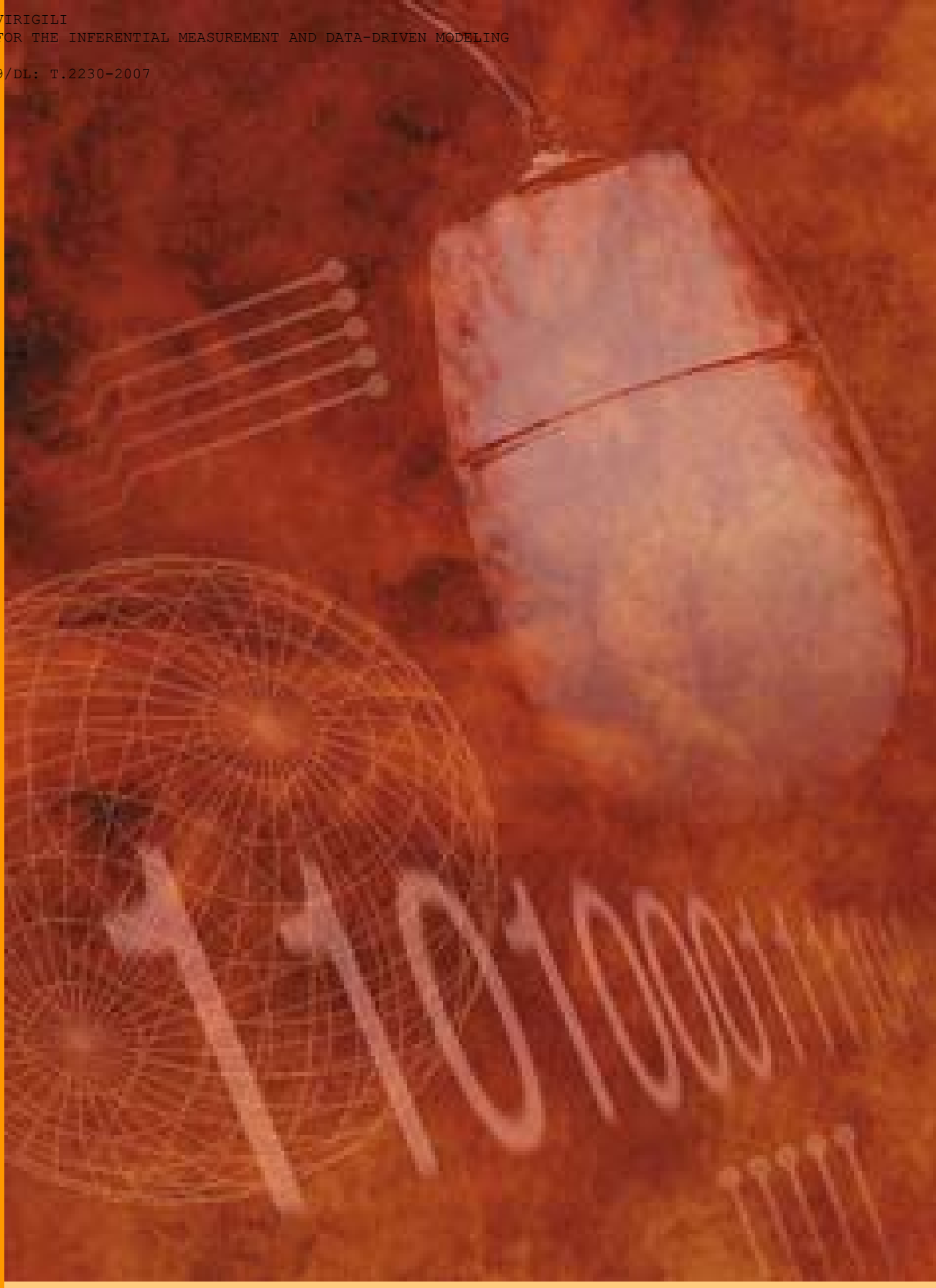
UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



Multi-Tier Framework for the Inferential Measurement and Data-driven Modeling

Robert Rallo Moya

Tarragona, juliol 2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Multi-tier Framework for the Inferential measurement and Data-driven Modeling

Memòria presentada per

Robert Rallo i Moya

Per optar al grau de Doctor per la Universitat Rovira i Virgili

Tarragona, juliol 2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

El Dr. Francesc Giralt i Prat, Catedràtic d'Universitat del Departament d'Enginyeria Química de la Universitat Rovira i Virgili, i el Dr. Joan M. Ferrer i Gener, Professor Titular d'Universitat del Departament d'Enginyeria Informàtica i Matemàtiques de la Universitat Rovira i Virgili,

FAN CONSTAR:

Que el present treball que porta per títol

MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING

Que presenta el Sr. Robert Rallo i Moya per optar al grau de Doctor per la Universitat Rovira i Virgili, ha estat realitzat sota la seva immediata direcció, i que tots els resultats obtinguts són fruit de la recerca realitzada per l'esmentat doctorand.

I perquè es faci saber i tingui els efectes que correspongui, signen aquesta certificació

Tarragona, 6 de juny de 2007

Dr. Francesc Giralt i Prat

Dr. Joan M. Ferrer i Gener

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

A Mercè

Als pares, iaios i germans

A Alba, a Laura i a Roc

Agraïments

A Francesc i a Joan per ser primer amics i després directors. Als membres del tribunal per acceptar de judicar aquest treball. Als companys del grup de recerca, departament i escola.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Summary

A framework for the inferential measurement and data-driven modeling has been proposed and assessed in several real-world application domains. The architecture of the framework has been structured in multiple tiers to facilitate extensibility and the integration of new components. Each of the proposed four tiers has been assessed in an uncoupled way to verify their suitability. The first tier, dealing with exploratory data analysis, has been assessed with the characterization of the chemical space related to the biodegradation of organic chemicals. This analysis has established relationships between physicochemical variables and biodegradation rates that have been used for model development. At the preprocessing level, a novel method for feature selection based on dissimilarity measures between Self-Organizing maps (SOM) has been developed and assessed. The proposed method selected more features than others published in literature but leads to models with improved predictive power. Single and multiple data imputation techniques based on the SOM have also been used to recover missing data in a Waste Water Treatment Plant benchmark. A new dynamic method to adjust the centers and widths of in Radial basis Function networks has been proposed to predict water quality. The proposed method outperformed other neural networks.

The proposed modeling components have also been assessed in the development of prediction and classification models for biodegradation rates in different media. The results obtained proved the suitability of this approach to develop data-driven models when the complex dynamics of the process prevents the formulation of mechanistic models. The use of rule generation algorithms and Bayesian dependency models has been preliminary screened to provide the framework with interpretation capabilities. Preliminary results obtained from the classification of Modes of Toxic Action (MOA) indicate that this could be a promising approach to use MOAs as proxy indicators of human health effects of chemicals.

Finally, the complete framework has been applied to three different modeling scenarios. A virtual sensor system, capable of inferring product quality indices from primary process variables has been developed and assessed. The system was integrated with the control system in a real chemical plant outperforming multi-linear correlation models usually adopted by chemical manufacturers. A model to predict carcinogenicity from molecular structure for a set of aromatic compounds has been developed and tested. Results obtained after the application of the SOM-dissimilarity feature selection method yielded better results than models published in the literature. Finally, the framework has been used to facilitate a new approach for environmental modeling and risk management within geographical information systems (GIS). The SOM has been successfully used to characterize exposure

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

scenarios and to provide estimations of missing data through geographic interpolation. The combination of SOM and Gaussian Mixture models facilitated the formulation of a new probabilistic risk assessment approach.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Resum

Aquesta tesi proposa i avalua en diverses aplicacions reals, un marc general de treball per al desenvolupament de sistemes de mesurament inferencial i de modelat basats en dades. L'arquitectura d'aquest marc de treball s'organitza en diverses capes que faciliten la seva extensibilitat així com la integració de nous components. Cadascun dels quatre nivells en que s'estructura la proposta de marc de treball ha estat avaluat de forma independent per a verificar la seva funcionalitat. El primer que nivell s'ocupa de l'anàlisi exploratòria de dades ha esta avaluat a partir de la caracterització de l'espai químic corresponent a la biodegradació de certs compostos orgànics. Fruit d'aquest anàlisi s'han establert relacions entre diverses variables físico-químiques que han estat emprades posteriorment per al desenvolupament de models de biodegradació. A nivell del preprocés de les dades s'ha desenvolupat i avaluat una nova metodologia per a la selecció de variables basada en l'ús del Mapes Autoorganitzats (SOM). Tot i que el mètode proposat selecciona, en general, un major nombre de variables que altres mètodes proposats a la literatura, els models resultants mostren una millor capacitat predictiva. S'han avaluat també tot un conjunt de tècniques d'imputació de dades basades en el SOM amb un conjunt de dades estàndard corresponent als paràmetres d'operació d'una planta de tractament d'aigües residuals. Es proposa i avalua en un problema de predicció de qualitat en aigua un nou model dinàmic per a ajustar el centre i la dispersió en xarxes de funcions de base radial. El mètode proposat millora els resultats obtinguts amb altres arquitectures neuronals.

Els components de modelat proposat s'han aplicat també al desenvolupament de models predictius i de classificació de les velocitats de biodegradació de compostos orgànics en diferents medis. Els resultats obtinguts demostren la viabilitat d'aquesta aproximació per a desenvolupar models basats en dades en aquells casos en els que la complexitat de dinàmica del procés impedeix formular models mecanicistes. S'ha dut a terme un estudi preliminar de l'ús de algorismes de generació de regles i de grafs de dependència bayesiana per a introduir una nova capa que faciliti la interpretació dels models. Els resultats preliminars obtinguts a partir de la classificació dels Modes d'acció Tòxica (MOA) apunten a que l'ús dels MOA com a indicadors intermediaris dels efectes dels compostos químics en la salut és una aproximació factible.

Finalment, el marc de treball proposat s'ha aplicat en tres escenaris de modelat diferents. En primer lloc, s'ha desenvolupat i avaluat un sensor virtual capaç d'inferir índexs de qualitat a partir de variables primàries de procés. El sensor resultant ha estat implementat en una planta química real millorant els resultats de les correlacions multilineals emprades habitualment. S'ha desenvolupat i avaluat un

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

model per a predir els efectes carcinògens d'un grup de compostos aromàtics a partir de la seva estructura molecular. Els resultats obtinguts després d'aplicar el mètode de selecció de variables basat en el SOM milloren els resultats prèviament publicats. Aquest marc de treball s'ha usat també per a proporcionar una nova aproximació al modelat ambiental i l'anàlisi de risc amb sistemes d'informació geogràfica (GIS). S'ha usat el SOM per a caracteritzar escenaris d'exposició i per a desenvolupar un nou mètode d'interpolació geogràfica. La combinació del SOM amb els models de mescla de gaussianes dona una nova formulació al problema de l'anàlisi de risc des d'un punt de vista probabilístic.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table of Contents

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

List of Figures

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

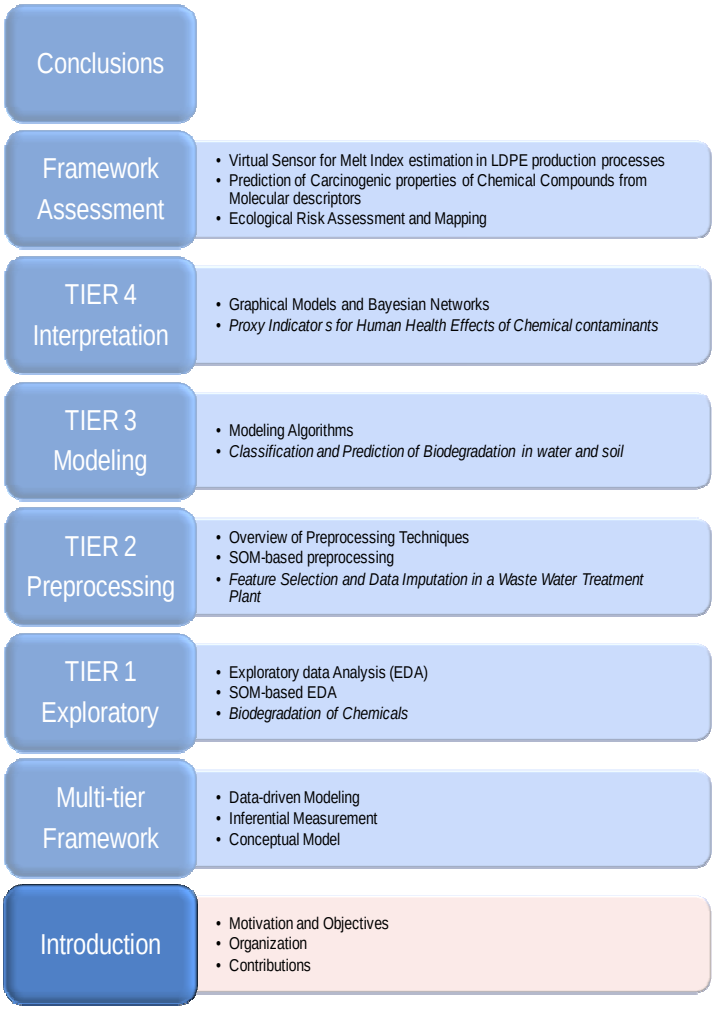
List of Tables

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 1

Introduction

This thesis focuses on the design and assessment of a multi-tier framework for data-driven modeling and inferential measurement. For this purpose a proof of concept model is developed and applied to real world data in several engineering related domains.

1.1 Motivation and Objectives

In science and engineering we are often involved in the study of physical systems to acquire a better understanding of the rules that govern their inner dynamics. Usually, due to scale related issues or to the inherent complexity of the studied process, a simplified representation of reality is needed. We refer to this alternate representation of reality using the term *model*. The modeling process includes a detailed study of the physical system under consideration, the proper statement of the problem, the collection, conditioning and preparation of experimental data, the construction of the model, its use, its assessment and, finally, the interpretation of the results obtained (Solomatine, 2002).

Traditionally, the description of the behavior of a physical system was grounded on a good understanding of their underlying processes (based on physic or first principles laws) resulting in mechanistic models. However, an accurate expression of many real-world processes through a rigorous mathematical formulation is difficult or even impossible. An alternative approach for these difficult problems would be to use data-driven modeling, which is especially attractive in processes for which adequate knowledge of their physics is limited. In data-driven modeling, a system is defined in terms of its state variables using only a limited amount of knowledge about the details of its physical behavior. Simple statistical models such as linear or multi-linear regression follow this approach. Inferential measurement is an extension of this model development paradigm that allows costly or difficult to measure parameters to be inferred from other more easily measured variables. A direct application of these concepts was software sensors which in contrast to hardware sensors use inferential techniques to overcome the aforementioned limitations.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

According to Breiman (2001) there exist two different scientific "cultures" in the development of experimental models from data, i.e., in process identification. Both cultures think about data as being generated by a black-box in which input variables enter at one side and system responses come out at the opposite side. Inside this box, an unknown nature mechanism associates inputs with their corresponding outputs. The main difference among these two cultures resides in the way on how they model this black-box. One, referred as *"data modeling culture"*, starts by assuming a stochastic data model for inside the box whose parameters are estimated from data. The other, known as *"algorithmic modeling culture"* considers the interior of the black-box as a complex set of unknown function and their modeling approach is based in finding an algorithm that operates on input data to mimic system responses.

The former methodology presents an inherent drawback in its formulation which states that the proposed model is fit to available data to draw conclusions about nature mechanisms. Users of these models often forget that these conclusions are only about model's mechanism, not about nature's mechanisms. In consequence, if the model is a poor emulation of nature the conclusions drawn would be wrong. The goodness-of-fit of these models is usually demonstrated by giving the value of the correlation coefficient R^2, which is often closer to zero than one and is sometimes overinflated by the use of too many parameters. Besides computing R^2 nothing else is done to verify whether the observed data could have been generated by the proposed model. An additional undesired effect of this modeling approach is the unexpected multiplicity of models. In fact, data may point with equal sensitivity to several possible models which in turn may lead to different interpretations of the relationships between input and output variables. Also, the use of simple parametric models imposed on data generated by complex systems involving unknown physical, chemical or biological mechanisms, usually results in a loss of both accuracy and information. As a consequence, the burden of data models increases (i.e. Bayesian methods combined with Markov Chain Montecarlo) loosing the advantage of presenting a simple and clear picture of nature's mechanisms.

In the mid-80's, new algorithms for fitting data became available, such as neural networks and decision trees. This was the origin of a new discipline known as Machine Learning (ML). ML constitutes a multidisciplinary field which draws ideas on results from Artificial Intelligence, probability and statistics, computational theory, control theory, information theory, philosophy, neurobiology and many other fields. Machine learning techniques facilitate the development of algorithmic modeling. Under the framework of this new approach, data fitting becomes more focused on the properties of algorithms by themselves than in the intrinsic characteristics of models. The goal is now to attain good predictive accuracy instead of the interpretability premises used in data modeling. The only assumption about the data model made in ML theory is that data are drawn independent and identically distributed (i.i.d.) from an unknown multivariate distribution.

Since then, the increase in predictive accuracy and the advances in machine learning theories have been substantial. Nevertheless there still remain some open issues which focalize most of current research in this field:

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

- Algorithmic modeling may result in multiple equivalent models for a certain data set, as in the data modeling approach. This effect leads to what it is known as instability and occurs when there are many different models crowded together that have about the same accuracy. In such cases a slight perturbation of data or in model generation will cause a skip from a model to another. The models may be close to each other in terms of predictive errors but can be very far in terms of its structure. This can be exploited to obtain new models with increased accuracy. It has been demonstrated that the aggregation over a large set of competing models improves the accuracy of predictions when compared with each single model estimates. Different techniques such as *bagging* or *boosting* exploit this property.

- William of Occam proposed in the fourteenth century a principle, known as Occam's razor, which states that "*Pluralitas non est ponenda sine neccesitate*", i.e., whenever we have two models (theories) with similar predictive accuracy, the simpler one should always prevail. This leads to a situation in which accuracy and model simplicity (interpretability) are always in conflict. The use of complex predictor models may be unpleasant to a part of the research community, but the soundest path to deal with this conflict is to first look for predictive accuracy, and then try to understand the mechanism driving the prediction.

- Richard Bellman coined the term "curse of dimensionality" to refer to the undesired effects of having a large number of variables in model's development. A commonly used approach to avoid this curse is to use feature selection and dimensionality reduction algorithms coupled with the modeling mechanism. Recent studies demonstrated that a large number of features is not always undesirable. In fact, techniques such as Support Vector Machines (Vapnik, 1998) increase the number of features to find optimal separating hyperplanes with the lowest classification errors.

Data-driven modeling is always a challenging task which requires the integration of multidisciplinary knowledge from different sources such as mathematics, statistics, engineering, and computer science. One can find in the literature multiple references to each of the techniques and algorithms needed to develop data-driven models and inferential measurement systems. However, there are fewer references on how to integrate all these components into a sound framework which provides all the necessary elements to successfully deploy these models for different application domains. In this context, our main objective will be the formalization of the architecture of a complete framework on top of which these models could be developed. The main hypothesis of this research work is that a complete framework to develop data-driven models could be constructed by the proper integration of a set of machine learning components on a multi-tier architecture in which each tier accounts for different abstraction levels. This integrated approach would give a complete and coherent vision of the modeling process at different levels. The framework should be general and applicable to diverse kinds of modeling tasks (multiple application domains). Thus, the assessment of the resulting architecture in different real-world application domains will be our second research objective.

The development of this framework poses multiple questions that will be studied in this dissertation.

- Are self-organizing algorithms such as the Self-Organizing Map a suitable technique to explore the data space that characterizes certain real-world phenomenon and to obtain an informative mapping of the relationships among these data?
- How does the preprocessing of data affect the development of data-driven models? Is it possible to design an integrated feature-example selection and data completion scheme using the Self-Organizing Map?
- How does the selection of classifiers or predictors affect the quality of the models developed? Are these models dependent on the application domain?
- Is it possible to overcome the limitations in the interpretability of algorithmic models by using probabilistic techniques such as dependency modeling or Bayesian networks? Can these probabilistic descriptions be integrated on the Self-Organizing Map?

1.2 Organization

The current document is organized as follows. Chapter two provides a short overview of the state of art of data-driven modeling techniques and introduces the problem of inferential measurement and the concept of virtual sensor as a natural extension of both procedures. A conceptual model for the integrated framework for inferential measurement is also presented and details at each abstraction level are given in subsequent chapters.

Chapter three introduces the first tier of the model which deals with the exploratory data analysis (EDA) as a preliminary step in data-driven and inferential modeling. The basis of this exploratory stage is the Self-Organizing Feature Map (SOM). The analysis of the *chemical space* of variables that permit the classification and the prediction of biodegradation of chemicals in water is presented as an example of the proposed techniques.

Chapter four deals with the model tier related to data preprocessing. It starts by introducing the main challenges at this level: the selection of relevant features, the selection of relevant examples and the reconstruction of missing data. A new feature selection algorithm based on dissimilarity measures over SOM maps is introduced and assessed. The use of the SOM to generate optimized train/test sets is presented and discussed afterwards. Finally, both single and multiple SOM imputation methods to deal with datasets containing missing data are considered in section 4.4. The proposed imputation techniques are assessed in section 4.5 in which these algorithms are applied to the prediction of biological oxygen demand (BOD) in effluents of a waste water treatment plants.

Chapter five goes one step further in the conceptual model with the integration of diverse modeling techniques into the proposed framework. Section 5.1 introduces three different approaches to develop the model; classification, prediction and

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

ensembles. Afterwards, section 5.2 presents an application of these approaches to the classification and forecasting of biodegradation rates in several environmental compartments.

The highest level of abstraction in the proposed framework is reached in chapter six with the interpretation of data. Deterministic models are complemented with probabilistic models such as graphical models, Bayesian networks and probabilistic models based on Gaussian mixtures developed on top of SOM. Section 6.3 outlines an application of this model tier in which the Modes of toxic Action (MOA) of different chemicals are classified and predicted using both deterministic and probabilistic approaches.

The goal of chapter seven is to assess the proposed framework in different application domains. To this end three case studies are analyzed. The first deals with the development of a virtual sensor for a product quality indicator in an industrial process. The second focuses on the development of data-driven models to classify and estimate eco-toxicological properties of chemicals using information of their molecular structure. Finally, the third concerns the development of a geographic inferential measurement system for exposure concentration functions and risk evaluation in a groundwater pollution scenario.

The main conclusions of this thesis and outlines future research lines are summarized in chapter eight.

1.3 Contributions

The main contributions of this dissertation can be summarized from two points of view. From the computer science perspective the major contributions are the following:

- Proposal, implementation and multi-domain assessment of a new multi-tier architecture for data-driven inferential modeling.
- Development of a new feature selection method established upon dissimilarity measures on Self-Organizing Maps capable of detecting relevant, irrelevant and redundant variables.
- Development and assessment of data imputation techniques using the Self-Organizing Map and ensembles of maps.
- Development and assessment of a data augmentation technique based on the use of Self-organizing map prototypes.
- Implementation of a dynamic method to construct Radial-basis function networks based on competitive learning algorithms to adjust the function centers and widths.

From the engineering perspective the above framework was applied to several chemical, process and environmental engineering related task such as:

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

- Development of a neural-based soft sensor system to forecast product quality from plant data in a low density polyethylene (LDPE) production process.
- Development of a data-driven model to predict carcinogenicity and drug activity of chemical compounds.
- Development of a data completion system using single imputation techniques based on the Self-organizing map as part of the development of an inferential measurement system to estimate the biological oxygen demand (BOD) in a Waste Water Treatment Plant benchmark.
- A data-driven model for the estimation of Modes of Toxic Action (MOA) for phenols within the proposed framework.
- A QSAR model for the prediction of Toxicological properties of chemical contaminants using an ensemble of neural networks.
- QSBR and SBR data-driven models for biodegradation of organic pollutants in water and soil.
- A novel methodology based on data-driven models to develop risk probability maps using SOM-based kriging and Gaussian Mixtures.

Moreover, we have conducted extensive computer simulations to verify the proposed multi-tier framework and each of its components. We have successfully used the proposed framework to develop software sensor systems to infer quality variables in real-world production processes. In addition the effectiveness of SOM-dissimilarity method for feature selection is confirmed by its ability to select the most appropriate input variables (process variables or molecular descriptors) in several real world problems.

The papers and conference presentations that have stemmed so far from this thesis are summarized in Table 1. This table explicitly mentions the relationships between the published material and each of the tiers in the proposed framework.

Table 1.1. Dissemination of the current dissertation with model tier identification

	Contributions	TIER 1: Exploratory	TIER 2: Preprocessing	TIER 3: Modeling	TIER 4: Interpretation
1	Giralt, F. , Arenas, A., Ferre-Giné, J., Rallo, R., Kopp, G.A. The simulation and Interpretation of free turbulence with a cognitive neural system. *Physics of Fluids*, 12, 1826, 2000			■	
2	Rallo. R., Ferré-Giné, J., Arenas, A., Giralt, F. A Message Oriented Middleware approach to HYSYS extensibility. *Proceedings of HYPROTECH'2000*, Amsterdam, NL, 2000	■		■	

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

3	Rallo, R., Ferré-Giné, J., Arenas, A., Giralt, F. Forecasting Product Quality in Industrial Processes with Virtual Sensors. *Proceedings of AICHE 2002*. Sensor Technology. 127-136, 2002	■	■	■	
4	Rallo,.R., Ferré-Giné, J., Arenas, A., Giralt, F. Neural Virtual Sensor for the Inferential Prediction of Product Quality from Process Variables. *Computers and Chemical Engineering* (26) 12, 1735-54, 2002	■	■	■	
5	Espinosa,G., Rallo, R., Arenas, Giralt, F., A., Carbó-Dorca, R., Cohen, Y. Carcinogenicity Prediction by using SOM and Fuzzy ARTMAP Neural Networks. *SETAC Europe 13th Annual Meeting*, Hamburg Germany 2003	■	■	■	
6	Giralt, F., Espinosa,G., Rallo, R., Arenas, A., Carbó-Dorca, R., Cohen, Y. A Predictive Methodology for Carcinogenicity and Drug Activity based on Neural Networks. *Proceeding of the AIChE Annual Conference*, San Francisco 2003	■	■	■	
7	Rallo, R., Ferré-Giné, J. and Giralt, F.Best Feature Selection and Data Completion for the Design of Soft Neural Sensors. *Proceedings of AIChE 2003, 2nd Topical Conference on Sensors*, San Francisco, November 2003		■	■	
8	Giralt, F., Espinosa,G., Arenas, A., Rallo, R., Besalú, E., Carbó-Dorca, R., Cohen, Y. Evaluation of Algorithms for Molecular descriptor selection: QSAR modeling of anti-HIV-1 activity *VI International Girona Seminars on Molecular Similarity*, Girona July 2003	■	■	■	

9	Giralt, F., Espinosa, G., Kohen, Y., Rallo, R. Feature Extraction and Quantitative Structure-Property Relations (QSPRs) with Integrated SOM-Fuzzy ARTMAP Neural Systems. *AIChE Annual Meeting*, Austin (Texas) USA, Nov. 2004	■	■	■	
10	Rallo, R., Ferré-Giné, J., Giralt, F. Design of soft sensors with multiple imputation of missing data by self-organizing map ensembles. *Proceedings of the 7th World Congress of Chemical Engineering* (CDROM WCCE'05)		■	■	
11	Rallo, R., Espinosa, G., Giralt, F., Using an Ensemble of Neural based QSPRs for the prediction of Toxicological properties of chemical contaminants. *Proceedings of the 7th World Congress of Chemical Engineering* (CDROM WCCE'05)		■	■	
12	Rallo, R., Espinosa, G., Giralt, F. Prediction of modes of toxic action and toxicity of phenols with feature selection algorithms coupled with fuzzy ARTMAP. *JMMC'05 Joint Meeting on Medicinal Chemistry*, Jun. 2005, Vienna (Austria)	■	■	■	
13	Rallo, R., Espinosa, G., and F. Giralt, Using an ensemble of neural based QSARs for the prediction of toxicological properties of chemical contaminants, *Trans IChemE Part B. Process Safety and Environmental Protection*, 83(B4), 387-392 (2005)		■	■	
14	Rallo, R., Espinosa, G., Ferrer-Gener, J., Giralt, J., Giralt, F. Ensemble methods to enhance the identification of MOA's with deterministic and probabilistic networks. *SETAC Europe 16th Annual Meeting*. The Hague, 7-11 May 2006		■	■	■

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

15	Espinosa, G., Rallo, R., Grifoll, J., Vogel, T., Giralt, F. Cognitive Neural Network Analysis of Organic Pollutants persitence in the environment. *SETAC Europe 16th Annual Meeting*. The Hague, 7-11 May 2006	■	■	■	
16	Rallo, R., Mujica, M., Climent, J., Espinosa, G., Grifoll, J., Giralt, F. (Ecological) Risk Mapping based on self-organizing maps. 1st Open International NoMiracle Workshop. Ecological and Human Health Risk Assessment. Ispra, Italy , June 8-9 2006	■	■	■	■
17	Rallo, R., Espinosa, G., Vogel, T.M., Cohen, Y., F. Giralt. Quantitative Structure Biodegradation Relationships for Organic Pollutants in Water and Soil *Advanced Computations for Environmental Applications. AICHE Annual Meeting*, San Francisco, November 2006	■	■	■	
18	Martínez, I., Espinosa, G. Rallo, R., Grifoll, J., Cohen, Y., F. Giralt. A Method for Modeling Chemical Multimedia Partitioning with Neural Networks and Classifiers Environmental Fate and Transport Processes. *AICHE Annual Meeting*, San Francisco, November 2006	■	■	■	
19	allo, R., Mujica, M., Climent, J., Espinosa, G., Giralt, F. Self-Organizing Maps and Gaussian Mixture Models for Environmental Risk Assessment and Mapping. SETAC Europe 17th Annual Meeting. Porto, 20-24 May 2007	■	■	■	■
20	Espinosa, G., Rallo, R., Vogel, T., Cohen, Y., Giralt, F. Structure Biodegradation Relationships for Organic Pollutants in Water and Soil. SETAC Europe 17th Annual Meeting. Porto, 20-24 May 2007	■	■	■	

21	Martínez, I., Espinosa, G., Rallo, R., Grifoll, J., Cohen, Y., Giralt, F. Estimation of Environmental Multimedia Partitioning of Pollutants from Molecular Descriptors using Artificial Neural Networks. SETAC Europe 17th Annual Meeting. Porto, 20-24 May 2007		■	■	
22	Espinosa, G., Rallo, R., Giralt, F., Thomsen, M. Self-organizing classification of Ecotoxicological and Molecular Information as a Proxy for Human Toxicity. SETAC Europe 17th Annual Meeting. Porto, 20-24 May 2007		■	■	■

1.4 References

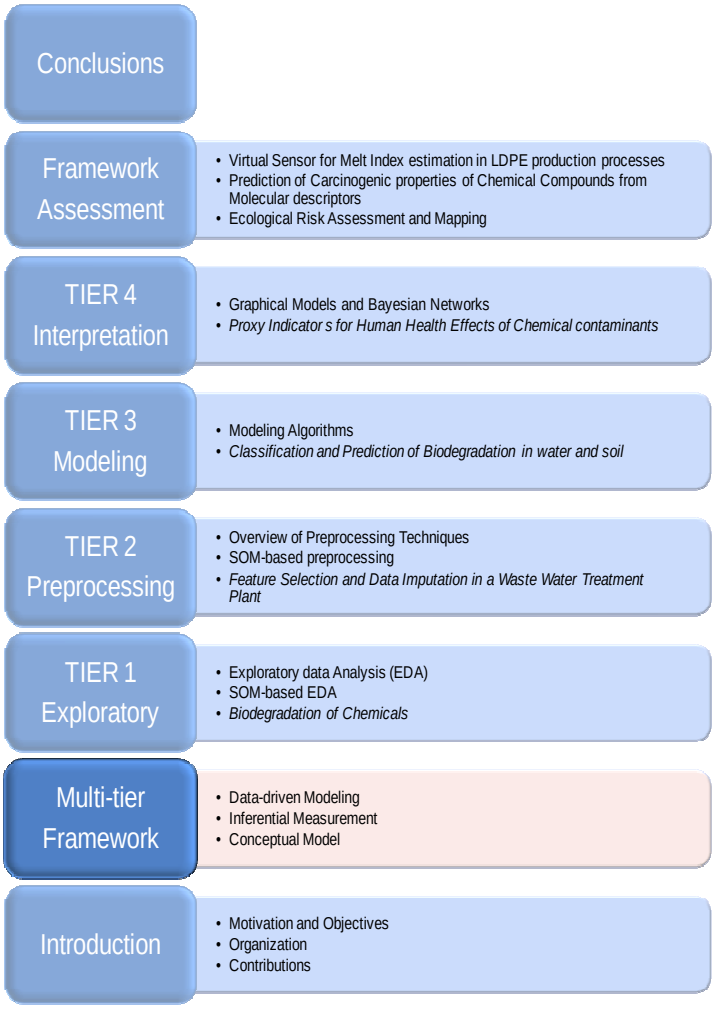
SOLOMATINE, D.P. Data-driven modeling: paradigm, methods, experiences. *Proc. 5th International Conference on Hydrodynamics*, Cardiff, UK. July 1-5, 2002.

BREIMAN, L. Statistical Modeling: The Two Cultures. *Statistical Science* **16**(3):199-231, 2001.

VAPNIK, V. *Statistical Learning Theory*. Wiley, New York, 1998.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 2

Multi-tier Framework for Inferential Measurement and Data-driven Modeling

The development of experimental models, i.e., models developed from data, has proven to be useful in situations where there is no prior knowledge of the complete set of equations governing the modeled phenomenon. In these situations, inferential measurement techniques such as software sensors emerge as a natural extension of data-driven modeling. This chapter introduces a conceptual framework specifically designed to integrate into a sound and coherent architecture all the necessary components that are needed to deploy these systems in different application areas.

2.1 Background

The development of experimental models usually starts by considering these data as being generated by a black box in which a vector **x** of independent variables is mapped to a vector **y** of response variables. Inside this black box the inner dynamics of the modeled process associates predictor variables (system inputs) with response variables (system outputs). These models are known in a generic way as *data-driven models* or *experimental models* because they don't take into account the physics of the phenomenon under study (Solomatine, 2002). A direct application of these models, which has emerged as an independent discipline, is *inferential measurement* (Montague et al., 1990). This technique is very useful in modeling response variables which are difficult to measure. In this section a brief overview of the main techniques and principles related to both data-driven modeling and inferential measurement are presented.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

2.1.1 Data-driven Modeling

One of the objectives of data analysis is to develop accurate predictive models while determining the inner dynamics of the analyzed process. Prediction aims to discover which will be the response values corresponding to never seen before input variables. In contrast, information on system dynamics is used to understand how nature associates input and response variables. Both, prediction and information require the development of models which should be consistent with the underlying processes and principles that govern the system under study. These models which use physical laws or first principles to describe the behavior of the analyzed phenomenon are known as *knowledge-driven models* (KDM) (also referred as behavioral, physical, process or simulation based models).

The first knowledge area used to estimate dependencies from data was statistics, mainly by means of the use of regression techniques. Afterwards, during the decades of the 60's and 70's, new techniques which don't require assumptions concerning the shape and behavior of the underlying statistical distributions started to emerge. Among these techniques were pattern recognition and cluster analysis, neural networks, fuzzy systems, and genetic algorithms. During the decade of the 90's these techniques became very popular due to the availability of new and powerful data capture and analysis mechanisms, as well as to the increased computational power of computers. Nowadays, new generations of algorithms known as machine learning methods have emerged. They constitute the main sources for the development of data-driven models. A machine learning technique (ML) can be defined as an algorithm that estimates an unknown mapping or relationship between the inputs and outputs of a certain system by using only the available data characterizing these inputs and outputs (Mitchell, 1997).

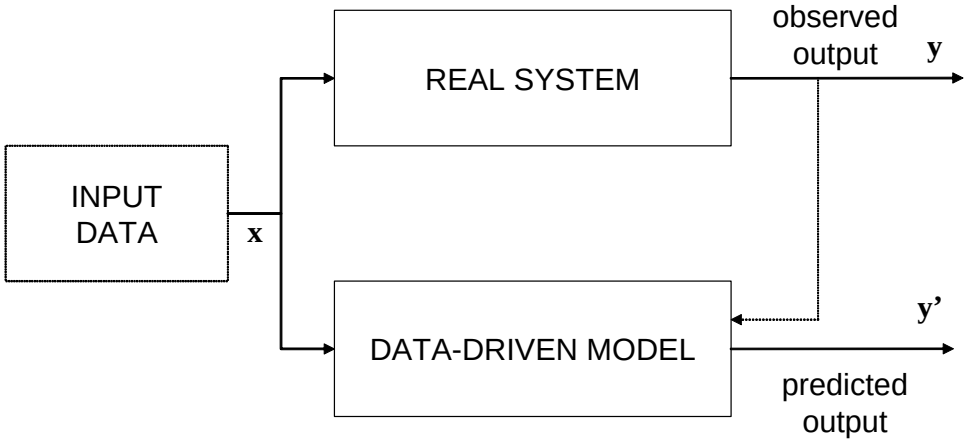


Figure 2.1. Scheme of the interaction between the data-driven model and the modeled physical system

In this context, data-driven modeling (DDM) can be seen as an approach to modeling that focuses on using machine learning techniques to build the models of the systems under study. The aim of DDM is to complement or completely replace knowledge-driven models (Hua et al., 2006). Data-driven models have been widely

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

used in several social and engineering disciplines ranging from the construction of domain specific data-driven simulators (Pidd, 1992) or data-driven predictive control (Wang et al., 2007; Kadali et al. 2003) to data-driven prediction of sports results (Joseph et al., 2006).

Figure 2.1 depicts a basic scheme for the development of data-driven models. Since these models mimic the behavior of the real physical system predicted outputs are eventually corrected by using the system responses in a feed-back loop.

The most important components in data-driven modeling tasks are data. Data are usually grouped into two main families, nominal data representing classes and real-valued data representing magnitudes of variables. Depending on the nature of data, the models developed will have different characteristics and the techniques applied to build these models should adapt to the intrinsic properties of data. In this context, two main kinds of models can be distinguished according to data:

- *Models driven by nominal data.* The main purpose of these models is to cluster or to classify data. Classification is a *supervised* process which consists in finding coherent groups (classes) of data points. The main characteristics of these classes are that data points belonging to the same class are closer to each other in some sense, while classes should be the farthest apart from each other. In contrast clustering is an *unsupervised* process which consists in partitioning a data set into subsets whose elements share common traits, often proximity, according to some distance measure. The most important methods currently used include partition-based clustering algorithms, Self-organizing maps, Bayesian classifiers, decision trees, and support vector classifiers.
- *Models driven by real-valued data.* Most problem formulations in engineering use real-valued data. The task of prediction of these data is often referred to as a regression problem. Since machine learning aims at finding a function that would best approximate a given set of data, it can also be seen as a function fitting problem. Thus, all the numerical fitting methods already available such as linear regression, polynomial functions or splines can be used. The current trends in DDM focus in the use of a combination of many simple functions to develop predictive models with good accuracy. Among the most currently used methods the following can be mentioned: artificial neural networks (radial basis functions (RBF) and multilayer perceptrons), regression trees, instance-based learning, locally weighted regression, and fuzzy-rule based systems.

