construct the linear predictor of the model as a formula object in R. Then, we call the `inla()` function while specifying the formula, the family, the data, and other options to run the model. Running the `inla()` function generates an object that encompasses information about the fitted model, such as various summaries and the posterior marginals of parameters, linear predictors, and fitted values. To further process these posteriors, R-INLA offers a range of functions. The library includes a list of priors. By default, the intercept of the model is assigned a Gaussian prior with mean and precision equal to 0. The rest of the fixed effects are assigned Gaussian priors with mean equal to 0 and precision equal to 0.001 (Lindgren and Rue, 2015; Rue et al., 2017). It is possible to modify the values of these priors by utilizing the `control.fixed` parameter in `inla()`. This involves assigning a list that contains the mean and precision values of the Gaussian distributions (Moraga, 2019). Simpson et al. (2017) proposed a method for constructing priors by penalizing model component complexity, which has been used in INLA-SPDE models. The R-INLA package offers a valuable framework for constructing priors as *penalized complexity (PC) priors*. These priors can be applied to individual components of a model, providing a flexible expansion of a simple and easily understood base model. PC priors are designed to penalize deviations from the base model, thereby controlling the degree of flexibility in the model and reducing overfitting, leading to improved predictive performance. The PC priors are determined by a single parameter that regulates the level of flexibility in the model. These priors are specified by setting values  $(U, \alpha)$  so that,

$$P(T(\xi) > U) = \alpha$$

where  $T(\xi)$  represents an interpretable transformation of the flexibility parameter  $\xi$ ,  $U$  is an upper bound that specifies a tail event, and  $\alpha$  is the probability of this event (Rue et al., 2009; Blangiardo and Cameletti, 2015). The package also includes estimates of different criteria to evaluate and compare Bayesian models, such as the model deviance information criterion (*DIC*) (Spiegelhalter et al., 2002), the Watanabe-Akaike information criterion (*WAIC*) (Watanabe and Opper, 2010), the marginal likelihood, and the conditional predictive ordinates (*CPO*) (Held et al., 2010).

## 4.5 Spatial Data

Spatial data can be mathematically represented as realizations of a stochastic process that is indexed by space. In particular, spatial data can be represented as a set of observations:

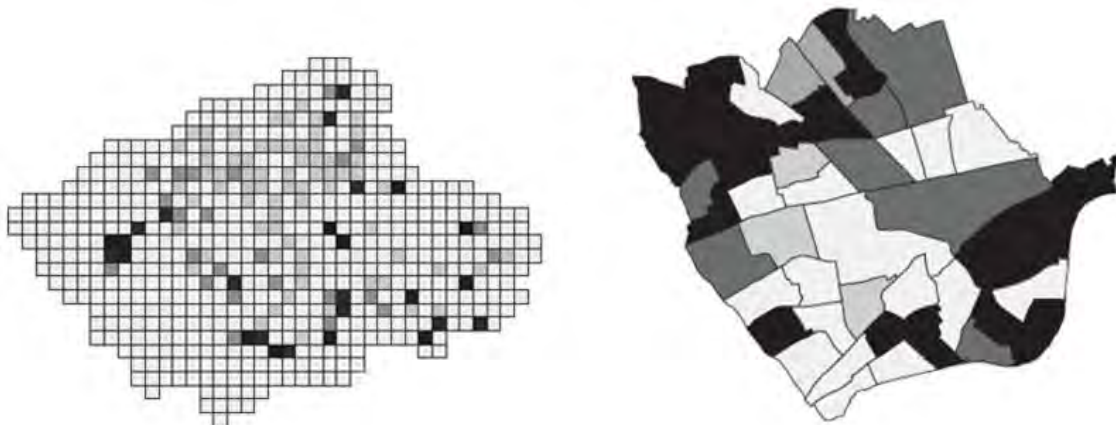
$$Y(s) \equiv y(s), s \in D$$

where  $D$  is a fixed subset of  $R^d$  and  $d = 2$  in this case. Following the specifications by Cressie (1993), Banerjee et al. (2004) and Gelfand et al. (2010), the spatial data is conventionally

classified into three principal domains, which are distinguished based on the nature of the problem and the data involved: areal or lattice data, point-referenced or, geostatistical data, and spatial point patterns. The observations can be thought of as a collection of  $n$  measurements,  $y = y(s_1), \dots, y(s_n)$ , where the set  $(s_1, \dots, s_n)$  denotes the spatial units where the measurements were obtained. Depending on whether  $D$  is a continuous surface or a countable collection of  $d$ -dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al., 2010).

### 4.5.1 Areal Data

Area or lattice data refers to a type of spatial data where a random aggregate value, denoted as  $y(s)$ , is associated with a well-defined areal unit  $s$  in a countable collection of  $d$ -dimensional spatial units. Examples of areal data might include the number of crimes reported in each neighborhood, the percentage of a population that is vaccinated in each county, or the incidence rate of a particular disease in each state. Areal data can be regular or irregular, depending on the shape and size of the geographic units. Regular lattice data is characterized by a fixed grid of equally sized cells, while irregular lattice data has cells of varying shapes and sizes. The main goal of analyzing area or lattice data is often to smooth or map an outcome over a study area.

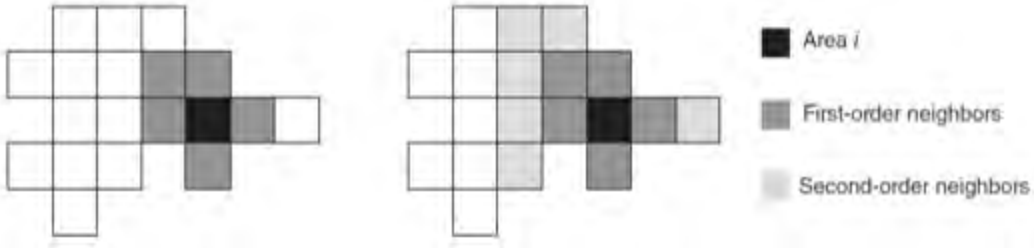


Source: Blangiardo and Cameletti (2015)

**Figure 1: Realizations of areal data  
regular lattice (left) and irregular lattice (right)**

Figure 1 shows two realizations of an areal process, where the left panel displays the proportion of children with respiratory illness in regular lattice and the right panel displays the standardized morbidity ratio of lung cancer in 44 London wards and representing irregular lattice structure. The figure and data presented in Figure 1 is obtained from Blangiardo and Cameletti (2015) and reproduced for this study. Higher tone of colors indicate higher proportion of respiratory illness in left panel and higher standardized morbidity ratio in right panel of Figure 1.

While dealing with area level data, the problem can be reformulated based on the neighborhood structure. The notation  $(s_1, \dots, s_n)$  can be simplified to  $(1, \dots, n)$  by numbering the areas from 1 to  $n$ . The neighbors of an area  $i$  are defined as the areas that share borders with it, including first-order neighbors (i.e., directly adjacent areas) and second-order neighbors (i.e., areas adjacent to the first-order neighbors). This is illustrated in Figure 2.



Source: [Blangiardo and Cameletti \(2015\)](#)

**Figure 2: Neighborhood structure in areal data**  
**first-order neighbors (left), first- and second-order neighbors (right)**

Assuming the Markovian property, which states that the parameter  $\theta_i$  for the  $i$ -th area is independent of all other parameters given its set of neighbors  $N(i)$ , then

$$\theta_i \prod \theta_{-i} \mid \theta_{N(i)}$$

where,  $\theta_{-i}$  denotes all elements in  $\theta$  except  $i$ -th element. This is known as the local Markov property. On the other hand, for any pair of elements  $(i, j)$  in  $\theta$  the nonzero pattern in the precision matrix is given by the neighborhood structure of the process, known as the pairwise Markov property and represented as:

$$\theta_i \prod \theta_j \mid \theta_{-ij} \Leftrightarrow Q_{ij} = 0$$

As a result,  $Q_{ij} \neq 0$  only if  $j \in \{i, N(i)\}$  which is again a GMRF ([Rue and Held, 2005](#)). It is worthy to note that, the independence of  $\theta_i$  from  $\theta_j$  is now not only conditional to the hyperparameters but also to the set of neighbors and we can specify the precision matrix as a function of the structure matrix  $\mathbf{R}$  as:

$$Q = \tau \mathbf{R}$$

where,  $(R_{ij} = N_i$  if  $i = j)$ ,  $(R_{ij} = 1$  if  $i \sim j)$  and  $(R_{ij} = 0)$  otherwise. Here,  $i \sim j$  indicates that  $i$  and  $j$  are neighbors ([Rue and Martino, 2007](#); [Rue et al., 2017](#)).

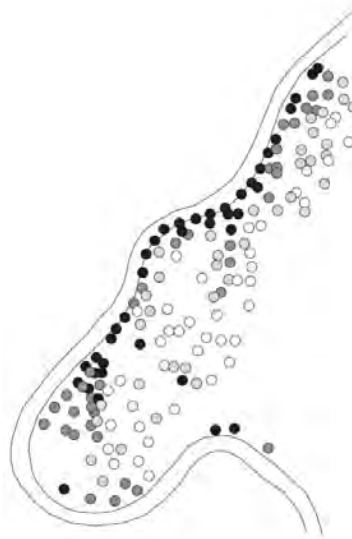
#### 4.5.2 Geostatistical data

Point-referenced or geostatistical data is a type of spatial data where a random outcome, denoted as  $y(s)$ , is associated with a specific location  $s$  in a fixed domain  $D$ . The location  $s$  is typically a two-dimensional vector with latitude and longitude but may also include altitude. The data are represented by a collection of observations  $y = (y(s_1), \dots, y(s_n))$ , where the set  $(s_1, \dots, s_n)$  indicates the locations at which the measurements are taken. The main goal of analyzing geostatistical data is to predict the outcome at unobserved locations in  $D$ . An example of geostatistical data is depicted in [Figure 3](#). The data set named *meuse*, it includes measurements of four heavy metals in the topsoil of a floodplain along the river Meuse, in west Europe. The distribution of heavy metals appears to be influenced by the transportation of polluted sediment by the river, with the majority of deposition occurring near the riverbank and in areas with low



elevation. Figure 3 shows zinc measurements at 155 locations near the river. Higher color tone indicates higher metal concentrations (ppm). Burrough and McDonnell in 1998 introduced this data (Burrough and McDonnell, 1998). This data was originally introduced by The figure and data presented in Figure 3 are obtained from Blangiardo and Cameletti (2015) and reproduced for this study.

Geostatistics is a statistical method used to analyze continuous processes in space. It involves estimation of variables of interest over a study region, based on observations made at a finite number of points. The geostatistical process is often represented as a continuous stochastic process  $y(x)$  with  $x \in D$ , where  $D$  is the study region. The Gaussian Process (GP) is a commonly used model for this process, assuming the stochastic process follows a Gaussian distribution.



Source: Blangiardo and Cameletti (2015)

Figure 3: Example of geostatistical data

The GP is often assumed to be stationary and isotropic, simplifying modeling by assuming that the covariance between two points only depends on their relative distance, rather than their actual positions. This method is useful for understanding the spatial distribution of various variables, such as temperature or pollutants in the air, and has a wide range of applications in fields such as environmental science and geology (Cressie, 2015).

### 4.5.3 Spatial Point Patterns

Spatial point patterns refer to a type of spatial data where the measurement  $y(s)$  represents the occurrence or not of an event, and the locations themselves are random. The spatial domain  $D$  is a set of points in  $R_d$  where events have occurred, such as the locations of trees of a species in a forest or addresses of persons with a particular disease. In this case, while locations  $s \in R_d$  are random, the measurement  $y(s)$  takes 0 or 1 values, depending on whether the event has occurred or not. If additional covariate information is available, it is called a marked point pattern process. The main goal of analyzing spatial point patterns is to evaluate possible clustering or inhibition behavior between observations. *This type of data is beyond the scope of the current research thesis.*

## 4.6 Extended Geostatistical Paradigm

The problem of dealing with exposure and health outcome data recorded at disparate spatial scales is known as the *modifiable areal unit problem* (MAUP) in geography, and as *spatial misalignment* in the statistical literature (Diggle et al., 2013). This problem has been addressed in the epidemiological setting by several authors. Mugglin et al. (2000) proposed a solution based on creating a single, finer partition that includes all nonzero intersections of subregions, where disease counts, and covariate information are recorded on different partitions. Best et al. (2000) assumed that covariate information on a risk factor of interest is available throughout the region of interest, and derived the distribution of observed counts from an underlying Cox process. Kelsall and Wakefield (2002) used a similar approach, except that they used a log-Gaussian latent stochastic process rather than a gamma random field. Low-rank models, such as the class of Gaussian predictive process models proposed by Banerjee et al. (2008) and further developed by Finley et al. (2009), can be used to simplify the technical and computational issues that arise when handling spatial integrals of stochastic processes. Gelfand (2012) provided a useful summary of this and related work.

All of these approaches can be included in a single modeling framework for multiple exposures and disease risk by considering them as a set of spatially continuous processes, irrespective of the spatial resolution at which data elements are recorded. A model for the spatial association between disease risk,  $R(x)$ , and  $m$  exposures  $T_k(x): k = 1, \dots, m$  can be obtained by treating individual case-locations as a log Gaussian Cox process (LGCP) with intensity represented as

$$R(x) = \exp \alpha + \sum_{k=1}^p \beta_k T_k(x) + S(x)$$

where  $S(x)$  denotes stochastic variation in risk that is not captured by the  $p$  covariate processes  $T_k(x)$ .

Let us consider, health outcome data are available in the form of area-level counts,  $Y_i: i = 1, \dots, n$ , in subregions  $A_i$ , while exposure data are collections of unbiased estimates,  $U_{ik}$ , of the  $T_k(x)$  at corresponding locations  $x_{ik}: i = 1, \dots, m_k$ . Consider, further that the  $U_{ik}$  are conditionally independent, with  $U_{ik} | T_k(\cdot) \sim N(T_k(x_{ik}), \tau_k^2)$ , the processes  $T_k(\cdot)$  are jointly Gaussian, and the process  $S(\cdot)$  is also Gaussian and independent of the  $T_k(\cdot)$ . One possible inferential goal is to evaluate the predictive distribution of the risk surface  $R(\cdot)$  given the data  $Y_i: i = 1, \dots, m$  and  $U_{ik}: i = 1, \dots, m_k; k = 1, \dots, p$  (Diggle et al., 2013). Using a simple and easily understandable way of expressing the concept and setting aside the matter of estimating parameters for the moment, the necessary forecasted distribution is denoted as  $[S, T|U, Y]$ . The joint distribution of  $S, T, U$ , and  $Y$  factorizes as

$$[S, T, U, Y] = [S][T][U|T][Y|S, T] \dots \dots \dots (2)$$

where  $[S]$  and  $[T]$  are multivariate Gaussian densities,  $[U|T]$  is a product of univariate Gaussian densities, and  $[Y|S, T]$  is a product of Poisson probability distributions having means  $\mu_i = \int(A_i)R(x)dx$ . The process of sampling from predictive distributions can be carried out using MCMC algorithms. In Bayesian parameter estimation, a suitable joint prior for the model parameters is augmented with the likelihood function (Equation 2) before designing the MCMC

algorithm. In this regard, [Diggle et al. \(2013\)](#) discusses two conventional methods for analyzing data from each subject as a time-sequence of binary responses with explanatory variables: generalized estimating equations ([Liang and Zeger, 1986](#)) and generalized linear mixed models ([Breslow and Clayton, 1993](#)). However, the article proposes an alternative modeling framework for multiple exposures and disease risk by treating them as spatially continuous processes. A model for the spatial association between disease risk and exposures can be obtained by treating individual case-locations as an LGCP with intensity. The LGCPs include time-constant and time-varying covariates, as well as a spatio-temporally continuous Gaussian process. The goal is to evaluate the predictive distribution of the risk surface given the data on health outcome and exposure, which can be carried out using MCMC algorithms.

The traditional approach in geostatistics is to use observations at a finite number of locations to estimate the parameters of a spatial model that can be used to predict values at unsampled locations. But [Diggle et al. \(2013\)](#) in this study propose that geostatistics is a suitable framework for addressing scientific problems that involve spatially continuous processes with spatially discrete observations at a finite number of locations. They suggest that geostatistics should be defined by the class of scientific problems that it addresses, rather than by specific models or data formats. This approach allows for a more flexible and adaptive use of geostatistical methods to address diverse scientific questions. This establishes the extended geostatistical paradigm.

## 4.7 Spatial and Spatiotemporal Modeling

The use of spatial data in inferential processes requires taking into account the spatial trend, which can provide valuable insights and neglecting it may lead to biased estimates. Bayesian approach is effective in handling such data and has been used in various applications like ecology, environmental studies, and infectious diseases. The selection of appropriate models depends on the nature of the data, such as aggregated counts, continuous underlying processes, or point locations. The hierarchical structure can be extended to account for similarities based on neighborhood or distance, and INLA approach can handle the computational challenges associated with the added complexity of spatial structure ([Blangiardo et al., 2013](#); [Bivand et al., 2015](#)).

In order to construct a spatial model within the Bayesian framework, the first step is to specify a probability distribution for the observed data. Typically, a distribution from the Exponential family is selected, with parameters  $\theta$  that account for spatial correlation. To simplify the notation, the subscript  $i$  is used to refer to a generic spatial point or region, rather than an indicator  $s_i$ . While dealing with geostatistical data, the parameters are expressed as a latent stationary Gaussian field (GF) which is a function of hyper-parameters  $\psi$  and associated with a prior distribution  $p(\psi)$  ([Blangiardo et al., 2013](#)). This assumption implies that the multivariate normal distribution of  $\theta$ , with mean  $\mu = (\mu_1, \dots, \mu_n)'$  and a spatially structured covariance matrix  $\Sigma$ , whose generic element is expressed as

$$\Sigma_{ij} = Cov(\theta_i, \theta_j) = \sigma_c^2 C(\Delta_{ij})$$

Here,  $\sigma_c^2$  is the variance component for  $i, j = 1, \dots, n$ .

### 4.7.1 Modeling for Areal Data

INLA can be used to fit a range of spatial models for areal or lattice data, including spatial regression and spatial autoregressive models. INLA provides estimates of the posterior distribution of the model parameters, as well as estimates of the posterior predictive distribution, which can be used to make predictions for new data.

For example, the spatial regression model for areal data using INLA can be written as:

$$y_i = X_i\beta + u_i + e_i$$

where  $y_i$  is the value of the response variable at location  $i$ ,  $X_i$  is a matrix of covariates,  $\beta$  is a vector of coefficients,  $u_i$  is a spatially correlated random effect, and  $e_i$  is a spatially uncorrelated error term. The spatially correlated random effect,  $u_i$ , is modeled as a GMRF, which is a discrete approximation of a continuous spatial process (Banerjee, Carlin, and Gelfand, 2014). Lattice data are rarely observed in regular grid. For example, administrative boundaries usually lead to an irregular lattice. For irregular lattice data, INLA can be used to fit a spatially varying coefficient (SVC) model. The SVC model can be written as:

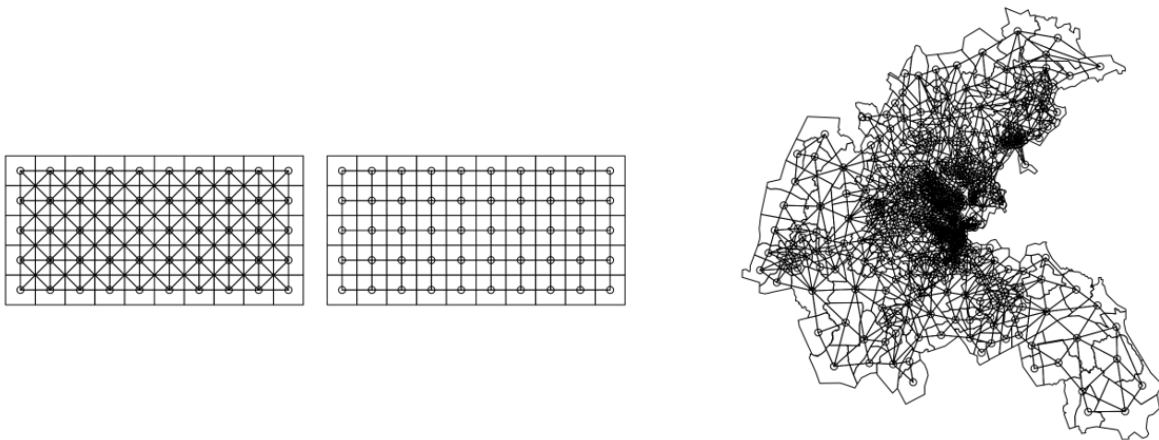
$$y_i = X_i\beta + \text{sum}(w_{ij}u_j) + e_i$$

where  $w_{ij}$  is the weight between location  $i$  and location  $j$ ,  $u_j$  is the spatially correlated random effect at location  $j$ , and the sum is taken over all neighboring locations. The weights can be determined based on the distance between the cells, or on some other measure of spatial similarity. Literature shows the use of spatial modeling for areal or lattice data. For example, Rue et al. (2009) provide a comprehensive overview of the INLA method and its application to spatial modeling. They describe the use of INLA for spatial regression analysis, which can be used to model the relationship between a response variable and a set of covariates, while accounting for spatial correlation among the observations. In addition, Bivand et al. (2013) give a summary of spatial modeling methods using areal data examples. While Baddeley et al. (2015) primarily discusses point pattern data analysis, they also cover areal data and spatial regression models. As a showcase example for regular lattice we demonstrate the use of the *bei* dataset from R-package *spatstat* (Baddeley et al., 2015), which contains the spatial coordinates of 3605 trees in a tropical rainforest. Typically, two regions are considered as neighbors if they share a common boundary, which may consist of a single point (queen adjacency) or a segment of at least some length (rook adjacency). Left panel of Figure 4 provides a visual representation of this concept for the present dataset. In case of regular lattice data, Simpson et al. (2017) applied Bayesian hierarchical model using INLA to model spatiotemporal trends in groundwater quality in Ireland. Similar work by Shaddick et al. (2018) analyze regular lattice data of air pollution levels in the UK.

On the other hand, Gelfand (2012) provide a comprehensive overview of spatial statistics, with a focus on irregular lattice data analysis. They demonstrate the use of INLA for modeling complex spatial structures, including non-stationary and anisotropic effects. The authors emphasize the importance of model selection and validation in spatial modeling and provide a framework for evaluating model performance. Martinez-Beneito et al. (2017) demonstrated the flexibility and

computational efficiency of INLA for handling irregular lattice data to model spatiotemporal variation in mortality rates in Valencia, Spain. To illustrate the concept of irregular lattice, we use the administrative boundaries of the *boston* dataset, which is available in R-package *spData* and contains housing values recorded for census tracts in Boston (Harrison and Rubinfeld, 1978). By default, a binary adjacency matrix is generated, such that two regions are considered neighbors only if they share at least one point along their common boundary (queen adjacency). Right panel of Figure 4 displays the census tracts in the boston dataset and the corresponding adjacency matrix.

In this context, it is worthy to mention about Besag-York-Mollié (BYM) model (Besag et al., 1991). It is a widely used spatial statistical model in areal spatial modeling. It is commonly used to model the spatial dependence in disease mapping, ecological modeling, and other spatial data applications.



Source: Blangiardo and Cameletti (2015)

**Figure 4: Regular and irregular lattice queen versus rook adjacency (left), irregular lattice: census tracts in the *boston* dataset and corresponding adjacency matrix (right)**

The BYM model assumes that the observed variable of interest in each areal unit is a combination of two components: a spatially structured component, which captures the spatial dependence among the units, and a random noise component, which accounts for the unexplained variation. The spatial structure is modeled using a conditional autoregressive (CAR) prior, which assumes that the observed variable in each unit is dependent on the values of the same variable in neighboring units.

Mathematically,

$$Y = X\beta + Zu + \varepsilon$$

where  $Y$  is a vector of the observed variable of interest,  $X$  is a design matrix of fixed effects,  $\beta$  is a vector of fixed effect coefficients,  $Z$  is a matrix that links the spatial structure of the data to the latent variable  $u$ , and  $\varepsilon$  is a vector of independent and identically distributed errors.

INLA provides an efficient and accurate method for Bayesian inference in the BYM model, by approximating the posterior distribution of the latent variable  $u$  using a combination of Gaussian quadrature and Laplace approximation techniques (Bakka et al., 2019). This allows for fast and accurate estimation of the model parameters, including the spatial dependence parameter and the

overall variance. BYM model can be used in regular lattice data as well as in irregular areal data. In regular lattice data, the lattice structure can be explicitly incorporated into the model by specifying a neighborhood structure among the lattice cells. For example, in a two-dimensional regular lattice, the neighborhood structure can be defined based on the spatial proximity of the cells, such that each cell is connected to its immediate neighbors (e.g., the four adjacent cells in a square lattice or the six adjacent cells in a hexagonal lattice) as shown in [Figure 4](#) (left). This neighborhood structure can be used to specify the spatial dependence structure of the data in the BYM model, by specifying a CAR prior on the latent variables. The BYM model with a CAR prior can be fitted to regular lattice data using the INLA method, which provides fast and accurate estimation of the model parameters ([Riebler et al., 2016](#)). The resulting estimates can be used to quantify the spatial dependence in the data and to make predictions at unsampled locations within the lattice ([Bivand et al., 2015](#)). Analyzing data on an irregular lattice presents additional challenges compared to analyzing data on a regular lattice, as the relationships between neighboring locations are not as straightforward ([Bivand et al., 2017](#)). Several studies in irregular lattice region using BYM are reported by [Blangiardo and Cameletti, \(2015\)](#) and [Moraga, \(2019\)](#).

Spatial analyses of aggregated data can be subject to two main problems: the *misaligned data problem* (MIDP) and the *modifiable areal unit problem* (MAUP). MIDP occurs when spatial data are analyzed at a different scale than that at which they were originally collected, leading to potential issues in the accuracy of spatial distribution and relationships among variables ([Banerjee et al., 2014](#)). MAUP refers to the variability in conclusions that can arise when the same underlying data is aggregated to different spatial scales or formations of areas, leading to the aggregation and zoning effects ([Openshaw, 1984](#)). Ecological studies, which are often based on aggregated data, can be particularly vulnerable to the ecological inference problem, which can be viewed as a special case of MAUP. Ecological bias, which can be caused by both the aggregation and specification effects, can lead to inaccurate conclusions when analyzing variables at the aggregated level ([Gotway and Young, 2002](#)).

## 4.7.2 Spatiotemporal Modeling

It is possible to expand the idea of spatial process to incorporate the dimension of time, resulting in a spatio-temporal framework.

Thus, a continuously indexed spatial process (random field) changing in time is denoted by

$$Y(s, t) \equiv \{y(s, t), (s', t') \in D \subseteq \mathbb{R}^2 \times \mathbb{R}\}$$

and are observed at  $n$  spatial locations or areas and at  $T$  time points. These observations are utilized for inferring information about the process and making predictions at specific locations. Typically, a Gaussian field (GF) is utilized, which is entirely determined by its mean and spatio-temporal covariance function  $\text{Cov}(y(s, t), y(s', t')) = \sigma^2 C((s, t), (s', t'))$ , defined for each  $(s, t)$  and  $(s', t')$  in  $\mathbb{R}^2 \times \mathbb{R}$ . Additionally, the process exhibits second-order stationarity if its mean remains constant and the spatio-temporal covariance function depends on the locations and time points only through the spatial distance vector  $h = (s - s') \in \mathbb{R}^2$  and the temporal lag  $l = (t - t') \in \mathbb{R}$  ([Gelfand et al. 2010](#); [Blangiardo and Cameletti, 2015](#)). Although a GF can be easily defined based on its first and second moments, its practical implementation is hindered by the

*big n problem* (Banerjee et al. 2008), particularly when dealing with large spatiotemporal datasets. This issue relates to the significant computational costs involved in performing linear algebra operations required for model fitting, spatial interpolation, and prediction. These computations necessitate the manipulation of dense covariance matrices, which are constructed using the spatio-temporal covariance function  $\sigma^2 C(\cdot, \cdot)$  and have dimensions equivalent to the number of observations across all spatial locations and time points (Balangiardo and Cameletti, 2015).

In the context of spatio-temporal geostatistical data (Gelfand et al., 2010), a valid spatio-temporal covariance function is needed, which can be represented as  $\text{Cov}(\theta_{it}, \theta_{ju}) = \sigma_c^2 (s_i, s_j; t, u)$  where stationarity is assumed in space and time. This means that the space-time covariance function can be expressed as a function of the spatial Euclidean distance  $\Delta_{ij}$  and the temporal lag  $\Lambda_{tu} = |t - u|$ . Valid non-separable space-time covariance functions are available in Cressie and Huang (1999) and Gneiting (2002). However, non-separable models can be computationally complex, and for this reason, some simplifications are made in practice. One approach is to assume separability, where the space-time covariance function is the sum or product of purely spatial and purely temporal terms, such as  $\text{Cov}(\theta_{it}, \theta_{ju}) = \sigma_c^2 C_1(\Delta_{ij}) C_2(\Lambda_{tu})$ , as described in Gneiting et al. (2006). Another approach is to assume constant spatial correlation in time, leading to a space-time covariance function that is purely spatial when  $t = u$ , i.e.  $\text{Cov}(\theta_{it}, \theta_{ju}) = \sigma_c^2 C \Delta_{ij}$ , and zero otherwise. In this case, temporal evolution can be introduced by assuming that the spatial process evolves over time according to autoregressive dynamics, as described in Harvill (2010).

Similar approaches can be applied to area level data. The GMRF framework can be extended to include a precision matrix defined in terms of time, assuming a neighborhood structure. If a space-time interaction is included, the precision can be obtained through the Kronecker product of the precision matrices for the space and time effects interacting, as explained in Knorr-Held (2000). We report that, GMRFs are a useful class of statistical models for spatial and spatiotemporal data that exhibit spatial or temporal correlation. The basic idea behind GMRF in spatiotemporal modeling is to represent the random variables as a multivariate Gaussian distribution, where the mean and covariance matrix are defined based on the spatiotemporal structure of the data. The covariance matrix is typically sparse, reflecting the fact that variables that are further apart in space or time are less correlated. The spatiotemporal structure of the data is represented by a graph structure, which is used to define the covariance matrix. In the case of spatial data, the graph is typically represented by a lattice or network of neighboring locations. In the case of spatiotemporal data, the graph may include both spatial and temporal neighbors, and the conditional independence structure may depend on both spatial and temporal distances. The edges between nodes represent the dependencies between the variables at those locations. In the context of spatiotemporal modeling, the GMRF is used to model the spatial and temporal random effects in a way that allows for efficient computation of the likelihood and posterior distribution.

## 4.8 Stochastic Partial Differential Equation (SPDE) Approach for Geostatistical Data

The  $f^{(i)}$  terms in Equation 1 in many models are often modeled using Gaussian Markov Random Fields (GMRFs), which have a range of applications, such as modeling smooth effects, random effects, measurement errors, and temporal dependencies (Rue and Held, 2005). GMRF models also exist for spatial dependence, including areal data, where models such as the CAR or BYM models have been proposed (Besag et al., 1991). The SPDE approach can be used for continuous spatial dependence by creating an approximation of the Matérn covariance field through stochastic partial differential equations (Lindgren et al., 2011; Rue et al., 2009). GMRFs are Gaussian models that have Markov properties, which are related to the precision matrix's non-zero structure. If two elements of the field are conditionally independent given all others, the corresponding entry of the precision matrix is zero. In practice, choosing GMRF priors for  $f^{(i)}$ , induces sparsity in the precision matrix  $Q(\theta)$  (Rue and Held, 2005).

Lindgren et al. (2011) proposed the SPDE method, which involves using a discretely indexed spatial random process (i.e., a GMRF) to represent a continuous spatial process (i.e., a GF). This method is based on the use of a linear fractional stochastic partial differential equation (SPDE)

$$(k^2 - \Delta)^{\alpha/2}(\tau\xi(s)) = W(s)$$

where  $s \in R^d$ ,  $k > 0$  is the scale parameter,  $\Delta$  is the Laplacian,  $\alpha$  controls the smoothness,  $\tau$  controls the variance and  $W(s)$  is a Gaussian spatial white noise process. The exact and stationary solution to this SPDE is the stationary GF  $\xi(s)$  with Matérn covariance function given by

$$Cov(\xi(s_i), \xi(s_j)) = Cov(\xi_i, \xi_j) = \frac{\sigma^2}{\Gamma(\lambda) 2^{(\lambda-1)}} (\kappa \|s_i - s_j\|)^\lambda K_\lambda(\kappa \|s_i - s_j\|)$$

Here,  $\|s_i - s_j\|$  is the Euclidean distance between two generic locations  $s_i, s_j \in R^d$  and  $\sigma^2$  is the marginal variance. The modified Bessel function of second kind and order  $\lambda > 0$  is denoted by  $K_\lambda$  (Abramowitz and Stegun, 1972). It measures the degree of smoothness of the process and is usually kept fixed. On the other hand,  $k > 0$  is the scaling parameter related to range  $r$ . The range is defined as the distance at which the spatial correlation becomes very low, close to 0.1. Mathematically, range  $r$  is expressed as  $r = \frac{\sqrt{8\lambda}}{\kappa}$  (Lindgren et al., 2011). The relationship between the Matérn parameters and SPDE can be expressed mathematically using the smoothness parameter  $\lambda$  and the marginal variance  $\sigma^2$  as:

$$\lambda = \alpha - d/2$$

$$\sigma^2 = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\lambda}\tau^2}$$

As in this case,  $s \in R^2$  ( $d = 2$ ) then the equations modified to:

$$\lambda = \alpha - 1$$



$$\sigma^2 = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)\kappa^{2\lambda}\tau^2}$$

Corresponding to smoothness parameter  $\alpha = 2$  when  $\lambda = 1$  the range and the variance are represented by  $r = \frac{\sqrt{8}}{\kappa}$  and  $\sigma^2 = \frac{1}{(4\pi\kappa^2\tau^2)}$  respectively. The solution to SPDE represented by the stationary and isotropic Matérn GF  $\xi(s)$  can be approximated using the finite element method through a basis function representation defined on a triangulation of the domain  $D$ :

$$\xi(s) = \sum_{g=1}^G \varphi_g(s) \tilde{\xi}_g \quad \dots\dots\dots (3)$$

where  $\varphi_g$  is the set of basis functions,  $G$  is the total number of vertices of the triangulation and  $\tilde{\xi}_g$  are zero mean Gaussian distributed weights. To achieve a Markovian framework, localized basis functions are selected, which exhibit a piecewise linear nature within each triangular region. Specifically,  $\varphi_g$  assumes a value of 1 at vertex  $g$  and 0 at all other vertices. For  $\alpha = 2$  using Neumann boundary conditions the precision matrix  $Q$  for Gaussian weight vector  $\tilde{\xi} = \{\tilde{\xi}_1, \dots, \tilde{\xi}_G\}$  is expressed as:

$$Q = \tau^2(\kappa^4 C + 2\kappa^2 G + G(C)^{-1}G)$$

Here,  $C$  and  $G$  are the diagonal and sparse matrix respectively. The generic element of  $C$  is represented as  $C_{ii} = \int \varphi_i(s)ds$  and for sparse matrix  $G$  is  $G_{ij} = \int \nabla\varphi_i(s)\nabla\varphi_j(s)ds$  where  $\nabla$  is the gradient. It is worthy mention that, the elements of the precision matrix  $Q$  depend on  $\tau$  and  $\kappa$ . Additionally,  $Q$  is sparse which makes  $\xi$  a GMRF with distribution  $N(0, Q^{-1})$  and it represents the approximated solution to the SPDE in a stochastically weak sense.

### ***Mesh Construction***

As mentioned in the beginning of this section, the SPDE approach is based on a triangulation of the spatial domain. The triangulation or the SPDE-mesh is an important component of the INLA-SPDE approach. Specifically, it is essential to establish a mesh across the study area, and it will be used to compute the approximation to the solution (i.e., the spatial process). In this method, the SPDE is used to model the spatial variation of the random field, and the triangular mesh is used to discretize the spatial domain (Krainski et al., 2018). The goal is to obtain a sparse linear system that can be solved to obtain the values of the random field over the mesh.

The process of triangulation involves dividing the spatial domain into a group of triangles that do not overlap with each other. Any two triangles meet in at most a common edge or corner. It is worthy to note that when defining the mesh, there is a balance to strike between the accuracy of the GMRF representation and computational expenses. Both factors rely on the number of vertices used in the triangulation. In other words, as the number of mesh triangles increases, the GF approximation becomes more precise but also incurs higher computational costs. Krainski et al. (2018) provide a comprehensive description of SPDE models using SPDE triangulation, in this context, we will provide an overview of the various steps necessary to fit these models.

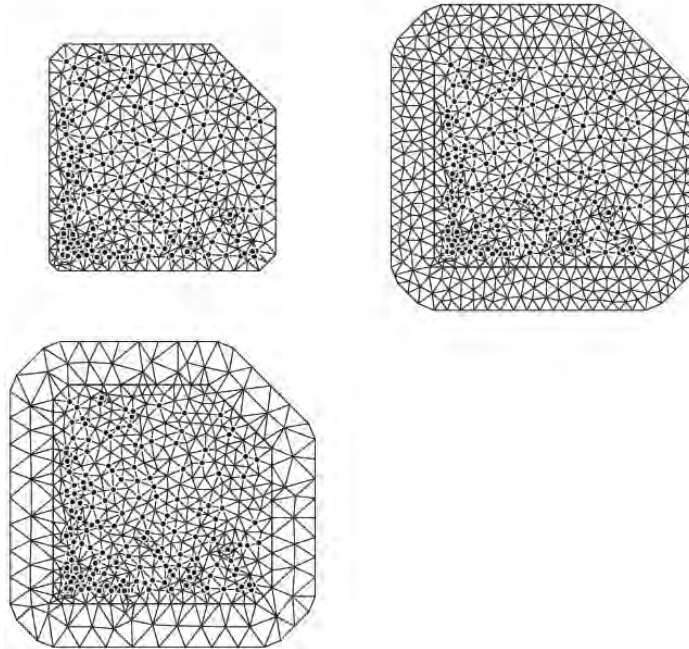
As the initial step, it is necessary to establish the limit or boundary of the study area. To accomplish this, the borders of the study region are estimated from the spatial shapefile of the

region or, individual pixels of the dataset will be merged to form a unified region, which will serve as a reasonable estimate of the boundary. Next, the set of basis functions will be determined by defining a two-dimensional mesh using the function `inla.mesh.2d()` from the R-INLA package (R-INLA Project, 2020).

```

coords <- -as.matrix(SPDEtoy[, 1:2])
mesh0 <- -inla.mesh.2d(loc = coords, max.edge = 0.1)
mesh1 <- -inla.mesh.2d(loc = coords, max.edge = c(0.1,0.1))
mesh2 <- -inla.mesh.2d(loc = coords, max.edge = c(0.1,0.2))

```



Source: [Blangiardo and Cameletti \(2015\)](#)

**Figure 5: Three triangulations for the SPDEtoy dataset  
Mesh0 (top left), Mesh1 (top right), and Mesh2 (bottom)**

The function requires input related to the spatial domain, which can be specified either by relevant spatial points (not necessarily where observations are available) or the domain extent, using the `loc` or `loc.domain` arguments, respectively. It is not compulsory to the locations of observations into the mesh. Instead, the `loc.domain` option allows for the provision of a point location matrix to specify the domain extension. The `max.edge` argument is also mandatory and represents the maximum allowable length of the triangle edge. When a vector of two values is provided, the spatial domain is divided into an inner and an outer area, and the triangle resolution for each is specified by `max.edge`. Higher values of `max.edge` result in lower resolution but greater accuracy. This domain extension can help avoid boundary effects associated with the SPDE approach, which includes an increase in variance near the boundary due to the Neumann boundary conditions used in R-INLA. [Lindgren and Rue \(2015\)](#) suggest extending the domain by a distance at least equal to the range  $r$  to avoid boundary effects. For example, `SPDEtoy` from R-INLA dataset ([Rue et al., 2009](#)) is triangulated using three different sets of parameters to create

three distinct meshes. The coordinates of the data locations are included as vertices in all three triangulations using the `loc` parameter, while different values for `max.edge` are used.

Figure 5 displays the resulting meshes. Black points denote the observation locations. Mesh0, which specifies only one `max.edge` value, does not exhibit an outer extension of the domain. A thick line separates the outer offset from the inner offset and the boundary of the study region. Mesh1 and Mesh2, on the other hand, both extend outside of the original domain. Mesh2 has a larger `max.edge` value, resulting in larger triangles and reduced accuracy. This approach allows for the extension of the original spatial domain to avoid boundary effects without significantly increasing computational costs. The `inla.mesh.2d` function provides an optional `offset` argument, which allows for the specification of the degree to which the domain is to be extended in the inner and outer regions. Another optional argument is `cutoff`, which prevents the creation of excessive small triangles in the vicinity of clustered data locations. By default, cut off value is set to 0. Thus, the process to discretize the spatial domain  $D$  into a mesh of triangles is achieved. The vertices of the triangles are referred to as nodes, and the edges of the triangles are referred as edges.

The next step is to build the projector matrix. Following basis function Equation 3 the linear predictor ( $\eta_i$ ) can be expressed as:

$$\eta_i = b_0 + \sum_{g=1}^G \varphi_g(s_i) \tilde{\xi}_g$$

where,  $b_0$  is the intercept,  $\varphi_g(s_i)$  is the value of the  $g$ -th basis function at  $s_i$ . This equation can be further generalized, and the linear predictor can be mathematically represented as:

$$\eta_i = b_0 + \sum_{g=1}^G A_{ig} \tilde{\xi}_g$$

Here,  $A_{ig} = \varphi_g(s_i)$  is the generic element of the sparse matrix ( $A$ ) which maps the GMRF  $\tilde{\xi}$  from the  $G$  mesh vertices to the  $n$  observed locations. The R-INLA function `inla.spde.make.A` creates the sparse weight matrix  $A$  by identifying the data locations in the mesh and organizing the corresponding values of the basis functions (Blangiardo and Cameletti, 2015). The dimension of the resulting matrix  $A$  can be determined by multiplying the number of data locations by the number of mesh nodes. It is important to highlight that the meshes created by including observation locations as mesh vertices, the projector matrix in this case has one non-zero value (equal to 1) for each row. On the other hand, each spatial location is placed within a triangle defined by three vertices, resulting in a projector matrix characterized by three non-zero elements for each row, with a sum equal to 1 (i.e.,  $\sum_{g=1}^G A_{ig} = 1$ ). Additionally, there are some columns with zero values corresponding to vertices not connected to points. All these simplify the computation of the estimates of the random effect at any given point because its estimate will be a linear combination of only three functions in the basis (Krainski et al., 2018). Thus, in this step `inla.spde.make.A` function is used to create the projector matrix  $A$  to map the projection of the SPDE to the observed points.

The final important step establishes the framework to estimate the model parameters using the triangulation and the projector matrix that have already been created and choosing the default prior specification for SPDE parameters provided by R-INLA package. First of all, a Matérn SPDE object is created using `inla.spde2.matern` function and another function `inla.spde.make.index` generates a necessary list of named index vectors for the SPDE model.

All these steps define how the solution to the SPDE that defined the spatial process with a Matérn covariance is computed using R-INLA package. The resulting solution will be used to define the spatial random effect using the function in the INLA model `formula`. When using a SPDE, the data passed to `inla()` must be in a particular format, which is provided by the `inla.stack` function. This function organizes data in a list format with named elements including a list of data vectors, a list of projector matrices, a list of effects (e.g., the SPDE index) or predictors, and a label for the data group usually denoted by `tag` (Krainski et al., 2018).

This is in brief the technical procedure to fit a model with an intercept and covariate(s) along with a spatial random effect defined with an SPDE object. However, since we are studying a continuous spatial process, it is important to estimate the variable of interest and the underlying spatial effect in the study region. To achieve this, we can define another set of data for prediction. The standard approach for making predictions is to add fake data rows containing the covariates and location we wish to predict, but with Not Available (*NA*) in place of response variable values (Krainski et al., 2018; Gomez-Rubio, 2020). Additionally, a different projector matrix will be necessary to map the estimates of the spatial process to the prediction points. We can utilize the function `inla.spde.make.A`, similar to before, by utilizing the points on the grid. In this case a new distinct tag should be assigned to enable identification of this specific portion of the data within the stack, facilitating the retrieval of the fitted values and other relevant quantities at the grid points. Subsequently, the two data stacks can be integrated into a unified object using `inla.stack` function once again. This resultant object will then be utilized as the input for the `inla` function during model fitting. To specify the model that will be fitted, the `f()` function must be utilized to incorporate spatial effects. The index that was previously created with `inla.spde.make.index`, will be passed as an argument to this function. Additionally, the model will be of the `spde` type, requiring the combination of the data passed to the function `inla()` with SPDE definition for spatial model fitting. This can be accomplished by means of the `inla.stack.data` function. Moreover, `control.predictor` argument will necessitate the usage of `inla.stack.A` to combine the projector matrix for the entire dataset (model fitting and prediction) in the combined stack (Krainski et al., 2018; Gomez-Rubio, 2020).

Finally, the posterior summaries of spatial parameters can be obtained by using the function `inla.spde2.result`, which efficiently extracts relevant information from the output list. This function further facilitates the transformation of internal parameter scales, offering posterior distributions for nominal variance ( $\sigma^2$ ) and nominal range ( $r$ ), along with internal results for  $\theta_1 = \log(\tau)$  and  $\theta_2 = \log(\kappa)$  (Krainski et al., 2018). It is noteworthy to indicate that the use of the `inlabru` package (Bachl et al. 2019) can streamline the process of defining and fitting the model. Additionally, the interactive function `meshbuilder` within the INLA package can aid in defining and evaluating the suitability of a mesh (Krainski et al., 2018).

### 4.8.1 Application of INLA-SPDE in Spatial and Spatiotemporal Modeling

The INLA-SPDE methodology has proven to be a powerful tool in statistical modeling due to its many advantages over other techniques. It provides low computation time, making it an attractive option for large datasets. Additionally, as the basic logic is Bayesian inference, it does not require only normally distributed data, enabling its application in a vast domain of fields. The methodology also allows for the implementation of both spatial and temporal effects, as well as the analysis of their significance in the model. INLA-SPDE permits the integration of a substantially high number of covariates and can also accommodate new covariates at a later stage of the process. Moreover, the level of significance for each covariate can be analyzed, which further strengthens its utility in statistical modeling.

Thus, INLA-SPDE combines the benefits of two powerful techniques: INLA, a fast and accurate Bayesian inference method, and SPDE, a flexible and scalable method for spatial modeling (Lindgren et al., 2011; Blangiardo et al., 2013; Blangiardo and Cameletti, 2015; Lindgren and Rue, 2015; Rue et al. 2017; Moraga 2019; Varga et al., 2019; Gomez-Rubio, 2020; Verdoy 2020). INLA-SPDE is particularly useful for spatiotemporal modeling in a range of fields, including meteorology, ecology and environmental science, epidemiology and public health, urban issues like, crime analysis, traffic management, traffic accidents, air pollution and its impact, disaster prevention and management, among others. A study by Bakka et al. 2018 review some recent publications that showcase the versatility and effectiveness of INLA-SPDE in these domains. In the study they highlight about the use of spatial models applied in several high-impact studies. For example, Jousimo et al. (2014) studied the effects of fragmentation on infectious disease dynamics, and Bhatt et al. (2015) assessed the effectiveness of malaria control efforts in Africa, while Golding et al. (2017) modeled mortality rates across various age groups in multiple countries. Additionally, Shaddick et al. (2018) estimated global exposure to PM<sub>2.5</sub>, which was utilized in the Global Burden of Disease study (Gakidou et al., 2017) and the World Health Organization's evaluation of health risks associated with air pollution (World Health Organization, 2016). Huang et al. (2017) compared spatial models in R-INLA with *REML-LMM* to perform environmental mapping of soil. Etxeberria et al. (2017) modeled pancreatic cancer mortality in Spain using a spatial gender-age-period-cohort model. In addition, Pereira et al. (2017) developed a spatial model of unemployment. Mejia et al. (2020) computed probabilistic activation regions in cortical surface fMRI data. Similar studies highlight the versatility and utility of spatial modeling in ecological and environmental research (Juan et al., 2012; Serra et al., 2014b; Rutten et al., 2017; Moraga et al., 2017; Gortázar et al., 2017; López-Abente et al., 2018; Barceló et al., 2021; Niekerk et al., 2021; Wright et al., 2021; Saez and Barceló, 2022). Other examples depict the applications in spatial econometrics (Bivand, Gómez-Rubio, and Rue, 2014; Gómez-Rubio, Bivand, and Rue, 2014; Gómez-Rubio, Bivand, and Rue, 2015). In their recent paper, Lindgren et al. (2022) present a comprehensive compilation of publications that employ the INLA-SPDE approach for modeling Gaussian and non-Gaussian fields in diverse fields, such as health, engineering, environmetrics, econometrics, urban planning, pollution, and several others. The paper sheds light on the wide range of applications of SPDEs in various disciplines and highlights their potential as a versatile tool for spatial statistical inference and prediction.

## 4.9 Challenges in Traditional SPDE Triangulation Approach

### 4.9.1 Complex Distributed Spatial Regions

To gain a deeper understanding of the issues and challenges in this domain and to determine whether they are unique to INLA-SPDE, we conducted a preliminary investigation using geospatial data without utilizing the INLA-SPDE approach. Specifically, we utilized spatial interpolation techniques such as kriging to analyze the complex island structure of the Maldives and explore the precise climate patterns within the nation (Chaudhuri et al., 2021a).

It is worthy to note that, from the left panel of Figure 6 it seems that atolls with enclosed lagoon and land on reefs are continuous land structure. But each atoll is originally a collection of number of distributed islands as depicted in the same figure right panel. The complex spatial structure of the study region motivated us to conduct research in this area, providing a unique opportunity to explore the performance of other spatial statistical techniques, such as kriging.



Source: Chaudhuri et al. (2021)

**Figure 6: Example of complex spatial region  
atoll with both lagoon and reef areas (left) with only land on reefs areas (right)**

The methodology used in the study involved an exploratory data analysis followed by geostatistical techniques of kriging to estimate and predict the spatial variability of meteorological variables throughout the nation. The study uses a set of observations of a spatial attribute to predict values for other locations using a linear regression estimate. The error variance in this expression is minimized under the constraint of unbiasedness. The study utilizes the geostatistical technique of kriging by employing variograms to describe spatial variation in terms of size and general shape. The best variogram model for the data used in the study is found to be the spherical model (Nouck, 2019), which is defined by a range, prior variance, and nugget effect. The study highlights the limited literature on the application of geostatistical tools like kriging in complex island structures and suggests that an increase in the number of meteorological stations could improve the kriging performance and help in precise prediction. Several studies show the application of kriging in complex coastal regions or, in dispersed islands (Irl t al., 2015; McKenzie et al., 2021).

From a scientific perspective, our interest lies in investigating how similar research studies utilizing INLA-SPDE have been carried out to model complex land structures in coastal regions and islands. Current literature shows that even for complex and distributed spatial regions, researchers have utilized a traditional continuous region concept to design the SPDE triangulation. This approach involves generating an SPDE mesh for the entire study region, despite the presence of physical barriers that make the study area complex and distributed. For example, [Lezama-Ochoa et al. \(2020\)](#) used this approach to predict the occurrence of spinytail devil ray species in the eastern Pacific Ocean. [Bi et al. \(2021\)](#) conducted a similar study to estimate seabird bycatch variations in the mid-Atlantic bight and northeast coast, and [Cosandey-Godín et al. \(2014\)](#) applied this approach to analyze spatiotemporal patterns of accidental bycatch in fisheries located in the Baffin Bay of the Atlantic Ocean.

In our review of the literature, we have found several studies that have used the same approach to model complex land structures. For example, [Crespo et al. \(2019\)](#) studied flood protection ecosystem services in the coast of Puerto Rico, [Myer et al. \(2017\)](#) used a spatiotemporal model to examine the ecological and sociological factors that predict the presence of West Nile virus in mosquitoes in Suffolk County, New York, [Paradinas et al. \(2015\)](#) employed a spatiotemporal approach to validate persistence areas and identify fish nurseries in the western Mediterranean Sea, and [Silva et al. \(2011\)](#) estimated the potential distribution of invasive and native trees in the Azores islands, Portugal. We aim to build upon these studies and further explore the application of INLA-SPDE in complex land structures, particularly in coastal regions and islands.

Another serious concern to model observations in complex island structures is the anomaly related to the polygon structure of the coastlines. Coastlines are often considered as fractal structure, in the sense that any finite approximation will not be accurate ([Bakka et al., 2019](#)). For the same coastline polygons, different researchers may use varying approximations which can lead to conflicting interpretations and predictions. In that case, the model loses its scientific credibility. It is worthy to mention that a stationary model cannot be aware of the coastline structure and will inappropriately smooth over the features. In spatial modeling, classical models become unrealistic when they fail to account for holes or physical barriers in the landscape. This can lead to further unrealistic assumptions.

## 4.9.2 Modeling on Linear Networks

Recent research has highlighted the growing trend of using point pattern techniques to model spatiotemporal events on linear networks. In this regard, to investigate the impact of spatiotemporal modeling on linear networks we conducted a study titled *On the Trend Detection of Time-Ordered Intensity Images of Point Processes on Linear Networks* ([Chaudhuri et al., 2021b](#)). Our study focused on the application of spatial point processes on linear networks to analyze time-ordered point patterns in traffic accidents and street crime analysis. To identify potential monotonic trends, we used the Mann-Kendall trend test ([Mann 1945; Kendall 1948](#)) and applied the analysis to monthly time-ordered point patterns of fatal traffic accidents and street crimes in London from January 2013 to December 2017.

However, when modeling random spatial events like crime or traffic accidents, it is a common practice to aggregate data for data protection and security reasons. Aggregation can be performed

in both spatial and temporal dimensions, but it is not necessarily related to statistical or numerical methodologies (Miaou et al., 1992; Abdel-Aty et al., 2000; Khattak et al., 2021). A common approach is to divide the spatiotemporal region into regular grids and count the number of events in each subregion. In the case of events on linear networks, such as road networks, analyzing individual road segments is a common practice. We found it interesting and more real-time applicable to model spatial events as aggregated values (using Poisson, binomial, or binary logistic models) on linear networks. This approach is particularly important for predictive modeling, where authorities and policy makers may need to identify risk factors related to road safety and criminal activities by predicting the exact number of incidences on a particular street or individual segments of a street (Weisburd et al., 2009; Hadayeghi et al., 2010; Santhosh et al., 2020). Count data has been widely used to estimate risk factors in such cases.

Based on our analysis of spatiotemporal events such as traffic accidents, street crimes, and issues in water and electric connection networks in cities that occur exclusively on linear networks, it has been observed that conventional INLA-SPDE techniques are frequently used to model these events, despite the fact that they are strictly confined to linear networks. When applying the INLA-SPDE method to linear networks, creating a triangulation for the entire region enables fitting of the INLA model in the study area. However, a significant problem arises while predicting events, as the observed events are discrete spatial points located precisely on the road network, whereas models fitted with a region mesh cover the entire study area. This implies that the locations of predicted events can be placed in any area with or without road networks, which is not realistic. Traditional methods of model prediction using a region mesh are, therefore, not appropriate in this context from a scientific perspective.

To address this issue, we propose the use of alternative modeling techniques, such as network-based spatial statistical models, that are specifically designed to analyze events on linear networks. These models take into account the unique characteristics of linear networks and are capable of accurately predicting events at discrete spatial points located precisely on the road network. By incorporating the spatial structure of the linear network into the modeling process, these models can provide more realistic and accurate predictions of events on the network.

Designing the INLA-SPDE triangulation specifically on road networks can offer several advantages for modeling random spatial events like crime or traffic accidents. One of the primary advantages of this approach is improved computational efficiency. By focusing the triangulation specifically on the road network, researchers can limit the number of nodes and edges that need to be modeled, reducing the computational burden associated with modeling data over an entire study area. Another advantage of designing the triangulation specifically on road networks is the potential for more precise modeling of spatial relationships along the road network. In contexts like traffic accidents or crime hotspots, spatial relationships along the road network are often critical for understanding patterns and trends in the data. By designing the triangulation specifically on the road network, researchers can more accurately capture these relationships, potentially leading to more accurate and informative models. Additionally, designing the triangulation specifically on road networks can help to reduce bias and improve the accuracy of the resulting models. By focusing on the road network, researchers can more effectively control for confounding variables that may be present in other areas of the study region, leading to more precise and accurate estimates of the relationships between variables.



While there are several potential advantages to design the INLA-SPDE triangulation specifically on road networks for modeling random spatial events like crime or traffic accidents, it is important to carefully consider the potential limitations and trade-offs associated with this approach.

## 4.10 Proposed Methodologies

In this section we discuss about the potential solutions that we have proposed and are presently developing to address the challenges to model spatiotemporal events on linear networks and on complex spatial structures like coastal areas or, islands. These solutions are grounded in scientific principles and methods and are designed to offer effective and robust approaches to address the identified issues. Additionally, we will analyze the related limitations and trade-offs of the suggested methodologies. This evaluation will enable us to gain a more comprehensive understanding of the strengths and weaknesses of each approach and make informed decisions about their suitability for different research applications.

In all the projects of the thesis, R programming language (version R 4.0.4 to R 4.2.2) has been used for statistical computing and graphical analysis. All computations are conducted on a quad-core Intel i7-4790 (3.60 GHz) processor with 32GB (DDR3-1600) RAM.

### 4.10.1 Network Triangulation

We introduced the novel concept of designing SPDE triangulation precisely on road networks. Our process involves several steps: first, we create a buffer region for each road segment, ensuring that the width of the buffer is selected to include the maximum number of points within a standard buffer area for all road segments. Next, we construct a clipped buffer polygon comprising only the area covered by the road network, and then apply SPDE triangulation on the clipped polygon to construct the SPDE Network Mesh.

We examined traffic accident locations on a road network in an urban environment and noted that many events were located away from road segments. Left panel of [Figure 7](#) depicts a sample of traffic accident locations (marked as red points) on a sample road network. We note that many events are located away from the road segments.



Source: [Chaudhuri et al. \(2022b\)](#)

**Figure 7: Traffic accident locations on buffered road segments without buffer (left), with buffer (right)**



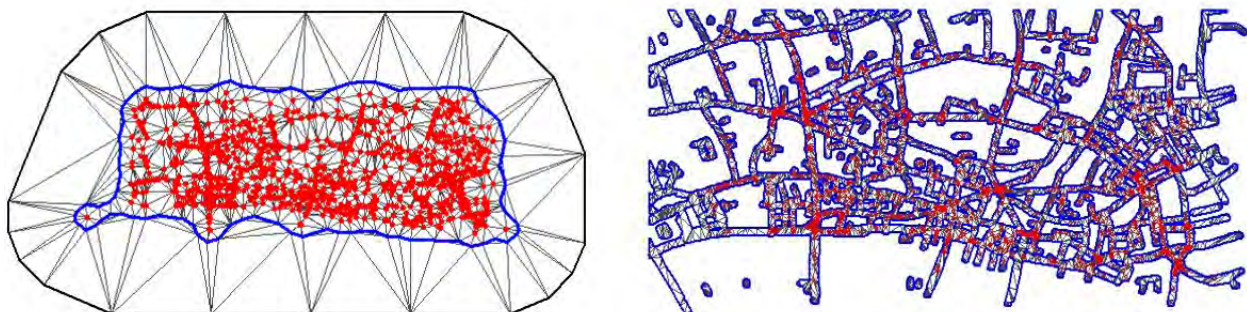
Source: Chaudhuri et al. (2022b)

**Figure 8: Clipped polygon of buffered road segments**

Initially, the buffer width is selected in such a manner to get maximum points within a standard buffer area for all road segments. We found that a common buffer width for all road segments was the most effective in achieving this. In Figure 7 (right panel) we show the built buffers on the same road network with 20 meters. We merged individual buffer segments into a single polygon clipped within a bounding box covering the study area.

Figure 8 illustrates the clipped polygon of the buffered segments in grey. According to Verdoy (2021), the best fitting mesh should have enough vertices for effective prediction, but the number should be within a limit to have control over computational time. With this concept we have fine-tuned the *inla.mesh.2d* function in R-INLA to control the largest allowed triangle edge length (*max.edge*) and minimum allowed distance between points (*cutoff*), regulating the number of vertices in the SPDE mesh to identify the best fitted mesh.

Figure 9 in the left panel depicts the SPDE triangulation for the entire study area while the right panel depicts the proposed network mesh, in both cases the accident events are highlighted in red as depicted in Chaudhuri et al., 2022b.



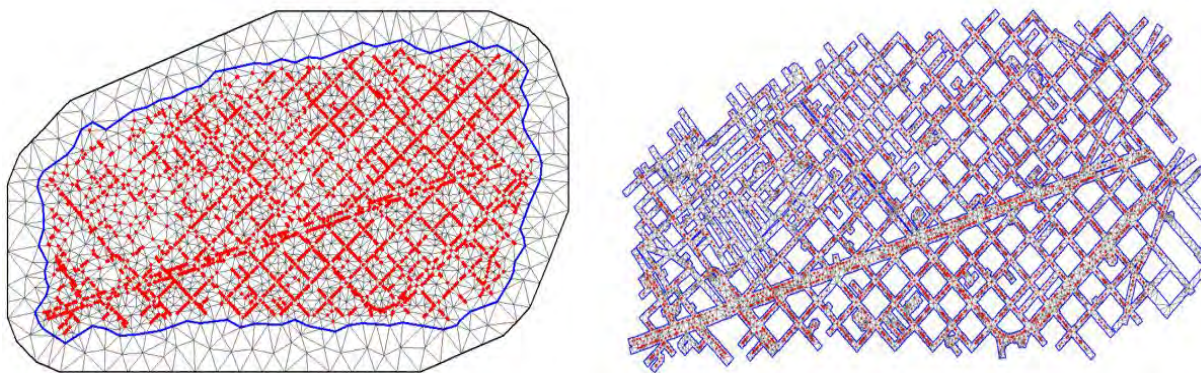
Source: Chaudhuri et al. (2022b)

**Figure 9: SPDE triangulation for entire study area and network mesh (London, UK)**

After aggregating and counting accident events within the buffer area of each road segment, we use the centroid of each segment as initial triangulation nodes applied on the clipped polygon. This approach allows us to analyze accident risk factors in each road segment and generate risk maps that provide information on safe routes between source and destination points. Our

approach, introduced in the publication titled *Spatio-Temporal Modeling of Traffic Accidents Incidence on Urban Road Networks Based on An Explicit Network Triangulation*, is a novelty in estimating the spatial autocorrelation of traffic accidents restricted to the linear network (Chaudhuri et al., 2022b). The resulting risk maps can be useful for accident prevention and multi-disciplinary road safety measures.

We conducted similar research work titled *Spatiotemporal Modeling of Traffic Risk Mapping: A Study of Urban Road Networks in Barcelona, Spain* that analyzed ten years of traffic accident data to investigate spatial and temporal variation in accidents and related injuries (Chaudhuri et al., 2023). Our proposed spatiotemporal model enabled us to predict the number of injuries that might occur on individual road segments. Figure 10 in the left panel depicts the SPDE triangulation for the entire study area while the right panel depicts the proposed network mesh, in both cases the accident events are highlighted in red as depicted in Chaudhuri et al., 2023.



Source: Chaudhuri et al. (2023)

**Figure 10: SPDE triangulation for entire study area and network mesh (Barcelona, Spain)**

To generate a predicted risk map for the entire road network, we used Bayesian methodology with INLA and SPDE. In contrast to traditional SPDE triangulations on entire region, our study applied the INLA-SPDE modeling approach to selected areas, specifically on road networks. The resulting risk maps can serve as a baseline for identifying safe routes within a spatiotemporal context. Moreover, this methodology can be adapted and applied to enhanced INLA-SPDE modeling precisely on road networks. The novelty of our study lies in the introduction of SPDE network triangulation to estimate the spatial auto-correlation of discrete events. By doing so, we took a new step in INLA-SPDE modeling to perform spatiotemporal predictive analysis only on selected areas (in this case, road networks). Our study contributes to the relatively small amount of literature on spatiotemporal analysis using INLA-SPDE of spatial events precisely on road networks. The methodology is dynamic and can be adapted and applied to other locations globally.

However, while using network triangulation for modeling spatial relationships can be effective in capturing relationships along the road network, there are some potential limitations and trade-offs that should be considered. For instance, the approach may not be able to capture important spatial relationships outside of the road network, particularly in contexts where traffic accidents can occur in adjacent areas or neighborhoods. Furthermore, accurately modeling spatial relationships along the road network can be challenging in areas with complex road networks or where the network is subject to frequent changes or updates, which may require frequent updates

to the triangulation to accurately capture changes in the road network. Another significant limitation we observed in the proposed methodology is the boundary effect, which can lead to biased estimates and prediction errors near the boundary if the mesh does not cover the entire domain. In the following section, we provide a brief discussion of the potential causes and effects of the observed boundary effects in our two publications proposing the novel concept of SPDE network triangulation.

### ***Discussion on Boundary Effects***

Spatial Gaussian fields (SGFs) are commonly utilized as model components in the construction of spatial or spatio-temporal models for various applications, including the Generalized Additive Model (GAM) framework, to represent the residual spatial structure resulting from unmeasured spatial covariates, spatial aggregation, and spatial noise. The use of a buffer road network in current studies adds complexity to the boundary regions, which can influence the spatial effect of the model. [Krainski et al. \(2018\)](#) propose creating a mesh to represent the spatial process as the first step in fitting a SPDE model. Building an SPDE mesh for a continuous region is relatively straightforward, but the creation of an SPDE network mesh requires fine-tuning to identify the best fit values for minimum allowed distance between vertices and maximum permissible triangle edge length for the inner (and outer) regions. Careful selection of additional points around the boundary or outer extension is also necessary. As a general rule, the variance near the boundary is inflated by a factor of two along straight boundaries and by a factor of four near right-angled corners ([Lindgren and Rue, 2015](#)). The complex boundary region of the buffer road network with several right-angled corners makes the process critical. The boundaries in the proposed mesh are located inside the mesh and not outside, as in a standard mesh, which creates fictitious spatial structures. Due to the complex boundary nature, it is necessary to reduce the high boundary effect that may cause a variance twice or four times as great at the border as within the domain ([Lindgren et al., 2011](#); [Lindgren and Rue, 2015](#)). Although the residual diagnostics and predicted risk maps produced by the model match the original observed records, the correlation values of the model indicate the need for improvement.

In addition, we recommend that researchers in this field carefully consider the assumptions and limitations of various modeling techniques before selecting the most appropriate one for their specific research question. It is crucial to choose a method that is well-suited to the unique characteristics of the spatiotemporal events being studied and can provide dependable and accurate predictions of these events on the linear network. Furthermore, for a more detailed understanding of the model performance, it may be advantageous to further analyze the model fitting phase using INLA-SPDE with a diverse set of spatial and temporal covariates, spatial and temporal structures, and space-time interactions. Researchers should thoroughly assess the computational efficiency, precision, and accuracy of this approach against the potential limitations associated with capturing essential spatial relationships outside of the road network or accurately modeling spatial relationships along the network.

#### **4.10.2 Barrier Model**

Spatial models often assume isotropy and stationarity, implying that spatial dependence is direction invariant and uniform throughout the study area. However, these assumptions are

violated when dispersal barriers are present in the form of geographical features or disease control interventions.

In response to this problem, [Bakka et al. \(2019\)](#) introduced the *Barrier model* as a new approach for modeling complex spatial regions. Unlike existing methods that rely on the shortest distance around physical barriers or boundary conditions, the Barrier model is based on the Matérn correlation viewed as a collection of paths through a simultaneous autoregressive (SAR) model. By manipulating local dependencies, the model can effectively cut off paths that cross physical barriers. To ensure that the new SAR model is well-behaved, the researchers formulated it as a SPDE, which can be discretized to represent a Gaussian field with a sparse precision matrix that is always positive definite. One of the principal advantages of the barrier model is that the computational cost is the same as for the stationary models.

### 4.10.3 Application of Barrier Model in Disjoint Spatial Regions

In general, SPDE triangulations that assume stationarity and isotropy, where the autocorrelation between two locations depends only on their Euclidean distance. However, when modeling events on dispersed island structures, physical barriers such as coastlines, road networks, power lines, categorical health sectors, and different land uses can pose a problem. To handle the coastline problem, several studies have proposed solutions, such as computing the shortest distance in water ([Wang, 2007](#); [ScottHayward, 2014](#), [Miller, 2014](#)), defining boundary conditions using a smoothing penalty together with Neumann boundary condition ([Ramsay, 2002](#)), or using the Dirichlet boundary condition ([Wood et al., 2008](#); [Sangalli et al., 2013](#)). However, these methods may not be suitable for complex archipelago structures with physical barriers. To handle nonstationary and anisotropic spatial processes in such cases, [Bakka et al. \(2019\)](#) proposed a finite element method-based approach that uses a system of two SPDEs. The example of an archipelago on the south-west coast of Finland is used to motivate the need for non-stationary SGFs in cases with physical barriers. A system of two SPDEs is presented in this case, one for the barrier area, and the other for the remaining area. The solution to the system is a nonstationary spatial effect, denoted by  $u(s)$ . The stochastic differential equations for the system are:

$$u(s) - \nabla \cdot \frac{r_b^2}{8} \nabla u(s) = r_b \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_b$$

and

$$u(s) - \nabla \cdot \frac{r^2}{8} \nabla u(s) = r \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_n$$

where  $u(s)$  is the spatial effect,  $\Omega_b$  is the barrier area, and  $\Omega_n$  is the remaining area, and their disjoint union gives the whole study area  $\Omega$ . Variables  $r$  and  $r_b$  represent the ranges for the remaining and barrier areas, respectively,  $\sigma_u$  is the marginal standard deviation.  $\nabla$  is the gradient operator and is equal to  $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$ , and  $W(s)$  stands for white noise. The barrier model is based on viewing the Matérn correlation as a collection of paths through a SAR model, rather than as a correlation function on the shortest distance between two points ([Bakka et al., 2019](#)). The local dependencies are manipulated to cut off paths crossing the physical barriers. In the next step, the

new SAR model is formulated to SPDE format to represent the Gaussian field, with a sparse precision matrix that is automatically positive definite.

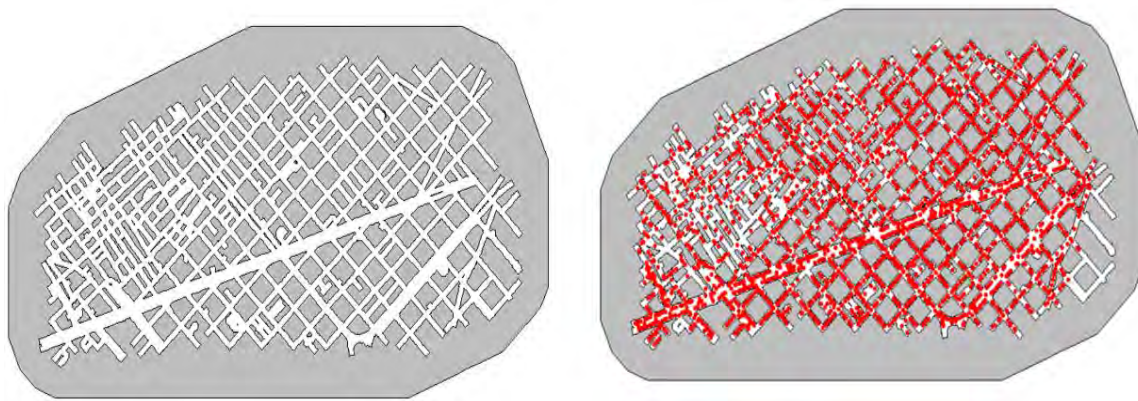
A recent scientific study conducted by [Cendoya et al. \(2021\)](#) has investigated the impact of different types of barriers on the spatial distribution of *Xylella fastidiosa*, a plant pathogenic bacterium, in a demarcated area in Alicante, Spain. The study utilizes occurrence data from official surveys conducted in 2018 and employed four spatial Bayesian hierarchical models. The first model represents a scenario without any control interventions or geographical features, whereas the second model incorporates mountains as physical barriers. The third and fourth models includes continuous and discontinuous perimeter interventions, respectively, as physical barriers surrounding the infested area. It is important to note that these barriers are assumed to be completely impermeable, implying that they do not allow infected vectors or propagating plant material to pass through. To perform inference and prediction, the study utilizes INLA-SPDE approach. Overall, the findings of the study sheds light on the effectiveness of barrier model in complex spatial regions having physical barriers.

We observed that classical stationary models in spatial statistics often assume isotropy and stationarity. It causes inappropriate smoothing over features having boundaries, holes, or physical barriers. Despite this, nonstationary models like barrier model have been little explored in the context of spatial and spatiotemporal modeling in complex spatial regions. We are currently working on a similar topic titled *A Nonstationary Approach with Barriers: Modeling Spatial Dependencies of Natural Hazards in Islands (under review)*. The principal objective of the current study is to evaluate the influence of barrier models compared to classical stationary models using tsunami data from the island nation of Maldives. For seven atolls across the nation, we have applied three distinct meshes, two stationary and one that corresponds the barrier concept. The results show that when assessing the spatial variance of tsunami incidence at the island scale, the barrier model outperforms the other two models. Moreover, it has the same computational cost as the stationary models, which facilitates to explore nonstationary spatial models in complex land structures. In the broader picture, this research work contributes to the relatively new field of barrier models as well as to initiate and develop scientific research works on the unique island nation of Maldives. A recent study by [Li et al. \(2023\)](#) proposes the multi-barriers model as an extension to the barrier model for characterizing areas of interest with multiple obstacles. The model divides the area of interest into general and obstacle regions and uses Gaussian random fields and SPDEs to construct continuous Gaussian fields. INLA is employed to calculate the posterior mean and parameters for spatial regression. Real data sets of burglaries in a certain area are used to compare the performance of the stationary Gaussian model, barrier model and Multi-Barriers Model. The comparison results suggest that the three models achieve similar performance in the posterior mean and posterior distribution of the parameters, as well as the deviance information criteria (DIC) value. However, the Multi-Barriers Model can better interpret the spatial model established based on the spatial data of the research areas with multiple types of obstacles, and it is closer to reality.

#### **4.10.4 Application of Barrier Model on Linear Network**

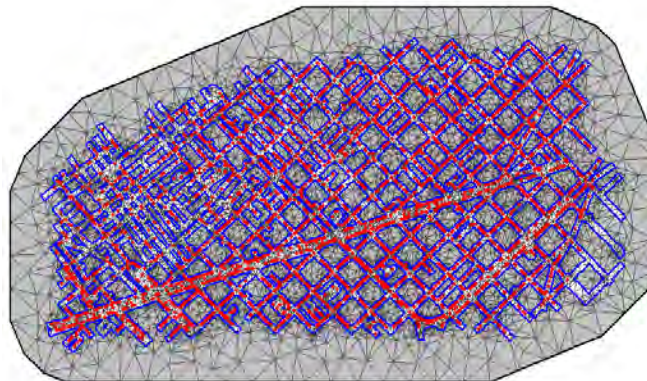
Another study by [Dawkins et al. \(2021\)](#) demonstrates novel approach for personalized decision-making for air quality using Bayesian methods. A hierarchical spatiotemporal model is

developed for city air quality that incorporates buildings as physical barriers and covariate information. High-resolution PM<sub>2.5</sub> data is used to train the model, which is then fit using R-INLA for computation at operational timescales. A method is proposed for eliciting multi-attribute utility for individual journeys within a city, providing Bayes-optimal journey decision support. The methodology is demonstrated using air quality data and a set of journeys in Brisbane city centre, Australia. Although the barrier model was not designed specifically for linear road networks, the study by [Dawkins et al. \(2021\)](#) inspired to take a similar approach to model traffic accidents by utilizing a barrier model in a linear network topology. In their study, [Bakka et al. \(2019\)](#), considered water body as normal terrain and distinct coastlines and boundaries are used as physical barriers. In contrast, in our study, we have defined polygons of individual road segments with a buffer as our study area and the remaining land areas that do not include roads serve as the physical barriers.



**Figure 11: Barrier objects**  
**Barrier object (left), barrier object with event locations highlighted as red points (right)**

The creation of the clipped buffer region and aggregation of the number of minor injuries (which serve as the response variable in the model) at the centroids of each road segment have been accomplished using the same approach outlined in [Section 4.10.1](#).



**Figure 12: Mesh with barrier object**

Figure 11 (left panel) depicts the barrier object where, region in grey indicates the physical barrier and white area indicates the buffered road segments combined together as clipped polygon where the spatial dependency will be analysed.

The right panel of Figure 11 is the same barrier object with the locations of events (here traffic accidents) highlighted in red. SPDE Triangulation is designed using barrier model where clipped polygon is considered as normal terrain and the regions without roads act as physical barriers. Figure 12 illustrates the triangulation along with the physical barrier in grey.

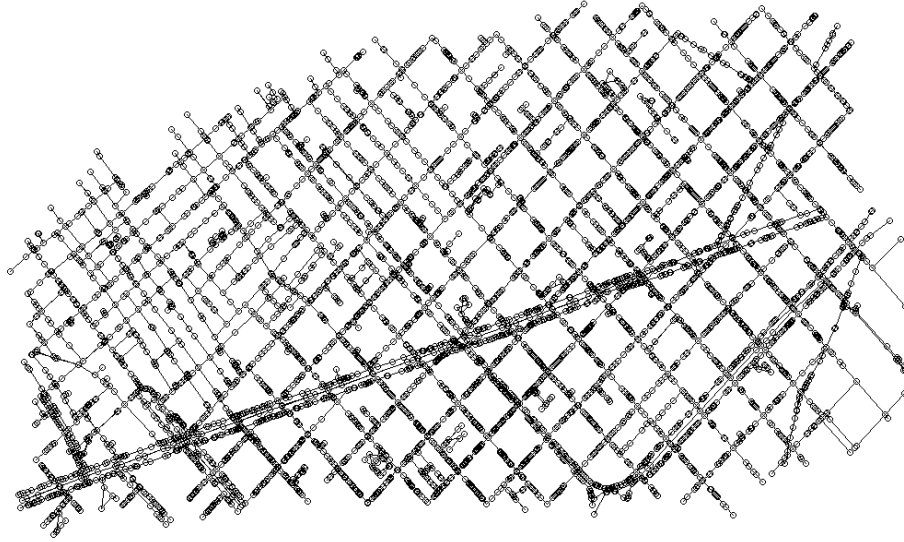
#### 4.10.5 Exponential Graph Model for Linear Network Problems

Our previous articles on explicit network triangulation and ongoing research on barrier models for complex land structures have highlighted issues related to boundary effects, including the creation of artifact spatial dependencies on the boundary. In standard meshes, boundaries are typically outside the spatial domain of interest, allowing for identification and elimination of these dependencies. However, in more complex meshes like network triangulation or barrier models, boundaries lie within the spatial domain, making it challenging to identify and eliminate these dependencies. Despite these difficulties, in these works we have managed to identify them. However, a different approximation is needed in which the SPDE-INLA approximation does not cause these fictitious spatial dependencies.

In this context, Bolin et al. (2022) presented an alternative to using the Euclidean distance by defining similar models with a non-Euclidean metric on a graph. It presents a novel class of Gaussian processes, called Whittle-Matérn fields (Whittle, 1963), which are defined on compact metric graphs such as street or river networks. However, it can be challenging to find a class of positive definite functions suitable for creating Gaussian fields on metric graphs when using a non-Euclidean metric. It is also difficult to apply the SPDE approach to metric graphs as it is uncertain how to define the differential operator and what kind of covariance functions would result. The study proposes a novel approach of a new and valid differentiable Gaussian field on general compact metric graphs.

We are currently working on that approach. These models are an extension of Gaussian fields with Matérn covariance functions on Euclidean domains to the non-Euclidean metric graph setting and are constructed via a fractional stochastic partial differential equation on the graph. The study establishes the existence of these processes and their sample path regularity properties, including differentiable Gaussian processes. It also shows that a model subclass contains processes with Markov properties and provide a computationally efficient method for evaluating their finite dimensional distributions (Bolin et al., 2020; Bolin et al., 2022). The proposed models can be used for statistical inference without the need for any approximations, and can derive several statistical properties, including consistency of maximum likelihood estimators and asymptotic optimality properties of linear prediction based on the model with mis-specified parameters.





**Figure 13: Graph data structure of traffic accident (Barcelona, Spain)**

In the current study we have used the traffic accident dataset of Barcelona city, from January 2010 to December 2019. The selected time frame was chosen in order to minimize bias in the analysis of traffic accident data. This decision is based on the fact that the COVID-19 pandemic has a significant impact on traffic patterns and behavior, which could skew the results of an analysis that includes data from the next few years. By excluding this period, we can more accurately capture the underlying trends and patterns in traffic accidents. The Barcelona city road network along with traffic accident locations are converted into a two-dimensional graph. [Figure 13](#) shows the graph structure where, individual accident locations, start and end points of each road segments and the intersecting points of road segments are depicted as the nodes (or, vertices) and the connecting road segments for the nodes are represented as the edges. Euclidean distances between each node have been calculated and used in the exponential graph model.

We have introduced the Gaussian Whittle-Matérn random fields on metric graphs and have provided a comprehensive characterization of their regularity properties and statistical properties ([Bolin et al., 2020](#)). We argue that this class of models is a natural choice for applications where Gaussian random fields are needed to model data on metric graphs. Of particular importance here are the Markov cases ([Bolin et al., 2020](#); [Bolin et al., 2022](#)). We derived explicit densities for the finite dimensional distributions in the exponential case  $\nu = 1$ , where we can note that the model has a conditional autoregressive structure of the precision matrix ([Besag, 1974](#)).

For the differentiable cases, such as  $\nu = 2$ , we derived a semi-explicit precision matrix formulated in terms of conditioning on vertex conditions. In both cases, we obtain sparse precision matrices that facilitate the use in real applications to big datasets via computationally efficient implementations based on sparse matrices ([Rue and Held, 2005](#)). There are several extensions that can be considered to this work. The most interesting in the applied direction is

to use the models in log-Gaussian Cox processes to model count data on networks, where it is likely that most applications of this work can be found. For example, log-Gaussian Cox processes on linear networks can be suitable for modeling crimes in cities or accidents on road networks or to identify faults in complex water pipeline system or, electrical connections in buildings (Bolin et al., 2020; Bolin et al., 2022). Another interesting extension is to consider Type-G extensions of the Gaussian Whittle–Matérn fields similarly to the Type-G random fields in (Bolin and Wallin, 2020). An interesting property in such a construction is that the process on each edge could be represented as a subordinated Gaussian process (Bolin et al., 2020; Bolin et al., 2022). One of the interesting outputs of this study will be the implementation of *exponential graph model* as a plugin in the INLA package to execute spatiotemporal modeling precisely for complex linear networks. Details about Gaussian Whittle–Matérn random fields on metric graphs have been discussed in Section 5.5.

## 5. Results

---

In this chapter, we highlight a collection of publications that showcase original contributions to the field of spatiotemporal modeling in complex spatial regions. These works represent advancement in our understanding of this research domain. These publications deepen our understanding of the underlying phenomena and pave the way for further research and innovation in this area.

### *List of Publications*

Publication	Quality criteria	Status
Chaudhuri, S., Giménez-Adsuar, G., Saez, M., & Barceló, M. A. (2022). PandemonCAT: Monitoring the COVID-19 Pandemic in Catalonia, Spain. <i>International Journal of Environmental Research and Public Health</i> , 19(8), 4783. <a href="https://doi.org/10.3390/ijerph19084783">https://doi.org/10.3390/ijerph19084783</a>	Impact Factor (2021): 4.614 Public, Environmental & Occupational Health, position 45 out of 182 (Q1).	Published
Chaudhuri, S., Juan, P., & Mateu, J. (2022). Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation. <i>Journal of Applied Statistics</i> , 1-22. <a href="https://doi.org/10.1080/02664763.2022.2104822">https://doi.org/10.1080/02664763.2022.2104822</a>	Impact Factor (2021): 1.416 Statistics & Probability, position 73 out of 125 (Q3).	Published
Chaudhuri, S., Juan, P., Varga, D., & Saez, M. (2023). Spatiotemporal modeling of traffic risk mapping: A study of urban road networks in Barcelona, Spain. <i>Spatial Statistics</i> , 53, 100722. <a href="https://doi.org/10.1016/j.spasta.2022.100722">https://doi.org/10.1016/j.spasta.2022.100722</a>	Impact Factor (2021): 2.125 Statistics & Probability, position 41 out of 125 (Q2).	Published

## **5.1 Article 1: PandemonCAT**

### **PandemonCAT: Monitoring the COVID-19 Pandemic in Catalonia, Spain**

Somnath Chaudhuri<sup>1,2</sup>, Gerard Giménez-Adsuar<sup>1</sup>, Marc Saez<sup>1,2</sup> and Maria A. Barceló<sup>1,2</sup>

1. Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, 17004 Girona, Spain.
2. CIBER of Epidemiology and Public Health (CIBERESP), 17003 Madrid, Spain.

**International Journal of Environmental Research and Public Health, 19(8), 4783**

Impact Factor (2021): 4.614

Public, Environmental & Occupational Health, Position 45 out of 182 (Q1)



Article

# PandemonCAT: Monitoring the COVID-19 Pandemic in Catalonia, Spain

Somnath Chaudhuri <sup>1,2</sup>, Gerard Giménez-Adsuar <sup>1</sup>, Marc Saez <sup>1,2</sup> and Maria A. Barceló <sup>1,2,\*</sup>

<sup>1</sup> Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, 17004 Girona, Spain; chaudhuri.somnath@udg.edu (S.C.); ggadsuar@gmail.com (G.G.-A.); marc.saez@udg.edu (M.S.)

<sup>2</sup> CIBER of Epidemiology and Public Health (CIBERESP), 17003 Madrid, Spain

\* Correspondence: antonia.barcelo@udg.edu

**Abstract:** Background: The principal objective of this paper is to introduce an online interactive application that helps in real-time monitoring of the COVID-19 pandemic in Catalonia, Spain (PandemonCAT). Methods: This application is designed as a collection of user-friendly dashboards using open-source R software supported by the Shiny package. Results: PandemonCAT reports accumulated weekly updates of COVID-19 dynamics in a geospatial interactive platform for individual basic health areas (ABSs) of Catalonia. It also shows on a georeferenced map the evolution of vaccination campaigns representing the share of population with either one or two shots of the vaccine, for populations of different age groups. In addition, the application reports information about environmental and socioeconomic variables and also provides an interactive interface to visualize monthly public mobility before, during, and after the lockdown phases. Finally, we report the smoothed standardized COVID-19 infected cases and mortality rates on maps of basic health areas ABSs and regions of Catalonia. These smoothed rates allow the user to explore geographic patterns in incidence and mortality rates. The visualization of the variables that could have some influence on the spatiotemporal dynamics of the pandemic is demonstrated. Conclusions: We believe the addition of these new dimensions, which is the key innovation of our project, will improve the current understanding of the spread and the impact of COVID-19 in the community. This application can be used as an open tool for consultation by the public of Catalonia and Spain in general. It could also have implications in facilitating the visualization of public health data, allowing timely interpretation due to the unpredictable nature of the pandemic.

**Keywords:** COVID-19; mobility; spatiotemporal; shiny



**Citation:** Chaudhuri, S.; Giménez-Adsuar, G.; Saez, M.; Barceló, M.A. PandemonCAT: Monitoring the COVID-19 Pandemic in Catalonia, Spain. *Int. J. Environ. Res. Public Health* **2022**, *19*, 4783. <https://doi.org/10.3390/ijerph19084783>

Academic Editor: Paul B. Tchounwou

Received: 22 March 2022

Accepted: 12 April 2022

Published: 14 April 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Catalonia, the second most populous autonomous community in Spain with 7.6 million inhabitants (Figure 1), as of 10 January 2022, has been the first most affected by the COVID-19 pandemic (the Madrid region being the second most affected), by number of cases (1,333,517 cases, 18.61% of all cases in Spain, 17,625 cases per 100,000 inhabitants, compared to 15,132 cases per 100,000 in Spain), and the second by number of deaths (16,144 deaths, 17.95% of all deaths in Spain, 213 deaths per 100,000 inhabitants, compared to 190 deaths per 100,000 in Spain) [1,2]. The geographical distribution of the spread of the pandemic has not been spatially homogeneous in the Catalan territory and important differences at the small-area level have been observed.

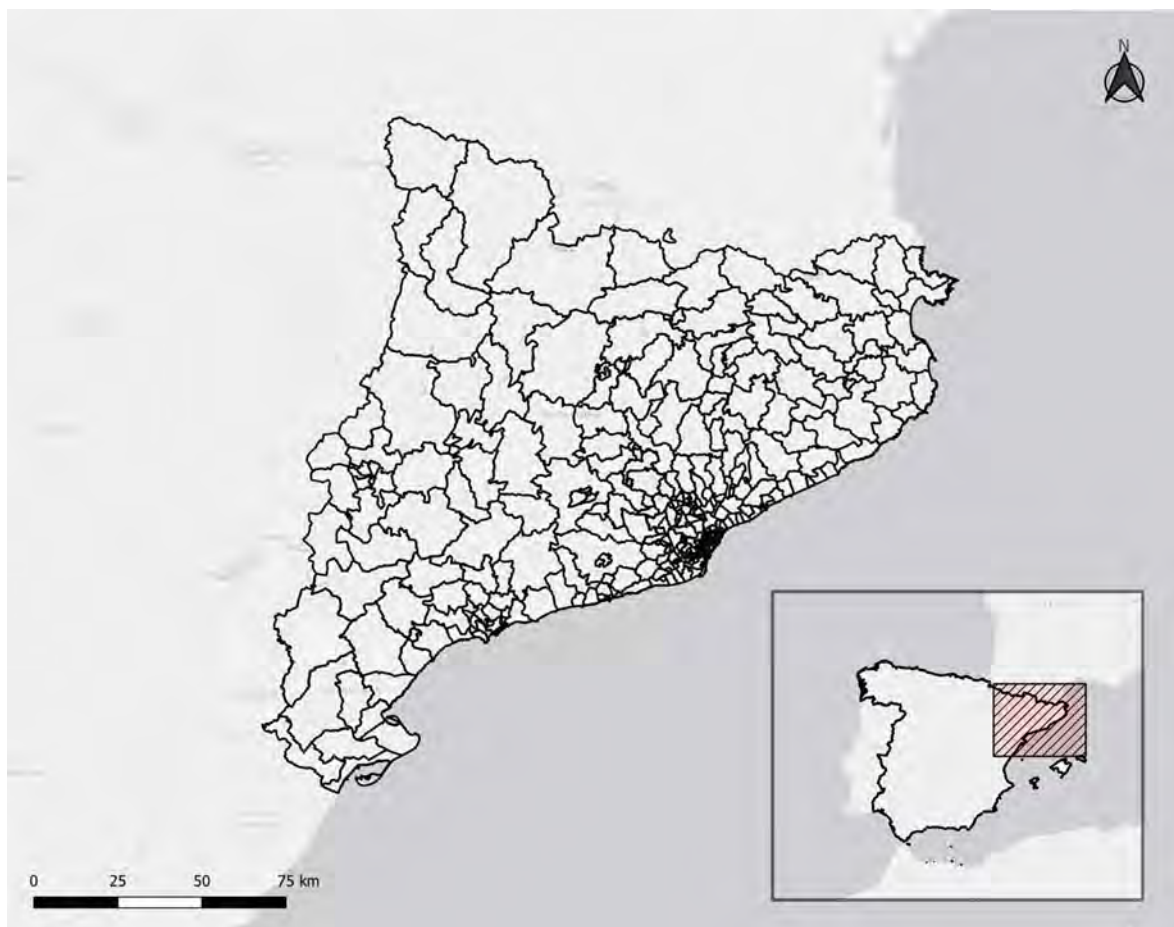


Figure 1. Geographic location of 343 ABSs of Catalonia.

Catalonia is basically an urban region. Sixty percent of the population resides in 23 cities with more than 50,000 inhabitants, and 52% in 14 cities with more than 100,000 inhabitants [3]. These include the second-largest city in Spain, Barcelona, and 36 adjacent municipalities making up the Barcelona Metropolitan Area.

The administrative aggregation levels in Catalonia are the following, from highest to lowest: autonomous community (Catalonia), 42 ‘comarcas’ (county-like regions, hereafter referred to as regions), and 947 municipalities. In addition, the health aggregation levels are the following, from highest to lowest: autonomous community, 8 health regions, 20 health sectors, 34 health basic areas (ABAs), and 343 Basic Health Areas (ABSs, in the Catalan language).

According to the definition by Catalan health planning, an ABS is defined as a basic geographical unit through which primary health care services are coordinated and [4]. Catalan practitioners, pediatricians, dentists, nurses and nursing assistants, social workers, and administrative support professionals make up each ABS. In the year 2020, the 343 ABSs in Catalonia had populations ranging from 32,326 to 202,666 people (mean 60,266 inhabitants, standard deviation 13,391, 16th and 84th percentiles 45,718 and 118,157 inhabitants, first quartile Q1 5540, 54th and 46th percentiles Q2 27,529, 52th the population density per square kilometer ranges from 0.31–34,590.58 (mean 3486.36, standard deviation 6719.23, median 309.18, Q1 44.83, Q3 3752.54) [3].

Our objective in this paper is to introduce an online interactive application that helps in real-time monitoring of the COVID-19 pandemic in Catalonia, Spain. Since the outbreak of the pandemic, real-time interactive web applications have gained attention, especially in the domain of monitoring and modeling the dynamics of COVID-19, throughout the globe. Several shiny dashboards are currently deployed (e.g., managed by individuals or as government health organizations). Much attention in their recent study explored and

lyzed a series of novel and interesting web-based applications that have been specifically developed during the pandemic that can act as tools for the health professional community to help in advancing their analysis and research [5]. For example, Fernandez-Lozano et al. created an interactive dashboard to visualize all data related to the pandemic (cases, hospitalizations, and deaths) and its temporal evolution [6]. However, their tool does not display estimates or other variables (environmental, vaccine status, etc.). The app created by Galván-Tejada et al. offers greater interactivity and completeness than the former but falls short of reporting updated vaccination data as well as other variables, even though they provide valuable demographic information on top of a very thorough analysis of the pandemic in Mexico [7]. Another useful online web application for updated country-specific analysis and visualization is “COVID19-World” by Tebé et al. used for basic epidemiological surveillance covering time trends and projections, population fatality rate, case fatality rate, and basic reproduction number [8]. Wissel et al. developed “COVID-19 Watcher”, a similar web resource that displays COVID-19 data from every county and 188 metropolitan areas in the United States [9]. It provides the rankings of the worst-affected areas along with auto-generating plots depicting temporal changes in testing capacity, cases, and deaths. It is important to mention the “covid19.Explorer” R package and web application by Revell that has been designed to explore and analyze United States COVID-19 infection, death, and relative risk for different age groups with emphasis on geographic progress of the pandemic and effectiveness of lockdowns [10]. A similar interactive Python-based analytical tool to compare data and monitor trends across geographical areas related to the COVID-19 pandemic across counties in the United States and worldwide was developed by Zohner et al. [11]. “Mortality Tracker” is another interesting in-browser application developed by Almeida et al. mainly focused on the visualization of public time series of COVID-19 mortality in the United States [12]. It was developed in response to requests by epidemiologists to access the effect of COVID-19 on other causes of death by comparing 2020 real time values with observations from the same week in the previous 5 years, thus facilitating modeling of the interdependence between its causes. The literature shows a similar web application named “COVID19-Tracker” for Spain developed by Tobías et al. [13]. It produces daily updated data visualization and analysis of the COVID-19 diagnosed cases, and mortality in Spain. It also explores several analyses to estimate the case fatality rate, assessing the impact of lockdown measures on incident data patterns, estimating infection time and the fundamental reproduction number, and analyzing the mortality excess. An attempt at real-time statistical analysis in a user-friendly dashboard for researchers as well as the general public is made by Salehi et al. [14]. It includes two mathematical methods (pandemic logistic and Gompertz growth models) to predict the dynamics of COVID-19, as well as the Moran’s index metric, which provides a geographical perspective via heat maps and can help in the identification of latent reactions and behavioral patterns. Literature shows similar applications being implemented and maintained by researchers from various domains and different countries [15–18].

All these apps for respective regions systematically produce daily updated COVID-19 data visualizations and analysis. But a robust app with a combination of relevant socioeconomic and environmental risk factors and their interrelation with the dynamics of the pandemic has been less explored. Thus, this complete project represents eight interactive dashboards which collectively enable monitoring of the pandemic in Catalonia (PandemonCAT) and explore environmental and socioeconomic factors in the spatiotemporal evolution of the pandemic.

This is a multicenter project, led by the Research Group on Statistics, Econometrics and Health (GRECS) of the University of Girona, Spain, in which the Andalusian School of Public Health (EASP) (Granada, Spain) and the University of Granada also participate. The aim of the PandemonCAT project is to provide a web application that allows the monitoring of the COVID-19 pandemic in Catalonia. Its results include, in addition to vaccination, the results of diagnostic tests, transmission (reproductive number), hospitalization, ICU admissions, and the number of deaths. It also provides a visualization of

those variables that could influence the spatiotemporal dynamics of the pandemic. Thus, in an interactive interface, the environmental variables (air pollutants and meteorological variables), the socio-economic variables, the points of interest where there may be a greater risk and the data of public mobility are shown.

The principal aim of the project is to provide a web application that allows the monitoring of the COVID-19 pandemic in Catalonia. Its results include, in addition to vaccination, the results of diagnostic tests, transmission (reproductive number), hospitalization, ICU admissions, and the number of deaths. It also provides a visualization of those variables that could influence the spatiotemporal dynamics of the pandemic. Thus, in an interactive interface, environmental variables (air pollutants and meteorological variables), socioeconomic variables, points of interest where there may be a greater risk, and data of public mobility are shown. The interactive web application reports a comprehensive list of all key variables with respect to the disease: new cases, hospitalizations, ICU's, and deaths. We have also depicted the vaccination flow (both first and second doses) for different age groups of the population in individual health zones of the community. Finally, we report the smoothed standardized COVID-19 infected cases and mortality rates on maps of ABSs and regions of Catalonia. Table A1 in Appendix A reports information about individual dashboards of the application along with respective components and brief descriptions.

The rest of the article is organized as follows. In Section 2 we present an overview of the methodology followed to design PandemonCAT. The subsections report detailed descriptions of individual components of the complete methodology. Section 3 is devoted to presenting the results of individual components of PandemonCAT. In Section 4 we briefly discuss possible implications in other fields of study, as well as enhancements that may be implemented to further develop the current study. The article ends with a conclusion in Section 5.

## 2. Methods

PandemonCAT has been developed in the RStudio Shiny framework [19]. The application uses R packages to execute all analysis and plots internally. The key R packages used in the tool implementation include dplyr [20] and tidyverse [21] for data management. Packages like rgdal [22], sf [23], raster [24], maptools [25], and flowmap.blue [26] are used for spatial data analysis and visualizations. Interactive charts are generated with plotly package [27], while static graphical displays are designed using ggplot2 package [28]. Leaflet [29] along with leafpop [30] packages are used to generate interactive geospatial maps. The Shiny package [19] with Shiny flexdashboard [31] and rmarkdown [32] are used extensively for application enhancement and implementation as interactive web apps directly from R.

Figure 2 depicts the workflow diagram of the PandemonCAT application. Since 1 January 2020, we have retrieved, and curated data and it is being updated weekly with the new data reported by different data sources. The raw data accessed from multiple open data sources (referred in Section 2.1) are initially cleaned and preprocessed to ensure consistency and reliability. To speed up the analysis and visualization process, initial data wrangling techniques such as merging multiple heterogeneous data sources and discarding redundant variables and duplicate observations are performed. We automated the weekly data wrangling process because the raw datasets are extremely large and unstructured. The local databases are updated automatically every week in the dedicated cloud server. Section 2.2 provides a complete outline of the data remediation process. In the next phase, spatial, temporal, and spatiotemporal analysis are performed on the periodically updated datasets. All the analyses have been carried out using R version 4.0.1 [33]. The results of these analyses are displayed in interactive Shiny dashboards. Finally, all these dashboards are combined as a single application with brief information about each dashboard and individual access links. The integrated application runs on shared cloud servers shinyapps.io [34] that are operated by RStudio [35]. The site is maintained by the Research Group on Statistics, Econometrics and Health (GRECS), at the University of Girona, Spain.



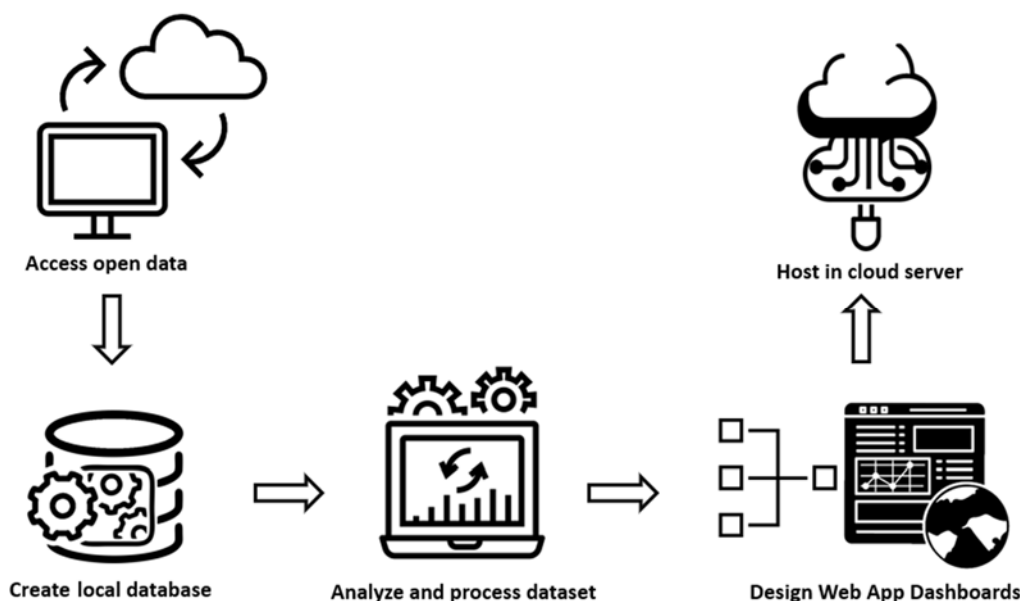


Figure 2. Workflow diagram of PandemonCAT.

2.1. Data Collection

All data displayed in PandemonCAT comes directly from official government sources. Specifically, we leverage the existence of open datasets as part of the increasing effort of the government to become more transparent. Early on during the COVID-19 pandemic, new datasets were created and shared, starting with the spatiotemporal incidence of both the number of cases and deaths by region and health zones (ABS). The existence of daily counts per day and geographic delimitation made a spatiotemporal representation possible. Unlike many autonomous communities, which only shared aggregated data (i.e., total counts per region, without the temporal evolution), Catalonia stood out in this regard and made apps such as PandemonCAT possible early on.

As the lockdown was ending (June 2020), more datasets were readily available such as the number of tests performed, the number of hospitalized (ICU) and in the beginning of 2021, the number of vaccinated individuals. This allowed PandemonCAT to add all relevant variables to be displayed in our visualizations.

The current process uses daily updates from the following open COVID-19 datasets:

- The current process uses daily updates from the following open COVID-19 datasets:
- Number of: Cases per ABS
- Cases per ABS Hospitalizations (and hospitalized) per region
- Hospitalizations (and hospitalized) per region
- ICU per region
- ICU per region Test per ABS
- Test per ABS Vaccinated individuals per ABS
- Vaccinated individuals per ABS
- Official data sources are as follows:
- Official data sources are as follows: the open datasets provided from the government: <https://analisi.transparenciacatalunya.cat/> (accessed on 11 April 2022)
- General link with all the open datasets (accessed on 11 April 2022): <https://analisi.transparenciacatalunya.cat/api/views/c7sd-zy9j/rows.csv?accessType=DOWNLOAD&sorting=true> (accessed on 11 April 2022)
- Regions: <https://analisi.transparenciacatalunya.cat/api/views/065f-nl6a/rows.csv?accessType=DOWNLOAD&sorting=true> (accessed on 11 April 2022)
- ABS: <https://analisi.transparenciacatalunya.cat/api/views/xuwf-dxjd/rows.csv?accessType=DOWNLOAD&sorting=true> (accessed on 11 April 2022)
- Digitized cartography of the ABS (accessed on 11 April 2022).
- Digitized cartography of the ABS. Cartography [https://salutweb.gencat.cat/ca/el\\_departament/estadistiques\\_sanitaries/cartografia/](https://salutweb.gencat.cat/ca/el_departament/estadistiques_sanitaries/cartografia/) (accessed on 11 April 2022)
- Vaccine:

<https://analisi.transparenciacatalunya.cat/api/views/tp23-dey4/rows.csv?accessType=DOWNLOAD&sorting=true> (accessed on 11 April 2022).

Regarding the meteorological and air pollutant variables, the same official government datasets have been used. Due to their nature, they are not updated daily. Data shown in PandemonCAT is limited to the 2020 period, coinciding with the onset of the pandemic.

Meteorological variables:

METEOCAT, Generalitat de Catalunya, Meteorological data from XEMA

<https://analisi.transparenciacatalunya.cat/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee> (accessed on 11 April 2022).

Air pollution:

<https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu> (accessed on 11 April 2022).

Socioeconomic variables:

We have used various sources for the socioeconomic variables: total population, percentage of population 65 years or more, percentage of population 0–25 years, and percentage of foreigners in 2020 from countries with medium and low human development index [2,36]

[https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica\\_C&cid=1254736177012&menu=resultados&secc=1254736195461&idp=1254734710990#!tabs-1254736195557](https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736177012&menu=resultados&secc=1254736195461&idp=1254734710990#!tabs-1254736195557) (accessed on 11 April 2022).

Average income per person (in Euros):

Average of the years 2015, 2016, 2017 and 2018 [37]

[https://www.ine.es/en/experimental/atlas/exp\\_atlas\\_tab\\_en.htm](https://www.ine.es/en/experimental/atlas/exp_atlas_tab_en.htm) (accessed on 11 April 2022).

Unemployment rate [38]:

[http://www.ine.es/censos2011\\_datos/cen11\\_datos\\_resultados\\_seccen.htm](http://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm) (accessed on 11 April 2022).

As in the case of pollutants, the data is limited to 2020, unless otherwise stated (for example, average income per person and unemployment rate).

## 2.2. Data Settings

Open data is an excellent source of information; however, raw data cannot be directly represented to convey important information regarding the state of the pandemic. Specifically, all figures need to be adjusted by population size (i.e., representing figures by 100,000 inhabitants). This adjustment is made possible due to the existence of updated demographic datasets with a low level of aggregation (for both regions and ABSs). The combination of both datasets (mainly thanks to the `dplyr` package [20] in R) has allowed us to create the following variables, which are the absolute reference for assessing the pandemic:

Weekly cases per 100,000 inhabitants: New cases adjusted for population on a 7-day window period. A new case is defined as those people who have received at least one positive PCR or antigenic test result during that period. The new case is allocated to the place of residence of such a person. That is if a person is registered to live in Barcelona, for instance, but gets a positive result from a hospital in other municipality, the new case is attributed to Barcelona.

Empiric 7-day  $R_t$ : indicates the rate of change of the new cases and is calculated as the ratio of the cumulative sum of weekly cases between  $t$  and  $t - 5$ .

Weekly tests per 100,000 inhabitants: PCR and antigenic tests performed on a 7-day period, adjusted for population, regardless of their results.

Weekly deaths per 100,000 inhabitants: new deaths attributed to COVID-19, adjusted for population, on a 7-day period.

Currently hospitalized per 100,000 inhabitants: number of people currently hospitalized due to COVID-19, adjusted for population.

Currently ICU per 100,000 inhabitants: number of people currently in intensive care unit (ICU) due to COVID-19, adjusted for population.

On the vaccination front, the same procedure of adjusting by population is performed at the health zone level. This has proved to be critical, especially since the inter-health zone population differences are large (and thus, absolute numbers don't give an accurate picture of the progress of the vaccination campaign).

### 2.3. Spatial Prediction of Air Pollutant Levels

In this section we provide the spatial predictions of the levels of atmospheric pollutants for each ABS in Catalonia. The problem is that the air pollution monitoring stations are not distributed homogeneously throughout the territory of Catalonia, but rather are concentrated, mainly in the Barcelona region. Therefore, we follow our previous work [39].

Specifically, our objective there was to perform spatial predictions of air pollution levels using a hierarchical Bayesian spatiotemporal model [39,40] that allowed us to perform the predictions in an effective way and with very few computational costs [39]. We used the Stochastic Partial Differential Equations (SPDE) representation [41] of the integrated nested Laplace approximations (INLA) approach [42,43] to spatially predict, in the territory of Catalonia, the levels of the four pollutants for which there is the most evidence of an adverse health effect: coarse particles, nitrogen dioxide, ozone, and carbon monoxide (pollutants of interest) [39]. We performed the spatial predictions at a point level (defined by its UTM coordinates), allowing them to be aggregated later in any spatial unit required (ABSs in our case). We were especially interested in the long-term exposure to air pollutants. That is, by living in a certain area an individual is exposed to a mix of pollutants that have lasting effects on their health.

We obtained information on the levels of air pollutants for 2011–2020 from the 143 monitoring stations from the Catalan Network for Pollution Control and Prevention (XVPCA) (open data) [44], located throughout Catalonia and that were active during that period. The pollutants we were interested in for making spatial predictions were coarse particles, those with a diameter of 10  $\mu\text{m}$  ( $\mu\text{m}$ ) or less ( $\text{PM}_{10}$ ), nitrogen dioxide ( $\text{NO}_2$ ), ozone ( $\text{O}_3$ ) (all of them expressed as  $\mu\text{m}/\text{m}^3$ ) and carbon monoxide, CO (all of them expressed as  $\text{mg}/\text{m}^3$ ) (air pollutants of interest, hereinafter). However, the monitoring stations also measured other pollutants: fine particles, those with a diameter of 2.5  $\mu\text{m}$  or less ( $\text{PM}_{2.5}$ ), nitrogen monoxide (NO), sulphur dioxide ( $\text{SO}_2$ ), benzene ( $\text{C}_6\text{H}_6$ ), hydrogen sulphide ( $\text{H}_2\text{S}$ ), dichloride ( $\text{Cl}_2$ ), and heavy metals (mercury, arsenic, nickel, cadmium, and lead).

We specified a hierarchical spatiotemporal model:

$$Z(s_i, t) = Y(s_i, t) + \varepsilon(s_i, t)$$

where  $i$  denotes the air pollution monitoring station where the pollutant was observed;  $t$  is the time unit;  $s_i$  the location of the station;  $Y(\cdot, \cdot)$  the spatiotemporal process, the realization of which corresponds to the pollutant measurements (at station  $i$  and time unit  $t$ ); and  $\varepsilon(\cdot, \cdot)$  the measurement error defined by a Gaussian white-noise process (i.e., spatially and temporally uncorrelated).

The spatiotemporal process,  $Y(\cdot, \cdot)$ , is a spatiotemporal Gaussian field that changes in time according to an autoregressive of order one (AR(1)).

The measurement equation was specified as:

$$y(s_i, t) = \mu(s_i, t) + \eta(s_i, t)$$

where  $\mu(\cdot, \cdot)$ , denotes a large scale component and  $\eta(\cdot, \cdot)$  the realization of a spatiotemporal process, specified as,

$$\eta(s_i, t) = \phi\eta(s_i, t - 1) + \omega(s_i, t) \text{ where } |\phi| < 1.$$

$\omega(s_i, t)$ , which was assumed to be a zero mean Gaussian and a Matérn covariance function:

$$\text{Cov}(\eta(s_i, t), \eta(s'_i, t)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa\|s_i - s'_i\|)^{\nu} \text{K}_{\nu}(\kappa\|s_i - s'_i\|)$$

where  $K_\nu$  is the modified Bessel function of the second type and order  $\nu > 0$ ,  $\nu$  is a parameter controlling the smoothness of the GF,  $\sigma^2$  is the variance, and  $\kappa > 0$  is a scaling parameter related to the range,  $\rho$ , the distance to which the spatial correlation becomes small.

The linear predictor of the GLMM specification of the large-scale component,  $\mu(\cdot, \cdot)$  was,

$$\mu_{i,t} = \beta_0 + \sum_{j=1}^{14} \beta_j \text{pollutant}_{j,it} + \beta_{15} \text{altitude}_i + \beta_{16} \text{area}_i + \eta_i + S_i + \tau_{\text{month}}$$

We included as covariates: (1) *pollutant*: the pollutants of interest other than the pollutant for which the spatial prediction was made and, second, the rest of the pollutants that are measured in each monitoring station (i.e., PM<sub>2.5</sub>, NO, SO<sub>2</sub>, C<sub>6</sub>H<sub>6</sub>, H<sub>2</sub>S, Cl<sub>2</sub>, mercury, arsenic, nickel, cadmium, and lead); (2) *altitude*: the altitude of the air pollution monitoring station; and (3) *area*: the area of the ABS. On the other hand, including random effects, we controlled for heterogeneity (those unobservable factors that could be associated with the levels of the pollutant)  $\eta_i$  (unstructured random effect indexed on the ABS), spatial dependence (that is, the existence of geographic patterns),  $S_i$  (structured random effect according a Matérn); and temporal dependence (trend and seasonality),  $\tau_{\text{year}}$  and  $\tau_{\text{month}}$ , respectively (structured random effects indexed on year and month, respectively).

#### 2.4. Smoothing of the Rates of the Outcomes from COVID-19

The simplest disease (or mortality) maps represent the cases or deaths observed in each geographic area. However, any interpretation of the geographical structure of the cases is limited by the lack of information on the spatial distribution of the population that could be 'at risk' and, consequently, has generated these observed cases. Therefore, the representation of rates that allow incorporating the effect of the population at risk is preferred, instead of representing gross cases. However, the direct use of crude rates does not allow comparison between different areas, since the differences observed between them may be due to risk factors that have not been considered, such as age. One measure that considers the age structure is the age-standardized rate. There are two methods for age standardization, which are known as direct and indirect standardization. In the representation of disease maps, the use of the indirect method is preferred, which consists of comparing the observed cases of the disease in an area with the expected cases if the risks for each age group were the same as in a certain area reference population. The observed/expected ratio is called the standardized incidence (or mortality) rate (SIR or SMR), which is nothing more than an estimator of the relative risk of the area, that is, of the risk of illness (or death) in relation to the reference group [45,46].

SIRs (or SMRs), even though they have been widely used, have some limitations. They depend to a great extent on the population size, since the variance of the standardized rates is inversely proportional to the expected values; thus, areas with little population will present estimators with great variability. In this sense, the extreme standardized rates and, therefore, dominant in the apparent geographic pattern, are those estimated with the least precision in areas with few cases. In addition, the variability of the observed cases is usually much greater than expected, producing what is called 'extra variability'. In fact, when spatial data are available it is important to distinguish two sources of extra variation. In the first place, the most important source is usually the so-called 'spatial dependence', which is a consequence of the correlation of the spatial unit with neighboring spatial units, generally those that are geographically contiguous. Thus, the standardized rates of contiguous, or close, areas are more similar than the standardized rates of spatially distant areas. Part of this dependency is not really a structural dependency but is mainly due to the existence of uncontrolled variables, i.e., those not included in the analysis. Regarding the second source, the existence of extra independent and spatially unrelated variation, called 'heterogeneity', due to unobserved variables without spatial structure that could influence the risk must be assumed [45,46].

To solve the problems derived from the direct use of SIRs (or SMRs), several alternatives have been proposed to “smooth” them, that is, to reduce the extra variation. Specifically, to estimate disease risks it is preferable to use models (known as ‘disease mapping models’) since they allow incorporating explanatory variables and borrowing information from neighboring areas to improve local estimators, smoothing the extreme values because of small sample sizes [45,46].

Here, to smooth the SIRs, we used a log Gaussian Cox (LGCP) model. The LGCP model is the analogue of the Gaussian linear model used for geostatistical data when data is modelled in the form of point processes. However, this model is currently being used to approximate spatial data of any type (that is, areal data, geostatistical data, and point processes) [47].

First, we assessed the existence of a geographic pattern, as well as clusters of cases in the incidence and mortality of COVID-10. To do this, we specified an LGCP, in which we did not include explanatory variables but only random effects that captured: (1) individual heterogeneity not spatially structured, that is, it collects those unobservable confounders associated with each ABS that do not vary over time; (2) the time trend of the risk (in a non-linear way); and (3) the spatial dependence. In our case, the LGCP model had three peculiarities. First, we included as an offset (denominator) the expected number of cases and deaths from COVID-19 in the ABS. In this way we smoothed the SIRs. Second, since there were ABSs that some weeks did not have any cases or deaths, we allowed the dependent variables to have an excess of zeros, assuming that they are distributed according to a negative binomial. Third, in addition to controlling for heterogeneity, spatial dependence, and temporal dependence using random effects, we allowed the spatial pattern of incidence and mortality to vary over time, including a random effect for the interaction between the spatial and the temporal components [48].

Second, in the previous model we included those variables that could have explained the risk of incidence and mortality and, therefore, also the possible geographic patterns and the existence of clusters, if any. As explanatory variables we included socioeconomic variables net income per person (average 2015 to 2018), unemployment rate, population density, percentage of the population aged 65 years or more (average 2015 to 2018), percentage of slums (with surface area smaller than 40 m<sup>2</sup>), and percentage of residents born in low-income countries); meteorological variables (net effective temperature—a thermal index that combines temperature, relative humidity and wind speed—and atmospheric pressure); long-term exposure (from 2009 to 2019) to air pollutants (PM<sub>10</sub>, NO<sub>2</sub>, and O<sub>3</sub>); mobility variables (exits, entrances, and internal movements) and the accumulated weekly percentage of those vaccinated with two doses. All variables were included at the ABS level. The socioeconomic variables (except for density) were collected at the census tract level and for this reason the values for each ABS were obtained by means of a weighted average of the census tracts contained in them, using the ABS population as weights. The values of the meteorological variables and atmospheric pollutants in each ABS were spatially predicted using a hierarchical Bayesian spatiotemporal model. With the exception of the accumulated percentage of those vaccinated with two doses, the rest of the time-dependent variables, that is, for which we had weekly values (meteorological, atmospheric pollutants, and mobility) were included in the model as the average of the values of the previous two weeks (since their effect on incidence and mortality, if any, was not immediate). Finally, we allowed the relationship between incidence and mortality and the explanatory variables to be non-linear.

### 3. Results

The spatiotemporal and visual analytics capabilities included in PandemonCAT can be useful to explore associations and trends among meteorological and air quality variables and COVID-19 indicators, with a layer of socioeconomic and public mobility information. The app demonstrates visualization of these variables that could have some influence on the spatiotemporal dynamics of the pandemic. We believe the addition of these new dimen-

not linear. Finally, we also considered the relationship between incidence and mortality and the explanatory variables to be non-linear.

### 3. Results

The spatiotemporal and visual analytics capabilities included in PandemonCAT can be useful to explore associations and trends among meteorological and air quality variables and COVID-19 indicators, with a layer of socioeconomic and public mobility information. The app demonstrates visualization of these variables that could have some influence on the spatiotemporal dynamics of the pandemic. We believe the addition of these new dimensions, which is the key innovation of our project, will improve the current understanding of the spread and impact of COVID-19. The application can be accessed online (<https://www.udg.edu/pandemoncat>, accessed on 11 April 2022). The application is a dynamic and interactive dashboard, as illustrated in Figure 3, which allows the user to get an overview of the entire application and its different modules. The user can get detailed information and links for the component modules by clicking on individual tabs.

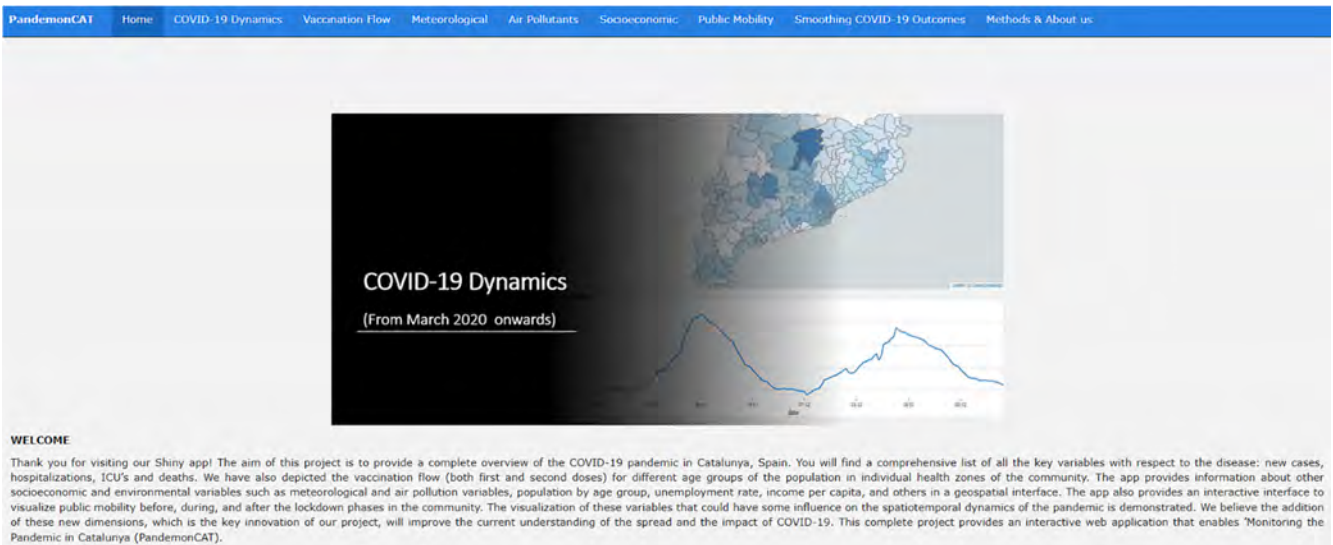
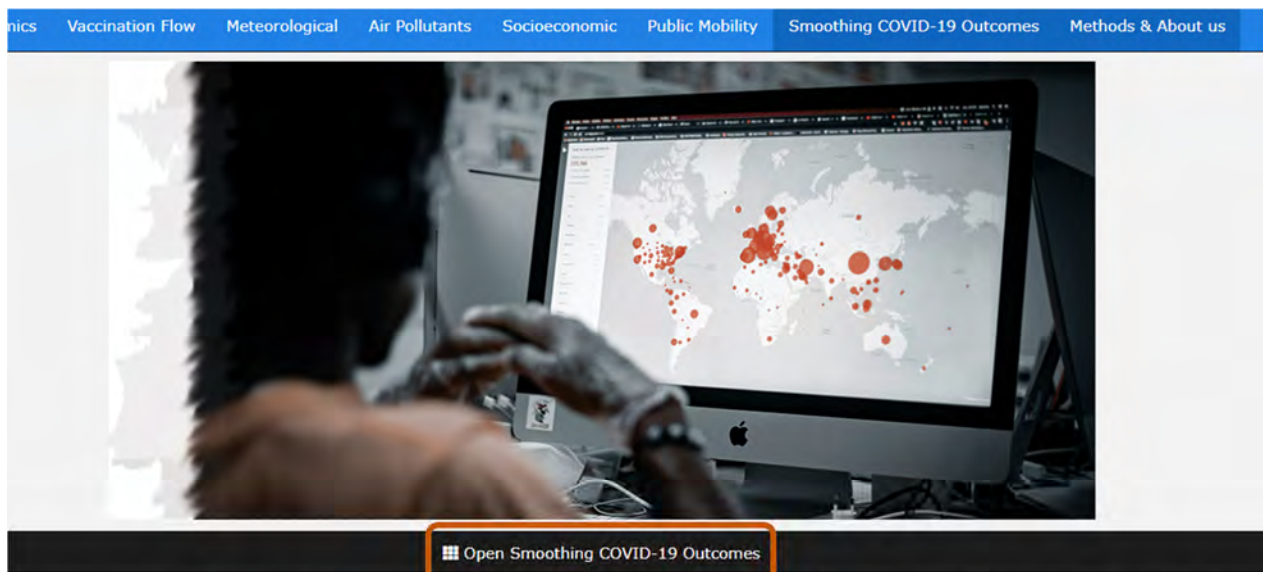


Figure 3. Overview of the PandemonCAT application.

Figure 3. Overview of the PandemonCAT application.

Figure 4 shows the information tab for one component module of PandemonCAT. It provides an outline concept of the particular dashboard, and its unique functionalities along with data type and sources. It also provides the weblink of the component module as highlighted in Figure 4.

Int. J. Environ. Res. Public Health 2022, 19, 4783



incidence (positive cases) and mortality rates by COVID-19 on maps of the Catalonia by Basic Health Areas (ABSs) and comarcas. These smoothed rates allow Those ABSs with smoothed rates higher than unity will have a risk of incidence higher than expected and those with smoothed rates lower than the unit, a lower is of excess cases (i.e., clusters), we also show the 'exceedance probabilities', which are the probability that the smoothed rates were above 1. Those ABSs or cc e or mortality) and those with a probability less than 20% of low risk. Smoothed rates are shown without including explanatory variables and adjusting for varian dulated percentage of people vaccinated with the two doses) and including vaccination and socioeconomic, environmental variables (meteorological and air pollutat

Figure 4. Dashboard for component module of PandemonCAT.

In the following sections we demonstrate the characteristics of individual components of PandemonCAT through various figures supported by relevant functional explanations.

In the following sections we demonstrate the characteristics of individual components of PandemonCAT through various figures supported by relevant functional explanations.

### 3.1. COVID-19 Dynamics

The “COVID-19 Dynamics” app depicts the principal parameters to track the COVID-19 pandemic in Catalonia (Figure 5). In the first section, we present a spatiotemporal map at the ABS level which is the lowest administrative aggregation level for data collection from official open data portals. The interactive map has the option to display the results categorized by the parameters, namely, empirical 7-day reproductive number and weekly records per 100,000 inhabitants for infected cases, tests performed, and deaths. Moreover, it also provides options to check the currently hospitalized and intensive care unit (ICU) patients per 100,000 inhabitants for individual ABSs. Details of these variables are discussed in Section 2.2. The app also provides the option to select any particular date starting from 20 March 2020 to explore spatial distribution of any variables mentioned above. In the next section, a time plot for each of the same variables is available to review its evolution for individual ABS compared with Catalonia as a whole. In this section, the user will have the option to select a particular time period mentioning a start and end date. It is worth noting that, though the majority of the populations are between 20,000 and 40,000, there exists a particular heterogeneity in the population size of each ABS. This fact may be relevant since outliers are more often found in ABSs with low populations that experience serious outbreaks. Figure 5 (left) depicts the parameter options to control the spatiotemporal visualizations. Right (top) map shows the spatial variation of weekly infected cases per 100,000 population for a particular selected date in different ABSs of Catalonia. While the plot on right (bottom) presents the temporal trend of the same variable for a particular ABS compared with the entirety of Catalonia for a selected range of time. In both map and line plots the user can get detailed information for any spatial and temporal resolution with a click (as shown on the map pop-up information window).

Int. J. Environ. Res. Public Health 2022, 19, 4783, Figure 5. Dashboard of PandemonCAT that allows to analyze the spatiotemporal dynamics of COVID-19 in Catalonia.

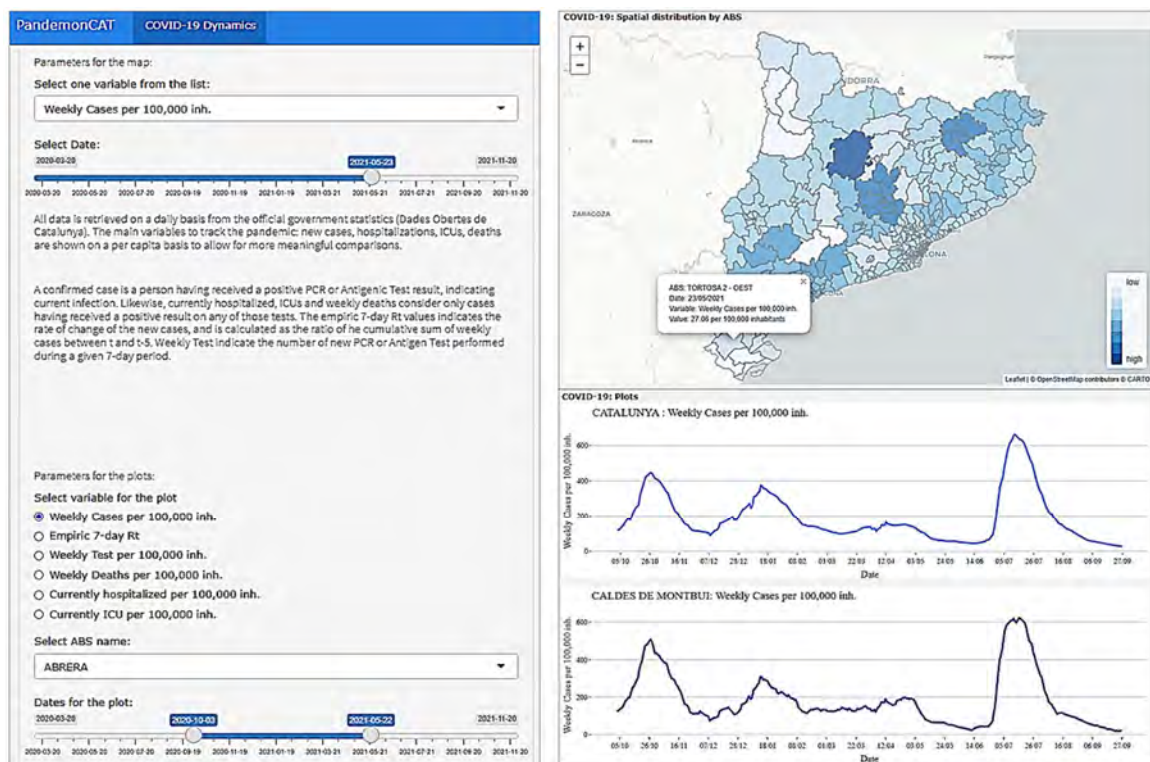


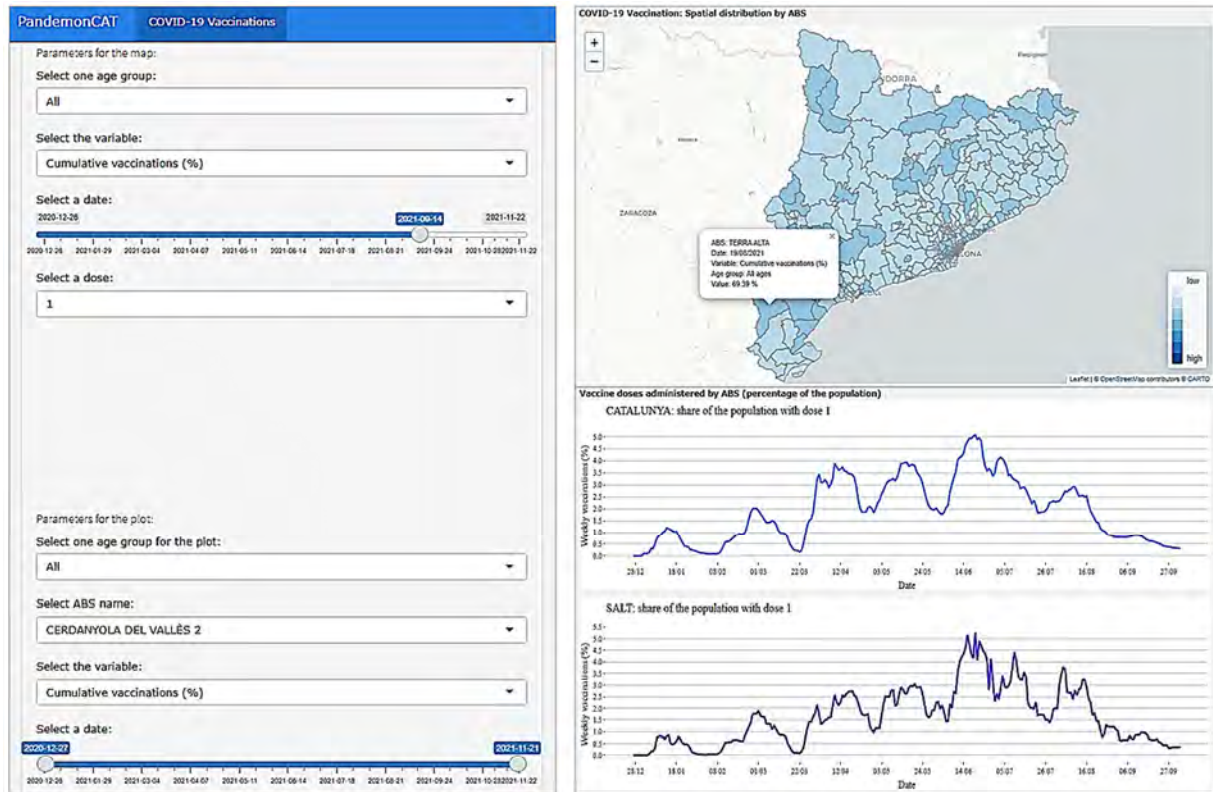
Figure 5. Dashboard of PandemonCAT that allows to analyze the spatiotemporal dynamics of COVID-19 in Catalonia.

### 3.2. Vaccine Rates

The “Vaccination” app displays the progress being made in the vaccination campaign against COVID-19 in Catalonia (Figure 6).

### 3.2. Vaccine Rates

The COVID-19 Vaccination app displays the progress being made in the vaccination campaign against COVID-19 in Catalonia (Figure 6).



**Figure 6.** The dashboard of PandemonCAT enables analysis of the spatiotemporal dynamics of COVID-19 vaccination in Catalonia.

In the first section, we present a spatiotemporal map at the ABS level representing the share of the population with either one or two doses of vaccine for both cumulative and the weekly proportion. In addition to the overall share of population, specific shares per age group can also be explored. The age groups are all 10-year periods, except for 0–9 years old and the second section, in the spirit of the same variables available elsewhere. The plot for Catalonia is also displayed and provides some context to the specific ABS plot. Related to similar issues in ABSs with their population ABSs also mentioned in previous sections (3.1) can be heterogeneity in the population size of the ABSs (left table ABS Figure 6 (left) depicts options to control the spatiotemporal dynamics of the population. Right (top) shows the spatial variation of the population for all age groups of all geographic units on a selected particular ABS of Catalonia. The distribution depicts the distribution of vaccine doses only. While the plot only the weekly (top) presents the trend of the temporal trend of the particular ABS compared with the entire Catalonia. With the map and the plot, the user can get detailed information for any spatial unit with a click or a resolution with a plot (as shown on the map pop-up information window).

### 3.3. Meteorological Variables

The ‘Meteorological’ app displays daily average records of six meteorological components for individual weather stations in Catalonia (Figure 7). The six components included in the current project are atmospheric pressure, precipitation, relative humidity, solar irradiance, temperature, and wind velocity.



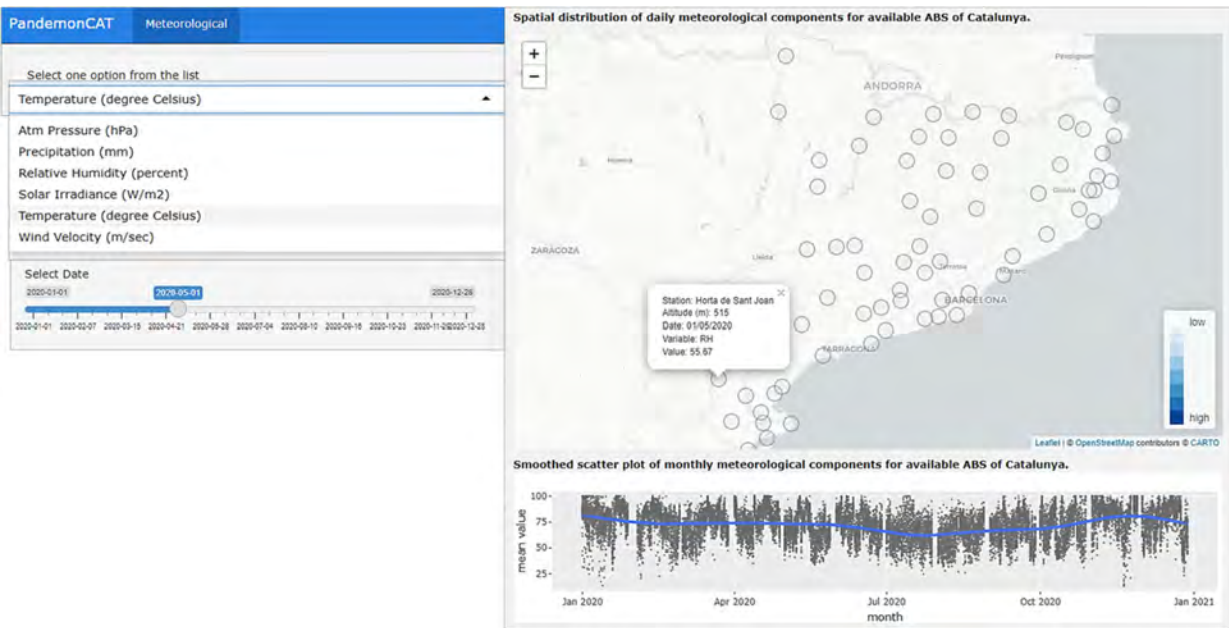


Figure 7. Dashboard of PandemonCAT displaying daily average records of meteorological components from different weather stations in Catalonia.

In the first section, we present the spatial distribution of daily average meteorological records from 75 weather stations located in different ABS of Catalonia. This section depicts a smoothed scatter plot of average monthly records of individual meteorological components for each ABS of Catalonia (Figure 7). In the upper part of the interface, the user can select the type of meteorological component to be displayed. In the right-top part of Figure 7, on a map of Catalonia with the location of the weather stations selected, clicking that particular station displays detailed values of the weather component. Clicking on any station displays detailed values of the weather component for the selected day that particular station. A smoothed scatter plot is displayed below.

3.4. Air Pollutants  
 This app focuses specifically on the daily average concentration of coarse particles (PM<sub>10</sub>), nitrogen dioxide (NO<sub>2</sub>), ozone (O<sub>3</sub>), and carbon monoxide (CO) recorded at 75 pollution monitoring stations in the region (Figure 8).  
 This app focuses specifically on the daily average concentration of coarse particles (PM<sub>10</sub>), nitrogen dioxide (NO<sub>2</sub>), ozone (O<sub>3</sub>), and carbon monoxide (CO) recorded at 75 pollution monitoring stations in the region (Figure 8).

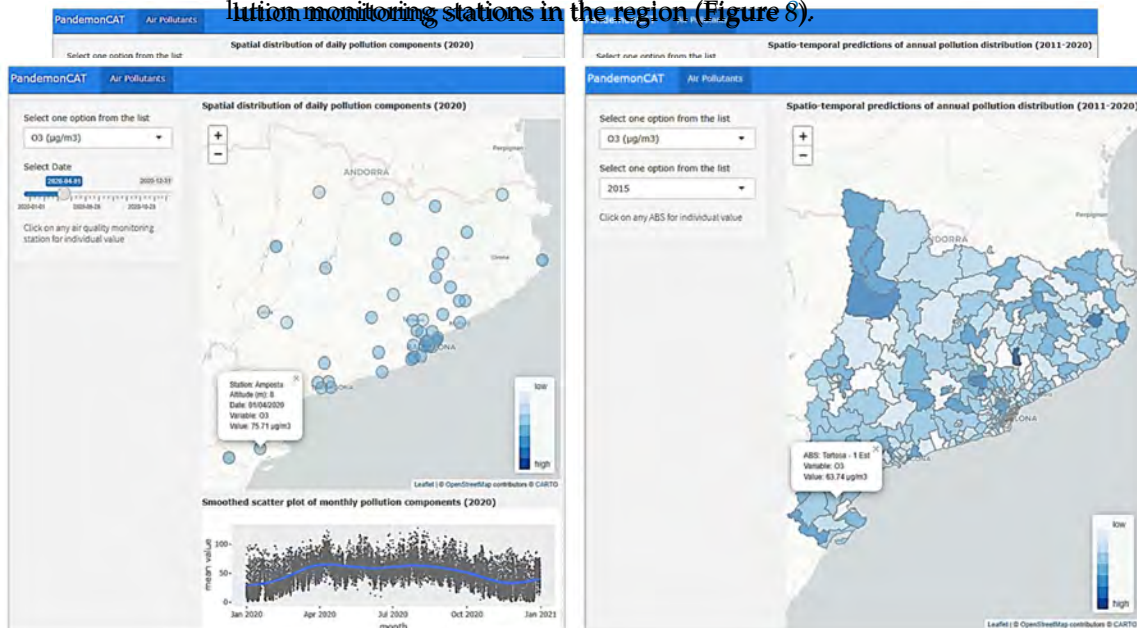


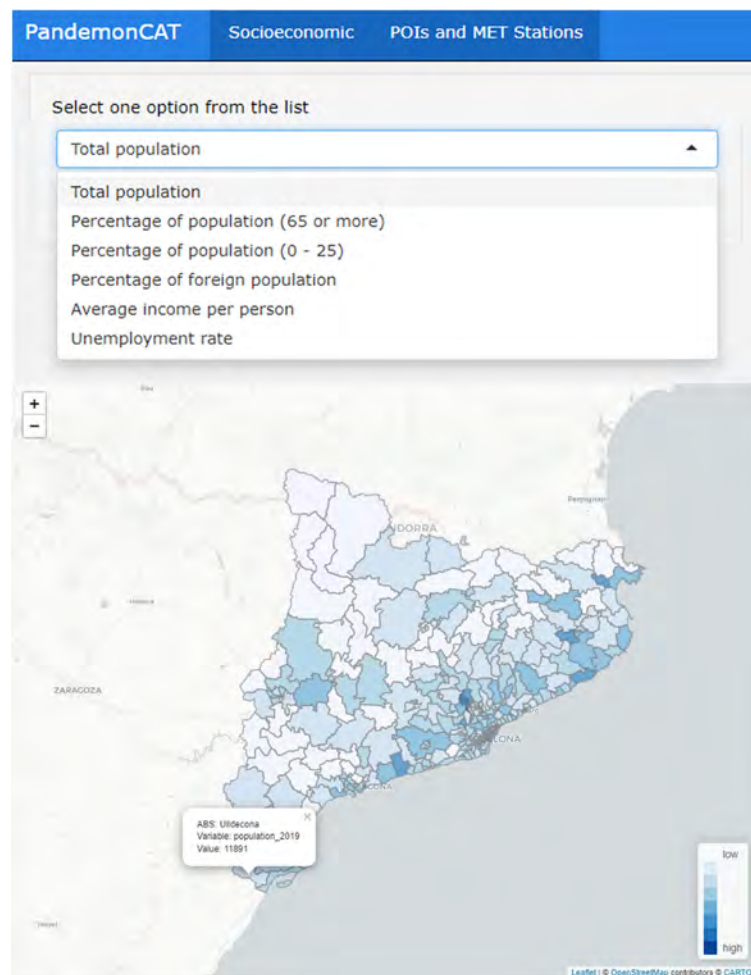
Figure 8. Dashboard of PandemonCAT that provides visualization of the spatiotemporal variation of principal air pollutants in Catalonia.

**Figure 8.** Dashboard of PandemonCAT that provides visualization of the spatiotemporal variation of principal air pollutants in Catalonia.

In the first map we report the daily average concentration of air pollutants for the pollution monitoring stations distributed over individual ABSs of Catalonia. The smooth scatter plot below displays the overall behavior as a monthly average for the pollutants in Catalonia. The set of ABSs in Catalonia is depicted in the scatter plot below, displaying the overall behavior and concentration. The first map provides a selection of pollutants for Catalonia and for 2019 and 2020. Figure 8 (right) shows the prediction for selected type of pollutant and the date. The map provides the options to select a pollutant for 2019 and 2020. Figure 8 (left) shows the options to select the type of air pollutant. Figure 8 (right) presents the spatial distribution of average selected pollutants for different pollution recording stations of selected pollutants and for a selected year. Figure 8 (right) presents the spatial distribution of average annual prediction for the average annual concentration of a selected pollutant and for a selected year.

3.5. Socioeconomic Variables

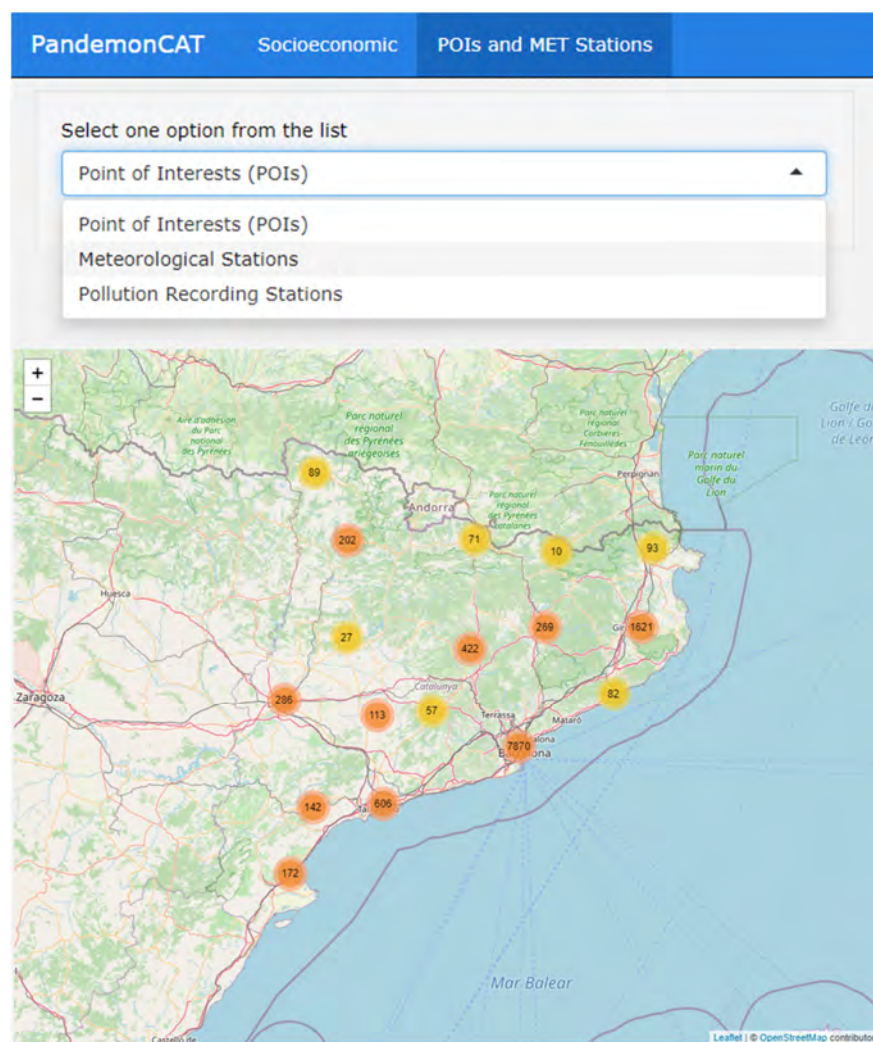
The first section of this app reports the spatial distribution of variables such as income per capita, percentage of population (65 or more), percentage of population (0–25), percentage of foreign population, and unemployment rate (Figure 9). The spatial resolution of the variables is ABSs of Catalonia and it reflects substantial inequalities across regions.



**Figure 9.** Dashboard of PandemonCAT displaying spatial distribution of socioeconomic variables.

The second map displays the geographic locations of points of interest (POIs), weather stations and pollution recording stations in an interactive map interface (Figure 10). POIs are potential COVID-19 contamination hotspots like restaurants, night clubs, bars, and other similar public aggregation hotspots. Figure 9 displays the spatial distribution of total population in the 343 ABSs of Catalonia. The map depicts a wide heterogeneity of population in

The second map displays the geographic locations of points of interest (POIs), weather stations and pollution recording stations in an interactive map interface (Figure 10). POIs are potential COVID-19 contamination hotspots like restaurants, night clubs, bars, and other similar public aggregation hotspots. Figure 9 displays the spatial distribution of total population in the 343 ABSs of Catalonia. The map depicts a wide heterogeneity of population in the entire region. On the other hand, Figure 10 displays clustered POIs distributed over the entire study region.



**Figure 10.** Dashboard of PandemonCAT displaying spatial distribution of POIs. The number indicated in the circle represent the number of units in the cluster.

### 3.6. Public Mobility

In Catalonia during late February 2020 and early March 2020, there were no strong actions or precautions taken by the government warning of the seriousness of the pandemic. Community transmission started in mid-February and by 13 March, confirmed cases of COVID-19 had been recorded in almost all the 343 ABSs of the region. This led to the implementation of nation-wide lockdown in Spain on 14 March 2020 which was also effective in Catalonia. The lockdown continued for more than 3 months. In the beginning of June 2020, with daily decreasing trends in the number of infections and deaths, the government started lifting some restrictions and relaxing the lockdown to some extent. Leveraging the recently available public mobility open data, in the Public Mobility tab we are able to provide exact figures for every municipality in Catalonia, including long trips and shorter ones (such as the daily commute to work).

This new dimension is key for understanding and quantifying the impact of non-pharmaceutical interventions (NPIs) throughout the pandemic. Never before has an open dataset provided so much insight into the daily mobility of the population, and thanks to it, we can easily spot the stay-at-home period of March–April 2020 in comparison with the following months, proving once again the high degree of compliance with that specific

Leveraging the recently available public mobility open data, in the Public Mobility tab we are able to provide exact figures for every municipality in Catalonia, including long trips and shorter ones (such as the daily commute to work).

This new dimension is key for understanding and quantifying the impact of non-pharmaceutical interventions (NPIs) throughout the pandemic. Never before has an open dataset provided so much insight into the daily mobility of the population, and thanks to it, we can easily spot the stay-at-home period of March–April 2020 in comparison with the following months. The dynamic flow map in Figure 11 represents average inter- and intra-ABS public mobility during October 2020. The user has the option to select the months from March to November 2020 which covers all the phases—before, during, and after the lockdown period.

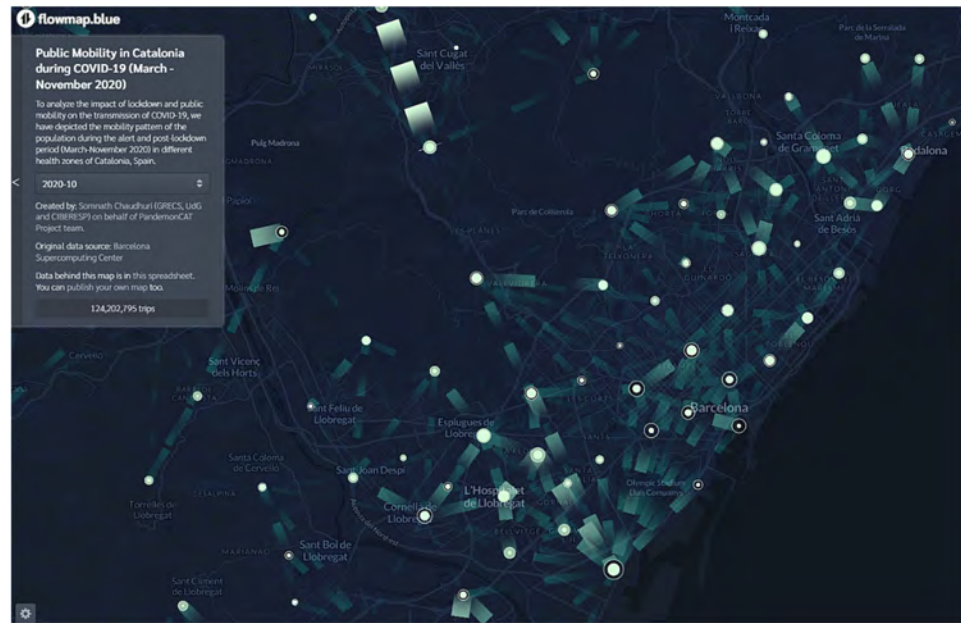


Figure 11. Dashboard of PandemonCAT to visualize inter- and intra-ABS monthly public mobility.

### 3.7. Smoothing

We show the weekly smoothed standardized incidence (positive cases) and mortality rates by COVID-19 on maps of Catalonia by ABS and region (Figure 12). These smoothed rates allow the user to glimpse the existence of geographic patterns in incidence and mortality.

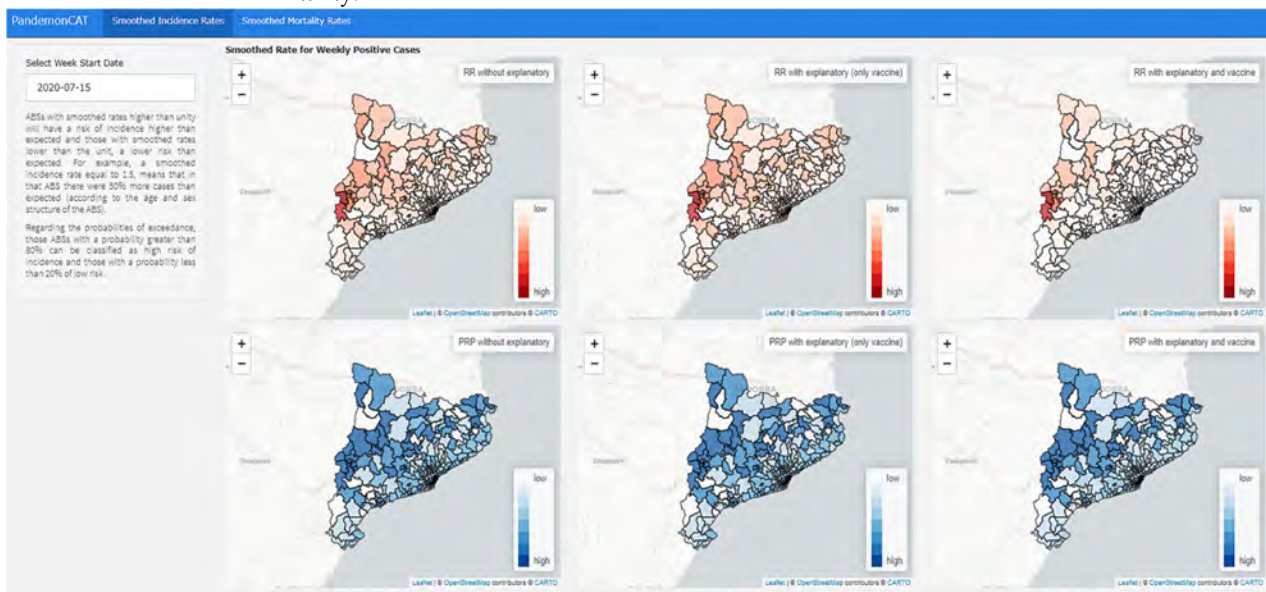


Figure 12. Dashboard of PandemonCAT that allows analysis of weekly smoothed standardized incidence (positive cases) of COVID-19.

Those ABSs with smoothed rates higher than unity have a risk of incidence higher than expected and those with smoothed rates lower than unity have a risk of incidence lower than expected.

To help evaluate the existence of agglomerations of excess cases (i.e., clusters), we also show the ‘exceedance probability’, which is the probability that the smoothed rate is above 1. Those ABSs or regions with a probability greater than 80% can be classified as high risk (of incidence or mortality) and those with a probability less than 20% of low risk.

We show smoothed rates without including explanatory variables and adjusting for

To help evaluate the existence of agglomerations of excess cases (i.e., clusters), we also show the ‘exceedance probability’, which is the probability that the smoothed rate is above 1. Those ABSs or regions with a probability greater than 80% can be classified as high risk (of incidence or mortality) and those with a probability less than 20% of low risk.

We show smoothed rates without including explanatory variables and adjusting for various explanatory variables. In the latter case, including only vaccination (weekly accumulated percentage of people vaccinated with the two doses) and including vaccination, socioeconomic and environmental variables (meteorological and air pollutants), and mobility.