2.1.2 Inferential measurement and software sensors

Measurements are used to monitor and ultimately, to control and optimize processes. In this context, a specific application of data-driven modeling aimed to infer measurements in situations in which it is difficult or even impossible to determine the value of the response variable arises. These circumstances are frequent in product quality control where quality indicator measurements are mostly available off-line as a result of complex laboratory analysis. In such situations, it would be useful to have a fast and reliable estimation of these indicators in order to reduce the off-specification products, particularly during changes in the operation region of the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

production process. As a result, inferential measurement techniques could play a key role to model, control and optimize these processes. Measurement difficulties can be caused by a variety of reasons being the most common, the lack of appropriate on-line instrumentations and/or its low reliability due to external factors such as fouling, drifts or long delays in response.

The behavior of any process is indicated by the states of certain output variables, which are dependent on the operating conditions and the adjustments made to the state of the process. However, productivity is quantified using only a subset of these output variables; normally the specifications upon which the product are sold, e.g. purity, physical or chemical properties, etc. These so called *primary variables* are often the ones that are difficult to measure on-line. Inferential measurement systems are designed to overcome such measurement problems. On the other hand, the remaining system outputs (for instance temperatures, flows and pressures), named *secondary variables,* can be easily measured on-line. The success of inferential measurement systems thus depends on how well the relationship of a primary output can be modeled with the corresponding secondary outputs and system inputs. The model can then be used to generate estimates of the primary output variables at the rate at which the inputs and secondary variables are measured. For example, an inferential measurement system could be returning estimates of product compositions every few seconds by using direct measurements of temperatures and flow rates instead of waiting 30 minutes for a gas chromatograph to complete its analysis.

The procedure for building an inferential measurement system is essentially that of developing a model that relates a primary or quality variable to other, more easily measured secondary variables. Thus any modeling paradigm may be employed, including the development of first principles models. However, in the present work we will focus only on data-based modeling methods, since the development of first principles models for most of industrial and natural scale processes or biological systems would be very difficult and time-consuming. Data-based inferential measurement systems have been traditionally developed using statistical based methods, time-series analysis or machine learning algorithms (artificial neural networks, genetic programming, and fuzzy systems). Inferential models developed using machine learning paradigms are capable of capturing non-linear process characteristics (Zadeh, 1997).

Inferential measurement techniques induced the development of the concept of software (or virtual) sensor (Tham et al. 1991; Martin, 1997). A software sensor can be conceived as the association of a sensor (hardware), which measures on-line some process variables, with an estimation algorithm (software) which delivers real-time estimates of unmeasured process variables (Cheruy, 1997). The hardware sensors that measure on-line secondary variables have to provide measurements that are reliable and informative enough to deduce unmeasured primary process variables. The estimation algorithm has also to be designed to predict on-line values which quickly converge towards the real values of the unmeasured variables, whatever the process disturbances are. The synthesis of an algorithm offering such guarantees is not always possible, and heavily depends on the available process knowledge and in the reliability of the data used for training.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Software sensors should be conceived at the highest cognitive level of abstraction so that a sufficiently accurate characterization of the global system behavior could be attained in terms of errors between the validated or measured data and the corresponding predicted outputs. Artificial neural networks constitute an adequate choice to develop such systems because, in addition to the above, are able to improve its performance with time, i.e., are capable of learning real cause-effect relations between sensor's stimulus and its response when historical databases of the whole process are used for training. Figure 2.1 presents a generic virtual sensor implementation integrated in a manufacturing process. It can receive real-time readings of several process variables as well as feedback signals of downstream on-line analyzers for the target property. Both sets of data are needed for the initial training of the virtual sensor. Once trained, this virtual device uses only real time measurements of the selected process variables obtained by process sensors at certain times to infer the value of the product target property. The output can be redirected as information to the plant operator or to the control system to maintain optimal plant operation for a given product quality.

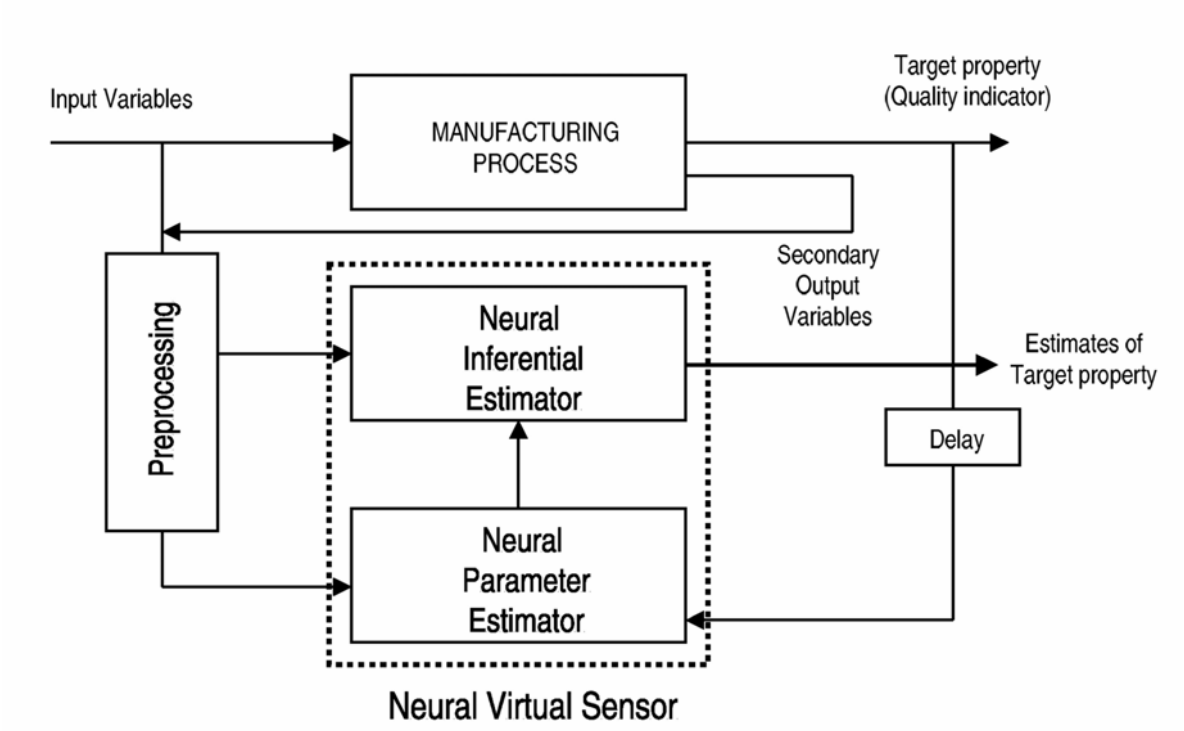


Figure 2.2. Architecture of an inferential measurement system using software sensors

For most industrial applications, software sensor estimates will not be as accurate as measurements provided by a carefully tuned physical sensor. If the virtual sensor is designed to replace the real sensor the user should be prepared to encounter an accuracy loss (Masson et al., 1999). The main purpose of a software sensor is to give predictions of laboratory data, to provide estimates when data are missing and to act as a diagnostic system to detect failures in the physical sensor. Since a software

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

sensor is mainly developed using a data-driven modeling paradigm, it does not learn the physics of the system but the behavior of the physical sensor which must be first installed and tuned to provide the initial training samples.

Soft sensors are mainly developed using machine learning algorithms and following the so called black-box modeling approach. Under this assumption the explained variable y for any plausible value of the explicative variables $\boldsymbol{x}$ is approximated by some function g($\boldsymbol{x}$). To achieve this, a data set $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, and a loss function l are needed. The data set acts as the training set from which the inference process is carried out, and the loss function provides a quantitative measure of the cost of errors. In this context, generalizing means achieving a small average loss on values not present in the training set. Since predictions are supposed to be obtained from examples drawn from the data distribution of the whole learning set p, generalization is measured by the mean loss (or prediction error, PE),

$$PE(g) = \int l\big(y, g(\boldsymbol{x})\big) dp(\boldsymbol{x}, y) \tag{2.1}$$

As the distribution p is unknown, the prediction error in Eq. (2.1) cannot be computed. The minimization of Eq. (2.1) using the empirical risk minimization principle is equivalent to minimize,

$$Risk(g) = \frac{1}{l}\sum_{i=1}^{l} l\big(y, g(\boldsymbol{x}_i)\big) \tag{2.2}$$

which is constructed on the basis of the training set S. If the loss function l is chosen to be quadratic, the least squares method is obtained,

$$Risk(g) = \frac{1}{l}\sum_{i=1}^{l}\big(y_i - g(\boldsymbol{x}_i)\big)^2 \tag{2.3}$$

Minimizing Eq. (2.3) is equivalent to estimate the expected value of y given $\boldsymbol{x}$ through some unknown regression function. As a consequence, the function g should be flexible enough to be able to approximate a broad family of functions. Usually g is modeled using kernel methods, spline smoothing or neural networks. The choice of one particular method is motivated by the intrinsic properties of the problem. For instance, kernel methods or spline smoothing are used when there are only a few explicative variables. In contrast, neural networks are a convenient approach if both the dimension of the input space and the size of the training set are large.

The major pitfall of these adaptive methods resides in the misuse of their flexibility. The more flexible the model is, the greater is its ability to approach any function (universal approximation properties), but simultaneously the more unstable is the estimation from a finite set of data. This is known as the bias/variance trade-off (or stability-plasticity dilemma) which is addressed by using techniques to control the model complexity, i.e., the number of model parameters. Complexity tuning avoids undesired effects such as over fitting the training data which results in the lack of model's generalization capabilities. This process is generally carried out by estimating the prediction or generalization error in Eq. (2.1). The empirical risk of Eq. (2.2) constitutes a biased estimate of the prediction error. The error on an independent

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

validation set results in an unbiased estimate of PE, but its computation requires putting aside a part of the training set for complexity tuning and this is not always possible, particularly when the available number of training cases is small. To overcome this limitation, the use of resampling techniques provides a large validation set for complexity tuning, while the whole training set can still be used to calibrate the software sensor. The two most commonly used resampling techniques are cross-validation (leave-one-out or leave-many-out) and bootstrap methods. It should be noted that the validation set is a part of the training set used to optimize the parameters of the model and should not be confused with the test set which is used to assess the generalization capabilities of the model for unseen data.

It should be finally be noted that in practical applications software sensors can not only provide a point wise estimate of the system response, but also an accuracy index such as a confidence interval or some uncertainty measure of its inferred value. Additionally, a software sensor should incorporate a self-diagnosis system that should alert the user when the operating conditions of the sensor are outside its training range, i.e., when the soft sensor is extrapolating outside the domain of the input space. This is mainly due to the fact that a faithful extrapolation cannot be guaranteed in data-driven models and its predictions should not be considered valid when input data are far from previously seen cases.

2.2 Conceptual Model

The first step in the design of a general purpose framework to implement inferential measurement systems consists in the design of a conceptual model including all the required elements to develop models from data. The conceptual model proposed in this thesis is inspired in the tiered scheme used in data mining frameworks (Fayyad et al., 1996). In these models, each component of the framework specializes in certain data processing tasks, ranging from low level knowledge extraction for raw data management (data conditioning and preprocessing to higher levels of abstraction where the extracted knowledge is interpreted in a form suitable for its use.

Figure 2.3 depicts the four tiers of the proposed conceptual model ordered according to abstraction level. The data that feeds the model are the raw values which describe the modeled process. They include both explanatory variables (inputs) and process response variables (outputs). The lowest level of abstraction corresponds to the tier responsible for the exploratory analysis of the input data space, including its mapping and visual representation. The second tier in abstraction level deals with the preprocessing of data which is a necessary step before attempting to build any experimental model. Data need to be normalized and scaled. Also, the most suitable input variables and examples must be selected to produce accurate models with good generalization capabilities. Moving upwards in the architecture of the system modeling tasks are included at the third level of abstraction. This modeling tier includes both classification and prediction algorithms. Finally, at the highest level of abstraction, the outputs of the inferential model are elaborated in order to obtain useful knowledge about the principles or rules that govern the inner dynamics of the process under examination. This is accomplished by using a

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

probabilistic reasoning component capable of extracting cause and effect relationships between model variables and to provide explanations for these causal relationships.

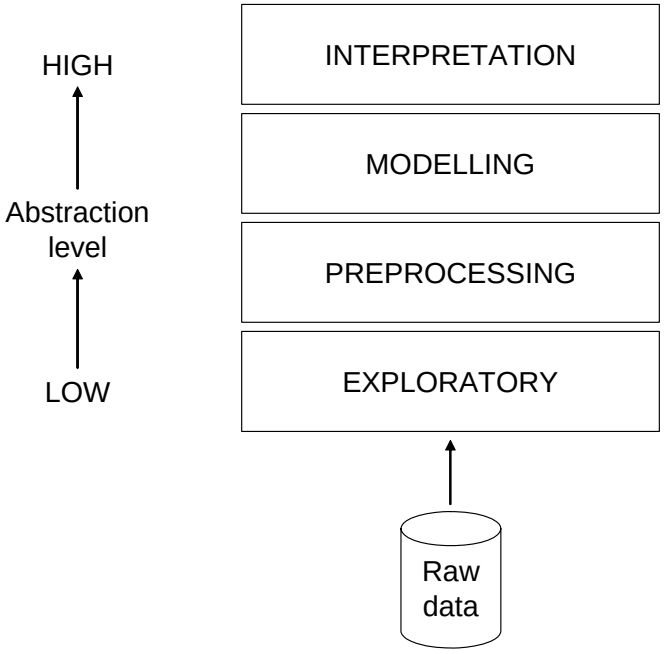


Figure 2.3. Conceptual model for the proposed multi-tier framework for inferential measurement and data-driven modeling

The proposed model architecture of Figure 2.3 has been implemented and validated using a twofold approach. First each tier has been characterized and assessed in terms of some real world application. Second, the whole framework has been assessed on three different application domains to prove its robustness and adaptability for engineering and scientific applications.

Figure 2.4 summarizes the aforementioned conceptual model focusing on the techniques and applications proposed at each tier level. The functional description of the exploratory level includes the mapping of the input space and the visualization and monitoring of system states. The main tasks performed at this tier are normalization and scaling of data, detection of redundant information and the visualization of relationships between system variables. The techniques integrated at this level are exploratory statistics methods and self-organizing maps. The preprocessing tier deals with two main functional tasks, the detection of relevant features and patterns, and the reconstruction of missing data. The tasks performed at this level include the generation of optimized train and test sets, the detection of relevant and irrelevant information, and techniques for data augmentation and imputation. The algorithms used at this level include well-known machine learning techniques for feature selection, a new SOM-dissimilarity based feature selection algorithm and SOM-based techniques for single and multiple data imputation. The next tier is related to the development of data-driven models using both classification and prediction algorithms. The functions of this tier include the development of software sensors and specific data-driven models. The techniques

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

used are a new dynamic radial basis function algorithm which uses a competitive learning module to adapt the centers and widths of the Gaussians, a modified Fuzzy ARTMAP network, kernel methods based on the use of support vector machines, and ensembles of all these models. Each of these techniques is assessed to evaluate its capabilities in diverse application domains.

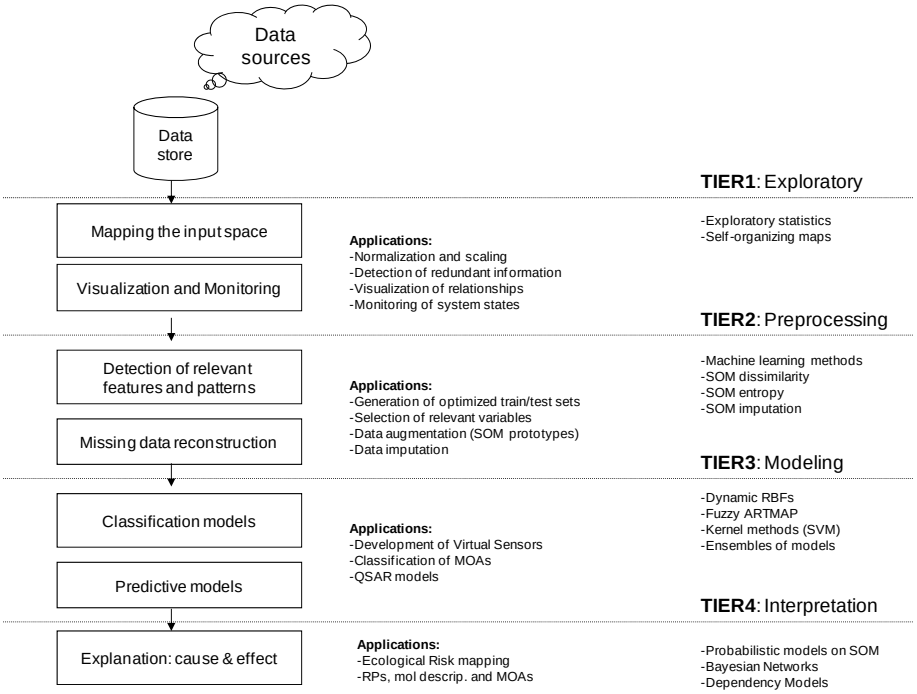


Figure 2.4. Detailed conceptual model including the functional description of each tier and the machine learning techniques used and the applications proposed for its assessment

Finally, the fourth tier covers the interpretation of parameters and structure of the models with the purpose of extracting robust cause and effect relationships from model outputs. The approach followed includes probabilistic dependency analysis, graphical models such as Bayesian networks, and probabilistic models on top of SOM with Gaussian mixture modeling.

2.3 References

BHARTIYA, S., WHITELEY, J.R. Development of inferential measurements using neural networks. *ISA Transactions*, **40**(4):307-323, 2001.

CHAN C.W. (2003) Editorial: Special issue on data-driven modelling methods and their applications. *International Journal of Systems Science*, **34**(14-15):731-732, 2003.

CHÉRUY, A. Software sensors in bioprocess engineering. *Journal of Biotechnology*, **52**:193-199, 1997.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

FAYYAD, U., PIATETSKY-SHAPIRO, G. SMYTH, P. From Data mining to Knowledge Discovery in Databases. *AI Magazine*, **17**:37-54, 1996.

HUA F., HAUTANIEMI S., YOKOO R., LAUFFENBURGER D.A. Integrated mechanistic and data-driven modelling for multivariate analysis of signalling pathways. *Journal of the Royal Society Interface* **3**(9):515-526, 2006.

JOSEPH A., FENTON N.E., NEIL M. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-based Systems* **19**(7):544-553, 2006.

KADALI R., HUANG B., ROSSITER A. A data driven subspace approach to predictive controller design. *Control Engineering Practice* **11**(3):261-278, 2003.

MARTIN, G. Consider Soft Sensors. *Chem. Eng. Progress*, **7**:66-70, 1977.

MASSON, M.H., CANU, S., GRANDVALET, Y., LYNGGAARD-JENSEN, A. Software sensor design based on empirical data. *Ecological Modelling*, 120:131-139, 1999.

MITCHELL, T. *Machine Learning*. McGraw-Hill. 1997.

MONTAGUE, G.A., THAM, M.T. AND LANT, P.A. Estimating the immeasurable without mechanistic models. *Trends in Biotechnology*, **8**(3):82-83, 1990.

PIDD, M. Guidelines for the Design of Data-driven generic simulator for specific domains. *Simulation*, **59**(4):237-243, 1992.

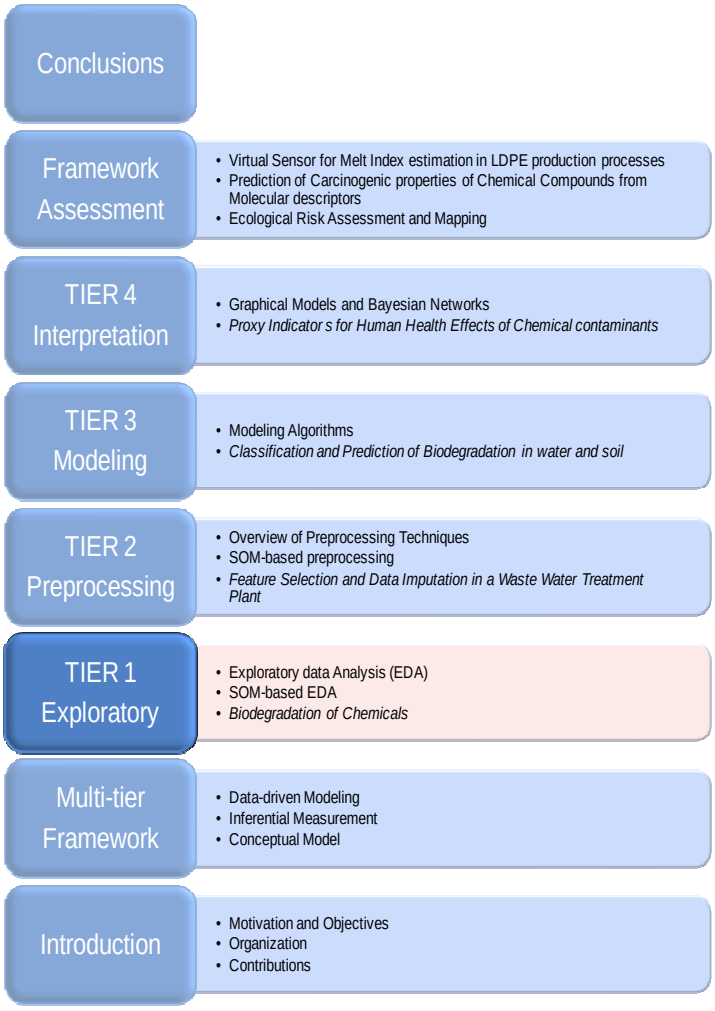
THAM, M.T., MONTAGUE, G.A., MORRIS, A.J., LANT, P.A. Soft-sensors for process estimation and inferential control. *Journal of Process Control*, **1**:3-14, 1991.

WANG, X., HUANG, B., CHEN, T. Data-driven predictive control for solid oxide fuel cells. *Journal of Process Control*, **17**(2):103-114, 2007.

ZADEH L. A. What is soft computing? *Soft Computing*, **1**:1–2, 1997.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 3

Tier 1: Exploratory Level

Exploratory analysis of data is a crucial step for the development of algorithmic models. This chapter covers the basic aspects of data analysis and visualization and introduces a set of techniques that apply the Self-organizing Map (SOM) to explore and visualize relationships among data in high dimensional spaces. This set of techniques conform the core of the first tier of the proposed framework.

3.1. Exploratory Data Analysis (EDA)

Algorithmic modeling as a process identification technique also relies heavily on data. For that reason it is essential to explore the information space in order to gain knowledge about the structure and relationships among the variables describing the process being analyzed. In this context, exploratory data analysis (EDA) plays a central role to extract knowledge from data (Tukey, 1977; Velleman and Hoaglin, 1981). EDA techniques are a specific part of statistics which is mainly concerned with reviewing, communicating and using data in systems where there is a low level of mechanistic knowledge. The EDA term was coined by John Tukey (1997) and its main objectives can be summarized as:

- state hypotheses about the causes of observed phenomena;
- assess assumptions on which statistical inference will be based;
- support the selection of appropriate statistical tools and techniques; and
- provide a basis for further data collection through experiment design.

EDA aims at providing tools to facilitate the extraction of meaningful information from data, i.e., detecting and uncovering the underlying structure and relationships present in data. EDA is in its nature a descriptive technique which provides a qualitative “feel” about data which comes almost exclusively from the application of different graphical techniques. To get this "feel", it is not enough to know what is contained in the data; it would be also important to know what is not in the data. The most efficient way to accomplish this task is to use our own human pattern-

ISBN:978-84-691-1008-9/DL: T.2230-2007

recognition and comparative abilities by means of the graphical techniques provided by EDA.

3.1.1 The process of EDA

Exploratory data analysis usually involves several processes such as: (i) identification of inappropriate and suspicious attributes, (ii) selection of the most appropriate attribute representation, (iii) generation of derived attributes, and (iv) the selection of the optimal subset of attributes (Becher et al., 2000).

Inappropriate attributes should be removed prior to the analysis of data. These attributes can be classified depending into the following four types according to their characteristics.

- **Constant**: only a single value is absent.
- **Null**: all values of the attribute are missing.
- **Near null**: the fraction of missing values is larger than a specified threshold.
- **Many values**: the fraction or unique values is larger than a specified threshold.

In contrast, suspicious attributes should be kept unless the modeler decides to remove them. Suspicious attributes fall into one of the following categories:

- **Artifact**. Their dependency on the target variable is larger than a threshold value. These attributes, if included, may lead to artificially good models with poor generalization capabilities.
- **Poor predictor**. Their dependency with the target variable is less than a threshold value. These attributes don't contribute much by themselves but their combinations may increase the predictive power of the whole model. Attribute selection techniques are the ultimate way to decide whether or not these variables should be included into the model.
- **Near constant**. A single value of an attribute covers more than a specified fraction of cases.
- **Few values**. The attribute has less than a specified number of distinct values.
- **Few cases**. The attribute has less than a specified number of distinct non-null cases.

After this preliminary filtering step, the remaining attributes must be processed to determine their most appropriate representation. This procedure includes the treatment of outliers and missing values, and a data encoding suitable for the subsequent modeling task. Continuous attributes may be discretized to transform native values into a small number of value ranges. On the other hand, categorical attributes can be grouped by merging several categories together. As these transformations can cause the loss of some of the detail contained in the original

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

attributes a balance among the encoding efficiency and the information loss is always needed. This is usually accomplished by using Target Dependency Analysis (TDA) techniques. The main goal of TDA is to find generic measures of associations between input data and the corresponding target attributes, looking at the strength and variation of these relationships. These measures can be considered as a generalization of the linear concept of correlation and may include, among other measures, mutual information, and Cramer's V and Goodman-Kruskal indices.

Formally the mutual information of two discrete random variables X and Y is defined as,

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \qquad (3.1)$$

where $p(x,y)$ is the joint probability distribution function for X and Y, and p(x) and p(y) are the marginal distribution functions for X and Y respectively. Intuitively, mutual information measures the information that X and Y share; it measures how much knowing one of these variables reduces our uncertainty about the other. For example, if X and Y are independent, knowing X does not give any information about Y and vice versa, and their mutual information is zero. In the opposite case, when the two variables are equivalent, knowing one of them determines de value of the other.

Cramer's V (Cramer, 1999) is a statistic measure of the strength of association or dependency between two (nominal) categorical variables in a contingency table. Goodman and Kruskal (Goodman and Kruskal, 1954) defined the lambda index as a measure of association between variables in a cross-tabulation table.

Finally, new attributes can be derived using a variety of univariate or multivariate transformations. In some situations these derived attributes may be more beneficial than the original ones. Some of the most usual transformations include quadratic, inverse, exponent, logarithm, power and square root functions. If a given transformation increases the dependency with the target beyond a specified threshold the derived attributes are used instead of the original ones.

3.1.2 Overview of EDA Graphical techniques.

Most EDA techniques are graphical in nature and a few quantitative. The reason for the heavy reliance on graphics is that the main role of EDA is to open-mindedly explore data. The use of graphics helps revealing the structural relationships of data and offers new, often unsuspected, insight. The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques.

- Plotting native data (such as data traces, histograms, bi-histograms, probability plots, lag plots, block plots, and Youden plots).
- Plotting simple statistics such as the mean, standard deviation, etc. of native data.

ISBN:978-84-691-1008-9/DL: T.2230-2007

- Positioning such plots to enhance pattern-recognition abilities, e.g., including multiple plots per page.

The most used graphic representations in EDA are (Chambers, 1983):

Run-sequence Plot. These plots constitute an easy way to graphically summarize a univariate data set. A common assumption of univariate data sets is that they behave like random drawings from a fixed distribution with a common location (mean) and with a common scale (variance). With run sequence plots, shifts in location and scale are typically quite evident. Also, outliers can easily be detected.

Lag plot. Data are plotted against previous values in the data set. A lag plot checks whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot and vice versa.

Histogram. The purpose of a histogram is to graphically summarize the distribution in univariate data sets. An histogram of data graphically shows their central location, spread or scale, skewness, presence of outliers, and presence of multiple modes. These features provide strong indications of the proper distributional model for the data. The probability plot or a goodness-of-fit test can be used to verify the distributional model. The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then for each bin, the number of points from the data set that fall into each bin is counted. The cumulative histogram is a variation of the histogram in which the vertical axis gives not just the counts for a single bin, but rather the counts for that bin plus all bins for smaller values of the response variable. Also, the counts can be replaced by normalized counts in both the histogram and cumulative histogram. The names for these two variants are the relative histogram and the relative cumulative histogram. A commonly employed normalization technique consists in setting the area (integral) under the histogram equal to one. In this case, the normalized count is the count in the class divided by the number of observations times the class width. From a probabilistic point of view, this normalization results in a relative histogram that is similar to the probability density function and a relative cumulative histogram that is analogous to the cumulative distribution function.

Normal Probability Plots. The normal probability plot is a graphical technique for assessing whether or not a data set is approximately normally distributed. Data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. One advantage of this method of computing probability plots is that the intercept and slope estimates of the fitted line are in fact estimates for the location and scale parameters of the distribution. Although this is not too important for the normal distribution since the location and scale are estimated by the mean and standard deviation, respectively, it can be useful for many other distributions. The correlation coefficient of the points on the normal probability plot can be compared to a table of critical values to provide a formal test of the hypothesis that the data are normally distributed.

4-Plot. This plot constitutes the most common tool used in EDA. It summarizes the main characteristics of data by grouping a combination of the above four plots.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Figure 3.1 shows a 4-plot graph corresponding to the exploratory analysis of biodegradation data introduced in section 3.3. The run sequence plot confirms that data spans in the range [0,100] and doesn't detect the presence of outliers. In addition there aren't identifiable structures in the lag plot corroborating that BOD values corresponding to all chemicals in the MITI-1 data set are independent. The histogram depicts a strongly skewed distribution with most of the values in the low range of biodegradation. Also, the presence of a two mode distribution can be identified in this graph. Finally, the normal probability plot confirms that these data are not normally distributed.

Bootstrap plots. To generate a bootstrap uncertainty estimate for a given statistic from a set of data, a subsample of a size less than or equal to the size of the data set is generated from the data and the statistic calculated. This subsample is generated with replacement so that any data point can be sampled multiple times or not sampled at all. This process is repeated for many subsamples, typically between 500 and 1000. The computed values for the statistic form an estimate of the sampling distribution of the statistic.

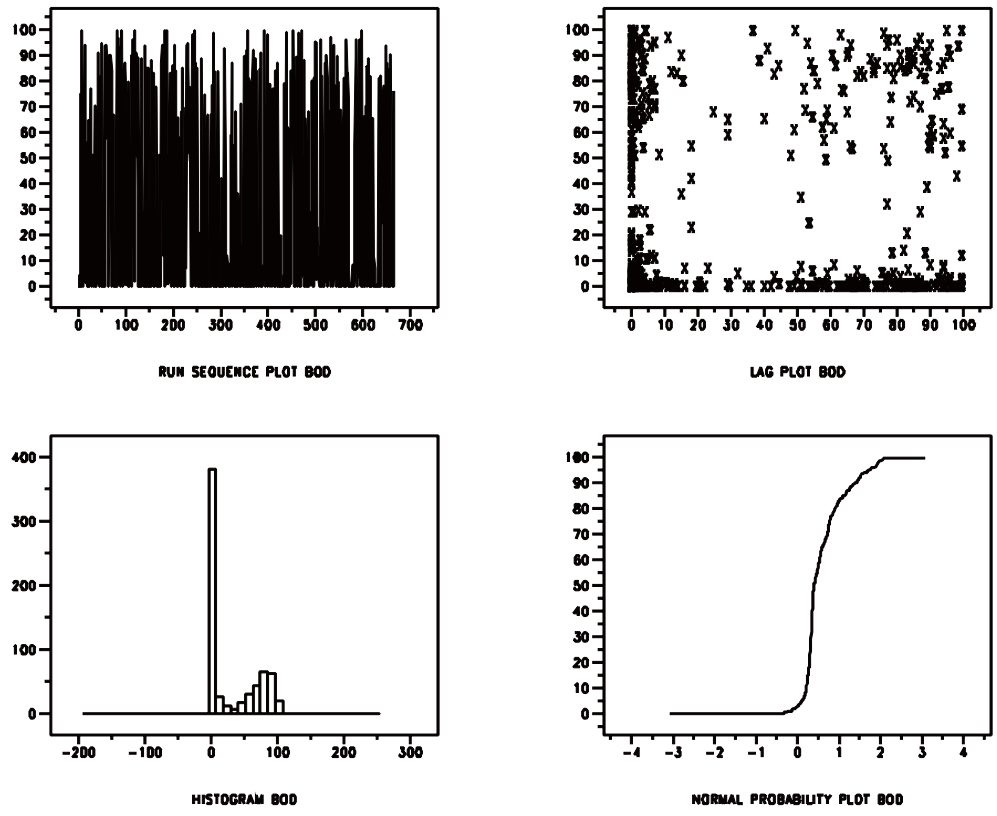


Figure 3.1. 4-plot corresponding to the exploratory data analysis of MITI-1 biodegradation data (see subsection 3 for details)

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

3.2 Self-Organizing Maps as a tool for Exploratory Data Analysis

The Self-Organizing Map (Kohonen, 1990) is with the backpropagation algorithm one of the most popular and widely used neural network architecture. It is a powerful tool for visualization and data analysis that is used in various application domains ranging from engineering to economics and social sciences. While EDA is mainly a visualization approach to data analysis, the self-organizing maps (SOM) fits perfectly in this context. In this section we give an overview of the SOM algorithm and its visualization capabilities.

3.2.1 The Self–Organizing Map

The Self-Organizing Map algorithm performs a topology preserving mapping from a high-dimensional input space onto a low dimensional output space formed by a regular grid of map units. This neural model is biologically plausible (Kohonen, 1993) and is present in various brains' structures to provide an ordered low-dimension internal model of the external environment. From a functional point-of-view, SOM resembles the *Vector Quantization* (VQ) algorithms (Linde et al., 1980) which approximates, in an unsupervised way, the probability density functions of a vector of input variables by a finite set of reference vectors with the only purpose of describing its class borders by using a nearest-neighbor rule. In contrast, SOM's units are organized over the space spanned by a regular grid of processing units in which the adaptation process affects some predefined topological neighborhood producing both, a vector quantization and an ordered representation of the original input data. In addition, since each map unit has well-defined low-dimensional coordinates over the map grid, the SOM can also be considered as a projection algorithm.

3.2.1.1 Fundamentals of the SOM algorithm

The mapping implemented by the SOM algorithm can be formalized in the following way. Assume that a set of input variables $\{\xi_j\}$ is defined as a real vector $x = [\xi_1, \xi_2, \dots, \xi_n]^T \in \Re^n$, and that each element in the SOM array has associated a parametric reference vector $m_i = [\mu_{i1}, \mu_{i2}, \dots, \mu_{in}]^T \in \Re^n$. Using this mapping, the image of an input vector $\mathbf{x}$ on the SOM array is defined by a function which is dependent on a measure of distance between $\mathbf{x}$ and $\mathbf{m}_i$ that can be defined as,

$$c = arg\min_i\{d(x, m_i)\} \qquad (3.2)$$

where $d(x, m_i)$, denotes a general distance measure and c is the index of a unit in the SOM array. The goal is to define $\mathbf{m}_i$ in such a way that the mapping is ordered and descriptive of the distribution of $\mathbf{x}$.

A common approach used to determine the proper set of $\mathbf{m}_i$ values follows an optimization process which uses the concept of "*vector quantization*" (VQ) (Gray,

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

1984). In this approach, a finite set of codebook vectors $\{m_i\}$ is placed into the space of the x input patterns to approximate them by minimizing some reconstruction error measure. Let $p(x)$ be the probability density function of x, and let m_c be the codebook vector which is *closest* to x in the input space, i.e., the one for which $d(x, m_c)$ is smallest. The VQ procedure minimizes the average expected quantization error (reconstruction error) which can be expressed by,

$$E = \int f[d(x, m_c)]p(x)d(x) \qquad (3.3)$$

where f is some monotonically increasing function of distance d. It should be noted that the index c is a function of $\mathbf{x}$ and m_i, whereby the integrand part of Eq. (3.3) is not continuously differentiable, i.e., c changes abruptly when crossing a border in the signal space where two codebook vectors have the same value for its distance function.

The set of vectors $\{m_i\}$ which minimizes E in Eq. (3.3) is the solution for a VQ problem, and the input space is mapped to this set of codebook vectors. If f is chosen as $f[d(x, m_c)] = \|x - m_c\|^r$, then the optimal placement of vectors $\{m_i\}$ results in their point density which is proportional to $[p(x)]^{\frac{n}{n+r}}$, where n is the dimension of the input space and $x \in \mathcal{R}^n$ (Zador, 1982).

It should be noted, however, that the indexing of these values is made in an arbitrary way and results in an unordered representation. A new formulation of the average quantization error function in Eq. (3.3) was introduced by Kohonen (1991) by adding a smoothing kernel h_{ci} which is a function of the distance between units i and c:

$$E' = \int \sum_i h_{ci} f[d(x, m_i)]p(x)\, dx \qquad (3.4)$$

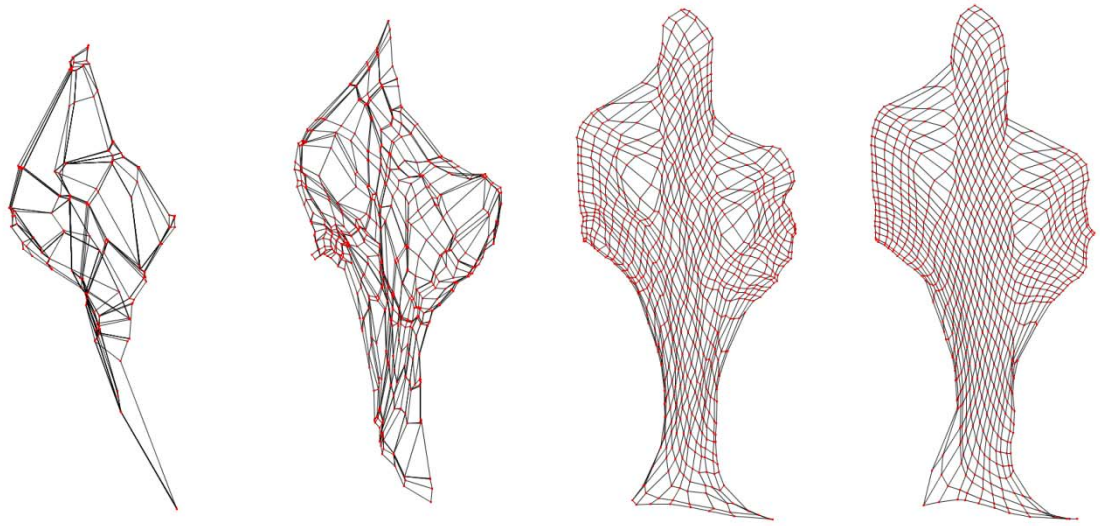


Figure 3.2. Elastic net corresponding to a self-organizing map. Adaptation on a 2D benchmark corresponding to a point distribution with a cactus-like shape

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The minimization of Eq. (3.4) imposes an ordering relation on the values of m_i as if these vectors where lying at the nodes of an elastic net (see Figure 3.2) fitted to the density probability distribution $p(x)$ of the input space.

Even in the most fundamental cases, the minimization of Eq. (3.4) constitutes a complicated non-linear optimization problem for which in most cases no closed solutions are known. Hence, approximation algorithms must be used.

One of the algorithms used to minimize E', which gives a fairly good approximation for the set of $\{m_i\}$, is the so-called *stochastic approximation* method (Robbins and Monro, 1951). Let us consider $x = x(t)$ at a discrete iteration time step t. Let $m_i(t)$ be the approximation of m_i at a time t, and consider a sample function $E''(t)$ defined as,

$$E''(t) = \sum_i h_{ci}\, f[d(x(t), m_i(t))] \tag{3.5}$$

An approximate optimization using gradient descent methods is used to minimize Eq. (3.5). Starting with initial values $m_i(0)$, all reference vectors are updated according to,

$$m_i(t+1) = m_i(t) - \left(\frac{1}{2}\right)\lambda(t)\frac{\partial E''}{\partial m_i(t)} \tag{3.6}$$

where $\lambda(t)$ is a small positive scalar factor that determines the size of the gradient step at time *t*. If this function is chosen appropriately, the sequence of $m_i(t)$ will always converge to a set of $\{m_i^*\}$ values which will approximate the solution $\{m_i\}$. Although this procedure does not guarantee that a global minimum is achieved, a local minimum reached provides a sufficiently good approximation in many applications. If required, better local minima can be found by repeating this procedure with different starting values or with the help of advanced optimization techniques such as "*simulated annealing*".

Many different forms of SOM algorithms can be expressed by Eq. (3.6). If *d* is defined by the Euclidean norm $d(x, m_i) = \|x - m_i\|$ and $f(d) = d^2$ the traditional SOM algorithm is obtained, which can be expressed with the following update rule:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{3.7}$$

The term $\lambda(t)$ in Eq. 3.6 can be introduced in the neighborhood kernel $h_{ci}(t)$ around the winner unit *c* at time *t*. This kernel which is a non-increasing function of iteration time and the distance of unit *i* to the best matching unit *c* , defines the region of influence that the input pattern has on the SOM. In the original Kohonen's formulation, the kernel was formed by two parts, a neighborhood function $h(d,t)$ and a *learning rate* term α(t),

$$h_{ci}(t) = h(\|r_c - r_i\|, t)\alpha(t) \tag{3.8}$$

where r_i is the location of unit *i* on the map grid and $0 < \alpha(t) < 1$.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The simplest definition of $h_{ci}(t)$ corresponds to the *bubble neighborhood function* which is constant over the whole neighborhood of the winner unit and zero elsewhere. In this case, $h_{ci}(t) = \alpha(t)$ if *i* and *c* are neighboring units, and $h_{ci}(t) = 0$ otherwise. Let's denote the set of these neighboring units as $N_c(t)$. Then, SOM's fundamental updating rules can be expressed as,

$$\begin{cases} m_i(t+1) = m_i(t) + \alpha(t)[x(t) - m_i(t)], & \text{if } i \in N_c(t) \\ m_i(t+1) = m_i(t), & \text{if } i \notin N_c(t) \end{cases} \quad (3.9)$$

Other commonly used neighborhood functions are the Gaussians, which produces slightly better maps while being computationally more expensive. During the training process, the SOM forms an elastic net that folds onto the cloud formed by the input data, trying to approximate the probability density function of the original data by placing more codebook vectors where the data are dense and few units where are sparse.

A variant of the basic SOM which is invariant to the scale of input variables is the so-called *dot product map*. In this approach the dot product $\eta_i = m_i^T x$ between the input vector **x** and each codebook m_i is used as similarity measure. The best matching unit is then selected as the one that maximizes the product, $\eta_c = \max_i\{\eta_i\}$. The updating rule must normalize at each time step the value of the new codebook vector by,

$$m_i(t+1) = \frac{m_i(t) + \alpha(t)x}{\|m_i(t) + \alpha(t)x\|} \quad (3.10)$$

The SOM algorithm can be further simplified considering the following equilibrium condition at the convergence limit:

$$\mathrm{E}\{h_{ci}(x - m_i^*)\} = 0 \quad (3.11)$$

where E is the mathematical expectation operator. This expression can be rewritten as,

$$m_i^* = \frac{\int h_{ci} x p(x)\, dx}{\int h_{ci} p(x)\, dx} \quad (3.12)$$

For instance, if h_{ci} is defined by Eq. (3.9),

$$m_i^* = \frac{\int_{V_i} x p(x)\, dx}{\int_{V_i} p(x)\, dx} \quad (3.13)$$

where V_i represents the domain of vectors **x**, whose nearest codebook vector belongs to the neighborhood set N_i of unit *i*, also known as the Voronoi region. This produces a variant of the SOM algorithm known as *Batch Map*, where the whole data set is presented at once to the map before any adjustment is made. The updating is done by simply replacing the prototype vector with a weighted average over the samples, where the weighting factors are the neighborhood function values h_{ci} by using the following update rule:

ISBN:978-84-691-1008-9/DL: T.2230-2007

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ci(j)}(t) x_j}{\sum_{j=1}^{n} h_{ci(j)}(t)} \qquad (3.14)$$

There exist many other variants of the basic SOM reported in the literature. Possible variations include the use of neuron specific learning rates and neighborhood sizes, and growing map structures. The goal of all these variations is to enable the SOM to follow the topology of the underlying data set better and to achieve good quantization results (Vesanto, 1999). **The most important SOM variants are:**

- **The *Tree Structured SOM* (Koikkalainen** and Oja, 1990) is a fast version of the SOM which consists in a set of layers that perform a complete quantization of the data space. The difference between these layers is the number of codebook vectors which increases exponentially as the tree is traversed downwards. Data from upper layers is used to train lower layers reducing the amount of distance calculations needed to find the winner unit. Each layer provides a more detailed interpretation of the data space.
- The *Minimum Spanning Tree SOM* (Kangas et al. 1990) uses a tree structure as neighborhood function which defines the minimal set of connections needed to link together a related set of codebook vectors. From the quantization point of view the modified algorithm is more stable than the traditional SOM. In contrast, the position of units in the lower dimension space is not fixed and visualization becomes more difficult.
- The *Neural Gas* (Martinetz et al., 1993) is a variant which uses a dynamic neighborhood that changes during the training process.
- The *Growing Cell Structures* (Fritzke, 1994) adds or removes map units as needed during the training process instead of working with a predefined number of codebook vectors.

3.2.1.2 Mathematical Properties of SOM

Although the SOM algorithm is by itself very simple, the mathematical proof of its properties is rather difficult. For instance, the ordering of the map has only been proven for the 1-dimensional case (Erwin et al., 1992). The most important mathematical properties of SOM can be summarized as:

Voronoi Region. The SOM partitions the input space into convex Voronoi regions, each of which corresponds to the influence area of one unit in the map. Figure 3.3 shows the segmentation of the input space into these regions. The Voronoi region of a map unit *i* is formed by the union of all the closest vectors *x*,

$$V_i = \{x \mid \|m_i - x\| < \|m_j - x\|, i \neq j\} \qquad (3.15)$$

Inside this region the reference vectors are placed according to the density distribution of the data weighted by the neighborhood function, i.e., according to equation (3.12).

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

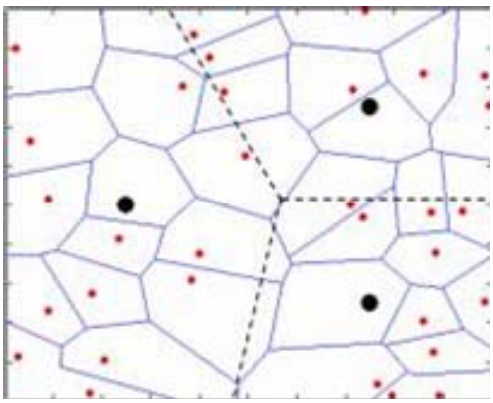


Figure 3.3. Partition of the input space into Voronoi regions. The reference vectors are placed according to the conditional expectation of data weighted by the neighborhood kernel

Convergence. The SOM algorithm has been demonstrated to converge in the 1-dimensional case. In a more realistic application scenario involving a large number of units and a final neighborhood radius R, the asymptotic point density of the reference vectors of the SOM has been shown to be proportional to $p(x)^{\frac{2}{3}-\frac{1}{3R^2+3(R+1)^2}}$, where *p(x)* is the density function of the input patterns **x**. When the dimension of the input vectors increases, the exponent of the density function approaches unity and the distribution of weight vectors estimates that of training data.

Energy function. The original SOM algorithm cannot be directly derived from an energy function in contrast with most artificial neural network algorithms. However, in the case of a discrete data set and a fixed neighborhood function the SOM has been shown to have the following energy function,

$$E = \sum_k \sum_i h_{ci} \|x_k - m_i\|^2 \qquad (3.16)$$

which resembles the energy function corresponding to the k-means vector quantization but takes into account the distance of each input vector to all the reference vectors weighted by the neighborhood kernel. This function (3.16) can be expressed in two parts,

$$E = \sum_k \|x_k - n_c\|^2 + \sum_i \sum_j h_{ij} N_i \|n_i - m_j\|^2 \qquad (3.17)$$

where N_i is the number of data items into the Voronoi region of m_i, and n_i is their centroid defined as $\frac{1}{N_i}\sum_{x \in V_i} x$. The first term in Eq. (3.17) corresponds to the vector quantization quality of the map and is equivalent to the energy function of the k-means algorithm. The second term reflects the ordering quality of the map. It is minimized when nearby map units have weight vectors close to each other in the input space and.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

3.2.1.3 Initialization and Training procedure

The SOM algorithm is based on an unsupervised competitive learning approach. Thus, the training process is entirely data-driven and map units compete to become specific detectors of certain data features. Each map unit is represented by a *n*-dimensional weight vector, where *n* is equal to the dimension of the input space. As in vector quantization, every weight vector describing a class is called a *codebook*. Each unit *i* has a topological neighborhood N_i determined by the form of the SOM grid lattice which can be either rectangular or hexagonal. The number of units as well as their topological relations are defined (and fixed) during the initialization phase. The granularity (size) of the map will determine its subsequent accuracy and generalization capabilities.

During the initialization process four parameters have to be selected: the number of units, dimension of the map grid, map lattice and shape. The number of units should usually be selected to as big enough to accommodate all training data, with the neighborhood size controlling the smoothness and generalization of the mapping. It should be noted that the computational complexity of the algorithm is proportional to $O(n^2)$ being n the number of map units. The use of a hexagonal lattice is usually recommended, because all 6 neighbors of a unit are at the same distance, as opposed to the 8 neighbors in a rectangular lattice configuration. The shape of the map grid should correspond to the shape of the data manifold whenever possible. To avoid the border effects in the mapping process, i.e., units with a reduced neighborhood, a periodic shape such as a torus is used. Figure 3.4 depicts the SOM grid corresponding to a toroidal topology. It can be seen that the clusters (same color spots) are organized in a coherent and continuous way across the grid boundaries.

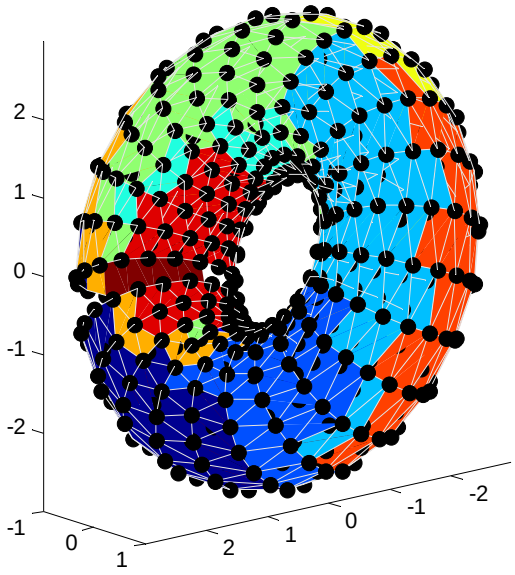


Figure 3.4. Toroidal SOM configuration to minimize "border effects"

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Initial values are assigned to the prototype vectors before training starts. The SOM is very robust with respect to the initialization conditions, but a properly selected starting point facilitates convergence to a good solution. Typically the initialization process can be performed in three different ways:

- random initialization;
- randomly selected samples drawn from the input data set;
- linear approach inspired in Principal Component Analysis (PCA), where initial codebook vectors lie in the same input space that is spanned by the two eigenvectors corresponding to the largest eigenvalues of the input data. This has the effect of stretching the map to the same orientation as the data having the most significant amount of energy.

The training algorithm proceeds as follows. At each training step, one sample input vector **x** is randomly chosen from the training data. Then, similarities (distances) between **x** and the codebook vectors are computed (usually, the Euclidean distance is used) by looking for the *"best matching unit"* (BMU). This similarity matching can be expressed as,

$$\|x - m_{bmu}\| = \min_i\{\|x - m_i\|\} \tag{3.18}$$

After finding the BMU and its topological neighbor cells, their degree of matching is increased by moving their codebook vectors in the proper direction in the input space. Figure 3.5 depicts this adaptive process.

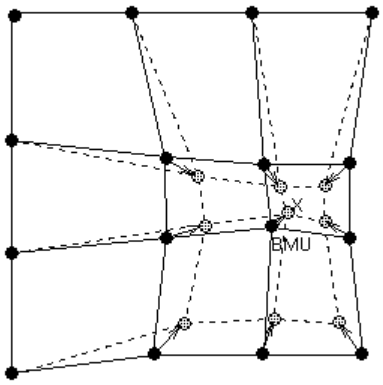


Figure 3.5. Adaptation of the best-matching unit and its neighborhood towards the input sample **x** which is marked with *x*. The solid and dashed lines correspond to the state of the map before and after updating the affected weights with respect to **x**

The codebook update during this competitive learning process follows a winner-takes-all approach described by Eq. (3.9). The tuning parameters of this update rule are the learning rate and the radius of the neighborhood, both being training time dependent and monotonically decreasing.

During the adaptation process two modes of operation should be distinguished. An initial *ordering phase*, in which the map is formed using a relatively large learning rate and a neighborhood radius, and a later *convergence phase*, where fine tuning of

ISBN:978-84-691-1008-9/DL: T.2230-2007

codebook vectors is performed by using smaller parameter values. This approach corresponds to a *Hebbian learning* on the topological neighborhood and active forgetting in the rest of units (Hebb, 1949).

3.2.1.4 SOM Quality Measures

The quality of a SOM can be evaluated from the resolution of the map and from the preservation of the topology of the native data set. The most important issue regarding the accuracy of the SOM projection is the "true" dimension of data. If it is larger than the dimension of the map grid the SOM may not follow the distribution of the data set. In this case topology preservation and map resolution become contradictory goals and cannot be simultaneously attained. When this occurs, a map with high resolution folds into itself and topology is broken.

Topology preservation can be measured by using the topographic error measure proposed by Kiviluoto (1996). It is defined as the percentage of sample vectors for which the two best matching reference vectors are not in adjacent units,

$$\epsilon_t = \frac{1}{N}\sum_{i=1}^{N} u(x_i) \tag{3.19}$$

where N is the cardinality of the data set, and $u(x_i) = 1$ if the first and second BMUs of x_i are not adjacent units and zero otherwise. Other metrics have been reported in the literature that measure topology preservation by using the local smoothness of the SOM.

SOM resolution is measured by means of the average quantization error over the complete data set. Other measures which are independent from the data set used can be defined by computing the average distance of map units from their neighbors.

SOM quality can also be estimated with a combined strategy that mixes both topology and resolution. An example of such approach is the SOM energy function given in Eq. (3.16). Other measures following this approach have been proposed in literature (Kaski and Lagus, 1996) such as the *topological quantization error* measure which combines both strategies to form a measure of the discontinuities in the representation of local data relationships. This measure is based on the computation of the shortest path between the first and second best matching units for all data sample vectors in the map.

3.2.2 Visualization applications of the Self–Organizing Map

SOM can be used to visualize relationships among data once the reference vectors have been ordered in the map grid. There exist a variety of SOM-based visualization techniques that can be used. Most of them exploit the ability to represent data in two dimensions. Figure 3.6 shows the layered organization of a SOM, where every unit provides a partial view of the whole data set for each variable at the corresponding horizontal layers.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

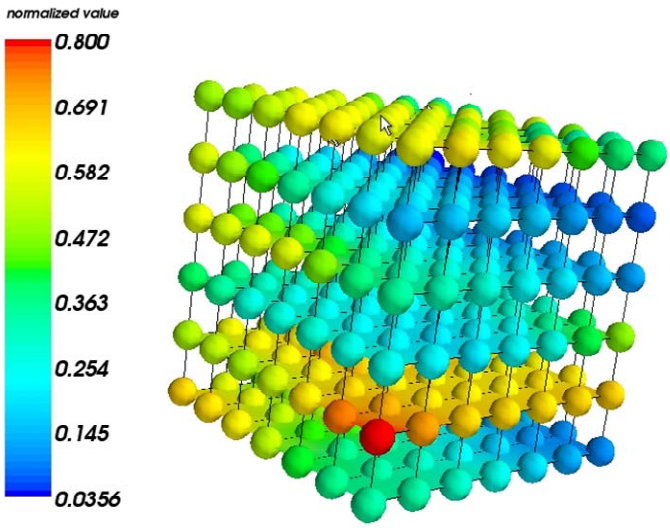


Figure 3.6. Layered SOM organization. Each horizontal layer corresponds to a single variable

The main SOM visualization techniques can be summarized as:

- **Vector Components**. The codebook vectors of the SOM can be visualized using the so-called component planes (or *c-planes*), i.e., horizontal layers in Figure 3.6. In this representation, the SOM is considered as formed by a set of stacked layers. Each component plane forms a horizontal layer in this structure while each codebook vector corresponds to column in Figure 3.6. The component planes are visualized by taking the values of each component from all codebook vectors and plotting them with a color code over the SOM grid. This representation provides valuable information on the distribution of the component values. By visualizing several component planes simultaneously it is possible to infer relationships between input variables. Figure 3.7 depicts a stacked SOM showing the distribution of component values at each *c-plane*.

- **Cluster structure**. Clusters are formed by groups of codebook vectors which are close to each other compared with their distance to other vectors. The clustering structure of the input space can be visualized over the SOM grid by displaying the distances between these reference vectors. Several methods have been proposed to display the clustering structure; the most common is the unified distance matrix (*U-matrix*), i.e., the matrix of distances between each codebook vector and its neighbors. By showing this matrix as a gray-level picture or using a 3D representation, the relative distances between adjacent units on the map can be seen. Usually a 2D gray-level encoding representation is used in which dark colors represent high distance values while lighter colors correspond to low ones. By using this encoding, light gray areas on the map represent clusters while dark gray regions

ISBN:978-84-691-1008-9/DL: T.2230-2007

correspond to cluster boundaries. The visualization of these clusters could be enhanced by labeling the map with auxiliary data. A sample representation of a U-matrix is depicted at the top of Figure 3.7.

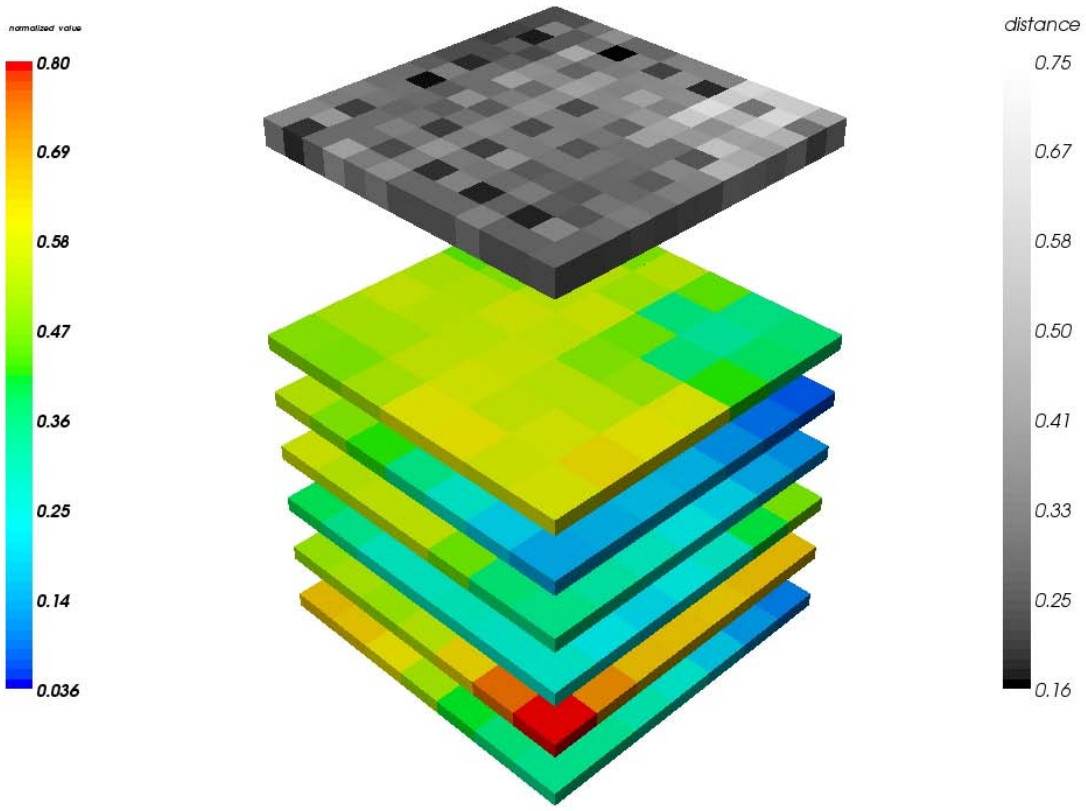


Figure 3.7. Layered organization of a Self-organizing map. A set of stacked component planes is plotted using a color scale proportional to the normalized value. The U-matrix is represented on top of these layers using a gray color code in which high distances between units (borders) appear in dark gray while low distances (clusters) appear in light gray colors

- **Clustering of the SOM**. Since it is difficult to detect clusters by visual inspection of the U-matrix, SOM's reference vectors can in turn be clustered to detect coherent sets of units with similar structural characteristics. Simpler clustering algorithms such as the K-means procedure are used to cluster SOM vectors. The clustering parameters are optimized using some clustering quality criteria such as the minimization of the Davies-Bouldin index. This index is a function of the ratio between the sum of cluster compactness and inter-cluster separations (Davies and Bouldin, 1979) and permits the selection of the optimal number of clusters to obtain a good cluster of clusters identification and partitioning.

- **SOM in the input space**. The U-matrix doesn't give any indication about the overall shape of the native data set because the visualization is tied to the map grid. To visualize the SOM codebook vectors in the input space a projection method is needed. The most commonly used transformation is

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

the Sammon's Mapping procedure (Sammon, 1969). This algorithm performs a non-linear projection from a high dimensional input space to a lower dimensional space, typically 2D. The algorithm tries to preserve the distances between data points, emphasizing local distances. When applied to SOM reference vectors and to enhance the proper look of the resulting projection, neighboring map units are inter-connected with lines to show their topological relations. Since the SOM tends to approximate the probability density of the input data, the Sammon's mapping of the SOM codebooks can be used as a rough approximation to the shape of the input data. This procedure could be applied directly to the input data set, but as it is computationally very expensive, it is too slow for large data sets. The SOM projection quantizes the input dataset into a small number of reference vectors. This lightens the burden of Sammon's computation to acceptable levels. Other projection techniques based on Principal Component Analysis (PCA) or in Curvilinear Component Analysis (CCA) can also be applied.

- **Input vectors on the SOM**. Data vectors can be projected over the SOM grid by using the BMUs. By taking a set of data vectors and projecting them onto the map, a histogram is obtained which shows how many input vectors belong to the clusters defined by each map unit. The histogram is computed by determining the BMU for each data set vector and increasing by one a "hit counter" on this unit. Different data sets can be contrasted by comparing their corresponding histograms.

- **Response Surfaces**. The responses of all units are used to obtain the global response of the whole map to certain input data. A response surface is a visualization technique which shows the goodness of each map unit in representing the data. The simplest approach is to use the quantization error $qe_i = \|x - m_i\|$ as a goodness indicator by using the following equation:

$$g(qe_i) = \frac{1}{1 + qe_i^2} \tag{3.20}$$

From the mathematical point of view, this goodness function has several beneficial properties such as $g(0) = 1$, $g(qe \to \infty) = 0$, and $g'(qe)$ is monotonically decreasing. Using this last property one can see that good matches produce sharp responses on the map while poor matches produce an even response over the whole map. To dismiss the scale effects in the evaluation of g, the quantization error must be normalized using the average quantization error. If a data independent measure is required, the average distance of each codebook vector to its neighbors can be used. The response surfaces obtained by using this approach are comparable even between maps trained using different datasets.

- **Transient data**. Dynamic processes can also be analyzed and visualized by the SOM. The procedure to study the behavior of a transient process uses the locations of the BMUs corresponding to measurements of the *operation point conditions* over the process dynamics. The location of these points in the SOM grid determines the current process state. The time evolution of the

ISBN:978-84-691-1008-9/DL: T.2230-2007

process can be characterized by a sequence of operation points displayed on the SOM as a *process trajectory*. Labeling of the SOM permits the characterization of these trajectories and the identification of interesting areas from the process operation point of view.

3.3 SOM-based EDA for Biodegradation of chemicals

Chemicals can be characterized by a wide range of descriptors such as molecular weight, lipophilicity, topological, electronic or quantum features. *Chemical space* (Lipinski and Hopkins, 2004) is a term used instead of multidimensional descriptor space and applies to a region defined by a particular choice of descriptors and their limits of applicability. Measured in terms of their physicochemical properties and molecular descriptors, families of similar compounds appear to cluster together in some regions of the chemical space. An important question is whether these clusters are evenly and sparsely distributed and therefore hard to find, or if most of the chemical space is empty (with no compounds of interest), with clusters of interesting chemicals scattered far apart. The concept of *chemography* has been proposed to navigate the diversity of chemical space (Oprea and Gottfries, 2001).

Chemicals of commercial interest may also be of environmental concern due to their persistence in the environment, bioaccumulative propensity and/or toxicity. Degradation mechanisms are of diverse nature, biotic and abiotic, and hence can be grouped as physical, chemical or biologically driven. Metabolism by microorganisms is one of the most important processes determining the fate of chemicals in the environment. However, laboratory analyses in which the chemicals of concern are used as the sole source of carbon and dietary energy are difficult to extrapolate to real field situations. When released to the environment, these compounds are usually degraded by multispecies microbial communities whose population dynamics and metabolic status are very difficult to model. In addition, biodegradation rates have a highly non-linear dependency with the rates of a series of biological processes such as uptake mechanism and transport within the cell, binding to an active enzymatic site or enzymatic transformations. There are numerous private and public research efforts aimed at developing biodegradation models with mechanistic approaches based on metabolic pathways, or in activity-structure relationships. The new European REACH directive that regulates the registration and authorization of chemicals since June 2007 will shed new light into the processes involved in the evaluation of chemicals.

Prior to any attempt to develop biodegradation models in terms of physicochemical properties and molecular descriptors, it is necessary to understand the relationship between all affected features. It would be very useful to have a complete *cartography* of the chemical space of organic compounds in terms of their biodegradation rates (or half-lives) to identify the regions of the chemical space where useful chemicals are located. This would also provide a better understanding of the mechanisms governing the degradation process and their relationships with chemical structure.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Such cartography could also guide the design of new chemicals towards portions of the chemical space to fulfill both high biodegradation rates and commercial interest.

To this end, the exploratory data analysis methods proposed in this chapter have been applied to two different sets of organic chemical compounds whose persistence in the environment are known. This section aims to illustrate the use of SOM-based visualization techniques for EDA and to assess its suitability as a tool for the first tier of the current model.

3.3.1 SOM-based EDA for the Persistence of organic pollutants

Chemicals with high persistence in air, water, soil or sediments have a high potential for bioaccumulation in these media, i.e., for uptake by living organisms. Thus, they are candidates to severe regulations. The data set selected for the current SOM-based EDA comprises 101 organic compounds from which experimental degradation rates in different media are known (Mackay et al. 1992). For this set, a group of 5 physicochemical properties (molecular weight, melting point, vapor pressure, solubility and the octanol-water partition coefficient, K_{ow}) and 30 molecular descriptors were used to characterize the chemical space.

The most important step in exploratory data analysis consists in the proper selection of the mechanisms to scale and transform the data to obtain the most informative projection over the target variable space. Machine learning algorithms require the proper scaling of data to avoid the undesired effects caused by differences in data magnitudes. Commonly used techniques include both, normalization of variables and non-linear transformations aimed to emphasize some variables. Figure 3.8 compares the effects that produce diverse normalization methods and transformations on the shape of the data distribution function. For simplicity, in this first step only data related to the current target properties (degradation rates) are presented and discussed.

A quick examination of the leftmost column in Figure 3.8a, which corresponds to degradation in air, reveals that this variable shows a skewed distribution with a higher population in the low degradation rate range. The application of different normalization techniques to Figure 3.8a maintains the biased distribution except in the case of a transform procedure based on the normalization of the data histogram. This technique produces smoother and regularly shaped distribution functions. The effect of these transformations is worse for degradation rates in water (see Figure 3.8c). The resulting multimodal distributions, showing three main modes, cannot be smoothed even with the histogram normalization technique. A similar situation is observed in Figures 3.8e and 3.8g for soil and sediments, respectively. However, in both situations, normalization by using data histogram equalization results in smooth unimodal distributions. The main drawback in the application of the histogram normalization technique is that the use of different bin sizes to equalize the histogram tends to emphasize parts of the data range and to moderate some others. The result is a lost in resolution for the data grouped in bigger bins with respect to data grouped in the smaller ones. If the exploration process is qualitative, this

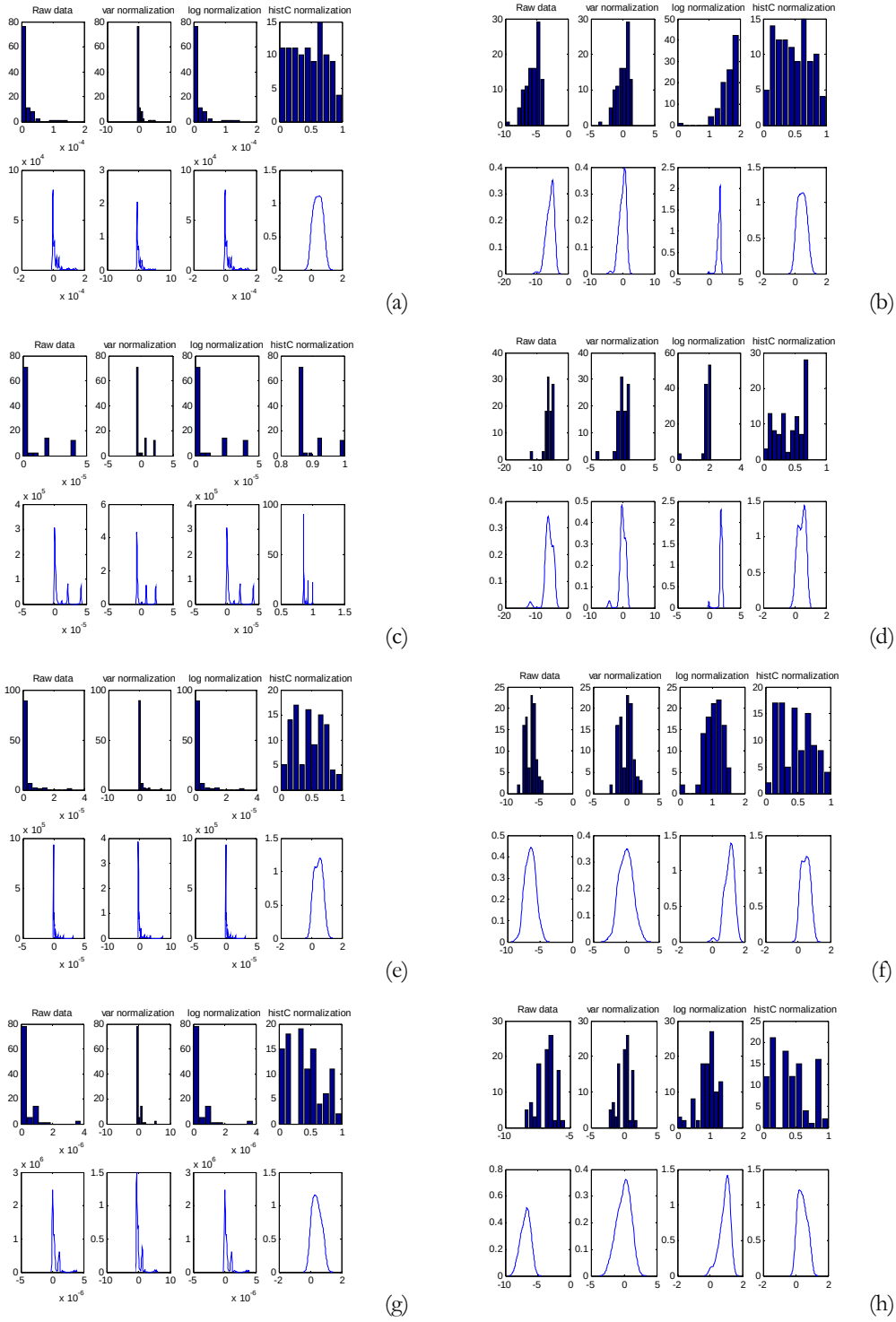


Figure 3.8. Effect of the normalization procedure on the distribution of the biodegradation rates over the whole data set. Biodegradation rate in air (a); logarithm of the biodegradation rate in air (b); biodegradation rate in water (c); logarithm of the biodegradation rate in water (d); biodegradation rate in soil (e); logarithm of the biodegradation rate in soil (f); biodegradation rate in sediments (g); logarithm of the biodegradation rate in sediments (h)

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

approach will be appropriate; if the focus is in quantitative exploration, these magnification effects could be undesirable and should be avoided.

An alternative path to avoid these effects consists in the transformation of the data using some non-linear scaling such as the logarithm. Figures 3.8b-g present the histograms and distribution functions for the logarithm of the degradation rates in the four media mentioned above. It can be observed that after the logarithmic transformation, the shapes of the data distributions are more even and their corresponding ranges are emphasized. Transformed data are more suitable for the subsequent use of machine learning algorithms to develop data-driven models.

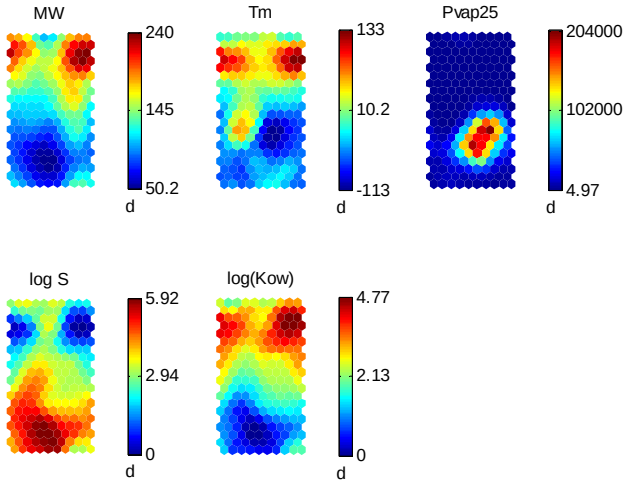
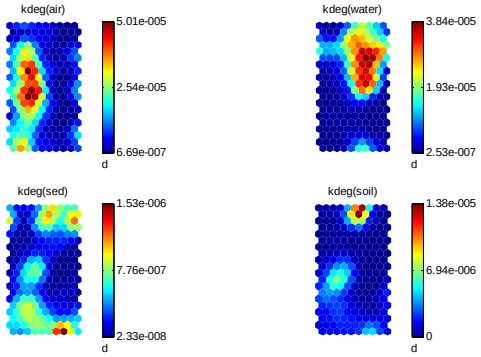


Figure 3.9. Component planes corresponding to the physicochemical properties used in the exploration of the chemical space for persistence rates

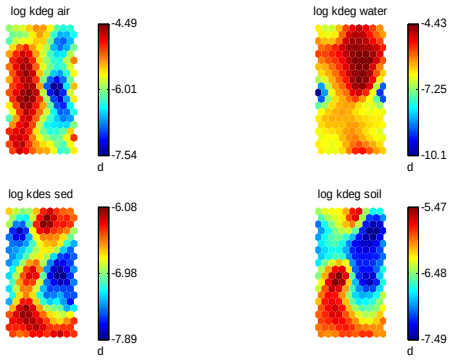
A 200 units SOM was adapted to visualize the relationships among the variables of the chemical space using the complete data set and including in the analysis both the values of the degradation rates and their logarithms. Figure 3.9 depicts the structure of the component planes corresponding to the physicochemical descriptors. The color coding is based on the native (not normalized) values of these variables. It can be observed that molecular weight (MW) and melting point (Tm) are closely related. The upper side of their respective component planes accounts for those chemicals with high MW which also have an elevated melt point temperature. The octanol-water partition coefficient represents the ratio of the concentration of a chemical in octanol and in water at equilibrium and at a specified temperature. Octanol is a surrogate for natural organic matter; this parameter, which measures hydrophobicity, is used in many environmental studies to determine the fate of chemicals in the environment. The c-planes in Figure 3.9 show that the logarithm of the octanol-water partition coefficient relates to MW and Tm. This partition coefficient is also known to be correlated with water solubility; in fact there exist many correlations in the literature to estimate this partition coefficient from aqueous solubility. Component planes in Figure 3.9 clearly point to this relationship. It can be observed that the SOM space occupied by chemicals with high water solubility corresponds to chemicals with low values of the K_{ow} constant. Additional relationships relate vapor

pressure with melt point. The high vapor pressure spot shown in Figure 3.9 is related to chemicals with low melt point.

Similar techniques can be used to examine the distribution of the degradation rates over the SOM. Figure 3.10 shows the component planes corresponding to these degradation rate data. The direct comparison of raw degradation constant values and their corresponding logarithmic transform indicates that the use of logarithms emphasizes the visualization of its distribution. Figure 3.10 indicates the existence of a similar pattern in the distribution of the degradation rates in soil and sediments. In fact, some biodegradation models consider sediments as a mixture of soil and water.