#### 3.7.1. Weekly Positive Cases

Those ABSs with smoothed rates higher than unity have a risk of incidence higher than expected and those with smoothed rates lower than the unit, a lower risk than expected. For example, a smoothed incidence rate equal to 1.5 means that in that ABS there were 50% more cases than expected (according to the age and sex structure of the ABS).

Regarding the probabilities of exceedance, those ABSs with a probability greater than 80% can be classified as high risk of incidence and those with a probability less than 20% of low risk.

#### 3.7.2. Weekly Deaths

The smoothed rates allow the user to view geographic patterns in mortality. For example, regions with a smoothed mortality rate equal to 1.2 means that the number of deaths from COVID-19 was 20% higher than expected.

Regarding the probabilities of exceedance, those regions with a probability greater than 80% can be classified as high risk of mortality and those with a probability less than 20%, as low risk.

### 4. Discussion

It is important for the health professionals and policymakers to have access to the most relevant, reliable, and real-time information that can be used in their day-to-day tasks of COVID-19 research and analysis.

In this context, all apps referred to in Section 1 produce daily updated COVID-19 data visualizations and analyses. The results of our current study illustrate that, PandemonCAT is a novel interactive web application which acts as a collective monitoring package for daily COVID-19 updates along with regional vaccination flow and several environmental and socioeconomic variables that could have some influence on the spatiotemporal dynamics of the pandemic. The app explores variables such as meteorological and air pollution variables, population by age group, unemployment rate, income per capita, and others in a geospatial interface. It also provides an interactive interface to visualize public mobility before, during, and after the lockdown phases in the community. The visualization of these variables could have some influence on the spatiotemporal dynamics of the pandemic.

On the other hand, linking the pandemic severity with environmental factors such as air pollution, we find the article from Martorell-Marugán et al. which may be the closest to our study in that regard [49]. Combining great visualization capabilities with sound and rigorous statistical analysis of the possible contributing factors of the pandemic goes hand in hand with our aim. We tried to take it a step forward, albeit limiting it to Catalonia due to data availability, and relate many other relevant variables that enrich the overall picture. Related to the social contact data sharing initiative, an interactive tool (SOCRATES) to assess mitigation strategies for COVID-19 was developed by Willem et al. [50]. It implements location-specific physical distancing measures (e.g., schools or at work) and captures their impact on the transmission dynamics.

To the best of our knowledge, this is the most complete, open, and free source of public health information regarding the pandemic and its possible contributing factors in Catalonia. Many other applications have been developed throughout this period; however,

none offer the end user the option to explore with a single click other key variables such as social mobility, meteorological and air pollution statistics. Another unique aspect of the current application is the level of spatial resolution. Limited online applications provide dynamic spatial resolution up to the level of ABSs. In addition, it is the only interactive application which provides a visualization of human mobility and highlights its influence on the transmission of COVID-19 for individual ABSs and regions of Catalonia. To explore the link between COVID-19 transmission and air pollution, PandemonCAT is one of the few online applications to provide spatial predictions of pollutants for the entire time frame which covers all the phases before, during, and after the lockdown periods. We also report that, to the best of our knowledge, no other application provides dynamic, smoothed, standardized COVID-19 infected cases and mortality rates. These smoothed rates allow the user to explore the existence of geographic patterns in incidence and mortality rates. Moreover, the datasets used as input for the application are collected from official open data portals. This makes the data collection process smoother and on the other hand allows additional individuals to analyze and interpret the data, making it transparent and reproducible. The entire application can be easily replicated using open data from any other region or country.

Although the sources of our data are very complete and informative, there are several limitations, which are common to many countries. First and foremost, during the first COVID-19 wave, testing for the disease was very limited, making it hard to estimate the true number of infected during that period. Data on hospitalizations and deaths was also constrained by the testing capacity, but it did so to a lesser degree since they were greatly prioritized. This fact is emphasized as total COVID-19 deaths in Catalonia account for almost all the excess deaths for the first epidemic wave.

However, this limitation became virtually eliminated during the month of July 2020, when the testing capacity was greatly expanded. This is reflected by much lower positive rates, even when the next several waves were at their highest, and by a much lower share of cases that ended up in hospitalization or death.

Another limitation is the lack of distinction between the type of vaccine being administered to the population. We do have information available regarding the total numbers administered for each manufacturer (either Pfizer/BioNTech, ModeRNA, Janssen or AstraZeneca), but given that it's likely that each vaccine offers a different protection profile, especially with the expected waning immunity, we may have a certain degree of heterogeneity among the fully vaccinated cohort.

## 5. Conclusions

The complex nature of the COVID-19 epidemic and its dynamics of spread and transmission in the global population demands that researchers and health professionals embrace a multidisciplinary approach in addressing the challenges raised by the pandemic. Thus, it is essential to have efficient web-based applications or, portals that can provide the most relevant, reliable, and up-to-date information with a single click. In this context, the dynamic web-application we have developed offers a tool to scientists and others in the broader community to visualize the spatiotemporal trends of COVID-19 and enables comparisons at the ABS level in Catalonia, Spain. The features we incorporated in our open-source web application provide a comprehensive picture of public mobility, environmental, and other socioeconomic aspects that may have an impact on the spatiotemporal dynamics of the pandemic. The visualization of spatial predictions of pollutants related to COVID-19 is another novel feature of PandemonCAT. Finally, the interactive functionality to depict dynamic, smoothed, standardized COVID-19 infected cases and mortality rates help in providing an insight for the policymakers in developing public health strategies and control measures related to the ongoing pandemic.

**Author Contributions:** M.A.B. and M.S. had the original idea for the paper. M.A.B. and M.S. designed the study. The bibliographic search and the writing of the introduction were carried out by all the authors. The methods and statistical analysis were chosen and performed by S.C. and G.G.-A.

S.C. and G.G.-A. created the tables and figures. All authors wrote the results and the discussion. The writing and final editing was performed by all authors. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially financed by the SUPERA COVID-19 Fund, from SAUN: Santander Universidades, CRUE and CSIC, and by the COVID-19 Competitive Grant Program from Pfizer Global Medical Grants. It also received funding, in the form of a free transfer of data, from the AEMET. The funding sources did not participate in the design or conduct of the study, the collection, management, analysis, or interpretation of the data, or the preparation, review, or approval of the manuscript.

**Data Availability Statement:** We used open data with free access. All data sources are referred in Section 2.1. Code will be available at [www.researchprojects.es](http://www.researchprojects.es).

**Acknowledgments:** This study was carried out within the ‘Cohort-Real World Data’ subprogram of CIBER of Epidemiology and Public Health (CIBERESP). We appreciate the comments of three anonymous reviewers of a previous version of this work who, without doubt, helped us to improve our work. The usual disclaimer applies.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

**Table A1.** Description of PandemonCAT component apps.

App Name	Components	Description
COVID-19 Dynamics	Weekly Cases per 100,000 inhabitants, Empiric 7-day Rt, Weekly Tests per 100,000 inhabitants, Weekly Deaths per 100,000 inhabitants, Currently hospitalized per 100,000 inhabitants, Currently ICU per 100,000 inhabitants (Both spatial and linear plots)	Displays the main parameters to track the COVID-19 pandemic in Catalonia.
Vaccination Flow	Cumulative vaccination percentage Weekly vaccination percentage (Both spatial and linear plots and for different age groups)	Displays the progress being made in the vaccination campaign against COVID-19 in Catalonia.
Meteorological	Daily records of temperature, atmospheric pressure, precipitation, relative humidity, solar irradiance, wind velocity	Displays information of meteorological components of individual weather stations for 2020.
Air Pollutants	Daily records of concentration for ozone, carbon monoxide, nitrogen dioxide, PM10 Daily records of concentration for ozone, carbon monoxide, nitrogen dioxide, PM10 for the years 2011 to 2020, both included.	Displays the concentration of ozone, carbon monoxide, nitrogen dioxide and PM10 available from several pollution monitoring stations for 2020. Second map displays the spatial prediction for each Basic Health Areas (ABSs) for the above-mentioned variables for the years 2011 to 2020, both included.

Table A1. Cont.

App Name	Components	Description
Socioeconomic	Total population Income per capita, Percentage of population (65 or more), Percentage of population (0–25), Percentage of foreign population, Unemployment rate Locations of Point of Interest, Weather stations, Pollution monitoring stations	Displays spatial distribution of socioeconomic components like population, unemployment rate etc. for 2019. Displays “Point of Interest” defined as restaurants, night clubs, bars, among others, whereas MET stations locate all weather and pollution monitoring stations across the region.
Public Mobility	Monthly intra and inter ABSs’ public mobility.	Displays inter and intra ABSs’ monthly public mobility before, during and after lockdown phases in Catalonia.
Smoothing COVID-19 Outcomes	Weekly Smoothed incidence rates Smoothed mortality rates	Displays weekly smoothed standardized incidence (positive cases) and mortality rates by COVID-19 by ABSs and comarcas of Catalonia.
Smoothing Methods	Smoothing methods Smoothing standardized incidence and mortality rates	Reports about the Bayesian spatiotemporal model used in the smoothing process.

## References

1. Secretaría General de Sanidad; Dirección General de Salud Pública; Calidad e Innovación; Ministerio de Sanidad; Gobierno de España. Coronavirus Disease (COVID-19). 7 January 2022. Available online: <https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/situacionActual.htm> (accessed on 10 January 2022). (In Spanish)
2. INE. Instituto Nacional de Estadística. Continuous Register Statistics 2022. Available online: [https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica\\_C&cid=1254736177012&menu=resultados&secc=1254736195461&idp=1254734710990#!tabs-1254736195557](https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736177012&menu=resultados&secc=1254736195461&idp=1254734710990#!tabs-1254736195557) (accessed on 10 January 2022).
3. DESCAT. Statistical Institute of Catalonia. Available online: <https://www.idescat.cat/?lang=en> (accessed on 10 January 2022).
4. Atenció Primària Girona; Institut Català de la Salut. Basic Health Areas (ABS). Available online: <http://www.icsgirona.cat/ca/cointingut/primaria/370> (accessed on 10 January 2022). (In Catalan)
5. Mukhtar, H.; Ahmad, H.F.; Khan, M.Z.; Ullah, N. Analysis and evaluation of COVID-19 web applications for health professionals: Challenges and opportunities. *Healthcare* **2020**, *8*, 466. [[CrossRef](#)] [[PubMed](#)]
6. Fernández-Lozano, C.; Cedrón, F. Shiny dashboard for monitoring the COVID-19 pandemic in Spain. *Proceedings* **2020**, *54*, 54023. [[CrossRef](#)]
7. Galván-Tejada, C.E.; Zanella-Calzada, L.A.; Villagrana-Bañuelos, K.E.; Moreno-Báez, A.; Luna-García, H.; Celaya-Padilla, J.M.; Galván-Tejada, J.I.; Gamboa-Rosales, H. Demographic and comorbidities data description of population in Mexico with SARS-CoV-2 infected patients (COVID19): An online tool analysis. *Int J. Environ. Res. Public Health* **2020**, *17*, 5173. [[CrossRef](#)] [[PubMed](#)]
8. Tebé, C.; Valls, J.; Satorra, P.; Tobías, A. COVID19-world: A shiny application to perform comprehensive country-specific data visualization for SARS-CoV-2 epidemic. *BMC Med. Res. Methodol* **2020**, *20*, 235. [[CrossRef](#)] [[PubMed](#)]
9. Wissel, B.D.; Van Camp, P.J.; Kouril, M.; Weis, C.; Glauser, T.A.; White, P.S.; Kohane, I.S.; Dexheimer, J.W. An interactive online dashboard for tracking COVID-19 in U.S. counties, cities, and states in real time. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1121–1125. [[CrossRef](#)]
10. Revell, L.J. COVID-19. Explorer: A web application and R package to explore United States COVID-19 data. *PeerJ* **2021**, *9*, e11489. [[CrossRef](#)]
11. Zohner, Y.E.; Morris, J.S. COVID-track: World and USA SARS-COV-2 testing and COVID-19 tracking. *BioData Min.* **2021**, *14*, 4. [[CrossRef](#)]
12. Almeida, J.S.; Shiels, M.; Bhawsar, P.; Patel, B.; Nemeth, E.; Moffit, R.; García-Closas, M.; Freedman, N.; Berrington, A. Mortality tracker: The COVID-19 case for real time web APIs as epidemiology commons. *Bioinformatics* **2021**, *37*, 2073–2074. [[CrossRef](#)]
13. Tobías, A.; Valls, J.; Satorra, P.; Tebé, C. COVID-19-Tracker: A shiny app to perform comprehensive data visualisation for SARS-CoV-2 epidemic in Spain. *medRxiv* **2020**, *35*, 99–101. [[CrossRef](#)]
14. Salehi, M.; Arashi, M.; Bekker, A.; Ferreira, J.; Chen, D.G.; Esmaili, F.; Frances, M. A synergetic R-Shiny portal for modeling and tracking of COVID-19 data. *Front. Public Health* **2021**, *8*, 623624. [[CrossRef](#)]



15. Berry, I.; Soucy, J.-P.R.; Tuite, A.; Fisman, D.; COVID-19 Canada Open DataWorking Group. Open Access Epidemiologic Data and an Interactive Dashboard to Monitor the COVID-19 Outbreak in Canada. *Can. Med. Assoc. J.* **2020**, *192*, E420. [CrossRef]
16. Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534. [CrossRef]
17. Florez, H.; Singh, S. Online dashboard and data analysis approach for assessing COVID-19 case and death data. *F1000Research* **2020**, *9*, 570. [CrossRef] [PubMed]
18. Marques da Costa, N.; Mileu, N.; Alves, A. Dashboard COMPRIME\_COMPRI\_MOv: Multiscalar Spatio-Temporal Monitoring of the COVID-19 Pandemic in Portugal. *Future Internet* **2021**, *13*, 45. [CrossRef]
19. Chang, W.; Cheng, J.; Allaire, J.J.; Sievert, C.; Schloerke, B.; Xie, Y.; Allen, J.; McPherson, J.; Dipert, A.; Borges, B. Shiny: Web Application Framework for R. 2021. R Package Version 1.6.0. Available online: <https://CRAN.R-project.org/package=shiny> (accessed on 10 January 2022).
20. Wickham, H.; François, R.; Henry, L. Dplyr: A Grammar of Data Manipulation. 2018. Available online: <https://CRAN.R-project.org/package=dplyr> (accessed on 10 January 2022).
21. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.; François, R.; Grolemond, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]
22. Bivand, R.; Keitt, T.; Rowlingson, B. Rgdal: Bindings for the ‘Geospatial’ Data. Abstraction Library. 2021. R Package Version 1.5-23. Available online: <https://CRAN.R-project.org/package=rgdal> (accessed on 10 January 2022).
23. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *R J.* **2018**, *10*, 439–446. [CrossRef]
24. Hijmans, R.J. Geographic Data Analysis and Modeling. 2021. R Package Raster Version 3.5-2. Available online: <https://CRAN.R-project.org/package=raster> (accessed on 10 January 2022).
25. Bivand, R.; Lewin-Koh, N. Maptools: Tools for Reading and Handling Spatial Objects. 2021. R Package Version 1.1-2. Available online: <https://CRAN.R-project.org/package=maptools> (accessed on 10 January 2022).
26. Flowmap.blue. Available online: <https://flowmap.blue> (accessed on 10 January 2022).
27. Sievert, C. Interactive Web-Based Data Visualization with R, Plotly, and Shiny. Chapman and Hall/CRC. 2020. Available online: <https://plotly-r.com> (accessed on 10 January 2022).
28. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; Available online: <https://ggplot2.tidyverse.org> (accessed on 10 January 2022).
29. Cheng, J.; Karambelkar, B.; Xie, Y. Leaflet: Create Interactive Web Maps with the JavaScript ‘Leaflet’ Library. 2021. R Package Version 2.0.4.1. Available online: <https://CRAN.R-project.org/package=leaflet> (accessed on 10 January 2022).
30. Appelhans, T.; Detsch, F. Leafpop: Include Tables, Images and Graphs in Leaflet Pop-Ups. 2021. R Package Version 0.1.0. Available online: <https://CRAN.R-project.org/package=leafpop> (accessed on 20 January 2022).
31. Flexdashboard. Available online: <https://pkgs.rstudio.com/flexdashboard/> (accessed on 10 January 2022).
32. Xie, Y.; Dervieux, C.; Riederer, E. R Markdown Cookbook. 2022. Available online: <https://bookdown.org/yihui/rmarkdown-cookbook> (accessed on 10 January 2022).
33. Team R.C. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: <https://www.r-project.org> (accessed on 10 January 2022).
34. Shinyapps.io. Available online: <https://www.shinyapps.io> (accessed on 10 January 2022).
35. RStudio, Open Source & Professional Software for Data Science Teams. Available online: <https://www.rstudio.com> (accessed on 10 January 2022).
36. UNDP (United Nations Development Program). Human Development Report 2019. Available online: <http://hdr.undp.org/en/content/human-development-report-2019-readers-guide> (accessed on 10 January 2022).
37. INE (Instituto Nacional de Estadística). Household Income Distribution Atlas 2022. Available online: [https://www.ine.es/en/experimental/atlas/exp\\_atlas\\_tab\\_en.htm](https://www.ine.es/en/experimental/atlas/exp_atlas_tab_en.htm) (accessed on 10 January 2022).
38. INE (Instituto Nacional de Estadística). Indicators for Census Tracks. *2011 Spanish Census of Population and Housing*. Available online: [http://www.ine.es/censos2011\\_datos/cen11\\_datos\\_resultados\\_seccen.htm](http://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm) (accessed on 10 January 2022). (In Spanish)
39. Saez, M.; Barceló, M.A. Spatial prediction of air pollution levels using a hierarchical Bayesian spatio-temporal model in Catalonia, Spain. *Environ. Model. Softw.* **2022**, *151*, 105369. [CrossRef]
40. Cameletti, M.; Lindgren, F.; Simpson, D.; Rue, H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *ASta Adv. Stat. Anal.* **2013**, *97*, 109–131. [CrossRef]
41. Lindgren, F.K.; Rue, H.; Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2011**, *73*, 423–498. [CrossRef]
42. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2009**, *71*, 319–392. [CrossRef]
43. Rue, H.; Riebler, A.; Sørbye, H.; Illian, J.B.; Simpson, D.P.; Lindgren, F.K. Bayesian computing with INLA: A review. *Annu. Rev. Stat. Its Appl.* **2017**, *4*, 395–421. [CrossRef]
44. Departament de Territori i Sostenibilitat, Generalitat de Catalunya. Available online: <https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu> (accessed on 10 January 2022).

45. Barceló, M.A.; Saez, M.; Cano-Serral, G.; Martínez-Beneito, M.A.; Martínez, J.M.; Borrell, C.; Ocaña-Riola, R.; Montoya, I.; Calvo, M.; López-Abente, G.; et al. Methods to smooth mortality indicators: Application to analysis of inequalities (the MEDEA project). *Gac. Sanit.* **2008**, *22*, 596–608. (In Spanish) [[CrossRef](#)] [[PubMed](#)]
46. Barceló, M.A. Tutorials. Talks & Teaching. Available online: <https://www.antonibarcelo.com/en> (accessed on 11 April 2022).
47. Diggle, P.J.; Moraga, P.; Rowlingson, B.; Taylor, B.M. Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Stat. Sci.* **2013**, *28*, 542–563. [[CrossRef](#)]
48. Saez, M.; Tobias, A.; Barceló, M.A. Effects of long-term exposure to air pollutants on the spatial spread of COVID-19 in Catalonia, Spain. *Environ. Res.* **2020**, *191*, 110177. [[CrossRef](#)] [[PubMed](#)]
49. Martorell-Marugán, J.; Villatoro-García, J.A.; García-Moreno, A.; López-Domínguez, R.; Requena, F.; Merelo, J.J.; Lacasaña, M.; Luna, J.D.; Díaz-Mochón, J.J.; Lorente, J.A.; et al. DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain. *Sci. Total Environ.* **2021**, *750*, 141424. [[CrossRef](#)] [[PubMed](#)]
50. Willem, L.; Van Hoang, T.; Funk, S.; Coletti, P.; Beutels, P.; Hens, N. SOCRATES: An online tool leveraging a social contact data sharing initiative to assess mitigation strategies for COVID-19. *BMC Res. Notes* **2020**, *13*, 293. [[CrossRef](#)]

## **5.2 Article 2: Modeling Traffic Accidents in London, UK**

### **Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation**

Somnath Chaudhuri<sup>1,2</sup>, Pablo Juan<sup>1,3</sup> and Jorge Mateu<sup>4</sup>

1. Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, 17004 Girona, Spain.
2. CIBER of Epidemiology and Public Health (CIBERESP), 17003 Madrid, Spain
3. IMAC, University Jaume I, Castellón, Spain
4. Department of Mathematics, University Jaume I, Castellón, Spain



**Journal of Applied Statistics, 1-22**

Impact Factor (2021): 1.416

Statistics & Probability, Position 73 out of 125 (Q3)



# Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation

Somnath Chaudhuri <sup>a</sup>, Pablo Juan <sup>a,b</sup> and Jorge Mateu <sup>c</sup>

<sup>a</sup>Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Girona, Spain; <sup>b</sup>IMAC, University Jaume I, Castellón, Spain; <sup>c</sup>Department of Mathematics, University Jaume I, Castellón, Spain

## ABSTRACT

Traffic deaths and injuries are one of the major global public health concerns. The present study considers accident records in an urban environment to explore and analyze spatial and temporal in the incidence of road traffic accidents. We propose a spatio-temporal model to provide predictions of the number of traffic collisions on any given road segment, to further generate a risk map of the entire road network. A Bayesian methodology using Integrated nested Laplace approximations with stochastic partial differential equations (SPDE) has been applied in the modeling process. As a novelty, we have introduced SPDE network triangulation to estimate the spatial autocorrelation restricted to the linear network. The resulting risk maps provide information to identify safe routes between source and destination points, and can be useful for accident prevention and multi-disciplinary road safety measures.

## ARTICLE HISTORY


Received 22 December 2021  
Accepted 9 July 2022


## KEYWORDS

INLA; network triangulation; Poisson hurdle model; SPDE; traffic risk mapping

## 1. Introduction

Road traffic collisions is one of the serious issues in the modern world. According to 2018 global status report on road safety by the World Health Organisation, approximately 1.35 million people die each year as a result of traffic collisions [74]. The rate of occurrence along with severity of traffic crashes are the principal indicators of urban road safety measures [74]. Literature suggests that factors such as road infrastructure or types of roads (highways, double or, single carriage tracks) play a vital role in road safety measures [18]. Indeed, uncontrolled vehicle speed or street junctions without traffic signals increase accident risk [10], but temporal factors (time of the day or weekend nights) also act as decisive aspects in the count and impact of accidents [20,33]. Identifying such significant elements has been a central focus of research in the domain of road safety. Available map applications offered by larger corporations, such as Google Maps or collaborative geospatial projects (for example, OpenStreetMap (OSM)) can provide information about the fastest (shortest) route from source to destination points. The existing applications can suggest,

**CONTACT** Pablo Juan  [juan@uji.es](mailto:juan@uji.es)

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2022.2104822>

however, the shortest route without considering likely risk factors. Multi-disciplinary predictor aspects are not implemented in most of these applications. According to Williamson and Feyer [76], a particular road can be safe during mid-day, but the same road might not be safe during office hours. Relevant spatio-temporal factors play a significant role in identifying safe roads [52]. Traffic components such as street light, road type, or speed limits act as significant factors in determining safe routes [12,44]. Thus, a multi-disciplinary approach is essential to explore spatio-temporal effects on road collisions. Identifying significant components [19,60] while performing spatio-temporal modeling of traffic accidents [30,78] have gained an increasing interest in the domain of road safety management. Research works by [5,6,44] made notable contributions in identifying significant factors influencing traffic collisions. Bhawkar [8] explored and analyzed the leading factors causing road accidents on the streets of UK. Shahid *et al.* [63] mentioned that the causes of traffic collisions can be broadly classified into spatial and temporal components. A series of studies [3,20,26,62] analyze historic data to identify risk factors and assess likelihoods of crash-related events to categorize spatio-temporal factors affecting traffic accidents. These factors are considered as significant predictors in statistical analysis and prediction modeling.

Several statistical techniques such as Poisson model variations [13,37,42,49], negative binomial error structure [53], logistic [28] and linear regressions [1] have been applied to analyze spatial variability of traffic accidents. In this regard, Wang *et al.* [72] while analyzing factors influencing traffic accident frequencies on urban roads, mentioned that accidents occurring at different locations are related. It supports spatial autocorrelation of traffic accident events. Spatial methods are able to incorporate geographical correlation in the model fitting process and, in most of the cases, spatial methods outperform the non-spatial models [23,77]. In this line, a number of research works [27,29,30,35] suggest that stochastic spatial processes are one of the most appealing analytical tools to analyze the spatial and spatio-temporal distribution of traffic collisions. Karaganis and Mimis [29] used spatial point processes to evaluate the probability of traffic accident occurrence on the national roads of Greece. In this context, statistical inference comes along with Bayesian methodology. Cantillo *et al.* [12] used a combined GIS-empirical Bayesian approach in modeling traffic accidents on the urban roads of Colombia. A space-time multivariate Bayesian model was designed by Boulieri *et al.* [21] used Bayesian spatial modeling with INLA in predicting road traffic accidents based on unmeasured information at road segment levels. The use of INLA-SPDE for spatial data is now quite well established in a number of disciplines with a large number of contributions [7,25,70] and in particular the references therein. However, in the context to traffic accident event modeling, there are limited contributions implementing Bayesian methodology with INLA-SPDE approach. In addition, if we consider events with a network support, see [14,45,46] for a nice overview of spatial and spatio-temporal point pattern analysis on linear networks, then Bayesian methodology on road networks using INLA-SPDE is even far under explored.

The aim of this paper is two-fold. On one side we provide a modeling framework to explore and analyze the spatial and temporal variation in the incidence of road traffic accidents on individual road segments. The second aim roots in providing an advanced and realistic computational strategy to create the spatial triangulation restricted only onto the network topology. In this context, we propose the novel concept of multi-disciplinary road-safety analysis by introducing spatio-temporal risk modeling of traffic accidents using

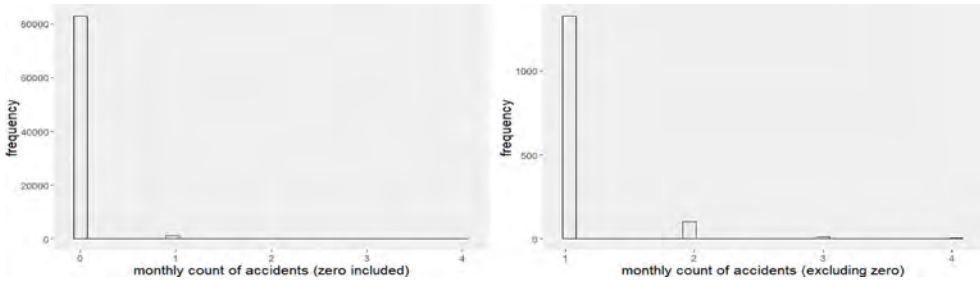
Bayesian methodology restricted entirely onto the road network. Our model acts as a comprehensive scoring system that can predict a risk index over individual road segments generating a categorized risk map of the entire road network. The study is conducted on five years road traffic accidents data from the city of London, UK. R programming language (version R 4.0.4) has been used for statistical computing and graphical analysis. All computations are conducted on a quad-core Intel i7-4790 (3.60 GHz) processor with 32 GB (DDR3-1600) RAM.

The rest of the paper is organized as follows. Section 2 presents the OSM street network data and provides some insights of the spatial distribution of traffic accidents in the city of London, UK. A description of the spatio-temporal modeling framework comes in Section 3. The design of the risk map algorithm is discussed in Section 3.2. Section 4 is devoted to present the results of model prediction and risk map analysis. Some discussion and concluding remarks come in Section 5.

## 2. Data settings

The Department for Transport of the Government of UK publishes road casualty statistics twice a year. Detailed data about the circumstances of road accidents on public roads reported to the police, and the corresponding casualties, are recorded using the STATS19 accident reporting form. The complete data set since 1979 is available in the UK Government open data repository [69]. The data is free and available under the *Open Government Licence v3.0* for public sector information. The dataset used in this paper contains detailed information of traffic accidents for five years, from January 2013 to December 2017, that have occurred in the city of London, UK. The city has an area of 2.90 km<sup>2</sup>, comprises six Lower Layer Super Output Area (LSOA) with an approximate population of around 90,000 citizens. The area is an important local district that contains the historic center and the primary Central Business District (CBD) of London.

According to Prasannakumar [52] the number of traffic collisions in each road segment plays a key role in designing predictive models that can reflect the influence of spatio-temporal factors on traffic accidents. The original traffic accidents dataset retrieved from [69] has records of daily accidents with geographical coordinates of individual occurrence. But one of the principal objectives of the current study is to measure the risk factor of individual road segments in the study area. As a result, we have applied up-scaling methods on both temporal and spatial resolutions. To identify the risk status of respective road segments rather than checking individual locations of accidents, the spatial resolution has been up-scaled to road networks, and the temporal unit is considered as month, to avoid having extreme number of zero-counts per segment. Thus, our target variable is the total number of accidents occurring in each road segment per month, from 2013 to 2017. An individual year will have  $12 \times 1406 = 16872$  events, where 1406 is the number of road segments in the entire study area. This results in  $5 \times 16872 = 84360$  events for five years of the study period on all road segments. We note that in 98% of the cases, we have no monthly traffic accidents on any road segment, and only 2% shows monthly accident records ranging from 1 to 4. Figure 1(left) illustrates the frequency distribution of instances with no accidents (depicted as zero) and more than one accident, and Figure 1(right) depicts the frequency distribution of only non-zero instances.



**Figure 1.** Frequency distribution of monthly instances with (*left*) no accidents (depicted as zero) and other values, and (*right*) only non-zero instances on all road segments in the study area.



**Figure 2.** OSM street network with locations of traffic accident (2013–2015) highlighted in gray.

The road network is accessed from OSM repository using R package *osmdata* [51]. OSM data is free and licensed under the *Open Data Commons Open Database License (ODbL)* by the OpenStreetMap Foundation. The OSM street network is illustrated in Figure 2, noting that OSM highway categories such as unclassified, bus\_guideway, raceway, path and bridleway are not included. Figure 2(right) also depicts individual accident locations (highlighted in red) over 1406 road segments in the OSM network.

We report that in the model fitting process we have used three covariates such as road type, road surface and months of a year. According to Transport for London [69], road surface has five unique categories such as dry, wet, snow, frost and flood (where surface water is over 3 cm deep). On the other hand, road types are roundabout, one way street, dual carriageway, single carriageway and slip road. The variable month ranges from 1 to

12. All three covariates are used as factors in the model. We note that for model fitting we have used traffic accident records for three years (January 2013 to December 2015) have been used while the records of following two years (January 2016 to December 2017) have been used for prediction purposes.

### 3. Spatio-temporal modeling

Random spatial events, such as traffic accidents, form irregularly scattered point patterns over regions of interest. In these cases, spatio-temporal point process models are useful tools to perform focused statistical analysis [27,29,35]. Moreover, we can consider that these events exist on a linear network, and we can find recent literature [21,45,46] on spatio-temporal point processes over networks that are able to identify spatial auto-correlations and interactions between points in the pattern. In this context, it is shown that the occurrence of traffic accidents depend on spatio-temporal interacting and triggering factors [34].

By aggregating data from locations to counts of events per segment, we open the door to consider Poisson regression models in combination with a Bayesian framework for the prediction of traffic accidents on individual road segments. A Bayesian approach with Markov Chain Monte Carlo (MCMC) simulation methods can be used to fit generalized linear mixed models (GLMM) [75]. MCMC methods provide multivariate distributions that can estimate the joint posterior distribution. As mentioned in Section 1, for latent Gaussian models and models having a large number of geo-locations, the performance of MCMC methods drops substantially [57,66,68]. As an alternative and computationally faster solution, prediction of marginal distributions by using a Laplace approximation for integrals was introduced by Rue *et al.* [57] with the integrated nested Laplace approximation (INLA) method. It focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF) [56].

We indeed follow this approach combining a spatio-temporal Poisson regression method within a Bayesian framework using INLA and SPDE. In particular, let  $Y_{it}$  and  $E_{it}$  be the observed and expected number of road traffic accidents on the  $i$ -th road segment and at the  $t$ -th month,  $t = 1, \dots, T$ . We assume that conditional on the relative risk,  $\rho_{it}$ , the number of observed events follows a Poisson distribution

$$Y_{it} | \rho_{it} \sim Po(\lambda_{it} = E_{it} \rho_{it})$$

where the log-risk is modeled as

$$\log(\rho_{it}) = \beta_0 + Z_i^T \beta_i + \xi_i + \zeta_t + \epsilon_i + \delta_{it} \tag{1}$$

Here,  $\xi_i$  and  $\zeta_t$  account for the spatially and temporally structured random effects, respectively,  $\delta_{it}$  represents spatio-temporal interaction between the two structured effects, and  $\epsilon_i$  stands for an unstructured zero mean Gaussian random effect and logGamma precision parameters 0.5 and 0.01, defined as penalized complexity (PC) priors [65].  $Z_i$  represents the spatial covariates. We assigned a vague prior to the vector of coefficients  $\beta = (\beta_0, \dots, \beta_p)$  which is a zero mean Gaussian distribution with precision 0.001. All parameters associated to log-precisions are assigned inverse Gamma distributions with parameters equal to 1 and 0.00005. In the current study, we have chosen to provide default prior distributions for all parameters in R-INLA. These have been chosen partly based on priors



commonly used in the literature [9,41,47,58]. We report that our results are robust against other alternative priors, as we run several cases with different priors obtaining the same results.

Description of the dataset in Section 2 suggests the current model can have problems of instability, especially with spatial random effects, which would be exacerbated due to zero inflation. Apart from the baseline Poisson model, both zero-inflated Poisson (ZIP) and Poisson hurdle models can be formulated for zero inflated discrete distributions. They provide mixtures of a Poisson and Bernoulli probability mass function to allow more flexibility in modeling the probability of a zero outcome [2]. According to Lambert [31], ZIP models add additional probability mass to the outcome of zero. Poisson hurdle models, on the other hand, are characterized as pure mixtures of zero and non-zero outcomes [24,55,61]. In a ZIP model, the response variable is  $Y_{it} = 0$  with probability  $\pi$ , and  $Po(\lambda_{it})$  with probability  $1 - \pi$ . In particular,

$$Y_{it} = \begin{cases} 0 & \text{with prob } \pi + (1 - \pi)e^{-\lambda_{it}} \\ k & \text{with prob } (1 - \pi) \frac{\lambda_{it}^k e^{-\lambda_{it}}}{k!}, \quad k \geq 1 \end{cases}$$

On the other hand, a Poisson hurdle model indicates that  $Y_{it} = 0$  with probability  $\pi$ , and a truncated Poisson distribution with parameter  $\lambda_{it}$  with probability  $1 - \pi$ . Thus, we have

$$Y_{it} = \begin{cases} 0 & \text{with probability } \pi \\ k & \text{with probability } \frac{(1 - \pi)}{1 - e^{-\lambda}} \left( \frac{\lambda_{it}^k e^{-\lambda_{it}}}{k!} \right), \quad k \geq 1 \end{cases}$$

In the model fitting process, we have explored three different distributions discussed above to fit the model in Equation (1). To compute the joint posterior distribution of the model parameters, we use an INLA-SPDE method, as introduced by Lindgren *et al.* [32]. SPDE consists in representing a continuous spatial process, such a Gaussian field (GF), using a discretely indexed spatial random process such as a Gaussian Markov random field (GMRF). In particular, the spatial random process  $\xi$ , here represented by  $\xi(\cdot)$  to explicitly denote dependence on the spatial field, follows a zero-mean Gaussian process with Matérn covariance function represented as

$$Cov(\xi(x_i), \xi(x_j)) = \frac{\sigma^2}{2^{\nu-1} \Gamma(\nu)} (\kappa \|x_i - x_j\|)^{\nu} K_{\nu}(\kappa \|x_i - x_j\|) \quad (2)$$

where  $K_{\nu}(\cdot)$  is the modified Bessel function of second order, and  $\nu > 0$  and  $\kappa > 0$  are the smoothness and scaling parameters, respectively. INLA approach constructs a Matérn SPDE model, with spatial range  $r$  and standard deviation parameter  $\sigma$ .

The parameterized model we follow is of the form

$$(\kappa^2 - \Delta)^{(\alpha/2)} (\tau S(x)) = W(x)$$

where  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$  is the Laplacian operator,  $\alpha = (\nu + d/2)$  is the smoothness parameter,  $\tau$  is inversely proportional to  $\sigma$ ,  $W(x)$  is a spatial white noise and  $\kappa > 0$  is the scale parameter, related to range  $r$ , defined as the distance at which the spatial correlation becomes negligible. For each  $\nu$ , we have  $r = \sqrt{8\nu}/\kappa$ , with  $r$  corresponding to the distance

where the spatial correlation is close to 0.1. Note that we have  $d = 2$  for a two-dimensional process, and we fix  $\nu = 1$ , so that  $\alpha = 2$  in our case [9].

INLA-SPDE requires a triangulation or mesh structure to interpolate discrete event locations to estimate a continuous process in space [59]. We use centroids of each road segment as the target locations over which we build the mesh. A detailed description of building a Delaunay’s triangulation with emphasis on a network mesh is shown in Section 3.1. Centroids of individual road segments and triangulations in the mesh are used to generate the projection matrix. Now we use `inla.spde2.pcmatern` function from R-INLA package to build SPDE model and specify PC priors for the parameters of the Matérn field. The parameters `prior.range` and `prior.sigma` control the joint prior on range and standard deviation of the spatial field [64,65]. According to Bakka *et al.* [7], the range value is selected based on the spatial distribution of event locations in the study area. In the current study, due to the proximity of accident locations we have decided to use a prior  $P(r < 0.01) = 0.01$ , which means that the probability that the range is less than 10 meters is very small. Parameter  $\sigma$  denotes the variability of the data. We specify the prior for this parameter as  $P(\sigma > 1) = 0.01$ .

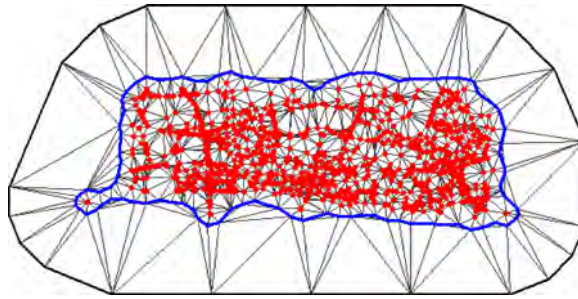
On the other hand, the temporal random effect ( $\zeta_t$ ) is assumed to be a smoothed function, in particular a random walk of order one (RW1) [57]. Using the specifications discussed above, we design a set of models for three distributions such as Poisson, Poisson hurdle and ZIP. Each of these models are explored having different combinations of three covariates (mentioned in Section 2) and several choices amongst PC priors and default priors for the parameters to create a SPDE model object in case of a Matérn field. Details of each model are shown in Table 1 in Appendix. As reported in Equation (1), we have also introduced a spatio-temporal interaction effect as an independent unobserved term for each combination of region and period ( $i, t$ ), thus without any structure  $\delta_{it} \sim Normal(0, 1/\tau_\delta)$ . However, if spatial and temporal main effects are present in the model, then this interaction only implies independence in the deviations from them. Note that it is a global space-time heterogeneity effect, and it is usually modeled as white noise [36]. See also Blangiardo and Cameletti [9]. Thus, a second set of models are designed using the three distribution types with all three covariates included in each case but with the choice of spatio-temporal interaction and PC priors. A summary of the considered competing models is depicted in Table 2 in Appendix.

As we have a battery of competing models, we compare them using the deviance information criterion (DIC) [67], which is a Bayesian model comparison criterion, represented as

$$DIC = \text{goodness of fit} + \text{complexity} = D(\bar{\theta}) + 2p_D$$

where  $D(\bar{\theta})$  is the deviance evaluated at the posterior mean of the parameters, and  $p_D$  denotes the *effective number of parameters*, which measures the complexity of the model [67]. When the model is true,  $D(\bar{\theta})$  should be approximately equal to the *effective degrees of freedom*,  $n - p_D$ . DIC may underpenalize complex models with many random effects.

An alternative is the Watanabe Akaike information criterion (WAIC) which follows a more strict Bayesian approach to construct a criterion [73]. Gelman *et al.* [22] claim that WAIC is more preferable over DIC. Likewise DIC, WAIC estimates the effective number of parameters to adjust over-fitting.  $pWAIC$  is similar to  $p_D$  in the original DIC. Gelman



**Figure 3.** Selected region mesh with non-convex hull boundary.

*et al.* [22] scales the WAIC of Watanabe [73] by a factor of 2 so that it is comparable to AIC and DIC. WAIC is then reported as  $-2(lppd - pWAIC)$  where  $lppd$  is the log pointwise predictive density and  $pWAIC$  is the effective number of parameters. Therefore, the lowest values of DIC and WAIC suggest the best model. A high number of parameters means more complexity. The best models are those with a high level of complexity and a high goodness-of-fit. In general, we choose that model showing lower DIC and WAIC.

### 3.1. SPDE triangulation design

Due to the densely distributed nature of the road segments in the study area, initially a continuous spatial structure is chosen for modeling, and triangulation is carried out on the entire study area. Triangle size is generated using a combination of maximum edge and cut-off. The size controls how precisely the equations will be tailored by the data. Using smaller triangles increases precision but also exponentially increases computational time [70]. The best fitting mesh should have enough vertices for effective prediction, but the number should be within a limit to have control over the computational time. Following this principle, a series of meshes with varied range in the number of vertices are created. Finally, the best fitting mesh without offset value and having non-convex hull boundary is selected. The number of vertices in the selected mesh is 1526. Figure 3 depicts the selected mesh with the locations of traffic accidents (in red) during the time period of January 2013 to December 2017.

**SPDE network triangulation:** The mesh created for the entire region can be used to fit INLA model in the study area. Prediction involves projecting fitted model into the mesh at precise spatial locations. However, while fitting the mesh (as depicted in Figure 3) a problem appears. The sampled traffic accidents are discrete spatial points located precisely on the road networks, but models fitted with a region mesh cover the entire study area. Therefore, the predicted locations of traffic accident can be placed in any area with or without road networks. It is not realistic that the model prediction provides results in locations without road network where there is no chance of traffic accidents to occur. Thus, the traditional methods of model prediction using a region mesh are not useful. We need to introduce the novel idea of designing SPDE triangulation precisely on road networks. The process is executed following sequential steps where a buffer region for each road segment is initially created, next a clipped buffer polygon is constructed which comprises only the



**Figure 4.** Traffic accident locations on road segments Without buffer (*left*), and With buffer (*right*).

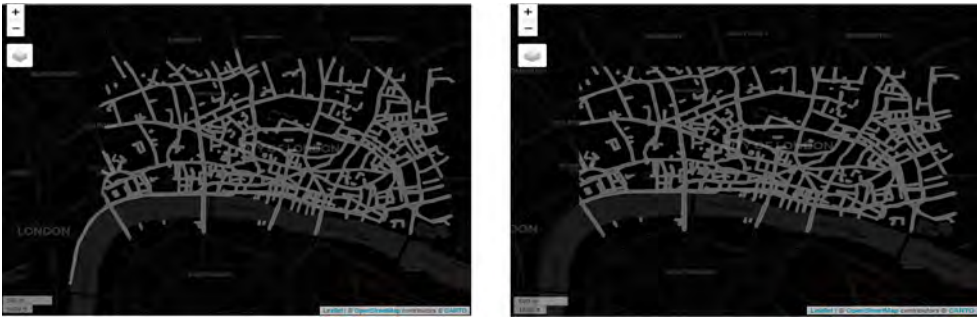
area covered by the road network, and finally SPDE triangulation is applied on the clipped polygon to construct the *SPDE Network Mesh*. Each step of the network triangulation process is discussed in brief as follows.

**Access OSM network:** OSM road network for the study region is accessed using R package *osmdata* [51].

**Buffer for each road segment:** A report by the National Academies of Sciences, Engineering, and Medicine (USA) on quality and accuracy of positional data in transportation [48] highlighted accuracy and reliability issues of positional data in transportation research works. There are instances where recorded data entries invariably introduce errors in both geometric and contextual attributes [43]. This happens also with our road traffic accident data when being positioned over the extracted OSM road network. Figure 4(left) depicts a sample of traffic accident locations (marked as red points) on the OSM road network. We note that many events are located away from the road segments. Initially, the buffer width is selected in such a manner to get maximum points within a standard buffer area for all road segments. We report that the GPS error is very similar for all road segments irrespective of their individual width. Thus, a common selected buffer width for all road segments served the best to get maximum points within buffer area. So, we check out with different buffer widths common for all the road segments. We have used several buffers widths, and selected a 20 meters buffer as the optimal one where the maximum points lie within buffer regions for each road segments. In Figure 4(right) we show the built buffers using the function *geo\_buffer* from R package *stplanr* [38]. OSM road network with 20 meters added buffer is depicted in Figure 5(left).

**Create Clipped Buffer Polygon:** Individual buffer segments are merged and converted into a single polygon clipped within a bounding box covering the study area. Figure 5(right) illustrates the clipped polygon of the buffered segments.

**Apply network triangulation:** As we need to analyze accident risk factor in each road segment, events within the buffer area of individual road segments are aggregated and counted. Then, the centroid of each segment is used as initial triangulation nodes applied on the clipped polygon. In relation to Delaunay's triangulation, it is worthy mentioning about the choice of buffer size and relevant parameters used to design SPDE mesh. Function *inla.mesh.2d* in R-INLA provides control for the largest allowed triangle edge length (*max.edge*) and minimum allowed distance between points (*cutoff*). The number of vertices in the SPDE mesh is regulated by both of these, as well as the boundary region of the study area. We report two issues while using buffer width proportional to street widths.



**Figure 5.** (Left) OSM road network with a 20 m buffer; (Right) Clipped polygon of buffered segments.

As mentioned in the previous section, erroneous GPS locations of accident sites leads to the first issue. In case of narrow streets, substantial numbers of accident points are found to be located outside the buffer area. The second issue is that when the buffer width goes below a threshold value, the entire structure of the mesh gets distorted. In contrast, if the buffer width is particularly large, owing to close proximity of road segments, two or more segments merge into one. This is not realistic in nature, especially while calculating the accident risk on individual road segments.

Thus, to avoid these technical issues we have identified a common threshold value for the buffer width for all road segments irrespective of their individual widths. According to Verdoj [70], we need to balance between number of vertices used to build the triangulated mesh and computational cost. The best fitting mesh should have enough vertices for effective prediction, but the number should be within a limit to have control over computational time. With this concept we have fine tuned *max.edge* and *cutoff* values with several models to identify the best fitted mesh. A series of SPDE-mesh are generated, and the best fitting mesh projected only on the road network, as illustrated in Figure 6, is selected. The number of vertices for the final selected mesh is 12666. Figure 6 depicts the network mesh together with 84360 accident events.

### 3.2. Risk map design

We discuss here how to build traffic accident risk maps onto the network structure coming from the fitted Poisson model. Coming from the predicted monthly accident occurrences on individual road segments, we build a *Risk Score* ( $R_{score}$ ). Initially, the raw risk score for any road segment is equal to the sum of the total number of expected monthly accident counts for that segment. Then, we design a dynamic normalization technique to convert this raw risk into categorical values defining what we call a *Risk Index*. Finally, the normalized risk indices are adapted to design the risk map over the entire road network. These steps are detailed as follows.

In the current study, we have calculated the raw risk score for a road segment as the sum of monthly accident counts on that segment. Literature on road safety suggests that a predefined category range has to be decided before modeling any risk map [16]. Thus, we consider some sort of dynamic normalization technique for the raw risk scores. Initially,



**Figure 6.** Selected network mesh with added traffic accident locations.

**Table 1.** Normalization for risk index values.

Normalize condition	Risk index	Safety measure
Segment having zero $R_{score}$	0	Low risk
$R_{score} < R_{range}$	0	Low risk
$R_{range} \leq R_{score} < 2 \times R_{range}$	1	Low-medium risk
$2 \times R_{range} \leq R_{score} < 3 \times R_{range}$	2	Medium risk
$3 \times R_{range} \leq R_{score} < 4 \times R_{range}$	3	Medium-high risk
$4 \times R_{range} \leq R_{score}$	4	High risk

the risk range is calculated as follows

$$R_{range} = \frac{(\max .R_{score} - \min .R_{score})}{\text{no. of risk categories}} \tag{3}$$

Next, we have used  $R_{range}$  to calculate the normalized values. As a relevant example, the values depicted in Table 1 show that the number of categories in the normalized scale is the same as the proposed number of risk categories. We note that the proposed dynamic normalization technique can be applied to similar risk index scales in road safety management.

We also highlight that the safety measure mentioned in Table 1 follows the European Road Assessment Programme (EuroRAP) standard to create the risk ratings of the motorways and other national roads in Europe [54]. The risk index algorithm implemented here has intended to categorize road segments based on the traffic accident records in each segment. As a result, segments having higher accident counts are categorized as accident-prone or high risk roads. A similar methodology can be adapted in other traffic risk modeling algorithms. The Risk index values of individual road segments are adapted to design final risk maps.

## 4. Results

This section presents results of the analysis and methodological approach developed in Section 3. In particular, we provide results on model fitting and prediction together with risk maps of accidents.

### 4.1. Model fitting

The proposed model, mentioned in Equation (1) has been fitted to the accident datasets for the years 2013 to 2015. The remaining accident records of 2016 and 2017 have been used for prediction purpose. R-INLA package [41] is used to fit all models mentioned in Section 3, by adapting and modifying existing coding for space-time analysis [79]. All models are executed separately for the same data set (January 2013 to December 2015).

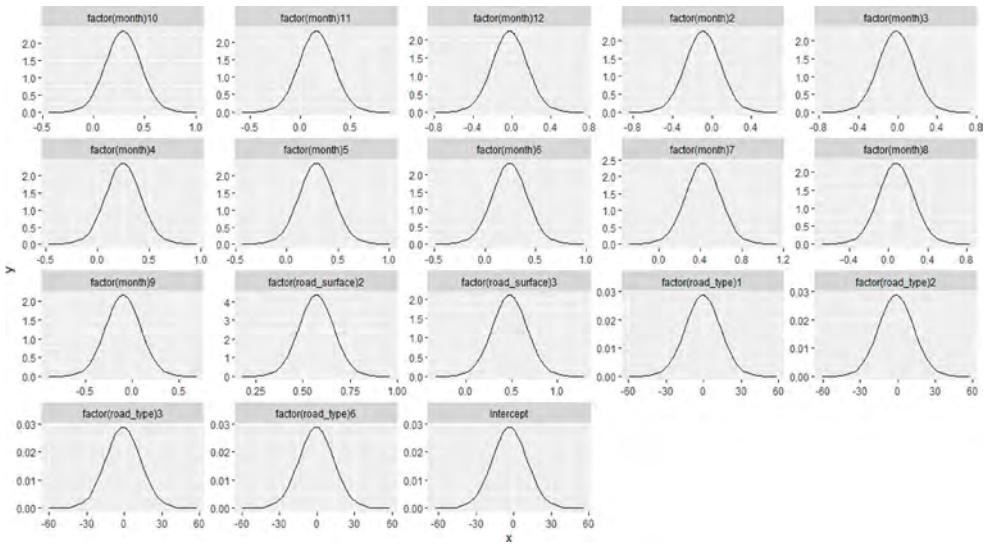
Deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC) are used to assess the performance of the models, and to select the best fitting model by balancing model accuracy against complexity [67]. Models having smaller DIC value, in spite of the added complexity, provide a more appropriate fit to the sampled data [9]. Summary results (DIC and WAIC) related to goodness-of-fit for all the fitted models are reported in Table 1 and Table 2 in Appendix. We note that, in each case, the computational time of non-interactive spatio-temporal models are found to be substantially low compared to the other interactive counterparts.

DIC values shown in Table 2 indicate that Poisson models (M1 to M4) provide the largest DIC values, while, in contrast, Poisson hurdle and ZIP show much better performances. Moreover, the zero-inflated models without spatio-temporal interaction (M5, M7, M9 and M11) provide a better fit than the corresponding spatio-temporal interactive pairs (M6, M8, M10 and M12). The values reported in Table 2 indicate that for model M7, DIC (9464.48) and WAIC (9471.32) are substantially lower compared to others.

Thus, to model the spatio-temporal structure of traffic accidents on London road networks, the Poisson hurdle model without a spatio-temporal interaction term is selected. We report that model M7 considers spatial and temporal effects together with three covariates (month, road type and surface) mentioned in Section 2. In each case, the models provide larger DIC and WAIC values when the covariates are not considered (see Table 1

**Table 2.** Competing models with DIC and WAIC values.

Model	DIC	WAIC
M1: Poisson	44137.41	44132.59
M2: Poisson	47251.88	47243.16
M3: Poisson	43433.75	43427.04
M4: Poisson	44041.93	44056.08
M5: Poisson hurdle	9571.01	9570.31
M6: Poisson hurdle	9932.40	9920.95
<b>M7: Poisson hurdle</b>	<b>9464.48</b>	<b>9471.32</b>
M8: Poisson hurdle	9490.83	9493.05
M9: Zero inflated Poisson	9683.70	9686.07
M10: Zero inflated Poisson	9896.62	9890.19
M11: Zero inflated Poisson	9491.44	9482.10
M12: Zero inflated Poisson	9511.15	9568.80



**Figure 7.** Marginal posterior distributions of covariate coefficients.

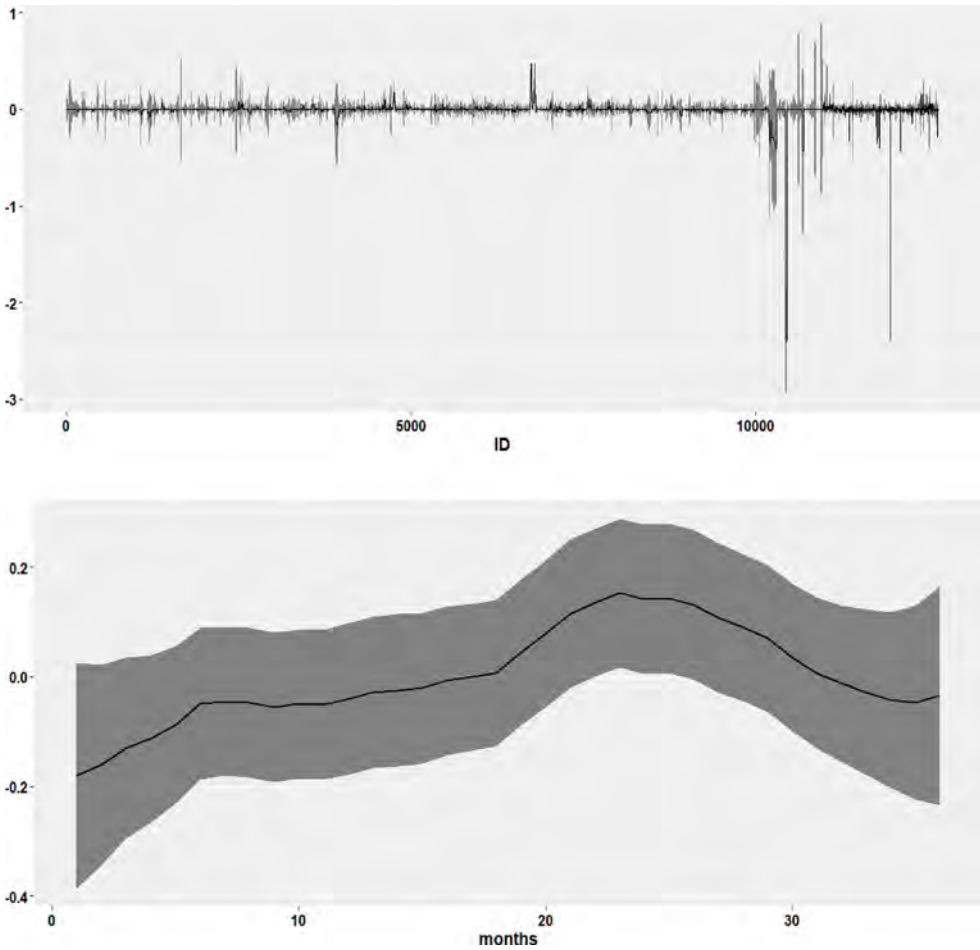
in Appendix). Additionally, the models perform better when PC priors are applied for the parameters to create the SPDE model object in case of Matérn model. As a note, in the current study, regardless of distribution type, the models show a better fit with the inclusion of all three covariates and PC priors under the case of no spatio-temporal interaction.

The posterior distribution of fixed and random effects included in the model are depicted in Figures 7 and 8. In particular, Figure 7 shows the marginal posterior distributions of all fixed effects related to covariates road type, road surface and month, confirming the Gaussian distribution centered at zero. Additionally, Table 3 in Appendix depicts the coefficients and credibility intervals of all fixed effects. We note that the covariate road type has no influence in our model. The positive mean values for the covariate road surface indicate positive influence in the model. However, in case of the month variable, only July shows positive significance while all other months have no influence in the model. In Section 5, we further detail the influence of variables on the model in further details.

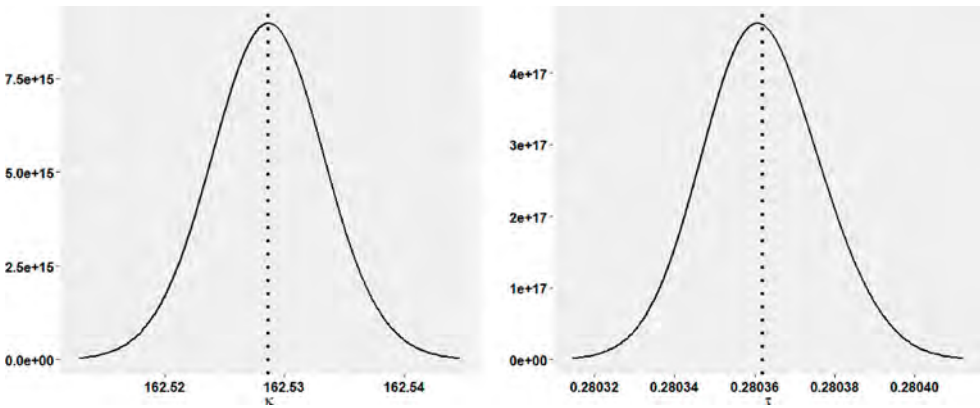
Figure 8 depicts the marginal posterior mean of the spatial  $\xi_i$  and temporal  $\zeta_t$  random effects. The horizontal axis of Figure 8 (top) represents 12666 triangulation nodes of SPDE network mesh used in the model. A stronger spatial effect is observed on the nodes of triangles on the road segments having higher accident counts (highlighted in Figure 6 as dark red patches). Similarly, Figure 8 (bottom) exhibits the variation of the marginal posterior mean of the temporal random effects over the 36 months for the model fitting years (2013 to 2015).

We finally note that spatial effect parameters  $\kappa$  and  $\tau$  have mean values 162.53 and 0.2804 as depicted in Figure 9 that shows the marginal posterior distributions of the two hyperparameters for the spatial random field. Using  $\kappa$  and  $\tau$  we can get the value of spatial range  $r = 0.0174$  km or 17.4 m.

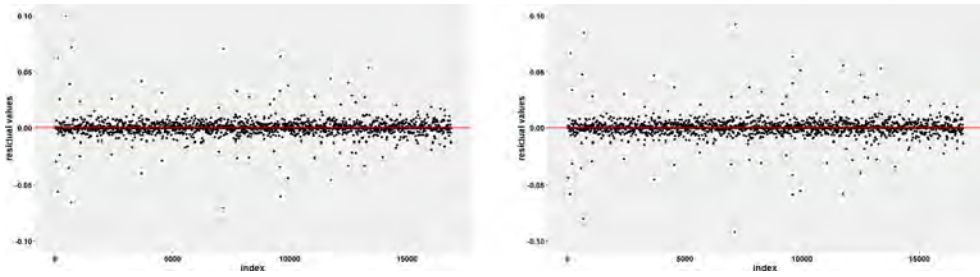




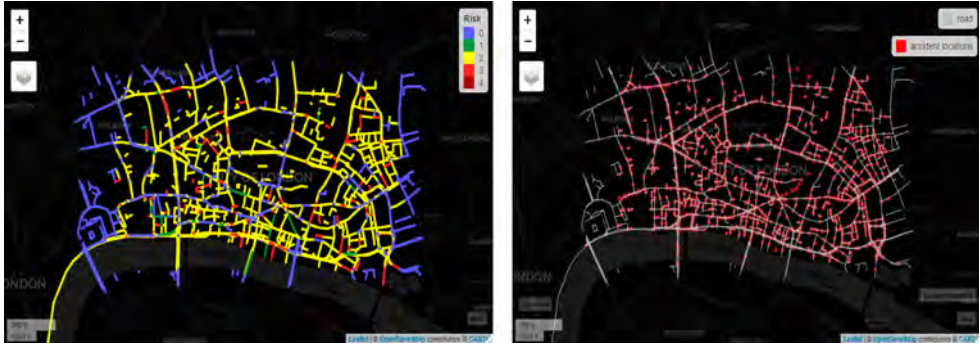
**Figure 8.** *Top:* marginal posterior mean of the spatial random effect  $\xi(\cdot)$ ; *Bottom:* marginal posterior mean of the temporal random effect  $\zeta_t$ .



**Figure 9.** Marginal posterior distributions of hyperparameters  $\kappa$  and  $\tau$  for the spatial random field  $\xi(\cdot)$ .



**Figure 10.** Residual (observed minus predicted) plots for: (Left) 2016; (Right) 2017.



**Figure 11.** Year 2016: (Left) Risk map; (Right) Original data of traffic accidents.



**Figure 12.** Year 2017: (Left) Risk map; (Right) Original data of traffic accidents.

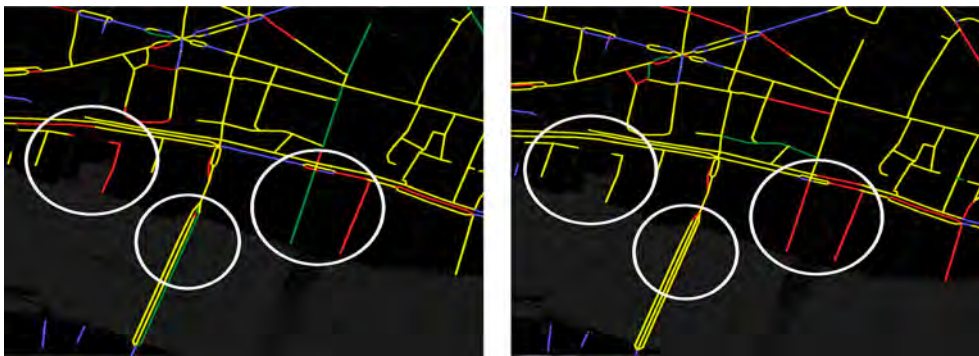
#### 4.2. Model prediction

Using the fitted model, we can analyze goodness-of-fit of the model by considering prediction over unsampled locations [79]. This prediction involves projecting the fitted model into the mesh at each road segments.

The proposed model is tested using test years (2016 and 2017) combined with the entire model fitting that used the years 2013 to 2015. From the final predicted result for both test years, we extract monthly predicted values for individual years. In each case, we calculate corresponding residuals of these predictions (observed minus predicted). Figure 10(a, b) depict such residuals; we note the residuals are generally close to zero and have no



**Figure 13.** Identified zones having consistent risk value for (Left) 2016 and (Right) 2017.



**Figure 14.** Highlighted streets with upward trend toward higher risk between 2016 (left) and 2017 (right).

particular structure. Root mean square error (RMSE) value acts as an indicator to assess the performance of a fitted model. We obtained  $RMSE = 0.0135$  for 2016, and  $RMSE = 0.0121$  for 2017, which are similar and particularly small. Further discussion on model performance is reviewed in Section 5.

### 4.3. Risk map

We calculate the risk index for individual road segments following the indications in Section 3.2, and using the safety measure scale shown in Table 1. The normalized risk index values are calculated using the predicted values for years 2016 and 2017. The risk maps are visualized in an interactive geospatial platform using R package `mapview` [4]. Figure 11(a) illustrates the risk map for 2016 and corresponding original traffic accident locations are depicted in Figure 11(b). Similar results for 2017 are presented in Figure 12(a,b). The color scale (0 through 4) used in each map follows the same safety measure scale used in Table 1.

The predicted risk maps are visually compared with original traffic accidents records during the same time span. Some interesting observations are noted. For both years, most of the roads in the outskirts of the city are predicted to be relatively safe than the city center. Indeed, during these years, roads near the city center are predicted with medium to

high risk levels. Figure 13 highlights three consistent risk zones for both years. The left-most highlighted area being outside the city center shows a steady low risk zone. While the other two highlighted zones represent consistent high risk roads. The identified zones show similar trends when we compare with original accident records during 2016 and 2017.

We note some interesting annual variations in particular road segments identified by our model predictions. If we look at Figure 14, with some streets highlighted by white circles, we see that the predictions from 2016 to 2017 are following an upward trend toward a higher risk. Indeed, the number of accidents in these streets increased from one year to the next.

## 5. Conclusions and discussion

The current study presents a spatio-temporal model predicting the occurrence of traffic accidents in an urban environment. The model is used to create dynamic risk maps for a road network. To balance computation time and accuracy, the present research work took advantage of the spatio-temporal nature of the data, and used Bayesian methodology by including INLA and SPDE in the modeling process.

Literature [11,39,40] suggests that model fitting using diverse subset combinations of variables provides opportunities to improve prediction accuracy. In the proposed model, we have included three covariates (see Section 2). Out of them, except variables road surface and one of the months (see Table 3 in Appendix) have no influence on the model. Thus, future research works can explain some of the noted variations on improving prediction accuracy by careful inclusion of significant exogenous variables related to traffic flow, traffic control and temporal variables such as time of accident occurrence. Furthermore, studies like [17,50] suggest future research works in exploring reliable and large training data set that can improve the performance of the proposed model.

In recent years, spatio-temporal modeling of road traffic accidents and risk mapping has gained attention, especially in the domain of multi-dimensional road safety management. Besides, travel risk maps are gaining popularity among business travellers, tourists and emergency service providers. Results and findings of the current study illustrate that the proposed model can generate predicted risk maps of the entire road network for any urban study area. In this sense, it is dynamic in nature. The model is flexible and general, and thus can be adapted to similar problems. It can handle different types of covariates in space or time, spatial and temporal structures and space-time interactions. The predicted risk maps of traffic accidents is one of our interesting outcomes. We can produce the road safety index of all road segments, including small details of each junctions or sharp turnings. In our particular problem, we can point to which elements authorities can take dedicated actions to control and reduce traffic accidents as our model identifies significant elements that can be controlled and modified by humans. This means we provide a real, pragmatic and realistic element for institutions to take actions on reducing the risk of traffic accidents.

Moreover, identification of potentially dangerous roads and regions can act as baseline information for geospatial analysis on road safety. The results can have strategic applications in developing GIS analytical tools to identify and depict possible safe routes. As the risk map provides information about the entire road network, it can be flexible enough to generate possible alternative safe route(s) between any source and destinations pairs.

Another important use of the model is analyzing the change and trend pattern of traffic accidents. We can find some literature suggesting this line of research in the city of London [8,15,71], and similar works in other countries [5,34]. As depicted in Figure 14, identification of gradual changes in risk values and their potential factors, are of interest for future research works on change point detection.

Consequently, the novelty of the study is the introduction of *SPDE network triangulation* or *SPDE network mesh* to estimate the spatial auto-correlation of discrete events. As such, it took a new step in INLA-SPDE modeling to perform spatio-temporal predictive analysis only on selected areas (specifically for road networks), instead of performing on entire continuous region. In a broader picture, the study contributes to the relatively small amount of literature on spatio-temporal analysis using INLA-SPDE of spatial events precisely on road networks. The methodology is dynamic and can be adapted and applied to other locations globally.

### Author contributions statement

Conceptualization, PJ and SC; Data curation, SC; Formal analysis, SC, PJ and JM; Methodology, SC, PJ and JM; Project administration, JM; Resources, SC; Software, PJ and SC; Validation, PJ and JM; Writing original draft, SC; Writing review & editing, PJ and JM. The authors declare that they have no conflict of interest.

### Ethical statement

Author and co-authors testify that, this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We know of no conflicts of interest associated with this publication and there has been no financial support for this work.

### Disclosure statement

No potential conflict of interest was reported by the author(s).

### ORCID

Somnath Chaudhuri  <http://orcid.org/0000-0003-4899-1870>

Pablo Juan  <http://orcid.org/0000-0002-2197-7502>

Jorge Mateu  <http://orcid.org/0000-0002-2868-7604>

### References

- [1] A. Abdel-Salam, F. Guo, A. Flintsch, M. Arafeh, and H. Rakha, *Linear regression crash prediction models, Efficient Transportation and Pavement Systems*, CRC Press, 2008.
- [2] M. I. Adarabioyo and R. A. Ipinyomi, *Comparing zero-inflated poisson, zero-inflated negative binomial and zero-inflated geometric in count data with excess zero*, *Asian J. Probab. Stat.* 4 (2019), pp. 1–10.
- [3] M. A. Aghajani, R. S. Dezfoulian, A. R. Arjroody, and M. Rezaei, *Applying GIS to identify the spatial and temporal patterns of road accidents using spatial statistics (case study: Ilam Province, Iran)*, *Transp. Res. Procedia* 25 (2017), pp. 2126–2138.
- [4] T. Appelhans, F. Detsch, C. Reudenbach, and S. Woellauer, *mapview – Interactive viewing of spatial data in R*, EGU General Assembly Conference Abstracts, 2016. Article EPSC2016–1832, EPSC2016–1832.
- [5] I. Ashraf, S. Hur, M. Shafiq, Y. Park, and Y. Guo, *Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis (Y. Guo, Ed.)*, *PLoS ONE*. 14 (2019), pp. e0223473.

- [6] E. Azuike, K. Okafor, and P. Okojie, *The causes and prevalence of road traffic accidents amongst commercial long distance drivers in Benin city, Edo state, Nigeria*, Nig. Hosp. Pract. 26 (2017), pp. 220–230.
- [7] H. Bakka, H. Rue, G. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren, *Spatial modeling with R-INLA: A review*, WIREs Comput. Stat. 10 (2018), pp. e1443.
- [8] A. Bhawkar, *Severe traffic accidents in United Kingdom*, Tech. rep., National College of Ireland, 2018. Available at [https://www.researchgate.net/publication/330676135\\_SevereTraffic\\_Accidents\\_in\\_United\\_Kingdom](https://www.researchgate.net/publication/330676135_SevereTraffic_Accidents_in_United_Kingdom).
- [9] M. Blangiardo and M. Cameletti, *Spatial and Spatio-temporal Bayesian Models with R-INLA*, John Wiley & Sons, Ltd, London, 2015.
- [10] Á. Briz-Redón, F. Martínez-Ruiz, and F. Montes, *Identification of differential risk hotspots for collision and vehicle type in a directed linear network*, Accid. Anal. Preven. 132 (2019), pp. 105278.
- [11] M. Cameletti, F. Lindgren, D. Simpson, and H. Rue, *Spatio-temporal modeling of particulate matter concentration through the SPDE approach*, AStA Adv. Stat. Anal. 97 (2013), pp. 109–131.
- [12] V. Cantillo, P. Garcés, and L. Márquez, *Factors influencing the occurrence of traffic accidents in urban roads: A combined GIS-Empirical Bayesian approach*, DYNA 83 (2016), pp. 21–28.
- [13] M. Castro, R. Paleti, and C. R. Bhat, *A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections*, Transp. Res. Part B: Methodol. 46 (2012), pp. 253–272.
- [14] S. Chaudhuri, M. Moradi, and J. Mateu, *On the trend detection of time-ordered intensity images of point processes on linear networks*, Commun. Stat. – Simul. Comput. (2021), pp. 1–13. doi: [10.1080/03610918.2021.1881116](https://doi.org/10.1080/03610918.2021.1881116).
- [15] R.P. Curiel, H.G. Ramírez, and S.R. Bishop, *A novel rare event approach to measure the randomness and concentration of road accidents* (Y. Deng, Ed.), PLoS. ONE. 13 (2018), pp. e0201890.
- [16] D. Curran-Everett, *Explorations in statistics: The analysis of ratios and normalized data*, Adv. Physiol. Educ. 37 (2013), pp. 213–219.
- [17] E.J. de Fortuny, D. Martens, and F. Provost, *Predictive modeling with big data: Is bigger really better?*, Big. Data. 1 (2013), pp. 215–226.
- [18] F. Demasi, G. Loprencipe, and L. Moretti, *Road safety analysis of urban roads: Case study of an Italian municipality*, Safety 4 (2018), pp. 58.
- [19] M. Deublein, M. Schubert, B. T. Adey, J. Köhler, and M. H. Faber, *Prediction of road accidents: A Bayesian hierarchical approach*, Accid. Anal. Preven. 51 (2013), pp. 274–291.
- [20] C.M. Farmer, *Temporal factors in motor vehicle crash deaths*, Inj. Prev. 11 (2005), pp. 18–23.
- [21] U. Galgamuwa, *Bayesian spatial modeling to incorporate unmeasured information at road segment levels with the INLA approach: A methodological advancement of estimating crash modification factors*, J. Traff. Transp. Eng. (English Edition), 2019.
- [22] A. Gelman, *Bayesian data analysis*, 2014. [OCLC: 909477393].
- [23] Y. Guo, A. Osama, and T. Sayed, *A cross-comparison of different techniques for modeling macro-level cyclist crashes*, Accid. Anal. Preven. 113 (2018), pp. 38–46.
- [24] M.-C. Hu, M. Pavlicova, and E.V. Nunes, *Zero-Inflated and hurdle models of count data with extra zeros: Examples from an HIV-Risk reduction intervention trial*, Am. J. Drug. Alcohol. Abuse. 37 (2011), pp. 367–375.
- [25] J. Huang, B. P. Malone, B. Minasny, A. B. McBratney, and J. Triantafyllis, *Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping*, Sci. Total Environ. 609 (2017), pp. 621–632.
- [26] F.J. Jegede, *Spatio-temporal analysis of road traffic accidents in Oyo state, Nigeria*, Accid. Anal. Preven. 20 (1988), pp. 227–243.
- [27] P. Juan, J. Mateu, and M. Saez, *Pinpointing spatio-temporal interactions in wildfire patterns*, Stoch. Environ. Res. Risk. Assess. 26 (2012), pp. 1131–1150.
- [28] M. Karacasu, B. Ergül, and A.A. Yavuz, *Limited. estimating the causes of traffic accidents using logistic regression and discriminant analysis*, Int. J. Inj. Contr. Saf. Promot. 21 (2013), pp. 305–313.

- [29] A. Karaganis, *A Spatial Point Process for Estimating the Probability of Occurrence of a Traffic Accident*, European Regional Science Association, ERSA conference papers, 2006.
- [30] D. Khulbe, *Modeling Severe Traffic Accidents With Spatial And Temporal Features*, ICONIP, 2019.
- [31] D. Lambert, *Zero-Inflated poisson regression, with an application to defects in manufacturing*, *Technometrics* 34 (1992), pp. 1–14.
- [32] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach*, *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 73 (2011), pp. 423–498.
- [33] C. Liu, *Exploring spatio-temporal effects in traffic crash trend analysis*, *Anal. Meth. Accid. Res.* 16 (2017), pp. 104–116.
- [34] C. Liu, S. Zhang, H. Wu, and Q. Fu, *A dynamic spatiotemporal analysis model for traffic incident influence prediction on urban road networks*, *ISPRS Int. J. Geo-Inform.* 6 (2017), pp. 362.
- [35] B.P.Y. Loo, S. Yao, and J. Wu, *Spatial point analysis of road crashes in Shanghai: A GIS-based network kernel density method*, 2011 19th International Conference on Geoinformatics, 2011.
- [36] A. Lopez-Quílez and F. Muñoz, *Review of spatio-temporal models for disease mapping*, Tech. rep., 2009. Available at <https://www.uv.es/famarmu/doc/Euroheis2-report.pdf>.
- [37] D. Lord and B. N. Persaud, *Accident prediction models with and without trend: Application of the generalized estimating equations procedure*, *Transp. Res. Rec.: J. Transp. Res. Board* 1717 (2000), pp. 102–108.
- [38] R. Lovelace and R. Ellison, *stplanr: A package for transport planning*, *The R Journal* 10 (2018), pp. 7–23.
- [39] J. Martínez-Minaya, F. Lindgren, A. López-Quílez, D. Simpson, and D. Conesa, *The Integrated nested Laplace approximation for fitting models with multivariate response*, 2021.
- [40] S. Martino and H. Rue, *Case studies in Bayesian computation using INLA*, in *Complex Data Modeling and Computationally Intensive Statistical Methods*, P. Mantovan, P. Secchi, eds., Springer, Milan, 2010. pp. 99–114.
- [41] T. G. Martins, D. Simpson, F. Lindgren, and H. Rue, *Bayesian computing with INLA: New features*, *Comput. Stat. Data Anal.* 67 (2013), pp. 68–83.
- [42] S.-P. Miaou, *The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions*, *Accid. Anal. Preven.* 26 (1994), pp. 471–482.
- [43] M. Miler, F. Todić, and M. Sevrović, *Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string matching technique*, *Transp. Res. Part C: Emerg. Technol.* 68 (2016), pp. 185–193.
- [44] M. Mohanty and A. Gupta, *Factors affecting road crash modeling*, *J. Transp. Lit.* 9 (2015), pp. 15–19.
- [45] M.M. Moradi, *Spatial and spatio-temporal point patterns on linear networks*, Doctoral diss., Universitat Jaume I, 2018.
- [46] M.M. Moradi and J. Mateu, *First- and second-order characteristics of spatio-temporal point processes on linear networks*, *J. Comput. Graph. Stat.* 0 (2019), pp. 1–21.
- [47] P. Moraga, *Geospatial Health Data : Modeling and Visualization with R-INLA and Shiny*, CRC Press, London, 2020.
- [48] NCHRP, *Quality and Accuracy of Positional Data in Transportation*, Transportation Research Board, 2003.
- [49] J. Oh, S. P. Washington, and D. Nam, *Accident prediction model for railway-highway interfaces*, *Accid. Anal. Preven.* 38 (2006), pp. 346–356.
- [50] M.C.M. Oo and T. Thein, *An efficient predictive analytics system for high dimensional big data*, *J. King Saud University – Comput. Inform. Sci.* 34 (2019), pp. 1521–1532.
- [51] M. Padgham, B. Rudis, R. Lovelace, and M. Salmon, *OSM Data*, *The J. Open Source Softw.* 2 (2017), 305.
- [52] V. Prasannakumar, H. Vijith, R. Charutha, and N. Geetha, *Spatio-Temporal clustering of road accidents: GIS based analysis and assessment*, *Procedia – Soc. Behav. Sci.* 21 (2011), pp. 317–325.
- [53] S.S. Pulugurtha and V.R. Sambhara, *Pedestrian crash estimation models for signalized intersections*, *Accid. Anal. Preven.* 43 (2011), pp. 439–446.