(a)



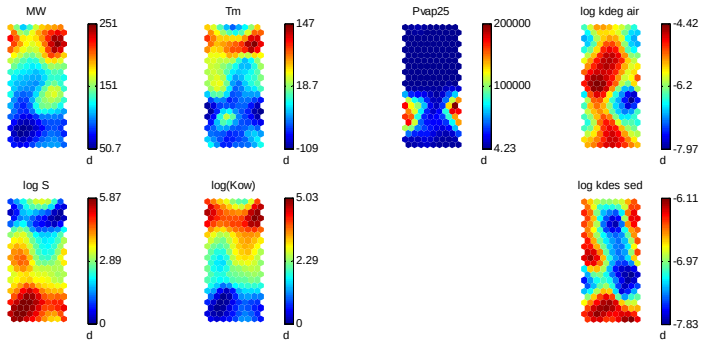
(b)

Figure 3.10. Component planes of degradation rates in air, water, soil and sediments. Native data values (a); logarithm of the degradation constants (b)

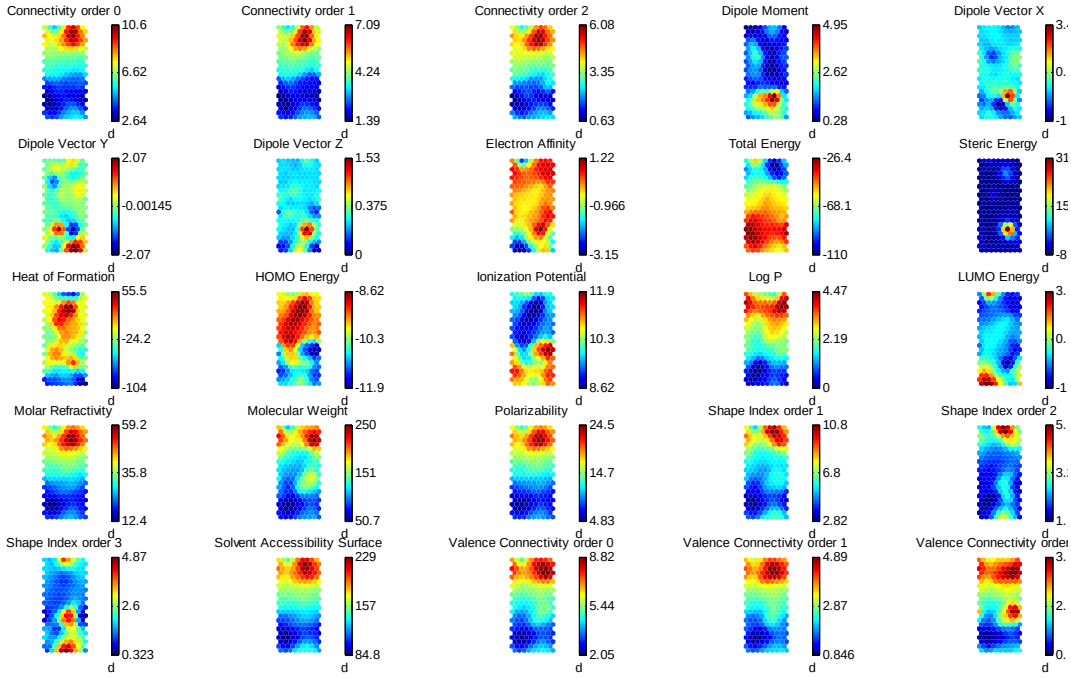
As has been discussed so far, the use of the logarithmic transform of degradation rates is preferred for machine learning purposes. A new SOM using only the log transformed degradation rates has been used to classify the input space. The resulting component planes are presented in Figure 3.11. Differences observed in the distribution of the physicochemical properties in the SOM planes with respect to those shown in Figure 3.9 are due to the effect of the removal of the raw degradation values and to data shifts in the toroidal surface of map. This alteration is more evident for the vapor pressure in which the high value spot observed in the center of the c-plane in Figure 3.19 is now divided in two areas by the effects of

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

periodic boundaries in the toroidal map topology. Figure 3.11c includes the component planes corresponding to the molecular descriptors computed for the current set of chemicals.



(a) (b)



(c)

Figure 3.11. Component planes corresponding to the complete chemical space for degradation rate constants. Physicochemical properties (a); logarithms of the degradation rates(b); molecular descriptors (c)

From the EDA point of view, component planes are useful to detect potentially redundant information. It can be observed that the three connectivity indices (order 0 to order 2) contribute with the same information to the clustering induced by the SOM, i.e., the component planes have the same distribution, despite their different

magnitudes. This indicates that the information added by these three variables is identical and two of them may be excluded from the analysis without any loss of significant information. A similar situation is observed for other descriptors, e.g., solvent accessibility surface and valence connectivity of order 0 and 1, molar refractivity, and polarizability. From the EDA perspective, the inspection of component planes provides useful insight on the relationships of all the variables involved in the description of the chemical space.

The adaptation performed during SOM training aims at reproducing the distances and point densities of the original data space. Therefore, if map dimensions are insufficient to accommodate all the input data, the SOM lattice appears folded and results in high topographic errors, i.e., the map is unable to preserve the distances of the original space. The 2D visualization of the SOM lattice in terms of the distances in the input space requires the use a projection method. Figure 3.12 shows two common techniques used to project the SOM. A color code is included to correlate the position of units in the SOM plane with their positions in each projection. These plots give visual clues about the regions of the map in which it is most difficult to preserve the topological relationships of the input space. The projection of the map lattice and the detection whether folding occurs or not is an indispensable step to confirm the accuracy of the maps obtained and the validity of the inferred evidences about the structure of the input space. Additional techniques such as feature selection are required to obtain regular maps, i.e., maps that are evenly distributed over the input space.

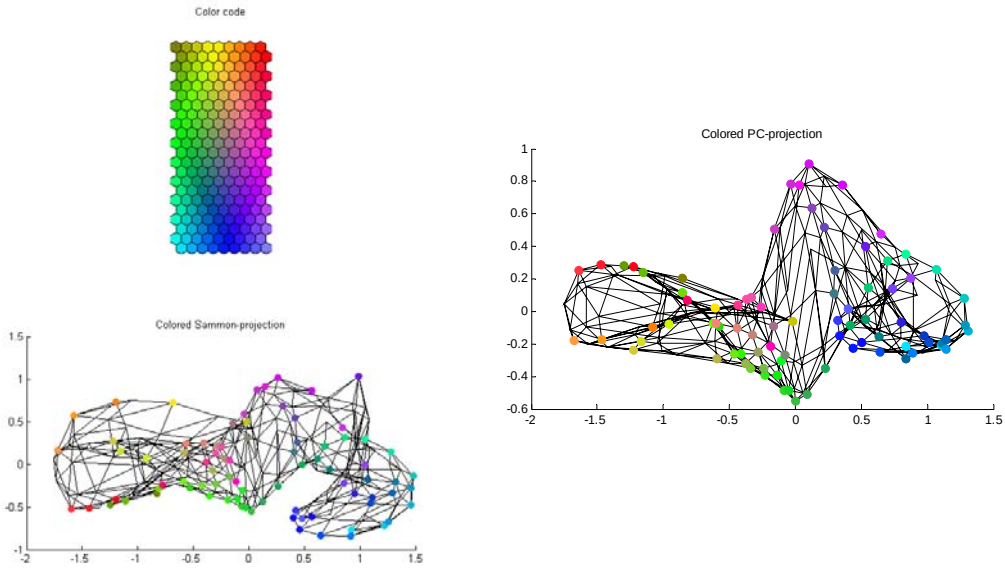


Figure 3.12. Comparison of two projection techniques based on Principal component Analysis and in the Sammon's mapping algorithm. A color coding scheme computed from unit distances on the SOM lattice is used to correlate the position of units in the SOM space and in each projection

Component planes only provide a partial view (projection) of the clustering induced by the map. Structural patterns in the chemical space can be detected using

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

additional techniques known as distance matrices. Distance matrices shown in Figure 3.13 represent a bi-dimensional projection of the chemical space corresponding to the current set of 101 chemicals according to their physicochemical properties, molecular descriptors and degradation rates. High distances represent cluster boundaries while small distances correspond to compact clusters of similar compounds. From the chemical space point of view, each cluster can be considered as a coherent portion of the chemical space where compounds have similar characteristics in terms of the variables used to generate the map. The internal structure of distance matrices in Figure 3.13 is quite complex and it is not easy to detect cluster boundaries. This is a common situation in most real-world clustering problems where the separation between homogeneous classes is not evident. These matrices can be used to detect areas of interest in the chemical space and to establish relationships with input properties trough the inspection of component planes.

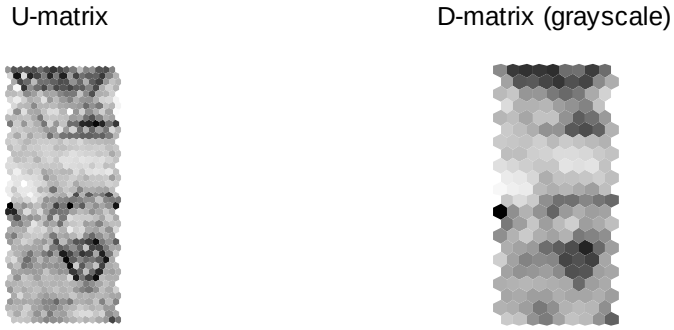


Figure 3.13. Visualization of the clustering structure of the SOM from the U-matrix (left) and the mean distance matrix (right) points of view. High distances are represented by dark gray colors, while lighter colors correspond to smaller distances

The internal structure of the SOM can be emphasized by clustering the SOM prototype vectors. Diverse clustering techniques can be used, being the K-means algorithm a usual choice. The optimal number of clusters, i.e., the K value, is found by optimizing cluster quality metrics such as the Davies-Bouldin index. The K-means partition shown in Figure 3.14 corresponds to the optimal clustering of the chemical space for the current data set. It can be observed that after clustering the number of classes and their boundaries are more evident than in previous distance matrices. The chemical space for this set is formed by 13 coherent chemical families according to the pre-selected set of properties. The complete and detailed exploration of the characteristics of each of these families provides a complete description of the chemical space in terms of the selected physicochemical and molecular information. Chemical families can be used to develop local models fitted to selected regions of the chemical space. It is important to note that the partitioning obtained after clustering the input data space differs from the clusters obtained by the classification of SOM reference vectors. The former approach produces fine-grained partitions where each SOM unit acts as cluster representatives, while clustering these cluster centers produces new coarser and more general class descriptions. The combined inspection of the clustering structure at the two

ISBN:978-84-691-1008-9/DL: T.2230-2007

partitioning levels highlights useful information about the presence of structurally similar groups of chemicals (analogous compounds).

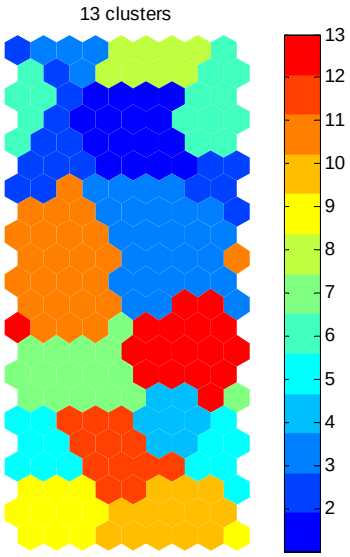


Figure 3.14. Classification of the chemical space into chemical families according to the physicochemical properties, molecular descriptors and degradation constants

3.3.2 SOM-based EDA of MITI-1 biodegradation rates in water

The data set used in this section comprises a heterogeneous group of 672 organic compounds whose biodegradation rates in water have been experimentally determined using the MITI-1 test. This analytical procedure was developed in Japan and constitutes a screening procedure for the determination of "ready" biodegradability in an aerobic aqueous medium. This method is one of the six standardized "ready" biodegradability tests described by EU and OECD regulations. For the MITI-1 test; 100 mg/l of test substance are inoculated and incubated with 30 mg/l of sludge. Biological oxygen demand (BOD) is measured continuously during a 28-day test period. The pass level for "ready" biodegradability is reached if BOD amounts to ≥60% of the theoretical oxygen demand (ThOD).

The MITI-1 experimental data have been complemented with calculated molecular descriptors for the 672 chemicals to include chemical structural information into the data set. Several types of descriptors have been computed, including constitutional (topological and valence information), electrostatic, geometric, and electro-topological. In addition, 3-dimensional (3D) quantum chemical descriptors were also generated with the semi-empirical PM3 method and included in the initial set of descriptors. Quantum chemical calculations are an attractive source of molecular information since they express the electronic and geometric properties of molecules and their interactions. Quantum chemical descriptors are able of characterizing the reactivity, shape and binding properties of a complete molecule.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 3.1. Descriptive statistics of the biodegradation data set, expressed as percentages

		Statistic	Std. Error
Mean		29.63	1.434
95% Confidence Interval for Mean	Lower Bound	26.82	
	Upper Bound	32.45	
5% Trimmed Mean		27.51	
Median		3.00	
Variance		1382.05	
Std. Deviation		37.18	
Minimum		0	
Maximum		100	
Range		100	
Interquartile Range		68	
Skewness		0.703	0.094
Kurtosis		-1.270	0.188

The analysis of the basic set of descriptive statistics given in Table 3.1 for biodegradation in water shows that the mean biodegradation is located around 30% which corresponds to non-ready biodegradable chemicals. The high values for the interquartile range, variance and standard deviation indicate that data are greatly dispersed. In addition, the skewness and kurtosis point to a non-symmetric skewed distribution far from a normal distribution.

To assess the SOM capabilities for exploratory analysis of the chemical space in this application domain, a reduced subset of variables related to the conditions of the MITI-1 experiment were first analyzed. This data set is formed by the total number of atoms in the molecule, the experimental biodegradation, the duration of the MITI experiments (2, 3 or 4 weeks) and the logarithm of the biodegradation half-life. Half-lives were computed from the experimental biodegradation rates assuming a first order kinetic mechanism. A SOM containing 130 units was adapted using the complete data set after data normalization in the range [0,1].

The examination of U-matrix in Figure 3.15 indicates the presence of five clusters which correspond to five homogeneous chemical families in terms of the MITI-1 test procedure conditions. The comparison of clusters present in the U-matrix by means of the BOD component plane shows that two of them, located in the upper left of the U-matrix, correspond to readily biodegradable chemicals. The main difference among these two families resides in the duration of the MITI experiment which is either reduced (less than 28 days) or complete. A similar situation is observed for the remaining clusters which correspond to non biodegradable chemicals. The partition observed is also related to the duration of the MITI experiment. Furthermore, an inverse dependency between BOD and the logarithm of half life is observed. It can also be inferred from the c-planes that the number of atoms in the molecule is not correlated with any other variables and doesn't

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

contribute significantly to the internal clustering structure of this subset of variables. As a conclusion, the inspection of the c-planes in Figure 3.15 indicates that MITI-1 biodegradation data are partitioned into two main groups for which the discriminating BOD value oscillates between 40% and 50%. Two clusters containing highly biodegradable chemicals are formed in the first group as a direct effect of the duration of the MITI-1 experiment. The second group contains persistent chemicals and is formed by three clusters generated by the combined effect of the duration of the experiment and half-life values. The effect of half-life is less important for highly biodegradable chemicals and doesn't contribute significantly to the clustering structure.

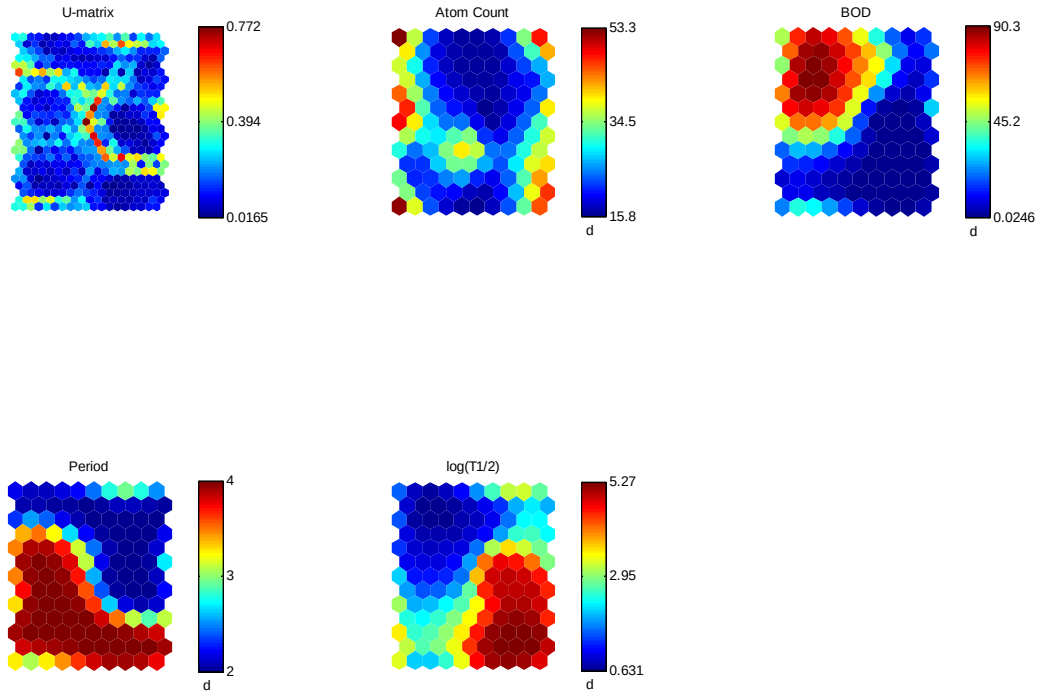


Figure 3.15. U-matrix and component planes corresponding to a preliminary analysis of biodegradation related parameters

The relative importance of each element or component of the reference vectors may be represented using different strategies. The simpler one is illustrated in Figure 3.16 and corresponds to a bar diagram. The height of bars is proportional to the values of components in the codebook vectors. It can be seen that this representation is consistent with the c-planes plotted in Figure 3.15 and provides an integrated picture of the relative distribution of all components of the codebook vectors.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

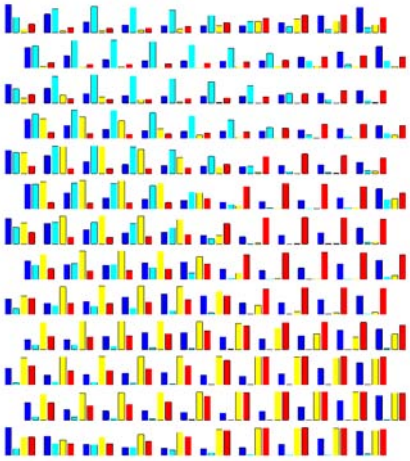


Figure 3.16. Bar representation of SOM codebooks. Each bar diagram represents the distribution of the values of each component of the codebook vector. For this example the components are atom count, BOD, period and half life

Similar techniques can be used to represent this information when the number of input variables is high. For high dimensional data, pie plots are used instead of bar diagrams. Also it is common to represent the "population" of each unit by making the size of the pie proportional to the number of elements associated to that unit. The right-hand side of Figure 3.17 shows that not all units are equally informative, i.e., few units account for most of the chemicals in the dataset. Units which don't represent any chemical are referred as "interpolative units". These nodes act as cluster boundaries and help to partition the SOM lattice into separate clusters. Interpolative units should be excluded when building local models as don't represent any real chemical. However, in some situations, the prototype vectors of interpolative units can incorporate unseen portions of the chemical space into the application domain.

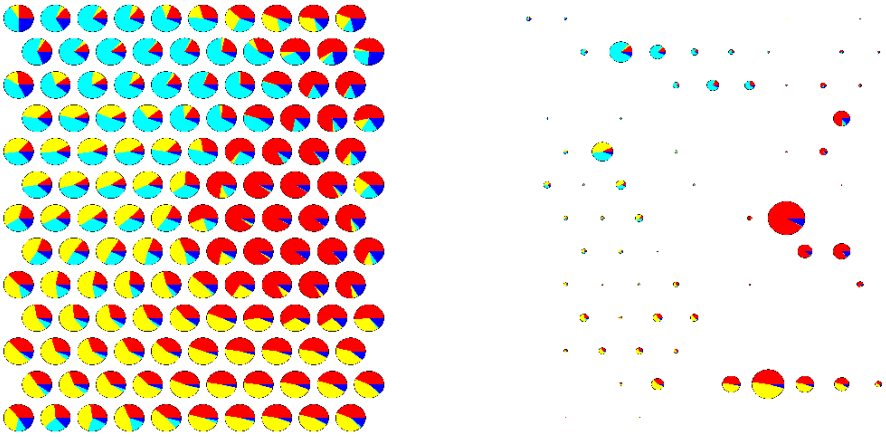


Figure 3.17. Pie plane representation of the SOM. (left) All pies are of the same size. (right) The size of pies is proportional to the number of *hits* of each unit. The variables represented and colors correspond to those in Figure 3.16

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The use of scatter-plot matrices is an alternative approach to the visual exploration of the relationships between the trained SOM and the initial data set. Figure 3.18 depicts the scatter plot matrix corresponding to the BOD dataset. This graph matrix constitutes a complex representation of data and contains large amounts of information. The upper right triangle accounts for the data corresponding to the reference vector after training. The lower left includes all cases contained in the data set used for training the SOM. The first row and the leftmost column facilitate the comparison of SOM prototype histograms with the histograms obtained using all data. The first element of the scatter-plot matrix diagonal defines a color coding scheme based on the relative position of each unit in the SOM grid. Subsequent elements of the diagonal represent each of the component planes. The remaining elements in the upper triangle represent pair-wise scatter-plots of each variable using the codebook vectors. Data points are linked together using the topological relations of the SOM lattice. Finally, the lower part of the matrix shows the pair-wise scatter-plots corresponding to the input data set.

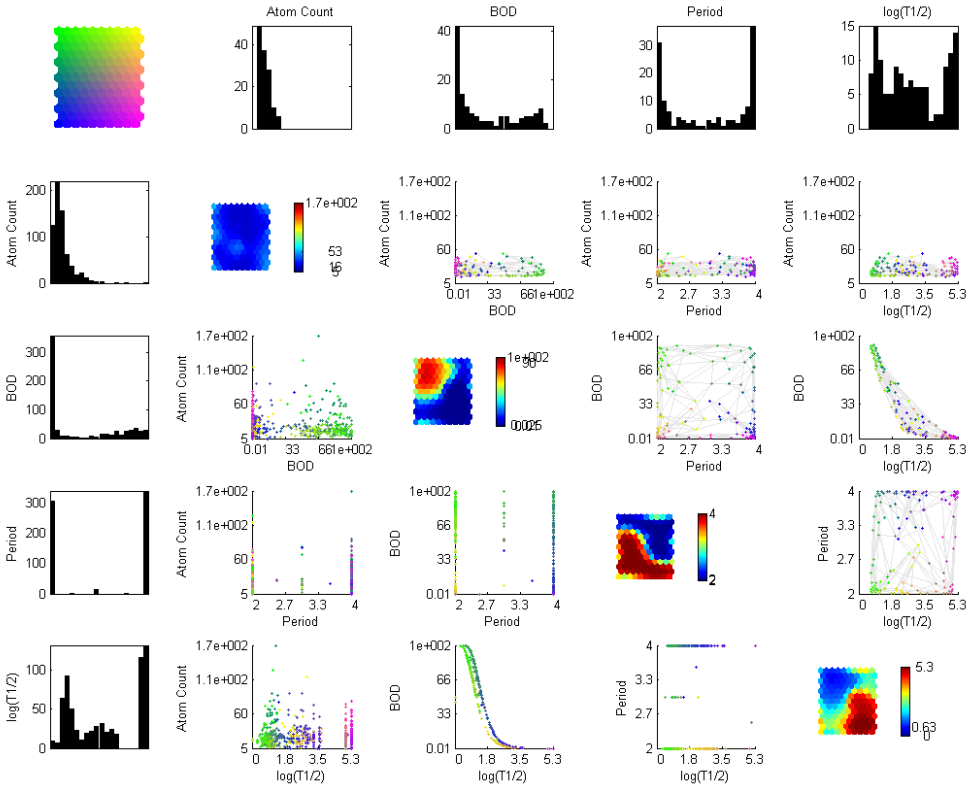


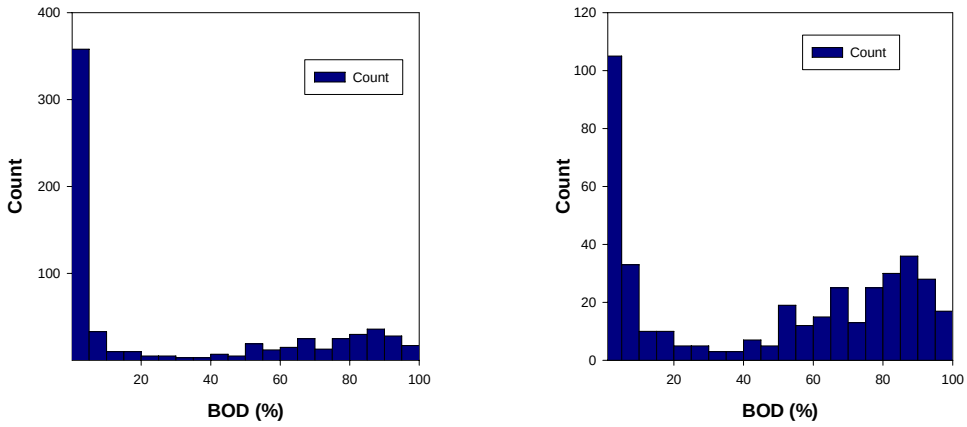
Figure 3.18. Scatter plot matrix representation of the trained SOM. The upper triangle contains data relative to the SOM codebook vectors, while the lower triangle represents the complete data set

The comparison of all these plots provides precise information about the relationships among variables and how these relationships are reproduced in the SOM space. For instance, examination of Figure 3.18 indicates that BOD and the logarithm of half life are clearly dependent. This is a consequence of the kinetic

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

approach used to compute the half-life from biodegradation. Also a color coding scheme is used in scatter plots to link these representations with the SOM ordering.

The histogram corresponding to biodegradation shows that the distribution of the BOD values is very skewed, resulting in a multimodal distribution in which most of the values have 0% BOD. A more detailed histogram of the values of this variable is depicted in Figures 3.19a and 3.19b, showing that the broad distribution of biodegradation rates is not homogeneous. Data distribution remains severely skewed even after the exclusion of chemicals with zero biodegradability. This suggests the need of considering biodegradation sub-ranges to obtain more equalized histograms which would facilitate the development of more accurate biodegradation models.



(a) (b)

Figure 3.19. Histogram of biodegradation percentage for the MITI-1 dataset. Range [0-100] (a); range [1-100] excluding chemicals with zero biodegradation rates (b)

The application of different normalization techniques helps to emphasize some parts of the data set, as in the previous case. Non-linear normalization techniques, such as logarithmic transformation and histogram normalization, show the presence of a strongly bimodal probability function with two main modes corresponding either to non-biodegradable or readily biodegradable chemicals. The exclusion of chemicals with 0% BOD doesn't eliminate the presence of the bimodal distribution. However, if the equalization of the histogram is used as normalization method, quasi-normal distributions are obtained. This transformation is useful to detect the presence of different structures in the data set. Nevertheless, if it is used to develop quantitative models may reduce their accuracy due to the non-linear expansions and contractions of the input space that it implies. The analysis of the low range of biodegradation [1-60%] reveals a similar structure to the one found for the whole dataset. Finally, if highly biodegradable chemicals [>60%] are considered separately, the probability distribution exhibits a smooth behavior which approaches a normal distribution. The main conclusion that can be drawn from these results is that it will be very difficult to build a single predictive model for the complete range of biodegradation rates. The the distributions obtained after removing chemicals with zero BOD indicate that this approach could be useful to develop classifiers for ready/non-ready

ISBN:978-84-691-1008-9/DL: T.2230-2007

biodegradable compounds. Comparison of distributions for low and high biodegradation ranges indicates that the development of models for low BOD will be more difficult than for highly biodegradable chemicals.

To confirm these relationships a subset of only 25 molecular descriptors (excluding biodegradation) was analyzed using the SOM-based EDA procedure. After the adaptation of the map, the resulting reference vectors were classified using the K-means algorithm combined with the optimization of the Davis-Bouldin index. Figure 3.20 shows the resulting chemical families labeled using the biodegradation ranges of each class.

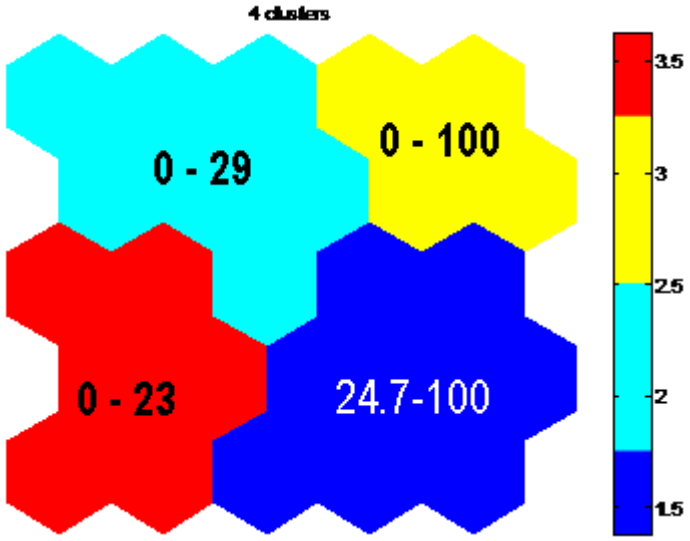


Figure 3.20. SOM trained with 25 molecular descriptors. The clustering shows 4 families of chemicals. Data used to build the map were normalized to zero mean and unity variance (standardized)

A close examination of the distributions given in Figure 3.19 suggests that a BOD value of 40% constitutes an appropriate cut-point to separate biodegradation into two well distributed subsets. This discriminating value was already detected from the examination of the borders of U-matrix in Figure 3.15. A subset of the non-readily biodegradable chemicals is grouped in two families containing compounds with BOD in the approximate range of 0 to 30%. The rest of compounds are grouped in two families containing both non-biodegradable and readily biodegradable chemicals.

In addition to these transformations, changes in parameters of the SOM algorithm could also influence the EDA process. The effects of changing the size of the SOM have been analyzed by using a selected subset of 11 molecular descriptors and training two maps of different sizes. The component planes and the optimal Davies-Bouldin clustering of the resulting SOM are shown in Figures 3.21 and 3.22, respectively. Comparison of c-planes shows that the distributions of variables in the map are consistent and independent of the size of the map. An increase in map size results in a higher resolution representation of the input space. The clustering

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

process of the reference vectors is also coherent with these observations. This process takes advantage of the resolution increase and produces improved cluster partitions. The portion of the chemical space corresponding to chemicals with high biodegradation rate, located in the lower part of the map, appears as a single cluster with a resolution of 25 map units.

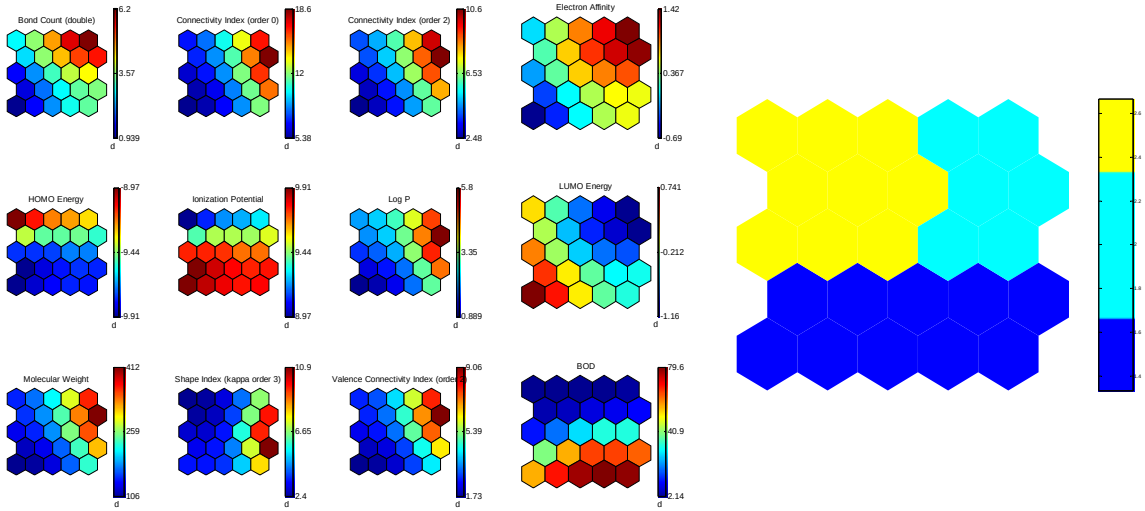


Figure 3.21. Component planes and Davies-Bouldin clustering of the SOM when using 11 molecular descriptors and [5x5] units in the map

The improved resolution attained after the increase in map size yields a better discrimination in the high biodegradation range. The single cluster observed in the lower part of Figure 3.21 is now divided in three smaller clusters.

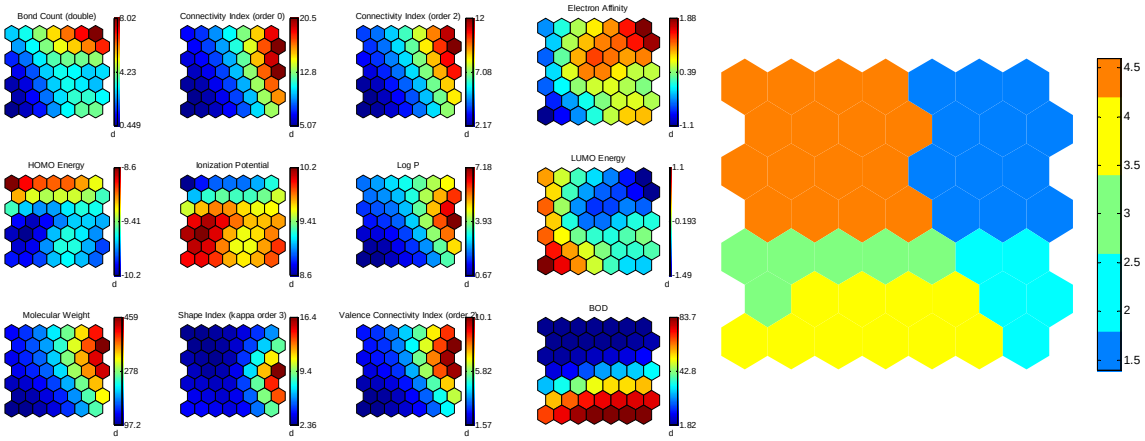


Figure 3.22. Component planes and Davies-Bouldin clustering of the SOM when using 11 molecular descriptors and [7x7] units in the map

An important feature that any EDA technique must include is interpretability. The use of hit maps combined with c-planes helps to understand and to explain the clustering generated by the SOM. To illustrate its application, a reduced subset of 6

ISBN:978-84-691-1008-9/DL: T.2230-2007

molecular descriptors, 3 physicochemical properties, and the experimental MITI-1 biodegradation rate were classified using a toroidal SOM formed by 330 units. Figure 3.23 depicts the results obtained after this analysis. An optimized threshold of 40% BOD (far from the standard value of 60%) is needed to discriminate between non-biodegradable and ready-biodegradable chemicals. It can be observed that chemicals with BOD ≤ 40% are mainly located in the central part of the map, while the ready-biodegradable are located in the lower part of the map. The size of the black inserts (hit counters) within the SOM units is proportional to the number of hits received by that unit.

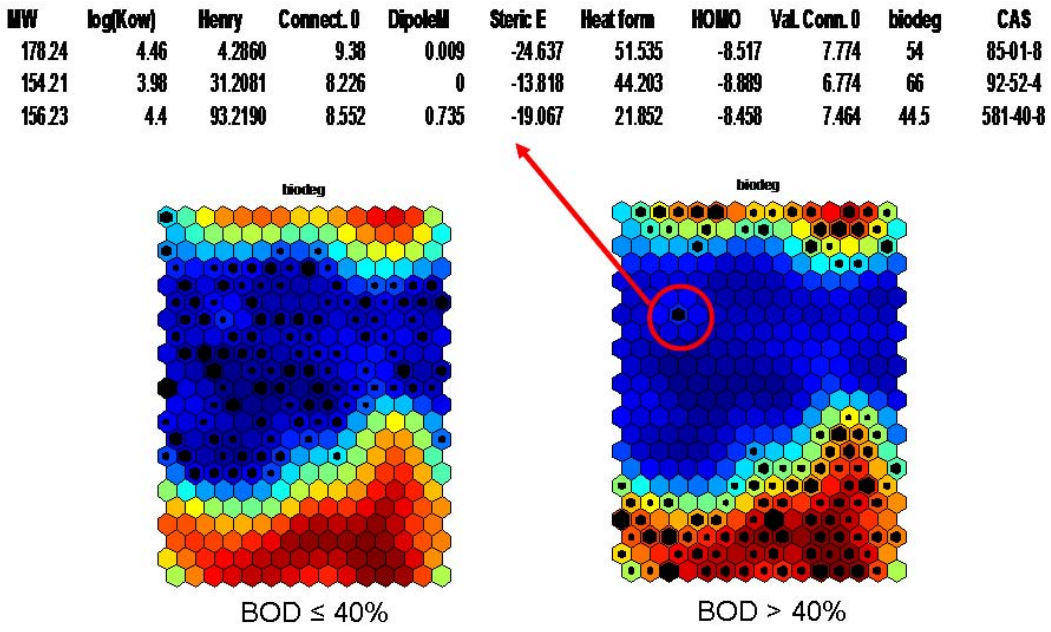


Figure 3.23. Exploration of the chemical space for biodegradation using MW, log(K_{ow})and Henry's law constant, together with 6 molecular descriptors. The threshold value used to discriminate between non-biodegradable and ready-biodegradable is 40%. The three chemicals circled in the map and described in the upper part are misclassified

The map in Figure 3.23 separates well these two biodegradation ranges below and above 40% BOD but with three misclassified chemicals. The examination of their BOD reveals that two of them have values of 44.5% and 54% which correspond to a low reliability zone for MITI data. This interval is in agreement with the values reported by Sabljic and Peijnenburg (2001) of 45 and 55% respectively. The third chemical, with a BOD of 66%, is also misclassified in the same SOM units, highlighting the difficulty of classifying biodegradation ranges using only this kind of input information.

EDA techniques are also useful to detect the presence of data relationships which a priori are not evident. This is illustrated by analyzing the effect of the inclusion of ecotoxicological information in the exploratory mapping of the chemical space of MITI-1 biodegradation. To this end, the experimental LC_{50} (lethal concentration for

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

half of the population in a 96 h test) toxicity index for the Orange-red killifish, for a subset of 330 MITI chemicals was added to the set of nine input variables used in Figure 3.23. A new toroidal SOM with a resolution of 432 units was used to classify these data. Figure 3.24a shows that there is not a clear relationship between acute toxicity (LC_{50}) and biodegradation. Nevertheless, the inclusion of LC_{50} induces better clustering in the chemical space of biodegradation as pointed by the U-matrix and response surfaces. The red colored spots in Figure 3.24b are proportional to the number of hits received by each unit. The SOM including ecotoxicological information discriminates well between these three biodegradation ranges. The implementation of accurate classification models of biodegradation would give best results when ecotox-related information is used to discriminate between the three biodegradation ranges identified.

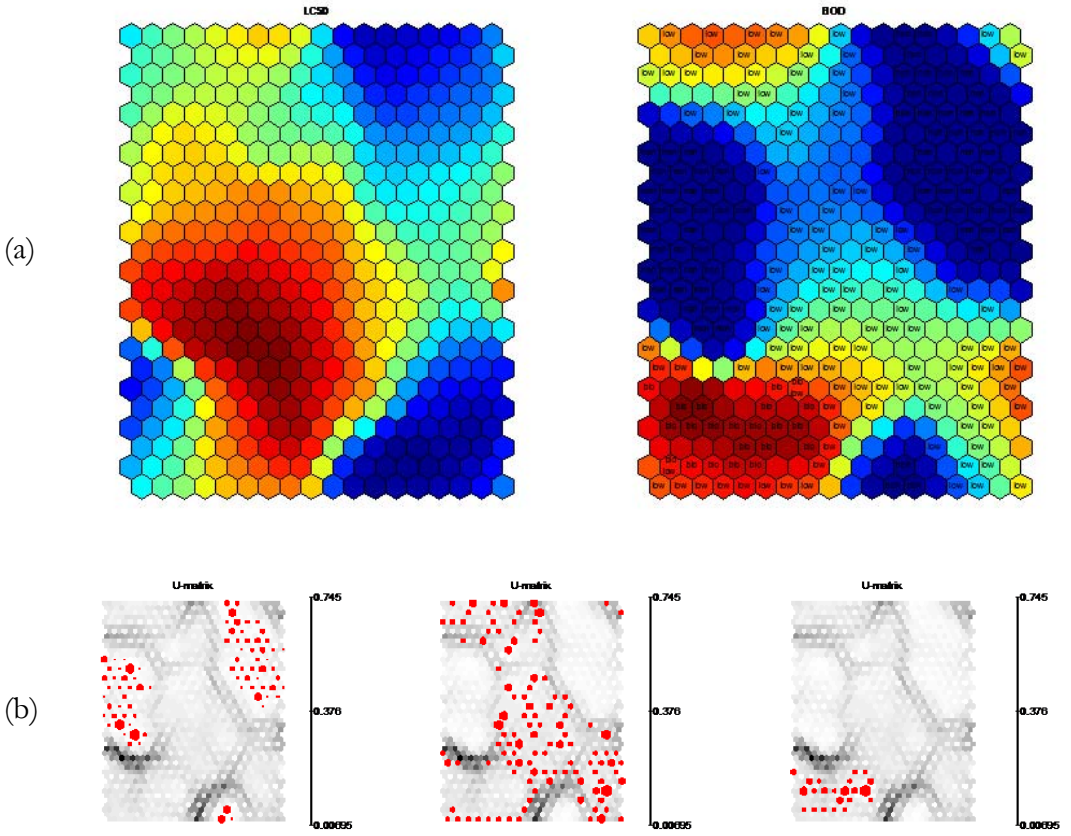


Figure 3.24. SOM for a subset of 330 MITI chemicals characterized with the nine input variables used in Figure 3.23 together with experimental LC50 (lethal concentration for half of the population in a 96 h test) toxicity index for the Orange-red killifish. (a) Component planes corresponding to LC_{50} (left) and BOD (right). (b) U-matrices and response surfaces (hit maps) for three ranges of BOD (left) [0%], (middle) [1-10%], (right) [>10%]

The partitioning into chemical families in Figure 3.24 after the inclusion of ecotoxicological information is more complex than the previous partitions shown in

Figures 3.21 and 3.22. The clustering of this new SOM indicates in Figure 3.24 the existence of 11 chemical families. Figure 3.25 depicts the distribution of these families over the SOM space. The labeling of each family uses a majority voting approach in which the label with most instances in the class is used as identifier. The use of ecotoxicological information is a clear example on how the addition of new information could result in a better partitioning of the chemical space. As a consequence it is of great importance the proper selection of information to obtain the *best view* of the chemical space. EDA techniques play an important role in this process by helping to visualize the contributions of each variable to the input space mapping.

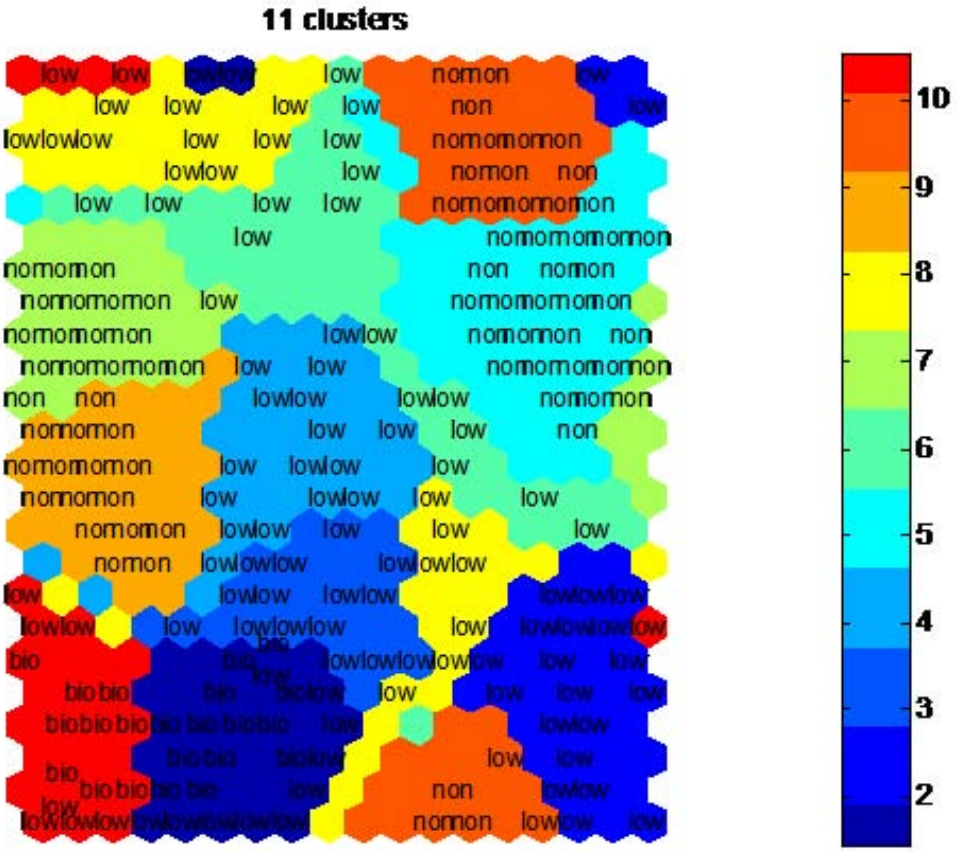


Figure 3.25. K-means clustering of the SOM in Figure 3.24 obtained with minimization of the Davies-Bouldin index. Each of the 11 chemical families detected is labeled by majority voting

3.4 Conclusions

This chapter has introduced the components of the first tier of the framework proposed in the current study and described in Chapter 2. Section 3.1 has shown that exploratory data analysis plays an important role in the characterization of the input space of variables. This exploratory analysis must be a prior step before attempting to develop data-driven models. The correct application of EDA

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

techniques results in the extraction of useful information about relationships between variables and the internal structure of the corresponding data space. In section 3.2 the use of the SOM algorithm as an EDA technique has been illustrated by first presenting the main properties of the algorithm and then introducing all the graphical techniques which provide useful insight into the input data and the SOM structure.

As an application example of the SOM-based EDA, the exploration of the chemical space corresponding to biodegradation in water which is complex to model has been performed in section 3.3. Two data sets containing physicochemical properties and molecular descriptors have been used to assess the EDA capabilities of the SOM. It has been shown that the inspection of the SOM component planes provides precise indications about the relationships between variables and permits the detection of redundant variables contributing to the structure of the chemical space with similar information. The information contained in the c-planes constitutes an important source for knowledge extraction from data. In addition, the use of distance matrices such as the U-matrix produces a clear picture of the internal structure of the data space. Clusters in the U-matrix induce a fine-grained partitioning of the chemical space that can be used to create local models fitted to the particular characteristics of each portion of the chemical space. Clustering of the SOM reference vectors coarsens the U-matrix clustering and permits the extraction of similar chemicals in terms of the variables describing their chemical space. The effects of the initial data transformations (normalization and scaling) as well as different parameters concerning both the SOM algorithm and the input data have also been analyzed.

3.5 References

BECHER, J. D., BERKHIN, P., FREEMAN, E. Automating exploratory data analysis for efficient data mining. In *Proceedings of the Sixth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*. KDD '00. ACM Press, New York, 424-429, 2000.

CHAMBERS, J., CLEVELAND, W., KLEINER, B., TUKEY, P. *Graphical Methods for Data Analysis*, Wadsworth, 1983.

CRAMER, H. *Mathematical Methods of Statistics*, Princeton University Press, 1999.

DAVIES, D.L.; BOULDIN, D.W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**(2):224-227, 1979.

ERWIN, E., OBERMAYER, K., SCHULTEN, K. Self-organizing maps: ordering, convergence properties and energy functions. *Biol. Cybern.*, **67**:47-55, 1992.

FRITZKE, B. Growing Cell Structures – A Self-Organizing Neural Network for Unsupervised and Supervised Learning, *Neural Networks*, **7**(9):1441-1460, 1994.

GOODMAN, L.A., KRUSKAL, W.H. Measures of association for crossclasification. J. Amer. Statist. Ass., **49**:732-764, 1954.

GRAY, R.M. Vector Quantization. *IEEE ASSP Mag.* **1**:4-29, 1984.

HEBB, D.O. *The organization of behavior*. Wiley, New York, 1949.

KANGAS, J., KOHONEN, T., LAAKSONEN, J. Variants of the Self-organizing map. *IEEE Transactions on Neural Networks*, **1**(1):93:99, 1990.

KASKI, S., LAGUS, K. Comparing self-organizing maps. In *Proc. of ICANN'96*, 809-814, 1996.

KIVILUOTO, K. Topology preservation in Self-organizing maps. In *Proc. of IEEE International Conference on Neural Networks*, **1**:294-299, 1996.

KOHONEN, T. Self-Organizing Maps: Optimization approaches, in *Artificial Neural Networks*, Kohonen, T., Mäkisara, K., Simula, O. and Kangas J. Eds. Amsterdam: North-Holland. **2**:981-990, 1991.

KOHONEN, T. The Self-Organizing Map. *Proc. IEEE*, **78**:1464-1480, 1990.

KOIKKALAINEN, P., OJA, E. Self-organizing hierarchical feature maps. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'90)*, 279:284, 1990.

LINDE, Y., BUZO, A., GRAY, R.M. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, **28**(1):84-95, 1980.

LIPINSKI, C., HOPKINS, A. Navigating Chemical Space for biology and medicine. *Nature*, **432**:855-861, 2004.

MACKAY, D., SHIU, W.Y., MA, K.C. *Illustrated Handbook of Physical chemical properties and environmental fate for organic chemicals.* 1st Ed. Lewis, Chelses, USA, 1992.

MARTINETZ, T.M., BERKOVICH, S.G., SCHULTEN, K.J. "Neural gas" network for vector quantization and its application to time series prediction. *IEEE Transactions on Neural Networks*, **4**(4):558-569, 1993.

OPREA, T.I., GOTTFRIES, J. Chemography: the art of navigating the chemical space. *J. Comb. Chem.* **3**:157-166, 2001.

ROBBINS, H., MONRO, S. A stochastic approximation method. *Ann. Math. Statist.* **22**:400-407, 1951.

SABLJIC A., PEIJNENBURG W. Modeling lifetime and degradability of organic compounds in air, soil, and water systems. *IUPAC Pure and Applied Chemistry*, **73**:1331-1348, 2001.

SAMMON, J.W. A nonlinear mapping for data structure analysis. *IEEE Transactions of Computers*, **C-18**(5):401-409, 1969.

TUKEY, J.W. *Exploratory Data Analysis.* Addison-Wesley, 1977.

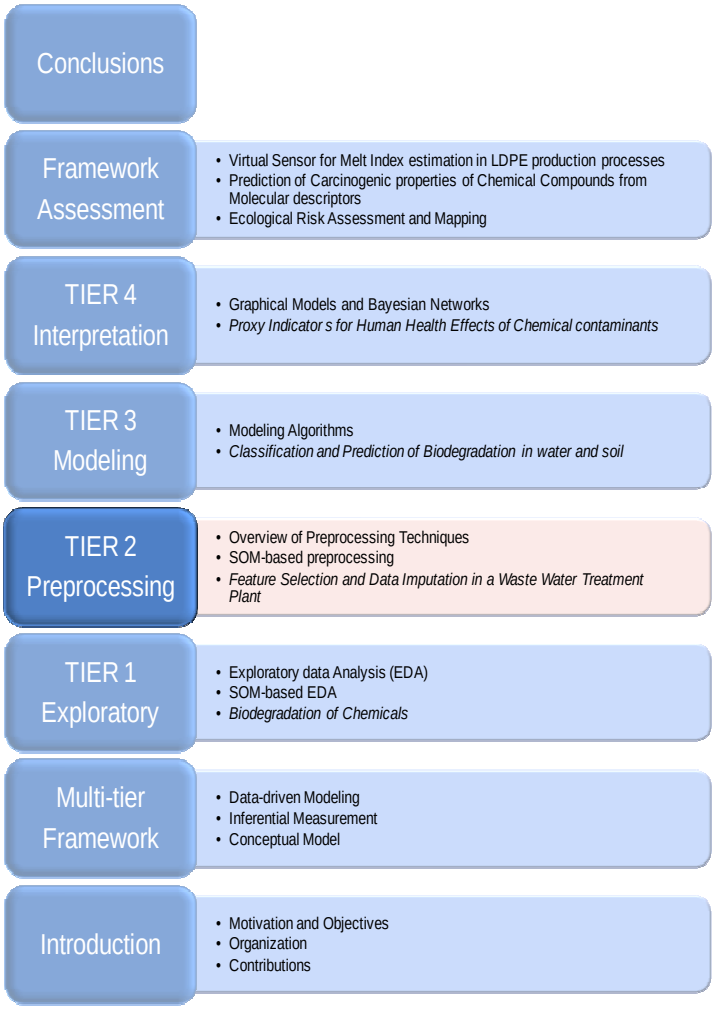
VELLEMAN, P., HOAGLIN, D. *The ABC's of EDA: Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury, 1981.

VESANTO, J. SOM-Based data visualization methods. *Intelligent Data Analysis,* **3**:111-126, 1999.

ZADOR, P. L. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inf. Theory*, **28**:139–149, 1982.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 4

Tier 2: Preprocessing Level

Data preprocessing techniques are required to deal with noisy, incomplete or inconsistent data. The goal of these procedures is to maximize the outcomes of subsequent inference models. Chapter four is focused in these preprocessing tasks and introduces the elements of the framework related to data transformations, detection of outliers, selection of the best features, and data imputation.

4.1 Background

Preprocessing techniques are required to ensure the quality of the models developed using a data-driven approach. According to Han and Kamber (2001), preprocessing techniques can be grouped as:

- data cleaning, to aim at removing noise and detecting inconsistencies;
- data integration, to merge data from multiple sources into a coherent data store;
- data transformations, which can be applied to improve the efficiency and accuracy of learning algorithms;
- data reduction, to decrease the size of the data set by aggregating variables or removing redundant information.

The development of models often requires the selection of the information that best represents the data space, i.e., the application of data selection techniques. Data completion techniques are also needed to deal with missing data in most real world modeling tasks. This section gives a short overview of all these preprocessing procedures.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

4.1.1 Basic preprocessing techniques

The set of basic transformations that transform data into a suitable representation are generally known as preprocessing operations. However, within the scope of this study we will use the term basic preprocessing to differentiate simple transformations from more elaborate treatments such as feature selection and automatic data completion.

The preliminary step in the characterization of the data space for any given data set is the analysis of its statistical properties. As pointed out in the previous chapter, this can be carried out by computing common univariate statistics and subsequently analyzing the shapes and properties of their distribution functions. In most modeling tasks this initial screening process provides a clear indication about the dispersion and variation ranges of each variable. This is of importance for the normalization of data as well as for the definition of the application domain of the corresponding models. The effects on the exploratory analysis of data transformations, e.g., the normalization techniques used, have been illustrated before. The preprocessing tier is the responsible of taking advantage of these effects to select the most appropriate representation to build good classifiers or predictors.

Lets us introduce some notation to describe these basic preprocessing procedures. Let *f* be a set of *n* features, $f = [f_1, f_2, ..., f_n]$. The preprocessing transformations aim at creating a new vector *f'* with a new dimension *n'* that will contain better features than in the native data set. This process is very domain specific and strongly depends on the quality and quantity of available data. The most commonly used preprocessing techniques are:

- *Standardization*. When different features that refer to a comparable object have different scales (for instance different units), centering and scaling methods must be applied. Features can be standardized by $f_i' = (f_i - \mu_i)/\sigma_i$, where μ_i and σ_i are respectively, the mean and standard deviation of f_i over the data set.

- *Vector Normalization*. Sometimes it may be necessary to make input features invariant on the size of the whole input data vector, i.e., to consider only the direction in the n-dimensional space ignoring magnitude. In this case, the input data vector is divided by its norm $f' = f/\|f\|$.

- *Linear and non-linear space embedding methods*. In cases where the dimensionality of input data is very high it may be convenient to project the data into a lower dimension space. Methods such as Principal Component Analysis (PCA), Multidimensional scaling or Random Projection can be applied.

- *Non linear expansions*. In the opposite situation as described above, it may be interesting to increase the dimensionality of data, especially when high order interactions between variables have to be considered to derive good models. In such cases, preprocessing transformations such as the inclusion of the products of certain variables may be applied.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

- *Feature discretization*. If the algorithms used to build the models are unable to handle continuous data or if the use of categorical data is more adequate in certain application domains (for instance in classification tasks), it would be helpful to discretize continuous values into a finite set of labels.

These preprocessing techniques may alter the dimensionality of the input space and even change the shape of the distribution of certain features.

Preprocessing techniques are also needed to detect the application domain defined by data, i.e., the domain where models can be safely built without hindering their validity. The application domain will depend on the intrinsic characteristics of the data used to build models. The effects due to the presence of "special" data samples, either in the form of outliers or as highly influential points, have to be detected and accounted for.

One of the techniques that can be used for this purpose is the Williams Plot (Ordinary Least Squares Outlier and Leverage Plot or OLS). This method is based in the plot of the studentized residuals (standardized cross-validated residuals) versus the leverages (hat diagonals). Each i^{th} studentized residual is computed by dividing the i^{th} residual by its standard deviation computed excluding the i^{th} observation.

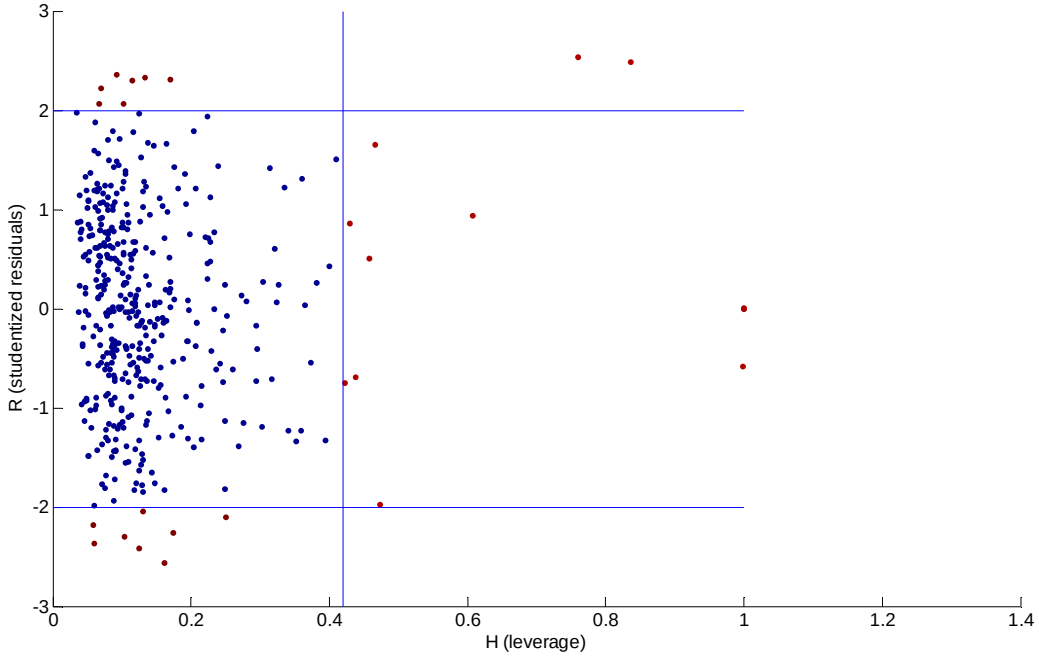


Figure 4.1. Williams plot or outlier and leverage plot (OLS). Points located above or below the horizontal lines represent outliers of the linear model. Points located to the right of the vertical line, exceeding the warning leverage, are highly influential for the linear model

The horizontal straight lines in the Williams plot depicted in Figure 4.1 indicate the presence of outliers. Usually these lines are located at 2 o 3 times the standard deviation. The vertical line indicates highly influential information (data) for the model. Usually this line is located at the so-called *warning leverage h** which is generally

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

fixed at $3k/n$, where k is the number of model parameters (number of input variables + 1 for OLS), and n is the cardinality of the training data set. Although the Williams plot is based on linear regression techniques it provides a convenient tool for the preliminary assessment of the applicability domain of a model. The simultaneous presence of both highly non-linear data relationships and non-linear modeling techniques require the use of additional analytical techniques to determine the application domain.

4.1.2 Feature Selection techniques

The problem of feature selection is a challenge shared by most fields of science and engineering. It is well known that data collected to analyze some real world phenomenon are not equally informative, mainly due to problems related to noise incorporated during the data acquisition process, i.e., acquired during the realization of the experiment. Also, information may be contradictory due to the inclusion of irrelevant or redundant variables. In fields such as environmental modeling this problem is an important issue due to the usually large amount of information needed to obtain good models. The selection of a good set of features (variables) to learn a given concept is a key issue, making this procedure one of the most important steps during the modeling process.

The problem of feature selection can be stated as: Given a set of attributes (features) V and a target variable T, find the minimum set of attributes that achieves maximum classification performance of T (Kira and Rendell, 1992; Piramuthu, 2004). The main challenges in this process are the following. If too few features are selected the information content in this set of features is low and models have poor performance. On the other hand, if too many irrelevant features are selected the effects of the noise present in data may overshadow relevant information content (Kohavi and Pfleger, 1994).

Artificial Intelligence and Machine learning techniques provide algorithms to perform the necessary data mining tasks. These algorithms, which act as preprocessors of a more complex learning task, can be represented as heuristic search problems. Under this paradigm any variable selection method can be classified in terms of four basic parameters that drive their operation mode.

- *The starting point (or points) in the search space*, which determines the components of the initial configuration of variables to perform the selection process. This leads to the following selection methods: (i) Forward selection where the process is started without variables or with only a small subset of them and proceeds by adding new variables after each iteration; (ii) backward elimination in which the process is started with the complete set of variables and continues with the deletion of some variables after each iteration; (iii) any combination of these alternating insertion and deletion steps. Figure 4.2 depicts the selection procedure.

- *Organization of the search process in the variable space*. Obviously an *exhaustive approach* will be usually impractical in the majority of real world applications, because for a given problem there exist 2^n possible subsets of n attributes. A more realistic

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

approach could be to adopt a greedy method, such as *heuristic* or *stochastic* search schemes, to traverse the space. In this situation only a subset of all possible configurations is explored. Examples of this approach are the *simulated annealing*, *hill-climbing* and *best-first search* (Kirkpatrick et al., 1983; Pearl, 1984; Russell and Norvig, 1995).

- *Strategy to evaluate the subsets of attributes*. Once the search over the space of variables starts, a metric to evaluate the goodness of an attribute is needed. Commonly used metrics are the attribute's ability to discriminate among classes present in the training data, or some other metrics based on information theory.
- *Criterion to stop the search procedure*. One may stop adding/removing attributes when none of the alternatives improves the accuracy of the previous set or continue generating candidate sets until reaching the other end of the search space to select the best candidate. A more robust alternative is to order variables according to some relevance measure and then apply some threshold value to determine the break point and select the best ones.

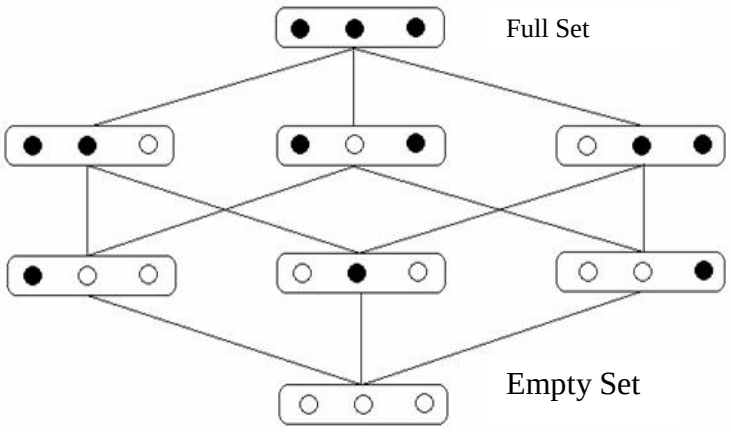


Figure 4.2. Exploration of the feature space, where forward selection proceeds by using a bottom-up scheme while backward elimination proceeds with a top-down approach

Feature selection facilitates a better understanding of the application domain, a more cost-effective predictor development and the improvement of prediction performance. According to the terminology proposed elsewhere (Guyon and Elisseeff, 2003) different approaches are available for feature selection. First, algorithms that embed the selection process within the basic learning algorithm. Second, algorithms that use variable selection to *filter* the variables passed to the learning step and, finally, algorithms that consider variable selection as a *wrapper* around the whole learning process.

Examples of the first approach are the methods for inducing logical descriptions, i.e., ID3, C4.5 and CART (Mitchell, 1997), which are characterized by the addition or deletion of features from the concept description based on some measure of prediction error after the presentation of patterns never seen before by the feature selection algorithm. The main drawback of this group of methods resides in the fact

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

that a substantial decrease in accuracy is observed when irrelevant variables are introduced into the target concepts. Also the presence of interactions between attributes (co-linearity) can lead a relevant feature to look no more discriminating than an irrelevant one.

The second group of methods (John et al., 1994) use characteristics of the dataset to filter out irrelevant attributes before the learning process begins. These methods are independent of the learning mechanism used. One of the simplest filtering schemes consists in the evaluation of the correlation of each individual feature with the desired target value, and choosing the k variables with the highest correlation. Other filtering methods construct sets of new variables (high-order features) as a combination of the original ones. Examples of these kinds of filters are techniques like principal component analysis (PCA) and independent component analysis (ICA). Generally, all these methods show some improvement over embedded methods because they are less sensitive to the effect of irrelevant features. Furthermore, since filters make their selection by ranking variables according to some measure, as shown in Figure 4.3, they reflect how well the classes separate from each other. Filters are relatively robust against over-fitting, although sometimes fail to select the most "useful" features.

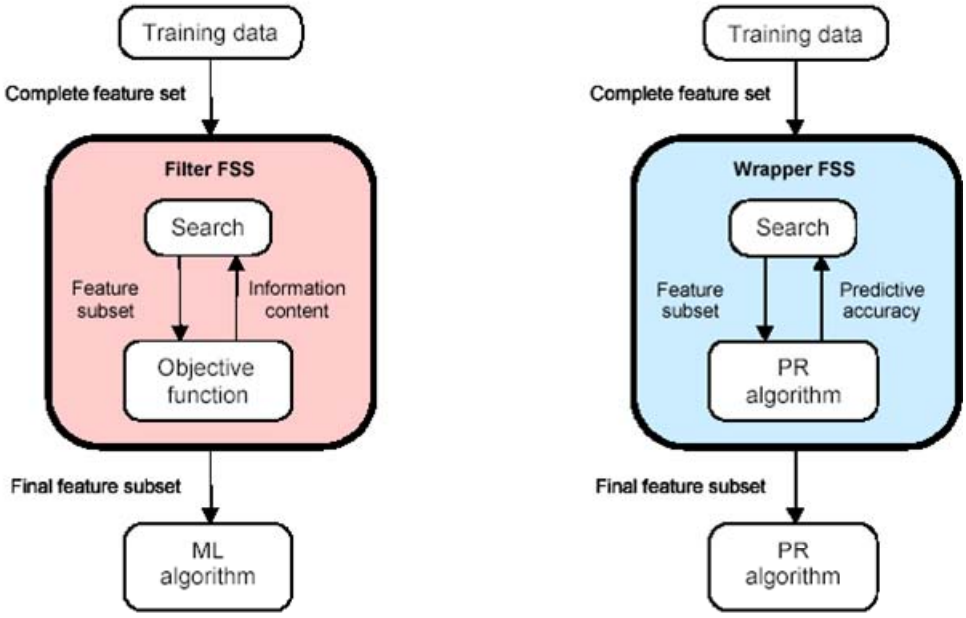


Figure 4.3. Filter and wrapper algorithms for feature selection

Finally, the task of variable selection with wrapper methods occurs outside of the learning method but uses the learning algorithm as a procedure to drive the selection process. Figure 4.3 summarizes the differences between wrappers and filters. The main drawback of wrappers resides in its computational complexity because after each selection of a candidate set of variables the learning process has to be performed to assess the 'goodness' of the selected set of variables. Nevertheless, wrappers yield good results in problems involving irrelevant and redundant variables. In addition, wrapper methods are prone to over fitting when finding the most

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

"useful" features since they assess subsets of variables according to their usefulness for a given prediction task.

There exists numerous feature selection algorithms reported in the literature. An extensive review is given by Guyon and Elisseeff (2003). Different feature selection algorithms using both filter and wrapper approaches have been applied in the current study to several examples of scientific and engineering interest. A new feature selection method based in Self-Organizing maps has also been developed and tested. This section describes these algorithms together with references for their implementation.

Correlation-based Feature Selection (CFS). The CFS algorithm (Hall, 1998) is a filter that uses a correlation-based heuristic to evaluate the merit of a subset of features. The hypothesis of the CFS method are: "*Good feature sets contain features highly correlated with the class, yet uncorrelated with each other.*" Ghiselli (1964) formalizes this concept by defining the merit score as,

$$merit_S = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k+1)\bar{r}_{ff}}} \tag{4.1}$$

where $merit_s$ is the merit of a feature subset S containing k features, $\bar{r}_{cf}$ is the mean feature-class correlation, and $\bar{r}_{ff}$ is the average feature-feature autocorrelation. The numerator of Eq. (4.1) can be considered as an indicator of how predictive the selected features are of the featured class. On the other hand, the denominator reflects how much redundancy there is among features. A measure based on conditional entropy is used to measure correlations between features and class, and between features. Continuous features are transformed to categorical features by using the discretization methods of Fayyad and Irani (1992). If X and Y are discrete random variables, equations (4.2) and (4.3) give the entropy of Y before and after observing X,

$$H(Y) = -\sum_{y \in Y} p(y) log_2 p(y) \tag{4.2}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2 p(y|x) \tag{4.3}$$

Equation (4.3) is also known as the information gain and accounts for the amount of information gained about Y after observing X, which is equal to the amount of information gained about X after observing Y (Quinlan, 1993).

ReliefF. The ReliefF algorithm (Kononenko and Simec, 1995) is a filter that uses the nearest neighbor algorithm to rank a complete set of features. For each input pattern, the closest pattern of the same class (*nearest hit*) and the closest example of a different class (*nearest miss*) are selected. The score $S(f)$ of the *f-th* variable is computed as the average over all examples of the difference between the distance to the nearest hit and the distance to the nearest miss, when projecting the *f-th* variable. The key idea of the ReliefF algorithm is to rank the quality of attributes according to

ISBN:978-84-691-1008-9/DL: T.2230-2007

how well their values discriminate between instances that are near to each other and belong to different classes.

ANNIGMA. This method follows a wrapper approach for feature selection. It uses a weight analysis heuristic approach named *Artificial Neural Net Input Gain Measurement Approximation (ANNIGMA)* (Hsu et al., 2002). The ANNIGMA algorithm ranks features according to their relevance, estimated by the weights of the networks associated with these features. The backpropagation neural network is used as the underlying machine-learning algorithm; network weights can be considered as an indication of the contribution (*gain*) of the input signal to the output target. Input signals that are irrelevant to the output will induce high errors in the prediction if the neural network tries to depend on them. Hence, the training algorithm will reduce the weights of these inputs so that they do not contribute to the output. Any of the three search strategies outlined in the previous section can be used with ANNIGMA. Forward selection performs particularly well for datasets that need a small number of features and runs the fastest. Backward elimination performs consistently well for all datasets, while backward stepwise elimination is effective for datasets with a large number of original attributes. Figure 4.4 outlines the main building block of the ANNIGMA algorithm.

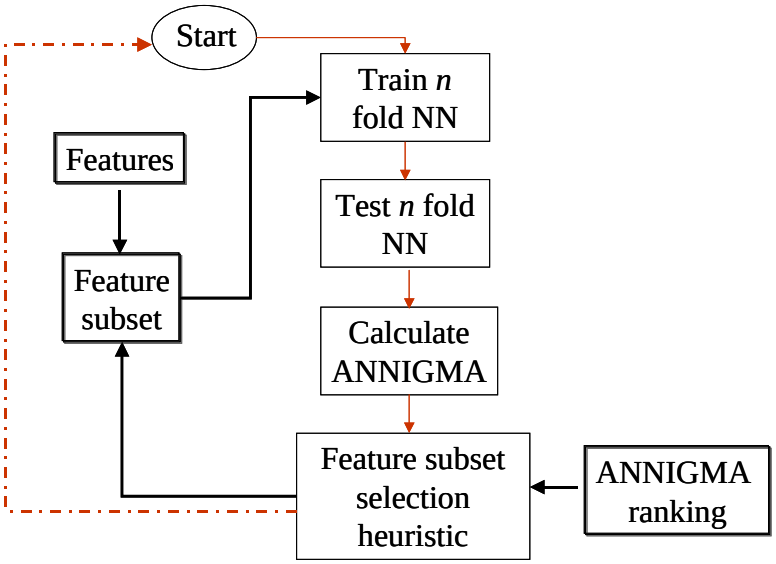


Figure 4.4. The ANNIGMA wrapper algorithm

4.1.3 Detection of redundant data

The detection of irrelevant or redundant information constitutes also an important step in the characterization of the data space. Usually redundant variables do not add a substantial amount of information to the model. Redundancy and co-linearity are both indicators of the presence of duplicated information that can be potentially dangerous for the development of good models. Redundant or repeated variables can be detected by using some kind of exploratory data analysis, e.g., by computing the cross-correlation matrix of the variables.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

A widely used indicator for the presence of co-linearity among variables is based on the K correlation index. The K theory was first developed by Todeschini (1997). This theory introduces the K correlation index which is aimed at evaluating multivariate correlation in data. The total correlation contained in a data set can be estimated from the eigenvalue distribution obtained from the eigenvalue decomposition of the corresponding correlation matrix.

Let $\lambda_1, \lambda_2, \ldots \lambda_p$ be the set of the p eigenvalues obtained from PCA applied to the correlation matrix of a data set. The K index is defined as,

$$K = \frac{\sum_{m=1}^{p} \left| EV_m - \frac{1}{p} \right|}{\frac{2(p-1)}{p}}, 0 \leq K \leq 1 \qquad (4.4)$$

where

$$EV_m = \frac{\lambda_m}{\sum_{m=1}^{p} \lambda_m} \qquad (4.5)$$

is the explained variance from the *m-th* principal component. The K correlation index is a redundancy index, that is equal to 1 when all variables are correlated and 0 otherwise. This index can be used to detect co-linearity situations in the development of regression models. Redundancy or overlapping among independent variables should be avoided to obtain robust prediction models. This can be measured by computing the K index for the independent variables (K_{xx}) and comparing their value with the global correlation including the target (K_{xy}). The difference, ΔK (defined as K_{xy}-K_{xx}) must always be greater than zero.

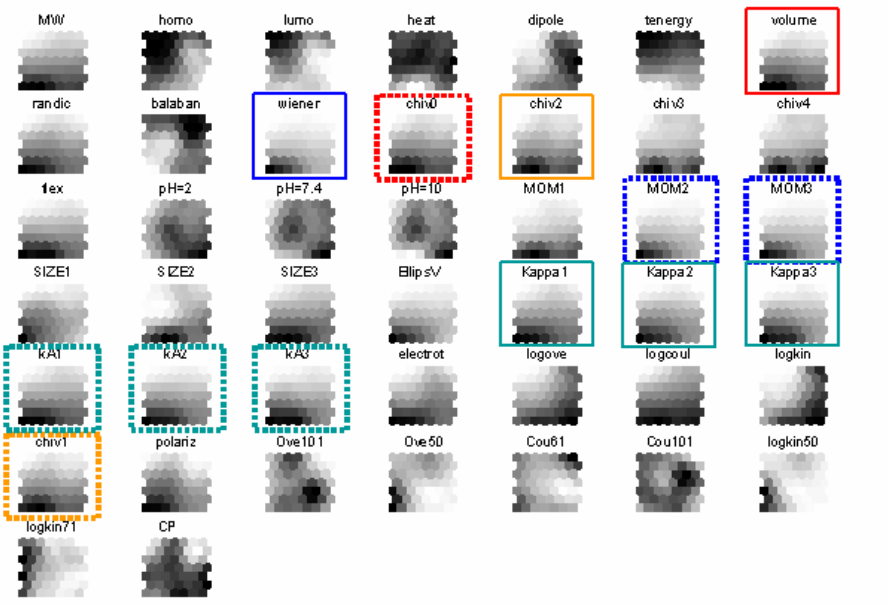


Figure 4.5. Detection of redundant variables by visual inspection of the SOM c-planes. The chemical input space correspond to the mammalian carcinogenic potency of 104 aromatic chemicals

ISBN:978-84-691-1008-9/DL: T.2230-2007

Other commonly used redundancy detection techniques are clustering methods since relationships between subsets of variables can easily be found; redundant or highly related variables can be identified by their membership to the same cluster. The use of self organizing maps in the current study has proved to be very useful for the detection of variables that account for similar information. The projection of input features over the feature space spanned by the SOM provides visual clues of their internal structure and relationships. Figure 4.5 shows the representation over the SOM space of a set of molecular descriptors that are relevant in the interpretation of mammalian carcinogenetic activity, measured as the carcinogenic potency CP = log[MW•1000/TD_{50} calculated from the toxic dose 50% (TD_{50}), which is triggered by 104 aromatic chemical with nitrogen containing substituent. Variables that account for the same type of information in relation to carcinogenic potency (CP) have a similar topological distribution of their values over the map. Examination of the component planes (*c-planes*) plotted in Figure 4.5 reveals the redundant variables.

4.1.4 Missing Data

An additional problem to deal with in the development of data-driven models is the recovery of missing data and to understand why some data are missing. This phenomenon can informally be thought of as being caused by certain combination of three main factors: random processes, processes that are measured, and processes that are not measured (Little and Rubin, 1987). These categories can be formalized as (Scheffer, 2002):

- *Missing Completely at Random (MCAR).* When data are MCAR, missing cases are not different from non-missing cases, in terms of the analysis being performed. In this case the missingness mechanism does not depend on the variable of interest or any other variable observed in the dataset.
- *Missing at Random (MAR).* Missing data depend on known values and, thus, they can be described fully by variables observed in the data set. In this situation the missingness mechanism depends on some variable of the dataset but not on the variable of interest.
- *Missing not at Random (MNAR).* The data are missing in an unmeasured fashion, i.e. depending on events or items that cannot be measured. In this case the missingness mechanism depends on the actual values of the missing data and constitutes the most difficult factor to deal with by the model.

In most real world tasks involving inferential measurement techniques, such as chemical processes and environmental modeling, a combination of all the above mechanisms are usually found to be responsible for sensor faults. Malfunctioning of field sensors pose an additional challenge to on-line inferential techniques since it may be very difficult to detect given the noise present under this real operating conditions. This important issue in not the subject of the current study.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Several preprocessing methods can be applied when the values of the missing data cannot be recovered. These techniques usually focus on "removing" the missing information, either by ignoring patterns with incomplete information (case deletion) or by substituting the missing variables with plausible values (mean, median, regression, etc.). The first option is not adequate for multivariate datasets containing a large number of attributes, as is usually the case in chemical processes, since it implies discarding an unacceptable high portion of input data; even if the per-variable rates of missingness are low, few patterns may have complete data for all variables. In addition case deletion procedures may bias the training of software (virtual) sensors if the patterns that provide complete inputs are unrepresentative of the entire sample. The second approach (imputation of missing values) is the most adequate for the development of data-driven models and inferential systems. Imputation techniques are divided into single and multiple imputations. Single imputation is used to recover the missing values by using statistical measures such are the mean, median or mode, or more elaborated techniques, i.e. regression, EM algorithms or hot deck substitution (Dempster et al. 1977). On the other hand, multiple imputations (Rubin, 1987) produce *n* estimations for each of the missing values by using different data models. In this situation, the imputed values are computed using statistical methods and the uncertainty associated to the imputed value can be estimated.