- [54] Risk mapping for the TEN-T in Croatia, Greece, Italy and Spain: Update. *EuroRAP* 2016. Available at <https://eurorap.org/risk-mapping-for-the-ten-t-in-croatia-greeceitaly-and-spain-update/>.
- [55] C.E. Rose, S.W. Martin, K.A. Wannemuehler, and B.D. Plikaytis, *On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data*, *J. Biopharm. Stat.* 16 (2006), pp. 463–481.
- [56] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications (Chapman & Hall/CRC Monographs on Statistics and Applied Probability)*, Chapman & Hall/CRC, London, 2005.
- [57] H. Rue, S. Martino, and N. Chopin, *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*, *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 71 (2009), pp. 319–392.
- [58] H. Rue, A. Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson, and F.K. Lindgren, *Bayesian Computing with INLA: A Review*, 2016.
- [59] H. Rue, A. Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson, and F.K. Lindgren, *Bayesian computing with INLA: A review*, *Annu. Rev. Stat. Appl.* 4 (2017), pp. 395–421.
- [60] M. Salifu, *Accident prediction models for unsignalised urban junctions in Ghana*, *IATSS Res.* 28 (2004), pp. 68–81.
- [61] L. Serra, M. Saez, P. Juan, D. Varga, and J. Mateu, *A spatio-temporal Poisson hurdle point process to model wildfires*, *Stoch. Environ. Res. Risk. Assess.* 28 (2013), pp. 1671–1684. doi: [10.1007/s00477-013-0823-x](https://doi.org/10.1007/s00477-013-0823-x).
- [62] G. A. Shafabakhsh, A. Famili, and M. S. Bahadori, *GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran*, *J. Traffic Transp. Eng. (English Edition)* 4 (2017), pp. 290–299.
- [63] S. Shahid, A. Minhans, O. Che Puan, S. A. Hasan, and T. Ismail, *Spatial and temporal pattern of road accidents and casualties in peninsular Malaysia*, *J. Teknol.* 76 (2015), pp. 000.
- [64] D.P. Simpson, H. Rue, A. Riebler, T.G. Martins, and S.H. Sørbye, *Penalising model component complexity: A principled, practical approach to constructing priors*, preprint (2014), arXiv doi: [10.48550/ARXIV.1403.4630](https://doi.org/10.48550/ARXIV.1403.4630).
- [65] D. Simpson, H. Rue, A. Riebler, T.G. Martins, and S.H. Sørbye, *Penalising model component complexity: A principled, practical approach to constructing priors*, *Stat. Sci.* 32 (2017), pp. 1–28. doi: [10.1214/16-sts576](https://doi.org/10.1214/16-sts576).
- [66] T.D. Smedt, *Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models*, *Arch. Public Health* 73 (2015), doi: [10.1186/2049-3258-73-s1-o2](https://doi.org/10.1186/2049-3258-73-s1-o2).
- [67] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, *Bayesian measures of model complexity and fit*, *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 64 (2002), pp. 583–639.
- [68] B.M. Taylor and P.J. Diggle, *INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes*, *J. Stat. Comput. Simul.* 84 (2014), pp. 2266–2284.
- [69] TFL, *Transport for London: Road Safety*, 2019. Retrieved April 30, 2019, Available at <https://tfl.gov.uk/corporate/publications-and-reports/road-safety>.
- [70] P.J. Verdoy, *Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. choosing the best mesh for each database*, *Commun. Stat. – Simul. Comput.* 50 (2021), pp. 2990–3030.
- [71] C. Wang, M. A. Quddus, and S. G. Ison, *Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England*, *Accid. Anal. Prev.* 41 (2009), pp. 798–808.
- [72] W. Wang, Z. Yuan, Y. Yang, X. Yang, Y. Liu, and Y. Guo, *Factors influencing traffic accident frequencies on urban roads: A spatial panel time-fixed effects error model (Y. Guo, Ed.)*, *PLoS ONE.* 14 (2019), pp. e0214539.
- [73] S. Watanabe, *Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory*, *J. Mach. Learn. Res.* 11 (2010), 3571–3594.
- [74] WHO, *Global Status Report on Road Safety 2018*, World Health Organization, (2019, Jan 10).
- [75] C. K. Wikle, L. M. Berliner, and N. Cressie, *Hierarchical Bayesian space-time models*, *Environ. Ecol. Stat.* 5 (1998), pp. 117–154.



- [76] A. M. Williamson and A.-M. Feyer, *Causes of accidents and the time of day*, Work & Stress 9 (1995), pp. 158–164.
- [77] P. Xu and H. Huang, *Modeling crash spatial heterogeneity: Random parameter versus geographically weighting*, Accid. Anal. Prev. 75 (2015), pp. 16–25.
- [78] F. Zhong-xiang, L. Shi-sheng, Z. Wei-hua, and Z. Nan-nan, *Combined prediction model of death toll for road traffic accidents based on independent and dependent variables*, Comput. Intell. Neurosci. 2014 (2014), pp. 1–7.
- [79] A.F. Zuur, *Beginners Guide to Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-INLA: Using GLM and GLMM (Vol. 1)*, Highland Statistics Ltd, Scotland, 2017.

## **5.3 Article 3: Risk Mapping Road Networks in Barcelona, Spain**

### **Spatiotemporal modeling of traffic risk mapping: A study of urban road networks in Barcelona, Spain**

Somnath Chaudhuri<sup>1,2</sup>, Marc Saez<sup>1,2</sup>, Diego Varga<sup>1,2,3</sup> and Pablo Juan<sup>1,4</sup>

1. Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain
2. CIBER of Epidemiology and Public Health (CIBERESP), Spain
3. Department of Geography, University of Girona, Spain
4. Department of Mathematics, University of Jaume I, Spain

**Spatial Statistics, 53, 100722**

Impact Factor (2021): 2.125

Statistics & Probability, Position 41 out of 125 (Q2)



Contents lists available at ScienceDirect

# Spatial Statistics

journal homepage: [www.elsevier.com/locate/spasta](http://www.elsevier.com/locate/spasta)

## Spatiotemporal modeling of traffic risk mapping: A study of urban road networks in Barcelona, Spain



Somnath Chaudhuri<sup>a,b</sup>, Marc Saez<sup>a,b,\*</sup>, Diego Varga<sup>a,b,c</sup>,  
Pablo Juan<sup>a,d</sup>

<sup>a</sup> Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain

<sup>b</sup> CIBER of Epidemiology and Public Health (CIBERESP), Spain

<sup>c</sup> Department of Geography, University of Girona, Spain

<sup>d</sup> Department of Mathematics, University of Jaume I, Spain

### ARTICLE INFO

#### Article history:

Received 21 May 2022

Received in revised form 4 December 2022

Accepted 9 December 2022

Available online 21 December 2022

#### Keywords:

INLA

Network triangulation

SPDE

Traffic accident

### ABSTRACT

Accidents on the road have always been a major concern in modern society. According to the World Health Organization, globally road traffic collisions are one of the leading and fastest growing causes of disability and death. The present research work is conducted on ten years of traffic accident data in an urban environment to explore and analyze spatial and temporal variation in the accidents and related injuries. The proposed spatiotemporal model can make predictions regarding the number of injuries incurred on individual road segments. Bayesian methodology using Integrated Nested Laplace Approximation (INLA) with Stochastic Partial Differential Equations (SPDE) has been applied to generate a predicted risk map for the entire road network. The current study introduces INLA-SPDE modeling to perform spatiotemporal predictive analysis on selected areas, precisely on road networks instead of traditional continuous regions. Additionally, the result risk maps act as a baseline to identify the safe routes in a spatiotemporal context. The methodology can be adapted and applied to enhanced INLA-SPDE modeling of spatial point processes precisely on road networks.

© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

\* Corresponding author at: Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain.  
E-mail address: [marc.saez@udg.edu](mailto:marc.saez@udg.edu) (M. Saez).

<https://doi.org/10.1016/j.spasta.2022.100722>

2211-6753/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

With the growth of population and economic development, the process of urbanization is accelerating, and the number of vehicles in urban cities is increasing. In recent years, road traffic collisions have become a major concern in modern society. According to 2018 global status report on road safety by the World Health Organisation, every year, around 1.2 million people die on the world's roadways, with another 20 to 50 million suffering non-fatal injuries. Traffic accident is declared as one of the leading causes of death for people of all ages (WHO, 2019). Literature suggests road infrastructure and uncontrolled vehicle speed increase accident risk (Briz-Redó et al., 2019). Additionally, temporal factors (time of the day or weekend nights) act as decisive aspects in the number and impact of accidents (Farmer, 2005; Liu and Sharma, 2017).

In the field of road safety, identifying relevant factors and spatio-temporal modeling of traffic accidents have been a major focus of research (Williamson and Feyer, 1995; Prasannakumar et al., 2011; Zhong-xiang et al., 2014; Cantillo et al., 2016; Khulbe and Sourav, 2019). A series of studies (Jegede, 1988; Farmer, 2005; Aghajani et al., 2017; Shafabakhsh et al., 2017) have been conducted to explore historical data in order to identify risk factors and assess the likelihoods of crash-related events in order to categorize spatiotemporal factors influencing traffic accidents. Research in suitable statistical methodologies to analyze traffic accident data is a fundamental line of research in the field of traffic safety analysis. We can focus on detecting areas with a high accident concentration (hot spot detection methods) and focus on modeling traffic accident risk (expressed by raw accident counts or accident rates) depending on a set of predictive covariates. In statistical analysis and prediction modeling, these factors are regarded as significant predictors.

Accessible, and sustainable transport systems in cities are a core target of 2030 sustainable development goals (SDGs) adopted by the United Nations (UNDP, 2021). Thus, there is an opportunity to apply advanced computational techniques to the problem of road safety and traffic management. Various models and techniques have been developed and explored in this domain (Karaganis and Mimis, 2006; Pulugurtha and Sambhara, 2011; Castro et al., 2012; Boulieri et al., 2016; Khulbe and Sourav, 2019). Analyzing traffic safety performance by understanding crash fatality rates and influencing factors is essential for developing well-informed policies and designing effective countermeasures. Understanding the causes of the crashes, identifying appropriate solutions, and, proactively adopting or using them helps improve traffic safety. Studies like, Xu and Huang (2015), Guo et al. (2018) and Wang et al. (2019), highlight the existence of spatial autocorrelation of traffic accident events. In this line to analyze spatial and spatio-temporal distribution of traffic collisions, statistical inference comes along with Bayesian methodology. Boulieri et al. (2016) designed a space-time multivariate Bayesian model to analyze road traffic accidents by severity in different cities of the UK. A Bayesian approach with Markov chain Monte Carlo (MCMC) simulation methods has traditionally been used to fit generalized linear mixed models (GLMM) in a spatial context (Wikle et al., 1998). However, the computation time for MCMC models is considerably high for big datasets (Rue et al., 2009; Smedt et al., 2015). As recommended by Rue et al. (2009), while processing spatial data we can utilize integrated nested Laplace approximations (INLA) in conjunction with SPDE to balance speed and accuracy of the models. But, in the context to traffic accident event modeling, there are limited contributions using INLA-SPDE approach. Recently, Galgamuwa et al. (2019) used Bayesian spatial modeling with INLA in predicting road traffic accidents based on unmeasured information at road segment levels. Similar study by Chaudhuri et al. (2022) explores spatiotemporal modeling of traffic accidents on the road network of London, UK based on an explicit network triangulation using INLA-SPDE.

The aim of this paper is to propose a multi-disciplinary road-safety analysis technique by introducing spatio-temporal modeling of traffic accidents using Bayesian methodology restricted entirely on to the road network. We introduce an advance and realistic computational strategy to construct spatial triangulation restricted only onto the network topology. The proposed model can predict a risk index over individual road segments generating a categorized risk map of the entire road network. The study is conducted on ten years road traffic accidents data from the city of Barcelona, Spain. R (version R 4.1.2) programming language (R Core Team, 2021) has been used for statistical computing and graphical analysis. All computations were conducted on a quad-core Intel i9-4790 (3.60 GHz) processor with 32 GB (DDR3-1333/1600) RAM.



**Fig. 1.** Location of Barcelona city and road network of the study area. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The rest of the paper is organized as follows. Section 2 presents the methods adopted in the current study. The beginning of the section reports about the study area and data settings along with official sources of data. A description of the spatio-temporal modeling framework with emphasis on SPDE triangulation designed precisely on road network is discussed in Section 2.2. Section 3 is devoted to present the results of model prediction and risk map analysis. Some discussions are highlighted in Section 4. The paper ends with some concluding remarks in Section 5.

## 2. Methods

### 2.1. Data settings

Barcelona is the largest and the capital city of the community of Catalonia, Spain. This second most populous municipality of Spain is located on the coast of northeastern of the country. Fig. 1 shows the location of Barcelona city, and the boundary of the municipality highlighted in red. According to Barcelona’s city hall open data service (Open Data BCN) (OpenDataBCN, 2021), the city has a population of 1.6 million and approximately 15,748 inhabitants per square km. Besides being one of the major cultural, economic, and financial center, Barcelona is also a transport hub for entire southwestern Europe. The city municipality maintains an extensive motorway network.

In the current study, we have considered a small area (4.4 square km) from the central part of Barcelona consisting of 2058 road segments as depicted in Fig. 1 inside the black circle. The road network is also accessed from Open Data BCN repository. The police department in the city maintains records of traffic accidents. Detailed records about the circumstances of road accidents on public roads, corresponding casualties and injuries are managed and published annually by Open Data BCN. The data is free and available under the *Creative Commons Attribution 4.0* for public sector information. During the period of January 2010 to December 2019 the study area has records of

11,067 traffic accidents. Fig. A.8 in the Appendix shows the road networks in the study area with traffic accident locations highlighted in red.

Five datasets from Open Data BCN have been accessed for the study, which are referred to the accident itself and interrelated by a record code from 2010 to 2019. The recorded common attributes are unique event id, district and neighborhood, location postal address and geographical coordinates, occurrence day and time, kind of day (working or holiday). Each of dataset contain the following temporal variables: year, month, and time of accident. Related to spatial variables we made few changes. In raw dataset, individual accident locations in most cases are not located exactly on the road segment. We have shifted individual locations to the nearest road segments. In addition, we calculated the on road network distance of nearest bus stop, municipality market, restaurant, school, street market for each accident locations. These distances are used as spatial covariates in the dataset. We report that the traffic intensity records for each road segment are collected from TomTom Traffic Stats (TomTom, 2021).

It is noteworthy to mention that values of traffic intensity for individual road segments is not directly available from TomTom dataset. We used three variables to calculate the traffic intensity such as, *road length* ranging from 3.69 to 186.25 meter, *road type* (values 1 to 7, higher the value less the traffic) and *road speed limit* (18 to 80 km per hour) where roads having 30, 35 and 50 km per hour cover 21%, 28% and 35% of the total. Our proposed formula to calculate the traffic intensity is:

$$\text{Traffic Intensity} = (\text{Speed Limit}/\text{Type of Road}) * \ln(\text{Road Length})$$

where,  $\ln$  stands for natural logarithm that is  $\log$  to the base of  $e$ . Fig. A.9 in the Appendix depicts the calculated traffic intensity for individual road segments of the study area.

We are using the road length in natural logarithm ( $\ln$ ) scale to reduce the range of road length values. Other variables available in the dataset are number of victims, vehicles involved, minor and major injured persons, and number of casualties for each accident records. We have used the number of minor injured person as the response variable in our models. Traffic accidents recorded with only one minor injury comprises the maximum percentage of records (74.8.76%) followed by two minor injuries (15.42%), and 3 or more minor injuries (3.42%). The record shows 6.4% of the accidents are without having any minor injuries. We note that most accidents (99.85%) are having no causality. The number of traffic accidents documented in each of the study years (2010–2019) is similar, with the highest number (1270) recorded in 2016 and the lowest number (847) recorded in 2011. In case of monthly records for the entire study period, January records the minimum accident counts (846) and July has the maximum value (1023). It is worth noting that, almost 50% of all accidents occur during office hours, which are from 8 a.m. to 11 a.m. and from 3 p.m. to 6 p.m.

## 2.2. Statistical analysis

Random spatial events, such as traffic accidents, form irregularly scattered point patterns over regions of interest. Literature shows, spatio-temporal point process models are useful tools for performing focused statistical analysis (Karaganis and Mimis, 2006; Loo et al., 2011; Juan et al., 2012). In this context, Liu et al. (2017) propose that the occurrence of traffic accidents depend on spatio-temporal interacting and triggering factors (Liu et al., 2017). Moreover, we can find recent studies (Galgamuwa et al., 2019; Moradi and Mateu, 2019; Chaudhuri et al., 2021) on spatio-temporal point processes over networks that are able to identify spatial autocorrelations and interactions between points in the pattern. We open the door to consider binomial regression models in combination with a Bayesian framework for the prediction of traffic accidents on individual road segments by aggregating data for the occurrence of accident injuries per road segment given the total traffic intensity. We have used a computationally faster solution for prediction of the marginal distributions for latent Gaussian models and models with a large number of geo-locations by using a Laplace approximation for the integrals with the integrated nested Laplace approximation (INLA) method. (Rue et al., 2009). It focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF) (Rue and Held, 2005). We follow this approach combining a spatio-temporal binomial regression method within a Bayesian framework using INLA and stochastic

partial differential equation (SPDE). In particular, specification of the binomial distribution in INLA for responses  $y = 0, 1, 2, \dots, n$  is represented as

$$Prob(y) = \binom{n}{y} p^y (1 - p)^{(n-y)} \tag{1}$$

where,  $n$  is the number of trials and  $p$  is the probability of success in each trial (Rue et al., 2009).

The mean and variance of  $y$  are respectively  $\mu = np$  and  $\sigma^2 = np(1 - p)$  and the probability  $p$  is linked to the linear predictor by

$$p(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

For the current study, let  $Y_{it}$  be the observed number of minor injuries in road traffic accidents on the  $i$ th road segment and at the  $t$ th day,  $t = 1, \dots, T$ . The average daily traffic intensity for individual road segment is represented by  $N_{it}$ . We assume that conditional on the relative risk,  $\rho_{it}$ , the number of observed events follows a binomial distribution

$$Y_{it} | \rho_{it} \sim \text{Binomial}(N_{it}, \rho_{it})$$

where the log-risk is modeled as

$$\text{logit}(\rho_{it}) = \beta_0 + Z_i^T \beta_i + S(x_i) + \delta_t + \epsilon_i \tag{2}$$

Here,  $S(x_i)$  and  $\delta_t$  account for the spatially and temporally structured random effects, respectively and  $\epsilon_i$  stands for an unstructured zero mean Gaussian random effect and logbeta precision parameters 0.5 and 0.01, defined as penalized complexity priors (Simpson et al., 2017).  $Z_i$  represents all covariates included in the model. We assigned a vague prior to the vector of coefficients  $\beta = (\beta_0, \dots, \beta_p)$  which is a zero mean Gaussian distribution with precision 0.001. All parameters associated to log-precisions are assigned inverse Gamma distributions with parameters equal to 1 and 0.00005.

To compute the joint posterior distribution of the model parameters, we use an INLA-SPDE method, as introduced by Lindgren et al. (2011). SPDE consists in representing a continuous spatial process, such a Gaussian field (GF), using a discretely indexed spatial random process such as a Gaussian Markov random field (GMRF). In particular, the spatial random process represented by  $S(\cdot)$  explicitly denote dependence on the spatial field, follows a zero-mean Gaussian process with Matérn covariance function represented as

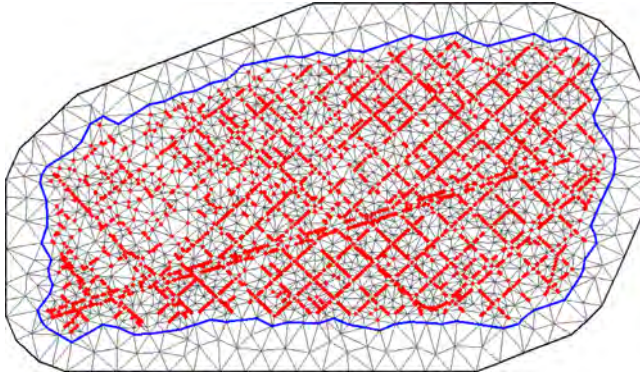
$$\text{Cov}(S(x_i), S(x_j)) = \frac{\sigma^2}{2^{\nu-1} \Gamma(\nu)} (\kappa \|x_i - x_j\|)^\nu K_\nu(\kappa \|x_i - x_j\|) \tag{3}$$

where  $K_\nu(\cdot)$  is the modified Bessel function of second order, and  $\nu > 0$  and  $\kappa > 0$  are the smoothness and scaling parameters, respectively. INLA approach constructs Matérn SPDE model, with spatial range  $r$  and standard deviation parameter  $\sigma$ . The parameterized model we follow is of the form

$$(k^2 - \Delta)^{(\alpha/2)} (\tau S(x)) = W(x)$$

where  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$  is the Laplacian operator,  $\alpha = (\nu + d/2)$  is the smoothness parameter,  $\tau$  is inversely proportional to  $\sigma$  and  $W(x)$  is a spatial white noise and  $\kappa > 0$  is the scale parameter, related to range  $r$ , defined as the distance at which the spatial correlation becomes small. For each  $\nu$ , empirically derived definition  $r = \sqrt{8\nu}/\kappa$  with  $r$  corresponding to the distance where the spatial correlation is close to 0.1 (Blangiardo and Cameletti, 2015). Note that we have  $d = 2$  for a two-dimensional process, and we fix  $\nu = 1$ , so that  $\alpha = 2$  in our case. We report that, heterogeneity, unobserved factors specific to each accident, although invariant over time, were captured by an identical and independently distributed random effect of zero mean and constant variance.

The temporal random effect ( $\delta_t$ ) is assumed to be a smoothed function, in particular a random walk of order one (RW1) (Rue et al., 2009). On the other hand, INLA-SPDE requires a triangulation or mesh structure to interpolate discrete event locations to estimate a continuous process in space (Rue et al., 2017). We use the centroids of each road segment as the target locations over which we



**Fig. 2.** Region mesh with non-convex hull boundary in blue and data locations highlighted as red points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

build the mesh. A detailed description of building a Delaunay's triangulation with emphasis on a network mesh is shown in Section 2.2. Centroids of individual road segments and the triangulations in the mesh are used to generate the projection matrix. We assign penalized complexity priors for the parameters to create INLA-SPDE model object for the Matérn model (Simpson et al., 2017). In the parametrization process we set prior according to hyperparameters for range as *prior.range* (0.01, 0.01) and standard deviation as *prior.sigma* (1, 0.1).

#### SPDE network triangulation:

To begin, we use the traditional SPDE method to triangulate the entire study area by considering the boundary for continuous spatial structure. According to Verdoy (2019), the best fitting mesh should have enough vertices for effective prediction, but the number should be within a limit to have control over the computational time. Following this principle, from a battery of meshes the best fitting mesh is selected having 2352 vertices. Fig. 2 depicts region SPDE mesh with 11,067 traffic accident locations highlighted as red points.

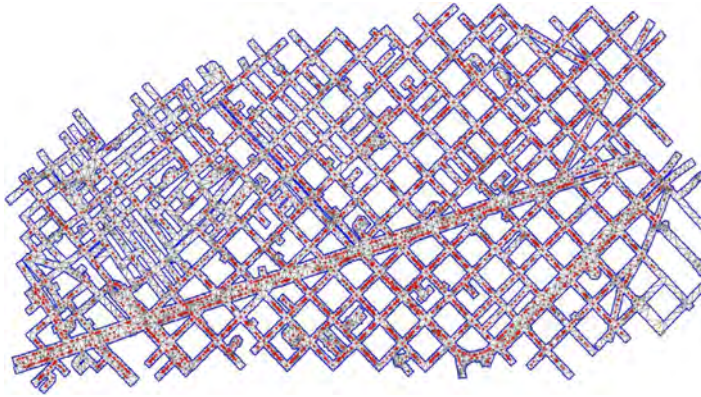
While fitting the mesh (as depicted in Fig. 2) a problem appears. Although the sampled traffic accidents are discrete spatial sites situated exclusively on the road networks, the models fitted with the mesh span the entire study area. So, it is not realistic and ambiguous for the model prediction to provide results in areas without a road network where traffic accidents are unlikely to occur. This leads to the motivation of designing SPDE triangulation precisely on road networks. The technique is carried out in three steps: first, a buffer region is generated for each road segment, then a clipped buffer polygon is created that only includes the area covered by the road network, and finally, SPDE triangulation is applied to the clipped polygon to create *SPDE Network Mesh*. We use R package rgeos (Bivand et al., 2017) to build buffers for individual road segment. While selecting the buffer size we need to balance between number of vertices used to build the triangulated mesh and computational cost (Krainski et al., 2018; Verdoy, 2019). We tested several buffers before deciding on a 15 meters buffer as the optimal one. In the next step, we merge individual buffer segments and convert them into a single polygon clipped within a bounding box covering the study area. The clipped polygon of the buffered segments is depicted in Fig. A.10, with accident locations highlighted as red points.

In the final step of building the proposed network mesh we use the centroids of each segment as initial Delaunay's triangulation nodes on the clipped polygon. Fig. 3 depicts the SPDE mesh precisely created on the road network, with accident locations highlighted in red.

#### Risk map design:

In this section we discuss about the process of designing the traffic accident risk map for the entire road network of the study area. The predicted daily number of minor injuries on individual





**Fig. 3.** Network mesh with data locations highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

road segment obtained using the binomial model has been considered as the initial Risk Score ( $R_{score}$ ) on that particular segment for that day. According to the literature on road safety before modeling any risk map, a predetermined category range must be determine (Curran-Everett, 2013). Keeping this in mind, we follow the European Road Assessment Programme (EuroRAP) standard to create the risk ratings of the motorways and other national roads in Europe (‘Risk mapping for the TEN-T in Croatia, Greece, Italy and Spain: Update’, 2016). We implement these 0 to 4 categorical values ranging from low risk to high risk defining what we call a Risk Index. We thus consider a normalization for the raw risk scores following a dynamic normalization technique. Initially, the risk range ( $R_{range}$ ) is calculated as

$$R_{range} = \frac{(max.R_{score} - min.R_{score})}{no. \text{ of risk categories}}$$

Next, we use  $R_{range}$  and  $R_{score}$  values in the proposed metric system to calculate the normalized risk index categories. As a relevant example in Table 1, we depict the values of categories in the normalized scale considered as the proposed risk categories to design the traffic risk map. We note that the metric system can be replicated using any other alternative risk index.

We depict an example of how risk index values are calculated using the proposed normalizing metric reported in Table 1. Consider the following scenario: a user wants to calculate the risk index for a specific road segment where four minor injury cases from traffic accidents have been recorded on a given day. So, the  $R_{score}$  of that segment is 4. Based on the traffic accident records for all the road segments on that particular day, assume that the maximum  $R_{score}$  is 15 and the minimum  $R_{score}$  is 0. In the present study, five risk categories have already been considered (i.e., the risk index values 0–4). Using Equation 4, we can calculate  $R_{range}$  is 3. Now, we can report that the risk index for the example road segment will match with the second normalize condition, indicating that it is a low–medium risk road segment with risk index value 1 for that particular day considered in the example. In this process, on the same day different segments can have diverse set of risk index values depending on their individual  $R_{score}$ . Alternatively, the risk index value for the same segment can vary on different days and months.

The risk-index algorithm implemented here has intended to categorize road segments based on the records of number of minor injuries incurred in each segment. As a result, segments having higher minor injury counts are categorized as accident-prone or high-risk roads. Other traffic risk modeling algorithms can use a similar concept.

**Table 1**  
Normalization metric for risk index values.

Normalize condition	Risk index	Safety measure
$R_{score} = 0$ or $R_{score} < R_{range}$	0	Low risk
$R_{range} \leq R_{score} < 2 \times R_{range}$	1	Low-medium risk
$2 \times R_{range} \leq R_{score} < 3 \times R_{range}$	2	Medium risk
$3 \times R_{range} \leq R_{score} < 4 \times R_{range}$	3	Medium-high risk
$4 \times R_{range} \leq R_{score}$	4	High risk

**Table 2**  
Fitted model DIC, WAIC and CPO values.

Model mesh	DIC	WAIC	CPO
Region mesh	23687.41	23674.25	0.3246
<b>Network mesh</b>	<b>23654.73</b>	<b>23647.06</b>	<b>0.3243</b>

*Inference:*

Inferences are made following a Bayesian perspective, using the INLA approach (Rue et al., 2009, 2017). We used priors that penalize complexity (called PC priors). These priors are robust in the sense that they do not have an impact on the results and in addition the notion of scale determines the magnitude of effects and simplifies interpretation of results (Simpson et al., 2017).

All analyses are carried out using the free software R (version 4.1.2) (R. Core Team, 2021), through the INLA package (Rue et al., 2009; Lindgren, 2012; Rue et al., 2017, ‘R. INLA Project’, 2020). Maps related to study area, Barcelona street network with accident locations and traffic intensity map are designed using ArcGIS Desktop Software (version 10.8. Redlands) (ESRI, 2021). Other maps to depict SPDE mesh generation technique and plotting risk maps are designed using R package mapview (Appelhans et al., 2016).

**3. Results**

In this section, we present the results of the methodological approach developed in Section 2. We provide results on model fitting, validation, and prediction along with risk maps of accidents. The proposed model (mentioned in Eq. (2)) is fitted using both region mesh as depicted in Fig. 2 and our proposed network mesh as depicted in Fig. 3. Both models are fitted to the daily accident records for the years 2010 to 2018. The remaining accident records of 2019 have been used to test the fitted model. It is worthy to mention that we have executed series of similar models for both categories of mesh, using different dimensions of SPDE meshes with different combinations of covariates. In each case, deviance information criterion (DIC) and the Watanabe–Akaike information criterion (WAIC) are used to assess the performance of the models, and to select the best fitting model by balancing model accuracy against complexity (Spiegelhalter et al., 2002). Models having smaller WAIC value, despite the added complexity, provide a more appropriate fit to sampled data (Blangiardo and Cameletti, 2015). In Table 2 we report the summary results (DIC, WAIC and CPO) related to goodness-of-fit along with computational time (in seconds) for only the best fitted models from individual region mesh and network mesh categories. We note the computational time (in seconds) of the best fit model with region mesh (336) is substantially lower compared to the best fit model with proposed network mesh (3187). This can be explained due to higher number vertices in network mesh (14,368) than the region mesh (2352).

DIC values shown in Table 2 indicate that the WAIC value of model with proposed network mesh (23,647.06) is lower compared to the other. Thus, to model the spatio-temporal structure of traffic accidents on road networks of Barcelona, the binomial model with SPDE network mesh is selected. We additionally note that the model is best by considering random spatial and temporal effects together with set of covariates mentioned in Section 2.1. When the covariates are not considered, the model provided larger DIC and WAIC values.

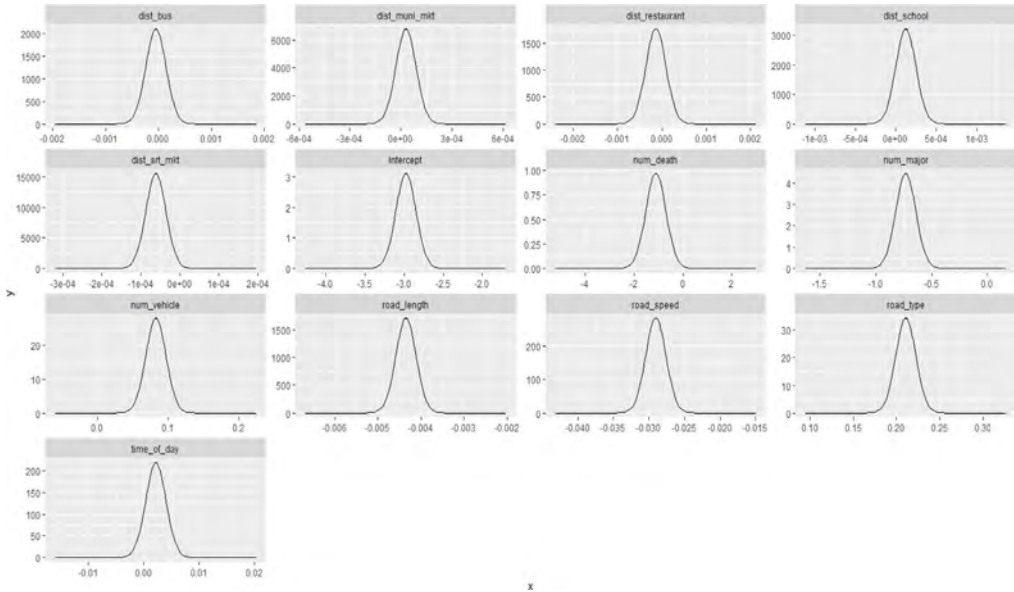


Fig. 4. Marginal posterior distributions of covariate coefficients.

The posterior distribution of fixed and random effects included in the model are depicted in Figs. 4 and A.11 (in Appendix). In particular, Fig. 4 shows the marginal posterior distributions of all fixed effects. Additionally, Table 3 depicts the coefficients and 95% of credibility intervals of all fixed effects.

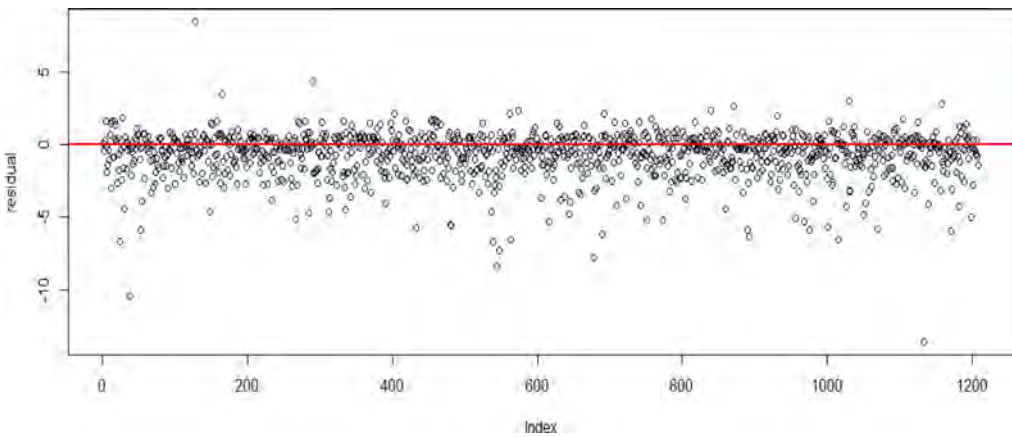
We note that six covariates, namely time of day, distances from nearest bus stop, school, restaurant, municipality, and street markets have no influence in our model. The positive mean values for covariates such as number of vehicles involved in accident and type of road indicate positive influence in the model. It is worthy to note that when the number of causalities and major injuries in individual traffic accident are high then the record of minor injuries is lower compared to other cases, thus there exists a negative association for these two covariates in our model. The covariate associated to number of causalities has the highest negative mean value which indicates strong negative influence on the model. Additionally, road length and speed limit of individual road show negative influence in our model. Indicating when the road length is high and speed limit is also high the number of minor injuries in a collision is comparatively lower than other short roads and low speed tracks.

Moreover, Fig. A.11 in the Appendix depicts the marginal posterior mean of spatial  $S(\cdot)$  and temporal  $\delta_t$  random effects with 95% credible intervals. The horizontal axis of Fig. A.11 (top) in the Appendix represents the 14,368 triangulation nodes of the SPDE network mesh used in the model. A stronger and significant spatial effect is observed on the vertices of triangles on road segments having higher traffic accident occurrence (highlighted in Fig. 3 as dark red patches). The vertices without accident events show no spatial effect. Similarly, Fig. A.11 (bottom) in the Appendix exhibits the variation of the marginal posterior mean of the daily temporal random effects over the entire study period (2010 to 2019).

We finally report that the spatial effect parameters  $\kappa$  and  $\tau$  have mean values 92.16 and 0.31 as depicted in Fig. A.12 (in the Appendix) that shows the marginal posterior distributions of the two hyperparameters for the spatial random field. Using  $\tau$  and  $\kappa$  we can get the value of spatial range  $r = 0.055$  km or 55 m. Using the fitted model, we can analyze the goodness-of-fit of the model by

**Table 3**  
Marginal posterior mean and credible interval of fixed effects.

Covariate	Mean	Credible interval
Nearest bus stop distance	0.002	−0.001, 0.003
Nearest municipality market distance	0.002	−0.001, 0.004
Nearest restaurant distance	0.000	0.000, 0.000
Nearest school distance	0.000	0.000, 0.000
Nearest street market distance	0.003	−0.001, 0.007
<b>Number of deaths</b>	<b>−1.108</b>	<b>−1.915, −0.301</b>
<b>Number of major injuries</b>	<b>−0.730</b>	<b>−0.906, −0.555</b>
<b>Number of vehicles</b>	<b>0.083</b>	<b>0.055, 0.110</b>
<b>Road length</b>	<b>−0.004</b>	<b>−0.005, −0.004</b>
<b>Road speed limit</b>	<b>−0.029</b>	<b>−0.032, −0.026</b>
<b>Road type</b>	<b>0.211</b>	<b>0.188, 0.234</b>
Time of day	0.002	−0.001, 0.006



**Fig. 5.** Residual (observed minus predicted) plots.

considering prediction over unsampled locations (Zuur et al., 2017). The fitted model is projected into the mesh at each road segment for this prediction.

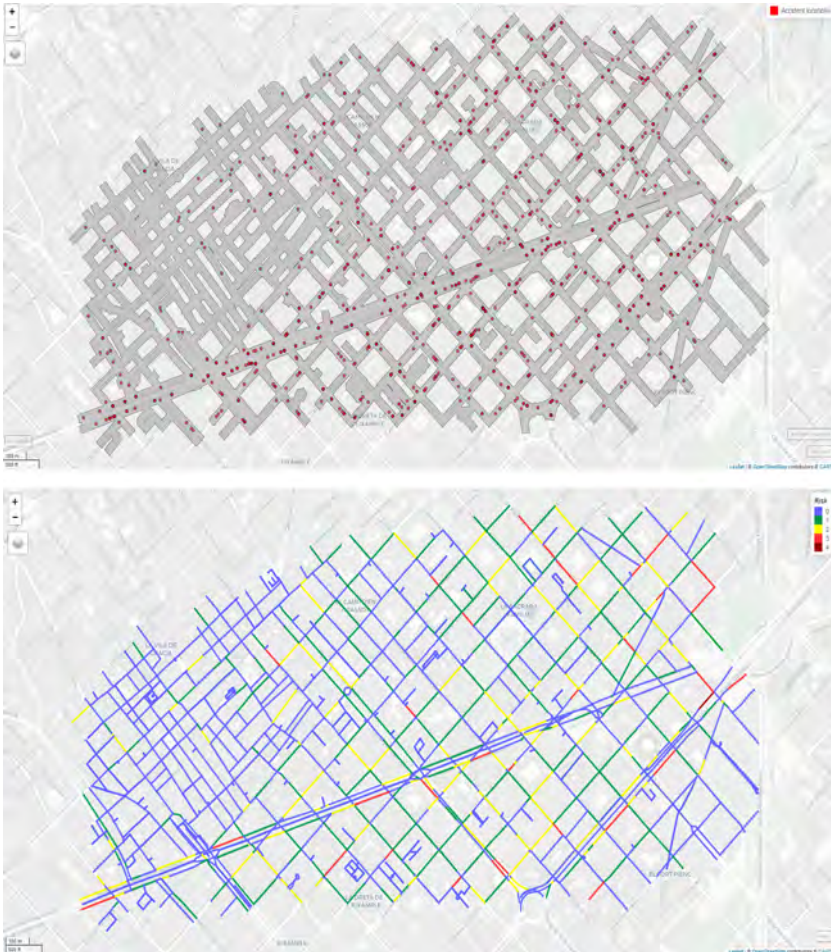
The proposed model is tested using accident records for the final study year 2019 combined with the entire model fitting 2010–2018 dataset. We compute predictions as well as their corresponding residuals (observed minus predicted). Fig. 5 depicts such residual plots. We note the residual values are generally close to zero and have no discernible structure.

*Risk map:*

We follow the risk map metrics in Section 2.2 and use safety measure scale shown in Table 1 to calculate the risk index for individual road segments. The predicted values for 2019 are used to construct the normalized risk index values.

Fig. 6 (top) shows the location of 1209 original road accident in 2019. The corresponding projected risk map for 2019 as a whole is shown in Fig. 6 (bottom). The color scales (0–4) used in the risk map follow the same safety measure scale used in Table 1. A visual comparison of the predicted risk map with the original road accident record shows that the road segment containing the observed cluster of accidents is correctly predicted by the risk map as medium to high-risk roads.

Similarly, roads that are predicted to have low or moderate risk are originally roads with no or very few incidents. We report that similar results are observed while comparing the predicted risk map for individual dates in 2019 with the corresponding original accident records.



**Fig. 6.** Top: Observed traffic accident events recorded in 2019.  
Bottom: Predicted risk map for 2019.

The current results show that the proposed model can produce the road safety index of all road segments, including small details of each junction or sharp turnings. In addition, identifying potentially dangerous roads can serve as baseline for geographic analysis of road safety management. The daily predicted risk maps can have strategic applications in developing GIS analytical tools to identify and depict possible safe routes. For example, in Fig. 7 (top) the start and destination points of a particular user is highlighted by green and red map pins. The user can choose path B which is considered to be shorter in length compared to path A. But in terms of safety measure the user should opt for path A as the cumulative risk index for this path is much less compared to the shorter path B. As the proposed dynamic risk map provides information about the entire road network, it is flexible enough to generate possible alternative safe route(s) between any source and destinations pairs as depicted in another similar example in Fig. 7 (bottom). In this example the length of both the roads between source and destination points are same but in terms of risk index path B is relatively unsafe compared to path A.

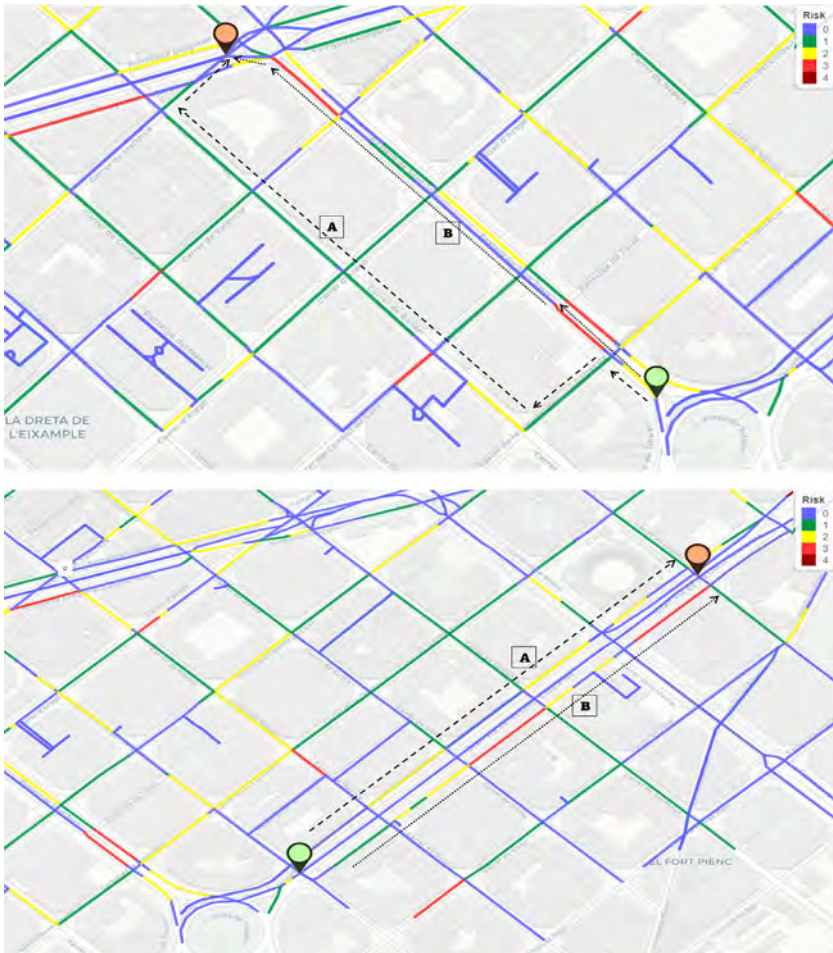


Fig. 7. Predicted risk map baseline to identify the safe route. Top: Example 1, Bottom: Example 2.

#### 4. Discussion

In recent years, spatiotemporal modeling of traffic accidents and risk mapping has gained attention, particularly in the domain of multi-dimensional road safety management. INLA-SPDE approach involves projecting the fitted model into the mesh at precise spatial locations, in this case the sampling points (here accidents locations) are located only on the road network. But traditional SPDE mesh is generated for the entire study area which includes road network as well as other regions. In that case, the output of the model may be unavoidably generalized because it will estimate predicted values for regions where there is no chance of an incident occurring. To avoid this ambiguity, we introduce the novel concept of designing the SPDE triangulation precisely on the road network. As a result, instead of performing spatio-temporal predictive analysis on the full continuous region, the proposed INLA-SPDE modeling took a new step by executing it only on selected sections (particularly for road networks).

In this context, Chaudhuri et al. (2022) recently proposed spatiotemporal modeling of road traffic accidents using explicit network triangulation. However, the study only used types and surfaces

of roads as covariates in the model. Significant exogenous variables related to traffic flow, traffic control and temporal variables such as time of accident occurrence and other spatial covariates are ignored in the modeling process. Furthermore, no analysis to compare and identify alternative safer roads is performed in the previous study. Whereas the results of the current study show that, when fitted with selected covariates, the proposed model can generate predicted risk maps of the entire road network for any urban study area. In that sense, the model proposed in this study is dynamic in nature. Additionally, Bayesian methodology is implemented using computationally faster INLA-SPDE approach where the number of covariates can be updated at any stage and the level of significance for individual covariates can be analyzed for further emphasis on the selection of significant factors causing traffic accidents.

Regarding the limitations and future works, we note that the buffer road network used in the current study has complex boundary regions which have some influence on the spatial effect of the model. According to Krainski et al. (2018), the first step in fitting an SPDE model is to create a mesh to represent the spatial process. Building SPDE mesh for a continuous region is relatively easier compared with the proposed SPDE network mesh. In the current study, fine tuning is required to identify the best fit values for minimum allowed distance between vertices and maximum permissible triangle edge length for the inner (and outer) regions. Moreover, additional points around the boundary, or outer extension, must be selected with care. As a rule of thumb, the variance near the boundary is inflated by a factor of 2 along straight boundaries and by a factor of 4 near right-angled corners (Lindgren and Rue, 2015). The complex boundary region of the buffer road network with several right-angled corners (as depicted in Fig. A.10) makes the process critical. The boundaries in the proposed mesh are located inside the mesh and not outside, as in a standard mesh and that creates fictitious spatial structures. Because of this complex boundary nature, it is unavoidable to reduce high boundary effect that might cause a variance twice or four times as great at the border as it is within the domain (Lindgren, 2012; Lindgren and Rue, 2015). Additionally, though the residual diagnostics and predicted risk maps produced by the model matches with the original observed records; but the correlation values of the model indicate room for improvement. Thus, for detailed understanding of the performance of the model, it may be beneficial to analyze further the model fitting phase using INLA-SPDE with a diverse set of spatial and temporal covariates, spatial and temporal structures, and space-time interactions. This paves the way for future research works in this domain and to reduce the boundary effects in the model results.

The final outcomes of the proposed model are predicted risk maps for the entire road network. Thus, using these maps, the road safety index of individual road segments, including small details of each junction point or sharp turn, can be obtained at a glance. This can act as baseline information for geospatial analysis on road safety metrics to design strategic geographical information system (GIS) analytical tools to identify and depict possible safe routes as depicted in Fig. 7. Similar research work by Hannah et al. (2018) considers only spatial traffic variables like speed limits, street junctions, and type of street. But the current study has proposed a more flexible and statistically convincing solution by implementing both spatial and temporal covariates in the predictive model. We note another crucial application of the model is in analyzing change and trend pattern of traffic accidents. Fig. 6 is the combined predicted traffic risk map for all days in 2019. As mentioned in Section 3 similar dynamic risk maps can be developed for individual months, weeks or even days. Using these maps trends in traffic accident risk can be identified for individual roads or, road junctions. A better understanding of these patterns may have implications for road safety measures. Identification of gradual changes in risk patterns and related potential factors, is of interest for future research works on change point detection.

Interestingly, the predicted risk maps can be used as an important guideline for traffic management authorities to identify potentially dangerous roads in any urban region and can take strategic measures and actions to prevent traffic accidents in advance. As a result, risk maps can be used to better understand accident hotspots, improve traffic safety measures, and conceivably can have an impact on public health by reducing traffic accidents. In addition, as the model is flexible and general, it can be applied to a wide range of related problems. The current methodology can be implemented for smart transportation systems by predicting traffic flow and reducing congestion

on roads. This would enable transport authorities to better manage the traffic condition during peak hours and would allow users to choose the best routes to their destinations. In context to smart cities, an intelligent traffic management system based on the proposed model can feasibly control air pollution caused by fine particulate matter emitted by transportation. It can have potential implications to achieve air quality levels for particles in suspension in line with the guideline value of the World Health Organization (WHO, 2019).

Consequently, we report that the proposed modeling approach could be a major step forward in the understanding of road safety measure and can act as a baseline in strategic decision making to control traffic collisions. The novel contribution of this work is that it is able to take advantage of INLA-SPDE approach precisely on road network rather than for continuous region. As a result, the model can predict risk factors for individual road segments and generate dynamic risk maps for the entire network. In conclusion, although it may be complicated to control the boundary effect in the complex network triangulation method, this work is able to present a model that can provide accurate predictions of accident-prone roads and help in identifying alternate safe routes between any source and destination pairs.

## 5. Conclusion

The current study implements Bayesian methodology by including INLA and SPDE to design a dynamic spatio-temporal analysis model predicting the occurrence of traffic accidents in the city of Barcelona, Spain. The use of SPDE network triangulation to estimate the spatial auto-correlation of discrete events is novel in this study. The methodology used in the study is a new step to perform spatio-temporal analysis precisely on road network and contributes to the relatively small amount of literature in this domain. Moreover, the dynamic risk maps of traffic accidents are one of the interesting outcomes. The risk maps can have strategic applications in road safety measures and designing travel risk maps for tourists, corporate travelers, and emergency service providers. The methodology to identify safe routes is dynamic and can be adapted and applied to other locations globally. Furthermore, the current study opens future research scopes to explore the influence of boundary effects on model performance and analyze the variation in spatial effects. We are investigating these anomalies of the spatial impact in a subsequent study project and working on a possible solution to the problem.

## CRedit authorship contribution statement

**Somnath Chaudhuri:** Original idea for the paper, Bibliographic search, Writing of the introduction, Data collection and cleaning, Statistical analysis, Created the tables and figures, Writing – review & editing. **Marc Saez:** Designed the study, Bibliographic search, Writing of the introduction, Statistical analysis, Writing – review & editing. **Diego Varga:** Bibliographic search, Writing of the introduction, Data collection and cleaning, Created the tables and figures, Writing – review & editing. **Pablo Juan:** Designed the study, Bibliographic search, Writing of the introduction, Statistical analysis, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The pre-processed data and R code for implementing the proposed model can be made available upon request.



## Acknowledgments

We appreciate constructive comments and suggestions from Håvard Rue and Elias T. Krainski. We appreciate the comments of two anonymous reviewers of a previous version of this work who, without doubt, helped us to improve our work. The usual disclaimer applies. All authors reviewed and approved the manuscript.

## Appendix

### A.1. Data settings

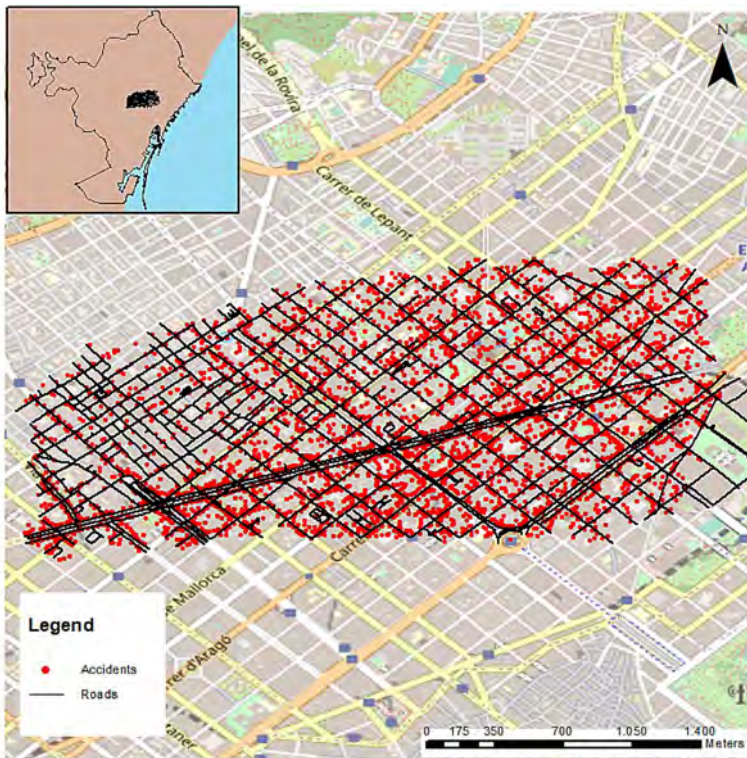
See Figs. A.8 and A.9.

### A.2. Methodology

See Fig. A.10.

### A.3. Results

See Figs. A.11 and A.12.



**Fig. A.8.** Location of road network of the study area with traffic accident locations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

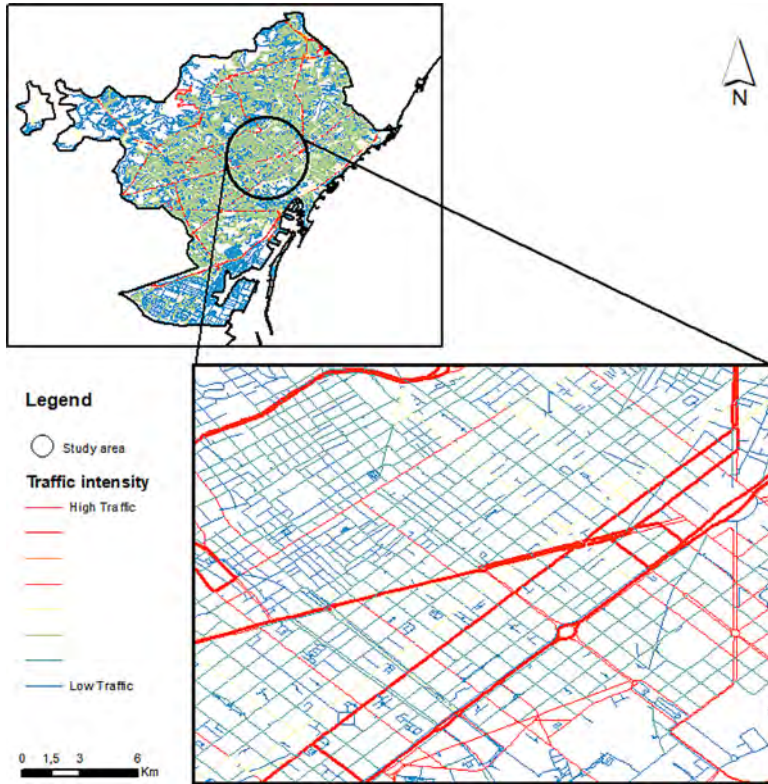


Fig. A.9. Traffic intensity on individual road segments of the study area.



Fig. A.10. Polygon of buffered road segments, red points indicate traffic accident locations on the road network. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

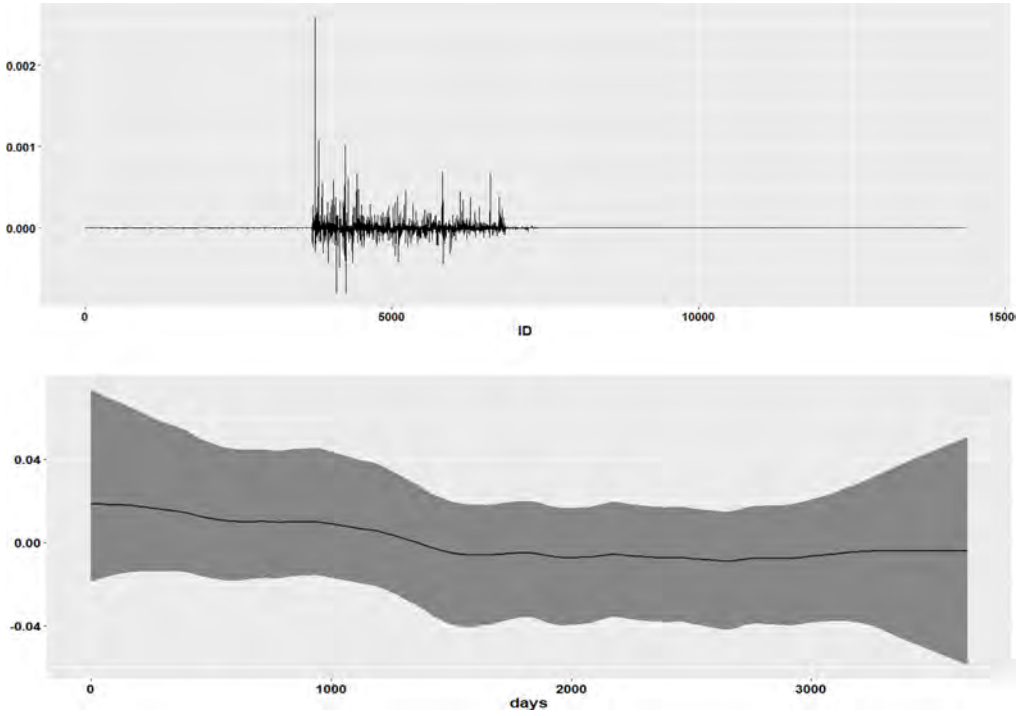


Fig. A.11. Top: Marginal posterior mean of the spatial random effect  $S(\cdot)$ . Bottom: Marginal posterior mean of the temporal random effect  $\delta_t$ .

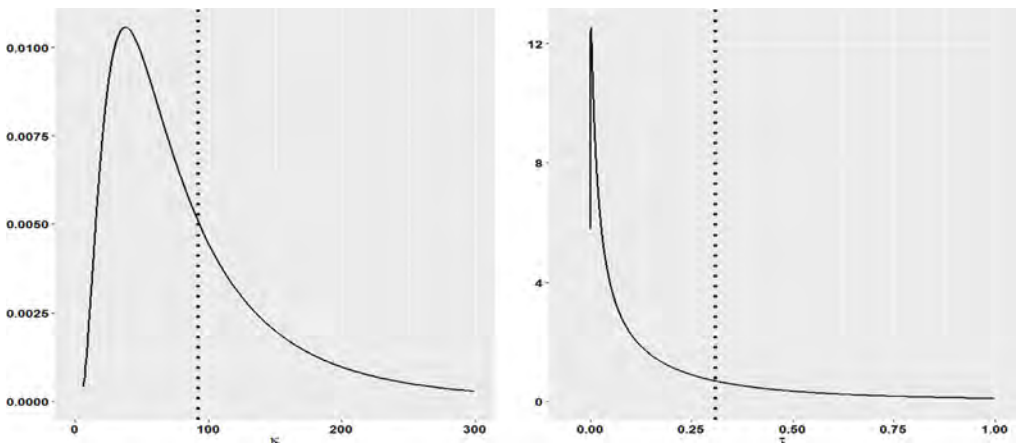


Fig. A.12. Marginal posterior distributions of hyperparameters  $\kappa$  and  $\tau$  for the spatial random field  $S(\cdot)$ .

## References

- Aghajani, M.A., Dezfoulian, R.S., Arjroody, A.R., Rezaei, M., 2017. Applying GIS to identify the spatial and temporal patterns of road accidents using spatial statistics (case study: Ilam province, Iran). *Transp. Res. Proc.* 25, 2126–2138.
- Appelhans, T., Detsch, F., Reudenbach, C., Woellauer, S., 2016. Mapview - Interactive viewing of spatial data in R. In: EGU General Assembly Conference Abstracts. EPSC2016-1832.
- Bivand, R., Rundel, C., Pebesma, E., 2017. RGEOS: Interface to geometry engine-open source (GEOS). R package version 0.3-26.
- Blangiardo, M., Cameletti, M., 2015. *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley & Sons Ltd.
- Boulieri, A., Liverani, S., Hoogh, K.de., Blangiardo, M., 2016. A space-time multivariate Bayesian model to analyse road traffic accidents by severity. *J.R. Stat. Soc. Ser. A (Stat. Soc.)* 180 (1), 119–139.
- Briz-Redó, Á., Martínez-Ruiz, F., Montes, F., 2019. Identification of differential risk hotspots for collision and vehicle type in a directed linear network. *Accid. Anal. Prev.* 132, 105278.
- Cantillo, V., Garcés, P., Márquez, L., 2016. Factors influencing the occurrence of traffic accidents in urban roads: A combined GIS-empirical Bayesian approach. *DYNA* 83 (195), 21–28.
- Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transp. Res. B* 46 (1), 253–272.
- Chaudhuri, S., Juan, P., Mateu, J., 2022. Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation. *J. Appl. Stat.* 1–22.
- Chaudhuri, S., Moradi, M., Mateu, J., 2021. On the trend detection of time-ordered intensity images of point processes on linear networks. *Comm. Statist. Simulation Comput.* 1–13.
- Curran-Everett, D., 2013. Explorations in statistics: The analysis of ratios and normalized data. *Adv. Physiol. Educ.* 37 (3), 213–219.
- ESRI, 2021. ArcGIS, Environmental System Research Institute (ESRI). <https://www.esri.com/>.
- Farmer, C.M., 2005. Temporal factors in motor vehicle crash deaths. *Inj. Prev.* 11 (1), 18–23.
- Galgamuwa, U., Du, J., Dissanayake, S., 2019. Bayesian spatial modeling to incorporate unmeasured information at road segment levels with the INLA approach: A methodological advancement of estimating crash modification factors. *J. Traffic Transp. Eng. (Engl. Ed.)*.
- Guo, Y., Osama, A., Sayed, T., 2018. A cross-comparison of different techniques for modeling macro-level cyclist crashes. *Accid. Anal. Prev.* 113, 38–46.
- Hannah, C., Spasić, I., Corcoran, P., 2018. A computational model of pedestrian road safety: The long way round is the safe way home. *Accid. Anal. Prev.* 121, 347–357.
- Jegade, F., 1988. Spatio-temporal analysis of road traffic accidents in Oyo state, Nigeria. *Accid. Anal. Prev.* 20 (3), 227–243.
- Juan, P., Mateu, J., Saez, M., 2012. Pinpointing spatio-temporal interactions in wildfire patterns. *Stoch. Environ. Res. Risk Assess.* 26, 1131–1150.
- Karaganis, A., Mimis, A., 2006. A spatial point process for estimating the probability of occurrence of a traffic accident. In: *ERSA Conference Papers*. European Regional Science Association.
- Khulbe, D., Sourav, S., 2019. Modeling severe traffic accidents with spatial and temporal features. In: *ICONIP*.
- Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H., 2018. *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. Chapman; Hall/CRC.
- Lindgren, F., 2012. Continuous domain spatial models in R-INLA. *ISBA Bull.* 19 (4), 14–20.
- Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* 63 (19), <http://dx.doi.org/10.18637/jss.v063.i19>.
- Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73 (4), 423–498.
- Liu, C., Sharma, A., 2017. Exploring spatio-temporal effects in traffic crash trend analysis. *Anal. Methods Accid. Res.* 16, 104–116.
- Liu, C., Zhang, S., Wu, H., Fu, Q., 2017. A dynamic spatiotemporal analysis model for traffic incident influence prediction on urban road networks. *ISPRS Int. J. Geo-Inf.* 6 (11), 362.
- Loo, B.P.Y., Yao, S., Wu, J., 2011. Spatial point analysis of road crashes in Shanghai: A GIS-based network kernel density method. In: *2011 19th International Conference on Geoinformatics*.
- Moradi, M.M., Mateu, J., 2019. First-and second-order characteristics of spatio-temporal point processes on linear networks. *J. Comput. Graph. Statist.* 1–21.
- OpenDataBCN, 2021. Open data BCN – Ajuntament de Barcelona's open data service. <https://www.opendata-ajuntament.barcelona.cat/en>.
- Prasannakumar, V., Vijith, H., Charutha, R., Geetha, N., 2011. Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia - Soc. Behav. Sci.* 21, 317–325.
- Pulugurtha, S.S., Sambhara, V.R., 2011. Pedestrian crash estimation models for signalized intersections. *Accid. Anal. Prev.* 43 (1), 439–446.
- R. Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- R. INLA Project, 2020. <https://www.r-inla.org>.
- Risk mapping for the TEN-T in Croatia, Greece, Italy and Spain: Update, 2016. <https://eurorap.org/risk-mapping-for-the-ten-t-in-Croatia-Greece-Italy-and-Spain-update/>.
- Rue, H., Held, L., 2005. *Gaussian Markov Random Fields: Theory and Applications* (Chapman & Hall. In: *CRC Monographs on Statistics and Applied Probability*), Chapman; Hall/CRC.

- Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71 (2), 319–392.
- Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K., 2017. Bayesian computing with INLA: A review. *Annu. Rev. Stat. Appl.* 4 (1), 395–421.
- Shafabakhsh, G.A., Famili, A., Bahadori, M.S., 2017. GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran. *J. Traffic Transp. Eng. (Engl. Ed.)* 4 (3), 290–299.
- Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. *Stat. Sci.* 32 (1), <http://dx.doi.org/10.1214/16-sts576>.
- Smedt, T.D., Simons, K., Nieuwenhuysse, A.V., Molenberghs, G., 2015. Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Arch. Public Health* 73 (S1), <http://dx.doi.org/10.1186/2049-3258-73-s1-o2>.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64 (4), 583–639.
- TomTom, 2021. Historical traffic data traffic stats. <https://www.tomtom.com/products/traffic-stats/>.
- UNDP, 2021. Sustainable development goals: United nations development programme. <https://www.undp.org/sustainable-development-goals>.
- Verdoy, P.J., 2019. Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. Choosing the best mesh for each database. *Commun. Stat.-Simul. Comput.* 1–34.
- Wang, W., Yuan, Z., Yang, Y., Yang, X., Liu, Y., 2019. Factors influencing traffic accident frequencies on urban roads: A spatial panel time-fixed effects error model (Y. Guo, Ed.). *PLoS One* 14 (4), e0214539.
- WHO, 2019. Global Status Report on Road Safety 2018. World Health Organization.
- Wikle, C.K., Berliner, L.M., Cressie, N., 1998. Hierarchical Bayesian space–time models. *Environ. Ecol. Stat.* 5, 117–154.
- Williamson, A.M., Feyer, A.-M., 1995. Causes of accidents and the time of day. *Work Stress* 9 (2–3), 158–164.
- Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accid. Anal. Prev.* 75, 16–25.
- Zhong-xiang, F., Shi-sheng, L., Wei-hua, Z., Nan-nan, Z., 2014. Combined prediction model of death toll for road traffic accidents based on independent and dependent variables. *Comput. Intell. Neurosci.* 2014, 1–7.
- Zuur, A.F., Ieno, E.N., Saveliev, A.A., 2017. *Beginners Guide to Spatial, Temporal and Spatial–Temporal Ecological Data Analysis with R-INLA: Using GLM and GLMM*, Vol. 1. Highland Statistics Ltd.

## 5.4 Natural Hazards in Islands. Nonstationary Approach with Barriers

### 5.4.1 Introduction

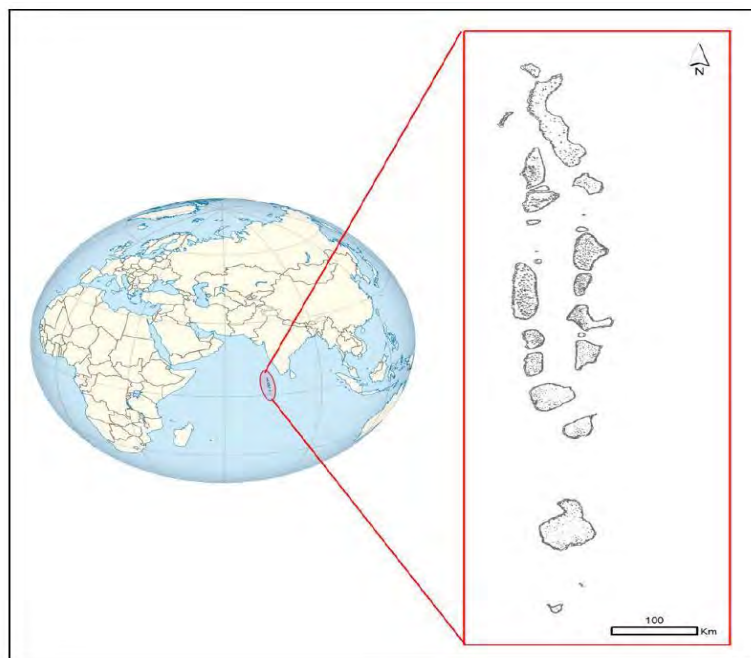
A 2019 United Nations report indicates that the world's population is expected to reach 11 billion by 2100 . By then, it is expected that nearly 75 percent of the population will live in urban areas. According to the scientific community, climate conditions would be very different from what we are currently experiencing. Even by 2030, there will be a 33 percent rise in the overall quantity of urban land in high-frequency flood zones compared to 2015 (Güneralp et al., 2015). From an environmental point of view, natural hazards represent a danger to ecosystems, directly affecting geomorphological and hydrological processes, as well as biodiversity. They also endanger human settlements with serious consequences for society (Zorn and Komac, 2013; Emmer, 2018). All of these facts demonstrate how crucial it is to assess natural hazard risks with an eye towards the future, in order to make well-informed decisions now about the spatial planning and risk prevention initiatives that will influence the coming decades. International organizations, such as the World Bank, the European Union (EU), and the United Nations (UN), among others, are aware of the necessity to take into account the long-term effects of natural disasters. Additionally, according to the intergovernmental panel on climate change (IPCC), "successful risk reduction and adaptation strategies consider the dynamics of vulnerability and exposure and their relationships with socioeconomic processes, sustainable development, and climate change" (IPCC, 2021).

A large number of existing studies in the broader literature have examined the impact of extreme climatic conditions on natural disasters (Sauerborn and Ebi, 2012; Phillips et al., 2015; Chaudhuri et al., 2021; Osberghaus and Fugger, 2022; Raju et al., 2022). As a result of unpredictable climate change, population growth, and increasingly urbanized societies, a better understanding and prediction of natural disasters has become undoubtedly important (SafarianZengir et al., 2019). Modeling natural disasters is crucial to characterize these phenomena and provide tools to overcome them. Estimating natural hazards, including spatial effects and local conditions, will help in management and even allow anticipation of events (Cutter and Finch, 2008). The 2015-2030 Sendai framework for disaster risk reduction recognizes this need and emphasizes the significance of having strategies in place to mitigate uncontrolled development in hazardous areas in order to better prepare for the disasters that our planet may experience in the future (UNDRR, 2015).

Several studies attempt to model natural hazards in order to examine global risk assessments (Morjani et al., 2007; Serra et al., 2013; Calkin and Mentis, 2015; Riley et al., 2016; Pittore et al., 2017; Sarkissian et al., 2020). Some of them address propagation of tsunami and its impact (Sarri et al., 2012; Hayashi et al., 2013; Sugawara, 2017; Shao et al., 2019; Rezaldi et al., 2021). In this line, to analyse spatial distribution of regions affected by natural hazards, statistical inference comes along with Bayesian methodology. Generally, Bayesian statistics are utilized in natural hazards engineering to deal with large-scale problems that involve different types of data inputs and explicitly handle uncertainties (Zheng et al., 2021). For example, studies like (Gaume

et al., 2010; Costa and Fernandes, 2017; Han and Coulibaly, 2017; Barbetta et al., 2018; Bolle et al., 2018) provide comprehensive review of the applications of Bayesian statistics in flood assessment and monitoring. Besides, Grezio et al. (2010) have presented the challenges in selecting proper models to quantify the uncertainties in the maximum tsunamigenic magnitudes. Literature shows application of Bayesian inference in analysing probabilistic tsunami hazards (Knighton and Bastidas, 2015; Risi and Goda, 2017; Smit et al., 2017).

A Bayesian approach with Markov chain Monte Carlo (MCMC) simulation methods has traditionally been used to fit generalized linear mixed models (GLMM) (Wikle et al., 1998). In this context, Shin et al. (2015) have explored the application of Bayesian MCMC method to estimate extreme magnitude of tsunamigenic earthquake. However, MCMC models require considerable computing time for large datasets (Rue et al., 2009; Smedt et al., 2015). Hence, instead of using MCMC, we can use Integrated Nested Laplace Approximation (INLA) methodology, developed by Rue et al. (2009), as it offers short computational time and is much easier to fit complex models (Ruiz-Cárdenas et al., 2012). As recommended by Rue et al. (2009), while processing spatial data we can utilize INLA in conjunction with stochastic partial differential equations (SPDE) to balance speed and accuracy of the models.



**Figure 14: Republic of Maldives geographical location and island structure**

However, there are limited contributions using INLASPDE approach in the context of natural disasters such as earthquakes or tsunamis. Recently, Wilson (2020) used Bayesian spatial modeling with INLA to analyse earthquake damages from geolocated cluster data. To date, no literature has documented applications of INLA-SPDE to tsunamis in the particular territory of the Maldives which involves a very advanced methodology to handle with its irregular and complex land structure (Riyaz and Suppasri, 2016). Analyzing tsunami propagation at the island scale is essential to develop well-informed policies for disaster management and to design

effective countermeasures. But due to the large domain and high resolution required for modelling, it is also challenging to study tsunami propagation across multiple atolls at the island scale ([Rasheed et al., 2022](#)).

The current study is conducted to model and estimate the spatial autocorrelation of tsunami data in the islands of Maldives. The initial point of this work is to explore the application of SPDE with INLA for Maldives tsunami data ([HDX, 2022](#)), including the mesh for spatial effect. Maldives consists of 1200 dispersed islands on both sides of the equator. Collection of these islands along with lagoon and reef areas form the complex atoll system of the country as depicted in [Figure 14](#). It is worthy to mention that the traditional SPDE method triangulates the entire study area based on continuous geographic boundaries ([Krainski et al., 2018](#)). A problem arises while designing the mesh for the entire country's boundary region or, for boundaries of individual atolls. Although the records of tsunami-affected regions are discrete spatial sites situated exclusively on land on reefs for individual islands, the mesh are generated for entire region inside the geographical boundary including the lagoon and sea surface. So, it is not realistic and ambiguous especially when we are interested to explore the spatial correlation of tsunami data precisely on the land regions. This leads to the motivation of designing SPDE triangulation precisely on land on reefs for individual islands. However, a stationary model cannot be aware of the coastline and the island boundaries and will inappropriately smooth over the features. In spatial modelling, classical models are unrealistic when they smooth over holes or physical barriers. This might result to another unrealistic assumption. In the research work by [Bakka et al. \(2019\)](#) a new nonstationary model has been constructed for INLA having syntax very similar to the stationary model. The model, named as barrier model is more realistic with both sparse data and complex barriers and computational cost is the same as for the stationary models ([Bakka et al., 2019](#)). To apply in complex island structures, barrier model has been designed considering water as normal terrain and it is aware of the distinct coastlines and boundaries considered as physical barriers. In the present study, we have explored the barrier model in a converse mode where water bodies (ocean and lagoons) act as barriers for the dispersed islands and natural hazards are the sample events considered precisely on the land area of the islands.

The motivation of this study is two-fold. On one side we provide a modeling framework to explore and analyze at the island scale the spatial variation in the incidence of tsunami. In particular, the occurrence of tsunami is analyzed with three spatial modeling scenarios using mesh for entire geographical boundary of atolls, mesh precisely on land on reefs for islands in individual atolls and barrier models for atolls. The second aim roots in providing an advanced and realistic computational strategy to design and customize meshes and nonstationary barrier model to examine the spatial dependencies of natural hazards in complex distributed land structure like the islands of Maldives. It can be applied in diverse sectors where complex physical barriers are present in road network ([Dawkins et al., 2021](#)), disease control interventions ([Cendoya et al., 2022](#)) and categorical areas with different land use.

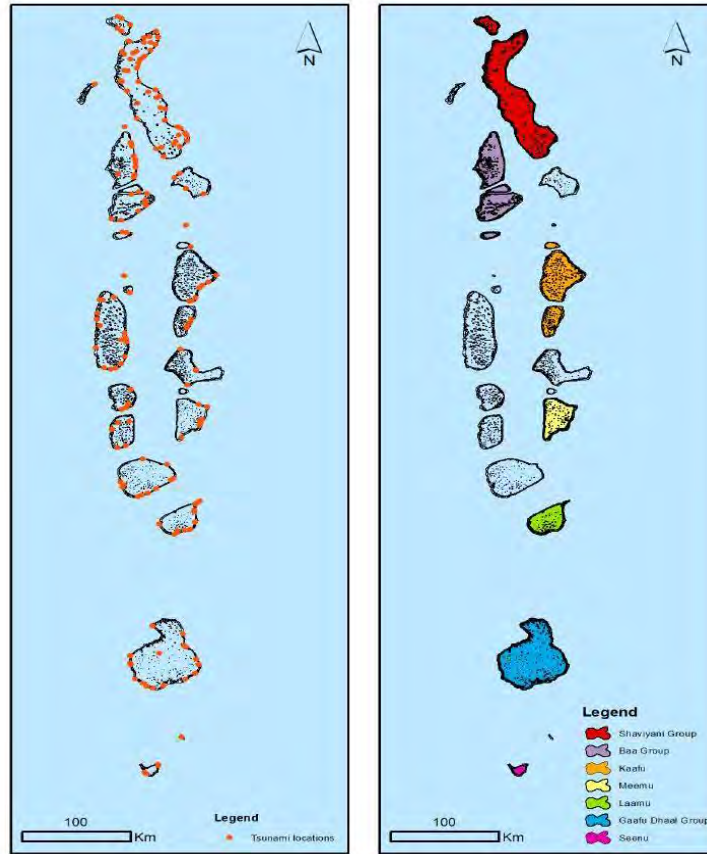
The study is conducted on 190 records of tsunami affected islands from the year 2004. R (version R 4.1.2) programming language ([R Core Team, 2022](#)) has been used for statistical computing and graphical analysis. As part of the data cleaning process and to design some maps,



we have used ArcGIS Pro (version 3.0.1) (Redlands, 2022). All computations are conducted on a quad-core Intel i9-4790 (3.60 GHz) processor with 32 GB (DDR3-1333/1600) RAM.

### 5.4.2 Data Settings

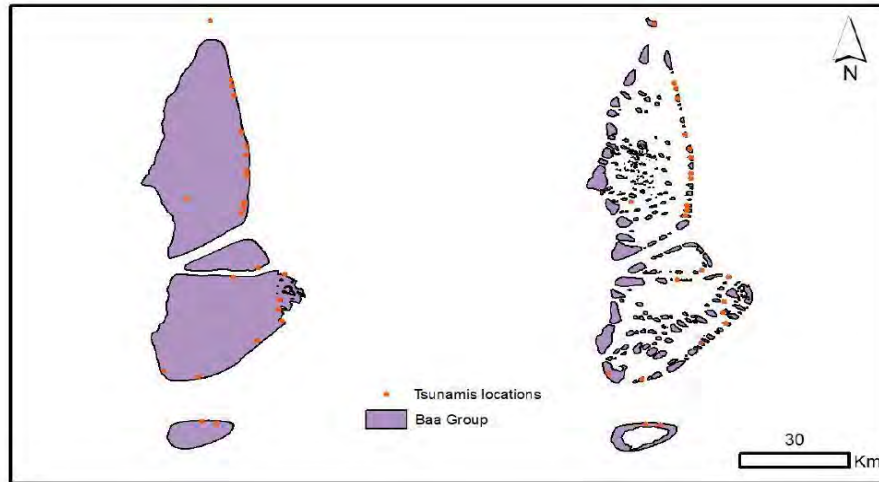
The Republic of Maldives is located in the south-western region off the coast of India in the Indian Ocean. This unique island nation is one of the smallest countries in Asia having a chain of coral islands across an archipelago more than 800 kilometers long and 130 kilometers wide. The archipelago consists of about 1190 coral islands grouped into 20 natural atolls. Out of which 189 islands are inhabited (Isles, 2022). Figure 14 illustrates the geographical location of the Maldives and the complex island structure for the atolls of the country. The natural hazards dataset of Maldives for the year 2004, contains 190 records of tsunami affected islands, all being inhabited islands and provides the number of direct and indirect affected people for individual island. The dataset is published by open data sharing platform, Humanitarian Data Exchange (HDX) managed by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) under a creative commons attribution 4.0 international license (HDX, 2022). We note that the shape files for atolls and islands of Maldives are also accessed from HDX open data portal for subnational administrative boundaries of Maldives (HDX, 2021). It is noteworthy to mention that natural hazards like cyclone, typhoon, storm, flood and water shortage can also be accessed from the same open portal. We have used tsunami data as a showcase for the current study. The dataset for 190 islands provides detailed information about deaths, injuries, destroyed houses, and also about people who were directly and indirectly affected by the tsunami.



**Figure 15: Locations of tsunami affected islands for individual atolls of Maldives**

In order to simplify the modeling process, we have used the number of indirectly affected people as the response variable. According to 2012 independent evaluation report by the Asian Development Bank, 2004 tsunami had a devastating impact on the island nation of Maldives. The country experienced a disaster of national proportions, with 39 islands severely damaged (Asian Development Bank, 2012). Figure 15 (left) depicts the locations of tsunami affected islands for individual atolls of Maldives which includes enclosed lagoon or basin, fore reef, subtidal reef, pass reef flat and land on reefs. An interactive map of the Maldivian island structure and tsunami affected locations is published in an ArcGIS online map. In connection to this, the present study covers 12 different atolls across the entire country from north to south namely, Haa Alifu, Haa Dhaalu, Shaviyani, Noonu, Raa, Baa, Kaafu, Meemu, Laamu, Gaafu Alifu, Gaafu Dhaal and Seenu as highlighted in Figure 15 (right). The current selected atolls ranges the entire geographical span of the country including all 39 highly affected islands and covers the capital city of Male', as well as other important cities in the Maldives. Atolls having a substantially lower number of tsunami affected islands are not included in the present study. As discussed in Section 1 the atolls of Maldives are collection of disjoint islands, similarly, almost all the atolls are disjoint land surfaces. But some atolls such as Haa Alifu, Haa Dhaalu, Shaviyani and Noonu atolls in the north, in the north-west Raa and Baa atolls and in the south Gaafu Alifu and Gaafu Dhaal, these three regional groups share common boundaries. In the current study, these eight atolls, based on their common geographical boundaries are considered as three

distinct combined spatial regions. Details about the regional integration is depicted in [Figure 19](#), [20](#) and [21](#) at the end of [Section 5.4.4](#). In the rest of the study, we have referred these three atolls groups as, Shaviyani group, Baa group and Gaafu Dhaal group respectively.



**Figure 16: Tsunami affected regions of Baa and Raa atolls boundaries of atolls (left), boundaries of land on reefs for component islands (right)**

These integrations allow us to explore a substantial number of tsunami hit islands in continuous spatial regions. We report that the present study examines 135 islands from 7 different atolls or groups of atolls as shown in [Figure 15](#) (right). This is approximately 71 percent of the total 190 tsunami affected islands of Maldives. Detail records of the number of affected islands and indirectly affected people for the selected atolls is reported in Table 1.

**Table 1: Records of tsunami effects in the selected 12 Maldivian atolls**

Atoll	Num of affected islands	Num of indirectly affected people
Haa Alifu	14	15711
Haa Dhaalu	13	7677
Shaviyani	14	12305
Noonu	13	9045
Raa	15	13539
Baa	13	13457
Kaafu	9	6591
Meemu	8	7780
Laamu	12	7790
Gaafu Alifu	9	6832
Gaafu Dhaal	9	4470

Figure 16 (left) shows the locations of tsunami affected regions of Baa group atolls. Figure 16 (right) depicts detailed island distribution on reefs areas for the same. In both cases, tsunami affected regions are highlighted in red. We note that, all datasets used in the current study are collected from sources without restrictions and that have open access.

### 5.4.3 Methodology

Random spatial events generate irregularly scattered point patterns over areas of interest. In these cases, spatial point process models are useful tools to perform precise statistical analysis (Karaganis and Mimis, 2006; Loo et al., 2011; Juan et al., 2012). Moreover, we can find recent studies (Verdoy, 2019; Opitz et al., 2020) on spatial point processes that are able to identify spatial auto-correlations and interactions between points in the pattern. From Figure 16 (left) it seems that atolls with enclosed lagoon and land on reefs are continuous land structure. But each atoll is originally a collection of number of distributed islands as depicted in Figure 16 (right). By considering the total number of indirectly affected people from individual tsunami-hit islands, we open the door to consider Poisson regression models in combination with a Bayesian framework. Instead of using MCMC, we have used computationally faster solution for latent Gaussian models by using a Laplace approximation for the integrals with the INLA method (Rue et al., 2009). It focuses on models that can be expressed as latent Gaussian Markov random fields (GMRF) (Rue and Held, 2005).

Our approach combines a spatial Poisson regression method with an INLA Bayesian framework. In particular, let  $Y_i$  and  $E_i$  be the observed and expected number of indirectly affected victims of tsunami on the  $i$ -th island. We assume that conditional on the relative risk,  $\rho_i$ , the number of observed events follows a Poisson distribution:

$$Y_i \mid \rho_i \sim \text{Po}(\lambda_i = E_i \rho_i)$$

where the log-risk is modeled as

$$\log(\rho_i) = \beta_0 + Z_i \beta_i + \xi_i + \epsilon_i$$

Here,  $\xi_i$  accounts for the spatially structured random effects and  $\epsilon_i$  stands for an unstructured zero mean Gaussian random effect and log Gamma precision parameters 0.5 and 0.01, defined as penalized complexity (PC) priors (Simpson et al., 2017).  $Z_i$  represents the covariates. It is worth noting that, no covariates are included in our study. But it is possible to incorporate relevant covariates into similar models in future studies (Rue et al., 2009). We assigned a vague prior to the vector of coefficients  $\beta = (\beta_0, \dots, \beta_p)$  which is a zero mean Gaussian distribution with precision 0.001. All parameters associated to log-precisions are assigned inverse Gamma distributions with parameters equal to 1 and 0.00005. In the current study, we have chosen to provide default prior distributions for all parameters in R-INLA. These have been chosen partly based on priors commonly used in the literature (Martins et al., 2013; Blangiardo and Cameletti,

2015; Rue et al., 2016; Moraga, 2019). As we run several cases with different priors, we find that our results are robust against other alternative priors.

To bypass the problem of inefficiency in the estimation under a general INLA approximation, we have used another computationally tractable approach based on SPDE models (Lindgren et al., 2011). On one hand, we used SPDE to transform the initial Gaussian Field (GF) with Matérn covariance function to a GMRF. In particular, the spatial random process  $\xi$ , here represented by  $\xi(\cdot)$  explicitly denote dependence on the spatial field, follows a zero-mean Gaussian process with Matérn covariance function represented as

$$\text{Cov} \left( \xi(x_i), \xi(x_j) \right) = \frac{\sigma^2}{2^{\nu-1} \Gamma(\nu)} (\kappa x_i - x_j)^\nu K_\nu(\kappa x_i - x_j)$$

where  $K_\nu(\cdot)$  is the modified Bessel function of second order, and  $\nu > 0$  and  $\kappa > 0$  are the smoothness and scaling parameters, respectively. INLA approach constructs a Matérn SPDE model, with spatial range  $r$  and standard deviation parameter  $\sigma$ .

The parameterized model we follow is of the form:

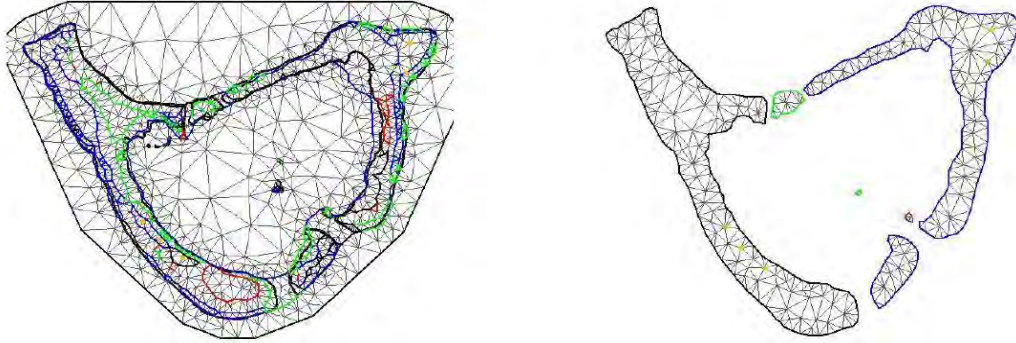
$$(\kappa^2 - \Delta)^{(\alpha/2)}(\tau S(x)) = W(x)$$

where  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$  is the Laplacian operator,  $\alpha = (\nu + d/2)$  is the smoothness parameter,  $\tau$  is inversely proportional to  $\sigma$ ,  $W(x)$  is a spatial white noise and  $\kappa > 0$  is the scale parameter, related to range  $r$ , defined as the distance at which the spatial correlation becomes small. For each  $\nu$ , empirically derived definition  $r = \sqrt{8\nu}/\kappa$  with  $r$  corresponding to the distance where the spatial correlation is close to 0.1 (Blangiardo and Cameletti, 2015). Note that we have  $d = 2$  for a two-dimensional process, and we fix  $\nu = 1$ , so that  $\alpha = 2$  in our case.

INLA-SPDE requires a triangulation or mesh structure to interpolate discrete event locations to estimate a continuous process in space (Krainski et al., 2018). In the current study, the spatial coordinates of each tsunami-affected island are employed as the target sites over which we constructed the mesh. We have designed SPDE mesh for the selected seven atoll groups. As a showcase, we have used Seenu atoll to discuss Delaunay's triangulation and barrier model techniques. Details mesh structure of stationary (for entire region and only for land on reefs) and non-stationary models for 7 atoll groups have been reported in Figures 22 to 28 at the end of Section 5.4.4. Because of the highly distributed nature of the island structure in each atoll, a continuous spatial structure is initially chosen for modeling, and triangulation is performed on the entire study area. According to Verdoy (2019), the best fitting mesh should have enough vertices for effective prediction, but the number should be within a limit to have control over the computational time. Based on this principle, a series of meshes with varying number of vertices are constructed. In the first case, we have ignored the internal complex island boundaries in each triangulation design and used the outline boundary of the entire atoll as the boundary parameter. Finally, from the battery of meshes, the best fitting mesh is selected.

Figure 17 (left) depicts the selected mesh with the locations of tsunami-hit islands highlighted in golden yellow. The number of vertices in the selected mesh is 1189. From Figure 17 (left) it is obvious that the SPDE mesh is generated for the entire geographic boundary of the atoll,

including its lagoon and sea surface. However, the records of the tsunami-affected areas are limited to discrete spatial sites that are located land on reefs for individual islands. Thus, the traditional stationary method of SPDE triangulation using the boundary of entire region is not realistic.

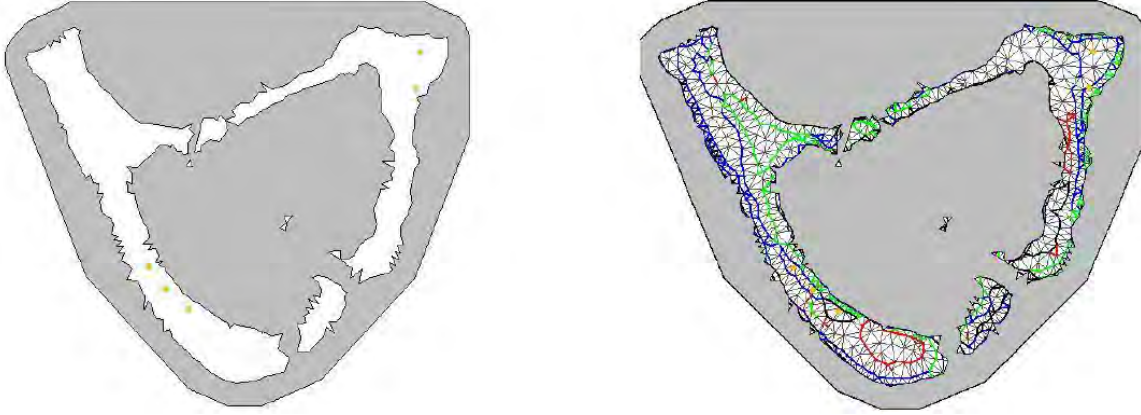


**Figure 17: SPDE triangulation with tsunami affected regions for the entire atoll boundary (left), triangulation generated only for land on reefs (right)**

As a result, we need to design triangulation precisely using boundaries of land on reefs for individual islands in the atoll. Recent study by [Chaudhuri et al. \(2022b\)](#) explores the application of explicit network triangulation where SPDE mesh is restricted to linear networks rather than the entire study area. Using similar methodology, we designed SPDE mesh only on land or reefs instead of defining the entire atoll boundary. As mentioned earlier, balancing between precision and computation time we have fine-tuned the number of vertices to identify the best fitted mesh. [Figure 17](#) (right) depicts the best fitting mesh having 427 vertices projected only on the land on reefs. According to [Wood et al. \(2008\)](#) and [Bakka et al. \(2019\)](#), this default method of designing triangulation only on the land on reefs has a serious drawback that the Neumann boundary condition is often unrealistic and severely impacts on the results.

#### 5.4.3.1 Barrier Model

The SPDE triangulations discussed in the previous subsection assume stationarity and isotropy that is, the autocorrelation between two locations depends solely on the Euclidean distance. However, while modeling events on dispersed island structure where there are physical barriers or, holes in the study area, stationarity is an unrealistic assumption ([Bakka et al., 2018](#)). Similar coastline problems are reported by [Ramsay \(2002\)](#), [Wood et al. \(2008\)](#) and [Scott-Hayward et al. \(2014\)](#). Moreover, stopping the triangulation at the coastline imposes the Neumann boundary conditions, also leading to unrealistic models ([Bakka et al., 2018](#)). This issue is common while exploring coastline and complex island problems. Other examples of physical barriers include road networks, power lines, categorical health sectors and areas with different land use. To deal with the coastline problem, several studies proposed solution by computing the shortest distance in water ([Wang and Ranalli, 2007](#); [Miller and Wood, 2014](#); [Scott-Hayward et al., 2014](#)). [Ramsay \(2002\)](#) proposed a methodology defining boundary conditions which uses a smoothing penalty together with Neumann boundary condition.



**Figure 18: Barrier object and barrier object with SPDE triangulation**

Furthermore, [Wood et al. \(2008\)](#) and [Sangalli et al. \(2013\)](#) demonstrate alternative solutions based on the Dirichlet boundary condition where a known value or, function is used along the boundary. In this line, [Bakka et al. \(2019\)](#) proposed an approach to handle nonstationary and anisotropic spatial processes with emphasis to handle complex archipelago structures in which the coastline is used as a physical barrier ([Bakka et al., 2019](#)). In their proposal, [Bakka et al. \(2019\)](#) approximated them using a finite element method based on the SPDE method. A system of two SPDEs is presented in this case, one for the barrier area, and the other for the remaining area.

The following system of stochastic differential equations has a solution that is specifically a nonstationary spatial effect, denoted by  $u(s)$ .

$$\begin{aligned}
 u(s) - \nabla \cdot \frac{r_b^2}{8} \nabla u(s) &= r_b \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_b, \\
 u(s) - \nabla \cdot \frac{r^2}{8} \nabla u(s) &= r \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_n,
 \end{aligned}$$

where  $u(s)$  is the spatial effect,  $\Omega_b$  the barrier area and  $\Omega_n$  is the remaining area and their disjoint union gives the whole study area  $\Omega$ . Ranges for the barrier and remaining areas are represented by  $r$  and  $r_b$  respectively.  $\sigma_u$  is the marginal standard deviation.  $\nabla$  is equal to  $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$  and  $W(s)$  stands for white noise.

It is worth noting that the barrier model is based on viewing the Matérn correlation as a collection of paths through a simultaneous autoregressive (SAR) model, rather than as a correlation function on the shortest distance between two points. The local dependencies are manipulated to cut off paths crossing the physical barriers. In the next step, the new SAR model is formulated to SPDE format to represent the Gaussian field, with a sparse precision matrix that is automatically positive definite ([Bakka et al., 2019](#)).

In the study by [Bakka et al. \(2019\)](#), water body is considered as normal terrain and distinct coastlines and boundaries are used as physical barriers. In contrast, in the current study, we have

defined boundary polygons of individual land on reefs for each island as our study area and the water body acts as the physical barrier. Besides, tsunami-hit locations are the sample events considered precisely on the land area of the islands. [Figure 18](#) (left) depicts the barrier object where the region in grey indicates the physical barrier and white area indicates the land on reefs where the spatial dependency will be analysed. Points in golden yellow color indicate the locations of tsunami-hit areas used as event locations in the model. Triangulation is designed using a barrier model where the land on reefs is considered as normal terrain and the water bodies (ocean and lagoons) act as physical barriers. [Figure 18](#) (right) illustrates the triangulation along with the physical barrier in grey. In this section, we discuss Seenu atoll. The mesh structure of other atoll groups has been reported in [Figures 22 to 28](#) at the end of [Section 5.4.4](#).

### 5.4.3.2 Model Fitting

Based on the discussions in the previous subsections we designed a set of hierarchical Bayesian models with Poisson likelihood and priors that penalise complexity. The steps for modeling the application includes the spatial effect created with the mesh using the spatial locations by Matérn covariance, and then implementing individual mesh structure. We have also considered independent and identically distributed Gaussian random effect, represented as  $i$  iid in the modeling process.

**Table 2: Competing models with choice of SPDE mesh and iid**

Model	iid	SPDE mesh (entire region)	SPDE mesh (only land)	Barrier model
M1		×		
M2			×	
M3			×	
M4	×			
M5	×	×	×	
M6	×			
M7	×			

In total, we have fitted 7 different models for each atoll group. Meshes used in both stationary and nonstationary models for individual atoll groups are reported in [Figures 22 to 28](#) at the end of [Section 5.4.4](#). The seven models are combinations of different types of mesh used along with iid random effects in both stationary and nonstationary scenarios and one model without any spatial effect. It is worth mentioning that no covariates are used in designing the models for the current study. Details of each model are shown in [Table 2](#).



As we have a battery of competing models, we compare them using the deviance information criterion (DIC) (Spiegelhalter et al., 2002), which is a Bayesian model comparison criterion, represented as

$$\text{DIC} = \text{goodness of fit} + \text{complexity} = D(\bar{\theta}) + 2p_D$$

where  $D(\bar{\theta})$  is the deviance evaluated at the posterior mean of the parameters, and  $p_D$  denotes the effective number of parameters, which measures the complexity of the model (Spiegelhalter et al., 2002). An alternative is the Watanabe Akaike information criterion (WAIC) which follows a more strict Bayesian approach to construct a criterion (Watanabe, 2010).  $pWAIC$  is similar to  $p_D$  in the original DIC (Gelman et al., 2014). The lowest values of DIC and WAIC suggest the best fitted model.

### 5.4.4 Results and discussions

As mentioned in the above sections we have applied different methodological approaches to 7 atolls selected from north to south. They have their particularities and differ from one to another because of their shape, extension of land, and number and size of component islands. Likewise, the number of nodes differs from the whole region and the only land in the mesh creation. Specifically, the nodes in the whole region ranges from 1189 in Seenu atoll to 3001 in Gaffu Dhaal. On the contrary, while considering only land, the numbers are much smaller, from 427 Seenu to 1283 Gaafu Dhaal. According to their characteristics, the mesh applied will be better or worst.

**Table 3: DIC and WAIC values according to the seven models and for the considered atolls**

Atoll	Model	DIC	WAIC
Shaviyani	M1	552.98	556.60
	M2	3917.20	4810.94
	M3	553.91	557.87
	M4	539.73	532.89
	M5	535.86	526.98
	M6	<b>534.90</b>	<b>524.98</b>
	M7	535.41	526.42
Baa	M1	329.8	427.19
	M2	4386.85	6046.55
	M3	329.88	427.51
	M4	263.83	262.44
	M5	259.25	256.30

	M6	<b>259.51</b>	<b>256.74</b>
	M7	259.60	257.21
	M1	229.88	1376.15
	M2	229.49	1375.69
Kaafu	M3	229.50	1375.57
	M5	133.02	133.78
	M6	131.75	132.66
	M7	131.24	132.15
	M1	73.46	71.42
	M2	73.46	71.42
Meemu	M4	73.44	71.37
	M5	73.46	71.40
	M6	<b>73.44</b>	<b>71.32</b>
	M7	73.46	71.41

We have applied three different meshes to verify that. In the first one, entire region meaning that each territory (without distinguishing between land and water) is considered a possibility for modeling. The other only considers land region, taking into account only land on reefs for individual component islands. The third mesh corresponds to the barrier model. It distinguishes between different component islands with water or lagoon in between, which means it can analyze the atoll as a whole but distinguishes between land and water because of the boundaries of the land on reefs for individual component islands. These characteristics are very important to understand the differences in the goodness of the models. The dependent variable is always the indirectly affected people by the tsunami.

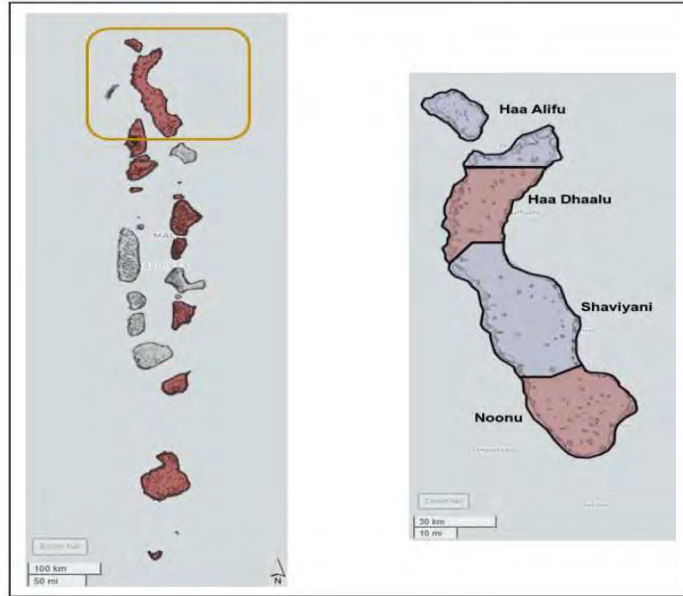
**Table 3: DIC and WAIC values according to the seven models and for the considered atolls (contd.)**

Atoll	Model	DIC	WAIC
	M1	363.83	1809.40
	M2	363.39	1810.52
	M3	363.25	1810.53
Laamu	M4	141.59	137.53
	M5	141.65	137.62
	M6	<b>141.65</b>	<b>137.61</b>
	M7	141.66	137.62

	M1	225.70	1708.00
	M2	225.55	1707.99
	M3	225.42	1708.09
Gaa Dhalu	M4	203.35	197.51
	M5	203.43	197.65
	M6	<b>203.43</b>	<b>197.48</b>
	M7	203.43	197.61
	M1	151.59	265.83
Seenu	M2	151.55	265.79
	M4	78.64	76.53
	M5	78.60	<b>76.35</b>
M6	78.64	76.48	
	M7	78.61	76.48

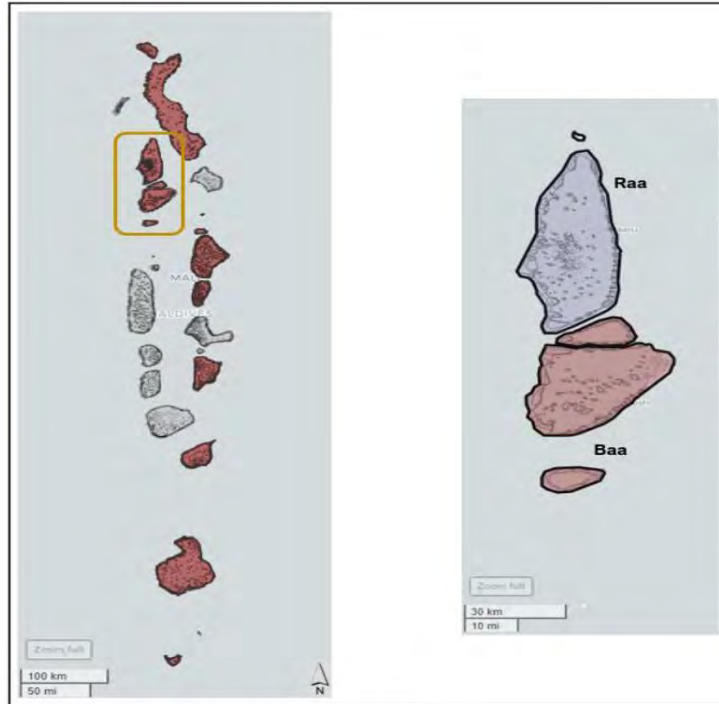
We provide and discuss the results of model fittings according to the aspect of the atolls. In particular, we are considering three different models but combining the introduction of the iid to the model. Thus, we have run 5 stationary models and 2 nonstationary models. More specifically, the models considered are reported in Table 2. On the other hand, Table 3 shows the DIC and WAIC values, which are the most commonly used diagnostics for knowing the model quality in a Bayesian setting (Rue et al., 2009). The smaller they are, the better the model will be. In this case, the nonstationary model including iid (model 6) is the best under the DIC criteria for almost all the atolls. That is, the proposed model with the iid, fits better in most of the cases. However, we observe some lack of resemblance when looking at specific atolls.

On the one hand, the Shaviyani and the Baa atolls have the same results in terms of the DIC and their structure is quite similar. These two atolls are characterized because the events occur both, on the coastline and also in the pieces of land inside the boulder reef. This distribution could explain why the model which applies SPDE, considering only the land region, doesn't work properly in these two cases as the mesh used doesn't take into account the space between the component islands. On the other hand, when applying the models on the Gaafu Dhaalu, Laamu and Kaafu atolls, when the iid is included, no differences are observed between the stationary models and the nonstationary ones. Nevertheless, it is worth noting that, the barrier model, the one that we propose in this study, has the same DIC value as the stationary models but not worse values.

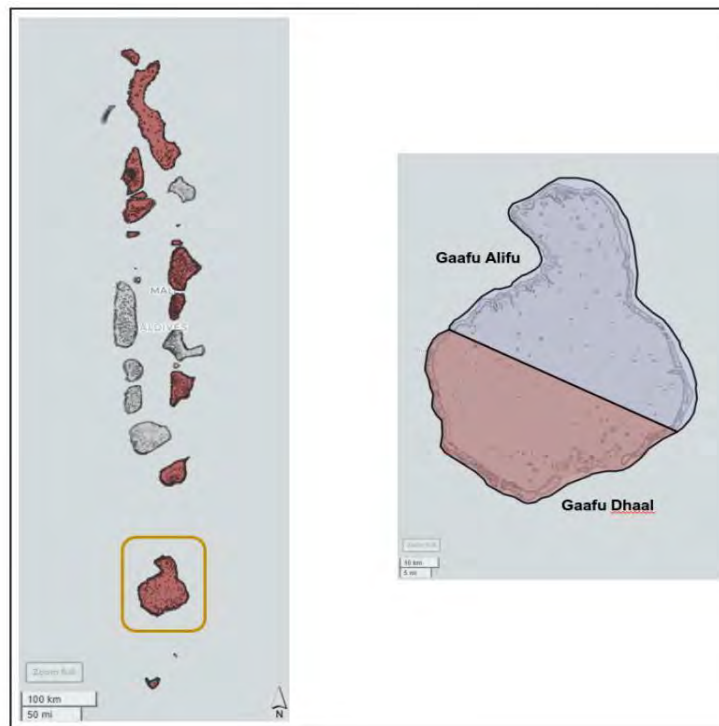


**Figure 19: Combined spatial region for Haa Alifu, Haa Dhaalu, Shaviyani and Noonu atolls**

Therefore, the proposed model (barrier model) works well but, perhaps due to the structure of this atoll, the stationary models also correctly fit the locations of the tsunamis that occurred in these atolls, which are mainly located on the boulder reef except for a maximum of one event that is located in the middle part of the atoll. This structure could explain the similarities between the three used models because the meshes, in this case, do not fault when modeling the events in these atolls as they are mainly concentrated on the boundaries and so it has less importance to consider the atoll as a whole distinguishing land and water (barrier model). Then, when looking at the atoll further south (Seenu) there is a clear difference when comparing the stationary models with the nonstationary model, mainly when the iid is not included.



**Figure 20: Combined spatial region for Raa and Baa atolls**



**Figure 21: Combined spatial region for Gaafu Alifu and Gaafu Dhaal atolls**



**Figure 22: Mesh for Shaviyani group atolls**



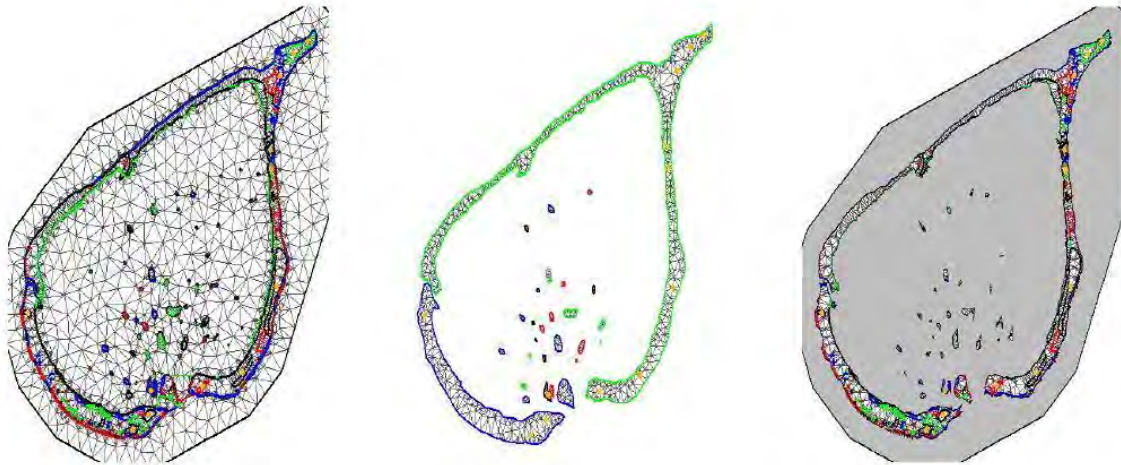
**Figure 23: Mesh for Baa group atolls**



**Figure 24: Mesh for Kaafu atoll**



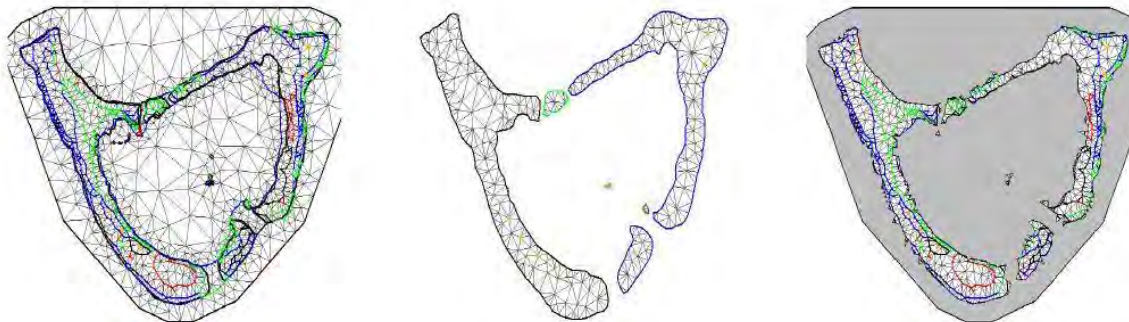
**Figure 25: Mesh for Meemu atoll**



**Figure 26: Mesh for Laamu atoll**



**Figure 27: Mesh for GDhaalu group atolls**



**Figure 28: Mesh for Seenu atoll**



These results are surprising because the structure of the territory of this atoll and the locations of the events are similar to that observed in the previously mentioned atolls. The only difference is that the Seenu atoll presents the smallest number of tsunamis compared to the rest of the studied atolls and so the number of events seems to be important in terms of the model fit. Finally, the models applied on the Meemu atoll do not present differences between them. This atoll includes only a piece of land, has all the tsunami events around the boulder reef and the number of tsunamis is not the smallest. Therefore, in this case, again, it is not relevant to apply a mesh that distinguishes between boundaries and water, since there are only events in one land region. Therefore, the fit of the model does not vary whether the applied mesh considers the atoll as a whole, if it considers the component islands as individual elements, or if it is used the one corresponding to the barrier model. In any case, the barrier model still offers very good results comparing the stationary models with the nonstationary models and so, the proposed model seems to be a very good option when we have to deal with sparse data and complex barriers.

The principal advantage of the barrier model is that the computational cost is the same as for the stationary model. In general, the model is easy to use, and can deal with both sparse data and very complex physical barriers. This work has some limitations. The most important one is related to boundary effects in the SPDE-INLA approximation. This approach creates artifact spatial dependencies on the boundary. In a standard mesh, as long as it is well constructed, the boundaries are in the outer limits of the spatial domain of interest and, therefore, those dependencies can be identified and eliminated. However, in a more complex mesh, such as barrier models, the boundaries lie within the spatial domain of interest. This fact makes it difficult, sometimes excessively, to identify and subsequently eliminate artifact spatial dependencies. Despite these difficulties, in this work we have managed to identify them. However, a different approximation is needed in which the SPDE-INLA approximation does not cause these fictitious spatial dependencies. We are working on that approach today. Second, these fictitious dependencies cause a very low predictive capacity of the model. This limitation cannot be resolved until that other approximation we are working on is achieved. However, this work also has its strengths. First of all, the methods we propose have the same computational cost as the stationary models, which simplifies their analysis. Second, the methodology we have proposed allows us to introduce the spatial effect in complex land structures.

## 5.5 Enhanced spatial modeling on linear networks using Gaussian Whittle-Matérn family

### 5.5.1 Introduction

Over the last few decades, advancements in computing and real-time data collection have enabled the collection of vast amounts of spatio-temporal data. As a result, statistical modeling of spatiotemporal data has gained more popularity and is now being utilized in various disciplines (Wood et al., 2004; Fuglstad and Castruccio, 2020). Applications range from the analysis of meteorological data, environmental data (Blangiardo et al., 2013), ecology (Zuur et al., 2017), and natural disasters such as forest fires (Serra et al., 2014), landslides (Lombardo et al., 2020), and earthquakes (Liu and Stein, 2016; Field et al., 2017). Additionally, spatiotemporal modeling is used in urban planning and strategic decision-making for issues such as traffic accidents (Prassanakumar et al., 2011; Liu and Sharma, 2018), criminal activities (Leong and Sung, 2015; Hossain et al., 2020), air pollution (Mota-Bertran et al., 2021, Saez and Barceló, 2022) and epidemiology and infectious disease dynamics (Schrödle et al., 2010; Moraga, 2019).

Depending on the objective of the study, various types of models are used with spatial and spatio-temporal data. With the development of Markov chain Monte Carlo (MCMC) simulation methods, researchers began to deal with these types of data using Bayesian methods (Gilks and Robert, 1996; Robert et al., 1999). To fit generalized linear mixed models (GLMM) in a spatial context, a Bayesian approach with MCMC simulation methods has traditionally been used. However, with the increase in data size and resolution, the computational burden of MCMC has become a critical issue (Rue et al., 2009; Rue et al., 2009; Taylor and Diggle 2014).

To address this issue, Rue et al. (2009), proposed significantly faster solution as integrated nested Laplace approximations (INLA) which focuses mainly on models that can be expressed as latent Gaussian Markov random fields (GMRF). Advancements in spatial statistics have made it possible to fit continuous spatial processes with a Matérn covariance function using INLA. Lindgren, Rue, and Lindström (2011) introduced a solution for the stochastic partial differential equation (SPDE) that provides a sparse representation of the solution fitting within the INLA framework. The solution for the spatial process can be represented as a sum of basis functions and associated coefficients, where the basis functions approximate the solution, and the coefficients follow a Gaussian distribution. This spatial model is implemented in INLA as the stochastic partial differential equation (SPDE) latent effect (Krainski et al., 2018). However, fitting this model with INLA requires the definition of a mesh over the study region to compute the approximation to the solution. Literature shows several research works where INLA along with SPDE construct spatio-temporal models through Kronecker products of a spatial Matérn model, and first- or second-order autoregressive models in time (Lindgren et al., 2015; Blangiardo and Cameletti, 2015; Bakka et al., 2018; Moraga, 2020; Lindgren et al., 2022).

### 5.5.1.1 Modeling on Complex Distributed Spatial Regions

Spatial models often assume isotropy and stationarity, implying that spatial dependence is direction invariant and uniform throughout the study area. However, these assumptions are violated when physical barriers are present in the form of geographical features as in case of complex island structures or in case of man-made barriers like disease control interventions and in animal species distribution problems. In these cases, the dependency among the observations should not be based on the shortest Euclidean distance between the locations but should take into account the effect of physical barriers and smooth them (Bakka 2019).

Traditional SPDE method triangulates the entire study area based on continuous geographic boundaries (Krainski et al., 2018). Problem arises in typical environmental research works such as modeling species distribution, where physical barriers such as mountains, roads or rivers could pose obstacles for the movement of species. Since propagation through those obstacles is not possible, spatial correlation should not follow the shortest path, but should travel around them. However, studies show that the meshes are usually generated for the entire study region, including the physical barriers. Similar approach is observed in case of complex archipelagos or coastlines. This approach involves generating an SPDE mesh for the entire study region, despite the presence of physical barriers that make the study area complex and distributed. For example, Lezama-Ochoa et al. (2020) used this approach to predict the occurrence of spine tail devil ray species in the eastern Pacific Ocean. Bi et al. (2021) conducted a similar study to estimate seabird bycatch variations in the mid-Atlantic bight and northeast coast, and Cosandey-Godin et al. (2014) applied this approach to analyze spatiotemporal patterns of accidental bycatch in fisheries located in the Baffin Bay of the Atlantic Ocean. In our review of the literature, we have found several studies that have used the same approach to model complex land structures. For example, R. De Jesus Crespo (2019) studied flood protection ecosystem services in the coast of Puerto Rico, Myer et al. (2017) used a spatiotemporal model to examine the ecological and sociological factors that predict the presence of West Nile virus in mosquitoes in Suffolk County, New York, Paradinas et al. (2015) employed a spatio-temporal approach to validate persistence areas and identify fish nurseries in the western Mediterranean Sea, and Lourenço et al. (2017) estimated the potential distribution of invasive and native trees in the Azores islands, Portugal. We aim to build upon these studies and further explore the application of INLA-SPDE in complex land structures, particularly in coastal regions and islands.

Another serious concern to model observations in complex island structures is the anomaly related to the polygon structure of the coastlines. Coastlines are often considered as fractal structure, in the sense that any finite approximation will not be accurate (Bakka et al., 2019). For the same coastline polygons, different researchers may use varying approximations which can lead to conflicting interpretations and predictions. In that case, the model loses its scientific credibility. It is worthy to mention that a stationary model cannot be aware of the coastline structure and will inappropriately smooth over the features. In spatial modeling, classical models become unrealistic when they fail to account for holes or physical barriers in the landscape. This can lead to further unrealistic assumptions.

Bakka et al. (2019) introduced the barrier model as a solution to the limitations of existing models. Unlike traditional models, this new model does not rely on the shortest distance around a physical barrier or specific boundary conditions. Instead, it provides a non-stationary Gaussian random field which can handle sparse data and complex barrier structures. The authors applied the model to study the distribution of fish larvae species in the *Finnish Archipelago Sea*, which is a particularly relevant example as the larvae live near the coast and the study area includes many barriers that should not be smoothed over. Additionally, the computational cost is comparable to that of a stationary model. In a recent study, Martínez-Minaya et al. (2019) have used the barrier model approach to design a Bayesian hierarchical species distribution model (SDM) to determine vulnerable habitats for bottlenose dolphins in the Northern Sardinia archipelago in Italy. Likewise, the use of barriers is also crucial in the control of infectious diseases. Cendoya et al. (2022) studied the impact of barriers on the spatial distribution of a quarantine plant pathogenic bacterium in Alicante, Spain. They compared the results of a traditional stationary model, a model with physical barriers, and models with both continuous and discontinuous perimeter barriers around the infected areas.

The simulation study in the Archipelago (Bakka et al., 2019) and other applications demonstrate that while barrier models have a similar computing cost to their corresponding stationary models, they are more flexible and realistic when used in complex spatial regions with physical barriers. However, some anomalies are found when the barriers are infinitely thin, in those cases artificially thicker barriers, such that the width is at least a mesh triangle, can make the model functional. Li et al. (2023) extended the barrier model by introducing a multi-barrier model that can characterize areas with different types of obstacles or physical barriers. Authors compared stationary Gaussian model, barrier model, and proposed multi-barrier model using real burglary data, and the results suggest that all three models have similar performance.

### 5.5.1.2 Modeling on Linear Networks

On the other hand, in many environmental applications such as urban road networks or stream systems it is essential to define statistical models on linear networks. A major focus of research in this field has been spatiotemporal modeling of traffic accidents on urban road networks (Karaganis and Mimis, 2006; Castro et al., 2012; Boulieri et al., 2016; Liu et al., 2019). Studies like, Xu and Huang (2015), Wang et al. (2019) and Eboli et al. (2020), effectively capture the spatial dependence and heterogeneity in traffic accident data, improving the accuracy and robustness of predictions compared to traditional regression models on road networks. Recently, a number of models on road safety have been proposed following Bayesian methodology. Cantilo et al. (2016) used a combined GIS-Empirical Bayesian approach in modeling traffic accidents in the urban roads of Columbia. A similar research work on urban road network of Florida by (Zeng and Huang, 2014) explored Bayesian spatial joint modeling of traffic crashes. A space-time multivariate Bayesian model was designed by (Boulieri et al., 2016) to analyze road traffic accidents by severity in different cities of UK. Recently, (Galgamuwa et al., 2019) used Bayesian spatial modeling using INLA in predicting road traffic accidents based on

unmeasured information at road segment levels. Due to densely distributed nature of the road segments, majority of these studies used continuous spatial structures and traditional spatial stationary models such as Matérn fields. As a result, though the sampling points (here the traffic accident locations) are mainly located on the road networks, the SPDE triangulations are designed on the entire study area, including the areas without road network. Thus, the model result might be unpreventably generalized as it is going to estimate predicted values for the regions where there is no chance of incident to occur. In this context, [Chaudhuri et al. \(2022b\)](#) recently proposed spatiotemporal modeling of road traffic accidents using explicit network triangulation on the road network of London, UK. In a similar study by [Chaudhuri et al. \(2023\)](#), SPDE triangulation has been designed precisely on linear road networks of Barcelona, Spain to generate dynamic traffic accident risk maps. The methodology used in these two studies is a novel approach to perform spatio-temporal analysis precisely on road network and contributes to the relatively small amount of literature in this domain. However, in both cases, the complex boundary regions of the buffer road network result in high boundary effect, that can influence the spatial effects of the models. This is a serious limitation of the SPDE network triangulation approach.

In general, Gaussian random fields with Matérn covariance functions are a popular choice but they have a stationary and isotropic covariance structure. [Dawkins et al. \(2019\)](#) made a novel attempt to apply a barrier model on linear road networks. This research showed a non-stationary approach to accurately estimate air quality levels on the roads of Brisbane, Australia using Bayesian methods. The study accounted for the topographical diversity of buildings in proximity to city roads by employing a non-stationary barrier model that extends upon the INLA framework.

On the other hand, [Bolin et al. \(2022\)](#) presented an alternative to using the Euclidean distance by defining similar models with a non-Euclidean metric on a graph. However, it can be challenging to find a class of positive definite functions suitable for creating Gaussian fields on metric graphs when using a non-Euclidean metric. It is also difficult to apply the SPDE approach to metric graphs as it is uncertain how to define the differential operator and what kind of covariance functions would result. The study proposes a novel approach of a new and valid differentiable Gaussian field on general compact metric graphs.

### **5.5.1.3 Motivating Example**

Accessible, and sustainable transport systems in cities are a core target of 2030 sustainable development goals (SDGs) adopted by the United Nations ([UNDP, 2021](#)). Thus, there is an opportunity to apply advanced computational techniques to model the spatial variation in the incidence of road traffic accidents in a linear road network system to aid in accident prevention and multi-disciplinary road safety measures. The motivating example we have used in this paper is ten-years (2010-2019) of daily traffic accident records on the road networks from the central part of Barcelona, Spain. The network is complex enough to motivate a general solution using the proposed non-Euclidean metric on graph model and also compare the results with SPDE

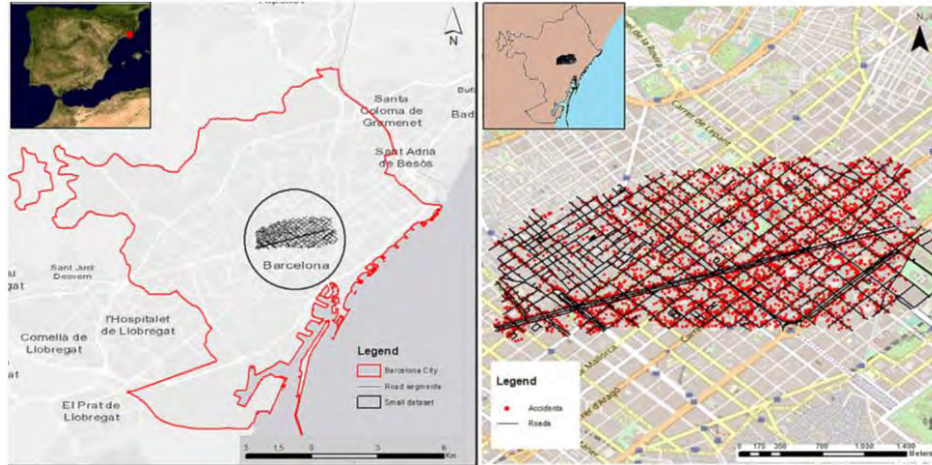
network model and barrier model. Further, the study region contains the road segments which observe the maximum daily records of traffic accidents as well as some road segments where there are no records of accidents during the entire study period.

The aim of this study is two-fold. The principal aim is to introduce a new class of Gaussian processes on compact metric graphs with Whittle-Matérn fields defined by a fractional SPDE on metric graph in R-INLA framework (Bollin et al., 2022). Secondly, to compare and contrast the performance of the proposed model with two other distinct approaches namely, the SPDE network triangulation models and barrier models on linear networks. R (version R 4.2.2) programming language (R Core Team, 2022) has been used for statistical computing and graphical analysis. All computations were conducted on a quad-core Intel i9-4790 (3.60 GHz) processor with 32 GB (DDR3-1333/1600) RAM.

### 5.5.2 Data Settings

Barcelona is the largest and capital city of Catalonia, Spain and is located on the northeastern coast of the country. With a population of 1.6 million and a density of 15,748 inhabitants per square km, it is the second most populous municipality in Spain (OpenDataBCN, 2021). The city is a major cultural, economic, and financial center, as well as a transportation hub for southwestern Europe with a well-developed motorway network. In this study, a small area of 4.4 square km in the central part of the city, consisting of 2058 road segments, has been considered as depicted in the left panel of Figure 29 inside the black circle. The road network data has been obtained from the Open Data BCN repository (OpenDataBCN, 2021). The police department in Barcelona keeps records of traffic accidents and related casualties and injuries, which are annually published by Open Data BCN under the *Creative Commons Attribution 4.0* for public sector information. The data is free and available for public sector information.

During the period from January 2010 to December 2019, there are 11,067 recorded traffic accidents in the study area. The locations of these accidents are shown in red on the road network map in the right panel of Figure 29. The study utilized five datasets from Open Data BCN, which are linked by a record code from 2010 to 2019. The common characteristics recorded in the data consists of a unique event ID, district and neighborhood, postal address and geographical coordinates, and the day and time of occurrence. We included three covariates in our models: road length (ranging from 3.69 to 186.25 meters) with a mean of 81.61 meters, road type (values 1 to 7, with higher values indicating lower traffic), and road speed limit (ranging from 18 to 80 km per hour). Notably, roads with speed limits of 30, 35, and 50 km per hour accounted for 21%, 28%, and 35% of the total sample, respectively. The datasets also include temporal variables such as year, month, and time of the accident. The individual accident locations are adjusted to the nearest road segments.



**Figure 29: Geographical location and road network with traffic accidents in Barcelona**

The number of minor injuries has been used as the response variable in the models. Most of the accidents (74.76%) have only one minor injury, followed by two minor injuries (15.42%) and 3 or more minor injuries (3.42%). There were 6.4% of accidents with no minor injuries, and 99.85% of the accidents resulted in no casualties. The number of accidents recorded in each year of the study are similar, with the highest number (1270) in 2016 and the lowest number (847) in 2011. It is worthy to mention that, in case of network mesh and barrier model the daily minor injuries for individual road segment have been aggregated and included in the centroid of that segment. This means that other temporal covariates related to each accident are not considered in the current study. In contrast, the proposed graph model converts road segments into the edges of a graph and considers accident locations, road network intersections, and the start and end nodes of each road segment as the vertices of the graph. In the first model, the distances to nearby facilities such as bus stops, municipal markets, restaurants, schools, and street markets are calculated from the centroid of each road segment and used as spatial covariates. But in the graph model, these distances are calculated from individual vertices of the graph. A detailed description of generating the vertices and edges of the graph is reported in Section 5.5.3.3.

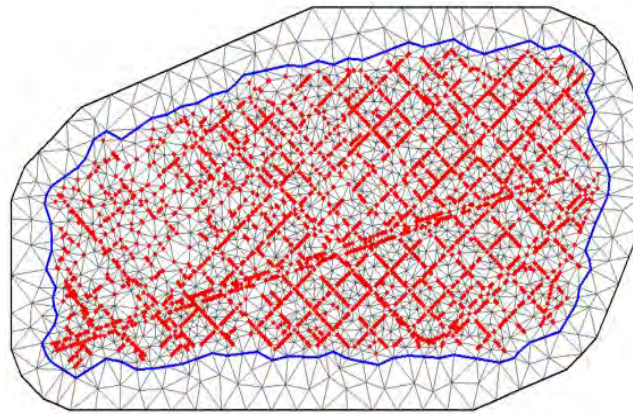
### 5.5.3 Methodology

Our discussion in this section initially covers two existing models, namely network mesh and barrier models on linear network, and in the third subsection we define the proposed exponential graph model and its application to the selected dataset.

#### 5.5.3.1 Network triangulation

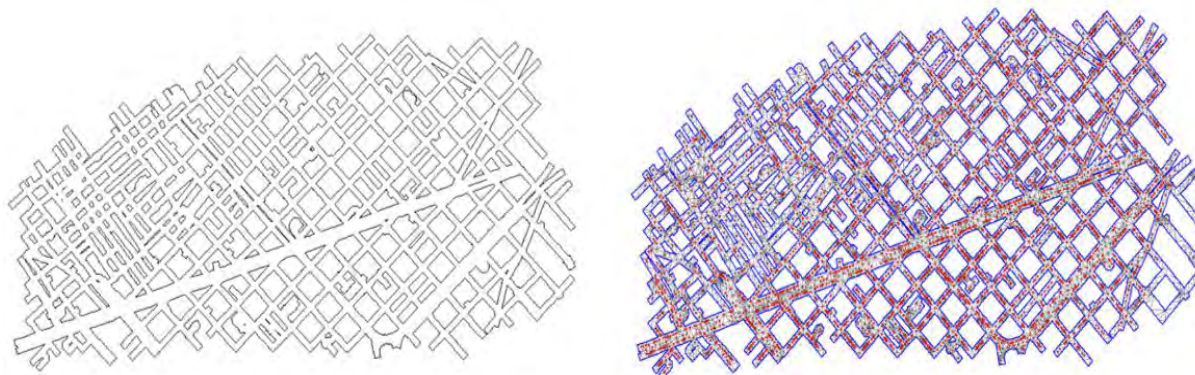
Analysis of spatiotemporal events such as traffic accidents, street crimes, and issues in water and electric connection networks in cities that occur exclusively on linear networks, it has been observed that conventional INLA-SPDE techniques are frequently used to model these events, despite the fact that they are strictly confined to linear networks. When applying the INLA-SPDE method to linear networks, creating a triangulation for the entire region enables fitting of

the INLA model in the study area. However, a significant problem arises while predicting events, as the observed events are discrete spatial points located precisely on the road network, whereas models fitted with a region mesh cover the entire study area. This implies that the locations of predicted events can be placed in any area with or without road networks, which is not realistic. Traditional methods of model prediction using a region mesh are, therefore, not appropriate in this context from a scientific perspective.



**Figure 30: Region mesh for Barcelona road network**

In the current study, due to close proximity of the road segments, initially a continuous spatial structure is selected for modeling, and triangulation is carried out on the entire study area. In this context, [Verdoy \(2019\)](#) argues that the best mesh for prediction should have a sufficient number of vertices for accuracy but also within a limit to reduce computational time. Following this principle, from a battery of meshes, the best fitted mesh is selected having 2352 vertices. [Figure 30](#) depicts the region SPDE mesh with 11,067 traffic accident locations highlighted as red points.



**Figure 31: Buffered road polygon and network mesh with event locations highlighted as red points**

However, the fitted mesh as shown in [Figure 30](#) has a problem when it covers the entire study area. It is unrealistic and ambiguous for the model predictions to cover areas without a road network where traffic accidents are unlikely to occur. This drives the need to design the SPDE



triangulation precisely on road networks. The process consists of three phases: generating a buffer region for each road segment, creating a clipped buffer polygon that covers only the road network, and designing SPDE triangulation on the clipped polygon to form an SPDE network mesh. Choosing the buffer size requires finding a balance between the number of vertices in the triangulated mesh and computational cost (Krainski et al., 2018; Verdoy, 2019). After evaluating various buffer sizes, a 15-meter buffer has been identified to be the best option. The left panel of Figure 31 illustrates the 2058 road segments with a 15-meter buffer around each segment. Following that, we merge individual buffer segments into a single polygon clipped within a bounding box covering the study area. In the final step, we use the centroids of each road segment as the target locations over which we build the initial Delaunay's triangulation.

It is worthy to note that, for each road segment, the total number of minor injuries has been aggregated daily and added at corresponding centroids as the response variable. The triangulation is created using the centroids. Figure 31 (right) depicts the SPDE mesh precisely designed on the road network, with accident locations highlighted in red. We report the number vertices in the network mesh is 14,368. By aggregating data from locations and converting it into event counts per segment, we can utilize Poisson regression models together with a Bayesian approach to model traffic accidents on individual road segments. In fact, we use a spatial Poisson regression method within a Bayesian framework using INLA and SPDE. Recent research conducted in the same study area and utilizing the same dataset, by Chaudhuri et al. in 2023, found that a network mesh model outperformed the SPDE mesh model for the entire study area. Therefore, in this section, we have focused solely on the more efficient network mesh model and compared it with the two other models discussed in following sections.

In particular, let  $Y_i$  and  $E_i$  be the observed and expected number of road traffic accidents on the  $i$ -th road segment. We assume that conditional on the relative risk,  $\rho_i$ , the number of observed events follows a Poisson distribution:

$$Y_i | \rho_i \sim PO(\lambda_i = E_i \rho_i)$$

where the log-risk is modeled as

$$\log(\rho_i) = \beta_0 + Z_i \beta_1 + S(x_i) + \epsilon_i \dots \dots \dots (4)$$

Here,  $S(x_i)$  account for the spatially structured random effects, and  $\epsilon_i$  stands for an unstructured zero mean Gaussian random effect and log Gamma precision parameters 0.5 and 0.01, defined as penalized complexity (PC) priors (Simpson et al., 2017).  $Z_i$  represents the spatial covariates. We assigned a vague prior to the vector of coefficients  $\beta = (\beta_0, \dots, \beta_p)$  which is a zero mean Gaussian distribution with precision 0.001. All parameters associated to log-precisions are assigned inverse Gamma distributions with parameters equal to 1 and 0.00005. The default prior distributions for all parameters in R-INLA were selected based on commonly used priors in previous studies (Martins et al., 2013; Blangiardo and Cameletti, 2015; Rue et al., 2016; Moraga, 2019). We report that our results are robust against other alternative priors, as we run several cases with different priors obtaining the same results.

To compute the joint posterior distribution of the model parameters, we use an INLA-SPDE method, as introduced by [Lindgren et al. \(2011\)](#). SPDE consists in representing a continuous spatial process, such a Gaussian field (GF), using a discretely indexed spatial random process such as a Gaussian Markov random field (GMRF). In particular, the spatial random process (represented by  $S(\cdot)$ ) explicitly denote dependence on the spatial field, follows a zero-mean Gaussian process with Matérn covariance function represented as:

$$\text{Cov}(S(x_i), S(x_j)) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa \|x_i - x_j\|)^\nu K_\nu(\kappa \|x_i - x_j\|)$$

where  $K_\nu(\cdot)$  is the modified Bessel function of second order, and  $\nu > 0$  and  $\kappa > 0$  are the smoothness and scaling parameters, respectively. INLA approach constructs a Matérn SPDE model, with spatial range  $r$  and standard deviation parameter  $\sigma$ .

The parameterized model we follow is of the form:

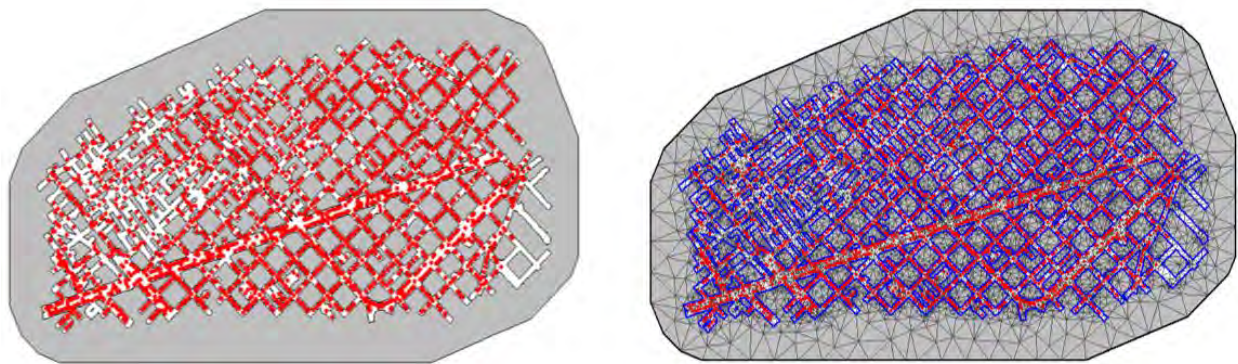
$$(\kappa^2 - \Delta)^{(\alpha/2)}(\tau S(x)) = W(x)$$

where  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$  is the Laplacian operator,  $\alpha = (\nu + d/2)$  is the smoothness parameter,  $\tau$  is inversely proportional to  $\sigma$ ,  $W(x)$  is a spatial white noise and  $\kappa > 0$  is the scale parameter, related to range  $r$ , defined as the distance at which the spatial correlation becomes negligible. For each  $\nu$ , we have  $r = \sqrt{8\nu}/\kappa$ , with  $r$  corresponding to the distance where the spatial correlation is close to 0.1. Note that we have  $d = 2$  for a two-dimensional process, and we fix  $\nu = 1$ , so that  $\alpha = 2$  in this case ([Blangiardo and Cameletti, 2015](#)). Next, to interpolate discrete event locations to estimate a continuous process in space we have used the SPDE network mesh as depicted in the right panel of [Figure 31](#). The projection matrix is generated using the centroids of individual road segments and triangulations in the mesh. [Bakka et al. \(2018\)](#) suggest that the range value should be determined based on the spatial distribution of events in the study area. In the current study, due to the proximity of accident locations we have decided to use a prior  $P(r < 0.01) = 0.01$ , meaning that it is highly unlikely the range is less than 10 meters. The parameter  $\sigma$  represents the variability of the data and has a prior specified as  $P(\sigma > 1) = 0.01$ .

### 5.5.3.2 Barrier model on linear network

The SPDE triangulations discussed in the previous subsection assume stationarity and isotropy that is, the autocorrelation between two locations depends solely on Euclidean distance. From a scientific perspective, it is problematic to use the previous approach due to the inclusion of additional assumptions, specifically the Neumann boundary conditions. [Bakka et al. \(2018\)](#) demonstrated that incorporating these assumptions can result in inferior outcomes compared to stationary models. While modeling events on dispersed spatial regions where there are physical barriers or, holes in the study area, stationarity is an unrealistic assumption ([Bakka et al., 2019](#)). This issue is common while exploring complex island structures. Similar coastline problems are

reported by Ramsay (2002), Wood et al. (2008) and Scott-Hayward et al. (2014). To handle the coastline problem, several studies have proposed solutions, such as computing the shortest distance in water (Wang and Ranalli, 2007; Scott-Hayward, 2014, Miller, 2014), defining boundary conditions using a smoothing penalty together with Neumann boundary condition (Ramsay, 2002), or using the Dirichlet boundary condition (Wood et al., 2008; Sangalli et al., 2013). However, these methods may not be suitable for complex archipelago structures with physical barriers. In addition to coastlines, other physical barriers include road networks, power lines, categorical health sectors, and areas with different land uses. Therefore, these models can be utilized in various domains, including geography, environmental science, and ecology, to examine the spread and migration of various entities, like air contamination, wildlife populations, and disease outbreaks.



**Figure 32: Barrier object and mesh with barrier object with event locations highlighted as red points**

Non-stationary Gaussian models with physical barriers are a type of statistical model used to analyze spatial data that vary over time or space and are influenced by physical or man-made barriers. Building on this, Bakka et al. (2019) introduced a methodology to deal with non-stationary and anisotropic spatial processes, with a focus on addressing complex archipelago structures where the coastline serves as a physical barrier. In the barrier model, the presence of barriers is represented by a latent variable that acts as a weight or factor for the predictors. The barriers are modeled as smooth functions of spatial or temporal variables, and the relationship between the barriers and the predictors is estimated using Bayesian methods. Although the barrier model was not designed specifically for linear road networks, Dawkins et al. (2019) made an effort to apply it to such networks. Another typical example presented by Krainski et al. (2019) is the use of barrier models in modeling anisotropic behavior, such as the propagation of noise in urban areas. In this study, noise data was collected from the city of Albacete, Spain, to analyze noise levels in a busy area of the city center with many bars and restaurants. The goal of the study is to analyze the fluctuations in noise levels in the city center if local noise regulations are being met. In their study, the buildings are used as physical barriers and the spatial process is considered on the road network of the study area.

In our current study, we have taken a similar approach to model traffic accidents by utilizing a barrier model in the road network of Barcelona. We have defined polygons of individual road

segments with a buffer as our study area and the remaining land areas that do not include roads serve as the physical barriers. The creation of the clipped buffer region and aggregation of the number of minor injuries (which serve as the response variable in the model) at the centroids of each road segment have been accomplished using the same approach outlined in Section 5.5.3.1. Similarly, the mesh for the entire study area, as shown in Figure 30, has been constructed using a method similar to that described in Section 5.5.3.1.

The *inla.over\_sp\_mesh()* function is utilized to determine which mesh triangles are contained within a polygon, while the *inla.barrier.polygon()* function has been employed to obtain the polygon surrounding the barrier. The barrier object is depicted in the left panel of Figure 32, where the grey area denotes the physical barrier, and the white area represents the road buffer polygons where spatial dependence will be analyzed. Points in red indicate the locations of traffic accidents used as event locations in the model. The triangulation will be created using a barrier model, in which the buffered road polygon serves as the normal terrain and areas without roads serve as physical barriers. The resulting mesh, with the polygon surrounding the barrier, as obtained using the *inla.barrier.polygon()* function (in blue) along with the event locations (in red) are displayed in the right panel of Figure 32.

As mentioned in Section 5.5.3.1, our approach involves aggregating event counts at the centroids of individual road segments. By using Poisson regression models and adopting a hierarchical Bayesian spatial model that accounts for barriers, we can model traffic accidents on individual road segments. Our response variable is the aggregate number of minor injuries recorded per day for each individual section of road. Following it, log-risk in Equation 4 can be modified to:

$$\log(\rho_i) = \beta_0 + Z_i\beta_1 + u(s_i) + \epsilon_i \dots\dots\dots (5)$$

Here,  $\beta_0$  corresponds to the intercept,  $Z_i$  represents the spatial covariates mentioned in Section 5.5.2 and  $\epsilon_i$  stands for an unstructured zero mean Gaussian random effect and log Gamma precision parameters 0.5 and 0.01, defined as penalized complexity (PC) priors (Simpson et al., 2017). We assigned default priors for all fixed-effect parameters to minimize their impact on the posterior distribution.  $u(\mathbf{s})$  is a non-stationary spatial random effect. Bakka et al. (2019) in their proposal suggested using a finite element method which is based on the SPDE approach. The proposed method involves a system of two SPDEs, where one is applied to the barrier region and the other to the rest of the area. The system of stochastic differential equations in question has a solution that exhibits a non-stationary spatial effect, represented as  $u(\mathbf{s})$ . The system can be mathematically modeled as a set of stochastic differential equations, which provide a continuous weak solution to the estimation problem:

$$u(s) - \nabla \cdot \frac{r_b^2}{8} \nabla u(s) = r_b \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_b \dots\dots\dots (6)$$

and

$$u(s) - \nabla \cdot \frac{r^2}{8} \nabla u(s) = r \sqrt{\frac{\pi}{2}} \sigma_u W(s), \text{ for } s \in \Omega_n \dots\dots\dots (7)$$

where  $u(\mathbf{s})$  is the spatial effect,  $\Omega_b$  the barrier area and  $\Omega_n$  is the remaining area and their disjoint union gives the whole study area  $\Omega$ . Ranges for the barrier and remaining areas are represented by  $r$  and  $r_b$  respectively.  $\sigma_u$  is the marginal standard deviation.  $\nabla$  is equal to  $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$  and  $W(\mathbf{s})$  stands for white noise. In contrast to stationary spatial effects, this method implies the creation of a GMRF at a local level, consisting of two governing equations - one for the normal area (buffered road polygon) and the other for the barrier area (areas without roads). The spatial effect prior is determined by two unknown hyperparameters, namely the standard deviation ( $\sigma_u$ ) and the range in the normal area ( $r$ ), while the range in the barrier area ( $r_b$ ) is maintained at a fixed, low value. Therefore, the system in Equation 6 and 7 represents a form of local averaging, with dependence on nearby values. This approach ensures that when two points are separated by a landmass, the small range in the barrier area prevents local averaging, forcing dependency to focus on movement around the barrier through local averaging in the buffered road polygon area. The system of differential equations in Equation 6 and 7 can be solved by constructing a Delaunay triangulation of the study area (as shown in Figure 32) and applying the finite element method, as described in Bakka et al. (2019). For the two hyperparameters in the model that define the covariance structure of  $u(\mathbf{s})$ , PC priors were assigned following the parametrization outlined in Simpson et al. (2017) and Fuglstad et al. (2019). These priors are designed to be minimally informative and to capture the uncertainty in the model.

### 5.5.3.3 Graph model on linear network

Previous subsection on explicit network triangulation and ongoing research on barrier models for complex land structures have highlighted issues related to boundary effects, including the creation of artifact spatial dependencies on the boundary. In standard meshes, boundaries are typically outside the spatial domain of interest, allowing for identification and elimination of these dependencies. However, in more complex meshes like network triangulation or barrier models, boundaries lie within the spatial domain, making it challenging to identify and eliminate these dependencies. Despite these difficulties, in these works we have managed to identify them. However, a different approximation is needed in which the SPDE-INLA approximation does not cause these fictitious spatial dependencies.

In this context, Bolin et al. (2022) presented an alternative to using the Euclidean distance by defining similar models with a non-Euclidean metric on a graph. Literature shows, statistical models are required to be defined on linear networks, such as connected river or street networks (Baddeley et al., 2017; Cronie et al., 2020). In such cases, it is necessary to define a model using a metric on the network rather than the Euclidean distance between points. However, constructing Gaussian fields over linear networks, or more generally on metric graphs, presents a challenge. This is due to the difficulty of finding flexible classes of functions that are positive definite when a non-Euclidean metric is used. While the geodesic metric, which calculates the shortest distance between two points, has gained much attention in research, it has been criticized for its unrealistic applicability to many real-world processes (Baddeley et al., 2017). Therefore, researchers often employ an alternative metric known as electrical resistance

distance (Okabe and Sugihara, 2012). Anderes et al. (2020) utilized this metric to create isotropic covariance functions for a specific type of metric graph with Euclidean edges. They demonstrated that, for graphs with Euclidean edges, it is possible to define a valid Gaussian field by using a Matérn type covariance function (Matérn, 1960):

$$r(s, t) = \frac{\Gamma(\nu)}{\tau^2 \Gamma(\nu + 1/2) \sqrt{4\pi\kappa^2\nu}} (\kappa d(s, t))^\nu K_\nu(\kappa d(s, t)) \dots\dots\dots (8)$$

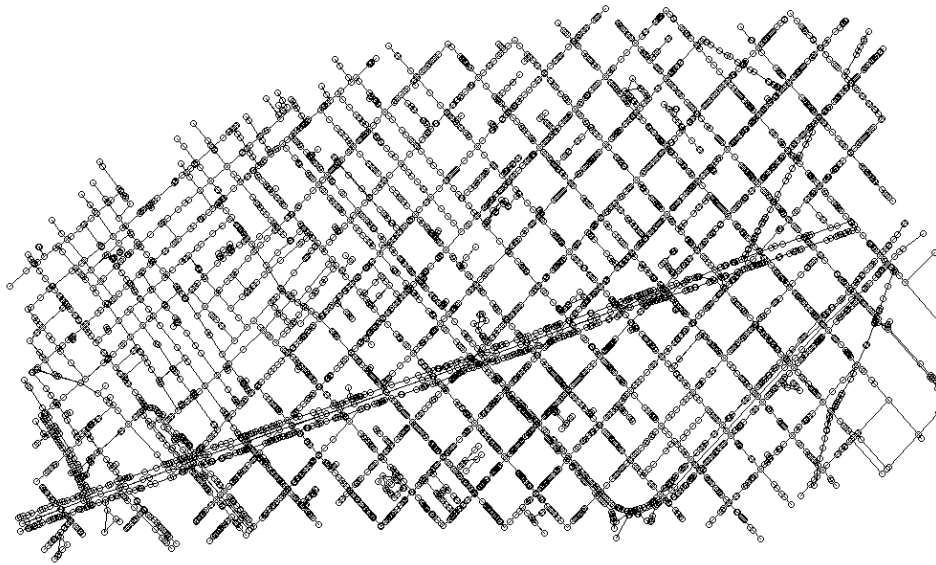
where  $d(\cdot, \cdot)$  is the resistance metric,  $\tau, \kappa > 0$  are parameters controlling the variance and practical correlation range, and  $0 < \nu \leq 1/2$  is a parameter controlling the sample path regularity. The limitation of  $\nu \leq 1/2$  means that we cannot use this approach to create differentiable Gaussian processes on metric graphs, even if they have Euclidean edges. Due to this constraint, as well as some other complex challenges in creating Gaussian fields via covariance functions on non-Euclidean spaces, Bolin et al. (2022) take a different approach in this work and focus on creating a Gaussian random field  $u$  on a compact metric graph  $\Gamma$  as a solution to a SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau u) = \mathcal{W}, \quad \text{on } \Gamma \dots\dots\dots (9)$$

where  $\alpha = \nu + 1/2$ ,  $\Delta$  is the Laplacian equipped with suitable boundary conditions in the vertices, and  $\mathcal{W}$  is Gaussian white noise. The advantage with this approach is that, if the solution exists, it automatically has a valid covariance function. The reason for considering this particular SPDE is that when Equation 9 is considered on  $R^d$ , it has Gaussian random fields with the covariance function (as in Equation 8) as stationary solutions (Whittle, 1963). Lindgren et al. (2011) proposed a technique for extending the Matérn fields to Riemannian manifolds by defining Whittle-Matérn fields as solutions to Equation 9 specified on the manifold. Since then, this method has been expanded to various scenarios (Lindgren et al., 2022), including non-stationary (Bakka et al., 2019; Hildeman et al., 2021) and non-Gaussian (Bolin, 2014; Bolin and Wallin, 2020) models.

However, one of the primary challenges of extending the SPDE approach to metric graphs is defining the differential operator and determining the type of covariance functions that would result. In this context, quantum graph theory plays a vital role (Bollin et al., 2022). A quantum graph is a combination of a metric graph and a differential operator, where the Laplacian is the most important operator (Berkolaiko and Kuchment, 2013). However, there are multiple ways to define the Laplacian on a metric graph due to the existence of various options for vertex conditions. The Laplacian can be defined as the second derivative on each edge. At the vertices, however, there are various options for defining the operator depending on the choice of boundary conditions or vertex conditions. The Kirchhoff conditions is one of the most popular choices. The Laplacian with these vertex conditions is often referred to as the Kirchhoff-Laplacian. According to (Berkolaiko and Kuchment, 2013), this Laplacian is self-adjoint. This Laplacian is the most natural form for defining Whittle-Matérn fields on metric graphs. The Whittle-Matérn covariance function is a popular choice for modeling spatial dependence in

Gaussian processes, as it provides a flexible framework for capturing different types of spatial dependence. At a high level, a Gaussian Whittle-Matérn random field on a metric graph is a collection of random variables, one for each node on the graph, that are jointly Gaussian with a Whittle-Matérn covariance function (Berkolaiko and Kuchment, 2013). The covariance function is parameterized by two hyperparameters namely, the smoothness parameter and the range parameter. In the context of a metric graph, the distance matrix of the graph is used to determine the covariance matrix of the random field. The distance matrix gives the shortest path distances between all pairs of nodes on the graph, which is used to compute the Whittle-Matérn covariance function. In their recent study, Bollin et al. (2022) utilize quantum graph theory to define an operator and prove that Equation 9 has a unique solution, from which they derive sample path regularity properties. They demonstrate that this solution has Markov properties when  $\alpha$  is a natural number, and in such cases, they derive the finite dimensional distributions of the process analytically. When  $\alpha = 1$ , the resulting process exhibits a covariance function that is similar to the exponential covariance function. Specifically, it corresponds to the case where  $\nu = 1/2$  in Equation 8, which was previously shown to be a valid covariance for metric graphs with Euclidean edges (Anderes et al., 2020).