4.2 SOM-based Preprocessing Techniques

The application of Self-Organizing maps (Kohonen, 1990) at the preprocessing level is introduced in this section. A new method based in the use of the SOM developed to perform feature selection using a metric based on the dissimilarity between maps is presented. The SOM is also used to select the optimal pair of train/test sets in order to assure that the external validation of models is performed into the application domain defined by available data. Finally, both, single and multiple imputation methods using SOM are discussed.

4.2.1 Feature Selection using SOM Dissimilarity

This filter methodology (Rallo et al., 2002) was proposed to select the most suitable subset of descriptors from a given initial pool of molecular information. The algorithm is based on the use of the self-organizing map to find the subset of descriptors which provides, in average, a similar amount of information than the complete set of variables. Accordingly, this constitutes the premise of this feature selection method which assumes that descriptors that cover the same region on the map essentially contribute with a similar type of information to the model (Espinosa et al., 2002).

The selection of the best set of variables is based on the projection of all the candidate subsets of variables onto the space generated by the SOM. Comparison of the resulting maps, using an adequate dissimilarity measure, provides an indicator of the relevance of each combination of variables with respect to a target variable. A dissimilarity measure is used here to compare the positions of the reference vectors

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

in different map structures. This measure is based on a *map goodness* measure (Kaski and Lagus, 1997) that combines an index of the continuity of the mapping from the dataset to the map grid with a measure of map accuracy. The dissimilarity of two maps L and M is defined as the average difference of its goodness,

$$D(L,M) = E\left[\frac{d_L(x) - d_M(x)}{d_L(x) + d_M(x)}\right] \tag{4.6}$$

In this equation (4.6) E is the average expectation, and *d(x)* the Euclidean distance over the map from the winner unit or best matching unit (BMU), denoted by $m_{bmu(x)}$, to the second best cluster or BMU, denoted by $m_{bmu'(x)}$. Of all possible paths between $m_{bmu(x)}$ and $m_{bmu'(x)}$ the shortest path passing continuously between neighbor units is selected,

$$d(x) = \|x - m_c(x)\| + \min_i \sum_{k=0}^{K_{c'(x)}-1} \left\|m_{I_i(k)} - m_{I_i(k+1)}\right\| \tag{4.7}$$

First all available data are used to self-organize a SOM. Next, c-planes are extracted and the U-matrix is computed. The process of selection of relevant variables starts by the identification of the redundant variables. A redundancy index (Rallo et. al., 2005) that takes into account the correlation between variables and their representation over the map (c-planes and U-matrix) is used. If the value of this index is greater than a certain threshold (threshold values vary from 0.95 to 0.98) the variable is discarded from the dataset because the information that it provides is redundant with that of other variables (see Figure 4.5). After removing redundant information, a new SOM is created. Each c-plane is then extracted creating a new dataset that contains the SOM internal representation of each variable. These planes are clustered again using a new SOM. By doing this, some of the SOM units will become the cluster representatives for the set of variables that are closer to these units.

The next step in the selection of the most suitable set of variables is the identification of an initial set to start the search procedure. This process is based on the clustering of the SOM and on the assumption that variables located in the same cluster contribute with similar information to the model. Thus, the starting point for the search is determined by choosing a representative variable for each cluster. The cluster descriptor is the variable with the highest correlation with the target variable. To avoid the inclusion of irrelevant variables in the initial set, only those variables with a correlation value with the target property higher than the average correlation for the whole set of variables are considered.

Since it is difficult to detect clusters by visual inspection of the U-matrix, an iterative clustering of the SOM prototypes using the K-nearest neighbor algorithm with different k values is used. The criterion of a minimum Davies-Bouldin index (Davies and Bouldin, 1979) has been applied to determine the optimal number of clusters. This index is a function of the ratio between the sum of cluster compactness and inter-cluster separations. Figure 4.6 depicts an example of the resulting clustering of C-planes for the selection of the best molecular descriptors that describe the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

chemical space in relation to the modeling of carcinogenic properties of organic chemicals. The optimal value of k is found by minimizing the Davies-Bouldin index.

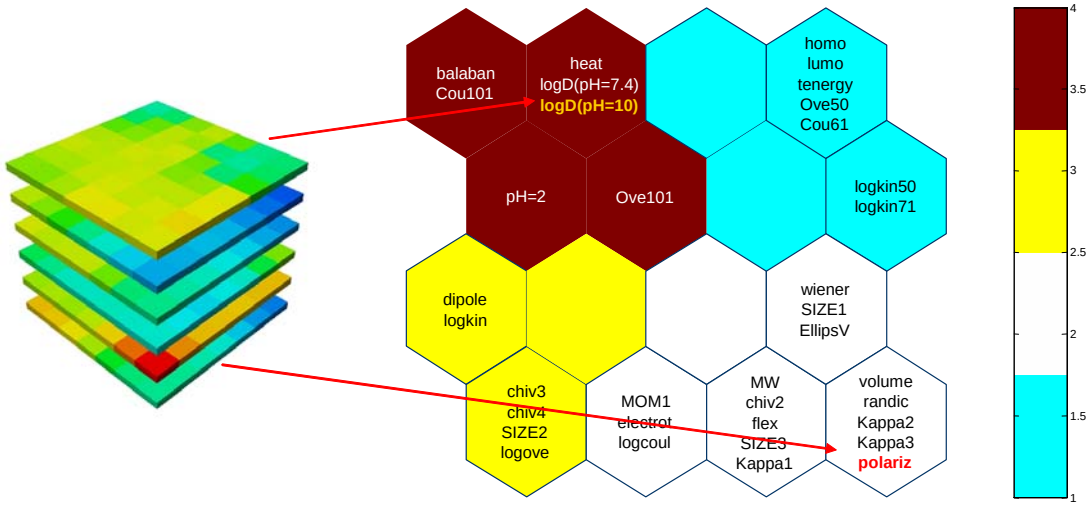


Figure 4.6. Classification of c-planes using the K-means algorithm and optimization with the Davies-Bouldin index to select the best set of chemical features to model carcinogenicity

The organization of the search proceeds from the initial set by building new subsets with the addition of the rest of variables, one by one, ranked by its correlation with the target property. For each of these subsets the dissimilarity with the remaining possible subsets is computed using Eq. (4.6). The process stops when the dissimilarity between all possible configurations has been computed. This procedure reduces the complexity of the search algorithm from $O(2^n)$ to $O(\frac{1}{2}n^2)$, where n is the number of input variables. The smallest average dissimilarity value for any given set of input variables indicates similarity in quality and quantity of the information presented by all maps. Thus, the process of including variables in the best set of input variables can be stopped when the dissimilarity measure stabilizes, i.e., the maps for these different variables are very similar. Note that any increase in dissimilarity, due to the inclusion of additional input variables in the previous subset, indicates that these additional variables do not provide any additional relevant information.

The effect of the inclusion of irrelevant information is tested by adding noise to the data used to train the map. The increase of noise produces more dissimilar maps and the value of dissimilarity index in Eq. (4.6) increases. Figure 4.7 corroborates that the inclusion of irrelevant information, i.e., data with added noise, increases the dissimilarity between maps for the feature selection problem dealt with in Figure 4.6.

The effect of the addition of variables using the ordering given by correlation with the target on dissimilarity values is illustrated in Figure 4.8. It can be seen that the average dissimilarity decreases after each new variable addition. This indicates that the successive addition of variables adds new relevant information to the subset of best features.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

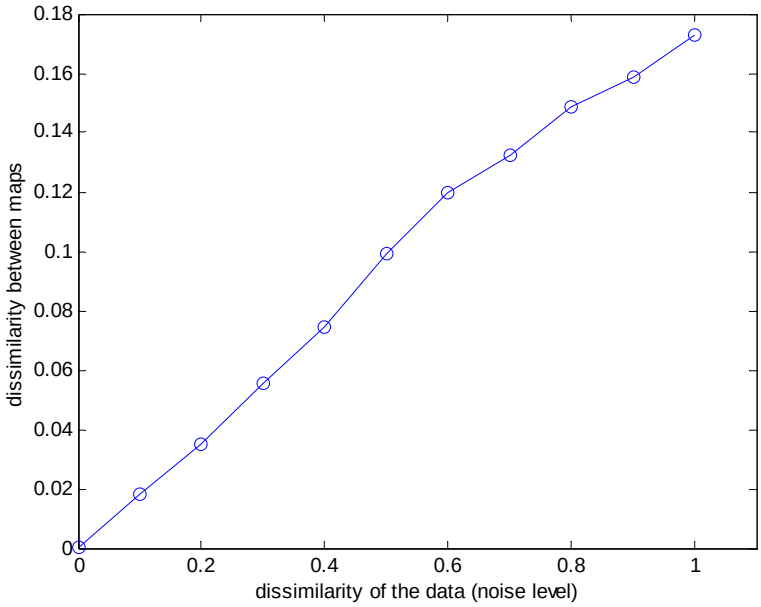


Figure 4.7. Evolution of the dissimilarity index after the addition of noise to data corresponding to the carcinogenicity modeling problem

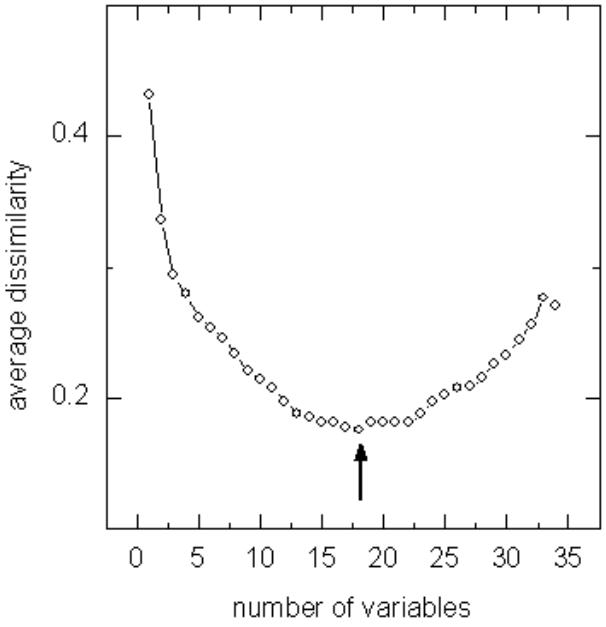


Figure 4.8. Variation of the average dissimilarity with the addition of variables for the problem of carcinogenicity prediction. The minimum in dissimilarity indicates the optimal subset of variables

The average dissimilarity between maps decreases until the minimum observed in Figure 4.8 is reached for variable no. 18, indicating that the addition of more variables beyond no. 18 doesn't add any further relevant information. In fact, the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

increase of the average dissimilarity indicates that the addition of more variables would add irrelevant information which will hinder subsequent modeling efforts.

4.2.2 SOM-based selection of Train/Test sets

An additional element to take into consideration for the implementation of data-driven models is the proper distribution of patterns into the train and test sets (Blum and Langley, 1997). The ideal splitting of data into train and test sets is one in which each of the members of the test set is close to at least one point in the training set. Furthermore, a representative selection of data spanning to the data space of interest should be included in these sets. There exist different approaches to build these sets. The easiest one is random sampling but its major drawback is that usually the resulting sets are not well balanced in terms of data space coverage. Other widely used splitting techniques are clustering algorithms. In fact the SOM algorithm can be used to obtain well balanced partitions from any dataset. This approach has been used in the present work to split the data space into consistent train and test sets for the proper evaluation of the predictive capabilities of all developed models by using external validation techniques.

The SOM-based splitting procedure can be stated as follows. The complete dataset including both input and target variables are used to build a self-organized map. It is expected that at the end of the SOM training, similar patterns will fall into the same SOM units. To select the training set it is assumed that the data point nearest to the centroid of each class is the most representative of that class. Thus the selection of data for the training set is performed by the inclusion of the nearest data point to the centroid for each SOM class. The remaining data are assigned to the test set. The coverage of the data space and the proper distribution of training and test sets can be visualized by using projection techniques such as PCA projection, multidimensional scaling or Sammon's maps.

4.2.3 SOM-based Data Imputation

The use of the SOM as a single data imputation tool has been explored in diverse application areas (Wang, 2002; Fessant et al., 2002; Rallo et al., 2003), ranging from chemical engineering to social sciences. The process of single imputation using the SOM is based on its capacity to deal with incomplete inputs since the presence of such patterns affects the calculation of distances. As missing components are excluded from the individual variable contributions to distance, the distances corresponding to incomplete patterns will be computed using fewer components. This can be interpreted as the error induced by the projection of the input data onto a lower dimensional space. Since the input components are normalized in the range [0,1] this error is upper-bound by the expression $\sqrt{n}-\sqrt{n-k}$, being n the dimension of the input space and k the number of missing variables. If the dimension of the input space is higher than the number of missing variables ($n>>k$) then the error induced in the calculation of the distance is low. It is important to note that after a proper self-organization process, similar patterns will be associated to neighboring units. This implies that if the error induced by missing variables in

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

the distance calculations leads to a miss assignment of the winner node, the new best matching unit (*bmu*) selected will be in the neighborhood of the correct one. Under these assumptions the SOM can be used as a data imputation system assuming that model degradation for current applications will not be linearly correlated with the rate of missing data (Samad and Harp, 1982).

Ensemble approaches to classification and regression have attracted a great deal of interest in recent years. These methods can be shown both theoretically and empirically to outperform single predictors in a wide range of tasks (Hansen and Salamon, 1990). The SOM-based single imputation model can be extended to multiple imputation schemes. The main idea resides in the concept of "model aggregation". Aggregation attempts to improve the quality of the imputed values by generating multiple versions of the imputation system $\phi_i(x)$, and combining their outputs in some prescribed way, usually by averaging,

$$\phi_{aggregated}(x) = \frac{1}{N}\sum_{i=1}^{N}\phi_i(x) \qquad (4.8)$$

where $\phi_{aggregated}(x)$ is the response of the aggregated imputation system and N is the cardinality of the ensemble. One of the elements required for accurate prediction when using an ensemble is recognized to be "model diversity", i.e., the disagreement between components in the ensemble. Diversity can be introduced by (i) manipulating the components of the input data (e.g., using feature selection), (ii) by randomizing the training procedure for each member of the ensemble (e.g., combining under-fitted with over-fitted models with different learning parameters and network topologies), (iii) by modifying the response with the addition of noise, and (iv) by re-designing the training set.

4.3 Feature selection and Data Imputation in a Waste Water Treatment Plant

The design and implementation of accurate and reliable models for chemical processes is a challenge to both chemical engineering and computer sciences. One of the emerging areas is that concerned with inferential prediction systems also known as "soft sensors". The inferential measurement systems based on neural networks are mostly developed using "data-based" methodologies, i.e., the models used to infer the value of target variables are developed using real-time plant data. This implies that inferential systems are heavily influenced by the quality of the data used to develop their internal models. In chemical processing plants, the number of process variables that can be measured is very large and the sampling rates used for these measurements are usually high, which imply the generation of large datasets containing lots of features. In those situations it is very useful to have an "intelligent system" capable of selecting the most relevant features needed to build an accurate and reliable experimental model for the process. Furthermore, the field sensors used to collect all these data are susceptible to damage and failure due to the hard environment around them. If one of these field sensors fails and failure is detected,

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

the data needed by the soft sensor are incomplete making the whole inferential system unusable. Thus, the preprocessing techniques for soft sensor development have to be capable of also recovering missing data. Finally, only truly relevant examples of the whole input space should be used for training the soft sensor. This application example focuses more in illustrating the proposed techniques for feature selection and data imputation than in the development of process models.

4.3.1 Problem statement and overview.

The current work presents a complete preprocessing methodology based on the use of Self-Organizing Maps (SOM) that covers both, feature selection and missing data imputation. The proposed method is evaluated using the Waste Water Treatment Plant (WWTP) benchmark available from the UCI Machine Learning Repository (Belanche et al., 1992; Blake and Merz, 1998). The list of identifiers for the measured and computed variables in this WWTP is given in Table 4.1.

Table 4.1. List of process variables for the WWTP in Figure 4.9

Symbol	Description	Units
Q	Flow	m^3/day
ZN	Concentration of Zinc	mg/l
PH	pH	
BOD	Biological Oxigen demand	mg/l
COD	Chemical Oxigen demand	mg/l
SS	Suspended solids	mg/l
SSV	Volatile suspended solids	mg/l
SED	Sedimentable solids	mg/l
COND	Electric Conductivity	mg/l

It should be noted that one of the most difficult problems in the modeling and control of these treatment plants is the construction of reliable process models since the development of detailed models based on fundamental principles and kinetic relationships is very difficult, expensive, and time consuming. Activated sludge is a common example of an industrial wastewater treatment process. In this process the inlet flow rate and composition are variable, the population of microorganisms (acting as living catalyst) varies over time (both in quantity and number of species), process knowledge is very limited, and the few available on-line analyzers tend to be unreliable. The amount of organic matter present is measured as the biological oxygen demand (BOD). This magnitude constitutes a key indicator of processed water quality. It is very convenient to have a reasonably accurate inferential model for BOD prediction because there is a five-day time delay that is inherent in any laboratory measurement for this quality variable. In addition the aeration tank also has a significant hydraulic time delay. Consequently, the experimental BOD data are not useful for purposes of process control.

The data used to test the proposed preprocessing framework correspond to daily measurements of the WWTP operation over a two-year period. The plant, whose layout is given in Figure 4.9, consists of three units: pretreatment, primary treatment by clarification, and secondary treatment by means of activated sludge. A total of

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

527 data records, each consisting of 38 process variables have been used. Of these, 29 correspond to measurements taken at different points in the plant (indicated in Figure 4.9), and the remaining 9 variables correspond to calculated performance measures. This benchmark contains 1480 missing values, corresponding to the 7.3% of the whole dataset. This constitutes one of the major drawbacks since the large amount of missing data that contains makes it unusable as training data for the development of a soft sensor system to infer water quality at plant's output.

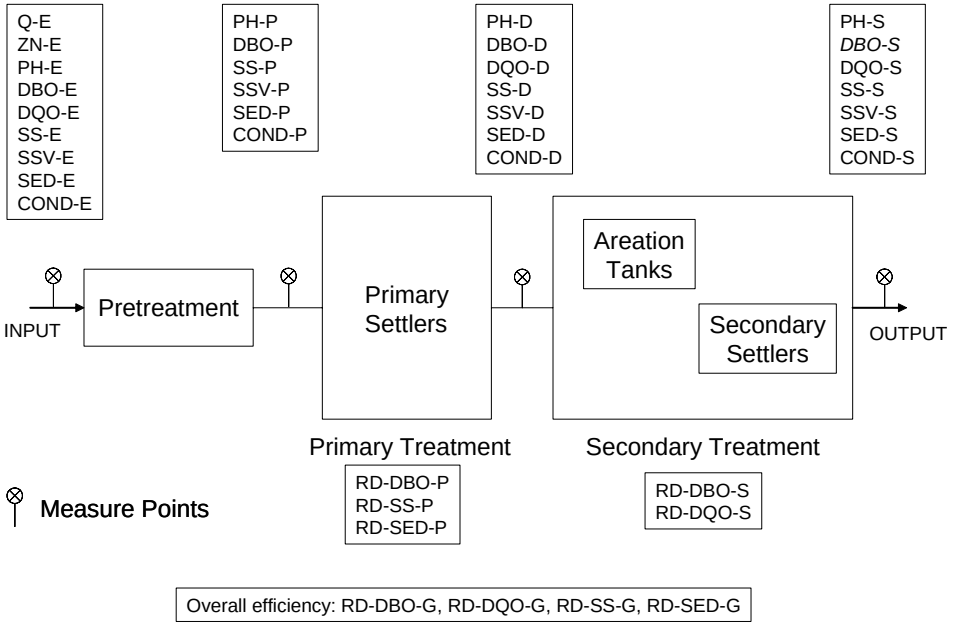


Figure 4.9. Layout of the WWTP with points of measurement and measured variables

The problem of missing data is solved by using the proposed SOM procedure for data imputation. A single soft sensor based in Radial Basis Functions (Moody and Darken, 1989) is first designed and implemented using the proposed preprocessing procedure. The target variable is the Biological Oxygen Demand (BOD) at the output of the treatment plant. In a second stage, an imputation system using an ensemble of SOMs is developed. Using the reconstructed data three virtual sensors for pH, chemical oxygen demand (COD) and biological oxygen demand (BOD) for the effluent water were implemented and trained using feed-forward neural networks.

4.3.2 Single Imputation Model

The model of single imputation using SOM is based in the estimation of the values of missing variables using the prototypes in the corresponding SOM clusters, as mentioned in the previous section. These prototypes can be selected and combined using several approaches. Two approaches have been applied in the current work: (i) Direct substitution with the corresponding component taken from the prototype vector of the *bmu*; (ii) substitution by a mean value obtained by averaging the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

corresponding components in the prototype vector of the *bmu* and those of a certain number of its neighboring units. The size of the neighborhood was chosen as three since the inclusion of more distant units did not improve the quality of the imputed data. Furthermore, an n-fold cross-validation with n=10 was performed for training to obtain a more accurate representation of the maps.

Figure 4.10 depicts the data corresponding to COD and BOD at the WWTP input during a two-month period. It is interesting to note that both series have been successfully reconstructed with the current approach; recovered values, highlighted with circles, follow smoothly the tendency of the plant data series. The imputation system was applied to the complete dataset and all the missing values were recovered. The original dataset contained 1480 missing values that represent the 7.3% of the total data. The average upper-bound for the error of distance calculations defined before is 0.40, which is higher than the average quantization error of 0.13 corresponding to the SOM map used to perform the data imputation. Thus, the effect of missing data in distance calculations can be neglected. Once the data set was reconstructed the techniques proposed in subsection 4.2.1 were applied to select the best subset of variables to infer BOD at the effluent water.

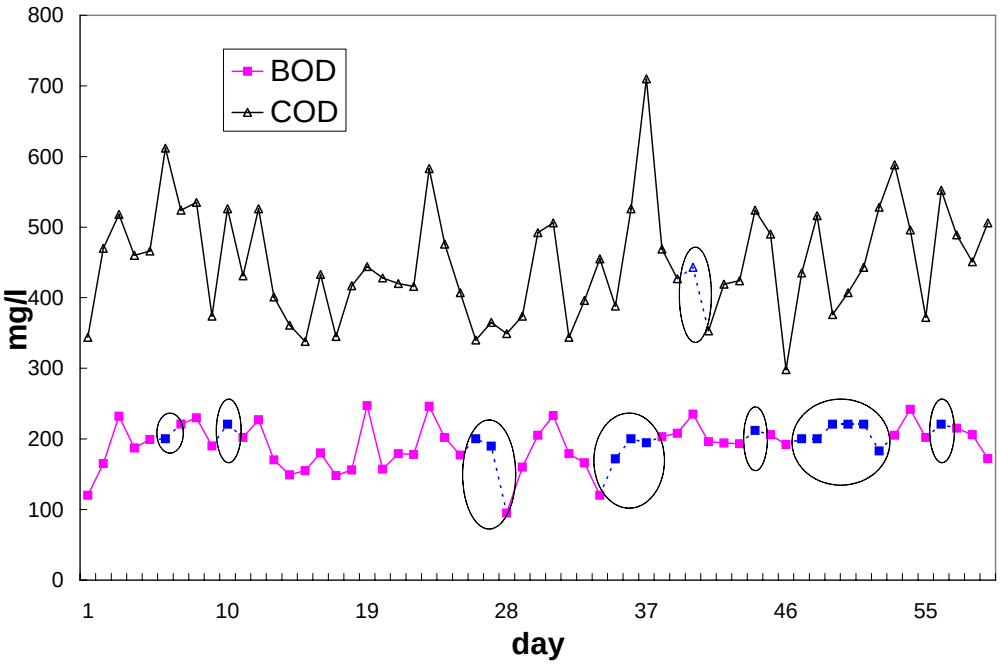


Figure 4.10. Imputed values of BOD and COD for a 60 days period at the input of the WWTP. Blue color and dashed lines represent imputed data and trends, respectively

Redundant variables were first identified. There were three redundant variables: input conductivity (COND-E), input conductivity to primary settler (COND-P), and input conductivity to secondary settler (COND-D). The last two were deleted and only COND-E was retained. Hence, the dimension was reduced to 35 input variables plus the target BOD.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The SOM dissimilarity feature selection method was applied to the remaining 35 variables. Figure 4.11 shows the clusters detected in the analysis of the WWTP data space. Each cluster is labeled with the names of the variables that belong to it. The initial candidate subset was built from these four clusters by selecting from each class the variable mostly correlated with the target (Biological Oxigen Demand at plant's output; DBO-S). A minimum correlation threshold was used to hinder the inclusion of irrelevant information in the initial data set. As a result only one variable was retained and included in the initial set (amount of suspended solids at plant's output; SS-S). It can be noted from the clustering in Figure 4.11 that the selected variable belongs both to the same SOM unit and Davies-Bouldin cluster as the target.

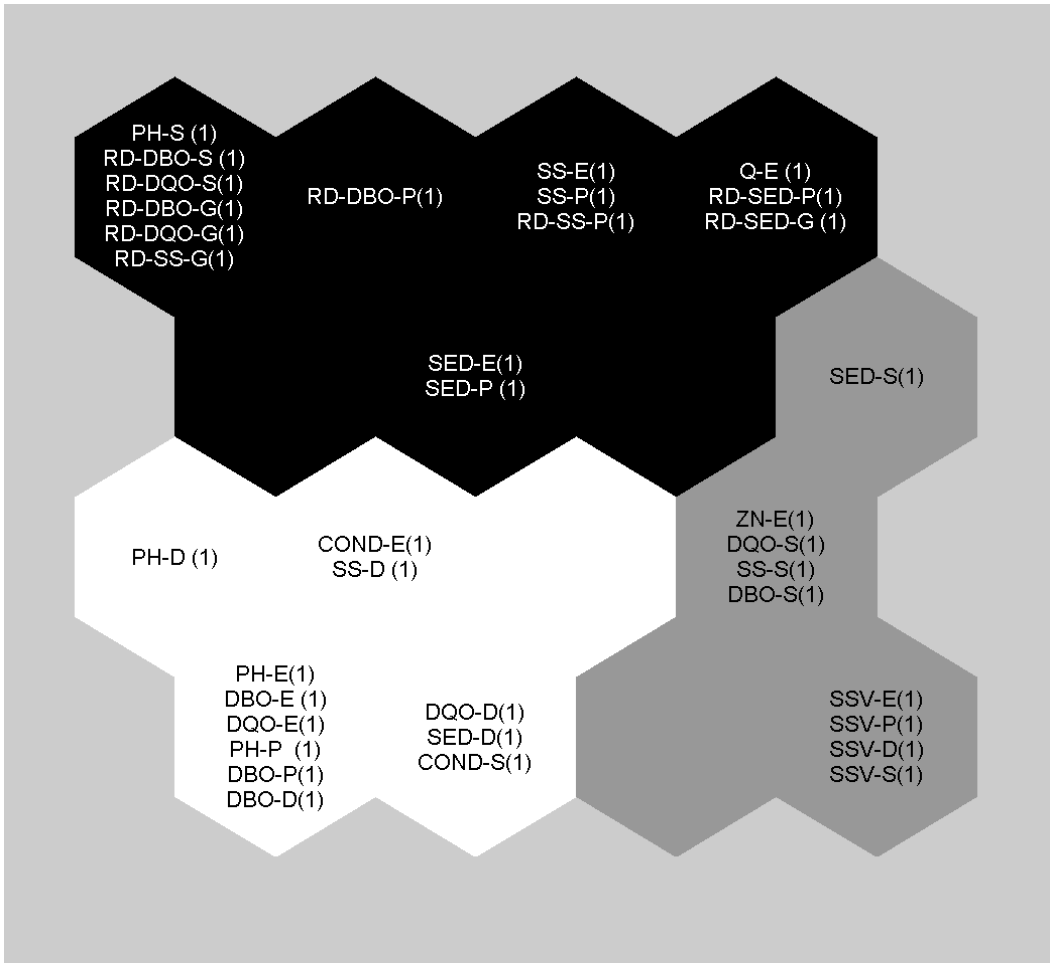


Figure 4.11. Clustering of the SOM that minimizes the Davies-Bouldin index. Each of the four clusters formed are labeled with the clustered variable names

Figure 4.12 depicts the average dissimilarity curve obtained for the WWTP data. The dissimilarity stabilizes after the inclusion of the additional 19th variable and then increases slightly indicating that further addition of variables does not improve the information contained in the best set (1+19 = 20 variables).

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

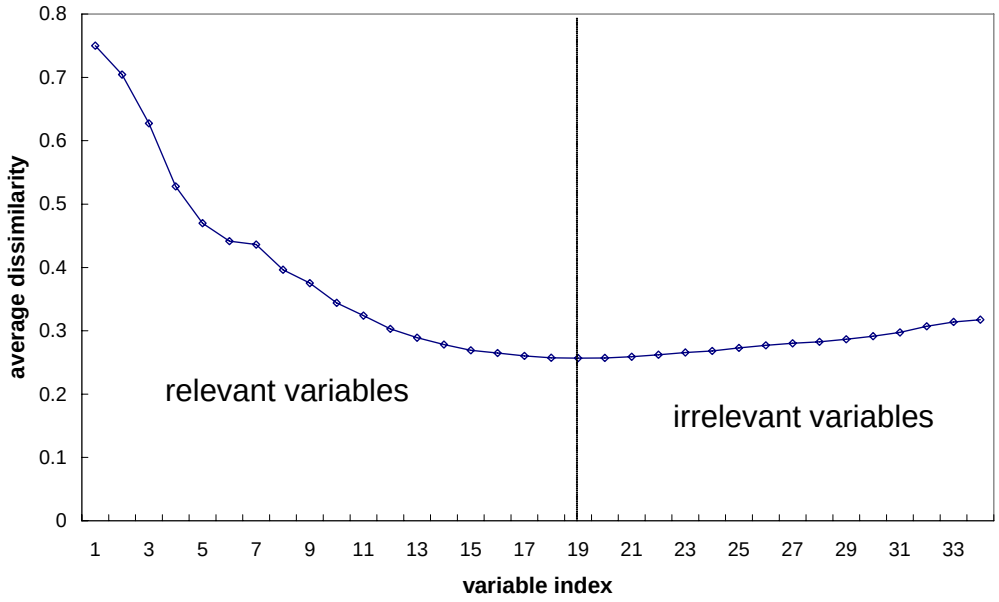


Figure 4.12. Average dissimilarity curve for the WWTP case. The initial set contains only one variable and 19 variables are subsequently added to form the best set

The best set of variables was formed by 20 variables, one selected as a relevant class representative (the amount of suspended solids at process output, SS-S) and nineteen from the dissimilarity measure analysis summarized in Figure 4.12. These nineteen variable are:

- **Calculated Efficiency Measures (8 vars):** Overall Plant: sedimentable solids (RD-SED-G), suspended solids (RD-SS-G), Biological oxygen demand (RD-DBO-G), and chemical oxygen demand (RD-DQO-G). Primary settler: Biological oxygen Demand (RD-DBO-P), and suspended solids (RD-SS-P). Secondary settler: Biological oxygen demand (RD-DBO-S), and Chemical oxygen demand (RD-DQO-S).
- **WWTP Input (2 vars):** DBO, and DQO
- **Primary Treatment (1 var):** DBO
- **Secondary Treatment (5 vars):** DQO, DBO, SS, PH, and SED
- **Output (3 vars):** DQO-S, SED-S, and PH-S

It is important to note that the SOM feature selection algorithm tends to select the calculated efficiency measures from the global process and process variables from the primary and secondary treatment units, since these variables contain information that explain the operation of the whole WWTP process. Also measures concerning the biological and chemical oxygen demand at the input and in each of the treatment units are selected.

A Radial basis Functions-based soft sensor was trained to infer the biological oxygen demand of plant effluents by using this reduced set of relevant variables and the selected training dataset. The soft sensor was trained using the first 506 input

patterns and the remaining 21 patterns were used for testing purposes. Figure 4.13 depicts the measured and predicted BOD values. It can be observed in this figure that predicted values follow accurately the measured ones. The absolute mean error for the predicted values is 1.17 (3.16%). The modeled soft sensor is capable of accurately predicting the BOD at the output of a wastewater treatment plant, despite the amount of missing data in the training set and the data reduction in the dimension of the input space from 37 to 20 variables.

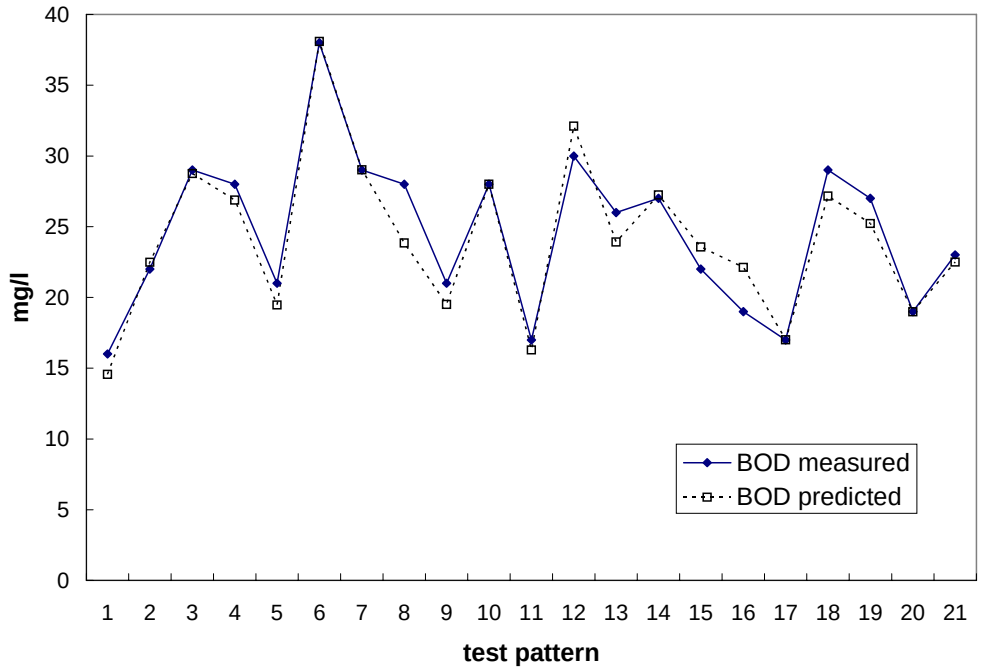


Figure 4.13. Predicted BOD values at WWTP output with Radial basis Functions by using the imputed dataset and the best features selected

4.3.3 Ensemble-based data imputation

The WWTP data set was also used to assess the suitability of SOM to perform multiple data imputations and to assess its efficiency when compared with mean-substitution and single imputation techniques. In this subsection only results corresponding to the WWTP data are presented. A more detailed discussion using different techniques and data sets is given in Rallo et al. (2005).

Two different approaches have been used to introduce diversity in the ensemble components. Diversity was first introduced in each of the single imputation models by changing the size of the maps. This leads to an ensemble where under-trained models, i.e., models with great generalization capabilities, coexist with others that were over-fitted, i.e., very accurate and adapted to certain regions of the training data. Second, the training set was manipulated using "bagging" techniques (Breiman, 1996), i.e., resampling techniques called "bootstrapping" to generate multiple versions of the training set. The procedure started by considering the training set TR formed by N patterns, each labeled with a probability 1/N. A new training set TR_{bag}

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

was created by sampling with replacement N times from the original training set, using these probabilities. Using this procedure some cases in TR may not appear in TR_{bag} while others may appear multiple times. The resampled training set is used to train an imputation model. The process is repeated several times and the results of each individual model are combined.

Single (SI) and multiple (MI) imputation models for this dataset were based only in the response of the *bmu*. The full dataset was reconstructed using mean substitution, SOM-based single imputation, and SOM-based multiple imputation with bagging. Using these data (without missing information) three virtual sensors for pH, chemical oxygen demand (COD) and biological oxygen demand (BOD) for the effluent water were implemented and trained. All three virtual sensors were based in a neural network trained using the backpropagation algorithm. The whole dataset (512 patterns) was separated into a training set (500 data vectors) and the remaining 21 data vectors were used for testing purposes. These input vectors were formed only by 22 variables corresponding to measurement points located in the pretreatment, primary settler and secondary settler. The best topology of the neural network was 22-13-1, i.e., 22 input variables, 13 hidden nodes and 1 output.

Table 4.2. Absolute Mean Error (AME) for the prediction of pH, COD and BOD at the effluent of the WWTP. Comparison of backpropagation predictive models resulting from different imputation techniques. The number of missing cases in the original dataset is shown for each target

	Imputation Model	AME
pH Missing: 1	Mean	0.029
	SI_{bmu}	0.032
	$MI^{bagging}_{bmu}$	0.031
	MI^{dim}_{bmu}	0.036
	MI^{hybrid}_{bmu}	**0.026**
COD Missing: 18	Mean	0.075
	SI_{bmu}	0.065
	$MI^{bagging}_{bmu}$	0.064
	MI^{dim}_{bmu}	0.060
	MI^{hybrid}_{bmu}	**0.058**
BOD Missing: 23	Mean	0.017
	SI_{bmu}	0.021
	$MI^{bagging}_{bmu}$	**0.015**
	MI^{dim}_{bmu}	0.016
	MI^{hybrid}_{bmu}	**0.015**

The results obtained are presented in Table 4.2. It can be seen that the predictions obtained from all the three virtual sensors have a similar value for their absolute mean errors. This could be due to the fact that the prediction model is not too sensitive to the quality of the imputations. Nevertheless, it is important to note that the prediction process is possible due to the prior imputation of data. The lowest errors are obtained for BOD, which is the target variable with most missing data in the original dataset (28). A new aggregation model (referred as MI^{hybrid} in the Table

4.2) for the multiple imputation system was used with this dataset. This model is based in the combination of multiple imputation models based in bagging with those based in maps of different sizes. This improved slightly the performance in terms of absolute mean errors.

4.4 Conclusions

This chapter has introduced the main components that form the second tier of the framework for data-driven and inferential modeling. Section 4.1 has presented an overview of the main procedures involved in data preprocessing. In section 4.2 the use of the SOM algorithm at the preprocessing level has been introduced. The SOM constitutes a valid alternative to perform most of the tasks needed at the preprocessing level. A new method for feature selection using dissimilarity measures has been developed and assessed. The SOM has been used to detect redundant information by developing a redundancy index which takes into account the correlation between variables and their representation over the map (c-planes and U-matrix). In addition, the use of SOM clustering capabilities for the optimal selection of training and test examples has been tested. The use of this approach ensures that data used for model training and testing belong to the same application domain. Finally the use of SOM as both a single and a multiple data imputation system has been exemplified.

To illustrate these methods, section 4.3 has applied the proposed methodology to the development of virtual sensors to infer quality indicators in the effluent waters of a Waste Water Treatment Plant. Data reconstructed using single imputation techniques have been evaluated by developing an inferential model for the biological oxygen demand of effluent water. The model has been developed using Radial basis Functions. Multiple imputations have also been assessed by developing SOM ensembles in which diversity has been introduced by topology changes and bagging. Results obtained in the two cases indicate that the SOM constitutes a valid approach to deal with data sets containing an important amount of missing information. Also the feature selection method proposed in 4.2.1 has been used to discard irrelevant information for the development of inferential models.

4.5 References

BELANCHE, LL., CORTES, U., SÁNCHEZ, M. A knowledge-based system for the diagnosis of waste-water treatment plant. *Proceedings of the 5th international conference of industrial and engineering applications of AI and Expert Systems IEA/AIE-92*. Ed. Springer-Verlag. Paderborn, Germany, June 1992.

BLAKE, C.L., MERZ, C.J. UCI Repository of machine learning databases, Irvine, CA: University of California, Department of Information and Computer Science. [http://www.ics.uci.edu/~mlearn/MLRepository.html], 1998.

BLUM, A., LANGLEY, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**:245-271, 1997.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

BREIMAN, L. Bagging Predictors. *Machine Learning*, **24**:123-140, 1996.

DAVIES, D.L., BOULDIN, D.W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**:224-227, 1979.

DEMPSTER, A., LAIRD, N., RUBIN, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1):1–38, 1977.

ESPINOSA, G., ARENAS, A., GIRALT, F. (2002). An integrated SOM-fuzzy ARTMAP system for the evaluation of toxicity. *J. Chem. Inf. Comput. Sci.* **42**:343-359, 2002.

FAYYAD, U.M., IRANI, K.B. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. *Machine Learning*, **8**:87-102, 1992.

FESSANT, F., MIDENET S. Self-Organising Map for Data Imputation and Correction in Surveys. *Neural Comput. and Applic.*, **10**:300-310, 2002.

GHISELLI, E. *Theory of Psychological Measurement*. McGraw-Hill, 1964.

GUYON, I. ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**:1157-1182, 2003.

HALL, M. A. Correlation-based Feature Selection for Machine Learning. Ph.D. diss. Dept. of Computer Science, Waikato University, 1998.

HAN, J., KAMBER, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Pub. , Academic Press, 2001.

HANSEN, L.K., SALAMON, P. Neural network ensembles, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **12**(10):993-1001, 1990.

HSU, CH., HUANG, H., SCHUSCHEL, D. The ANNIGMA-wrapper approach to fast feature selection for neural nets. *IEEE Transactions on Systems, Man, and Cybernetics Part B*, **32**:207-212, 2002.

JOHN, G. H., KOHAVI, R., PFLEGER, K. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, 121-129, 1994.

KASKI, S., LAGUS, K. Comparing Self-organizing Maps. *Proceedings of ICANN'96, International Conference on Artificial Neural Networks*, 809-814, 1997.

KIRA, K. RENDELL, L. A. A practical approach to feature selection. *Proceedings of the Ninth International Conference on Machine Learning*, 249-256, 1992.

KIRKPATRICK, S., GELATT, C.D., VECCHI, M.P. Optimization by Simulated Annealing. *Science*, **220**(4598): 671-680, 1983.

KOHAVI, J. PFLEGER, K. Irrelevant features and the subset selection problem. *Proceedings of the 11th International Conference on Machine Learning*, 121-129, 1994.

KOHONEN, T. The Self-Organizing Map. *Proc. IEEE*. **78**:1443-1464, 1990.

KONONENKO, I., SIMEC, E. Induction of decision trees using RELIEFF. In: Della Riccia, G., Kruse, R., Viertl, R., (eds.): *Mathematical and statistical methods in artificial intelligence*, Springer Verlag, 1995.

LITTLE, R.J.A., RUBIN, D.B. *Statistical Analysis with Missing Data.* New York: J. Wiley & Sons, 1987.

MITCHELL, T.M. *Machine Learning*, McGraw-Hill, 1997.

MOODY, J., DARKEN, C. J. Fast learning in networks of locally-tuned processing units. *Neural computation*, **1**:281-294, 1989.

PEARL, J. *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Addison-Wesley, 1984.

PIRAMUTHU, S. Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research*, **56**:483-494, 2004.

QUILAN, J. R. *C4.5. Programs for Machine Learning.* Morgan Kaufmann, 1993.

RALLO, R., ESPINOSA, G., GIRALT, F. Using an ensemble of neural based QSARs for the prediction of toxicological properties of chemical contaminants, *Trans. IChemE Part B. Process Safety and Environmental Protection*, **83**(B4):387-392, 2005.

RALLO, R., FERRE-GINÉ, J., ARENAS, A., GIRALT, F. Neural Virtual Sensor for the inferential prediction of product quality from process variables. *Computers & Chemical Engineering*, **26**:1735-1754, 2002.

RALLO, R., FERRÉ-GINÉ, J., GIRALT, F. Best Feature Selection and Data Completion for the Design of Soft Neural Sensors. *Proceedings of AIChE 2003, 2nd Topical Conference on Sensors,* San Francisco, 2003.

RUBIN, D.B. *Multiple Imputation for Non response in Surveys.* New York: J. Wiley & Sons, 1987.

RUSSELL, S., NORVIG, P. *Artificial Intelligence: A Modern Approach.* Prentice-Hall, 1995.

SAMAD, T., HARP, S. Self organisation with partial data. *Network*, **3**:205–212, 1992.

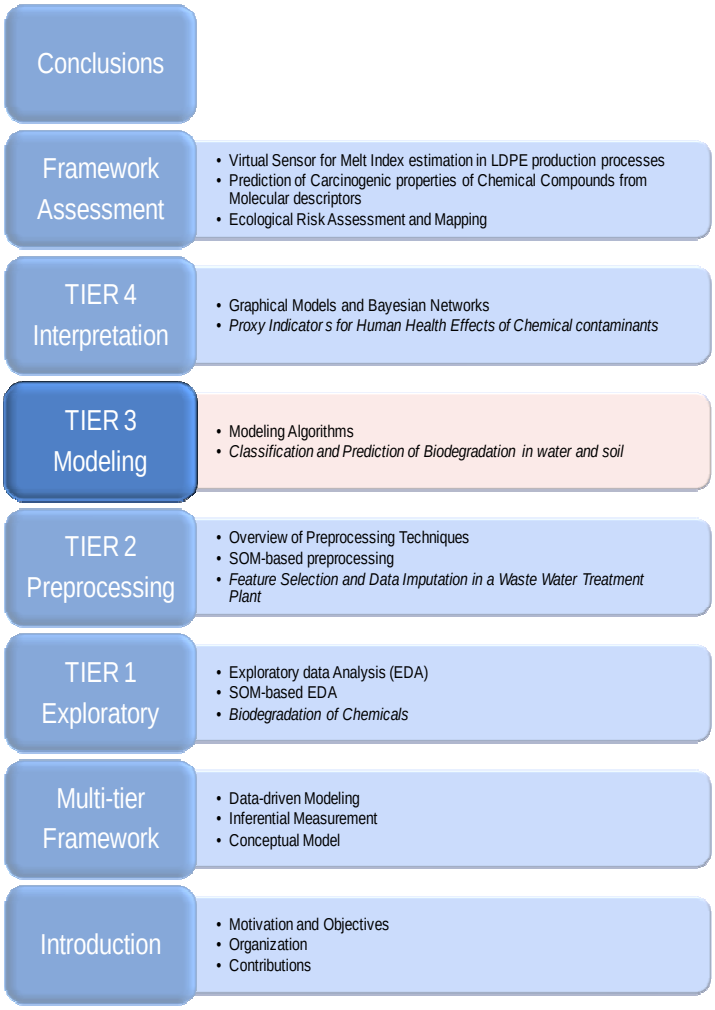
SCHEFFER, J. Dealing with Missing Data. *Res. Lett. Inf. Math. Sci.*, **3**:153-160, 2002.

TODESCHINI, R. Data correlation, number of significant principal components and shape of molecules. The K correlation index. *Analytica Chimica Acta*, **348**:419-430, 1997.

WANG, S. Application of Self-organising maps for data mining with incomplete data sets. *Neural Comput. and Applic.*, **12**:42-48, 2003.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 5

Tier 3: Modeling

The core components of the proposed framework are modeling algorithms. These are used to build the internal representation of the application domain. Chapter 5 presents an overview of different machine learning methodologies that have been used through the current study to develop data-driven and inferential models. The development of a model to classify and predict the biodegradation of chemicals in different media is used to illustrate the use of the framework at this tier 3.

5.1 Models

Classification and prediction are two different terms that in most modeling situations refer to the same learning task. There is agreement that this dual nomenclature is a matter of definition. In fact the same set of Machine Learning algorithms can be used to perform both, classification and prediction tasks. According to Han and Kamber (2001), predicting class labels is classification, and predicting values, e.g., using regression techniques, is prediction. In general, the former term refers to building models to classify class labels which can be either, discrete o nominal. Prediction, in contrast, refers to the application of the model to obtain class labels for unknown data to. Some authors use the term estimation to refer to numerical prediction. In the present study the term classifier will be used to refer to algorithms that build models in which the target is a discrete class label, while predictive models will be applied to estimate continuous numerical data. Algorithms to classify or to predict target values will be referred to as "modeling algorithms".

The task of classification occurs in a wide range of human activities. In a broad sense the term could cover any context in which some decision or forecast is made on the basis of currently available information. Hence, a classification procedure is some formal method for repeatedly making such judgments in new situations. The construction of a classification procedure from a set of data for which the true

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

classes are known has also been termed pattern recognition, discrimination, or supervised learning to distinguish it from unsupervised learning or clustering in which the class structures are inferred solely from the data. Contexts in which a classification task is fundamental include, for example, mechanical procedures for sorting letters on the basis of machine-read postcodes, assigning individuals to credit status on the basis of financial and other personal information, and the preliminary diagnosis of a patient's disease in order to select immediate treatment while awaiting definitive test results. In fact, some of the most urgent problems arising in science, industry and commerce can be regarded as decision problems by means of the classification of complex and often very extensive data.

A wide variety of approaches can be used to build classifiers. Three main historical threads of research can be identified: statistical, machine learning and neural networks. Their common goal is to attempt to derive procedures that could be able to equal, if not exceed, a human decision-maker behavior. These three approaches have the advantage of consistency and, to a variable extent, explicitness when handling a wide variety of problems. They have proven to be extremely general when enough data is available, as is the case in practical settings.

- **Statistical approaches.** Two main phases of progress on statistical classification can be identified. The first "classical" phase originates from the early work on linear discrimination carried out by Fisher (1936). The second "modern" phase exploits more flexible types of models conceived and designed to estimate the joint probability distribution of the features within each class, which can in turn provide a classification rule (McLachlan, 2004). Statistical approaches have an explicitly underlying probability model to estimate the probability of belonging to each class. In addition, it is usually assumed that these techniques will be used by statisticians and, hence, some human intervention is assumed with regard to variable selection and transformation, and to the overall structuring of the problem.

- **Machine learning.** Machine Learning generally encompasses automatic computing procedures based on logical or binary operations designed to learn a task from a series of examples. Attention has focused on decision-tree approaches where classification stems from a sequence of logical steps, which are capable of representing very complex problems if sufficient data is provided (Breiman et al., 1984). Other techniques, such as genetic algorithms (Holland, 1975) and inductive logic procedures (Quinlan, 1990) are currently under active development. They should allow working with more general types of data, including cases where the number and type of attributes vary, and where additional layers of learning are superimposed with a hierarchical structure of attributes and classes, and so on. Machine Learning aims to generate classifying expressions simple enough to be understood easily by humans. They must mimic human reasoning sufficiently well to provide insight into the decision process. Like statistical approaches, background knowledge may be exploited in development, but operation is assumed without human intervention.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

- **Neural networks.** The field of Neural Networks (Minsky and Papert, 1969; Rosenblatt, 1962; Rumelhart and McClelland, 1986) has arisen from diverse sources, ranging from the fascination of mankind with understanding and emulating the human brain parallel performance, to broader issues of copying human abilities such as speech and the use of language, to the practical commercial, scientific, and engineering disciplines of pattern recognition, modeling, and prediction.

The pursuit of technology development in computational sciences is a strong driving force and challenge for researchers, both in academia and industry, in many fields of science and engineering. Both in neural networks research as in Machine Learning the excitement of technological progress is supplemented by the challenge of producing abiotic intelligence itself. Neural networks generally consist of layers of interconnected nodes, each node producing a non-linear function from its input. The input to a node may come from other nodes or directly from an input data layer. Also, some nodes can be identified with the output layer of the network. Therefore, the complete network represents a very complex set of interdependencies which may incorporate any degree of nonlinearity, allowing very general functions to be modeled. In the simplest networks, the output from one node is fed into another node in such a way as to propagate "messages" through layers of interconnecting nodes. More complex behavior can be modeled by networks where the output nodes are connected with previous nodes, resulting in a highly nonlinear system with feedback. It has been argued that neural networks mirror the behavior of networks of neurons in the brain to a certain extent. Neural network approaches combine the complexity of some of the statistical techniques with the machine learning objective of imitating human intelligence. However, this attained at a more "unconscious" level and hence there is no accompanying ability to make learned concepts transparent to the user.

There are three essential components involved in a classification problem:

- The relative frequency by which classes occur in the population of interest, expressed formally as the prior probability distribution.
- An implicit or explicit criterion for separating the classes: we may think of an underlying input/output relationship that uses observed attributes to distinguish a random individual from each class.
- The cost associated with making a wrong classification.

5.1.1 Modeling algorithms

In this subsection the learning algorithms used in the current study for model development are briefly described and references for their implementation are given. Algorithms are grouped as stemming from machine learning, neural or statistical learning theory approaches.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

5.1.1.1 Machine Learning Approaches

Machine learning treats modeling tasks from a search-like perspective. In this context, model development consists in the search, through a space of possible hypothesis, of the hypothesis that best fits available training data. A summary overview of features corresponding to main families of Machine Learning algorithms is given below.

Instance-based Learning. Instance-based learning methods, such as nearest neighbor and locally weighted regression, are conceptually straightforward approaches to approximate real or discrete valued target functions. Learning in these algorithms simply consists in storing the presented training data. When a new query instance is encountered, a set of similar related instances are retrieved and used to classify the presented input. Instance-based approaches construct a different approximation to the target function for each distinct query instance. This has significant advantages when the target function is very complex but can still be described by a finite collection of less complex local approximations. One disadvantage of these methods is that the cost of classifying new instances can be high since all computations take place at classification time rather than when the training examples are first presented. There exist several methods that use this approach. The instance based learning IBK (Aha & Kibler, 1991) uses a K-nearest neighbor approach in which the training set is incrementally processed and missing values are ignored. In IBK is possible to define the number of nearest neighbors used to classify data. However, the widening in the number of used instances may result in memory intensive implementations when large neighborhoods are considered. The K-star algorithm is a variation that uses entropy as distance measure in a K-nearest neighbor transformation (Clearly and Trigg, 1995). As a consequence, it performs well in managing data sets that contain either, missing values, real valued attributes or symbolic data.

Decision Trees. A popular machine learning method is the decision tree, also known as recursive partitioning (Breiman et al., 1984; Quinlan, 1993). Decision trees classify instances by sorting them in a tree structure where data are classified at the leaf nodes. Each node in the tree specifies a test of some attribute of the instance and every branch descending from a node corresponds to subsets of the possible values of the attribute. Decision trees represent a supervised approach to classification in which non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes (Bauer et al., 1999). Practical issues of the learning process in decision trees include the determination of how deeply has to grow the decision tree, the discretization of continuous attributes, choosing an appropriate attribute selection measure, management of data with missing values, and the improvement of computational efficiency. Most algorithms used to implement decision trees use variations of a core algorithm based on a top-down greedy search through the space of all possible decision trees. This approach was first used in the ID3 algorithm (Quinlan, 1986) and its C4.5 decision tree successor (Quinlan, 1993). Variations of these basic tree algorithms lead to diverse types of tree implementations, for instance:

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

- *Logistic Model Trees* (LMT) construct a tree-structured classifier with logistic regression functions at the leaves. The classic logistic regression approach models log(p/(1−p) as a linear function of the features, where p represents the probability of a feature vector x belonging to class i. LMT approaches use the divide and conquer principle, i.e., a complex set of data is divided into a sufficiently large number of subsets such that a simple linear logistic regression model can adequately fit the data in each subset.
- *Model trees* are a technique for dealing with continuous class problems that provide a structural representation of the data. Model trees have a conventional decision tree structure but use linear functions at the leaves instead of discrete class labels. Like conventional decision tree learners, M5 builds a tree by splitting the data based on the values of predictive attributes. Instead of selecting attributes by an information theoretical metric, the M5 chooses attributes that minimize intra-subset variation in the class values of instances that go down each branch. After constructing a tree, the M5 algorithm computes a linear model for each node; the tree is then pruned back from the leaves, so long as the expected estimated error decreases. The expected error for each node is calculated by averaging the absolute difference between the predicted value and the actual class value of each training example that reaches the node. To compensate for an optimistic expected error from the training data, this average is multiplied by a factor that takes into account the number of training examples that reach the node and the number of parameters in the model that represent the class value at that node. The M5 Rules algorithm is a method for generating rules from these model trees. To generate rules a tree is applied to the full training data set and a pruned tree is learned. Next, the best leaf (according to some heuristic) is made into a rule and the tree is discarded. All instances covered by the rule are removed from the data set. The process is applied recursively to the remaining instances and terminates when all instances are covered by one or more rules.
- *Random Forest* was developed by Breiman (2001) as an ensemble method that combines several individual classification trees. Prediction is made by aggregating (for example by majority voting) the predictions given by the whole ensemble. Two types of randomness induction techniques, bootstrap sampling and random selection of input variables, are used in the algorithm to make sure that the classification trees grown in the forest are dissimilar and uncorrelated from each other. Growing a forest of trees and using randomness in building each classification tree in the forest leads to better predictions compared to a single classification tree and makes the algorithm more robust to noise present in data. Random Forests, as most classifiers, can suffer from the curse of learning from an extremely imbalanced training data set. Since it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class.

ISBN:978-84-691-1008-9/DL: T.2230-2007

5.1.1.2 Neural Network Approaches

Artificial Neural Networks (ANN) constitutes an effective tool to build nonlinear models from data. An ANN is formed by a set of adaptable nodes which, through a process of learning from examples, store experimental knowledge to make it available when needed. Neural networks learn by using supervised (both input and output information are used for training) or unsupervised (only input information is available for training) learning. For example, backpropagation and probabilistic neural networks use supervised learning, while most of the neural network classifiers, such as Self-Organizing Maps, Learning Vector Quantization, or Adaptive Resonance Theory, use unsupervised learning. The flexibility and general applicability of neural systems have been demonstrated by diverse applications across many fields in science and engineering. Let us briefly review some of the neural models used in this study.

Feed-forward Networks. The most commonly used ANN model is a multilayer feed-forward network architecture known as backpropagation neural network. This network gets its name from the technique used for its training; propagation of errors through the network layers using gradient descent or conjugate gradient algorithms (Masters, 1993). With a conventional mean square error, the backpropagation weight update rule can be defined as,

$$\Delta w_{ij}(n+1) = \eta o_i \delta_j + \alpha \Delta w_{ij}(n) \qquad (5.1)$$

where Δw_{ij} is the change in the weight connecting unit i to unit j, η is the learning rate, o_i is the output of the unit i, α is the momentum term, n indicates the epoch of the pattern presentation sequence, and δ_j is the error associated with unit j and is calculated as,

$$\delta_j = \begin{cases} o'_j(t_j - o_j) \text{ if unit } j \text{ is an output unit} \\ o'_j \sum_k w_{kj}\delta_k \text{ if unit } j \text{ is a hidden unit} \end{cases} \qquad (5.2)$$

where o'_j represents the derivative of the output and t_j is the target output of unit j.

A data preprocessing step is necessary in these ANNs before starting the training process. The most common technique is known as min/max interval, and consist in the linear transforms of the [min max] ranges of input/output data values in the training set into values normalized in the range [0,1] or [-1,1]. Internal validation techniques such as cross-validation should be carried out to properly estimate the robustness and generalization ability of the model.

Radial Basis Function Networks. Radial basis functions, commonly known as RBF, consist of a three layer feed forward architecture with input, hidden and output layers (Yingwei et al., 1997; Schwenker et al., 2001; Rivas et al., 2004). Their main characteristic is that the transfer functions of the hidden layer nodes are radial basis functions. These functions are typically implemented using kernel functions (usually Gaussians), which operate over a localized area of the input space. The effective

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

range of these kernels is determined by the values of its center and width. The output of each hidden neuron, $h_i(x)$, is computed using its kernel function by,

$$h_i(x) = \phi\left(\|x - c_i\|^2 / r_i^2\right) \tag{5.3}$$

where ϕ is the kernel of the RBF, c_i is the center of the *ith* hidden neuron, r_i is its radius, and $\|\bullet\|^2$ is the Euclidean distance. The response of the output units, o_j, is computed by the expression

$$o_j(x) = \sum_{i=0}^{n-1} w_{ij} h_i(x) + w_{oj} \tag{5.4}$$

where w represents a weight matrix and w_{oj} is a bias term for the *j-th* output neuron and n is the number of hidden neurons.

Fuzzy ARTMAP. Fuzzy ARTMAP (Carpenter et al. 1992) is a cognitive artificial neural network based on Adaptive Resonance Theory (Carpenter et al., 1991). The basic learning rule of fuzzy ARTMAP (FAM) consist in the creation of a new category when an unfamiliar input event is encountered during the process of updating the connection weights of an old category when a similar input to that category is presented to the network. The level of similarity is determined by a vigilance parameter (ρ).

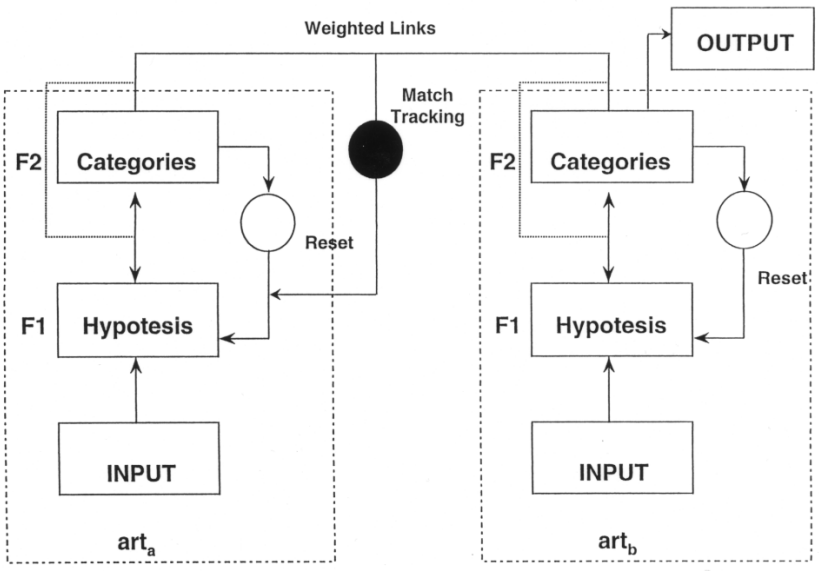


Figure 5.1. Block diagram of the fuzzy ARTMAP architecture

Figure 5.1 illustrates the fuzzy ARTMAP architecture. Each fuzzy ARTMAP system includes a pair of ART modules (art_a and art_b) that create stable recognition

categories in response to arbitrary sequences of input patterns. During supervised learning, art_a receives a stream of input patterns, e.g., molecular descriptors, and art_b also receives a stream of output patterns, e.g., biodegradation rates. An associative leaning network and an internal controller linking both art modules ensure autonomous system operation in real time. The controller is designed to create the minimal number of art_a categories needed to meet a predefined accuracy criterion. The vigilance parameter, ρ_a, calibrates the minimum confidence that art_a must have in a recognition category. It acts as an automatically controlled process of hypothesis testing. Lower values of ρ_a enable larger categories to form and lead to broader generalization and a higher degree of code compression. A predictive failure of art_b recognition increases ρ_a by the minimum amount needed to trigger hypothesis testing at art_a by using a mechanism called *match tracking*. Match tracking causes the minimum reduction in generalization that is necessary to correct a predictive error. Hypothesis testing leads to the selection of a new art_a category, which focuses attention on a new cluster of input features that is more able to predict the corresponding output. Match tracking facilitates the learning of a different prediction for a rare event instead of producing the same output as other events of its class (Anagnostopoulos et al., 2002).

To formalize the Fuzzy ARTMAP learning algorithm it is convenient to consider the elements of the set of n-dimensional data vectors $\{\xi^1,..., \xi^p\}$, where p is the number of vectors to be classified, as the pattern of values showing the extent to which each feature is present in a native data set. Every pattern must be normalized to satisfy the following conditions:

$$\begin{aligned} \xi^i &\in [0,1]^n \\ \sum_{j=1}^{n} \xi_j^i &= k \qquad \forall \, \mathrm{i} = 1,..., \mathrm{p} \end{aligned} \tag{5.5}$$

Classification in fuzzy ART takes place according to *Fuzzy Set Theory* (Zadeh, 1965). The similarity between two vectors can be established by the grade of the membership function, which for two sets (l, m) of generic vectors can be easily calculated as:

$$grade\,(\xi^1 \subset \xi^m) = \frac{\left|\xi^l \wedge \xi^m\right|}{\left|\xi^l\right|} \tag{5.6}$$

The fuzzy AND operator $\wedge$ in Eq. (5.6) is defined by,

$$\begin{aligned} \wedge : [0,1]^n \times [0,1]^n &\rightarrow [0,1]^n \\ (\xi^l, \xi^m) &\rightarrow \xi^i \end{aligned} \tag{5.7}$$

The components of the image vector that results from the application defined by Eq. (5.7) are:

$$\xi_j^i = \min\left(\xi_j^l, \xi_j^m\right) \quad \forall \mathrm{j} = 1,...,\mathrm{n} \tag{5.8}$$

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The norm |•| in Eq. (5.6) is the sum of the components of the vector given by Eq. (5.8).

The classification algorithm clusters the data with a value for the grade of membership in Eq. (5.6) greater than the *vigilance parameter* ρ into groups or classes. The value of ρ controls the granularity of the classes and assures the fulfillment of the accuracy criterion in the classification procedure. A weight vector ω^{μ} represents each class μ. Classification starts by creating the first class from the first pattern presented to the network,

$$\omega^1 = \xi^1 \tag{5.9}$$

The rest of input patterns ξ^i (i=2,...,p) are presented to the network and if the similarity of ξ^i with any established class μ is greater than ρ then ξ^i is classified into this class, and the representative of this class is updated according to,

$$\omega^{\mu}_{new} = \omega^{\mu}_{old} \wedge \xi^i \tag{5.10}$$

Otherwise a new class represented by ξ^i is created. Eq. (5.10) is the learning rule of the network. The mechanisms to speed up the process and to conduct the classification properly can be found elsewhere (Carpenter et al., 1991).

The dynamics of Fuzzy ARTMAP is essentially the same as two separate Fuzzy ART networks, each one working with a part of the training vector. The first part could be interpreted as the input pattern and the second one as the desired classification output (supervisor). The associative memory records the link between the classes corresponding to the input pattern and the desired classification. The internal controller supervises if a new link is in contradiction with any other one previously recorded. If no contradiction is found, the link is recorded; otherwise the pattern is re-classified with a larger vigilance parameter. Once the network has been trained it can be used to classify input vectors without any additional information.

The output of Fuzzy ARTMAP is restricted to a single class label. Intersecting distributions are common in real problems due to classification uncertainty and/or use of sub-optimal sets of features (Blume at al., 1996). When clouds of data points associated with several classes overlap, Fuzzy ARTMAP generates too many clusters (some times as many as data points); this is known as category proliferation. It has been shown by Carpenter (1991) that a pattern located inside the intersection of some categories will choose the category with smallest size. As a result, while the classification for training vectors can work quite well for such a situation, performance on test data is often worse than if fewer clusters had formed. Finally, the number and shape of categories created during the training phase is a function of network parameter values (ρ, α, β), as well as of the order in which training patterns are presented to the network. FAM classifiers exhibit stable learning under off-line training using the fast learning rule. However, achieving 100% correct classification performance on the training set implies over-fitting and hindering generalization for the test set. The choice parameter α largely affects the formation of decision regions. While small values of α favor categories of larger size, in the conservative limit

$\alpha \to 0$, patterns tend to choose the category of smallest size. Thus, in the conservative limit it is possible to extract from a trained FAM classifier some meaningful IF-THEN rules that partially describe its classification decisions (Anagnostopoulos et al., 2002).

The Fuzzy ARTMAP architecture, which has been successfully applied to educe the different classes of large-scale events present in free turbulence (Ferre-Gine et al., 1996), was designed to classify data and, thus, cannot generate an output pattern after the training stage. To implement predictive capabilities the categories identified by the system from the learned information can be linked to the desired outputs, as depicted in Figure. 3. This is mathematically equivalent to defining an application from the space of categories to that of output patterns, the image of the application being defined by examples of patterns provided to the neural system in a supervised manner. The accuracy of the procedure increases asymptotically towards a constant value with the number of examples used for training, i.e., when the space of outputs is accurately mapped. In the predictive mode, only the category layer of ART_b in Figure 5.1 is active and linked to ART_a to provide an output for each input vector presented to this module (Giralt et al., 2000).

Dynamic Unsupervised Layer. The success of predictive fuzzy ARTMAP in difficult forecasting problems (Giralt et al., 2000), together with its limitations when interpreting information with underlying periodicity and limited training (Carpenter et al., 1992), has motivated the current search for other potentially suitable systems for sensor development, such as dynamic unsupervised RBFs layers. The challenges related with the application of this neural system are the design of a node generation mechanism and of a clustering process that are simple enough and simultaneously compatible with different supervised output generation procedures.

The current approach to construct the unsupervised RBF layer is performed in five steps: (i) Set-up an initial configuration with the center of the first cluster formed by one pattern chosen randomly from the training dataset; (ii) determine the minimal mean distance between patterns; (iii) use this distance as the constant maximum attention radius d_{max} to control the generation of new nodes over the node-generation process; (iv) present a new input pattern, ξ, to the network and compute the Euclidean distance between this pattern and all nodes; and (v) adapt the structure of the network according to the following two rules. If the input pattern is located inside the region of influence of any node *i*, the pattern is classified in *i* and its center adapted using a winner takes-all approach based on Kohonen's learning rule,

$$c_i(n+1) = c_i(n) + \alpha(n) \cdot \left[\xi - c_i(n)\right] \qquad (5.11)$$

with *n* denoting the training epoch, $c_i(n)$ the center vector of the selected node, and $\alpha(n)$ a monotonically decreasing learning rate computed as $\alpha(n) = \alpha_0 \cdot 1/\sqrt{n}$, which controls the adaptation or upgrading of the center of the cluster. Otherwise, if the input pattern is located outside the region of influence of all the nodes, a new node is created with the center located at the point that defines the input pattern. The procedure is repeated until the number of nodes stabilizes and either the classification or the number of iterations reaches a predetermined minimum or

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

maximum value, respectively. A similar approach using K-means was applied by Hwang et al. (1997) to speed up the training of RBF layers.

The current algorithm tends to create an appropriate number of clusters since it determines the attention radius based on the distribution of the training patterns. This yields the minimal clustering necessary to achieve a good classification in accordance with the attention radius chosen. The process of complying with a given classification accuracy has to be tested for generalization by trial and error, which is a customary practice when working with neural systems.

Once the classifier is built it is necessary to produce an output from the unsupervised layer. This procedure could consist in a hybrid approach that combines the current dynamic unsupervised classifier with a supervised learning engine. In this work two techniques for producing this output have been adopted. The first is a *clustering average* based on the labeling of the unsupervised layer using the values of the target variable. One of the most common labeling processes consists in averaging the target value for each of the training patterns belonging to a given cluster, like in the k-means algorithm. This averaged value subsequently becomes the output of the network. This labeling algorithm can be summarized as follows: (i) Obtain a pattern from the training set; (ii) compute the winner node; (iii) add its value to the output value of the winner node; (iv) increase the pattern counter for this node; (v) repeat (i) until all patterns have been processed; (vi) compute the average output value for each node. Once the dynamic unsupervised layer is labeled the network is ready to infer the target property values, i.e., act as an inferential measurement system.

The second technique used in the current study to obtain an output from the unsupervised layer consist in the placement of Radial Basis Functions (RBF) over the cluster centers with supervised training to adjust the output. This neural network is hereinafter identified as *Dynamic Radial Basis Function network* (DYNARBF; Rallo et al., 2002). RBF neural networks facilitate the parameterization of any function *f(x)* as a linear combination of non-linear basis functions (Powell, 1987,1992; Broomhead and Lowe, 1988; Lee and Kil, 1988; Moody and Darken, 1989; Poggio and Girosi, 1990a, b; Musavi et al., 1992),

$$f(x) = p + \sum q_j G\left(\left\| x - x_j \right\|\right) \tag{5.12}$$

where j is the function index, the norm $|\bullet|$ is the Euclidean distance (Park and Sandberg, 1991), x_j are the centers of the proposed basis functions, p and q are adjustable parameters and G is a radial kernel function. In the current model a Gaussian activation function is used,

$$G(r_j) = \exp(-r_j^2 / 2\sigma_j) \tag{5.13}$$

where r_j is the Euclidean distance to the center of the *j-class* and σ_j is the dispersion of the Gaussian. For each activation function in Eq. (5.13), the center position x_j and its width σ_j must be determined to define a receptive field around the node. Both can be determined using an unsupervised learning process, in which data are clustered

into "classes" or nodes. The idea is to pave the input space (or the part of it where the input vectors lie) with the receptive field of these nodes (Gaussians in this case).

A map between the RBFs outputs and the desired process outputs is then constructed in a second supervised training stage. It should be noted that the RBF neural network needs some a-priori hypothesis concerning the number of nodes that will be used for any particular problem. This is a major drawback in the application of RBF's because the approximation error is highly dependent on the number of nodes. Usually, more nodes imply more accuracy in the mapping of the predicted target values over the training dataset. Nevertheless, "over-fitting" during training could in some cases imply a loss of network generalization capabilities during testing. There is no a priori methodology to estimate rigorously the number of nodes for optimal generalization. These issues have been the subjects of research in the past (Platt, 1991; Fritzke, 1994; Berthold, 1995). These studies share the common strategy of extending on-line the structure of the neural network to reduce a given measure of the classification error. In this study the Dynamic Unsupervised Layer algorithm explained above is used to overcome this drawback.

The neural system is trained in two separate phases. First, the unsupervised layer that defines the number of radial functions in the hidden layer as well as their position in input space is constructed. The width of each radial function σ_j is usually calculated *ad hoc* as the mean Euclidean distance between the k-nearest neighbors of node *j*,

$$\sigma_j = \sigma_0 \cdot \sum_{i=1}^{K} d(x_j, x_i) \tag{5.14}$$

The center of the i-Gaussian is x_i, σ_0 is a constant for width scaling, K is the cardinality of the neighborhood and d the Euclidean distance.

The activation of the functions once placed over the hidden layer is accomplished by

$$\begin{cases} f_j(\xi) = \exp\left(\dfrac{-d^2(\xi, x_j)}{2 \cdot \sigma_j^2}\right) \\ A_j(\xi) = \dfrac{f_j(\xi)}{\sum_{k=1}^{N} f_k(\xi)} \end{cases} \tag{5.15}$$

where ξ is the pattern presented to the network, A_j the activation of node *j* and N the total number of Gaussians.

In a second supervised learning stage the activation of the RBF layer for a given input pattern ξ is related to the desired output $\tilde{\theta}$. First, the output of the network is calculated as,

$$\theta_i = \sum_{j=1}^{N} w_{ij} \cdot A_j(\xi) \tag{5.16}$$

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The weights are updated afterwards using the common Delta rule minimization error procedure,

$$\Delta w_{ij} = \beta \cdot \left(\theta_i - \tilde{\theta}_i\right) \cdot A_j(\xi) \qquad (5.17)$$

Here, the activation of node *j* is the value of the Gaussian function when the input *i* is presented with a constant learning rate β.

The combination of Equations (5.15) and (5.16), together with the information contained at the centers (x_i) and in the widths (σ_i) of the RBFs, and with the weights connecting the activation of each radial function with the output node (w_i), yields an analytical model relating the target property with the parameters of the neural network:

$$P = \sum_{i=1}^{N} w_i \cdot \frac{\exp\left(-d^2(\xi, x_i)/2\sigma_i^2\right)}{\sum_{j=1}^{N} \exp\left(-d^2\left(\xi, x_j\right)/2\sigma_j^2\right)} \qquad (5.18)$$

In this equation N is the number of RBF nodes, d the Euclidean distance and ξ is the input vector of the data under investigation presented to the network.

5.1.1.3 Statistical Learning Theory Approaches

One of the major drawbacks of ANNs, which is shared by all types of *black-box* models, is that the resultant model and its parameters are difficult to interpret from the point of view of the data used to build the experimental model.

Support Vector Machines (SVM) appeared as an alternative formulation to the computational learning problem discussed above. The SVM algorithm stems from the statistical learning theory formalism (Vapnik, 1998) and some of its properties are very convenient: (i) Good generalization ability; (ii) robustness and sparseness of the solution; and (iii) automatic control of model complexity. In addition, SVM provides an explicit knowledge of data points, which is useful to define classifiers or regression functions. This feature facilitates the interpretation of the SVM-based model in terms of the data in the training set.

The general problem of regression in the framework of statistical learning theory can be stated as follows. Let us consider a set of measurements (or training data), $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{p}$, where $\mathbf{x}_i \in \Re^n$ is a vector of the input data and $y_i \in \Re$, represents the corresponding target. The goal of the regression analysis is to determine a function $f(\mathbf{x})$ that predicts accurately the target outputs, $\{y\}$, for a never seen before set of input-output data, $\{(\mathbf{x}, y)\}$, drawn from the same underlying joint probability distribution, $P(\mathbf{x}, y)$ as that of the training set, D. In essence, the task is to find a function, f, that minimizes the expected risk, $R[f]$, which is defined as,

ISBN:978-84-691-1008-9/DL: T.2230-2007

$$R[f] = \int L(f(\mathbf{x}) - y)dP(\mathbf{x}, y) \tag{5.19}$$

In this equation L denotes a loss function. For a given function, $f(\mathbf{x})$, the expected risk (or test error) is the average error associated with the prediction of outputs from unknown examples. The loss function in Eq. (5.19) indicates how this error is penalized.

In practice the probability distribution $P(\mathbf{x}, y)$ is not known and Eq. (5.19) cannot be evaluated. However, a stochastic approximation to $R[f]$, known as empirical risk (R_{emp}), can be computed as,

$$R_{emp}[f] = \frac{1}{p}\sum_{i=1}^{p} L(f(\mathbf{x}_i) - y_i) \tag{5.20}$$

The empirical risk is a measure of prediction errors for the training set. It approaches the expected risk as the number of samples increases to infinity, i.e. $R_{emp}[f]_{p\to\infty} = R[f]$. This implies that for a small size training set the minimization of R_{emp} does not ensure the minimization of $R[f]$. As a consequence, the selection of a predictive model based only on the basis of empirical risk minimization does not guarantee a good generalization performance and often produces a phenomenon known as over-fitting. Over-fitting occurs when the complexity of the model is extremely high and the function fits not only the mechanism underlying the training data but also the noise of data. This undesirable effect is controlled by introducing a term known as *capacity control* which penalizes the complexity of the model. The underlying idea in statistical learning is that if we build a model with low capacity that yields a small empirical risk, then the true risk is also likely to be small.

Support Vector Regression (SVR). This method is an adaptation of the Support Vector Machines Theory to solve regression problems (Schölkopf and Smola 2002). In SVR, the inputs are first mapped into a high dimensional feature space (***F***) in which they are linearly correlated with the targets. Support vector regression follows a principle know as structural risk minimization (SRM) instead of the commonly empirical risk minimization (ERM) used in machine learning and artificial neural networks. SRM generates models with improved generalization performance by minimizing simultaneously prediction errors and model complexity.