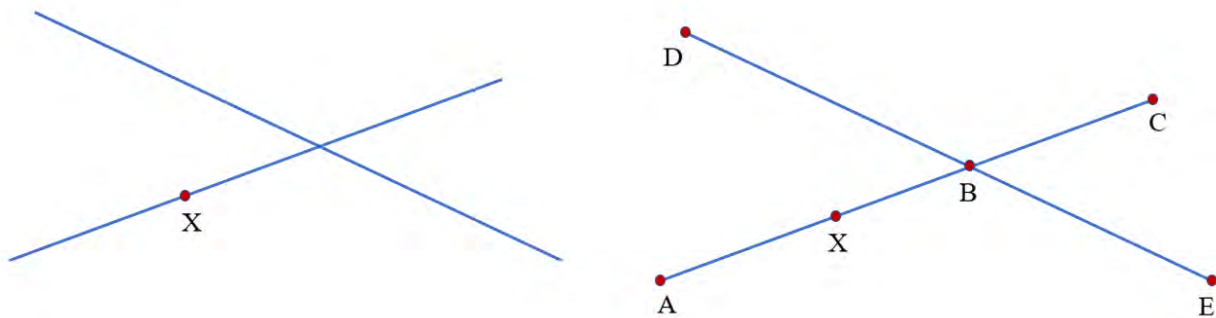


**Figure 33: Graph data structure of the traffic accident locations as nodes and road networks as edges**

In the current study, we have implemented Whittle-Matérn fields that are defined using a fractional SPDE on a compact metric graph within the R-INLA interface. This implementation serves as a natural extension of Gaussian fields with Matérn covariance functions on Euclidean domains to the non-Euclidean metric graph settings. We have used the same traffic accident dataset of Barcelona city spanning from January 2010 to December 2019 reported in Section 5.5.2. We focused on the number of minor injuries as the response variable for our modeling process. To employ a graph model, we first converted the dataset into a graph data structure that is compatible with the model. In this structure, we have represented individual accident

locations, start and end points of road segments, and intersecting points of road segments as nodes or vertices, while the connecting road segments for the nodes are represented as edges.

Figure 33 depicts the resulting graph comprised of 7401 vertices and 7937 edges. It is important to note that we have aggregated the number of minor injuries for different time instances at the same location and considered it as a single vertex. Thus, in this model and also in the previous two models (mentioned in Section 5.5.3.1 and Section 5.5.3.2) no temporal covariates are considered. Moreover, we have included the start and end points of each road segment and the intersection points of two or more road segments as new vertices in the graph data structure. This process resulted in a different number of vertices and edges than the original dataset of 2058 road segments and 11067 traffic accidents records (represented as red points) illustrated in the right panel of Figure 29. The vertices that have been added new to the network are assigned a value of zero for minor injuries. In situations where the intersection point between two road segments has already recorded one or more accidents throughout the entire study period, the total number of minor injuries is aggregated and assigned to that junction point, which is considered as a single vertex. On the other hand, distances to nearby facilities such as bus stops, municipal markets, restaurants, schools, and street markets are computed from each vertex and incorporated as spatial covariates in the model. Additionally, other covariates such as road length, road type, and road speed limit are determined for each vertex based on its position in the specific road segment.



**Figure 34: Road network conversion to graph data structure**  
two road segments having one traffic accident location highlighted in red (left), same road segments with start and end points along with intersection point and accident location highlighted in red (right)

As a show case example, we illustrate two road segments and a single traffic accident location represented by a red point (X) in the left panel of Figure 34. We introduced the start and end points (A, C, and D, E) of the two road segments and their intersecting point (B) as vertices. The connecting lines between these vertices, including the accident location, were considered as edges. This resulted in a graph with six vertices (A, X, B, C, D, and E) and five edges (A-X, X-B, B-C, D-B, and B-E) as depicted in the right panel of Figure 34. In summary, we have converted the traffic accident dataset into a graph data structure that considers the road network and traffic accident locations as vertices and their connections as edges. This process resulted in



a graph model that differs in the number of vertices and edges from the original dataset, and no temporal covariates are considered in our model.

In the next step, Euclidean distances between each vertex have been calculated and used in the exponential graph model. These distances are now considered as the length of each edge of the graph having a mean value of 10.92 meters with a maximum value of 183.29 meters. We have introduced the Gaussian Whittle-Matérn random fields on metric graphs and have provided a comprehensive characterization of their regularity properties and statistical properties (Bolin et al., 2020). We argue that this class of models is a natural choice for applications where Gaussian random fields are needed to model data on metric graphs. Of particular importance here are the Markov cases (Bolin et al., 2020). We derived explicit densities for the finite dimensional distributions in the exponential case where we can note that the model has a conditional autoregressive structure of the precision matrix (Besag, 1974). For the differentiable cases we derived a semi-explicit precision matrix formulated in terms of conditioning on vertex conditions. In both cases, we obtain sparse precision matrices that facilitates the use in real applications to big datasets via computationally efficient implementations based on sparse matrices (Rue and Held, 2005). Finally, while implementing it in R-INLA, a generic model is defined using the function `inla.rgeneric.define()`, which takes as input a function `rmodel` and additional variables or functions in that might be used in the `rmodel`. The resulting `inla.rgeneric` object can be used to define a normal model component in INLA using function `f()`. The function `rmodel` needs to provide the required features, including the *graph*, the *precision matrix*  $Q(\theta)$ , the zero mean, the initial values of  $\theta$ , the log-normalizing constant, and the log-prior. For the proposed graph-model, two hyperparameters are used, and a good reparameterization is required for INLA to work well. Gaussian priors are used for both hyperparameters.

#### 5.5.4 Results and Discussion

This section presents the findings of the analysis and methodological approach developed in methodology section. We have used the same dataset and compare the performance of the three different modeling approaches. It is worth noting that we did not consider any temporal covariates in any of these modeling processes. For both the network mesh model and barrier model, we executed batteries of similar models based on the argument values to create SPDE triangulation. The default prior distributions for all parameters in R-INLA are selected based on commonly used priors in previous studies (Blangiardo and Cameletti, 2015; Rue et al., 2017; Moraga, 2019). Our results indicate that our findings are robust against alternative priors, as we ran several cases with different priors and obtained the same results. In the case of the graph model, we executed several models with different log-prior probability density for the model parameters and ultimately selected the best fitted model. We assessed the performance of the models from the three different approaches using deviance information criterion (DIC) and the Watanabe–Akaike information criterion (WAIC), balancing model accuracy against complexity (Spiegelhalter et al., 2002). We have used conditional predictive ordinate (CPO) value (Gelfand

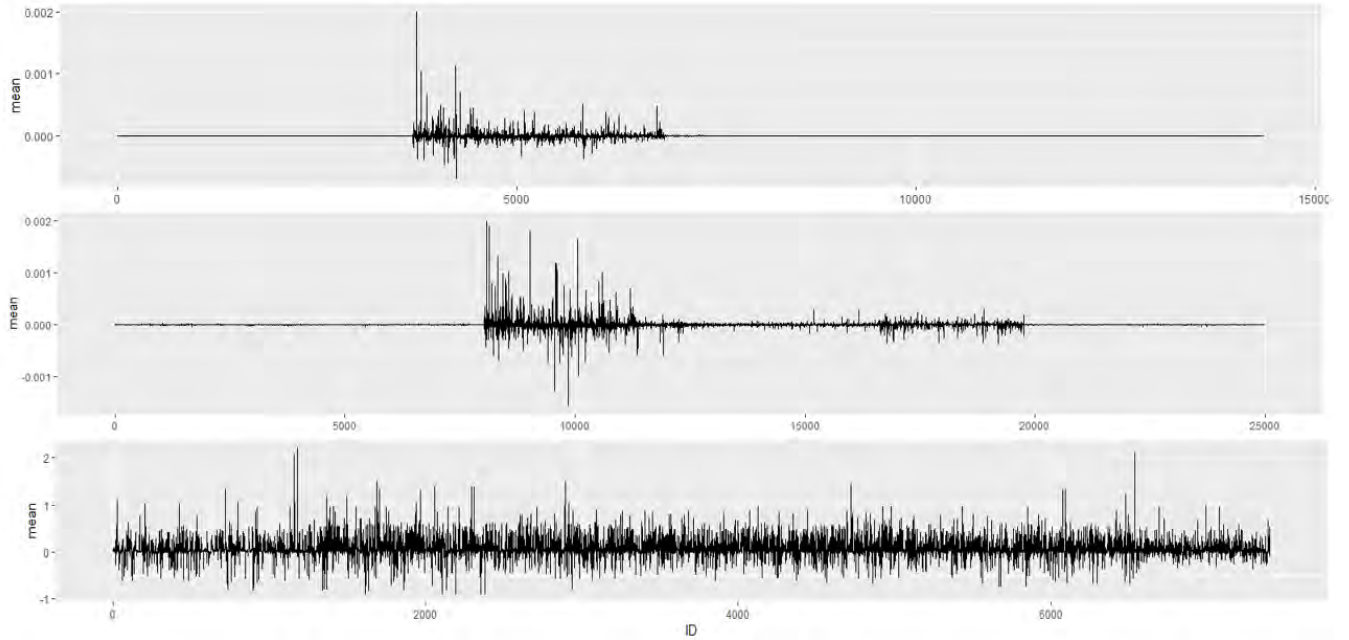
et al., 1992) which also acts as a selection measure; smaller value of CPO indicates a better prediction quality of the model. Execution time for each modeling approach has also been reported as a measure of comparison. In Table 4, we report the selected models with the lowest DIC, WAIC, CPO, and execution time for each of the three categories from their respective battery of models.

**Table 4: DIC, WAIC and CPO values of Fitted Models**

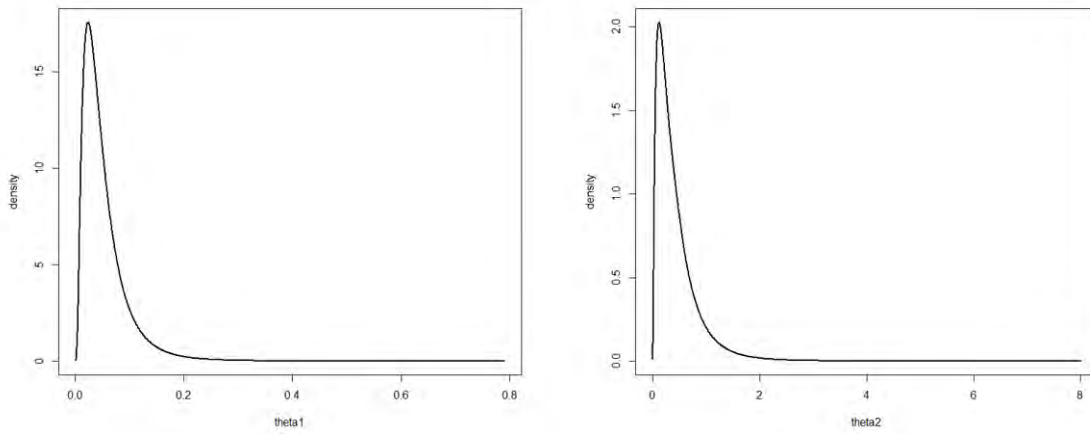
	<b>Network Mesh Model</b>	<b>Barrier Model</b>	<b>Graph Model</b>
<b>DIC</b>	23390.17	23390.40	15393.55
<b>WAIC</b>	23382.56	23382.68	15094.93
<b>CPO</b>	0.3252899	0.3252868	0.3288132
<b>Execution Time (Secs.)</b>	15.2	54.3	12.8

Looking at the DIC values, we can see that the graph model has the lowest DIC value (14980.69), followed by the two other models with very similar DIC values (23390.17 for the network mesh model and 23390.40 for the barrier model). A lower DIC value indicates a better fit, so we can conclude that the graph model is the best-fitted model among the three. Similarly, looking at the WAIC values, we can see that the graph model has the lowest value (14943.23), again indicating that it is the best-fitted model.

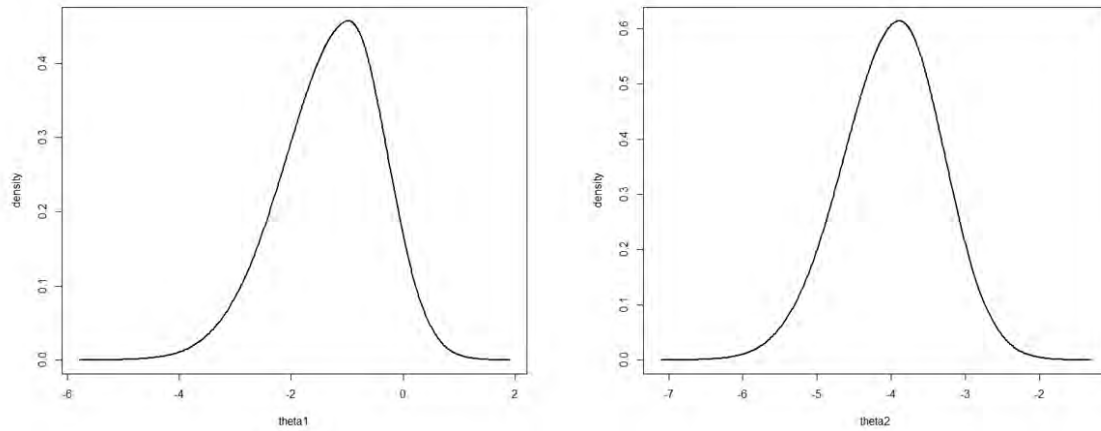
The difference between the WAIC values of the other two models is very small, suggesting that they have very similar performance in terms of fitting the data. Finally, the CPO values for all three models are close to each other, indicating that they all have similar predictive performance. Furthermore, the graph model has the shortest execution time among the three models, with an execution time of only 12.8 seconds. The barrier model has the longest execution time among the three models, taking 54.3 seconds to execute. This implies that the graph model is not only more accurate but also more computationally efficient than the other two models. In conclusion, the graph model has the best performance according to both DIC and WAIC, while all three models have similar predictive performance according to CPO. These results suggest that the graph model is the best model among the three fitted models for describing the data as well as computational efficiency.



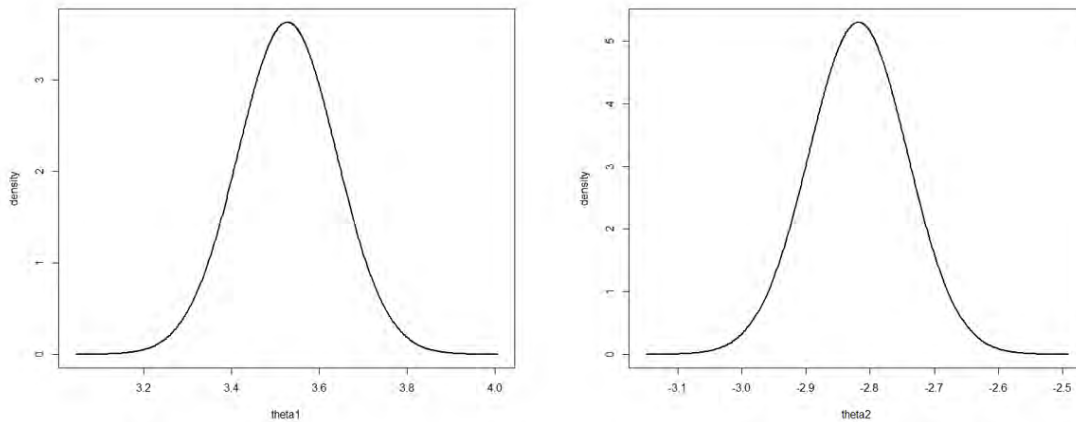
**Figure 35: Marginal posterior mean of the spatial random effect network mesh model (top panel), the barrier model (middle panel), and the graph model (bottom panel)**



**Figure 36: Marginal posterior distributions of network mesh model hyperparameters  $\theta_1$  (left) and  $\theta_2$  (right)**



**Figure 37: Marginal posterior distributions of barrier model hyperparameters  $\theta_1$  (left) and  $\theta_2$  (right)**



**Figure 38: Marginal posterior distributions of graph model hyperparameters  $\theta_1$  (left) and  $\theta_2$  (right)**

While comparing the significance of the fixed effects for the three modeling techniques, we have observed that the covariates included in all models do not exhibit a statistically significant influence on the outcome.

Furthermore, [Figure 35](#) displays the marginal posterior mean of the spatial random effect for three different models: the network model (top panel), the barrier model (middle panel), and the graph model (bottom panel). The horizontal axis of [Figure 35](#) represents the nodes or vertices used in each model: 14,368 triangulation nodes of the SPDE network mesh (top panel), 24,993 nodes of the barrier model (middle panel), and 7401 vertices of the graph model (bottom panel). It is worth noting that the vertices of triangles located on road segments with higher traffic accident occurrences (represented as dark red patches in [Figure 31](#) and [Figure 32](#)) exhibit a more prominent and statistically significant spatial effect. Conversely, vertices without any accident

events do not display any spatial effect. In the case of the graph model, the spatial effect is significant in all the vertices included in the modeling process.

The Appendix includes marginal posterior distributions of model hyperparameters for three different models: network mesh model in [Figure 36](#), barrier model in [Figure 37](#), and graph model in [Figure 38](#). The spatial range values for each model are reported as follows: 0.0523 Km (equivalent to 52.3 meters) for the network mesh model, 0.0183 Km (equivalent to 18.3 meters) for the barrier model, and 0.0596 Km (equivalent to 59.6 meters) for the graph model.

Designing the INLA-SPDE triangulation specifically on road networks can offer several advantages for modeling random spatial events like crime or traffic accidents. One of the primary advantages of this approach is improved computational efficiency. By focusing the triangulation specifically on the road network, researchers can limit the number of nodes and edges that need to be modeled, reducing the computational burden associated with modeling data over an entire study area. Another advantage of designing the triangulation specifically on road networks is the potential for more precise modeling of spatial relationships along the road network. In contexts like traffic accidents or crime hotspots, spatial relationships along the road network are often critical for understanding patterns and trends in the data. By designing the triangulation specifically on the road network, researchers can more accurately capture these relationships, potentially leading to more accurate and informative models. Additionally, designing the triangulation specifically on road networks can help to reduce bias and improve the accuracy of the resulting models. By focusing on the road network, researchers can more effectively control for confounding variables that may be present in other areas of the study region, leading to more precise and accurate estimates of the relationships between variables. While there are several potential advantages to design the INLA-SPDE triangulation specifically on road networks for modeling random spatial events like crime or traffic accidents, it is important to carefully consider the potential limitations and trade-offs associated with this approach. At the same time, the principal advantage with the barrier model is that the computational cost is the same as for the stationary model. In general, the model is easy to use, and can deal with both sparse data and very complex physical barriers.

However, while using these two techniques for modeling spatial relationships can be effective in capturing relationships along the road network, there are some potential limitations and trade-offs that should be considered.

For instance, the approach may not be able to capture important spatial relationships outside of the road network, particularly in contexts where traffic accidents can occur in adjacent areas or neighborhoods. Furthermore, accurately modeling spatial relationships along the road network can be challenging in areas with complex road networks or where the network is subject to frequent changes or updates, which may require frequent updates to the triangulation to accurately capture changes in the road network.

Another significant limitation we observed in the proposed methodology is the boundary effect, which can lead to biased estimates and prediction errors near the boundary if the mesh does not cover the entire domain. Spatial Gaussian fields (SGFs) are commonly utilized as model

components in the construction of spatial or spatio-temporal models for various applications, including the Generalized Additive Model (GAM) framework, to represent the residual spatial structure resulting from unmeasured spatial covariates, spatial aggregation, and spatial noise. The use of a buffer road network in current studies adds complexity to the boundary regions, which can influence the spatial effect of the model. [Krainski et al. \(2018\)](#) propose creating a mesh to represent the spatial process as the first step in fitting a SPDE model. Building an SPDE mesh for a continuous region is relatively straightforward, but the creation of an SPDE network mesh requires fine-tuning to identify the best fit values for minimum allowed distance between vertices and maximum permissible triangle edge length for the inner (and outer) regions. Careful selection of additional points around the boundary or outer extension is also necessary. As a general rule, the variance near the boundary is inflated by a factor of two along straight boundaries and by a factor of four near right-angled corners ([Lindgren et al., 2011](#); [Lindgren and Rue, 2015](#)). The complex boundary region of the buffer road network with several right-angled corners makes the process critical. The boundaries in the proposed mesh are located inside the mesh and not outside, as in a standard mesh, which creates fictitious spatial structures. Due to the complex boundary nature, it is necessary to reduce the high boundary effect that may cause a variance twice or four times as great at the border as within the domain ([Lindgren and Rue, 2015](#)). Although the residual diagnostics and predicted risk maps produced by the model match the original observed records, the correlation values of the model indicate the need for improvement.

On the other hand, the proposed construction by [Bolin et al. \(2022\)](#) has two significant advantages. Firstly, it has Markov properties, implying that the precision matrices of the finite dimensional distributions of the process will be sparse, which simplifies the use of the model for big datasets. Secondly, the model is well-defined for any compact metric graph, not just the subclass with Euclidean edges. Additionally, the authors derive an explicit density for the finite dimensional distributions of the processes for higher values of  $\alpha$ , which is not possible using the corresponding Matérn covariance function to construct a valid Gaussian process in general, even for graphs with Euclidean edges ([Anderes et al., 2020](#)). Therefore, this construction provides a covariance function for differentiable random fields on compact metric graphs ([Bollin et al., 2022](#)).

## 6. DISCUSSION

---

Spatiotemporal analysis is crucial for gaining insight into the dynamic behavior of complex systems and developing predictive models in various scientific fields. The INLA-SPDE methodology has proven to be a powerful tool in statistical modeling due to its many advantages over other techniques. It provides low computation time, making it an attractive option for large datasets. Additionally, as the basic logic is Bayesian inference, it does not require only normally distributed data, enabling its application in a vast domain of fields. The methodology also allows for the implementation of both spatial and temporal effects, as well as the analysis of their significance in the model. INLA-SPDE permits the integration of a substantially high number of covariates and can also accommodate new covariates at a later stage of the process. Moreover, the level of significance for each covariate can be analyzed, which further strengthens its utility in statistical modeling.

It is worthy to mention, that in any data analysis project, the quality and accuracy of the data is vital. Therefore, we spent a considerable amount of time accessing huge datasets from various open-source portals and ensuring that the data was accurate, complete, and reliable. Once we had obtained the data, we then proceeded to clean and validate it. This involved identifying and correcting any errors, inconsistencies, or missing values in the data. After the data cleaning and validation process, we moved on to performing basic exploratory analysis. This step involved examining the data to identify patterns, trends, and relationships between different variables. We used R code to perform these analyses, which allowed us to explore the data and generate visualizations that helped us to better understand the data quickly and efficiently. In addition to the data cleaning and exploratory analysis, we also performed data twinning and added new spatial variables to the dataset. This task involved combining datasets from different sources and merging them into a single, unified dataset. We also incorporated new spatial variables, such as geographic boundaries, Euclidian distances of events from points of interests into the dataset to enable spatial analysis. By investing significant time and effort in data preparation and exploration, we were able to ensure that the subsequent analysis and inference we performed was based on reliable and accurate data. This approach helped us to avoid potential errors or biases that may have been introduced by analyzing incomplete or inaccurate data. Overall, this rigorous approach to data preparation and exploration was essential in enabling us to draw robust conclusions and make accurate predictions based on the data.

Our research projects on the relationship between pollution and COVID-19 infection rates in Catalonia and some selected counties of New York utilized spatiotemporal modeling using the traditional SPDE triangulation technique applied for continuous spatial regions ([Chaudhuri et al., 2022a](#); [Díaz-Avalos et al., 2020](#)). However, during the triangulation process for the New York study, we encountered challenges and observed the presence of boundary effects. From a scientific perspective, our interest lies in investigating how similar research studies utilizing INLA-SPDE have been carried out to model complex land structures in coastal regions and islands. Current literature shows that even for complex and distributed spatial regions, researchers have utilized a traditional continuous region concept to design the SPDE

triangulation. This approach involves generating an SPDE mesh for the entire study region, despite the presence of physical barriers that make the study area complex and distributed.

To address this issue, we propose the use of alternative modeling techniques, such as network-based spatial statistical models, that are specifically designed to analyze events on linear networks. These models take into account the unique characteristics of linear networks and can accurately predict events at specific spatial points located on the road network. By incorporating the spatial structure of the linear network into the modeling process, these models provide more realistic and accurate predictions of events on the network. Using this approach allows researchers to focus the triangulation specifically on the road network, limiting the number of nodes and edges that need to be modeled and reducing the computational burden associated with modeling data over an entire study area. Additionally, researchers can more effectively control for confounding variables that may be present in other areas of the study region, leading to more precise and accurate estimates of the relationships between variables. In our recent research project [Chaudhuri et al. \(2022b\)](#) we proposed spatiotemporal modeling of road traffic accidents using explicit network triangulation on the road network of London, UK. In our following project [Chaudhuri et al. \(2023\)](#), SPDE triangulation has been designed precisely on linear road networks of Barcelona, Spain to generate dynamic traffic accident risk maps. The methodology used in these two studies is a novel approach to perform spatiotemporal analysis precisely on road network and contributes to the relatively small amount of literature in this domain. However, in both cases, the complex boundary regions of the buffer road network result in high boundary effect, that can influence the spatial effects of the models. This is a serious limitation of the SPDE network triangulation approach. Furthermore, accurately modeling spatial relationships along the road network can be challenging in areas with complex road networks or where the network is subject to frequent changes or updates, which may require frequent updates to the triangulation to accurately capture changes in the road network.

The SPDE triangulations proposed in network triangulation assume stationarity and isotropy, meaning that the autocorrelation between two locations only depends on Euclidean distance. However, using this approach can be problematic because it includes additional assumptions, specifically the Neumann boundary conditions. Typically, spatial models assume isotropy and stationarity, which means that spatial dependence is uniform throughout the study area and direction invariant. To analyze spatial data that vary over time or space and are influenced by physical or man-made barriers, non-stationary Gaussian models with physical barriers are often used. [Bakka et al. \(2019\)](#) introduced a methodology to deal with non-stationary and anisotropic spatial processes, with a focus on complex archipelago structures where the coastline serves as a physical barrier. [Dawkins et al. \(2021\)](#) attempted to apply this barrier model to linear road networks, while [Krainski et al. \(2018\)](#) used it to model anisotropic behavior such as noise propagation in urban areas. In our current study, we employed a similar approach to model traffic accidents in the road network of Barcelona. We defined polygons of individual road segments with a buffer as our study area, and the remaining land areas that do not include roads served as the physical barriers.



Although using network-based spatial statistical models and triangulation techniques can be useful for modeling spatial relationships along road networks, there are potential limitations and trade-offs that need to be considered. One major limitation of the proposed methodology is the boundary effect, which can result in biased estimates and prediction errors near the edge of the domain if the mesh does not cover the entire area. SGFs are commonly used to model residual spatial structure resulting from unmeasured spatial covariates, spatial aggregation, and spatial noise in spatial or spatiotemporal models. The use of a buffer road network in the current studies adds complexity to the boundary regions, which can affect the spatial effect of the model. These challenges have motivated us to investigate the feasibility of implementing spatial modeling techniques in complex or distributed spatial regions, such as islands, road networks, or areas demarcated by boundaries. In addition, our current study paves the way for future research to explore the impact of boundary effects on model performance and analyze variations in spatial effects.

We are currently analyzing these problems in a subsequent study and exploring potential solutions. In collaboration with [Bolin et al. \(2022\)](#), we are working on an alternative method that does not rely on Euclidean distance. Instead, we aim to define models with a non-Euclidean metric on a graph, which requires using a metric on the network rather than the Euclidean distance between points. However, constructing Gaussian fields over linear networks or metric graphs poses a challenge because finding flexible classes of functions that are positive definite when a non-Euclidean metric is used is difficult. We are presently working on this approach, which extends Gaussian fields with Matérn covariance functions on Euclidean domains to the non-Euclidean metric graph setting, via a fractional stochastic partial differential equation on the graph. The study demonstrates that these processes exist and have sample path regularity properties, including differentiable Gaussian processes. Additionally, the study establishes that a model subclass contains processes with Markov properties and provides a computationally efficient method for assessing their finite dimensional distributions. These proposed models can be utilized for statistical inference without the need for approximations, and several statistical properties can be derived, including consistency of maximum likelihood estimators and asymptotic optimality properties of linear prediction based on the model with mis-specified parameters.

In the current study, we have implemented Whittle-Matérn fields that are defined using a fractional SPDE on a compact metric graph within the R-INLA interface and named it as graph model. In our recent research project, we compared the efficacy of the proposed graph model to two other modeling approaches - network mesh and barrier model - using the same dataset. No temporal covariates were included in any of the models. We evaluated the performance of the models using the deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC), which balance model accuracy against complexity. We also used the conditional predictive ordinate (CPO) value as a selection measure, with a smaller value indicating better prediction quality. Additionally, we measured the execution time for each modeling approach. The results indicated that the graph model performed the best according to

both DIC and WAIC. While all three models had similar predictive performance according to CPO, the graph model was the most efficient among the three. The results suggest that the graph model is the best model among the three fitted models for describing the data as well as computational efficiency.

The proposed technique is more flexible and statistically convincing, and its construction of the graph model has two main advantages. Firstly, it has Markov properties, which means that the precision matrices of the process's finite dimensional distributions will be sparse. This simplifies the use of the model for large datasets. Secondly, the model is valid for any compact metric graph, not just the subclass with Euclidean edges. Additionally, the authors have derived an explicit density for the finite dimensional distributions of the processes for higher values of  $\alpha$ . This is not possible using the corresponding Matérn covariance function to construct a valid Gaussian process, even for graphs with Euclidean edges. Therefore, this construction provides a covariance function for differentiable random fields on compact metric graphs, as described by [Bollin et al. \(2022\)](#).

We are currently investigating the use of a graph model with a battery log-prior probability density for the model parameters to assess its performance under various conditions. We are validating and testing the model with both simulated and real-time datasets. It is important to note that we are combining the response variable (number of minor injuries) for different time instances at the same location and treating it as a single vertex in the model. We are not including any temporal covariates in this model. We have added start and end points of each road segment and intersection points of two or more road segments as new vertices in the graph data structure. As a result, the number of vertices and edges in the graph data structure differs from the original dataset of road segments and traffic accidents. We have assigned a value of zero for minor injuries to the newly added vertices, while for intersection points that have already recorded accidents, we have combined the number of minor injuries and considered it as a single vertex. In addition, we have incorporated spatial covariates like distances to nearby facilities as well as other covariates such as road length, road type, and road speed limit for each vertex based on its position in the specific road segment. However, we are still exploring ways to incorporate temporal covariates in the graph vertices. We found that the model performance needs improvement in predicting future events accurately on the road network, particularly for events like traffic accidents or street crimes where temporal covariates are significant.

We are continuing our research to enhance the model performance in a spatiotemporal context. This field of spatial modeling using INLA and SPDE is particularly exciting for complex spatial regions that have physical barriers or linear networks such as roads or river systems. It presents an offers a unique opportunity for research that may lead to new avenues for investigation in the field. As a result, we are committed to pursue further research in this area for our thesis work.

## 7. CONCLUSIONS

---

The current study discusses the challenges of using INLA and traditional SPDE method in implementing Bayesian spatiotemporal modeling in complex or distributed spatial regions demarcated by boundaries and linear networks like road networks. These challenges call for a comprehensive and novel approach to address them. This may involve improving the SPDE triangulation approach, especially for linear networks, or developing a generalized approach to model spatial and spatiotemporal events in complex land structures. The study proposes to develop an innovative and realistic computational strategy for constructing spatial triangulations constrained to linear network topologies. Additionally, it intends to establish a modeling framework to explore spatiotemporal phenomena in complex spatial regions with physical barriers.

In the initial phase, the novel concept of designing the SPDE triangulation precisely on linear networks have been introduced. But the presence of complex boundary regions resulted in artificial spatial structures and dependencies leading to unavoidable boundary effects in the model. To address this, as an alternative computational strategy, nonstationary barrier models have been implemented but the boundaries remained within the spatial domain of interest, which hindered the reduction of high boundary effects. Finally, a new class of Gaussian processes on compact metric graphs, incorporating Whittle-Matérn fields defined by a fractional SPDE on a metric graph has been introduced. This approach extends Gaussian fields with Matérn covariance functions on Euclidean domains to non-Euclidean metric graph settings. The proposed technique uses a graph model that is more flexible and statistically convincing, with advantages of Markov properties and validity for any compact metric graph.

At present, the model is equipped with spatial covariates, hence its capability to predict future events is suboptimal, specifically for real-time scenarios where temporal covariates can play a vital role. We are currently exploring the graph model with a battery of log-prior probability density for the model parameters and conducting validation and testing with simulated and real-time datasets. The objective is to enhance the model performance in the spatiotemporal context, thereby opening up new possibilities for investigation in the field of Bayesian spatiotemporal modeling for complex spatial regions having physical barriers and for linear networks.

Scientific advancements in spatiotemporal modeling in complex spatial regions and regions having geographical or man-made physical barriers can provide opportunities for effective modeling real-time events. Furthermore, the development of new statistical models and techniques can improve the accuracy of the predictions, leading to a range of enhanced applications and improved management and control of real-time environmental issues, marine phenomena and hazards, urban problems such as traffic accidents, traffic congestion, antisocial activities, and air pollution. These advancements can have a profound impact on our understanding of the dynamics of epidemics and other chronic and infectious diseases, enabling us to make strategic decisions and improve public health management. The applications of these developments are wide-ranging and can extend to fields such as environmental science, economics, epidemiology, and ecology.

## References

- Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5), 633-642.
- Abramowitz, M., & Stegun, I. A. (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. *National Bureau of Standards Applied Mathematics Series 55*. Tenth Printing.
- Adler, P. B., & Lauenroth, W. K. (2003). The power of time: spatiotemporal scaling of species diversity. *Ecology letters*, 6(8), 749-756.
- Afrin, T., & Yodo, N. (2021). A probabilistic estimation of traffic congestion using Bayesian network. *Measurement*, 174, 109051.
- Aguilar-Palacio, I., Martínez-Beneito, M. A., Rabanaque, M. J., ... & Martos, C. (2017). Diabetes mellitus mortality in Spanish cities: trends and geographical inequalities. *Primary Care Diabetes*, 11(5), 453-460.
- Ajelli, M., Merler, S., Fumanelli, L., Pastore y Piontti, A., Dean, N. E., Longini, I. M., ... & Vespignani, A. (2016). Spatiotemporal dynamics of the Ebola epidemic in Guinea and implications for vaccination and disease elimination: a computational modeling analysis. *BMC medicine*, 14(1), 1-10.
- Al-Kindi, K. M., Alkharusi, A., Alshukaili, D., ... & El Kenawy, A. M. (2020). Spatiotemporal assessment of COVID-19 spread over Oman using GIS techniques. *Earth Systems and Environment*, 4, 797-811.
- Allepuz, A., Casal, J., Napp, S., Saez, M., ... & Saez, J. L. (2011). Analysis of the spatial variation of bovine tuberculosis disease risk in Spain (2006-2009). *Preventive Veterinary Medicine*, 100, 44-52.
- Allepuz, A., García-Bocanegra, I., Napp, S., Casal, J., Arenas, A., Saez, M., & González, M. A. (2010). Monitoring Bluetongue disease (BTV-1) epidemic in southern Spain during 2007. *Preventive Veterinary Medicine*, 96(3-4), 263-271.
- Allepuz, A., Saez, M., Solymosi, N., Napp, S., & Casal, J. (2009). The role of spatial factors on the success of an Aujeszky's disease eradication programme in a high pig density area (northeast Spain, 2003-2007). *Preventive Veterinary Medicine*, 91(2-4), 153-160.
- An, Y., Tsou, J. Y., Wong, K., Zhang, Y., Liu, D., & Li, Y. (2018). Detecting land use changes in a rapidly developing city during 1990–2017 using satellite imagery: A case study in Hangzhou Urban area, China. *Sustainability*, 10(9), 3303.
- Anderes, E., J. Møller, J. G. Rasmussen, et al. (2020). Isotropic covariance functions on graphs and their edges. *Ann. Stat.* 48(4), 2478–2503.
- Apostolopoulos, D. N., Avramidis, P., & Nikolakopoulos, K. G. (2022). Estimating quantitative morphometric parameters and spatiotemporal evolution of the Prokopos Lagoon using remote sensing techniques. *Journal of Marine Science and Engineering*, 10(7), 931.
- Asian Development Bank. (2012). *Maldives: Tsunami emergency assistance project*. Retrieved October 12, 2021, from <https://www.adb.org/documents/maldives-tsunami-emergency-assistance-project>

- Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, *10*(6), 760-766.
- Baddeley, A., G. Nair, S. Rakshit, and G. McSwiggan (2017). Stationary point processes are uncommon on linear networks. *Stat* *6*(1), 68–78.
- Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC press.
- Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis* (Vol. 413, No. 8). Essex: Longman Scientific & Technical.
- Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Illian, J., Krainski, E., Simpson, D., & Lindgren, F. (2018). Spatial modeling with R-INLA: A review. *WIREs Computational Statistics*, *10*(6).
- Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D., & Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial statistics*, *29*, 268-288.
- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.
- Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(4), 825-848.
- Barbetta, S., Coccia, G., Moramarco, T., & Todini, E. (2018). Real-time flood forecasting downstream river confluences using a Bayesian approach. *Journal of Hydrology*, *565*, 516–523.
- Barceló, M. A., Povedano, M., Vázquez-Costa, J. F., Franquet, A., Solans, M., & Saez, M. (2021). Estimation of the prevalence and incidence of motor neuron diseases in two Spanish regions: Catalonia and Valencia. *Scientific Reports*, *11*, 6207.
- Barceló, M. A., Saez, M., & Saurina, C. (2009). Spatial variability in mortality inequalities, socioeconomic deprivation and air pollution in small areas of the Barcelona Metropolitan Region, Spain. *Science of the Total Environment*, *407*, 5501-5523.
- Barceló, M. A., Varga, D., Tobias, A., Díaz, J., Linares, C., & Saez, M. (2016). Long term effects of traffic noise on mortality in the city of Barcelona, 2004-2007. *Environmental Research*, *147*, 193-206.
- Bednarik, M., Yilmaz, I., & Marschalko, M. (2012). Landslide hazard and risk assessment: a case study from the Hlohovec–Sered’landslide area in south-west Slovakia. *Natural hazards*, *64*, 547-575.
- Bennett, J. E., Tamura-Wicks, H., Parks, R. M., Burnett, R. T., Pope III, C. A., Bechle, M. J., ... & Ezzati, M. (2019). Particulate matter air pollution and national and county life expectancy loss in the USA: A spatiotemporal analysis. *PLoS medicine*, *16*(7), e1002856.
- Berkolaiko, G. and P. Kuchment (2013). *Introduction to quantum graphs, Volume 186 of Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192-225.

- Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43, 1-20.
- Best, N. G., Ickstadt, K., & Wolpert, R. L. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American statistical association*, 95(452), 1076-1088.
- Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., ... & Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504-507.
- Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., ... & Gething, P. W. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526(7572), 207-211.
- Bi, R., Jiao, Y., & Browder, J. A. (2021). Climate driven spatiotemporal variations in seabird bycatch hotspots and implications for seabird bycatch mitigation. *Scientific Reports*, 11(1), 20704.
- Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., ... & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 173, 144-154.
- Bivand, R., Gómez-Rubio, V., & Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *Journal of statistical software*, 63, 1-31.
- Bivand, R., Rundel, C., Pebesma, E., Stuetz, R., Hufthammer, K. O., & Bivand, M. R. (2017). Package 'rgeos'. *The Comprehensive R Archive Network (CRAN)*.
- Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, Ltd.
- Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, 4, 33-49.
- Blangiardo, M., Finazzi, F., & Cameletti, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and spatio-temporal epidemiology*, 18, 1-12.
- Blaustein, C., Hertz, N., & Vettor, R. (2022). A comparative spatiotemporal analysis of crime patterns in European and American urban areas. *European Journal of Criminology*, 19(2), 178-197.
- Bolin, D. (2014). Spatial Mat'ern fields driven by non-Gaussian noise. *Scand. J. Stat.* 41(3), 557-579.
- Bolin, D., & Wallin, J. (2020). Multivariate type G Matern stochastic partial differential equation random fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 82(1), 215-239.
- Bolin, D., Kirchner, K., & Kovacs, M. (2020). Numerical solution of fractional elliptic stochastic PDEs with spatial white noise. *IMA Journal of Numerical Analysis*, 40(2), 1051-1073.
- Bolin, D., Simas, A. B., & Wallin, J. (2022). Gaussian Whittle- Matérn fields on metric graphs. *arXiv preprint arXiv:2205.06163*.
- Bolle, A., das Neves, L., Smets, S., Mollaert, J., & Buitrago, S. (2018). An impact-oriented early warning and Bayesian-based decision support system for flood risks in Zeebrugge harbour. *Coastal Engineering*, 134, 191-202.

- Boqué, P., Saez, M., & Serra, L. (2022). Need to go further: Using INLA to discover limits and chances of burglaries' spatiotemporal prediction in heterogeneous environments. *Crime Science, 11*, 7.
- Boqué, P., Serra, L., & Saez, M. (2020). 'Surfing' burglaries with forced entry in Catalonia. *European Journal of Criminology*. Advance online publication.
- Borrell, C., Mari-Dell'olmo, M., Palència, L., ... & Díez, E. (2014). Socioeconomic inequalities in mortality in 16 European cities. *Scandinavian Journal of Public Health, 42*, 245-254.
- Borrell, C., Mari-Dell'olmo, M., Serral, G., Martínez-Beneito, M., Gotsens, M., Barceló, M. A., Saez, M., & other MEDEA Members. (2010). Inequalities in mortality in small areas of eleven Spanish cities (the multicenter MEDEA project). *Health & Place, 16*, 703-711.
- Boulieri, A., Liverani, S., Hoogh, K., & Blangiardo, M. (2017). A space-time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society Series A: Statistics in Society, 180*(1), 119-139.
- Bovea, M. D., Ibáñez-Forés, V., Juan, P., Pérez-Belis, V., & Braulio-Gonzalo, M. (2018b). Variables that affect the environmental performance of small electrical and electronic equipment. Methodology and case study. *Journal of Cleaner Production, 203*, 1067-1084.
- Bovea, M. D., Ibanez-Fores, V., Perez-Belis, V., & Juan, P. (2018a). A survey on consumers' attitude towards storing and end of life strategies of small information and communication technology devices in Spain. *Waste management, 71*, 589-602.
- Braulio-Gonzalo, M., Bovea, M. D., Jorge-Ortiz, A., & Juan, P. (2021a). Contribution of households' occupant profile in predictions of energy consumption in residential buildings: A statistical approach from Mediterranean survey data. *Energy and Buildings, 241*, 110939.
- Braulio-Gonzalo, M., Bovea, M. D., Jorge-Ortiz, A., & Juan, P. (2021b). Which is the best-fit response variable for modelling the energy consumption of households? An analysis based on survey data. *Energy, 231*, 120835.
- Braulio-Gonzalo, M., Bovea, M. D., Ruá, M. J., & Juan, P. (2016). A methodology for predicting the energy performance and indoor thermal comfort of residential stocks on the neighbourhood and city scales. A case study in Spain. *Journal of Cleaner Production, 139*, 646-665.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association, 88*(421), 9-25.
- Bresson, G., Lacroix, G., & Rahman, M. A. (2021). Bayesian panel quantile regression for binary outcomes with correlated random effects: an application on crime recidivism in Canada. *Empirical Economics, 60*, 227-259.
- Briz-Redón, Á., & Serrano-Aroca, Á. (2020). A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Science of the total environment, 728*, 138811.
- Brooks, Steve, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Burrough, P. A. (1986). Principles of geographical. *Information systems for land resource assessment*. Clarendon Press, Oxford.

- Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principles of geographical information systems*. Oxford university press.
- Calkin, D. E., & Mentis, M. (2015). Opinion: The use of natural hazard modeling for decision making under uncertainty. *Forest Ecosystems*, 2(1).
- Cantillo, V., Garcés, P., & Márquez, L. (2016). Factors influencing the occurrence of traffic accidents in urban roads: A combined GIS-Empirical Bayesian approach. *Dyna*, 83(195), 21-28.
- Carr, T., Mkuhlani, S., Segnon, A. C., Ali, Z., Zougmore, R., Dangour, A. D., ... & Scheelbeek, P. F. (2022). Climate change impacts and adaptation strategies for crops in West Africa: a systematic review. *Environmental Research Letters*.
- Castro, M., Paleti, R., & Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation research part B: methodological*, 46(1), 253-272.
- Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J., & Boëlle, P. (2004). A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in medicine*, 23(22), 3469-3487.
- Ceccato, V., Kahn, T., Herrmann, C., & Östlund, A. (2022). Pandemic restrictions and spatiotemporal crime patterns in New York, São Paulo, and Stockholm. *Journal of Contemporary Criminal Justice*, 38(1), 120-149.
- Cendoya, M., Hubel, A., Conesa, D., & Vicent, A. (2022). Modeling the spatial distribution of *Xylella fastidiosa*: A nonstationary approach with dispersal barriers. *Phytopathology*, 112(5), 1036–1045.
- Chaudhuri, S., Giménez-Adsuar, G., Saez, M., & Barceló, M. A. (2022a). PandemonCAT: Monitoring the COVID-19 Pandemic in Catalonia, Spain. *International Journal of Environmental Research and Public Health*, 19(8), 4783.
- Chaudhuri, S., Juan, P., & Mateu, J. (2022b). Spatio-temporal modeling of traffic accidents incidence on urban road networks based on an explicit network triangulation. *Journal of Applied Statistics*, 1– 22.
- Chaudhuri, S., Juan, P., & Serra, L. (2021). Analysis of precise climate pattern of Maldives. a complex island structure. *Regional Studies in Marine Science*, 44, 101789.
- Chaudhuri, S., Saez, M., Varga, D., & Juan, P. (2023). Spatiotemporal modeling of traffic risk mapping: A study of urban road networks in Barcelona, Spain. *Spatial Statistics*, 53, 100722.
- Coles, S. G., & Powell, E. A. (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review/Revue Internationale de Statistique*, 119-136.
- Comber, A., & Wulder, M. (2019). Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. *Transactions in GIS*, 23(5), 879-891.
- Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., ... & Worley, S. J. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137(654), 1-28.
- Contreras, C., & Hipp, J. R. (2020). Drugs, crime, space, and time: A spatiotemporal examination of drug activity and crime rates. *Justice Quarterly*, 37(2), 187-209.



- Cosandey-Godin, A., Krainski, E. T., Worm, B., & Flemming, J. M. (2014). Applying Bayesian spatio-temporal models to fisheries bycatch in the Canadian Arctic. *Canadian Journal of Fisheries and Aquatic Sciences*, *71*(9), 1322-1331.
- Cosandey-Godin, A., Krainski, E. T., Worm, B., & Flemming, J. M. (2014). Applying Bayesian spatio-temporal models to fisheries bycatch in the Canadian Arctic. *Canadian Journal of Fisheries and Aquatic Sciences*, *71*(9), 1322-1331.
- Costa, V., & Fernandes, W. (2017). Bayesian estimation of extreme flood quantiles using a rainfall-runoff model and a stochastic daily rainfall generator. *Journal of Hydrology*, *554*, 137–154.
- Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883-904.
- Crespo, R. D. J., Wu, J., Myer, M., Yee, S., & Fulford, R. (2019). Flood protection ecosystem services in the coast of Puerto Rico: Associations between extreme weather, flood hazard mitigation and gastrointestinal illness. *Science of the total environment*, *676*, 343-355.
- Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.
- Cressie, N., & Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical association*, *94*(448), 1330-1339.
- Cromley, E. K., & McLafferty, S. L. (2012). *GIS and Public Health* New York and London.
- Cronie, O., M. Moradi, and J. Mateu (2020). Inhomogeneous higher-order summary statistics for point processes on linear networks. *Statistics and Computing* *30*(5), 1221–1239.
- Cusimano, M., Marshall, S., Rinner, C., Jiang, D., & Chipman, M. (2010). Patterns of urban violent injury: a spatio-temporal analysis. *PLoS One*, *5*(1), e8669.
- Cutter, S. L., & Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the National Academy of Sciences*, *105*(7), 2301–2306.
- Davis, R. C., Mateu-Gelabert, P., & Miller, J. (2005). Can effective policing also be respectful? Two examples in the South Bronx. *Police Quarterly*, *8*(2), 229-247.
- Dawkins, L. C., Williamson, D. B., Mengersen, K. L., Morawska, L., Jayaratne, R., & Shaddick, G. (2021). Where Is the Clean Air? A Bayesian Decision Framework for Personalised Cyclist Route Selection Using R-INLA. *Bayesian Analysis*, *16*(1).
- de Figueiredo, L. P., Grana, D., Roisenberg, M., & Rodrigues, B. B. (2019). Multimodal Markov chain Monte Carlo method for nonlinear petrophysical seismic inversion Multimodal MCMC inversion. *Geophysics*, *84*(5), M1-M13.
- De Gruijter, J., Brus, D. J., Bierkens, M. F., & Knotters, M. (2006). *Sampling for natural resource monitoring* (Vol. 665). Berlin: Springer.
- Di Baldassarre, G., Castellarin, A., Montanari, A., & Brath, A. (2009). Probability-weighted hazard maps for comparing different flood risk management strategies: a case study. *Natural Hazards*, *50*, 479-496.
- Díaz-Avalos, C., & Juan, P. (2022). Modeling the spatial evolution wildfires using random spread process. *Environmetrics*, e2774.

- Díaz-Avalos, C., Juan, P., Chaudhuri, S., Sáez, M., & Serra, L. (2020). Association between the new COVID-19 cases and air pollution with meteorological elements in nine counties of New York state. *International Journal of Environmental Research and Public Health*, *17*(23), 9055.
- Diggle, P. J., Moraga, P., Rowlingson, B., & Taylor, B. M. (2013). Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. *Statistical Science*, *28*(4), 542-563.
- Dimakopoulou, K., Samoli, E., Analitis, A., Schwartz, J., Beevers, S., Kitwiroon, N., ... & Katsouyanni, K. (2022). Development and Evaluation of Spatio-Temporal Air Pollution Exposure Models and Their Combinations in the Greater London Area, UK. *International Journal of Environmental Research and Public Health*, *19*(9), 5401.
- Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., & Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. *Advances in water resources*, *94*, 87-102.
- Duan, P., Mao, G., Liang, W., & Zhang, D. (2018). A unified spatio-temporal model for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, *20*(9), 3212-3223.
- Dunn, D. C., Boustany, A. M., & Halpin, P. N. (2011). Spatio-temporal management of fisheries to reduce by-catch and increase fishing selectivity. *Fish and Fisheries*, *12*(1), 110-119.
- Eboli, L., Forciniti, C., & Mazzulla, G. (2020). Factors influencing accident severity: an analysis by road accident type. *Transportation research procedia*, *47*, 449-456.
- Elith, J., Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677-697.
- Emmer, A. (2018). Geographies and scientometrics of research on natural hazards. *Geosciences*, *8*(10), 382.
- Etcheberria, J., Goicoa, T., Lopez-Abente, G., Riebler, A., & Ugarte, M. D. (2017). Spatial gender-age-period-cohort analysis of pancreatic cancer mortality in Spain (1990–2013). *PloS one*, *12*(2), e0169751.
- Field, E. H., Milner, K. R., Hardebeck, J. L., Page, M. T., van der Elst, N., Jordan, T. H., ... & Werner, M. J. (2017). A spatiotemporal clustering model for the third Uniform California Earthquake Rupture Forecast (UCERF3-ETAS): Toward an operational earthquake forecast. *Bulletin of the Seismological Society of America*, *107*(3), 1049-1081.
- Finley, A. O., Sang, H., Banerjee, S., & Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, *53*(8), 2873-2884.
- Finney Rutten, L. J., Wilson, P. M., Jacobson, D. J., Agunwamba, A. A., Radecki Breitkopf, C., Jacobson, R. M., & St. Sauver, J. L. (2017). A population-based study of sociodemographic and geographic variation in HPV vaccination. *Cancer Epidemiology, Biomarkers & Prevention*, *26*(4), 533-540.
- Franch-Pardo, I., Desjardins, M. R., Barea-Navarro, I., & Cerdà, A. (2021). A review of GIS methodologies to analyze the dynamics of COVID-19 in the second half of 2020. *Transactions in GIS*, *25*(5), 2191-2239.
- Franci, F., Bitelli, G., Mandanici, E., Hadjimitsis, D., & Agapiou, A. (2016). Satellite remote sensing and GIS-based multi-criteria analysis for flood hazard mapping. *Natural Hazards*, *83*, 31-51.

- Frazier, A. E., Bagchi-Sen, S., & Knight, J. (2013). The spatio-temporal impacts of demolition land use policy and crime in a shrinking city. *Applied Geography*, *41*, 55-64.
- Fuglstad, G. A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, *114*(525), 445-452.
- Fuglstad, G.-A. and Castruccio, S. (2020) Compression of climate simulations with a nonstationary global spatiotemporal spde model. *Ann. Appl. Stat.*, *14*, 542–559.
- Gakidou, E., Afshin, A., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., ... & Duncan, S. (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, *390*(10100), 1345-1422.
- Galgamuwa, U., Du, J., & Dissanayake, S. (2021). Bayesian spatial modeling to incorporate unmeasured information at road segment levels with the INLA approach: A methodological advancement of estimating crash modification factors. *Journal of traffic and transportation engineering (English edition)*, *8*(1), 95-106.
- Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M., & Stephenson, J. (2009). Markov chain Monte Carlo (MCMC) sampling methods to determine optimal models, model resolution and model choice for Earth Science problems. *Marine and Petroleum Geology*, *26*(4), 525-535.
- Gaume, E., Gaál, L., Viglione, A., Szolgay, J., Kohnová, S., & Blöschl, G. (2010). Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of hydrology*, *394*(1-2), 101-117.
- Gelfand, A. E. (2012). Hierarchical modeling for spatial data problems. *Spatial statistics*, *1*, 30-39.
- Gelfand, A. E., Smith, A. F., & Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523-532.
- Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P. (Eds.), 2010. *Handbook of Spatial Statistics*. Chapman & Hall.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* [OCLC: 909477393].
- Geweke, John. 1992. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. New York: In J. Bernardo, J. Berger, A. Dawid, A. Smith (Eds.), *Bayesian Statistics 4*. Oxford University Press.
- Giardini, D., Danciu, L., Erdik, M., Şeşetyan, K., Demircioğlu Tümsa, M. B., Akkar, S., ... & Zare, M. (2018). Seismic hazard map of the Middle East. *Bulletin of Earthquake Engineering*, *16*, 3567-3570.
- Gielen, E., Riutort-Mayol, G., Palencia-Jiménez, J. S., & Cantarino, I. (2018). An urban sprawl index based on multivariate and Bayesian factor analysis with application at the municipality level in Valencia. *Environment and Planning B: Urban Analytics and City Science*, *45*(5), 888-914.
- Gilks, W. R., & Roberts, G. O. (1996). Strategies for improving MCMC. *Markov chain Monte Carlo in practice*, *6*, 89-114.

- Gilks, W. R., W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Boca Raton, Florida: Chapman & Hall.
- Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote sensing of Environment*, 80(1), 76-87.
- Glahn, H. R., & Lowry, D. A. (1972). The use of Model Output Statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, 11(8), 1203-1211.
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458), 590-600.
- Gneiting, T., Genton, M. G., & Guttorp, P. (2006). Geostatistical space-time models, stationarity, separability, and full symmetry. *Monographs On Statistics and Applied Probability*, 107, 151.
- Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., ... & Hay, S. I. (2017). Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *The Lancet*, 390(10108), 2171-2182.
- Gomez-Rubio, V. (2020). *Bayesian Inference with INLA*. Taylor & Francis Group.
- Gómez-Rubio, V., & Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28, 1033-1051.
- Goodchild, M. F. (1991). Geographic information systems. *Progress in Human geography*, 15(2), 194-200.
- Gortázar, C., Fernández-Calle, L. M., Collazos-Martínez, J. A., Mínguez-González, O., & Acevedo, P. (2017). Animal tuberculosis maintenance at low abundance of suitable wildlife reservoir hosts: A case study in northern Spain. *Preventive veterinary medicine*, 146, 150-157.
- Gotsens, M., Mari-Dell'olmo, M., Pérez, K., ... & Borrell, C. (2013). Socioeconomic inequalities in injury mortality in small areas of 15 European cities: A multicenter IneqCities project using spatial analysis. *Health & Place*, 24, 165-172.
- Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458), 632-648.
- Grauer, J., Löwen, H., & Liebchen, B. (2020). Strategic spatiotemporal vaccine distribution increases the survival rate in an infectious disease like Covid-19. *Scientific reports*, 10(1), 21594.
- Grezio, A., Marzocchi, W., Sandri, L., & Gasparini, P. (2009). A bayesian procedure for probabilistic tsunami hazard assessment. *Natural Hazards*, 53(1), 159–174.
- Gross, B., Zheng, Z., Liu, S., Chen, X., Sela, A., Li, J., ... & Havlin, S. (2020). Spatio-temporal propagation of COVID-19 pandemics. *Europhysics Letters*, 131(5), 58003.
- Grüss, A., Thorson, J. T., Babcock, E. A., & Tarnecki, J. H. (2018). Producing distribution maps for informing ecosystem-based fisheries management using a comprehensive survey database and spatio-temporal models. *ICES Journal of Marine Science*, 75(1), 158-177.
- Güneralp, B., Güneralp, İ., & Liu, Y. (2015). Changing global patterns of urban exposure to flood and drought hazards. *Global environmental change*, 31, 217-225.

- Guo, Y., Wu, W., Du, M., Liu, X., Wang, J., & Bryant, C. R. (2019). Modeling climate change impacts on rice growth and yield under global warming of 1.5 and 2.0 C in the Pearl River Delta, China. *Atmosphere*, 10(10), 567.
- Ha, H., Bui, Q. D., Nguyen, H. D., Pham, B. T., Lai, T. D., & Luu, C. (2023). A practical approach to flood hazard, vulnerability, and risk assessing and mapping for Quang Binh province, Vietnam. *Environment, Development and Sustainability*, 25(2), 1101-1130.
- Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis & Prevention*, 42(2), 676-688.
- Hagemeyer-Klose, M., & Wagner, K. (2009). Evaluation of flood hazard maps in print and web mapping services as information tools in flood risk communication. *Natural hazards and earth system sciences*, 9(2), 563-574.
- Hamra, G., MacLehose, R., & Richardson, D. (2013). Markov chain Monte Carlo: an introduction for epidemiologists. *International journal of epidemiology*, 42(2), 627-634.
- Han, S. Y., Tsou, M. H., Knaap, E., Rey, S., & Cao, G. (2019). How do cities flow in an emergency? Tracing human mobility patterns during a natural disaster with big data and geospatial data science. *Urban Science*, 3(2), 51.
- Han, S., & Coulibaly, P. (2017). Bayesian flood forecasting methods: A review. *Journal of Hydrology*, 551, 340–351.
- Handcock, M. S., & Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, 89(426), 368-378.
- Hanson, T., & Branscum, A. (2006). Bayesian Semiparametric Methods for Joint Modeling Longitudinal and Survival Data.
- Harries, K. (2006). Extreme spatial variations in crime density in Baltimore County, MD. *Geoforum*, 37(3), 404–416.
- Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1), 81-102.
- Harvill, J. L. (2010). Spatio-temporal processes. *Wiley interdisciplinary reviews: computational statistics*, 2(3), 375-382.
- Hayashi, S., Narita, Y., & Koshimura, S. (2013). Developing tsunami fragility curves from the surveyed data and numerical modeling of the 2011 Tohoku earthquake tsunami. *J. Jpn. Soc. Civ. Eng. Coast. Eng*, 69, 1–5.
- HDX. (2021). *Maldives - subnational administrative boundaries*. Retrieved January 22, 2021, from <https://data.humdata.org/dataset/cod-ab-mdv>
- HDX. (2022). *Maldives disaster records*. Retrieved January 12, 2022, from <https://data.humdata.org/dataset/509cd879-f937-4428-8868-5459938744d3>
- Held, L., & Rue, H. (2010). Conditional and intrinsic autoregressions. *Handbook of spatial statistics*, 201-216.

- Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validators predictive checks: a comparison of MCMC and INLA. *Statistical modelling and regression structures: Festschrift in honour of ludwig Fahrmeir*, 91-110.
- Hildeman, A., D. Bolin, and I. Rychlik (2021). Deformed SPDE models with an application to spatial modeling of significant wave height. *Spat. Stat.* 42, 100449.
- Hoffmann, R., Borsboom, G., Saez, M., ...& Borrell, C. (2014). Social differences in avoidable mortality between small areas of 15 European cities: An ecological study. *International Journal of Health Geographics*, 13, 8.
- Hossain, S., Abtahee, A., Kashem, I., Hoque, M. M., & Sarker, I. H. (2020). Crime prediction using spatio-temporal data. In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1* (pp. 277-289). Springer Singapore.
- Hsieh, Y. H., & Ma, S. (2009). Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. *The American journal of tropical medicine and hygiene*, 80(1), 66-71.
- Hu, Y., Wang, F., Guin, C., & Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied geography*, 99, 89-97.
- Huang, J., Malone, B. P., Minasny, B., McBratney, A. B., & Triantafyllis, J. (2017). Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. *Science of the Total Environment*, 609, 621-632.
- Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS international journal of geo-information*, 4(3), 1549-1568.
- Hubin, A., & Storvik, G. (2018). Mode jumping MCMC for Bayesian variable selection in GLMM. *Computational Statistics & Data Analysis*, 127, 281-297.
- Intergovernmental Panel on Climate Change. (2021). *Climate change 2021: The physical science basis*. Retrieved August 12, 2021, from <https://www.ipcc.ch/report/ar6/wg1/>
- Irl, S. D., Harter, D. E., Steinbauer, M. J., Gallego Puyol, D., Fernández-Palacios, J. M., Jentsch, A., & Beierkuhnlein, C. (2015). Climate vs. topography—spatial patterns of plant species diversity and endemism on a high-elevation island. *Journal of Ecology*, 103(6), 1621-1633.
- Irvin-Erickson, Y., & La Vigne, N. (2015). A spatio-temporal analysis of crime at Washington, DC metro rail: Stations' crime-generating and crime-attracting characteristics as transportation nodes and places. *Crime science*, 4, 1-13.
- Isles, The President's Office. (2022). *Maldives facts*. Retrieved February 5, 2022, from <https://isles.gov.mv/Home/en>
- Jousimo, J., Tack, A. J., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., & Laine, A. L. (2014). Ecological and evolutionary effects of fragmentation on infectious disease dynamics. *Science*, 344(6189), 1289-1293.
- Juan Verdoy, P. (2021). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. Choosing the best mesh for each database. *Communications in Statistics-Simulation and Computation*, 50(10), 2990-3030.

- Juan, P., Braulio-Gonzalo, M., Díaz-Ávalos, C., Bovea, M. D., & Serra, L. (2022). Bayesian and network models with covariate effects for predicting heating energy demand. *Spatial and Spatio-temporal Epidemiology*, *43*, 100547.
- Juan, P., Díaz-Avalos, C., Mejía-Domínguez, N. R., & Mateu, J. (2017). Hierarchical spatial modeling of the presence of Chagas disease insect vectors in Argentina. A comparative approach. *Stochastic Environmental Research and Risk Assessment*, *31*, 461-479.
- Juan, P., Mateu, J., & Saez, M. (2012). Pinpointing spatio-temporal interactions in wildfire patterns. *Stochastic Environmental Research and Risk Assessment*, *26*, 1131-1150.
- Karaganis, A., & Mimis, A. (2006). A spatial point process for estimating the probability of occurrence of a traffic accident. *European Regional Science Association, ERSA conference papers*.
- Karimiziarani, M., Jafarzadegan, K., Abbaszadeh, P., Shao, W., & Moradkhani, H. (2022). Hazard risk awareness and disaster management: Extracting the information content of twitter data. *Sustainable Cities and Society*, *77*, 103577.
- Kaygisiz, Ö., Düzgün, Ş., Yildiz, A., & Senbil, M. (2015). Spatio-temporal accident analysis for accident prevention in relation to behavioral factors in driving: The case of South Anatolian Motorway. *Transportation research part F: traffic psychology and behaviour*, *33*, 128-140.
- Kelsall, J., & Wakefield, J. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, *97*(459), 692-701.
- Kendall, M. G. 1948. *Rank correlation methods*. Oxford, England: Griffin.
- Khattak, M. W., Pirdavani, A., De Winne, P., Brijs, T., & De Backer, H. (2021). Estimation of safety performance functions for urban intersections using various functional forms of the negative binomial regression model and a generalized Poisson regression model. *Accident Analysis & Prevention*, *151*, 105964.
- Kiremidjian, A. S., & Shah, H. C. (1997). *Seismic Hazard Mapping of California*.
- Knighton, J., & Bastidas, L. A. (2015). A proposed probabilistic seismic tsunami hazard analysis methodology. *Natural Hazards*, *78*(1), 699-723.
- Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, *19*(17-18), 2555-2567.
- Kraemer, M. U., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J. T., ... & Pybus, O. G. (2021). Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B. 1.1. 7 emergence. *Science*, *373*(6557), 889-895.
- Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., ... & Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC press.
- Kumari, N., Srivastava, A., & Dumka, U. C. (2021). A long-term spatiotemporal analysis of vegetation greenness over the Himalayan Region using Google Earth Engine. *Climate*, *9*(7), 109.
- Lateltin, O., Haemmig, C., Raetzo, H., & Bonnard, C. (2005). Landslide risk management in Switzerland. *Landslides*, *2*(4), 313-320.

- Leong, K., & Sung, A. (2015). A review of spatio-temporal pattern analysis approaches on crime analysis. *International E-Journal of Criminal Sciences*, (9).
- Lertxundi-Manterola, A., & Saez, M. (2009). Modelling of dioxide nitrogen (NO<sub>2</sub>) and fine particulate matter (PM<sub>10</sub>) air pollution in the metropolitan areas of Barcelona and Bilbao, Spain. *Environmetrics*, 20(5), 477-493.
- Lezama-Ochoa, N., Pennino, M. G., Hall, M. A., Lopez, J., & Murua, H. (2020). Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of the Spinetail Devil Ray (Mobular mobular). *Scientific reports*, 10(1), 18822.
- Li, S., Ye, X., Lee, J., Gong, J., & Qin, C. (2017). Spatiotemporal analysis of housing prices in China: A big data perspective. *Applied Spatial Analysis and Policy*, 10(3), 421-433.
- Li, Y., Brown, P., Gesink, D. C., & Rue, H. (2012). Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical methods in medical research*, 21(5), 479-507.
- Li, Z., Liu, L., Wang, J., Lin, L., Dong, J., & Dong, Z. (2023). Design and Analysis of an Effective Multi-Barrier Model Based on Non-Stationary Gaussian Random Fields. *Electronics*, 12(2), 345.
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.
- Lin, Q., Wang, Y., Glade, T., Zhang, J., & Zhang, Y. (2020). Assessing the spatiotemporal impact of climate change on event rainfall characteristics influencing landslide occurrences based on multiple GCM projections in China. *Climatic Change*, 162, 761-779.
- Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of statistical software*, 63, 1-25.
- Lindgren, F., Bolin, D., & Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*, 50, 100599.
- Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4), 423-498.
- Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., & Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and ecological statistics*, 21, 411-433.
- Liu, C., & Sharma, A. (2017). Exploring spatio-temporal effects in traffic crash trend analysis. *Analytic methods in accident research*, 16, 104-116.
- Liu, C., & Sharma, A. (2018). Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. *Analytic methods in accident research*, 17, 14-31.
- Liu, J., Hainen, A., Li, X., Nie, Q., & Nambisan, S. (2019). Pedestrian injury severity in motor vehicle crashes: an integrated spatio-temporal modeling approach. *Accident Analysis & Prevention*, 132, 105272.
- Liu, M., & Stein, S. (2016). Mid-continental earthquakes: Spatiotemporal occurrences, causes, and hazards. *Earth-Science Reviews*, 162, 364-386.



- Lobell, D. B., Bänziger, M., Magorokosho, C., & Vivek, B. (2011). Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nature climate change*, *1*(1), 42-45.
- Lombardo, L., Opitz, T., Ardizzone, F., Guzzetti, F., & Huser, R. (2020). Space-time landslide predictive modelling. *Earth-science reviews*, *209*, 103318.
- Loo, B. P. Y., Yao, S., & Wu, J. (2011). Spatial point analysis of road crashes in Shanghai: A GIS-based network kernel density method. *2011 19th International Conference on Geoinformatics*.
- López-Abente, G., Núñez, O., Fernández-Navarro, P., Barros-Dios, J. M., Martín-Méndez, I., Bel-Lan, A., ... & Ruano-Ravina, A. (2018). Residential radon and cancer mortality in Galicia, Spain. *Science of the Total Environment*, *610*, 1125-1132.
- Lourenço, P., Medeiros, V., Gil, A., & Silva, L. (2011). Distribution, habitat and biomass of *Pittosporum undulatum*, the most important woody plant invader in the Azores Archipelago. *Forest Ecology and Management*, *262*(2), 178-187.
- Mahata, D., Narzary, P. K., & Govil, D. (2019). Spatio-temporal analysis of road traffic accidents in Indian large cities. *Clinical Epidemiology and Global Health*, *7*(4), 586-591.
- Malve, O., Laine, M., Haario, H., Kirkkala, T., & Sarvala, J. (2007). Bayesian modelling of algal mass occurrences—using adaptive MCMC methods with a lake water quality model. *Environmental Modelling & Software*, *22*(7), 966-977.
- Mann, H. B. 1945. Nonparametric tests against trend. *Econometrica* *13* (3):245–59.
- Marí-Dell'olmo, M., Gotsens, M., Palencia, L., & Borrell, C. (2015). Socioeconomic inequalities in cause-specific mortality in fifteen European cities. *Journal of Epidemiology and Community Health*, *69*(5), 432-441.
- Martinez-Beneito, M. A., Botella-Rocamora, P., & Banerjee, S. (2017). Towards a multidimensional approach to Bayesian disease mapping. *Bayesian analysis*, *12*(1), 239.
- Martínez-Minaya, J., Conesa, D., Bakka, H., & Pennino, M. G. (2019). Dealing with physical barriers in bottlenose dolphin (*Tursiops truncatus*) distribution. *Ecological Modelling*, *406*, 44-49.
- Martino, S., Akerkar, R., & Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, *38*(3), 514-528.
- Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, *67*, 68–83.
- Mata, F., Torres-Ruiz, M., Guzmán, G., ... & Loza, E. (2016). A mobile information system based on crowd-sensed and official crime data for finding safe routes: a case study of Mexico City. *Mobile Information Systems*, 2016.
- Matérn, B. (1960). *Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations* (Doctoral dissertation, Stockholm University).
- Maynou, L., & Saez, M. (2016). Economic crisis and health inequalities: evidence from the European Union. *International Journal for Equity in Health*, *15*, 135.

- Maynou, L., Saez, M., Bacaria, J., & López-Casasnovas, G. (2015). Health inequalities in the European Union: An empirical analysis of the dynamics of regional differences. *European Journal of Health Economics*, 16(5), 543-559.
- Maynou-Pujolràs, L., Saez, M., & López-Casasnovas, G. (2016b). Has the economic crisis widened the intra-urban socioeconomic inequalities in mortality? The case of Barcelona, Spain. *Journal of Epidemiology and Community Health*, 70(2), 114-124.
- Maynou-Pujolràs, L., Saez, M., Kyriacou, A., & Bacaria, J. (2016a). The impact of structural and cohesion funds on Eurozone convergence (1990-2010). *Regional Studies*, 50(7), 1127-1139.
- McKenzie, L. J., Yoshida, R. L., Aini, J. W., Andréfouet, S., Colin, P. L., Cullen-Unsworth, L. C., ... & Unsworth, R. K. (2021). Seagrass ecosystems of the Pacific Island Countries and Territories: A global bright spot. *Marine Pollution Bulletin*, 167, 112308.
- Mejia, A. F., Yue, Y., Bolin, D., Lindgren, F., & Lindquist, M. A. (2020). A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*, 115(530), 501-520.
- Meseguer Costa, S., Juan, P., Vicente, A. B., & Díaz Ávalos, C. (2016). A New Bayesian Inference Methodology for Modeling Geochemical Elements in Soil with Covariates. Characterization of Lithium in South Iberian Range (Spain).
- Miaou, S. P., Hu, P. S., Wright, T., Rathi, A. K., & Davis, S. C. (1992). Relationship between truck accidents and highway geometric design: a Poisson regression approach. *Transportation Research Record*, (1376).
- Miller, D. L., & Wood, S. N. (2014). Finite area smoothing with generalized distance splines. *Environmental and Ecological Statistics*, 21(4), 715-731.
- Mizuta, R., Kamae, Y., Yoshimura, K., Matsueda, M., Endo, H., Ose, T., ... & Kitoh, A. (2018). Future changes in precipitation over East Asia projected by a global atmospheric model with a 60-km grid. *SOLA*, 14, 40-44.
- Moraga, P. (2019). *Geospatial health data : Modeling and visualization with R-INLA and Shiny*. CRC Press.
- Moraga, P., Cramb, S. M., Mengersen, K. L., & Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21, 27-41.
- Morjani, Z. E. A. E., Ebener, S., Boos, J., Ghaffar, E. A., & Musani, A. (2007). Modelling the spatial distribution of five natural hazards in the context of the WHO/EMRO atlas of disaster risk as a step towards the reduction of the health impact related to disasters. *International Journal of Health Geographics*, 6(1).
- Morris, M. C., Marco, M., Maguire-Jack, K., Kouros, C. D., Im, W., White, C., ... & Garber, J. (2019). County-level socioeconomic and crime risk factors for substantiated child abuse and neglect. *Child Abuse & Neglect*, 90, 127-138.
- Mota-Bertran, A., Saez, M., & Coenders, G. (2021). Compositional and Bayesian inference analysis of the concentrations of air pollutants in Catalonia, Spain. *Environmental Research*, 204(Pt D), 112388.

- Mugglin, A. S., Carlin, B. P., & Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, *95*, 877-887.
- Muhammad, R., Zhang, W., Abbas, Z., Guo, F., & Gwiazdzinski, L. (2022). Spatiotemporal change analysis and prediction of future land use and land cover changes using QGIS MOLUSCE plugin and remote sensing big data: a case study of Linyi, China. *Land*, *11*(3), 419.
- Mustafa, A., Ebaid, A., Omrani, H., & McPhearson, T. (2021). A multi-objective Markov Chain Monte Carlo cellular automata model: Simulating multi-density urban expansion in NYC. *Computers, Environment and Urban Systems*, *87*, 101602.
- Myer, M. H., & Johnston, J. M. (2019). Spatiotemporal Bayesian modeling of West Nile virus: Identifying risk of infection in mosquitoes with local-scale predictors. *Science of the Total Environment*, *650*, 2818-2829.
- Nahayo, L., Mupenzi, C., Habiyaremye, G., Kalisa, E., Udahogora, M., Nzabarinda, V., & Li, L. (2019). Landslides hazard mapping in Rwanda using bivariate statistical index method. *Environmental Engineering Science*, *36*(8), 892-902.
- National Climate Assessment. (2014). Third National Climate Assessment. US Global Change Research Program. Accessed from [www.nca2014.globalchange.gov/](http://www.nca2014.globalchange.gov/)
- Niekerk, J. V., Bakka, H., & Rue, H. (2021). Competing risks joint models using R-INLA. *Statistical Modelling*, *21*(1-2), 56-71.
- Niekerk, J. V., Krainski, E. T., Rustand, D., & Rue, H. (2022). A new avenue for Bayesian inference with INLA.
- Niraula, P., Mateu, J., & Chaudhuri, S. (2022). A Bayesian machine learning approach for spatio-temporal prediction of COVID-19 cases. *Stochastic Environmental Research and Risk Assessment*, 1-19.
- Niu, K., Zhang, H., Zhou, T., Cheng, C., & Wang, C. (2019). A novel spatio-temporal model for city-scale traffic speed prediction. *IEEE Access*, *7*, 30050-30057.
- Nouck, P. N., Kenfack, C., Diab, A. D., Njeudjang, K., Meli, L. J., & Kamseu, R. (2013). A geostatistical re-interpretation of gravity surveys in the Yagoua, Cameroon region. *Geofisica internacional*, *52*(4), 365-373.
- Okabe, A. and K. Sugihara (2012). *Spatial analysis along networks: statistical and computational methods*. John Wiley & Sons.
- OpenDataBCN, 2021. Open data BCN — Ajuntament de Barcelona's open data service. <https://www.opendata-ajuntament.barcelona.cat/en>.
- Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and planning A*, *16*(1), 17-31.
- Opitz, T., Bonneau, F., & Gabriel, E. (2020). Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France. *Spatial Statistics*, *40*, 100429.
- Osberghaus, D., & Fugger, C. (2022). Natural disasters and climate change beliefs: The role of distance and prior beliefs. *Global Environmental Change*, *74*, 102515.

- Paoletti, E., De Marco, A., Beddows, D. C., Harrison, R. M., & Manning, W. J. (2014). Ozone levels in European and USA cities are increasing more than at rural sites, while peak values are decreasing. *Environmental Pollution*, *192*, 295-299.
- Paradinas, I., Conesa, D., López-Quílez, A., & Bellido, J. M. (2017). Spatio-temporal model structures with shared components for semi-continuous species distribution modelling. *Spatial Statistics*, *22*, 434-450.
- Paradinas, I., Conesa, D., Pennino, M. G., Bellido, J. M., Sáenz-Arroyo, A., Cisneros-Montemayor, A. M., & Sumaila, U. R. (2015). Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. *Marine Ecology Progress Series*, *529*, 223-238.
- Paul, M., Riebler, A., Bachmann, L. M., Rue, H., & Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested Laplace approximations. *Statistics in medicine*, *29*(12), 1325-1339.
- Phillips, M. C. K., Cinderich, A. B., Burrell, J. L., Ruper, J. L., Will, R. G., & Sheridan, S. C. (2015). The effect of climate change on natural disasters: A college student perspective. *Weather, Climate, and Society*, *7*(1), 60–68.
- Pittore, M., Wieland, M., & Fleming, K. (2017). Perspectives on global dynamic exposure modelling for geo-risk assessment. *Natural Hazards*, *86*(1), 7–30.
- Plug, C., Xia, J. C., & Caulfield, C. (2011). Spatial and temporal visualisation techniques for crash analysis. *Accident Analysis & Prevention*, *43*(6), 1937-1946.
- Poompavai, V., & Ramalingam, M. (2013). Geospatial analysis for coastal risk assessment to cyclones. *Journal of the Indian Society of Remote Sensing*, *41*, 157-176.
- Popa, M. C., Peptenatu, D., Drăghici, C. C., & Diaconu, D. C. (2019). Flood hazard mapping using the flood and flash-flood potential index in the Buzău River catchment, Romania. *Water*, *11*(10), 2116.
- Povedano, M., Saez, M., Martínez-Matos, J. A., & Barceló, M. A. (2018). Spatial assessment of the association between long-term exposure to environmental factors and the occurrence of amyotrophic lateral sclerosis (ALS) in Catalonia, Spain. A population-based nested case-control study. *Neuroepidemiology*, *51*(1-2), 33-49.
- Prasannakumar, V., Vijith, H., Charutha, R., & Geetha, N. (2011). Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia-social and behavioral sciences*, *21*, 317-325.
- Puigpinós-Riera, R., Mari-Dell'olmo, M., Gotsens, M., ... & Sánchez-Villegas, P. (2011). Cancer mortality inequalities in urban areas: A Bayesian small area analysis in Spanish cities. *International Journal of Health Geographics*, *10*, 6.
- Quesada-Román, A., Ballesteros-Cánovas, J. A., Granados-Bolaños, S., Birkel, C., & Stoffel, M. (2022). Improving regional flood risk assessment using flood frequency and dendrogeomorphic analyses in mountain catchments impacted by tropical cyclones. *Geomorphology*, *396*, 108000.
- Quick, M., Law, J., & Li, G. (2019). Time-varying relationships between land use and crime: A spatio-temporal analysis of small-area seasonal property crime trends. *Environment and Planning B: Urban Analytics and City Science*, *46*(6), 1018-1035.

- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- R. INLA Project, 2020. <https://www.r-inla.org>.
- Raju, E., Boyd, E., & Otto, F. (2022). Stop blaming the climate for disasters. *Communications Earth & Environment*, 3(1).
- Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 307–319.
- Rasheed, S., Warder, S. C., Plancherel, Y., & Piggott, M. D. (2022). Nearshore tsunami amplitudes across the Maldives archipelago due to worst case seismic scenarios in the Indian ocean.
- Redlands, C. E. S. R. I. (2022). Arcgis pro: Version 3.0.1.
- Renart-Vicens, G., Saez, M., Moreno-Crespi, J., Serdà, B., & Marcos-Gragera, R. (2014). Incidence variation of prostate and cervical cancer according to socioeconomic level in the Girona Health Region. *BMC Public Health*, 14, 1079.
- Rezaldi, M. Y., Nugroho, B., Kushadiani, S. K., Prasetyadi, A., Riyanto, A. M., Hanifa, N. R., & Yoganingrum, A. (2021). A Systematical Review of the Tsunami Hazards Modeling. *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 1–6.
- Riebler, A., Held, L., & Rue, H. (2012). Estimation and extrapolation of time trends in registry data—borrowing strength from related populations. *The Annals of Applied Statistics*, 304-333.
- Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4), 1145-1165.
- Riley, K., Thompson, M., Webley, P., & Hyde, K. D. (2016). Uncertainty in natural hazards, modeling and decision support. In *Natural hazard uncertainty assessment* (pp. 1–8). John Wiley & Sons, Inc.
- Risi, R. D., & Goda, K. (2017). Simulation-based probabilistic tsunami hazard analysis: Empirical and robust hazard predictions. *Pure and Applied Geophysics*, 174(8), 3083–3106.
- Riyaz, M., & Suppasri, A. (2016). Geological and geomorphological tsunami hazard analysis for the Maldives using an integrated WE method and a LR model. *Journal of Earthquake and Tsunami*, 10(01), 1650003.
- Robert, C. P., Casella, G., & Casella, G. (1999). *Monte Carlo statistical methods* (Vol. 2). New York: Springer.
- Roos, M., & Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data.
- Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. CRC press.
- Rue, H., & Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference*, 137(10), 3177-3192.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2), 319–392.

- Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4, 395-421.
- Ruiz-Cárdenas, R., Krainski, E. T., & Rue, H. (2012). Direct fitting of dynamic models using integrated nested Laplace approximations—INLA. *Computational Statistics & Data Analysis*, 56(6), 1808-1828.
- Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied geography*, 86, 255-261.
- Saez, M., & Barceló, M. A. (2022). Spatial prediction of air pollution levels using a hierarchical Bayesian spatiotemporal model in Catalonia, Spain. *Environmental Modelling & Software*, 151, 105369.
- Saez, M., & López-Casasnovas, G. (2019). Assessing the effects on health inequalities of differential exposure and differential susceptibility of air pollution and environmental noise in Barcelona, 2007-2014. *International Journal of Environmental Research and Public Health*, 16(18), E3470.
- Saez, M., Barceló, M. A., Farrerons, M., & López-Casasnovas, G. (2018). The association between exposure to environmental factors and the occurrence of attention-deficit/hyperactivity disorder (ADHD). A population-based retrospective cohort study. *Environmental Research*, 166, 205-214.
- Saez, M., Barceló, M. A., Martos, C., Saurina, C., Marcos-Gragera, R., Renart, G., Ocaña-Riola, R., Feja, C., & Alcalá, T. (2012). Spatial variability in relative survival from female breast cancer. *Journal of the Royal Statistical Society*, 175(1), 107-134.
- Saez, M., Tobias, A., & Barceló, M. A. (2020). Effects of long-term exposure to air pollution on the spatial spread of COVID-19 in Catalonia, Spain. *Environmental Research*, 191, 110177.
- SafarianZengir, V., Sobhani, B., & Asghari, S. (2019). Modeling and Monitoring of Drought for forecasting it, to Reduce Natural hazards Atmosphere in western and north western part of Iran, Iran. *Air Quality, Atmosphere & Health*, 13(1), 119–130.
- Safford, H., Zuniga-Montanez, R. E., Kim, M., Wu, X., Wei, L., Sharpnack, J., ... & Bischel, H. N. (2022). Wastewater-based epidemiology for covid-19: Handling qpcr nondetects and comparing spatially granular wastewater and clinical data trends. *ACS Es&t Water*, 2(11), 2114-2124.
- Sahoo, B., & Bhaskaran, P. K. (2018). Multi-hazard risk assessment of coastal vulnerability from tropical cyclones—A GIS based approach for the Odisha coast. *Journal of environmental management*, 206, 1166-1178.
- Sangalli, L. M., Ramsay, J. O., & Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681 703.
- Santhosh, A., Sam, E., & Bindhu, B. K. (2020). Pedestrian accident prediction modelling—A case study in Thiruvananthapuram City. In *Transportation Research: Proceedings of CTRG 2017* (pp. 637-645). Springer Singapore.
- Sarkissian, R. D., Abdallah, C., Zaninetti, J.-M., & Najem, S. (2020). Modelling intra-dependencies to assess road network resilience to natural hazards. *Natural Hazards*, 103(1), 121–137.
- Sarri, A., Guillas, S., & Dias, F. (2012). Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification. *Natural Hazards and Earth System Sciences*, 12(6), 2003–2018.

- Sauerborn, R., & Ebi, K. (2012). Climate change and natural disasters – integrating science and practice to protect health [PMID: 28140855]. *Global Health Action*, 5(1), 19295.
- Saurina, C., Bragulat, B., Saez, M., & López-Casasnovas, G. (2013). A conditional model for estimating the increase in suicides associated with the 2008-2010 economic recession in England. *Journal of Epidemiology and Community Health*, 67(9), 779-787.
- Saurina, C., Marzo, M., & Saez, M. (2015). Inequalities in suicide mortality rates and the economic recession in the municipalities of Catalonia, Spain. *International Journal for Equity in Health*, 14(1), 75.
- Saurina, C., Saez, M., Marcos-Gragera, R., Barceló, M. A., Renart, G., & Martos, C. (2010). Effects of deprivation on the geographical variability of larynx cancer incidence in men, Girona (Spain) 1994-2004. *Cancer Epidemiology*, 34(2), 109-115.
- Schrödle, B., & Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6), 725-734.
- Schutte, F. H., & Breetzke, G. D. (2018). The influence of extreme weather conditions on the magnitude and spatial distribution of crime in Tshwane (2001–2006). *South African Geographical Journal= Suid-Afrikaanse Geografiese Tydskrif*, 100(3), 364-377.
- Scott-Hayward, L. A. S., Mackenzie, M. L., Donovan, C. R., Walker, C. G., & Ashe, E. (2014). Complex region spatial smoother (CReSS). *Journal of Computational and Graphical Statistics*, 23(2), 340– 360.
- Senapati, N., Halford, N. G., & Semenov, M. A. (2021). Vulnerability of European wheat to extreme heat and drought around flowering under future climate. *Environmental Research Letters*, 16(2), 024052.
- Serra, L., Juan, P., Varga, D., Mateu, J., & Saez, M. (2012). Spatial pattern modelling of wildfires in Catalonia, Spain 2004-2008. *Environmental Modelling & Software*, 40, 235-244.
- Serra, L., Saez, M., Juan, P., Varga, D., & Mateu, J. (2014a). A spatio-temporal Poisson hurdle point process to model wildfires. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 28(7), 1671-1684.
- Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C., & Rue, H. (2014b). Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: The case of Catalonia, 1994-2008. *Environmental and Ecological Statistics*, 21(3), 531-563.
- Serra, L., Vall-Llosera, L., Varga, D., Saurina, C., Saez, M., & Renart, G. (2022). Analysis of the geographic pattern of the police reports for domestic violence in Girona (Spain). *BMC Public Health*, 22, 552.
- Shaddick, G., Thomas, M. L., Amini, H., Broday, D., Cohen, A., Frostad, J., ... & Brauer, M. (2018). Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. *Environmental science & technology*, 52(16), 9069-9078.
- Shao, K., Liu, W., Gao, Y., & Ning, Y. (2019). The influence of climate change on tsunami-like solitary wave inundation over fringing reefs. *Journal of Integrative Environmental Sciences*, 16(1), 71–88.
- Shariati, M., Mesgari, T., Kasraee, M., & Jahangiri-Rad, M. (2020). Spatiotemporal analysis and hotspots detection of COVID-19 using geographic information system (March and April, 2020). *Journal of Environmental Health Science and Engineering*, 18, 1499-1507.

- Shim, S. R., Kim, S. J., Lee, J., & Rücker, G. (2019). Network meta-analysis: application and practice using R software. *Epidemiology and health*, 41.
- Shin, J. Y., Chen, S., & Kim, T.-W. (2015). Application of Bayesian Markov chain monte Carlo method with mixed Gumbel distribution to estimate extreme magnitude of tsunamigenic earthquake. *KSCE Journal of Civil Engineering*, 19(2), 366–375.
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors.
- Singh, A. (2019). Remote sensing and GIS applications for municipal waste management. *Journal of environmental management*, 243, 22-29.
- Smedt, T. D., Simons, K., Nieuwenhuysse, A. V., & Molenberghs, G. (2015). Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models. *Archives of Public Health*, 73(S1).
- Smit, A., Kijko, A., & Stein, A. (2017). Probabilistic tsunami hazard assessment from incomplete and uncertain historical catalogues with application to tsunamigenic regions in the Pacific ocean. *Pure and Applied Geophysics*, 174(8), 3065–3081.
- Song, Y., Huang, B., He, Q., Chen, B., Wei, J., & Mahmood, R. (2019). Dynamic assessment of PM<sub>2.5</sub> exposure and health risk using remote sensing and geo-spatial big data. *Environmental Pollution*, 253, 288-296.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Sreejaya, K. P., Raghukanth, S. T. G., Gupta, I. D., Murty, C. V. R., & Srinagesh, D. (2022). Seismic hazard map of India and neighbouring regions. *Soil Dynamics and Earthquake Engineering*, 163, 107505.
- Stein, S., Geller, R. J., & Liu, M. (2012). Why earthquake hazard maps often fail and what to do about it. *Tectonophysics*, 562, 1-25.
- Sugawara, D. (2017). Evolution of Numerical Modeling as a Tool for Predicting Tsunami-Induced Morphological Changes in Coastal Areas: A Review Since the 2011 Tohoku Earthquake. In *Advances in natural and technological hazards research* (pp. 451–467). Springer International Publishing.
- Tascikaraoglu, A. (2018). Evaluation of spatio-temporal forecasting methods in various smart city applications. *Renewable and Sustainable Energy Reviews*, 82, 424-435.
- Taylor, B. M., & Diggle, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, 84(10), 2266-2284.
- Tierney, L., Kadane, J., 1986. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 393 (81), 82–86.
- Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1), 234-240.



- Trilles, S., Juan, P., Chaudhuri, S., & Fortea, A. B. V. (2021). Data on CO<sub>2</sub>, temperature and air humidity records in Spanish classrooms during the reopening of schools in the COVID-19 pandemic. *Data in Brief*, 39, 107489.
- Trilles, S., Vicente, A. B., Juan, P., Ramos, F., Meseguer, S., & Serra, L. (2019). Reliability validation of a low-cost particulate matter IoT sensor in indoor and outdoor environments using a reference sampler. *Sustainability*, 11(24), 7220.
- UNDP. (2015). Sustainable Development Goals| United Nations Development Programme.
- United Nations Office for Disaster Risk Reduction (UNDRR). (2015). *Sendai Framework for Disaster Risk Reduction 2015-2030*. Retrieved October 19, 2021, from <https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030>
- United Nations. (2019). *World population prospects 2019: Highlights*. Retrieved January 8, 2020, from <https://www.un.org/development/desa/publications/world-population-prospects-2019-highlights.html>
- Valente, R. (2019). Spatial and temporal patterns of violent crime in a Brazilian state capital: A quantitative analysis focusing on micro places and small units of time. *Applied geography*, 103, 90-97.
- Varga, D., Roigé, M., Pintó, J., & Saez, M. (2019). Assessing the spatial distribution of the biodiversity in a changing temperature pattern. The case of Catalonia, Spain. *International Journal of Environmental Research and Public Health*, 16, 4026.
- Verdoy, P. J. (2019). Enhancing the SPDE modeling of spatial point processes with INLA, applied to wildfires. choosing the best mesh for each database. *Communications in Statistics - Simulation and Computation*, 50(10), 2990–3030.
- Verdoy, P. J. (2020). Spatio-temporal hierarchical Bayesian analysis of wildfires with Stochastic Partial Differential Equations. A case study from Valencian Community (Spain). *Journal of applied statistics*, 47(5), 927-946.
- Vicente, A. B., Juan, P., Meseguer, S., Díaz-Avalos, C., & Serra, L. (2018). Variability of PM<sub>10</sub> in industrialized-urban areas. New coefficients to establish significant differences between sampling points. *Environmental Pollution*, 234, 969-978.
- Vicente, A. B., Juan, P., Meseguer, S., Serra, L., & Trilles, S. (2019). Air quality trend of PM<sub>10</sub>. statistical models for assessing the air quality impact of environmental policies. *Sustainability*, 11(20), 5857.
- Vlad, I. T., Diaz, C., Juan, P., & Chaudhuri, S. (2023). Analysis and description of crimes in Mexico City using point pattern analysis within networks. *Annals of GIS*, 1-13.
- Wang, C., Quddus, M., & Ison, S. (2013). A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. *Transportmetrica A: Transport Science*, 9(2), 124-148.
- Wang, H. & Ranalli, M. G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics*, 63(1):209 217.
- Wang, H., Zhu, Y., Qin, T., & Zhang, X. (2022). Study on the propagation probability characteristics and prediction model of meteorological drought to hydrological drought in basin based on copula function. *Frontiers in Earth Science*, 10.

- Wang, R., Derdouri, A., & Murayama, Y. (2018). Spatiotemporal simulation of future land use/cover change scenarios in the Tokyo metropolitan area. *Sustainability*, *10*(6), 2056.
- Wang, S., Zhong, Y., & Wang, E. (2019). An integrated GIS platform architecture for spatiotemporal big data. *Future Generation Computer Systems*, *94*, 160-172.
- Wang, W., Yuan, Z., Yang, Y., Yang, X., & Liu, Y. (2019). Factors influencing traffic accident frequencies on urban roads: A spatial panel time-fixed effects error model. *PLoS one*, *14*(4), e0214539.
- Wang, Y., Yuan, Z., Liu, H., Xing, Z., Ji, Y., Li, H., ... & Mo, C. (2022). A new scheme for probabilistic forecasting with an ensemble model based on CEEMDAN and AM-MCMC and its application in precipitation forecasting. *Expert Systems with Applications*, *187*, 115872.
- Wang, Z., Ye, X., & Tsou, M. H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*, 523-540.
- Wang, Z., Yue, Y., He, B., Nie, K., Tu, W., Du, Q., & Li, Q. (2021). A Bayesian spatio-temporal model to analyzing the stability of patterns of population distribution in an urban space using mobile phone data. *International Journal of Geographical Information Science*, *35*(1), 116-134.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*.
- Weisburd, D., Morris, N. A., & Groff, E. R. (2009). Hot spots of juvenile crime: A longitudinal study of arrest incidents at street segments in Seattle, Washington. *Journal of quantitative criminology*, *25*, 443-467.
- Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.* *40*, 974–994.
- Wikle, C. K., Berliner, L. M., & Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, *5*, 117–154.
- Wikle, C., 2003. Hierarchical models in environmental science. *International Statistical Review* *71* (2), 181–199.
- Wilson, B. (2020). Evaluating the INLA-SPDE approach for Bayesian modeling of earthquake damages from geolocated cluster data.
- Wood, A.W., Leung, L. R., Sridhar, V. and Lettenmaier, D. P. (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, *62*, 189–216.
- Wood, S. N., Bravington, M. V., & Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *70*(5), 931–955.
- World Health Organization. (2016). *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization.
- Wrenn, D. H., & Sam, A. G. (2014). Geographically and temporally weighted likelihood regression: Exploring the spatiotemporal determinants of land use change. *Regional Science and Urban Economics*, *44*, 60-74.

- Wright, N., Newell, K., Lam, K. B. H., Kurmi, O., Chen, Z., & Kartsonaki, C. (2021). Estimating ambient air pollutant levels in Suzhou through the SPDE approach with R-INLA. *International Journal of Hygiene and Environmental Health*, 235, 113766.
- Xu, P., & Huang, H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention*, 75, 16-25.
- Yang, W., Shen, L., & Shi, P. (2015). Mapping landslide risk of the world. *World atlas of natural disaster risk*, 57-66.
- Yanosky, J. D., Paciorek, C. J., Laden, F., Hart, J. E., Puett, R. C., Liao, D., & Suh, H. H. (2014). Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. *Environmental Health*, 13, 1-15.
- Yi, H., Güneralp, B., Kreuter, U. P., Güneralp, İ., & Filippi, A. M. (2018). Spatial and temporal changes in biodiversity and ecosystem services in the San Antonio River Basin, Texas, from 1984 to 2010. *Science of the total environment*, 619, 1259-1271.
- Yu, I. T., Li, Y., Wong, T. W., Tam, W., Chan, A. T., Lee, J. H., ... & Ho, T. (2004). Evidence of airborne transmission of the severe acute respiratory syndrome virus. *New England Journal of Medicine*, 350(17), 1731-1739.
- Zakaria, N. A., & Noor, N. M. (2018). Imputation methods for filling missing data in urban air pollution data formalaysia. *Urbanism. Arhitectura. Constructii*, 9(2), 159.
- Zapata-Cachafeiro, M., Prieto-Campo, Á., Portela-Romero, M., Carracedo-Martínez, E., Lema-Oreiro, M., Piñeiro-Lamas, M., ... & Figueiras, A. (2022). Effect of Previous Anticoagulant Treatment on Risk of COVID-19. *Drug Safety*, 1-9.
- Zeng, C., Liu, Y., Stein, A., & Jiao, L. (2015). Characterization and spatial modeling of urban sprawl in the Wuhan Metropolitan Area, China. *International Journal of Applied Earth Observation and Geoinformation*, 34, 10-24.
- Zeng, J., Xiong, Y., Liu, F., Ye, J., & Tang, J. (2022). Uncovering the spatiotemporal patterns of traffic congestion from large-scale trajectory data: A complex network approach. *Physica A: Statistical Mechanics and its Applications*, 604, 127871.
- Zeng, Q., & Huang, H. (2014). Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention*, 67, 105-112.
- Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., & Vannucci, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data.
- Zhang, W., Yu, Y., Qi, Y., Shu, F., & Wang, Y. (2019). Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning. *Transportmetrica A: Transport Science*, 15(2), 1688-1711.
- Zheng, Y., Xie, Y., & Long, X. (2021). A comprehensive review of Bayesian statistics in natural hazards engineering. *Natural Hazards*, 108(1), 63-91.
- Zhou, C., Wang, H., Wang, C., Hou, Z., Zheng, Z., Shen, S., ... & Zhu, Y. (2021). Geoscience knowledge graph in the big data era. *Science China Earth Sciences*, 64(7), 1105-1114.

- Zhu, Z., Chen, B., Zhao, Y., & Ji, Y. (2021). Multi-sensing paradigm based urban air quality monitoring and hazardous gas source analyzing: a review. *Journal of Safety Science and Resilience*, 2(3), 131-145.
- Zhuang, J., & Mateu, J. (2019). A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(3), 919-942.
- Zirschky, J. (1985). Geostatistics for environmental monitoring and survey design. *Environment international*, 11(6), 515-524.
- Zorn, M., & Komac, B. (2013). Contribution of Ivan Gams to Slovenian physical geography and geography of natural hazards. *Acta geographica Slovenica*, 53(1), 23–41.
- Zuur, A. F., Elena, N. I., & Anatoly, A. S. (2017). Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with R-INLA Volume I: using GLM and GLMM. Highland Statistics Ltd. *Newburgh United Kingdom*.

## 8. ANNEX

We hereby present additional publications that have been produced during the course of our research endeavors, aligned with the core objectives of the research topic.

### 8.1 List of Additional Publications

Publication	Status
<u>Chaudhuri, S., Juan, P., &amp; Serra, L. (2021).</u> Analysis of precise climate pattern of Maldives. A complex island structure. <i>Regional Studies in Marine Science</i> , 44, 101789.	Published
<u>Chaudhuri, S., Moradi, M., &amp; Mateu, J. (2021).</u> On the trend detection of time-ordered intensity images of point processes on linear networks. <i>Communications in Statistics-Simulation and Computation</i> , 1-13.	Published
Díaz-Avalos, C., Juan, P., <u>Chaudhuri, S.</u> , Sáez, M., & Serra, L. (2020). Association between the new COVID-19 cases and air pollution with meteorological elements in nine counties of New York state. <i>International Journal of Environmental Research and Public Health</i> , 17(23), 9055.	Published
Vicente, AB., <u>Chaudhuri, S.</u> , Juan, P., Díaz-Avalos, C., & Serra, L. (2020). <i>Relación entre contaminantes atmosféricos, variables meteorológicas y casos de COVID-19 en ciudades Europeas</i> . Congreso Nacional del Medio Ambiente (CONAMA), Madrid, Spain.	Published
Carbó, E., Juan, P., Añó, C., <u>Chaudhuri, S.</u> , Diaz-Avalos, C., & López-Baeza, E. (2021). Modeling Influence of Soil Properties in Different Gradients of Soil Moisture: The Case of the Valencia Anchor Station Validation Site, Spain. <i>Remote Sensing</i> , 13(24), 5155.	Published
Trilles, S., Juan, P., <u>Chaudhuri, S.</u> , & Fortea, A. B. V. (2021). Data on CO2, temperature and air humidity records in Spanish classrooms during the reopening of schools in the COVID-19 pandemic. <i>Data in Brief</i> , 39, 107489.	Published
Carbo, E., Juan, P., Añó, C., <u>Chaudhuri, S.</u> , Diaz-Avalos, C., & López-Baeza, E. (2022). <i>Modeling soil moisture at the Valencia Anchor Station (VAS), Eastern Spain</i> (No. EGU22-8625). Copernicus Meetings, Vienna, Austria.	Published
Zapata-Cachafeiro, M., Prieto-Campo, A., Portela-Romero, M., Carracedo-Martínez, E., Lema-Oreiro, M., Piñeiro-Lamas, <u>Chaudhuri, S.</u> , Salgado-Barreira, A., Figueiras, A. (2022). Effect of Previous Anticoagulant Treatment on Risk of COVID-19. <i>Drug Safety</i> 2022.	Published
Niraula, P., Mateu, J., & <u>Chaudhuri, S.</u> (2022). A Bayesian machine learning approach for spatio-temporal prediction of COVID-19 cases. <i>Stochastic Environmental Research and Risk Assessment</i> , 1-19.	Published
Vlad, Iulian T., Díaz-Avalos, C., Juan, P., <u>Chaudhuri, S.</u> (2023). Analysis and description of crimes in Mexico City using point pattern analysis within networks. <i>Annals of GIS</i> .	Published

## **8.2 Article 4: Climate Pattern in Complex Islands**

### **Analysis of precise climate pattern of Maldives. A complex island structure**

Somnath Chaudhuri<sup>1,3</sup>, Pablo Juan<sup>1,2</sup> and Laura Serra<sup>1,3</sup>

1. Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain.
2. Department of Mathematics, Universitat Jaume I, Castellón, Spain.
3. CIBER of Epidemiology and Public Health (CIBERESP), Spain.



# Analysis of precise climate pattern of Maldives. A complex island structure

Somnath Chaudhuri <sup>a,c</sup>, Pablo Juan <sup>a,b,\*</sup>, Laura Serra <sup>a,c</sup>

<sup>a</sup> Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain

<sup>b</sup> Department of Mathematics, Universitat Jaume I, Castellón, Spain

<sup>c</sup> CIBER of Epidemiology and Public Health (CIBERESP), Spain



## ARTICLE INFO

### Article history:

Received 12 May 2020

Received in revised form 25 February 2021

Accepted 12 April 2021

Available online 18 April 2021

### Keywords:

Climate pattern

Geostatistics

Kriging

Maldives islands

## ABSTRACT

Republic of Maldives is located on the south and south-western region of the coast of India. The country is one of the most geographically dispersed nations in the world, with 1192 coral islands grouped into 26 natural atolls in the middle of the Indian Ocean. A descriptive study of its climatic conditions is presented in this work. We have used geostatistical technique of kriging (described below) for the estimation of meteorological variables. Complexity related to the structure of the region, with diverse distribution of islands of varying sizes, is discussed in connection to the analysis. Climatic characteristics explored in the current work indicate the need for subsequent studies of seasonal patterns in climate change, especially temperatures and precipitation across the country, and also to identify the effect of extreme climatic conditions and natural disasters such as Tsunamis. The results do not show periodicity over the study period. It emphasizes that climatic patterns appearing in the study area must be analyzed more extensively over time, with the inclusion of a greater number of meteorological stations for precise spatio-temporal analysis.

© 2021 Elsevier B.V. All rights reserved.

## 1. Introduction

Republic of Maldives is one of the most geographically dispersed nations in the world, with 1192 coral islands grouped into 26 natural atolls in the middle of the Indian Ocean, as shown in Fig. 1. Atolls vary in size and number of islands. This variation is related to geographical grouping. Some atolls are huge, like Huvadhu atoll with 255 islands, while the atoll of Gnaviyani is based on only one island. Fig. 1 depicts the geographical location, the naturally occurring atolls and the complex island structure forming the atolls of Maldives. The islands are low lying and almost 80% of them are less than 3 ft above the sea level. The archipelago is more than 800 km long and 130 km wide (Ministry of Environment and Energy, 2017). Maldives stretches from north to south and the Equator crosses between Fuvahmulah and Gaafu Dhaalu Atoll in the southern parts of the nation. Ever since its beginning, the tourism industry has represented a growing portion of the GDP share of the country. The main image sold in tourism marketing has been the sunny side of life, representing the climate, beautiful sandy beaches, and the crystal-clear water around these small islands.

The coral reefs of Maldives represent the most diverse reefs in the Indian Ocean (Naseer, 2007). Due to the nature and structure

of the coral reefs, conventional definitions of coastal area applied to continental land do not apply to this island nation. A more appropriate notion of coastal land for Maldives can be seen in Fig. 2 which represents Addu atoll, the southernmost atoll of the country. The figure on the left depicts the atoll which includes enclosed lagoon or basin, forereef, subtidal reef, pass reef flat and land on reefs. It appears that the whole area enclosed by the reef and lagoon is land surface. But the figure on the right shows only land on reefs areas of the same atoll. In fact, only five percent of the total reef area of the Maldives is land. Most of the reefs are landless with vast expanses of shallow reef flats and are extremely small islands (size ranges from 0.1 to 5 square kilometers) (Naseer, 2007). As a result of this, only two hundred of the 1192 islands are inhabited.

More than 40 percent of the country's population resides in Male', the capital city of Maldives. Most of the other inhabited islands are sparsely populated. About 120 islands are assigned exclusively for tourism development as resort sites. The remaining uninhabited islands are used for agriculture and other commercial developments (Shifaza, 2018).

According to the Köppen system, India is characterized by 3 types of climates depending on their location: (1) arid deserts in the west, (2) alpine tundra and glaciers in the north, and (3) humid tropical regions in the tropical forests of the southwest as well as on the islands (Beck et al., 2018). Thus, the climate in the Maldives is characterized as tropical climate. The temperature is

\* Corresponding author.

E-mail address: [juan@uji.es](mailto:juan@uji.es) (P. Juan).



Fig. 1. Study area: Maldives geographical location and complex island structure of the atolls.



Fig. 2. Addu atoll with both lagoon and reef areas and with only land on reefs areas.

moderately high all year round, but influenced by the monsoon winds. Seasonal changes in the direction of the monsoon winds control the weather conditions of the islands. With a stretch of 800 km from north to south and the equator crossing the country, the weather conditions vary considerably in different parts of the archipelago (Climate of Maldives, 2020). In particular, the southwest monsoon (from late April to September), which includes wind, higher humidity, and more frequent cloud cover, is more intense on the northern islands than elsewhere. The north-east monsoon (October to December), located mainly on the southern atolls, is calmer and simply brings rain and thunderstorms in the afternoon or evening. Finally, the driest period, outside the monsoons, runs from January to April and is most felt in the northern atolls (Weather-atlas, 2020). Maldives, being located near the Equator, offers warm and stable temperature throughout the year and protection from cyclones. But there are specific particularities depending on the location of each atoll. In addition, the weather condition is generally humid with a high relative humidity of around 80 percent. Finally, as is common in tropical areas, the rains are torrential and often devastating, being short but intense (Climate of Maldives, 2020). Maldives has been occasionally been

hit by tropical cyclones, but the cyclones are not considered to be in any geological system at risk. Even the morphology of the coral reefs helps to prevent cyclonic catastrophe (Naylor, 2015). Due to this morphology of the sea-beds, the gigantic waves of tsunamis, as happened on the December 26, 2004, did not cause disastrous consequences, but only led to notable flood phenomena. In addition, the external eastern coral reef acts as a natural barrier absorbing the impact of the anomalous waves and protects the internal islands (Naylor, 2015).

There are some important objectives that we have discussed in the current study. Firstly, the limited number of descriptive studies on identifying the spatial and seasonal patterns of the climate in this region demands more extensive spatio-temporal research work. Since Maldives is a popular tourist destination, the seasonal patterns of weather conditions can cause recurrent fluctuations in tourism demand. A second objective is to identify the correlations among the meteorological parameters like precipitation, temperature, atmospheric pressure, and relative humidity (RH). This can help us to better understand how typical local atmospheric phenomena or extreme climatic conditions develop and influence the seasonal pattern of the climate. A



third objective is to analyze individual meteorological parameters to interpret the climatic patterns or seasonality in the region, and at the same time to identify the impact on the seasonality because of extreme climatic conditions and natural hazards like Tsunami. A final objective is to illustrate the complex island structure of the archipelago and the computational issues involved in geostatistical studies. This paper presents a geostatistical study on an island region that differs from other studies as the land structure and distribution of the islands are unique. There are a few existing useful works on the analysis and description of elements on islands (Bland et al., 2019; Brushett et al., 2014; Kabir et al., 2020; Riyas et al., 2020; Staniec and Vlahos, 2017). However, in the current study, it is important to highlight the complex archipelago structure of the Maldives since it hinders the usual methodological and technical development. Because of the advantages of kriging over traditional interpolation techniques, the current study can help future researchers to identify other study areas having similar or more complex land structures where similar methodology can be applied.

The rest of the paper is organized as follows. In Section 2 we present the locations of the five Meteorological (MET) Stations of Maldives and the meteorological parameters used in the current study. We briefly highlight some details of the exploratory data analysis and kriging method used to develop geostatistics in Section 3. Section 4 reports the results with graphical representations. Section 5 contains a discussion about the exploratory analysis and kriging as well as the comparison with other similar results in the literature. The paper ends with the principal conclusions on the absence of common seasonal periodicity in the study area during the study period in Section 6.

## 2. Data settings

The original data is collected from the Maldives Meteorological Service (MMS), Republic of Maldives ("Maldives Meteorological Service", 2020). MMS is responsible for the seismological and meteorological services in the country. The dataset provides monthly weather reports from five MET stations under MMS namely, Gan, Hanimaadhoo, Hulhule, Kaadedhdhoo and Kadhdhoo. The offices are located in different atolls covering the full stretch of Maldives from north to south as depicted in Fig. 3.

The current dataset from five individual MET stations includes monthly average records of four meteorological parameters: precipitation, temperature, atmospheric pressure (atm. pressure) and RH. The time frame of the present study is from January 2000 to December 2015. The sampled dataset is divided into four independent sets for each meteorological parameter having 960 observations in each case.

Table 1 reports the minimum, maximum and average value of each of the meteorological parameters from individual MET stations. It allows the identification of differences between the five stations as well as comparing the range of values for each parameter. For instance, the maximum value of precipitation was recorded in the month of November in Kaadedhdhoo office (624.9 mm) while Hanimaadhoo had the lowest precipitation value during the study period. Temperatures range from 25.1 to 30.3 degree Celsius in all the offices. In this case, Kaadedhdhoo is the one with the lowest values. The other two variables show little differences among the MET stations. Therefore, the data presented in Table 1 shows that there are considerable variations between the northern, central, and southern regions. For this reason, more methodological studies are needed in the area.

## 3. Methods

The methodology used in the current study is, first, an exploratory data analysis to inspect the datasets and summarizes the characteristics and distribution of the data, supported by graphical plots. Subsequently, we have implemented geostatistical techniques of kriging. The type of dataset used in the current study is clearly geostatistical data. Geostatistics is the science that studies phenomena that fluctuate in space and/or time and offers a collection of statistical tools for the description and modeling spatial (and temporal) variability (Boer et al., 2001; Bostan et al., 2012; Juan and Mateu, 2009). Geostatistical methods have a wide range of applications, for example, in soil science, meteorology, and ecology (Jordan et al., 2004; Serra et al., 2017). The objectives of a typical geostatistical analysis are estimation and prediction. Estimation refers to inference about the parameters of a stochastic model for the data, and prediction refers to inference about the realization of the unobserved signal. It was the South African mining engineer (Kriging, 2015) who pioneered work in the field of geostatistics and presented the basic equations for optimal linear interpolation of spatially correlated variables. The name Kriging was immortalized by Matheron (1962) in a series of books and papers that used the French term Kriging. An important tool in geostatistics is kriging, which refers to a least square linear predictor that, under certain stationarity assumptions, requires at least the knowledge of covariance parameters and the functional form for the mean of the underlying random function. The modern development of spatial methods came from Besag (1974). For the description of the climate data of the Maldives, the geostatistical technique of kriging has been used, employing variograms. The basic paradigm of predictive Geostatistics is both the characterization of the unknown value  $z$  as a continuous random variable  $Z$ , and the associated uncertainty defined by the corresponding probability distribution. In the context of Geostatistics, the random variable  $Z$  shows a significant dependence on the spatial location, denoted by  $Z(u)$ , where  $u$  is the spatial location. The cumulative distribution function (cdf) of this continuous random variable depends on the spatial location  $Z(u)$  and is defined by:

$$F(u; z) = P\{Z(u) \leq z\} \quad (1)$$

being  $P$  the associated likelihood function. In this case, we can work with the conditional cumulative distribution function (ccdf) given by:

$$F(u; z|n) = P\{Z(u) \leq z|n\} \quad (2)$$

In Geostatistics, it is important to model the correlation or dependence between a certain variable  $Z(u_i)$ ,  $i=1, \dots, n$ , and other potential covariates in other sampling points  $Y(v_j)$ ,  $j=1, \dots, n$ . A random function defines a set of random variables defined on the field of interest (Cressie, 1993). In most applications, only one sample is possible in each spatial location  $u$ , so you need to use the condition of repeatability, which is guaranteed under the assumption of stationarity. A random variable  $Z(u)$ ,  $u \in A$  is stationary in the region  $A$ , if the multivariate cdf is invariant under any translation  $C$  made about locations, this is:

$$F(u_1, \dots, u_k; Z_1, \dots, Z_k) = F(u_1 + C, \dots, u_k + C; Z_1, \dots, Z_k) \quad (3)$$

for any translation vector  $C$ .

Next, we focus briefly on semivariogram and kriging. The basic object we consider is a stochastic process  $\{Z(u), u \in D\}$  in which  $D$  is a subset of  $R^d$  (Euclidean space  $d$ -dimensional), although normally  $d = 2$ . Assuming that the average value in a location  $u$  expressed as  $\mu(u)$  is a constant, which we assume as zero without loss of generality, we can define:

$$2\gamma(u_1 - u_2) = \text{var}\{Z(u_1) - Z(u_2)\} \quad (4)$$



**Fig. 3.** Locations of five meteorological offices under MMS. Source: Locations of five Climate Observatories (MET stations) collected from Maldives Meteorological Service, "Maldives Meteorological Service", (2020).

**Table 1**

Principal descriptive values about the four meteorological parameters for each MET station. Source: Collected from Maldives Meteorological Service, "Maldives Meteorological Service," (2020).

MET station	Precipitation (mm)			Atm. pressure (hPa)			Temperature (deg. Cel)			RH (percentage)		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Gan	0.6	187.3	530.7	1008	1011	1014	27	28.17	29.3	75	80.72	86
Hanimaadhoo	0	142.09	511.7	1008	1010	1013	26.9	28.46	30	73	79.73	86
Hulhule	0	166.06	568.9	1008	1010	1013	27.4	28.71	30.2	71	79.05	85
Kaadedhdhoo	0	181.28	624.9	978.4	1010.6	1004.2	25.1	28.46	29.9	72	79.7	86
Kadhdhoo	0.1	174.67	542.5	1009	1011	1013	27.3	28.61	30.3	72	79.11	84
Summary	0	170.3	624.9	978.4	1010.7	1014.2	25.1	28.48	30.3	71	79.66	86

The function  $2\gamma(\cdot)$  is called variogram and  $\gamma(\cdot)$ , semivariogram. The semivariogram represents a rate of change showing a variable (attribute) with distance. Its shape pattern describes spatial variation in terms of size and general shape. The maximum value that a semivariogram reaches is called sill, or prior variance, and indicates the low-level data which defines a stationary second-order process. The lag or distance, for which the sill is reached, is called range and defines the limit of the spatial dependence. Finally, a semivariogram with a separate variance term defined is called nugget, and it defines the intrinsic variability in the data that has not been captured by the range of distances analyzed or any purely random variation (Juan and Mateu, 2009). After testing a variety of models, for our data, the best one was the Spherical Model for our data, defined by a current range  $a$ , a prior variance

(sill)  $c_1$  and a nugget effect  $c_0$ ,

$$\gamma(|h|) = \begin{cases} c_0 + c_1 \cdot 1.5 \frac{|h|}{a} - 0.5 \left(\frac{|h|}{a}\right)^3 & \text{if } |h| \leq a, \\ c_0 + c_1 & \text{if } |h| \geq a \end{cases} \quad (5)$$

Thereafter, we will focus on the main element in Geostatistics using spatial covariance models to predict and interpolate spatial processes, the Kriging. An important problem is the following: given a set of observations of a spatial attribute  $Z(u_1), Z(u_2), \dots, Z(u_n)$ , the goal is to predict the value  $Z(u_0)$  for some  $u_0 \notin \{u_0, \dots, u_n\}$ . All Kriging estimates are variants of the basic linear regression estimates which predict the value of  $Z$  in the location attribute  $u_0$ , denoted by  $Z^*(u_0)$ , and is defined by:

$$Z^*(u_0) - m(u_0) = \sum_{\alpha=1}^{n(u_0)} \lambda_{-\alpha}(u_0) [Z(u_{\alpha}) - m(u_{\alpha})] \quad (6)$$

where  $\lambda_{\alpha}$  defines the weighting assigned to the data involved in the summation weight, and  $m(u_0)$  and  $m(u_{\alpha})$  are the corresponding expected values of  $Z(u_0)$  and  $Z(u_{\alpha})$  respectively. Note that only those neighboring locations  $u_{\alpha}$  are required to operate the localization prediction  $u_0$ .

The error variance  $\sigma_E^2(u)$  in its general form is given by:

$$\sigma_E^2(u) = \text{Var}[Z^*(u) - Z(u)] \quad (7)$$

where the superscript \* indicates the estimated value for that location. Furthermore, in this expression the variance is minimized under the constraint of unbiasedness,  $[Z^*(u) - Z(u)] = 0$ .

According to the model considered for the trend, we can consider two variants of Kriging: simple kriging (SK) with known and constant trend over the entire area of study and ordinary kriging (OK) that considers the possible local fluctuations of the trend or average (Cressie, 1993). When we introduce more covariates, the cokriging regression methods take part in which multiple attributes are involved. Suppose then that we have two variables  $Z$  and  $Y$  defined in the same locations. The equation for estimating the value of the main variable in the location  $u_0$  is given as:

$$Z_{COK}^*(u_0) = \sum_{\alpha_1=1}^{n_1} \lambda_{\alpha_1}(u_0) Z(u_{\alpha_1}) + \sum_{\alpha_2=1}^{n_2} \lambda_{\alpha_2}^*(u_0) Y(u_{\alpha_2}^*) \quad (8)$$

It requires a model for the covariance matrix of features, including covariance of  $Z$ ,  $C_Z(h)$ , covariance of  $Y$ ,  $C_Y(h)$ , cross-covariance of  $Z - Y$ ,  $C_{ZY}(h) = \text{Cov}[Z(u), Y(u + h)]$ , and cross-covariance of  $Y - Z$ ,  $C_{YZ}(h)$ .

The final step is the cross-validation. General statistical modeling requires a posteriori validation of the results and re-estimation based on the known values under the same conditions the constructed models were subject to. These include variogram models, the type of kriging and the choice of the general modeling strategy. The cross-validation technique (CV) is used to compare estimated models with actual values. The idea consists in an iterative process where, each time real data is deleted, it is estimated with the remaining data.

In the current study, the R programming language (version R 3.6.1) ("R Core Team", 2020) has been used for the exploratory and graphical analysis and, for computing kriging plots, ArcGIS Pro (version 2.4.1) ("ArcGIS Pro Resources", 2020) geographic information system application. RStudio (version RStudio 1.2.1335) integrated development environment has been used to implement R ("RStudio", 2020).

#### 4. Results

In this section we first report the results of exploratory analysis. Annual and monthly data are displayed for the four meteorological parameters, and then for each MET station. In addition, the kriging maps include different time points that will be of interest to better understand the behavior of the meteorological parameters considered. The kriging results are divided into two categories: (1) three years in which Tsunami took place (from 2003 to 2005), and (2) another three observation years (2000, 2009 and 2015) covering the entire study period.

##### Annual and monthly plot of meteorological elements

###### • Average Annual Records (for all MET stations)

Individual plots depict the average annual records for four meteorological parameters combinedly for all MET stations. Thus, the value is comparable to the average annual climate records for the entire country of Maldives.

###### • Average Monthly Records (for all MET stations)

Individual plot depicts the average monthly records for four meteorological parameters combinedly for all MET stations. Thus,

the value is comparable to the average monthly climate records for the entire country of Maldives.

While analyzing the average annual climate data some variability is observed throughout the analyzed time window although the general trend is quite stable (Fig. 4). Furthermore, the variability seems to be important, although if we look at the scale of the y axis, we see that the differences over time are minimal. On the other hand, when we analyze the monthly behavior, the data shows practically no seasonality pattern (Fig. 5). This is because the equator crosses the study area, so the climate is quite stable throughout the year. However, it is possible to observe some unique trends in the meteorological parameters. For example, in case of precipitation, like the case of RH, it is observed that it increases as the year progresses, with December being the month with the highest amount of rainfall. The temperature, on the other hand, shows a peak in the month of April and then it decreases again until the end of the year. Finally, with respect to atmospheric pressure, the behavior is more variable throughout the year, decreasing to minimum values between April and June, increasing again to another peak in September and decreasing slightly in the last months of the monsoon.

###### • Annual Records (for individual MET station)

Individual plot depicts the annual records for four meteorological parameters records separately for each MET station.

Fig. 6 depicts annual precipitation records for each MET station. Clearly, in all cases, from Gan to Kadhdhoo, the values are very similar. In addition, 2002, 2005 and the last study years, 2014 and 2015 experience the highest rainfall with mean records close to 200 mm. On the contrary, Fig. 7 shows clear variations in terms of atmospheric pressure. In general, for all MET stations, there is a rise until 2002, a drop in pressure until 2011 and a rise again with the highest value recorded in 2015. Another clear distinction from precipitation is that there are years where wide variations of atmospheric pressures are observed among different MET stations. A clear example is 2014 in Gan, with distinct variation of data having very high values, which are not recorded in other MET stations.

Fig. 8 clearly depicts variations of temperature patterns in all the MET stations. In addition the mean value records are different in each case, much lower in Gan and Kadhdhoo and higher values in Hanimaadhoo and Hulhule. All this accompanied by a pattern of similarity in temperatures over the years in Gan with a slow gradual increase. In case of Kaadedhdhoo, a decreasing trend is noted, similar trend is observed in Kadhdhoo. While Hulhule experiences a slow but gradual increase in annual temperature. Fig. 9 illustrates the annual records of RH in each MET stations. The records clearly show a reverse trend with respect to temperature. It is noted that, the stations where the annual temperatures have been increasing, as in Hulhule, the RH is found to be decreasing. On the other hand, Kaadedhdhoo and Kadhdhoo experience gradual increase in RH values. In case of Gan, the temperature values have been regular, similar trend is observed for RH with some occasional outliers.

###### • Monthly Records (for individual MET stations)

Individual plot depicts the monthly records for four meteorological parameters records separately for each MET station.

Fig. 10 depicts the monthly precipitation records in individual MET stations. The patterns are similar, where early each year there are low precipitation values, an increase in the central months and high to intermediate values in the last months. Hanimaadhoo and Hulhule experience the driest months in the beginning of the year compared to other three MET stations. Except for Hanimaadhoo, all other MET stations record high monthly rainfall in the last two months of the year. Fig. 11 illustrates that the atmospheric pressure follows a sinusoidal pattern having higher value in the beginning of the year. It is noteworthy to

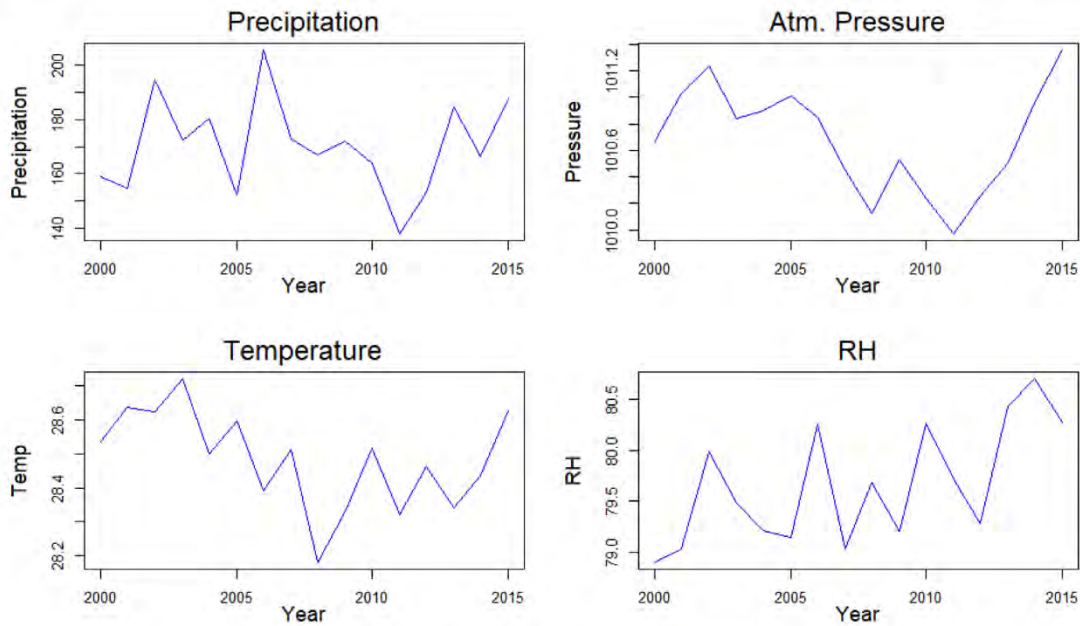


Fig. 4. Average annual meteorological records (for all MET stations).

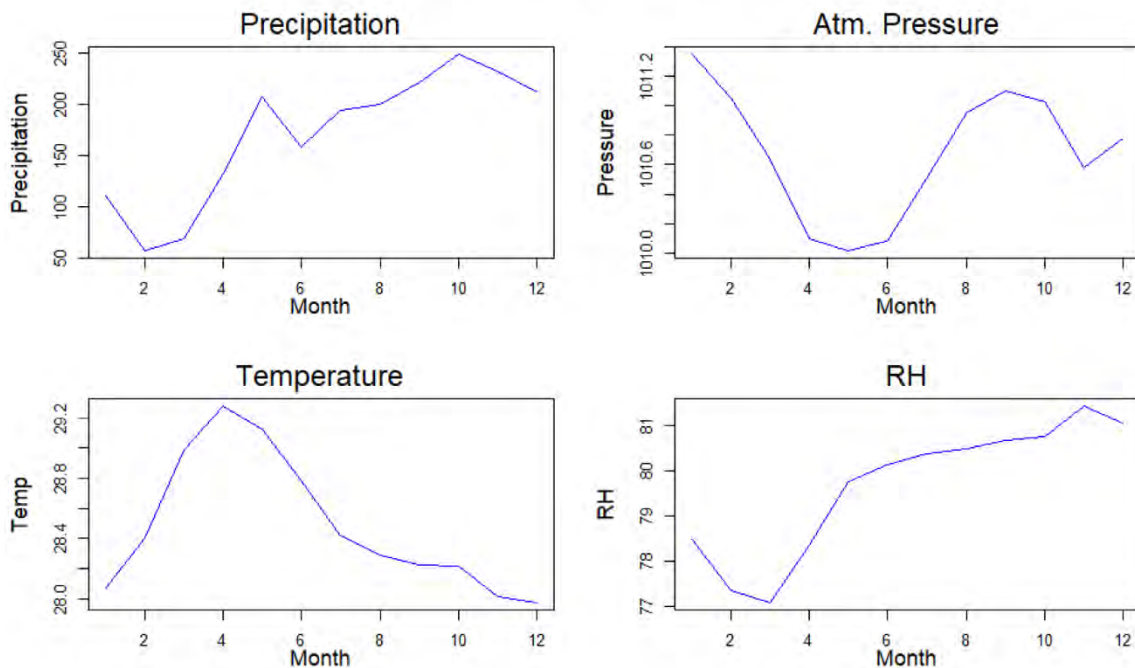


Fig. 5. Average monthly meteorological records (for all MET stations).

mention in this context that monthly variation in atmospheric pressure is noted to be relatively low in all MET stations.

In Fig. 12, the records of monthly temperatures follow the similar trend in all MET stations, an annual increase until April and May and a pronounced decrease until the last months of the year. Though Gan experiences a similar trend, its temperature values are comparatively low with respect to other four MET stations. Fig. 13 illustrates the pattern of RH for each MET station are not at all correlated with temperature, it occurs, but not in such pronounced way as in Fig. 12.

Next, a very important element in the description of similar datasets is the correlation values among four meteorological parameters overall and separately for each meteorological office data. Overall correlation among the four parameters for the entire

study period has been reported in Fig. 14. Individual MET stations correlation results are not illustrated in the current study. The results shown in Fig. 14 depicts a positive relationship between RH and precipitation. On the other hand, a negative relationship exists between RH and temperature. Also, there is a negative relationship between precipitation and temperature, but the value is relatively small. Among the rest of the parameters, no clear relationships are identified.

Figs. 15–18 illustrate the maps of the study area using kriging for the four meteorological parameters considered. In each figure different temporal moments of the kriging result have been depicted. On one hand, three years that coincide with the Tsunami that took place in the study area are presented. On the other hand, maps of another three years representing the complete study

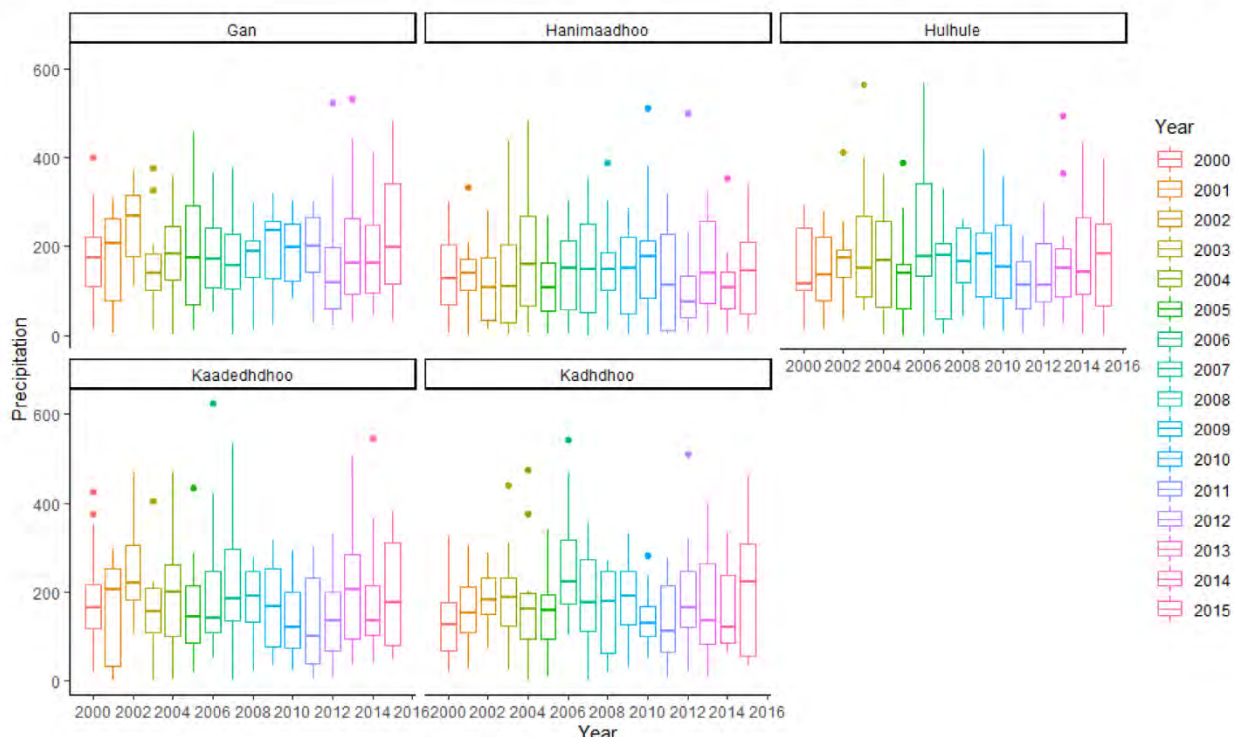


Fig. 6. Annual precipitation records (for individual MET stations).

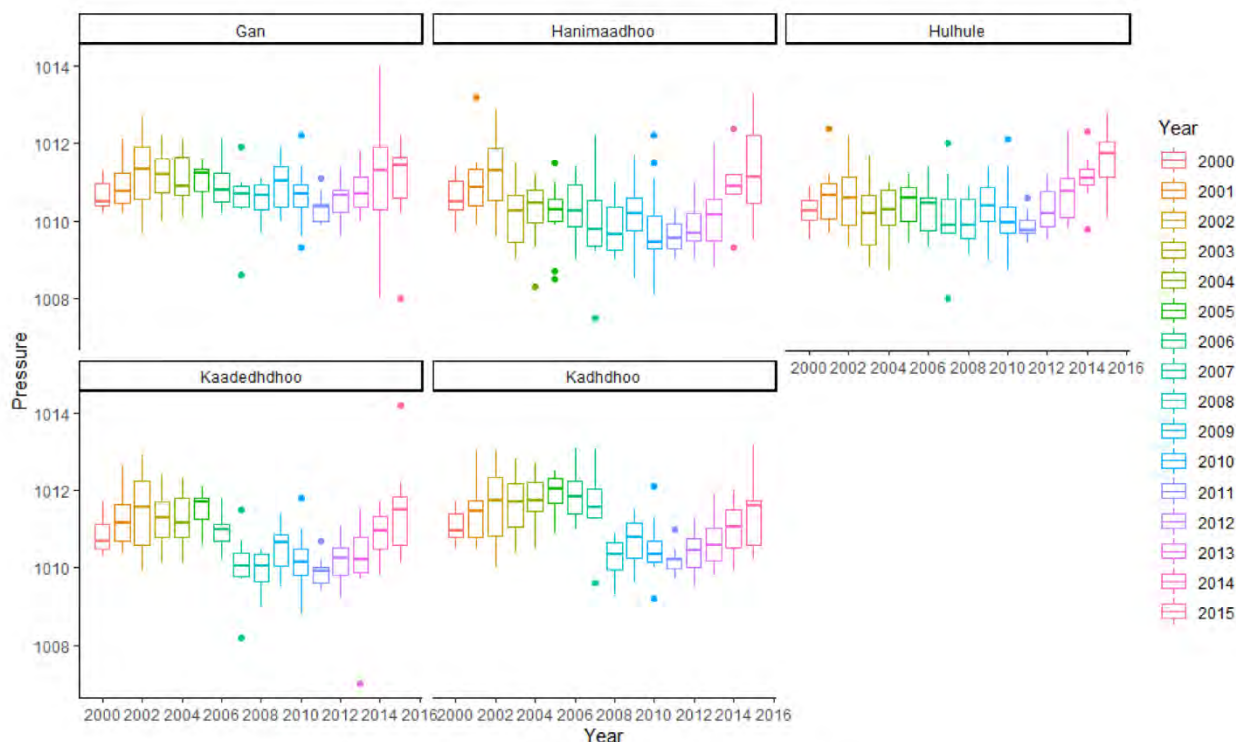


Fig. 7. Annual atmospheric pressure records (for individual MET stations).

period from initial year (2000), intermediate and post Tsunami year (2009) and final study year (2015) have been depicted.

In the case of precipitation as depicted in Fig. 15, the values are very low for the year 2000, with only slightly higher values in the south (Gan). In the years of the Tsunami, the values began to increase, especially in the central part of the study area in 2003, showing high values for the entire territory in 2004. In 2005 the

same pattern was repeated as in 2000. Finally, in the years 2009 and 2015, we again observed high values, mainly in the central and southern parts of the territory.

In relation to atmospheric pressure, maximum values stand out throughout the entire territory in 2015. During the time of the Tsunami, however, high values occurred only in the central-southern part of the study area (Fig. 16). While 2000 and 2009

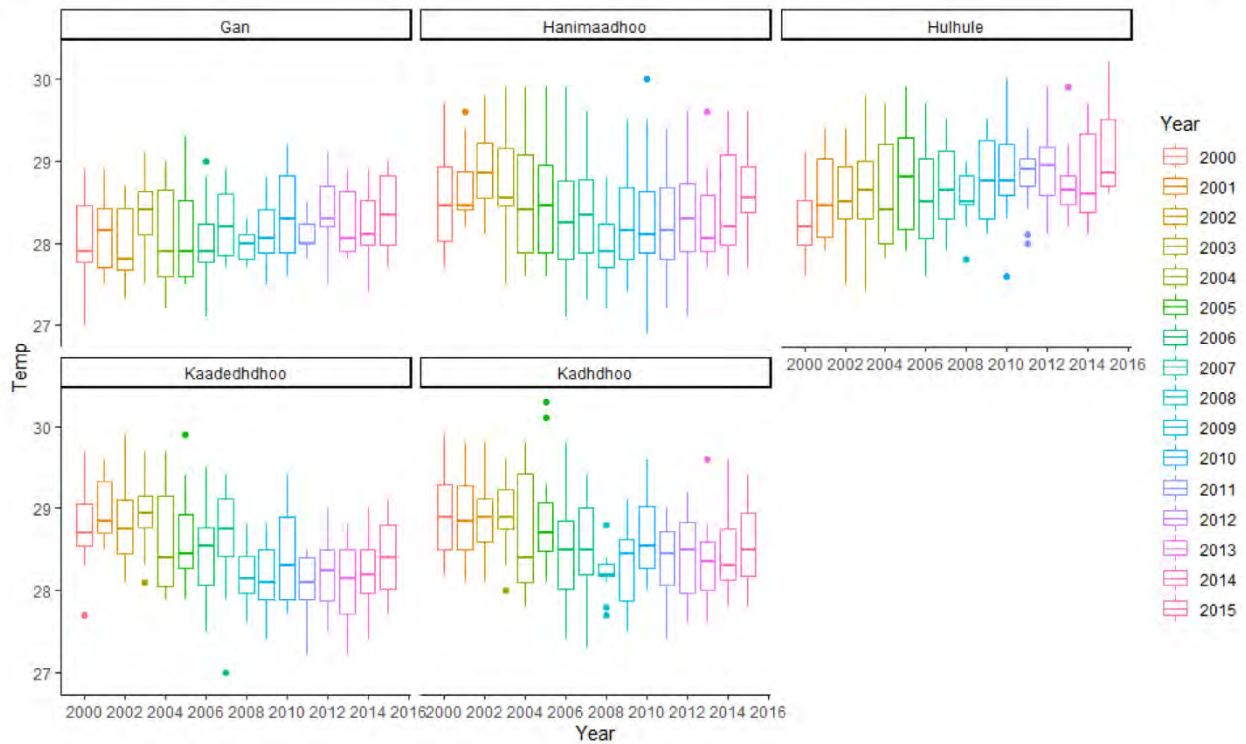


Fig. 8. Annual temperature records (for individual MET stations).

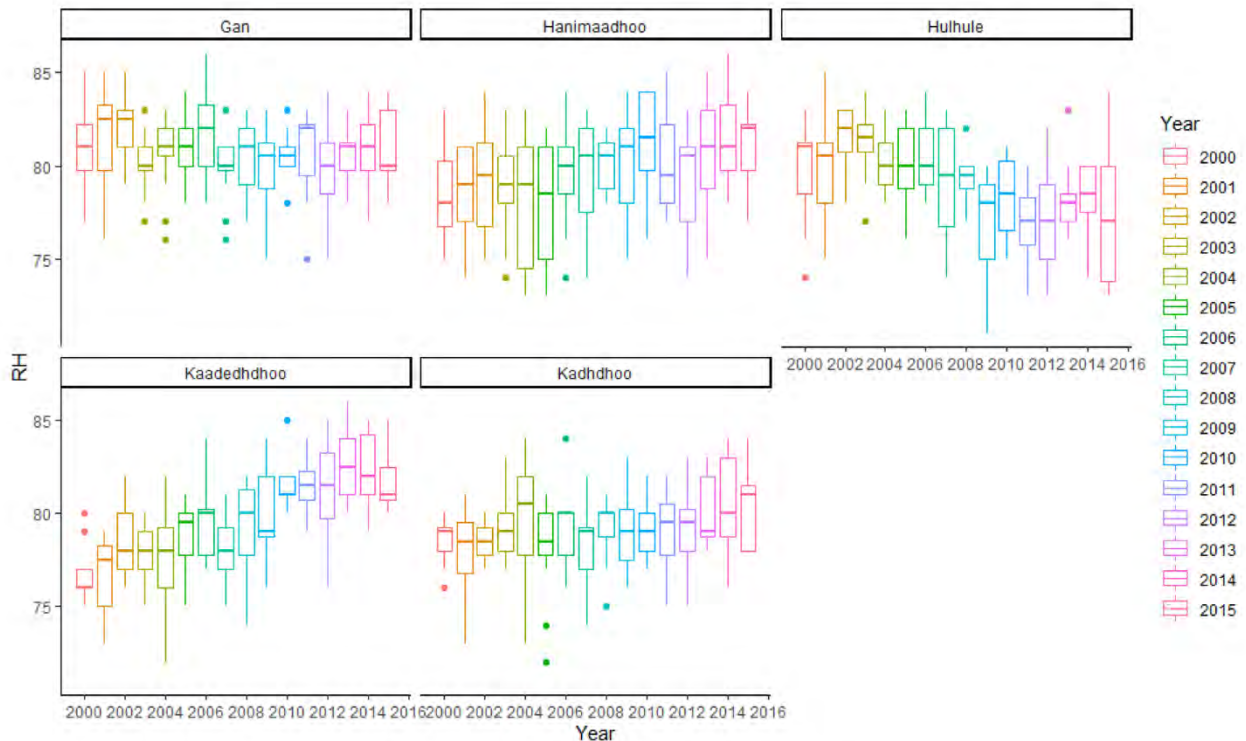


Fig. 9. Annual RH records (for individual MET stations).

experienced similar relatively low atmospheric pressure. It is noteworthy to mention in this context that annual variations are relatively low.

In the case of temperature, Fig. 17 shows an increase in temperature for the entire territory, at the beginning of the Tsunami period (2003) which returns to low values in 2004 which are maintained until 2009, when the central region of the country

recorded higher temperature compared to previous years. The same trend persists, and the central region experienced maximum values temperature in 2015.

Finally, for RH, records in 2000 and 2003 are similar. But in the Tsunami year an increase in value is observed throughout the study region. 2005 experienced a similar pattern as in 2000. In 2009 there is an increasing trend in the north and south of

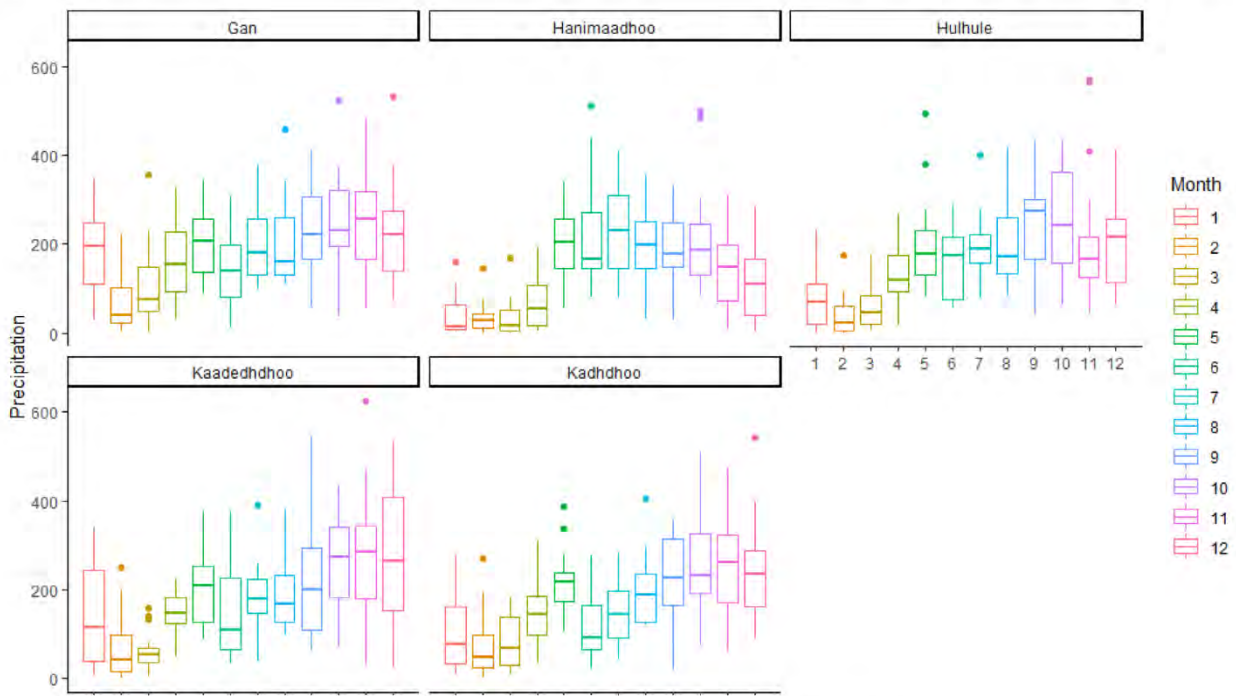


Fig. 10. Monthly precipitation records (for individual MET stations).

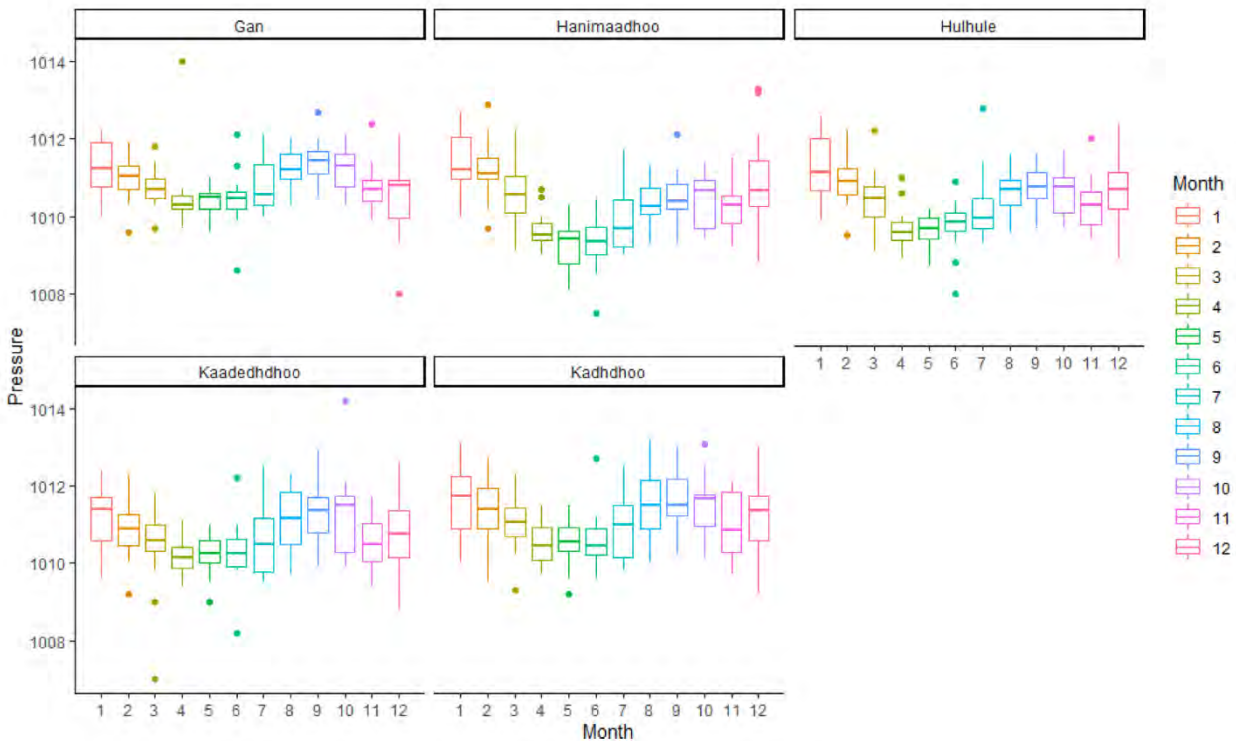


Fig. 11. Monthly atmospheric pressure records (for individual MET stations).

the territory with maximum values in these same areas in 2015 (Fig. 18).

In general, the maximum values appear in 2015 and the lowest in 2000. More specifically related to precipitation pattern and Tsunami, the study shows that the central islands of the Maldives have relatively less rainfall in the year of the Tsunami (2004) than the rest of the region. In addition, in the same year the islands most affected by this natural catastrophe (northern Maldives

region) (Naylor, 2015) experienced much more precipitation than the central region (Fig. 15).

### 5. Discussion

The results presented above allow us to explore the climatic variation in the island nation of Maldives and provide a foundation for further studies of this region. In fact, the current study contributes to the relatively small amount of literature on

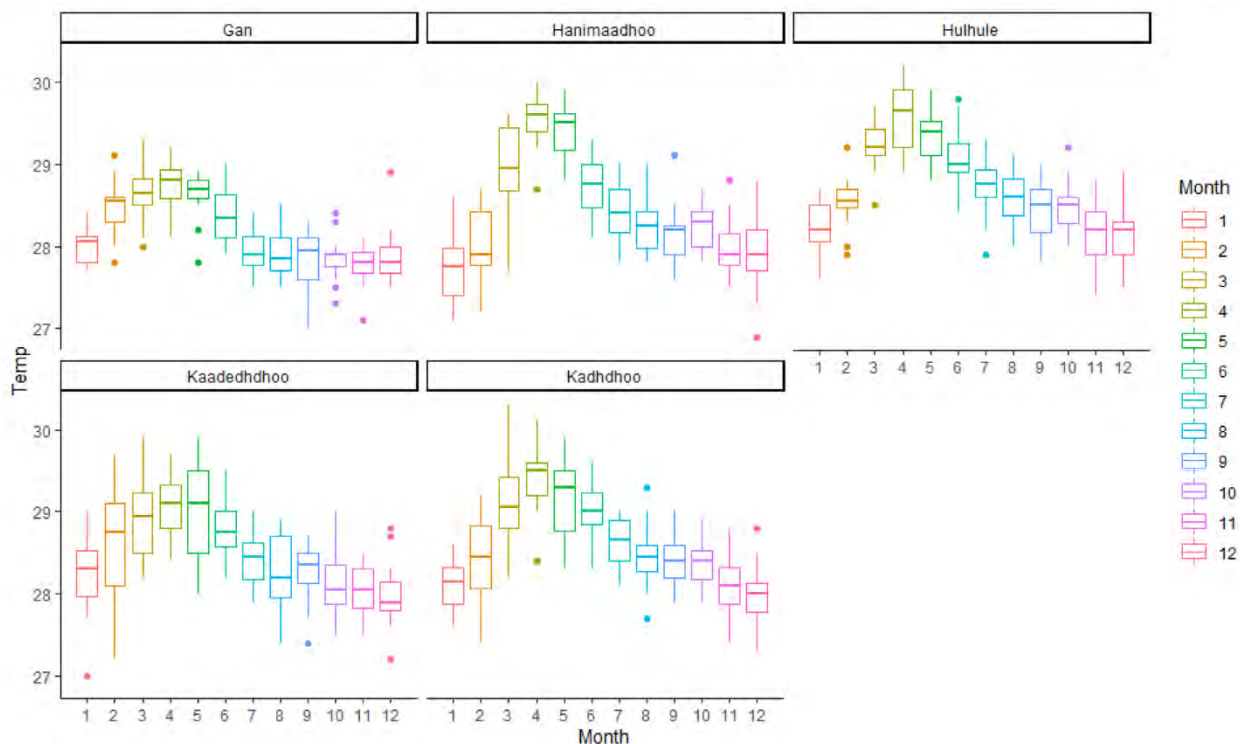


Fig. 12. Monthly temperature records (for individual MET stations).

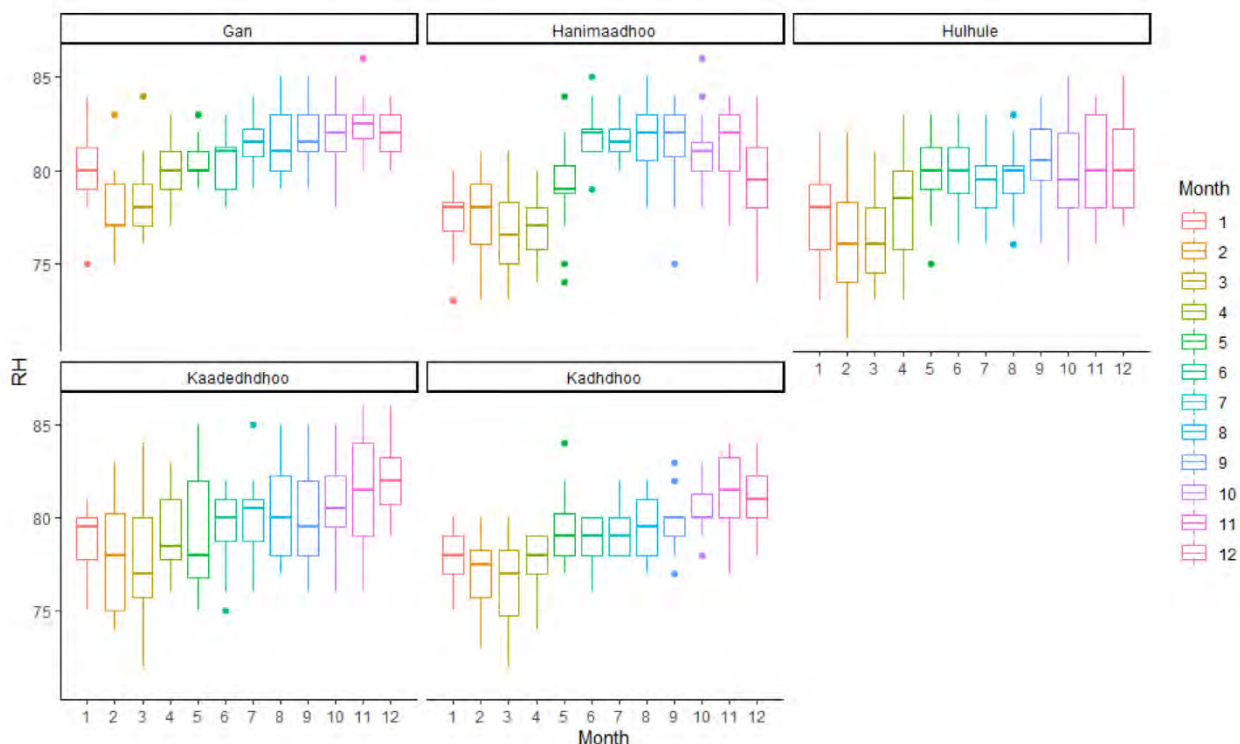


Fig. 13. Monthly RH (for individual MET stations).

seasonal patterns in climate for dispersed island nations, with emphasis on the effect of extreme climatic conditions and natural disasters like Tsunamis. In general, the application of geostatistical tools like kriging is more appropriate for implementation in areas with continuous land surfaces. There is limited research where kriging has been implemented in dispersed land

surface (Cantet, 2015). The current study suggests the appropriate application of geostatistical techniques to analyze the spatial distribution of meteorological parameters in complex island structures.

Since 1972, Maldives has been a major global tourist destination. Tourism in this pristine archipelago is categorized by water sports, sunbathing, and street shopping. Thus, an understudied



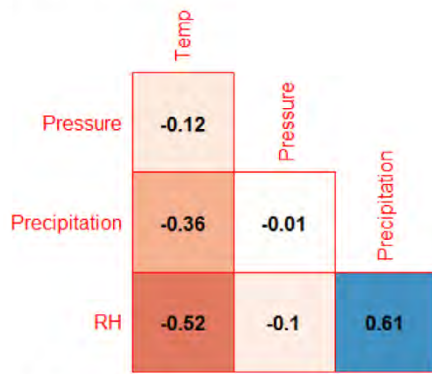


Fig. 14. Correlations among meteorological parameters.

and unpredicted seasonal climate pattern of the country can have an impact on the tourist inflow. As tourism is the largest economic industry in Maldives, it might affect the overall economy of the nation. The present study focuses on analyzing and identifying spatial and seasonal patterns in climate especially for temperature and precipitation across different atolls of Maldives. This can help in strategic tourism management and promotion for the country. Literature shows there are few research works using geostatistical tools to analyze spatial and temporal patterns of different meteorological factors in island nations (Agou et al., 2019; Longobardi et al., 2016). Thus, the current study illustrates that geostatistical techniques can be applied effectively in complex island settings. Also, this initial step is important for future precise analytical tools of clustering like finite mixture models (Scrucca et al., 2016) to further analyze and identify climatic patterns among different atolls.

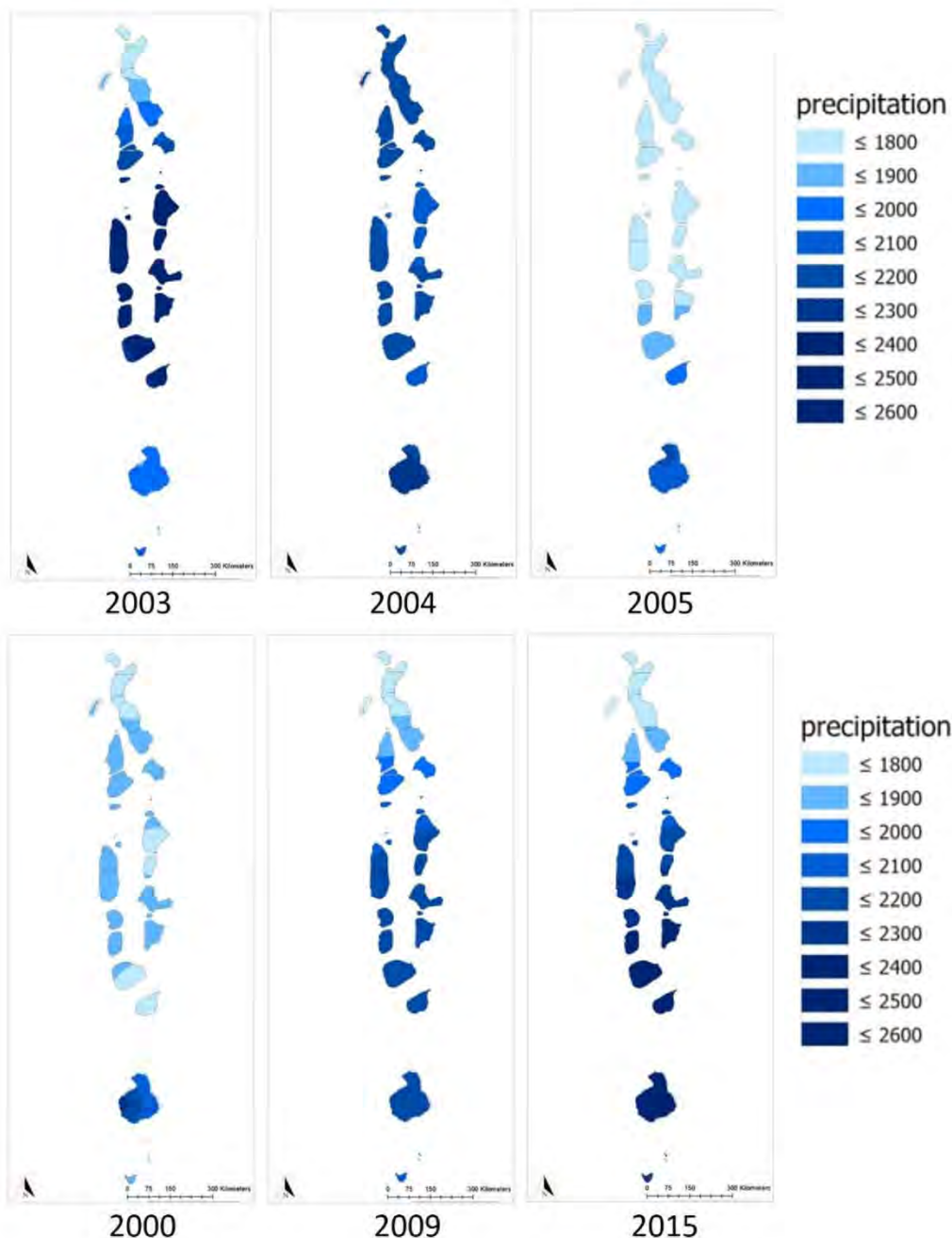


Fig. 15. Kriging result of precipitation. Top: Tsunami years, 2003, 2004, 2005. Down: 2000 (before Tsunami), 2009 and 2015 (after Tsunami).

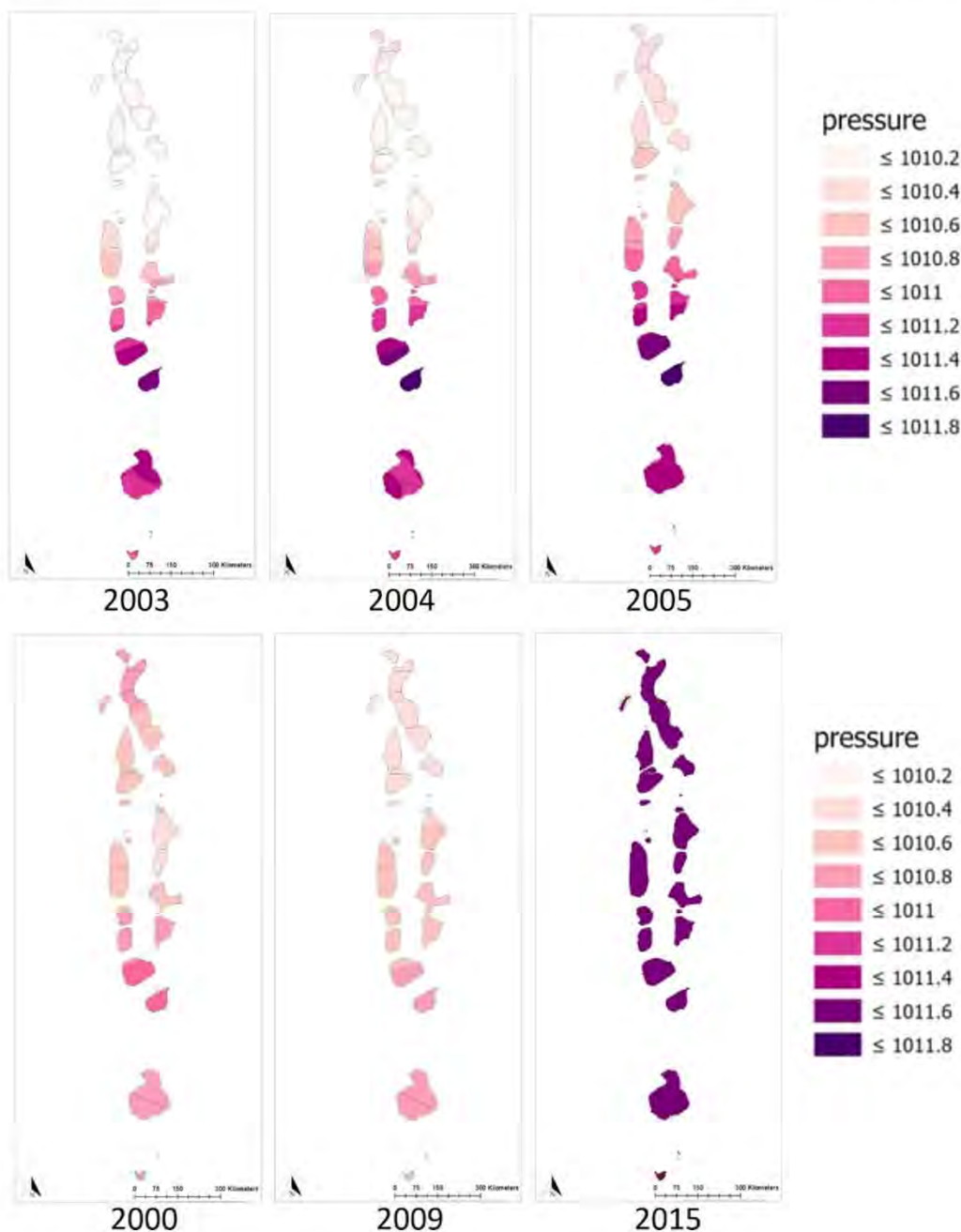


Fig. 16. Kriging result of atmospheric pressure. Top: Tsunami years, 2003, 2004, 2005. Down: 2000 (before Tsunami), 2009 and 2015 (after Tsunami).

A few limitations need to be recognized in the current study. Maldives has overall five meteorological stations under MMS located in different atolls covering the full stretch from north to south of the territory. To carry out this study we had to work with this limited number of meteorological stations. In general, kriging precision can be improved with increasing data density and more regular spatial distribution (Krige, 2015). However, in this case the limited number of meteorological stations indicates low spatial resolution for kriging. Higher spatial distribution would have helped to better analyze and identify the spatio-temporal variation of the meteorological parameters in the study region. However, analyzing the optimal number of stations and determining their ideal positions is beyond the objectives of the current study. Related literature (Lane et al., 2013; Nunn and Kumar, 2017) have analyzed dispersed islands having complex

land structures, but none of them have studied the optimal number or locations of weather stations. Thus, the limited number of meteorological stations is a limitation of this study. After analyzing the current results, it can be stated that an increase in the number of meteorological stations distributed over the entire study area would improve the kriging performance and help in precise prediction. Therefore, further research along these lines, to identify the optimal spatial distribution of meteorological stations to implement kriging and similar geostatistical tools, can provide better understanding of the climate pattern for all the atolls of Maldives. On the other hand, the temporal resolution of four meteorological parameters (precipitation, temperature, atmospheric pressure and RH) for a study period of sixteen years is sufficient to conduct the study. The substantial time frame of

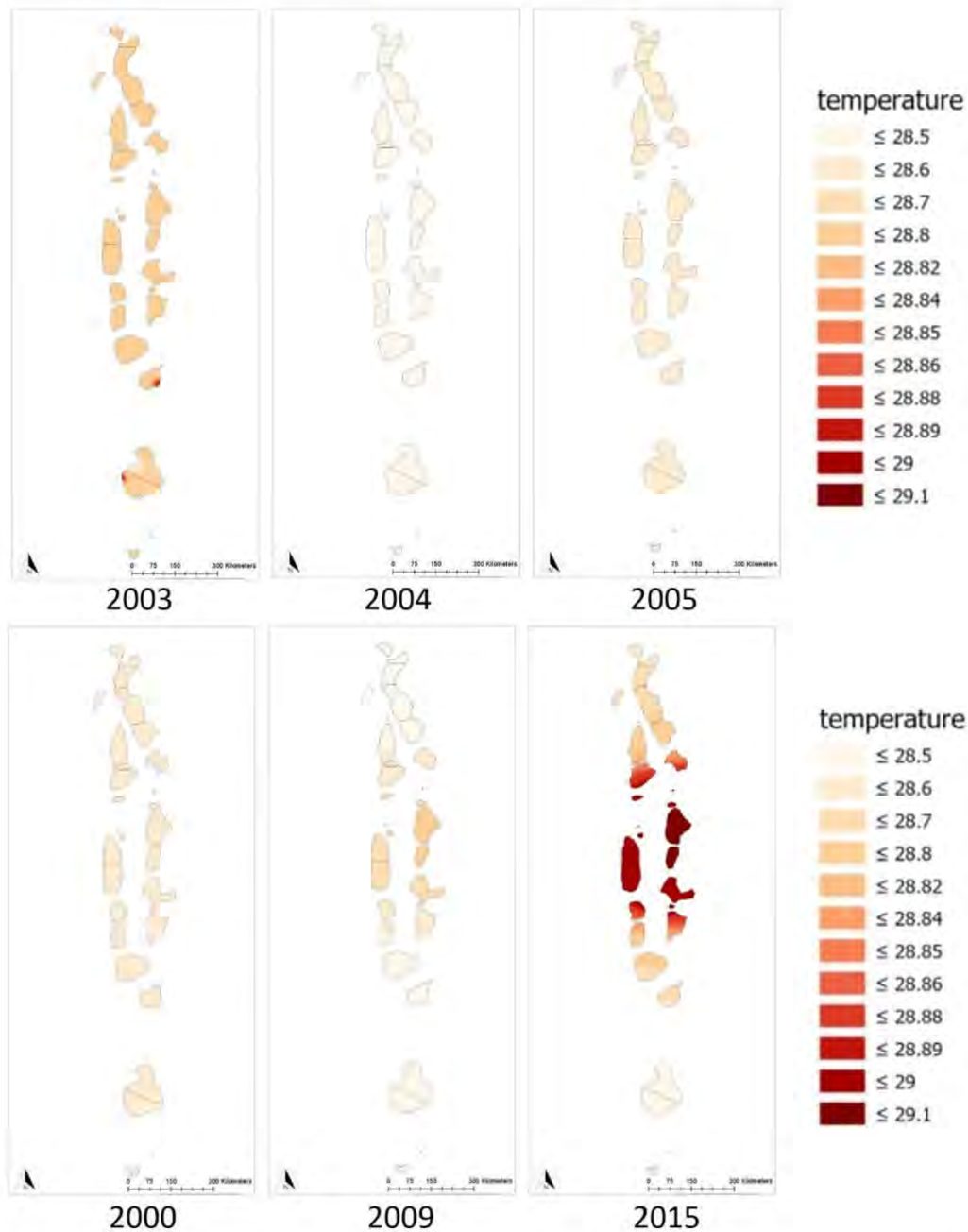


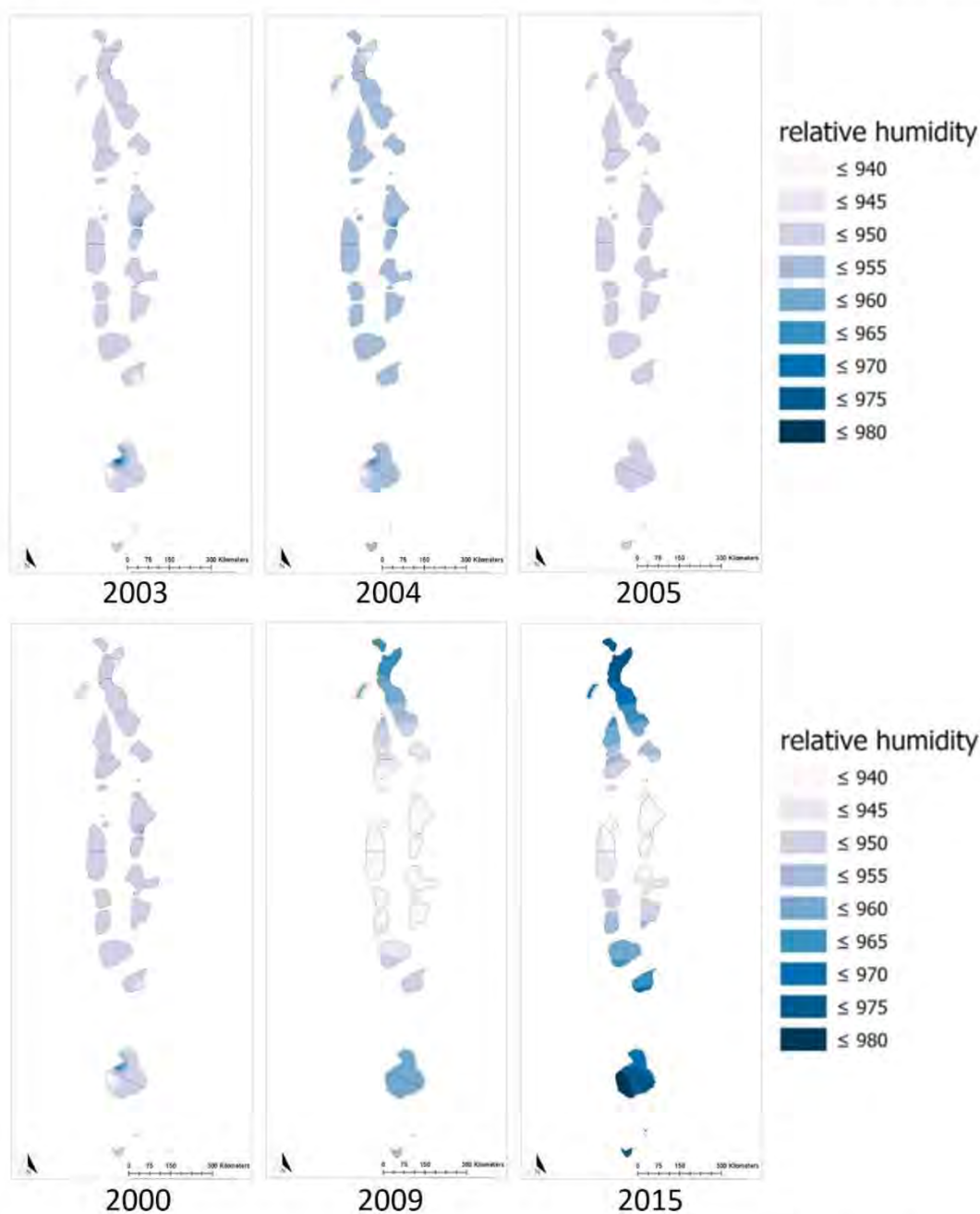
Fig. 17. Kriging result of temperature. Top: Tsunami years, 2003, 2004, 2005. Down: 2000 (before Tsunami), 2009 and 2015 (after Tsunami).

the dataset facilitates performing initial investigations to identify distinct patterns, to diagnose hypotheses and to examine assumptions.

## 6. Conclusions

Among all the elements discussed in the current study, it is worth noting the distinct spatio-temporal variability of the meteorological parameters presented. Maldives differs from the regular seasonal behavior observed in other areas where winters, summers and rainy seasons are distinctly identified annually. The common periodic seasonal pattern is clearly missing in the current study area. In other words, the periodicity is not observed during the study period of 2000 to 2015, making it difficult to predict the seasonal impact on socio-economic factors of Maldives,

especially on the tourism industry. For this reason, a clear and concise description of four important meteorological parameters with emphasis on precipitation and temperature of the region, is explored. In addition, possible prediction maps are depicted for the areas where there are no sampling nodes available. The analysis shows low seasonality in the region, something that future researchers can study in more detail to analyze its impact on tourism and the economy in the region. Moreover, the meteorological description in the time frame before, during and after the Tsunami is another addition for future researchers to analyze if there is any impact on the seasonality because of natural disasters like Tsunami. On the other hand, the complex island structure of the atolls of Maldives acts as a challenge to implement spatial and spatio-temporal geostatistical techniques in the region. The complication in finding precise climatic patterns in this dispersed



**Fig. 18.** Kriging result of RH. Top: Tsunami years, 2003, 2004, 2005. Down: 2000 (before tsunami), 2009 and 2015 (after Tsunami).

land structure, together with the drawback of having limited number of meteorological stations, especially while performing geostatistical analysis provides additional reasons for scientific researchers to conduct future research works using innovative analytical techniques. In summary, the current description of the Maldives islands region highlights the complexity in the land structure and the non-periodicity of the climatic patterns. This may complicate the process of strategic planning in terms of tourism management and development. Thus, from this point, an accurate cluster analysis would allow determining the spatially distributed regions grouped by similar seasonal patterns.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

The authors would like to thank Mr. Hussain Waheed, Assistant Meteorologist, Maldives Meteorological Service, Hulhule, Republic of Maldives for his invaluable help in providing the meteorological data for the study years. The authors gratefully acknowledge the Maldives Meteorological Service (MMS), Republic of Maldives for their kind support. The authors acknowledge and thank Dr. Douglas Schofield (Professor and Senior Advisor, Royal Thimphu College, Bhutan) for his valuable suggestions in technical editing of the manuscript. The authors are grateful to the editor and three referees for constructive comments.

#### References

Agou, V.D., Varouchakis, E.A., Hristopoulos, D.T., 2019. Geostatistical analysis of precipitation in the island of Crete (Greece) based on a sparse monitoring

- network. *Environ. Monit. Assess.* 191, 353.
- ArcGIS Pro Resources | Tutorials, Documentation, Videos & More, 2020. Retrieved April 26, 2020, from <https://www.esri.com/en-us/arcgis/products/arcgis-pro/resources>.
- Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future köppen-geiger climate classification maps at 1-km resolution. *Sci. Data* 5, 1–12.
- Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 36, 192–236.
- Bland, A., Konar, B., Edwards, M., 2019. Spatial trends and environmental drivers of epibenthic shelf community structure across the Aleutian Islands. *Cont. Shelf Res.* 175, 12–29.
- Boer, E.P.J., De Beurs, K.M., Hartkamp, A.D., 2001. Kriging and thin plate splines for mapping climate variables. *ITC J.* 3 (2), 146–154.
- Bostan, P.A., Heuvelink, G.B.M., Akyurek, S.Z., 2012. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *Int. J. Appl. Earth Obs. Geoinf.* 19 (1), 115–126.
- Brushett, B.A., Allen, A.A., Futch, V.C., King, B.A., Lemckert, C.H., 2014. Determining the leeway drift characteristics of tropical Pacific island craft. *Appl. Ocean Res.* 44, 92–101.
- Cantet, P., 2015. Mapping the mean monthly precipitation of a small island using kriging with external drifts. In: *Theoretical and Applied Climatology*. Springer, <http://dx.doi.org/10.1007/s00704-015-1610-z>. hal-01206851.
- Climate of Maldives, 2020. <https://www.meteorology.gov.mv/climateofmaldives/>. Accessed 24 September 2020.
- Cressie, N.A.C., 1993. *Statistics for spatial data*. *J. Agric. Biol. Environ. Stat.*
- Jordan, M.M., Navarro-Pedreño, J., García-Sánchez, E., Mateu, J., Juan, P., 2004. Spatial dynamics of soil salinity under arid and semi-arid conditions: geological and environmental implications. *Environ. Geol.* 45, 448–456.
- Juan, P., Mateu, J., 2009. *Geostatística Espacial. Técnicas Espectrales Con Aplicaciones*. Editorial VERLAG.
- Kabir, M.A., Salaudinn, M., Hossain, K.T., Tanim, I.A., Saddam, M.M.H., Ahmad, A.U., 2020. Assessing the shoreline dynamics of Hatiya Island of Meghna estuary in Bangladesh using multiband satellite imageries and hydro-meteorological data. *Reg. Stud. Mar. Sci.* 35.
- Krige, D.G., 2015. A statistical approach to some mine valuation and allied problems on the Witwatersrand.
- Lane, D., Clarke, C.M., Forbes, D.L., Watson, P., 2013. The Gathering Storm: managing adaptation to environmental change in coastal communities and small islands. *Sustain. Sci.* 8, 469–489.
- Longobardi, A., Buttafuoco, G., Caloiero, T., Coscarelli, R., 2016. Spatial and temporal distribution of precipitation in a mediterranean area (Southern Italy). *Environ. EarthSci.* 75, 189–208.
- Maldives Meteorological Service, 2020. Retrieved April 26, 2020, from <http://www.meteorology.gov.mv/>.
- Matheron, G., 1962. *Traité de Géostatistique Appliquée: Mémoires du Bureau de Recherches Géologiques et Minières*. Editions Technip, Paris, 14, 333.
- Ministry of Environment and Energy, 2017. *State of the environment 2016*. In: Ministry of Environment and Energy.
- Naseer, A., 2007. Pre-and post-tsunami coastal planning and land-use policies and issues in the Maldives.
- Naylor, A.K., 2015. Island morphology, reef resources, and development paths in the Maldives. *Prog. Phys. Geogr.: Earth Environ.* 39 (6), 728–749.
- Nunn, P., Kumar, R., 2017. Understanding climate-human interactions in Small Island Developing States (SIDS). *Int. J. Clim. Change Strateg. Manage.* 10 (2), 245–271.
- R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, Retrieved April 26, 2020, from <http://www.R-project.org/>.
- Riyas, C.A., Idreesbabu, K.K., Marimuthu, N., Sureshkumar, S., 2020. Impact of the tropical cyclone ockhi on ecological and geomorphological structures of the small low-lying Islands in the Central Indian Ocean. *Reg. Stud. Mar. Sci.* 33.
- RStudio | Open source & professional software for data science teams - RStudio, 2020. Retrieved April 26, 2020, from <https://rstudio.com/>.
- Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 8 (1), 289–317.
- Serra, L., Juan, P., Varga, D., 2017. *Spatial prediction and mapping temperature*. In: *Classical Kriging and INLA*. Lambert Academic Publishing.
- Shifaza, F., 2018. *Statistical pocketbook of Maldives 2018*. Retrieved from <http://statisticsmaldives.gov.mv/nbs/wp-content/uploads/2019/01/Statistical-Pocketbook-of-Maldives-2018-Printing.pdf>.
- Staniec, A., Vlahos, P., 2017. Timescales for determining temperature and dissolved oxygen trends in the Long Island Sound (LIS) estuary. *Cont. Shelf Res.* 151, 1–7.
- Weather-atlas, 2020. <https://www.weather-atlas.com/en/maldives-climate>. Accessed 24 September 2020.

## **8.3 Article 5: Trend Detection on Linear Networks**

### **On the trend detection of time-ordered intensity images of point processes on linear networks**

Somnath Chaudhuri<sup>1,2</sup>, Mehdi Moradi<sup>3</sup> and Jorge Mateu<sup>4</sup>

1. Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain.
2. CIBER of Epidemiology and Public Health (CIBERESP), Spain.
3. Department of Statistics, Computer Science, and Mathematics, and Institute of Advanced Materials and Mathematics (InaMat), Public University of Navarre, Pamplona, Spain.
4. Department of Mathematics, University Jaume I, Castellón, Spain.



## On the trend detection of time-ordered intensity images of point processes on linear networks

Somnath Chaudhuri, Mehdi Moradi & Jorge Mateu

To cite this article: Somnath Chaudhuri, Mehdi Moradi & Jorge Mateu (2021): On the trend detection of time-ordered intensity images of point processes on linear networks, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2021.1881116](https://doi.org/10.1080/03610918.2021.1881116)

To link to this article: <https://doi.org/10.1080/03610918.2021.1881116>



Published online: 09 Feb 2021.



Submit your article to this journal [↗](#)



Article views: 67



View related articles [↗](#)



View Crossmark data [↗](#)



# On the trend detection of time-ordered intensity images of point processes on linear networks

Somnath Chaudhuri<sup>a</sup>, Mehdi Moradi<sup>b</sup> , and Jorge Mateu<sup>c</sup>

<sup>a</sup>Institute of New Imaging Technologies (INIT), GEOTEC, University Jaume I, Castellón, Spain; <sup>b</sup>Department of Statistics, Computer Science, and Mathematics, and Institute of Advanced Materials and Mathematics (InaMat<sup>2</sup>), Public University of Navarre, Pamplona, Spain; <sup>c</sup>Department of Mathematics, University Jaume I, Castellón, Spain

## ABSTRACT

Spatial point processes on linear networks are increasingly getting attention in different disciplines such as traffic accidents and street crime analysis. Dealing with a set of time-ordered point patterns on a linear network over a period, helps in obtaining a time series of estimated intensity images. In this article, we combine the problem of estimating the intensity and relative risk of point patterns on linear networks with trend detection in time-ordered observations. Taking the temporal autocorrelation between consecutive time-ordered intensity and relative risk images into account, we make use of the Mann–Kendall trend test to look for potential locations in the network where the estimated intensity and/or relative risk show evidence of a monotonic trend. The monthly time-ordered spatial point patterns of fatal traffic accidents and street crimes in the city of London, UK, in the period of January 2013 to December 2017, are used as an application.

## ARTICLE HISTORY



Received 25 March 2020  
Accepted 20 January 2021

## KEYWORDS

Mann–Kendall trend test;  
Relative risk; Separability;  
Spatio-temporal data; Street  
crime; Traffic accident

## 1. Introduction

The analysis of spatial point patterns on linear networks, e.g. the location of traffic accidents or street crimes, is increasingly receiving scientific interest. Since such locations inherently only live on their corresponding network structure, considering such structure as the support of data instead of a general state space might result in defining a more realistic scenario (Yamada and Thill 2004). Nevertheless, geometrical complexities of linear networks give rise to different mathematical/computational challenges. Thus far, most of the attention is paid to estimating the intensity function of such point processes non-parametrically (Okabe, Satoh, and Sugihara 2009; McSwiggan, Baddeley, and Nair 2016; Moradi, 2018; Moradi, Rodriguez-Cortes, and Mateu 2018; Moradi et al., 2019; Rakshit et al., 2019). Regarding traffic accidents or street crimes data, such locations are usually recorded daily, and their incidence rate may be affected by external events such as different activities of the Town-hall or the Police department, and/or environmental characteristics like physical environment, weather, and so forth (Feng et al. 2016; Hipp, Kim, and Kane 2019). The density/intensity of traffic accidents and/or street crimes may face gradual/sudden changes over time. For instance, new strategies to reduce the crime/accident rate in a particular area might push the corresponding intensity down at a particular area. The efficiency of such strategies to reduce the rate of traffic accidents or street crimes might be then detectable when having a set of time-ordered realizations of the underlying point process.

**CONTACT** Mehdi Moradi  mehdi.moradi@unavarra.es; m2.moradi@yahoo.com  Department of Statistics, Computer Science, and Mathematics, and Institute of Advanced Materials and Mathematics (InaMat<sup>2</sup>), Public University of Navarre, Pamplona, Spain.

© 2021 Taylor & Francis Group, LLC



The problem of trend/change-point detection has been frequently raised within different disciplines such as agronomy, hydrology, geology, climatology, etc. Several proposals have been developed for detecting gradual/sudden distributional changes in time-ordered datasets including non-parametric, parametric, and regression-based methods (Mann 1945; Kendall 1948; Cox and Stuart 1955; Pettitt 1979; Zeileis et al. 2003; Matteson and James 2014; Grundy, Killick, and Mihaylov 2020). A selective review of several change-point detection methods is provided by Truong, Oudre, and Vayatis (2020). The developed proposals were initially considered for time series, and later they are examined for time series of satellite images (Verbesselt et al. 2010; Bullock, Woodcock, and Holden 2020; Militino, Moradi, and Ugarte 2020). In general, since time-ordered datasets usually experience a seasonal behavior, there also exists a technique to decompose time series into trend, seasonal, and remainder components looking for possible changes in both trend and seasonal components individually (Verbesselt et al. 2010). Although these methods demonstrate a reasonably high power of the test, the majority drastically suffer from a high rate of introducing false positives when dealing with highly autocorrelated data. Several modifications have been proposed to reduce the type I error probability of the Mann–Kendall test in the presence of temporal autocorrelation (Kulkarni and von Storch 1995; Hamed and Rao 1998; Von Storch 1999; Yue et al. 2002; Yue and Wang 2004; Hamed 2009). Nevertheless, their major drawback is to reduce the power of the test along with reducing the type I error probability. Note that reducing the power of the test means increasing the type II error probability. It is shown that taking a tradeoff between the type I and II error probabilities into account, the original Mann–Kendall method might still be a reliable technique (Militino, Moradi, and Ugarte 2020).

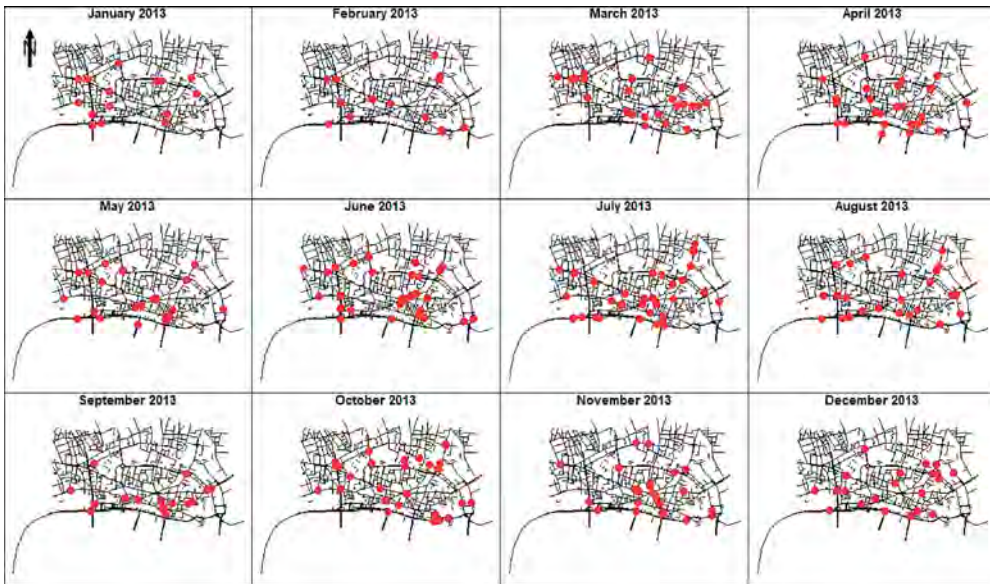
Although, in practice, this might often be the case to have a set of time-ordered point patterns on a linear network, to the best of our knowledge this field has not yet benefited from trend detection techniques. In this article, we focus on an application of trend detection in the time series of estimated intensities and relative risk images of spatial point patterns on linear networks. Two sets of monthly time-ordered point patterns of fatal traffic accidents and street crimes, in the period of January 2013 to December 2017, in the city of London, UK, are used for this purpose. Each point pattern represents the locations of the events in a particular month. Taking the temporal autocorrelation degree of such time-ordered estimated intensities and relative risk images into account, we make use of the multivariate/univariate Mann–Kendall test (Mann 1945; Kendall 1948; Militino, Moradi, and Ugarte 2020) to look for potential locations where the estimated intensity and/or relative risk show evidence of monotonic trend.

The rest of the article is organized as follows. In [Sec. 2](#) we present the time-ordered spatial point patterns of traffic accidents and street crimes in the city of London, UK. [Section 3](#) provides a summary about point processes on linear networks together with their intensity and relative risk estimators. In [Sec. 4](#) we briefly present some details of the Mann–Kendall trend detection test. [Section 5](#) is devoted to present the results of the traffic accidents and street crimes data analysis. The article ends with a summary in [Sec. 6](#).

## 2. Data

In this section we present two time-ordered sets of monthly spatial point patterns of traffic accidents and street crimes in the city of London, UK, from January 2013 to December 2017. The city of London has an area of 2.90 km<sup>2</sup>, with an approximate population of 8000 people, and comprises six lower layer super output area (LSOA). The number of people who commute into the city daily for work exceeds 5,00,000, with over 10 million visits as tourists yearly. The area is an important local government district of UK that contains the historic center and the primary Central Business District (CBD) of London.

The street crime data contain 18,908 records for the study period including antisocial behavior, bicycle theft, drug-related, public disorder and weapons, public order, robbery, shoplifting, theft



**Figure 1.** Monthly spatial point patterns of fatal traffic accidents in the city of London, UK, during 2013.

from the person, vehicle-related crime, violence and sexual offenses, and violent crime. Antisocial behavior comprises the maximum percentage of records (26.79%) followed by violence and sexual offenses (19.06%), and shoplifting (16.97%). Amongst all types of crimes, only antisocial behavior, shoplifting, vehicle-related, and drug-related crimes appeared in all months. Regarding the traffic accident data, it contains 1678 observations all being fatal accidents having at least one causality count. We note that most accidents (90.58%) are having only one causality. As showcases, the locations of traffic accidents and street crimes for the year 2013 are shown in [Figures 1](#) and [2](#).

We note that the road network is accessed from open street map (OSM) repository using the R package `osmdata` (Padgham et al. 2017). OSM data is free and licensed under the open data commons open database license (ODbL) by the OpenStreetMap Foundation (OSMF)<sup>1</sup>. Initially, complete OSM street network for the entire study area has been retrieved using primary tag `highway` (used for any category of streets). Then, less important OSM highway categories such as `unclassified`, `bus guideway`, `path`, `raceway`, `escape`, and `bridleway` are not included in the current study. In fact, these categories are not used for usual traffic, and thus they do not host any event. Both data retrieval and cleaning has been performed using the same R package `osmdata`.

The traffic accident dataset is published by the Department of Transport, government of UK, under the UK government open data project<sup>2</sup>. The street crime data is provided by 43 geographic police forces in the UK and Wales, the British Transport Police, the Police Service of Northern Ireland and the Ministry of Justice, and the government of UK. Both the traffic accidents and street crimes datasets are free and licensed under the Open Government License v3.0 for public sector information, government of UK<sup>3</sup>.

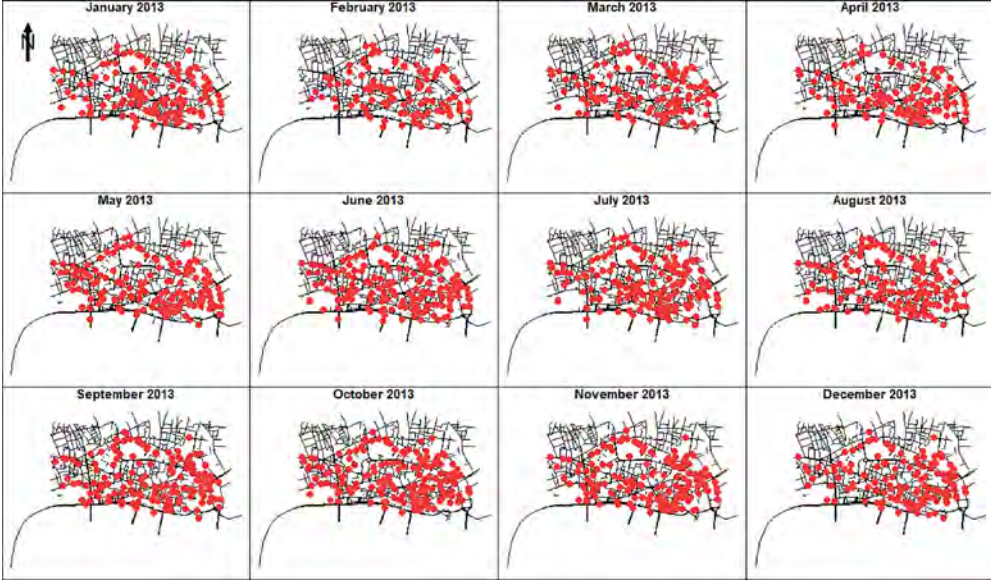
### 3. Point processes on linear networks

Throughout the article, we consider  $X$  as a simple spatial point process on the linear network  $L \subset \mathbb{R}^2$ , which is a union of some finite number of segments  $l_i = [u_i, v_i] = \{tu_i + (1-t)v_i : 0 \leq$

<sup>1</sup><https://www.openstreetmap.org/copyright>

<sup>2</sup><https://data.gov.uk>

<sup>3</sup><http://www.nationalarchives.gov.uk/doc/open-government-license/version/3>



**Figure 2.** Monthly spatial point patterns of street crimes in the city of London, UK, during 2013.

$t \leq 1\} \subset \mathbb{R}^2, 1 \leq i < \infty$ . We do not set any restriction regarding the connectivity of the network or the kind of intersection between different segments. The distance between any two points  $u, v \in L$  is denoted by  $d_L(u, v)$ . For any subnetwork  $A \subset L$ , its total length is obtained by summing the length of all its corresponding segments and is denoted by  $|A|$ . For any measurable function  $f : L \rightarrow [0, \infty)$ , the Campbell formula states that:

$$\mathbb{E} \sum_{x \in X} f(x) = \int_L f(u) \lambda(u) d_1 u,$$

where  $\lambda(\cdot)$  is called the intensity function of  $X$  governing its distribution over  $L$ , and  $d_1$  stands for integration with respect to arc length. In particular,

$$\mathbb{E}[\#(A \cap X)] = \int_A \lambda(u) d_1 u, \quad A \subset L,$$

where  $\#(A \cap X)$  denotes the number of points of  $X$  falling in  $A$ . If  $\lambda(u) \equiv \lambda$ , then  $X$  is called a homogeneous point process, otherwise it is said to be an inhomogeneous point process (Ang, Baddeley, and Nair 2012; Baddeley, Rubak, and Turner 2015).

Due to the geometrical complexities of linear networks, estimating the intensity function  $\lambda(\cdot)$  has been quite challenging. Nevertheless, several proposals have been developed including some network-distance kernel-based smoothing methods (Okabe, Satoh, and Sugihara 2009; McSwiggan, Baddeley, and Nair 2016; Moradi, 2018; Moradi, Rodriguez-Cortes, and Mateu 2018), the two-dimensional convolution-based kernel intensity estimators (Rakshit et al. 2019), and the resample-smoothed Voronoi intensity estimator (Moradi et al. 2019). Consider  $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$  as a realization of point process  $X$  on  $L$ , the two-dimensional convolution-based kernel intensity estimator, with uniform corrections, is of the form:

$$\hat{\lambda}^U(u) = \frac{1}{c_L(u)} \sum_{i=1}^n \kappa(u - x_i), \quad u \in L, \quad (1)$$

and with Jones-Diggle correction, it is given as

$$\hat{\lambda}^{\text{JD}}(u) = \sum_{i=1}^n \frac{1}{c_L(x_i)} \kappa(u - x_i), \quad u \in L, \quad (2)$$

where  $\kappa$  is a bivariate kernel function, and

$$c_L(u) = \int_L \kappa(u - v) d_1 v,$$

is an edge correction. The Eq. (1) is unbiased if the true intensity  $\lambda(\cdot)$  is constant, and the Eq. (2) provides mass conservation, i.e.  $\int_L \hat{\lambda}^{\text{JD}}(u) d_1 u = n$ . For further details regarding different statistical properties of the Eqs. (1) and (2), and additional details of relative risk see Rakshit et al. (2019).

It is common practice to estimate the spatially-varying relative frequency of each type of events, when there are several types of events occurring on the same network. Assume that two realizations  $\mathbf{x}$  and  $\mathbf{y}$  are observed, on the same network  $L$ , from two different point processes  $X$  and  $Y$ . The relative risk between the two types is then calculated by  $\rho(u) = \log(\lambda_X(u)/\lambda_Y(u))$ ,  $u \in L$ , in which  $\lambda_X(\cdot)$  and  $\lambda_Y(\cdot)$  stand for the intensity functions of  $X$  and  $Y$ , respectively. The literature recommends to estimate both  $\lambda_X(\cdot)$  and  $\lambda_Y(\cdot)$  using a common bandwidth (Kelsall and Diggle 1995; Hazelton 2008; Davies, Jones, and Hazelton 2016). Relative risk for point patterns on linear networks is substantially discussed by Rakshit et al. (2019) and McSwiggan, Baddeley, and Nair (2020).

#### 4. Mann–Kendall trend detection

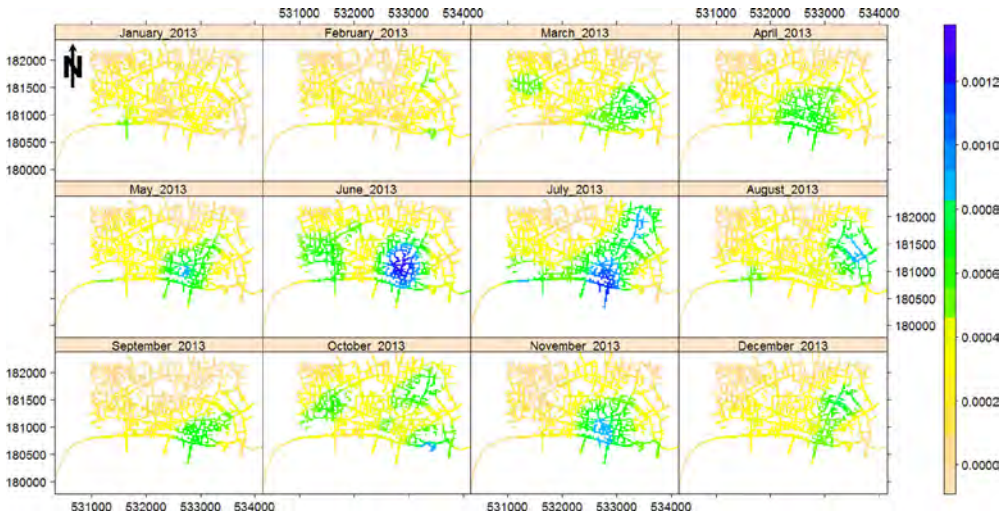
When dealing with observations that appear as ordered in time, a very first thing that might be of interest is to check whether, in the distribution of data, there is any gradual/sudden departure from its past norm. The importance of being aware of such departure, e.g. in model fitting and prediction, has led to the development of several proposals under different settings. Amongst all, the Mann–Kendall trend test has been one of the most frequently used trend tests in the literature (Mann 1945; Kendall 1948; Militino, Moradi, and Ugarte 2020). Generally, for trend detection methods, the null hypothesis  $\mathcal{H}_0$  is that data is independently and randomly ordered, whereas the alternative hypothesis  $\mathcal{H}_1$  claims the existence of a monotonic trend. In other words, the null hypothesis stands with no gradual change in data over time. Considering  $\mathbf{y} = \{y_1, y_2, \dots, y_m\}$ ,  $1 < m < \infty$ , as a finite set of numerical time-ordered observations, the test statistic of the univariate Mann–Kendall is given as

$$S = \sum_{i=1}^{m-1} \sum_{j=i+1}^m \text{sgn}\{y_j - y_i\}, \quad (3)$$

where

$$\text{sgn}\{y_j - y_i\} = \begin{cases} 1, & y_j - y_i > 0, \\ 0, & y_j - y_i = 0, \\ -1, & y_j - y_i < 0. \end{cases}$$

Under the null hypothesis, the expectation and variance of Eq. (3) are  $\mathbb{E}[S] = 0$  and  $\text{Var}[S] = m(m-1)(2m+5)/18$  subject to there being no ties. The test statistic Eq. (3) compares each data point to all data appeared at a later time, looking for any gradual growth/shrinking in the data. Moreover, the so-called (rank correlation) Kendall's  $\tau$  is in close relation with Eq. (3), being calculated as  $S/\binom{m}{2}$  if there is no tie in  $\mathbf{y}$ . Note that positive/negative values of  $S$  are used as indicators of upward/downward trend in  $\mathbf{y}$ . In practice, however, the standardized test statistic  $Z = [\text{sgn}\{S\}(|S| - 1)]/\sqrt{\text{Var}[S]}$  and its corresponding approximate  $p$  value



**Figure 3.** Monthly estimated intensities of the fatal traffic accident data in the city of London, UK, in 2013.

$$p = 2\min(0.5, P(X > |Z|)), \quad X \sim N(0, 1), \quad (4)$$

are used to whether accept or reject the null hypothesis  $\mathcal{H}_0$ . In addition, a multivariate version of the Mann–Kendall test is available for trend detection in a group of time-ordered datasets jointly. This takes information from all individual ones, combines the information and provides a corrected statistics based on the corresponding variance-covariance matrix (Libiseller and Grimvall 2002), and makes decision about the existence of trend in data without pointing to where dominance occurs in case a trend is detected (Pohlert 2018). Looking at the literature, the performance reduction of Mann–Kendall test in the presence of temporal autocorrelation has been frequently highlighted, suffering from a high rate of false positives. The higher the degree of autocorrelation, the higher the type I error probability (Yue et al. 2002). In order to remedy such an issue, several modifications have been developed including pre-whitening techniques (Kulkarni and von Storch 1995; Von Storch 1999; Yue et al. 2002; Hamed, 2009) and variance correction approaches (Hamed and Rao 1998; Yue and Wang 2004). Although these modifications generally reduce/moderate the type I error probability of the Mann–Kendall test, they inevitably decrease the power of the test which means increasing the type II error probability. However, in hypothesis testing a balance between the type I and type II error probabilities is needed. Under different settings, and through a comprehensive simulation study, it is shown that looking for a tradeoff between the type I error probability and the power of the test leads to the original Mann–Kendall test as a reliable and preferable test when data have experienced a monotonic trend (Militino, Moradi, and Ugarte 2020).

## 5. Results

This section is devoted to present the results of trend detection, based on the multivariate/univariate Mann–Kendall test, for the time series of estimated intensity images of fatal traffic accidents and street crimes in the city of London, UK, from January 2013 to December 2017, and also their corresponding time series of relative risk images. Prior to employ the Mann–Kendall test, we need to estimate the intensities and relative risk images. Since we are interested in the temporal evolution in the time series of estimated intensities of time-ordered spatial point patterns, and also to avoid undesirable halo artifacts, we make use of a common bandwidth for each time series of point patterns. Hence, for each such time series, we first select the bandwidth

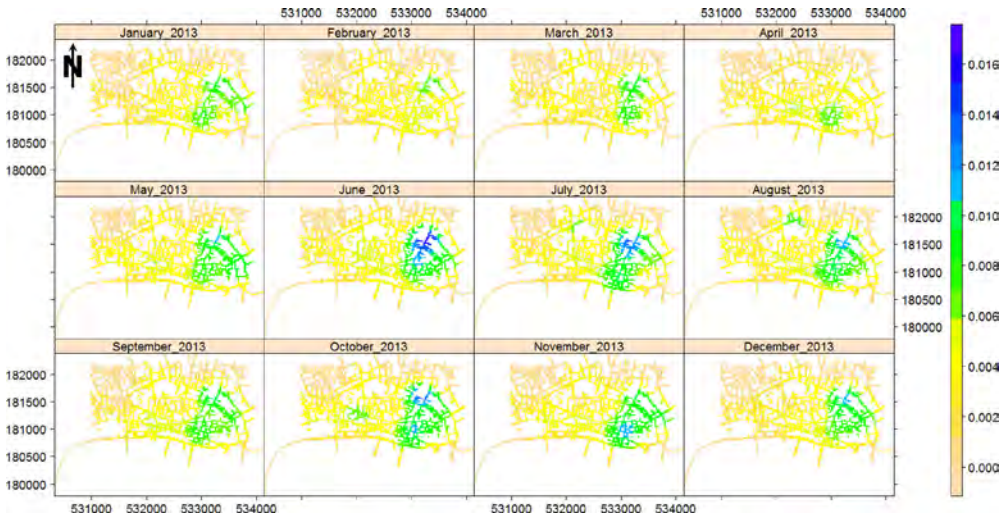


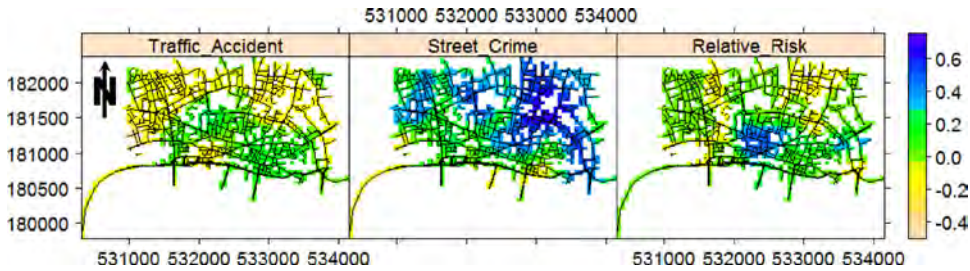
Figure 4. Monthly estimated intensities of the street crime data in the city of London, UK, in 2013.

parameter using the Scott's rule-of-thumb (Rakshit et al. 2019) for individual point patterns, and then use the geometrical mean of such selected bandwidths as a common choice. In the calculation of relative risk images we use the method of Davies (2013) to employ equal bandwidths for both the numerator and denominator for each individual risk, and again we make use of the geometrical mean of the selected bandwidths as a common choice in this case.

Each time series of point patterns contain 60 monthly patterns for which the common selected bandwidths for accidents data and street crime data are 277.19 and 179.12 m, respectively. The considered common bandwidth for relative risk calculation is 215.95 m. Figures 3 and 4 show the monthly estimated intensities, using the uniform edge correction, of fatal traffic accidents and street crimes in the city of London, UK, in 2013 respectively. Clearly, the estimated intensities in both cases vary across time, implying a change in the corresponding time of hot-spots and pointing to some spatial variation in the intensities. This indeed might be a sign of first-order non-separability. We did not overlay the network for a better visualization of spatial/temporal changes in the intensity images.

Before turning to the trend detection problem, we note that for each pixel in Figures 3 and 4, or their corresponding relative risks, there exists a time series of estimated values. We are now interested in trend detection in such time series. Thus, we first deseason the data by creating seasonal anomalies of data (Appelhans, Detsch, and Nauss 2015), and then aggregate it with factor 2. Note that aggregation might reduce the number of potential false positives by smoothing out the estimated intensity images locally.

In order to study the existence of potential trend/change in the time series of estimated intensities and their corresponding relative risk more precisely, we next call the Mann-Kendall trend detection method. Nevertheless, being aware of the effect of temporal autocorrelation in the performance of trend/change-point detection methods, we initially calculate the first lag partial autocorrelation for the (pixel) time series of estimated intensities and relative risk images by fitting autoregressive models to each (pixel) time series of such values. Figure 5 shows how the first lag partial autocorrelation of such time series of images vary over the region. Amongst all, the time series of estimated intensities of street crimes show the highest temporal autocorrelation reaching its maximum in the center and eastern part of the network. The time series of estimated intensities of traffic accidents and relative risk generally show a low degree of temporal autocorrelation having their maximum around the central part of the region. Moreover, their spatial variation does not necessarily follow the same distribution, e.g. a location with high temporal



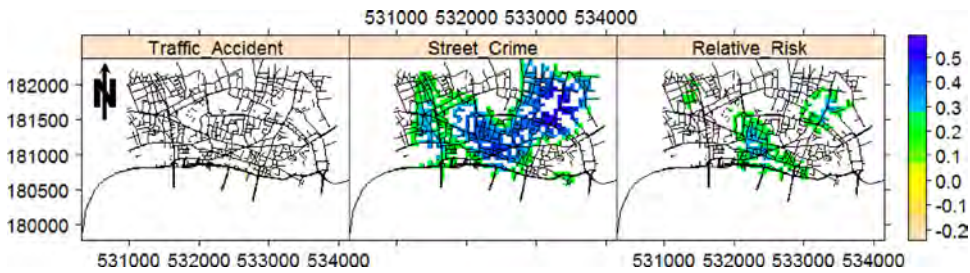
**Figure 5.** First lag partial autocorrelation for the time series of monthly estimated intensity and relative risk images of fatal traffic accident and street crime in the city of London, UK, in the period of January 2013 to December 2017.

autocorrelation in the time series of the estimated intensities of street crime does not necessarily show a high temporal autocorrelation in the time series of the estimated intensities of fatal traffic accidents. Looking back into the literature, areas with quite high temporal autocorrelation are vulnerable to introduce false positives in terms of trend/change-point detection (Serinaldi and Kilsby 2016; Militino, Moradi, and Ugarte 2020).

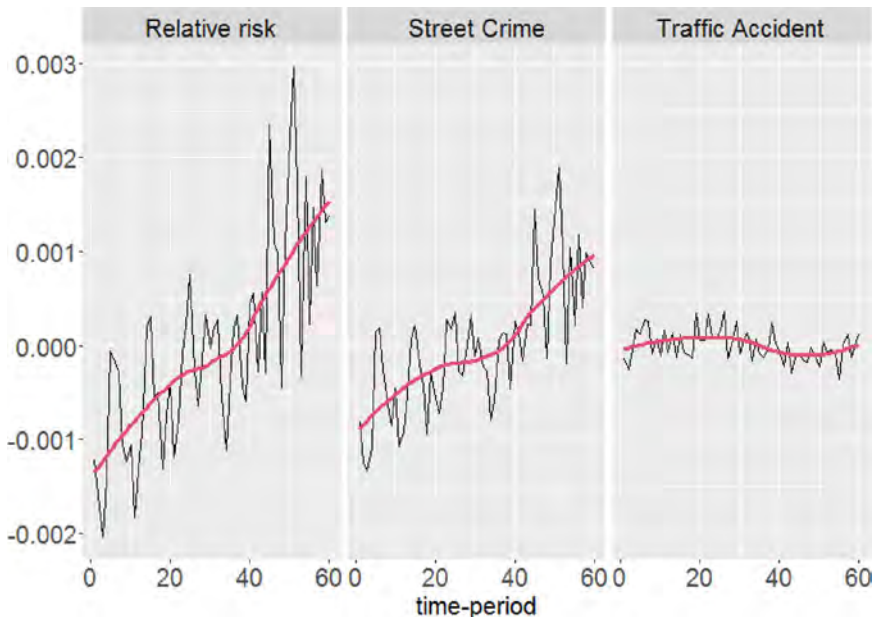
Turning to the trend detection problem, we first make use of the multivariate version of the Mann–Kendall method (Pohlert 2018) to check if there is any major area showing any significant monotonic trend that could dominate the behavior of data in question in the rest of the network. The obtained  $p$  values of multivariate Mann–Kendall for the corresponding time series of fatal traffic accident, street crime, and relative risk are 0.95,  $3.7 \times 10^{-6}$ , and 0.007, respectively. Therefore, the time series of estimated intensities of traffic accident data does not show any evidence of trend. Regarding the street crime data, there is a strong claim on the existence of a monotonic trend, and the time series of the estimated relative risk images also show a monotonic trend. Nevertheless, and in order to get an insight into where dominance occurs, we next employ the univariate Mann–Kendall method. Figure 6 shows the detected segments/pixels/locations in the network of the city of London, where the time series of the monthly estimated intensities of fatal traffic accidents, street crimes, and their corresponding relative risk show a monotonic trend in the period of January 2013 to December 2017. Apparently, the fatal traffic accident data does not show a particular trend in the network, apart from a very small area in the center of the southernmost street that shows a downward trend. The street crime dataset, however, generally shows an upward trend in many of the western, central, and northeastern streets. The time series of relative risk images shows three major areas with upward trend, in the (southern) center, northwest, and northeast of the network.

Looking into Figures 5 and 6 simultaneously, it is seen that detected areas with significant trend in Figure 6 somehow show a higher temporal autocorrelation than the rest of the network. Having this said, and being conscious of the adverse effect of temporal autocorrelation on the performance of Mann–Kendall method (Yue et al. 2002), we next aim at checking the behavior of individual pixel time series in the detected areas in Figure 6. However, since all detected pixels with significant trend, per each type in Figure 6, generally show similar trend, we look into their average behavior over time. Figure 7 shows the average time series of the detected pixels in Figure 6 in combination with their locally weighted smooth regression lines (Cleveland, Grosse, and Shyu 2017). The monotonic trend in the behavior of time series of the estimated intensities of street crime, and of the estimated relative risk of street crime with respect to traffic accident is clearly visible in Figure 7. Concerning the time series of the estimated intensities of traffic accidents, apparently there is a low slope downward trend from the middle of time series onwards.

In addition, we next look for trend in the monthly estimated intensity images for different types of street crime such as antisocial behavior, shoplifting, vehicle-related, and drug-related crimes individually. Note these are the only types of street crimes appeared in all months. Figures 8 and 9 show their corresponding first lag partial autocorrelation and detected pixels with



**Figure 6.** Detected pixels with significant trend based on the univariate Mann–Kendall test, at significance level 0.05, for the time series of monthly estimated intensity and relative risk images of fatal traffic accident and street crime in the city of London, UK, in the period of January 2013 to December 2017. Values represent the Kendall's  $\tau$ .

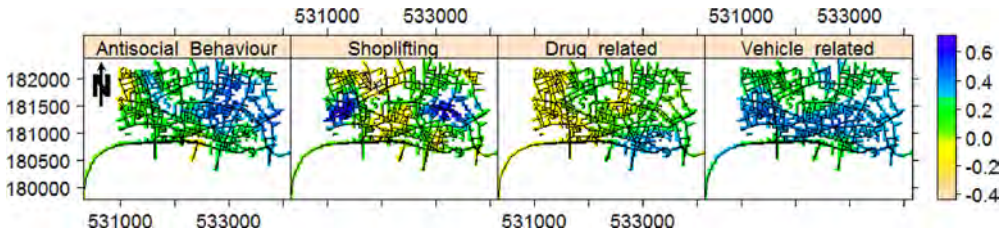


**Figure 7.** Average relative risk and estimated intensities, after aggregation and deseasoning, of the detected significant pixels by the Mann–Kendall method, at significance level 0.05, together with their locally weighted smooth regression lines.

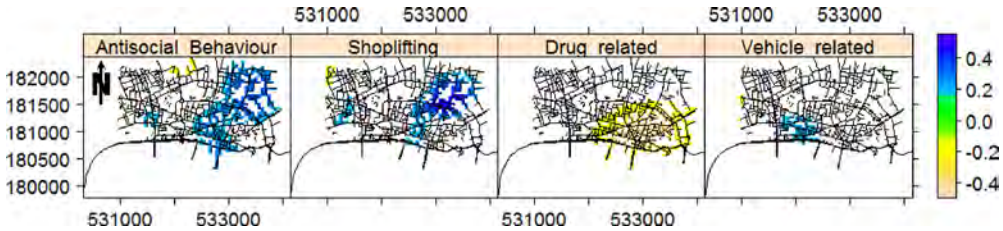
significant trend, respectively. It is clearly seen that these types of crimes show different behavior over the network. Further, from Figure 9 we can see that the estimated intensity of drug-related crimes has experienced a reduction over time in a big part of the network. Regarding other types of crimes, there are both upward and downward detected trends in different areas, where the major areas having upward trend belong to antisocial behavior and shoplifting, respectively. We add that the multivariate Mann–Kendall tests also gave rise to  $p$  values 0.05, 0.01,  $9.1 \times 10^{-5}$ , and 0.55 for antisocial behavior, shoplifting, drug-related, and vehicle-related crimes, respectively.

We further check the existence of any monotonic trend in the time-ordered relative risk images of different types of crimes with respect to each other. The multivariate Mann–Kendall test gave rise to  $p$  values 0.03 (antisocial behavior vs. drug-related),  $3.35 \times 10^{-5}$  (drug-related vs. shoplifting), 0.43 (vehicle-related vs. shoplifting), and 0.13 (vehicle-related vs. drug-related). We now employ the univariate Mann–Kendall test over each pixel time series to disclose the pixels/locations with trends. Figure 10 shows the locations where such relative risk time series have experienced monotonic trends. We have seen that the relative risk of antisocial behavior crimes with respect to drug-related crimes shows an increasing trend in the southeast of the network. The majority of the eastern part of the network shows a decreasing trend for the relative risk of





**Figure 8.** First lag partial autocorrelation for the time series of monthly estimated intensity images of different types of street crime in the city of London, UK, in the period of January 2013 to December 2017.

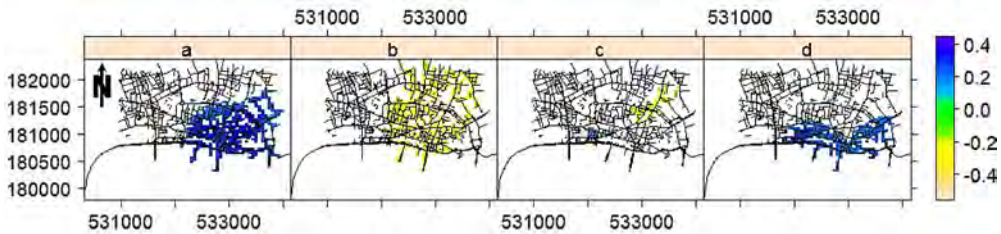


**Figure 9.** Detected pixels with significant trend based on the univariate Mann-Kendall test, at significance level 0.05, for the time series of monthly estimated intensity images of different types of street crime in the city of London, UK, in the period of January 2013 to December 2017.

drug-related crimes with respect to shoplifting. Also, the relative risk of vehicle-related crimes versus shoplifting shows a decreasing trend in a small area in the northeast of the network, together with an increasing trend in a few pixels in the south. Finally, the relative risk of vehicle-related versus drug-related crimes shows an increasing trend in most of the southern part of the network. These outputs show that the temporal changes in the relative risks between different types of crimes clearly varies over the network, there is no overall unique behavior, and moreover the slope of the trend varies among different risks. We did not find any significant trend for the relative risks between other combinations of crimes. We also add the first lag partial autocorrelation for the corresponding time series of the images displayed in Figure 10, in the detected pixels, is generally quite low with averages 0.03, 0.21, 0.28, and 0.08 for relative risks of antisocial behavior versus drug-related, drug-related versus shoplifting, vehicle-related against shoplifting, and vehicle-related against drug-related, respectively.

## 6. Summary

On the one hand, the problem of trend detection in time series has been often called within different fields such as remote sensing, agronomy, finance, etc, due to its important role in model fitting and prediction. On the other hand, spatial point patterns may also appear as a time series of realizations. However, the field of point processes has not yet benefited from trend detection methods. In this article, we have combined the well-known trend detection problem with the recently gained attention topic of spatial point processes on linear networks. We have focused on the time series of monthly estimated intensities and relative risk images of fatal traffic accident and street crime in the city of London, UK, from January 2013 to December 2017. We have obtained the intensity and relative risk images by using the non-parametric kernel-based estimator of Rakshit et al. (2019). In our results, the time series of estimated intensities has shown that, for both datasets, they go under significant changes temporally and spatially which is a sign of first-order non-separability (an assumption which is commonly considered when analyzing spatio-temporal point patterns). The time series of estimated intensity images of traffic accident data has not generally shown a strong evidence of trend anywhere in the network. Conversely, the



**Figure 10.** Detected pixels with significant trend based on the univariate Mann–Kendall test, at significance level 0.05, for the time series of monthly estimated relative risk images for antisocial vs. drug-related (a), drug-related vs. shoplifting (b), vehicle-related vs. shoplifting (c), and vehicle-related vs. drug-related (d) in the city of London, UK, in the period of January 2013 to December 2017.

time series of estimated intensity images of street crime, and consequently its relative risk with respect to fatal traffic accident, however, have notably experienced a strong upward trend in mostly western, central, and northeastern parts of the network. Further, we have seen that different types of crimes show different behavior over the network, and consequently different behavior in terms of upward/downward trend. Generally, we have observed that the temporal changes in the intensity/relative risk images clearly varies over the network, and there is no overall unique behavior. Furthermore, the relative risks between certain types of crimes experience different types of trends over different regions in the city of London. Although the average time series of the locations with significant monotonic trend show evidence of such detected trend, distinguishing true and false positives needs further detailed research.

Regarding the limitations and future works, we note that one may estimate the intensities by means of parametric estimation to also reveal the effect of the characteristics of network over intensities/relative-risks such as distances to crossings, roundabouts, etc. Trend detection based on parametric intensity/relative-risk estimation may not necessarily lead to similar results. In addition, one may aim to model the time-ordered non-parametrically estimated intensities/relative-risks values based on some given/collected covariates to disclose their effect over the evolution of intensities/relative-risks over time. Such parametric modeling can further reveal what actually causes the trend. Moreover, another relevant and interesting idea might be to investigate the influence of autocorrelation, and also to detect the time index when trend starts to grow using e.g. deep-learning-based methods such as Long-short term memory (LSTM), Recurrent Neural Networks (RNN), and Convolutional Neural Network (CNN).

Our data and R codes, to reproduce the results, are available at [https://github.com/Moradii/trend\\_intensity\\_images](https://github.com/Moradii/trend_intensity_images). Moreover, throughout the article, we have made use of the R packages `stats` (R Core Team 2020), `spatstat` (Baddeley and Turner 2005; Baddeley, Rubak, and Turner 2015), `sparr` (Davies, Marshall, and Hazelton 2018), `remote` (Appelhans, Detsch, and Nauss 2015), `raster` (Hijmans 2019), `trend` (Pohlert 2018), `gimms` (Detsch 2018), `sp` (Pebesma and Bivand 2005; Bivand, Pebesma, and Gomez-Rubio 2013), and `ggplot2` (Wickham 2016).

## Acknowledgments

The authors are grateful to the editor and three referees for constructive comments.

## Funding

Jorge Mateu has been partially funded by grants UJI-B2018-04 from Universitat Jaume I of Castellón (Spain), AICO/2019/198 from Generalitat Valenciana, and MTM2016-78917-R from Ministry of Science. Somnath Chaudhuri has been funded through the Erasmus Mundus programme by the European Commission under the Framework Partnership Agreement, FPA-2016-2054.

## ORCID

Mehdi Moradi  <http://orcid.org/0000-0003-3905-4498>

## References

- Ang, Q. W., A. Baddeley, and G. Nair. 2012. Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics* 39 (4):591–617. doi:10.1111/j.1467-9469.2011.00752.x.
- Appelhans, T., F. Detsch, and T. Nauss. 2015. remote: Empirical orthogonal teleconnections in R. *Journal of Statistical Software* 65 (10):1–19. doi:10.18637/jss.v065.i10.
- Baddeley, A., E. Rubak, and R. Turner. 2015. *Spatial point patterns: Methodology and applications with R*. London: Chapman and Hall/CRC.
- Baddeley, A., and R. Turner. 2005. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software* 12 (6):1–42. doi:10.18637/jss.v012.i06.
- Bivand, R., E. Pebesma, and V. Gomez-Rubio. 2013. *Applied spatial data analysis with R*. 2nd ed. NY: Springer.
- Bullock, E. L., C. E. Woodcock, and C. E. Holden. 2020. Improved change monitoring using an ensemble of time series algorithms. *Remote Sensing of Environment* 238:111165. doi:10.1016/j.rse.2019.04.018.
- Cleveland, W. S., E. Grosse, and W. M. Shyu. 1992. Local regression models. Chapter 8 of *Statistical models in S*, eds. J. M. Chambers and T. J. Hastie, 309–76. Wadsworth & Brooks/Cole.
- Cox, D. R., and A. Stuart. 1955. Some quick sign tests for trend in location and dispersion. *Biometrika* 42 (1-2): 80–95. doi:10.1093/biomet/42.1-2.80.
- Davies, T. M. 2013. Jointly optimal bandwidth selection for the planar kernel-smoothed density-ratio. *Spatial and Spatio-Temporal Epidemiology* 5:51–65. doi:10.1016/j.sste.2013.04.001.
- Davies, T. M., K. Jones, and M. L. Hazelton. 2016. Symmetric adaptive smoothing regimens for estimation of the spatial relative risk function. *Computational Statistics & Data Analysis* 101:12–28. doi:10.1016/j.csda.2016.02.008.
- Davies, T. M., J. C. Marshall, and M. L. Hazelton. 2018. Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Statistics in Medicine* 37 (7):1191–221. doi:10.1002/sim.7577.
- Detsch, F. 2018. *gimms: Download and process GIMMS NDVI3g data*. R package version 1.1.1.
- Feng, S., Z. Li, Y. Ci, and G. Zhang. 2016. Risk factors affecting fatal bus accident severity: Their impact on different types of bus drivers. *Accident; Analysis and Prevention* 86:29–39. doi:10.1016/j.aap.2015.09.025.
- Grundy, T., R. Killick, and G. Mihaylov. 2020. High-dimensional changepoint detection via a geometrically inspired mapping. *Statistics and Computing* 30 (4):1155–66. doi:10.1007/s11222-020-09940-y.
- Hamed, K. 2009. Enhancing the effectiveness of prewhitening in trend analysis of hydrologic data. *Journal of Hydrology* 368 (1-4):143–55. doi:10.1016/j.jhydrol.2009.01.040.
- Hamed, K., and A. R. Rao. 1998. A modified Mann–Kendall trend test for autocorrelated data. *Journal of Hydrology* 204 (1-4):182–96. doi:10.1016/S0022-1694(97)00125-X.
- Hazelton, M. L. 2008. Kernel estimation of risk surfaces without the need for edge correction. *Statistics in Medicine* 27 (12):2269–72. doi:10.1002/sim.3047.
- Hijmans, R. J. 2019. *raster: Geographic data analysis and modeling*. R Package Version 2.8–19.
- Hipp, J. R., Y. A. Kim, and K. Kane. 2019. The effect of the physical environment on crime rates: Capturing housing age and housing type at varying spatial scales. *Crime & Delinquency* 65 (11):1570–95. doi:10.1177/0011128718779569.
- Kelsall, J. E., and P. Diggle. 1995. Kernel estimation of relative risk. *Bernoulli* 1 (1/2):3–16. doi:10.2307/3318678.
- Kendall, M. G. 1948. *Rank correlation methods*. Oxford, England: Griffin.
- Kulkarni, A., and H. von Storch. 1995. Monte carlo experiments on the effect of serial correlation on the Mann–Kendall test of trend. *Meteorologische Zeitschrift* 4 (2):82–85. doi:10.1127/metz/4/1992/82.
- Libiseller, C., and A. Grimvall. 2002. Performance of partial mann–kendall tests for trend detection in the presence of covariates. *Environmetrics* 13 (1):71–84. doi:10.1002/env.507.
- Mann, H. B. 1945. Nonparametric tests against trend. *Econometrica* 13 (3):245–59. doi:10.2307/1907187.
- Matteson, D. S., and N. A. James. 2014. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* 109 (505):334–45. doi:10.1080/01621459.2013.849605.
- McSwiggan, G., A. Baddeley, and G. Nair. 2016. Kernel density estimation on a linear network. *Scandinavian Journal of Statistics* 44 (2):324–45. doi:10.1111/sjos.12255.
- McSwiggan, G., A. Baddeley, and G. Nair. 2020. Estimation of relative risk for events on a linear network. *Statistics and Computing* 30 (2):469–84. doi:10.1007/s11222-019-09889-7.
- Millitino, A. F., M. Moradi, and M. D. Ugarte. 2020. On the performance of trend and change-point detection methods for remote sensing data. *Remote Sensing* 12 (6):1008. doi:10.3390/rs12061008.
- Moradi, M. 2018. Spatial and spatio-temporal point patterns on linear networks. PhD diss., University Jaume I.

- Moradi, M., O. Cronie, E. Rubak, R. Lachize-Rey, J. Mateu, and A. Baddeley. 2019. Resample-smoothing of Voronoi intensity estimators. *Statistics and Computing* 29 (5):995–1010. doi:[10.1007/s11222-018-09850-0](https://doi.org/10.1007/s11222-018-09850-0).
- Moradi, M., F. Rodriguez-Cortes, and J. Mateu. 2018. On kernel-based intensity estimation of spatial point patterns on linear networks. *Journal of Computational and Graphical Statistics* 27 (2):302–11. doi:[10.1080/10618600.2017.1360782](https://doi.org/10.1080/10618600.2017.1360782).
- Okabe, A., T. Satoh, and K. Sugihara. 2009. A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science* 23 (1):7–32. doi:[10.1080/13658810802475491](https://doi.org/10.1080/13658810802475491).
- Padgham, M., B. Rudis, R. Lovelace, and M. Salmon. 2017. osmdata. *The Journal of Open Source Software* 2 (14): 305. doi:[10.21105/joss.00305](https://doi.org/10.21105/joss.00305).
- Pebesma, E., and R. Bivand. 2005. Classes and methods for spatial data in R. *R News* 5 (2):9–13.
- Pettitt, A. N. 1979. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (2):126–35. doi:[10.2307/2346729](https://doi.org/10.2307/2346729).
- Pohlert, T. 2018. *trend: Non-parametric trend tests and change-point detection*. R package version 1.1.1.
- Rakshit, S., T. M. Davies, M. Moradi, G. McSwiggan, G. Nair, J. Mateu, and A. Baddeley. 2019. Fast kernel smoothing of point patterns on a large network using two-dimensional convolution. *International Statistical Review* 87 (3):531–56. doi:[10.1111/insr.12327](https://doi.org/10.1111/insr.12327).
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Serinaldi, F., and C. G. Kilsby. 2016. The importance of prewhitening in change point analysis under persistence. *Stochastic Environmental Research and Risk Assessment* 30 (2):763–77. doi:[10.1007/s00477-015-1041-5](https://doi.org/10.1007/s00477-015-1041-5).
- Truong, C., L. Oudre, and N. Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing* 167:107299. doi:[10.1016/j.sigpro.2019.107299](https://doi.org/10.1016/j.sigpro.2019.107299).
- Verbesselt, J., R. Hyndman, G. Newnham, and D. Culvenor. 2010. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment* 114 (1):106–15. doi:[10.1016/j.rse.2009.08.014](https://doi.org/10.1016/j.rse.2009.08.014).
- Von Storch, H. 1995. Misuses of statistical analysis in climate research. In *Analysis of Climate Variability: Applications of Statistical Techniques*, ed. H. von Storch and A. Navarra, 11–26. Berlin, Germany: Springer-Verlag.
- Wickham, H. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
- Yamada, I., and J. C. Thill. 2004. Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography* 12 (2):149–58. doi:[10.1016/j.jtrangeo.2003.10.006](https://doi.org/10.1016/j.jtrangeo.2003.10.006).
- Yue, S., P. Pilon, B. Phinney, and G. Cavadias. 2002. The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes* 16 (9):1807–29. doi:[10.1002/hyp.1095](https://doi.org/10.1002/hyp.1095).
- Yue, S., and C. Wang. 2004. The mann-kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water Resources Management* 18 (3):201–18. doi:[10.1023/B:WARM.0000043140.61082.60](https://doi.org/10.1023/B:WARM.0000043140.61082.60).
- Zeileis, A., C. Kleiber, W. Krämer, and K. Hornik. 2003. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis* 44 (1-2):109–23. doi:[10.1016/S0167-9473\(03\)00030-6](https://doi.org/10.1016/S0167-9473(03)00030-6).