The SVR method considers the following linear estimator,

$$f(\mathbf{x}) = (\mathbf{w} \cdot \Phi(\mathbf{x})) + b \tag{5.21}$$

when solving a non-linear regression problem. In this equation, $\mathbf{w}$ is a weight vector, b is a constant term, $\Phi(\mathbf{x})$ represents a function named feature, and $(\mathbf{w} \cdot \Phi(\mathbf{x}))$ is the dot product in the feature space, ***F***, such that $\Phi : \mathbf{x} \to F, \mathbf{w} \in F$. In the SVR formalism the problem of non-linear regression in the lower dimensional input space, ($\mathbf{x}$), is transformed in a linear regression problem in a high dimensional feature space, ***F***. To avoid the over-fitting of the regression model, SVR minimizes a

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

regularized risk functional based on the empirical risk and a model complexity term $\|\mathbf{w}\|^2$

$$R_{reg}[f] = R_{emp}[f] + \frac{1}{2}\|\mathbf{w}\|^2 \tag{5.22}$$

where R_{reg} is the regularized regression risk and $\|\bullet\|^2$ is the Euclidean norm. The term $1/2\|\mathbf{w}\|^2$ in Eq. (5.22) controls the trade-off between model complexity and accuracy. In fact, the complexity of the linear estimator function defined by Eq. (5.21) is controlled by keeping **w** as small as possible. Some ANN models with good generalization properties also use this *weight decay* term in their cost functions. However, SVR and these ANN use conceptually different approaches to minimize these cost functions.

Several types of either linear or non-linear cost functions (loss) can be used in the SVR formulation. The most frequently used is the so-called ε-insensitive loss function which is defined as,

$$L_\varepsilon(f(\mathbf{x}) - y) = \begin{cases} |f(\mathbf{x}) - y| - \varepsilon & \text{for } |f(x) - y| \geq \varepsilon \\ 0 \text{ otherwise} \end{cases} \tag{5.23}$$

where ε is a model parameter which determines the radius of the tube located around the function *f*(**x**). Figure 5.2 depicts some of these loss functions.

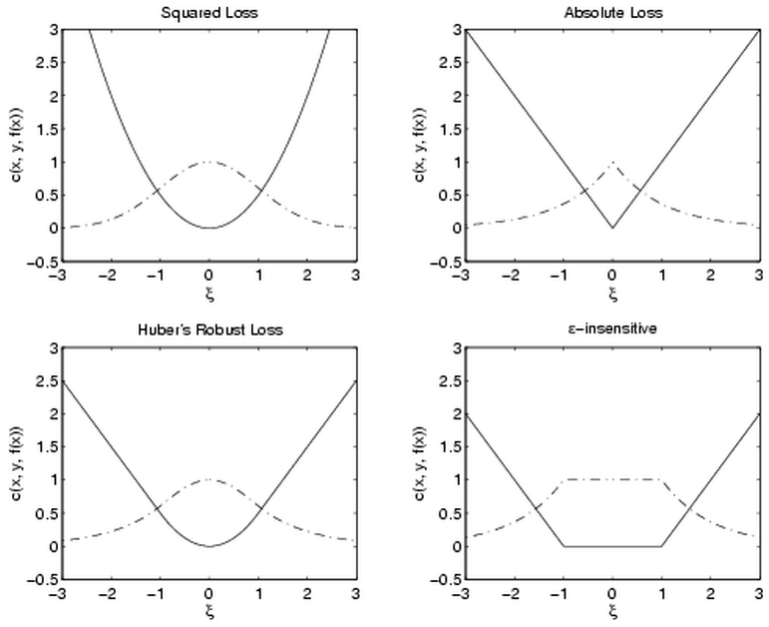


Figure 5.2. Examples of commonly used loss functions

while Figure 5.3 depicts this tube, also known as ε–insensitive zone. In this zone, the loss function is equal to zero and does not penalize prediction errors smaller than ε.

ISBN:978-84-691-1008-9/DL: T.2230-2007

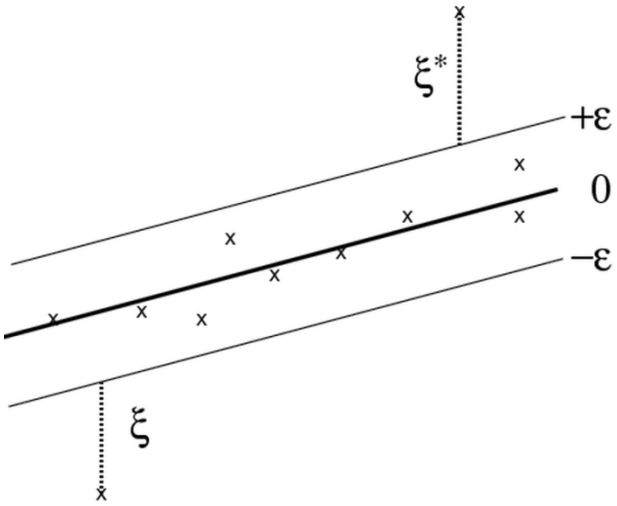


Figure 5.3. ε-insensitive zone and slack variables

The empirical risk minimization problem given by Eq. (5.21) can be reformulated by using Eq. (5.22) and adding two slack variables, ξ_i and $\xi_i^*, i = 1,..., p$, into the risk function together with a set of linear constrains. These slack variables (see figure 5.3) measure the deviation of $(y_i - f(\mathbf{x}_i))$ from the boundaries of the ε-insensitive zone. After the introduction of a regularization constant, C, the optimization problem in equation (5.21) becomes,

$$\text{Minimize}: \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{p}(\xi_i + \xi_i^*) \tag{5.24}$$

subject to the following set of linear constraints:

$$\begin{cases} (w \cdot \Phi(\mathbf{x}_i)) + b - y_i \le (\varepsilon + \xi_i^*) \\ y_i - (w \cdot \Phi(\mathbf{x}_i)) - b \le (\varepsilon + \xi_i) \\ \xi_i, \xi_i^* \ge 0 \quad \text{for } i = 1,..., p \end{cases} \tag{5.25}$$

SVR optimizes the position of the ε-tube around the training data when performing this minimization. Restrictions given by Eq. (5.25) penalize all the training points located at a distance greater than ε from the fitted model. To this end, SVR minimizes the training error by minimizing not only the deviations from the function (ξ_i, ξ_i^*), but also the complexity term ($\|\mathbf{w}\|^2$) with the objective of increasing the flatness of the function, i.e., decreasing its complexity.

Vapnik (1998) demonstrated that the regularized risk functional given by Eq. (5.24) is minimized by

$$f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \sum_{i=1}^{p}(\alpha_i - \alpha_i^*)(\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) + b \tag{5.26}$$

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

where α_i and α_i^* are coefficients equal or greater than zero. They are known as Lagrange multipliers, which pertain to the input data vector $\mathbf{x}_i$ and satisfy $\alpha_i \alpha_i^* = 0, i = 1,.., p$.

Equations (5.24), (5.25) and (5.26) involve the computation of a dot product in the high dimensional feature space, ***F***. To avoid these time consuming computations the technique known as "*kernel trick*" can be used. This method is based in the Mercer's condition which states that any positive semi-definite, symmetric kernel function, *K*, can be expressed as a dot product in an alternate high dimensional space. The main advantage of using this kernel function is that the dot product in the feature space can be computed without mapping the vectors $\mathbf{x}$ and $\mathbf{x}_i$ into that space; all computations can be carried out in the original input space. In the SVR formalism several kernels, either linear or nonlinear (polynomial, Gaussian, etc.) can be used. The most commonly used kernel is the radial basis function,

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{5.27}$$

where σ is the width (or radius) of the RBF. The substitution of the dot product in Eq. (5.26) by a kernel yields the general form of the SVR based regression model,

$$f(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \sum_{i=1}^{p} (\alpha_i - \alpha_i^*) K(\mathbf{x}, \mathbf{x}_i) + b \tag{5.28}$$

In Eq. (5.28) the original weight vector, ***w***, is expressed in terms of the Lagrange multipliers. The values of these multipliers are obtained by solving a convex quadratic programming problem (QP),

$$\text{Max: } -\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) - \varepsilon\sum_{i=1}^{p}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{p}(\alpha_i - \alpha_i^*) \tag{5.29}$$

subject to $\alpha_i, \alpha_i^* \in [0, C]$ and $\sum_{i=1}^{p}(\alpha_i - \alpha_i^*) = 0$. The bias parameter, b, in equation (5.26) is computed by using the Karush-Kuhn-Tucker (KKT) conditions,

$$b = \begin{cases} y_i - f(\mathbf{x}_i)_{b=0} - \varepsilon & \text{for } \alpha_i \in (0, C) \\ y_i - f(\mathbf{x}_i)_{b=0} + \varepsilon & \text{for } \alpha_i^* \in (0, C) \end{cases} \tag{5.30}$$

Each training data point is associated with pairs of Lagrange multipliers which can be interpreted as interactions pushing or pulling the model output towards target values. The solution of the QP problem in Eq. (5.29) assures that some of the coefficients $(\alpha_i - \alpha_i^*)$ have non-zero values. The set of training vectors with non-zero coefficients are named "*Support Vectors*" (SV) and represent important data examples for which $|f(\mathbf{x}_i) - y_i| \geq \varepsilon$. As the percentage of SVs decrease the model obtained is more general and lesser computations are necessary to evaluate new pairs of data. The points with zero values in the Lagrange multipliers have no influence on

ISBN:978-84-691-1008-9/DL: T.2230-2007

the model and the solution remains the same if removed. The final model can be defined as a combination of a relatively small number of input vectors; this is known as "sparseness".

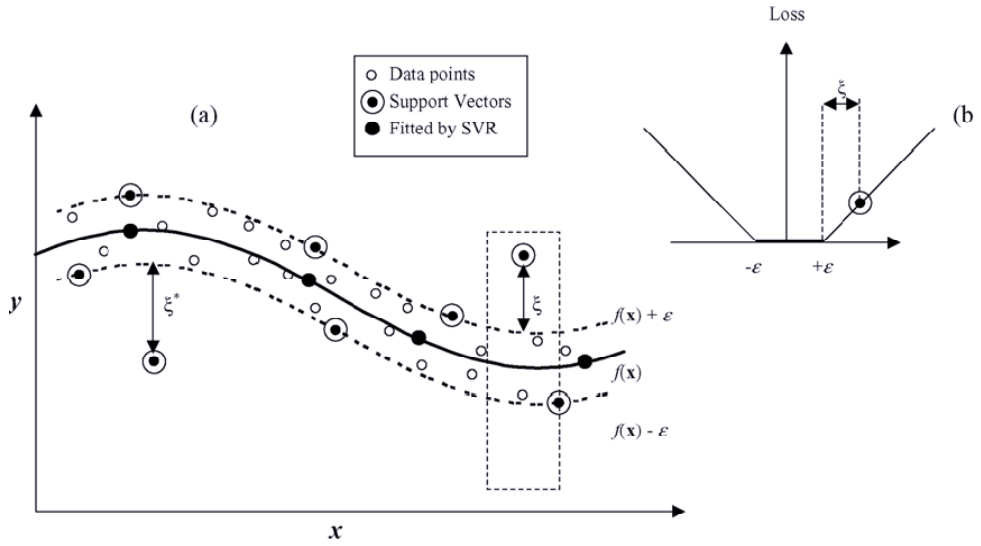


Figure 5.4. Representation of a SVR by using the ε-insensitive loss function (a); the linear ε-insensitive loss function (b). The capacity *C* determines the slope

Figure 5.4a depicts the samples used as SVs. The prediction accuracy and generalization capability is controlled by two parameters named *C* and ε. *C* controls the number of errors with values higher than $|\varepsilon|$ and determines the slope of the linear loss function (see Figure 5.4b). If the value of *C* is too high ($C \to \infty$) the SVR model minimizes only the empirical risk without considering model complexity. Low values of *C* yield considerable errors in the training set but increase the chance of obtaining models with good generalization capability. The size of the ε-tube controls the amount of SVs included in the model and therefore its complexity. If ε decreases the number of SVs increase and the complexity of the model also increases. As a consequence, these parameters must be carefully chosen to obtain well performing models.

Support Vector Classification (SVC). The same principles of Support Vector Machines and structural risk minimization can be applied to classification problems. In a general sense and without loss of generality, any classification problem can be restricted to a two-class problem. Let us consider a set of training data, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{p}$, where $\mathbf{x}_i \in \Re^n$ is a vector of the input data and $y_i \in \{-1,1\}$ represents the corresponding two output classes.

We may assume that there exist a separating hyperplane defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$ for linear support vector machines in this classification case. Let d_+ (*d*) be the shortest distance from the separating hyperplane to the closest positive (negative) example. The support vector algorithm looks for the separating hyperplane with largest margin. This is equivalent to the minimization of the risk functional given by Eq. (5.22) with a specific *class loss* function. A Lagrange formulation equivalent to the one in the previous case can be applied and a convex quadratic programming problem

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

solved; a similar solution is obtained. If data are non-linearly separable, positive slack variables are introduced to relax the constraints of the optimization problem.

If the decision function is a non-linear function of the training data, the kernel trick can be applied to solve the problem in a different feature space which becomes linearly separable. The only restrictions are that kernels must fulfill Mercer's condition. For these non-linear SVM the same Lagrange formulation is applied and the corresponding convex quadratic programming problem solved.

5.2 Classification and Prediction of Biodegradation in water and soil

Classification and prediction models for the biodegradation rates of chemicals have been used to illustrate the functionalities of the modeling tier. The results of the SOM-based EDA carried out for these data sets and discussed previously in section 3.3 have been used to analyze the structure of the chemical space. In addition, feature selection and train/test generation methods provided by the preprocessing tier in Chapter 4 have also been applied.

The development of models to estimate properties and activity of chemicals is mainly based on structural similarity. Under this assumption, chemicals with similar molecular structure are expected to have similar properties and activities. This modeling paradigm, known as Quantitative Structure-Activity Relationship (QSAR), was introduced by Hansch (1962) as an extension of the pioneering work of Hammett (1940). QSARs originate from the intuitive notion that the molecular structure of a given chemical determines its activity. The term "activity" may be a specific end-point measurement representing a biological, physicochemical, physiological, or chemical process. The QSAR approach relies on the development of models that relate variations in activity to changes in molecular structure by using molecular descriptors. These descriptors are the mathematical representation of the information content encoded in a molecule and may embody empirical, quantum chemical or non-empirical parameters. Empirical descriptors can be measured or estimated and include physicochemical properties such as hydrophobic, electronic, and steric terms. Non-empirical descriptors are typically structural properties based on 2-D topological or graph theoretical parameters. Quantum chemical descriptors are based on optimized 3-D structures of molecules.

QSARs constitute a quantitative modeling approach which has been extended to qualitative models known as Structure-Activity Relationships (SAR). The term *activity* can be replaced by the property being modeled. For example, in this section Quantitative Structure Biodegradation Relationship models (QSBR) and Structure-Biodegradation Relationships (SBR) are developed and discussed.

The validation of QSARs has been (and still is) the subject of intense debate within the academic, regulatory and regulated communities. The main point of discussion is how to define the criteria needed to establish the scientific validation of models to

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

permit its application for regulatory purposes. Nevertheless, all the involved parts agree in the need to establish *"appropriate measures of goodness-of-fit, robustness and predictivity"* for any QSAR model. Thus the two fundamental steps in QSAR modeling should be:

- model validation, both internal and external, which implies quantitative assessment of model robustness and its predictive power; and
- definition of the application domain of the model in the space of chemical descriptors used in the derivation.

Internal validation is an essential, but not sufficient, form of validation. It should ideally be supplemented by external validation if QSAR models have to be used for predictive purposes. In internal validation, the information related to each chemical in the data set is considered at least in one iteration of the validation process. Therefore, these chemicals are used for model development and their information is included in the model. Internal validation provides only a reasonable first approximation of the predictive ability of a QSAR model. Diverse techniques can be used to perform the internal validation process:

- **Cross-validation.** This method refers to the use of one or more statistical techniques for internal validation in which different proportions of chemicals are omitted from the training set to verify the "internal predictability", e.g., leave-one-out, leave-many-out, or bootstrapping. The QSAR is developed on the basis of training chemicals and used afterwards to predict activities or properties for the chemicals that were omitted, i.e., for the test set. This procedure is repeated a number of times so that statistics can be derived from the comparison of predicted data with the known data. The final model is the model developed for all chemicals. Cross-validation techniques provide an assessment of internal predictability as well as of the robustness of the model (stability of QSAR model parameters). In any case nothing is known regarding the predictability on new external chemicals.
- **Cross-validation by Leave-One-Out (LOO).** This procedure uses n training sets formed by excluding each time one chemical from the training set, i.e., models are developed by using training sets of $n-1$ compounds. For each model, the excluded compound is predicted and the cross-validated explained variance computed as the average value of all validation runs. The LOO approach is not sufficient to assess robustness and predictability of a QSAR model.
- **Cross-validation by the Leave-Many-Out (LMO).** LMO employs smaller training sets than the LOO procedure. In a typical LMO validation, n chemicals of the data set are divided in G cancellation groups of equal size, m_j (= n/G). G is generally selected between 2 and 10 depending on the number of available training examples. Models are developed for each of the $n-m_j$ objects in the training set and m_j objects in the validation set. For each corresponding model, m_j objects are predicted and the cross validated explained variance computed. When dealing with small data sets, LMO validation with too strong perturbation (up to 50% of data out of training,

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

each run) often under-estimates predictability because only a reduced part of the data is used each time for model calibration.

- **Boot-strapping.** This method requires the data set to be representative of the population from which it was drawn. Since there is only one data set, bootstrapping simulates what would happen if the population in the data set were randomly resampled to obtain a training set. In a typical bootstrap validation, *K* groups containing *n* compounds are generated by random selection with replacement of compounds from the original data set. Some of these objects may appear in the same training group more than once while others might never be present. The model obtained with the first selected training objects is used to predict the values for the excluded sample. This yields an ensemble of estimates which are subsequently used to obtain an estimate of the variance of prediction error.

Finding new experimentally tested compounds for external validation purposes is generally difficult. The new data should be in a statistically significant number (in fact, results on few data can give optimistic or unreliable information regarding to the model predictability) and belonging to the same chemical domain as the compounds used for model development. When additional data are not available, statistical external validation can be helpful in defining more precisely the actual predictive power of the model. This is done by an adequate splitting of the available input data set into training (for model development) and validation (for model predictive assessment) sets, using experimental design and other procedures such as the Self-Organizing Map. Splitting chemicals into two sets by random selection, while useful for internal validation, might yield very variable results when applied in external validation, depending strongly on the dimension and representativeness of the split sets. The optimal splitting should lead to a validation set in which each of its members is close to at least one point of the training set. The composition of the training and validation sets is of crucial importance. A representative selection of compounds spanning, to a good degree, the chemical domain of interest should be included in these sets.

The domain of application of robust, significant and validated QSAR must be defined since they cannot be expected to reliably predict the modeled property or activity for the entire universe of chemicals. Predictions for only those chemicals that fall in this domain can be considered reliable. The chemical domain of applicability is a theoretical region in the space defined by the modeled response and the descriptors for which a given QSAR should make reliable predictions. This region is defined by the nature of the chemicals in the training set, and can be characterized in various ways, e.g., the Williams plot described in chapter 4.

Once the QSAR has been internally and externally validated response permutation methods such as Y-scrambling can be applied to check robustness and statistical significance. In response permutation, the dependent variable vector, Y-vector, is randomly shuffled and a new QSAR model developed by using the original independent variable matrix. The process is repeated several times. It is expected that the resulting QSAR models yield low values for its correlation coefficients. If the new models developed from the data set with randomized responses have

significantly lower values for the correlation coefficient than the original model, then this is strong evidence that the proposed model is well founded, and performance is not the result of chance correlation. In contrast, if the QSAR models obtained in the Y-randomization test yield relatively high correlations, then the model is not acceptable for the used data set.

The statistical performance of SBR models can be characterized in terms of their contingency table (or confusion matrix,

		True class: p	True class: n
Hypothesized class	Y	True Positives	Fase Positives
	N	False Negatives	True Negatives
		P	N

In this table P and N represent positive and negative examples, respectively. Additional metrics used to report the performance of classifiers can be derived from the confusion matrix:

- FP rate = FP/N
- TP rate or Sensitivity: TP/P = Recall
- Specificity: TN/(FP + TN) = 1 – FP
- Precision = TP/(TP+FP)
- Accuracy = (TP+TN)/(P+N)

More informative statistics can be derived from the confusion matrix using association coefficients such as concordance, κ and λ_b. Concordance simply measures the proportion of chemicals where the predicted and experimental classifications agree. The main drawback with this measure is that ignores the fraction of agreement that might have been obtained by chance. In contrast, the κ index normalizes the level of agreement between predicted and actual categories with respect to the agreement that could have been achieved by allocating chemicals to classes at random but according to data distribution. Finally, λ_b quantifies the reduction in the prediction error achieved when using the classifier compared to that obtained when allocating chemicals to classes on the basis of their marginal totals. A complete discussion and details for the calculation of these indices is given by Schüürmann et al. (2003).

To compare different qualitative classification systems (SBR models) the ROC (Receiver Operating Characteristics) analysis is commonly used. ROC graphs are 2-dimensional representations of false positive rate (FP) versus true positive rate (TP). A ROC graph depicts the relative trade-off between benefits (true positives) and

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

costs (false positives). A discrete classifier is represented as a point in the ROC space. Informally, one point in the ROC space is considered better that another if the former is located to the northwest of the latter (TP rate higher, FP rate lower or both). Classifiers on the left-hand side of the ROC graph near the X axis can be thought as "conservative", i.e., make positive classification only with strong evidence so they make few false positive errors but often have low true positive rates. Classifiers on the upper right-hand side are considered as "liberal", i.e., make positive classifications with weak evidence so nearly all positives are classified correctly, but often have a high false positive rate.

5.2.1 Models for Biodegradation in Water

The data set used to develop these models contains 672 chemicals and is fully described in section 3.3. A pool of 900 molecular descriptors was used to characterize the chemical space of these compounds. The target property is the percentage of biodegradation (BOD) after the MITI-1 experiment.

5.2.1.1 Quantitative Models (QSBR)

The first set of models is developed for the complete range of biodegradation values. The SOM-based EDA carried out in section 3.3 revealed a very imbalance data set with a skewed data distribution in which most chemicals are non-biodegradable (BOD=0%). Despite these limitations, results obtained using the complete dataset are presented to confirm the modeling difficulties anticipated by EDA.

Prior to the development of the models, diverse feature selection algorithms were used to extract the best subsets of descriptors to predict BOD. In this section the subsets of descriptors which have generated QSBR models that yielded the best performances are discussed. For the complete range of BOD values [0-100] the feature selection methods which yield the best models are the filters CFS and ReliefF, described in section 4.1.

Table 5.1 shows the two sets of molecular descriptors selected by CFS and ReliefF, respectively. Although both techniques select a similar number of descriptors, only *ATS Geary1 AlogP98* is selected by both procedures. Despite the significant dimensionality reduction accomplished, the QSBR models developed using the selected descriptors yield acceptable error rates of 20-25% in agreement with quantitative models reported in the literature. Table 3 summarizes the errors obtained with these descriptors and QSBR models built with feed-forward and Fuzzy ARTMAP neural networks. The feed-forward networks used had an optimized architecture of 8-22-1 and were trained using the error backpropagation algorithm. External validation was performed on an independent dataset never seen before. Internal validation was performed by *leave-one-out* (LOO) cross-validation technique.

Table 5.2 shows that the magnitudes of the error for the external validation and for internal LOO validation are of the same order in all cases. This constitutes a clear

indication of the validity of the developed QSBR models for the whole application domain.

Table 5.1. Best subsets of descriptors selected using two filter algorithms, CFS and ReliefF, for the complete range of BOD values [0-100]

Method (number)	Descriptors
CFS (8)	Heat of Formation HOMO IP **ATSGeary 1 AlogP98** Bound charge index 4 Eccentric adjacency index V Chi4 path/cluster V Chi3 cluster
ReliefF (9)	ATS Geary3 AlogP98 **ATS Geary1 AlogP98** ATS Geary3 electronegativity ATS Geary2 AlogP98 ATS Geary 1 polarizability Fraction of 2D-VSA polar ATS Geary 2 mass LUMO

Table 5.2. Summary of the absolute mean errors and standard deviations obtained from QSBR models developed for the complete range of BOD values

			All	Train	Test	
BOD range	**Feature selection method**	**QSBR model**	**error (σ)**	**error (σ)**	**error (σ)**	**LOO error (σ)**
[0-100]	**CFS**	Fuzzy ARTMAP	5.09 (10.99)	1.32 (1.41)	**20.58** (17.64)	27.36 (30.13)
		feed-forward	9.75 (15.28)	6.19 (8.69)	24.39 (24.87)	29.01 (35.36)
	ReliefF	Fuzzy ARTMAP	11.12 (15.12)	7.63 (8.35)	25.71 (25.09)	26.25 (30.7)
		feed-forward	9.90 (15.30)	6.25 (6.96)	25.13 (26.89)	27.35 (31.7)

Figure 5.5a shows the results for the entire range of biodegradation [0-100%] for the best model obtained, i.e., for the Fuzzy ARTMAP based QSBR. The absolute mean error is 5.1% for the entire set with a standard deviation of around 11%. For the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

external validation the absolute mean error is higher (20.6%) but of the same magnitude as the error obtained in the LOO internal validation process. Figure 5.5b shows the error distribution for both training and test sets.

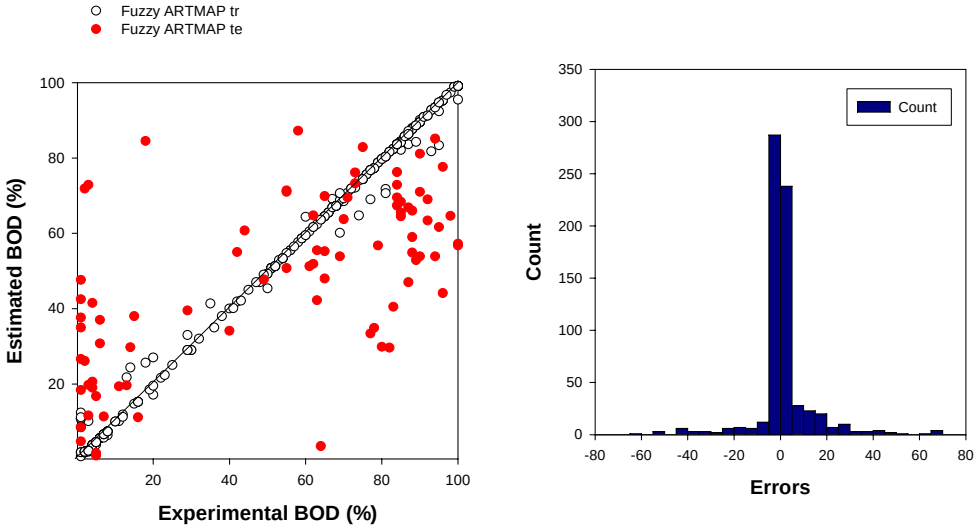


Figure 5.5. Fuzzy ARTMAP-based QSBR model with descriptors selected by the CFS filter algorithm. Experimental vs. predicted BODs for the train and test sets (a); histogram of error distribution for the train and test sets (b)

According to the structure of the distribution revealed during the exploratory stage, the dataset was split into two non-overlapping sets. A BOD threshold of 40% was adopted to discriminate between non-biodegradable and ready-biodegradable compounds. The lower range [0-40% BOD] is formed by 353 substances.

Training and test set were selected by using the SOM-based procedure presented in section 4.1 to guarantee the even distribution of compounds in the application domain. The training set was formed by 284 compounds selected as representative of the whole data. The remaining 69 chemicals were used for testing in the external validation procedure.

A feature selection process has been performed for the low range of biodegradation. Table 5.3 shows the descriptors selected by the methods which yielded the best QSBR models. It should be noted that CFS, ReliefF and the ANNIGMA wrapper select a relatively low number of descriptors method (6, 6, and 8 respectively) compared to the SOM dissimilarity procedure which selects a larger set of 25 descriptors. Nonetheless, all methods select descriptors that cover the complete chemical space with geometrical, electro-topological and quantum information. It should be noted that only ReliefF and SOM algorithms select the LC_{50} ecotoxicity index as relevant information.

ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 5.3. Best subsets of descriptors selected with filter and wrapper algorithms for the range of low (non-biodegradable) BOD values [0-40]

Method (number)	Descriptors
CFS (6)	Fraction of Rotatable bonds 2D-VSA hydrophobic_unsat Eccentric adjacency index Ring degree-distance index Path/walk3 ATS Moran 3 mass
ReliefF (6)	LC50 Fraction of Rotatable bonds SC-4cluster ATS Moran 1 mass Balaban index JY Balaban index JX
ANNIGMA [backward] (8)	Heat of Formation Log K_{ow} (calc.) HOMO LUMO Chi 4 path/cluster Diff. Chi 4 log P path walk 4
SOM Dissimilarity (25)	LC50 atom count Conf. Energy dipole moment heat of formation HOMO Shape Index 3 No.H-bond donors Fracc. Rotable bonds 2D-VSA hydrophobic 2D-VSA hydrophobic sat. 2D-VSA hydrophobic unsat. 2D-VSA other Eccentric adjacency index Ring degree-distance index Difference chi 0 Difference chi 2 Difference chi 3 Difference chi 4 Difference chi 5 ATS Moran 1 mass ATS Moran 3 mass ATS Moran 4 mass ATS Moran 4 AlogP98

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The SOM is the algorithm that selects the largest number of descriptors and that yields the best QSBR model, both in terms of internal and external validation. Despite the different number of molecular descriptors selected by each technique, the generalization capabilities of all QSBR models are quite equivalent.

The molecular descriptors selected by CFS for the range [0-40% BOD] are shown in Figure 5.6 in terms of the c-plane representation of SOM. Two well defined zones are observed in the component plane corresponding to biodegradation. A similar situation occurs with the rest of feature selection methods. This constitutes a clear indication that the selected descriptors cover the whole descriptor space and induce a coherent clustering structure in the target property.

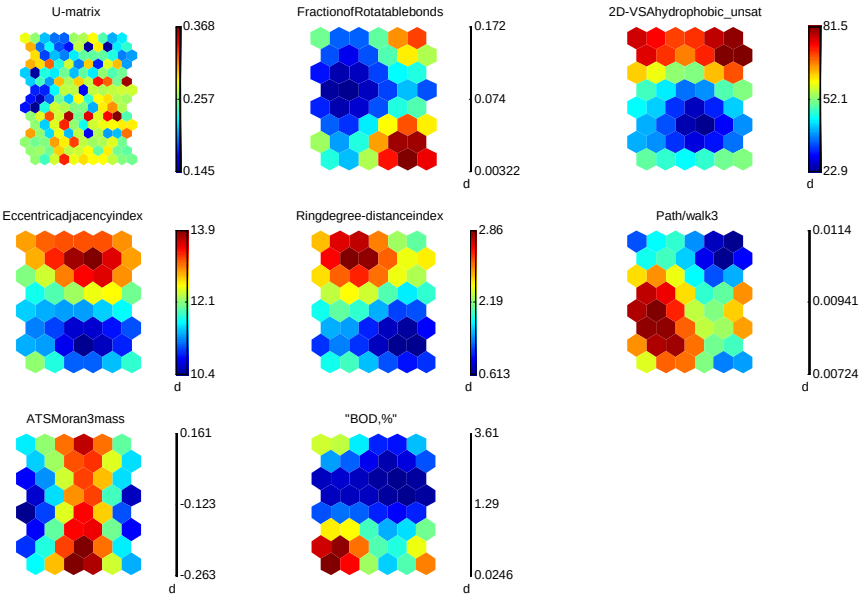


Figure 5.6. Visualization of the component planes corresponding to the set of molecular descriptors selected by CFS for the lower range of biodegradability data [0-40% BOD]

Table 5.4 summarizes the internal and external validation results for the best QSBR models. It can be observed that the models resulting from descriptors selected by CFS and ReliefF methods yield very close results. The fuzzy ARTMAP model with the most suitable set of descriptors selected by CFS was trained with the vigilance parameter set to ρ_a =0.9. This strategy yields models with improved generalization capabilities and has been used to train all the fuzzy ARTMAP based QSBRs. This model predicts the % BOD for the complete data set of 353 compounds with an average absolute error of 1.46 % and a standard deviation of 4.06%. For the validation sets performance decreases with errors increasing to 5.67% and 3.90% for external and internal validation, respectively. As expected, the performance of the Fuzzy ARTMAP-based QSBR models is significantly superior to that for feed-forward neural networks which were unable to predict the new data set.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 5.4 Summary of the absolute mean errors and standard deviations obtained from QSBR models developed for the low range of BOD values [0-40%]

			All	Train	Test	
BOD range	**Feature selection method**	**QSBR model**	**error (σ)**	**error (σ)**	**error (σ)**	**LOO Error (σ)**
[0-40]	**CFS**	Fuzzy ARTMAP	1.46 (4.06)	0.44 (1.82)	5.67 (7.05)	3.9 (6.39)
		feed-forward	1.66 (6.15)	0.52 (2.0)	6.42 (12.35)	3.16 (6.23)
	ReliefF	Fuzzy ARTMAP	1.47 (3.75)	0.42 (0.71)	5.56 (6.79)	4.59 (7.25)
		feed-forward	3.54 (5.60)	2.72 (4.38)	6.75 (8.17)	3.09 (5.68)
	ANNIGMA [backward]	Fuzzy ARTMAP	0.96 (2.12)	0.37 (0.65)	3.25 (3.72)	3.89 (5.62)
		feed-forward	2.37 (3.42)	2.20 (3.56)	3.03 (2.74)	2.85 (5.35)
	SOM	Fuzzy ARTMAP	0.86 (2.08)	0.35 (0.56)	**2.87** (3.88)	**2.83** (5.02)
		feed-forward	1.44 (4.72)	0.12 (0.51)	6.55 (8.71)	3.12 (5.82)

The most difficult to predict chemicals are those with BOD values close to or equal to zero. All the QSBR models exhibit a tendency to have anomalous responses for these data. This confirms what was observed in the exploration of the chemical space for biodegradation. The BOD = 0 mode of the distribution restricts the development of goog performing QSBR models. It should also be mentioned that the SOM based feature selection method yields a significantly better performance when it is used in conjunction with a neural classifier such as Fuzzy ARTMAP. Table 5.4 shows that in this case the validation error rates for the FAM-based QSBR are 2.87 and 2.83% for the external and internal validation procedures, respectively. In contrast for QSBRs based in feed-forward networks the error rates respectively increase to 6.55 and 3.12% for external and internal validation.

The procedure was repeated for the higher range of biodegradability, i.e., for compounds which will biodegrade fast. In this case, the BOD values cover the range [>40-100%]. Table 6 show the results obtained with the feature selection methods that yielded best results. In this case the number of descriptors selected range from a minimum of 4, corresponding to the ANNIGMA wrapper using forward selection, to a maximum of 10 corresponding to the ANNIGMA wrapper using backward elimination. The CFS and ReliefF filters select a similar number of 7-8 descriptors.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 5.5. Best subsets of descriptors selected with filter and wrapper algorithms for the range of high (ready-biodegradable) BOD values [>40-100]

Method (number)	Descriptors
CFS (7)	Log K_{ow} (calc.) ATS Geary 1 polarizability ATS Geary 3 AlogP98 ATS Geary 3 E-state E-state SaaaC E-state S_hydrophobic E-state S_none
ReliefF (8)	ATSGeary2mass 2D-VSAHbonddonor ATS Geary1 AlogP98 ATS Geary2 electronegativity ATS Geary1 VDWradius, ATS Geary2 VDWradius ATS Geary1 mass V Chi4 path/cluster
ANNIGMA [forward] (4)	E-state S_hbond_donor Valence bound chargeindex 0 I_adj_equ E-state S_hydrophobic_sat
ANNIGMA [backward] (10)	Ionization Potential 2-MTIprime ATS Geary 2 electronegativity ATS Geary 3 electronegativity E-state SH_hydrophobic E-state S_polar Valence bound charge index 7 Valence charge index 3 VChi4 path/cluster Vertex degree-distance index

Figure 5.7 shows the structure of the component planes corresponding to the molecular descriptor space selected by CFS. This feature extraction method yields a selection of descriptors which maintain the structure of two classes of BOD values observed in the case of low biodegradability (see figure 5.6). This constitutes an indication of the bimodal nature of the distribution of BOD values over this range, that may affect the performance of the corresponding QSBR models developed. Table 5.6 summarizes the results obtained for the best set of QSBR models. It can be observed that all models yield similar error rates. As for the other ranges of BOD, internal and external validation error rates are of the same magnitude.

ISBN:978-84-691-1008-9/DL: T.2230-2007

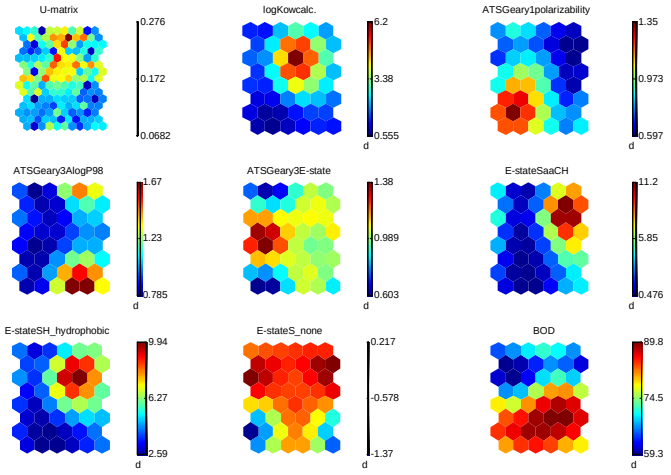


Figure 5.7. Visualization of the component planes corresponding to the set of molecular descriptors selected by CFS for the high range of biodegradability [>40-100% BOD]

Table 5.6. Absolute mean errors and standard deviations for QSBR models developed in the high range of BOD values [>40-100%]

			All	Train	Test	
BOD range	**Feature selection method**	**QSBR model**	**error (σ)**	**error (σ)**	**error (σ)**	**LOO Error (σ)**
[>40-100]	**CFS**	Fuzzy ARTMAP	3.11 (5.48)	1.13 (0.49)	**11.16** (8.47)	16.55 (12.21)
		feed-forward	6.20 (6.17)	4.46 (3.99)	13.24 (8.21)	16.04 (12.29)
	ReliefF	Fuzzy ARTMAP	4.76 (8.93)	1.40 (1.96)	18.38 (12.70)	16.49 (13.32)
		feed-forward	7.04 (8.29)	4.64 (4.96)	16.78 (11.52)	17.61 (14.33)
	ANNIGMA [forward]	Fuzzy ARTMAP	5.36 (8.37)	2.54 (3.93)	16.85 (11.48)	16.72 (6.17)
		feed-forward	7.11 (8.76)	4.47 (5.13)	17.85 (11.96)	17.96 (13.75)
	ANNIGMA [backward]	Fuzzy ARTMAP	4.17 (7.54)	1.12 (0.76)	16.59 (9.89)	18.69 (13.67)
		feed-forward	7.18 (7.73)	5.14 (4.46)	15.45 (11.74)	16.55 (12.90)

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The fuzzy ARTMAP-based QSBR model developed with the molecular descriptors selected by CFS shows the best generalization capability. The absolute mean error and the corresponding standard deviation of the test set are 11.16 and 8.47%, respectively. The performance slightly decreases when using the backpropagation algorithm as the error rate increases to 13.24%. Figure 5.8 compares both models.

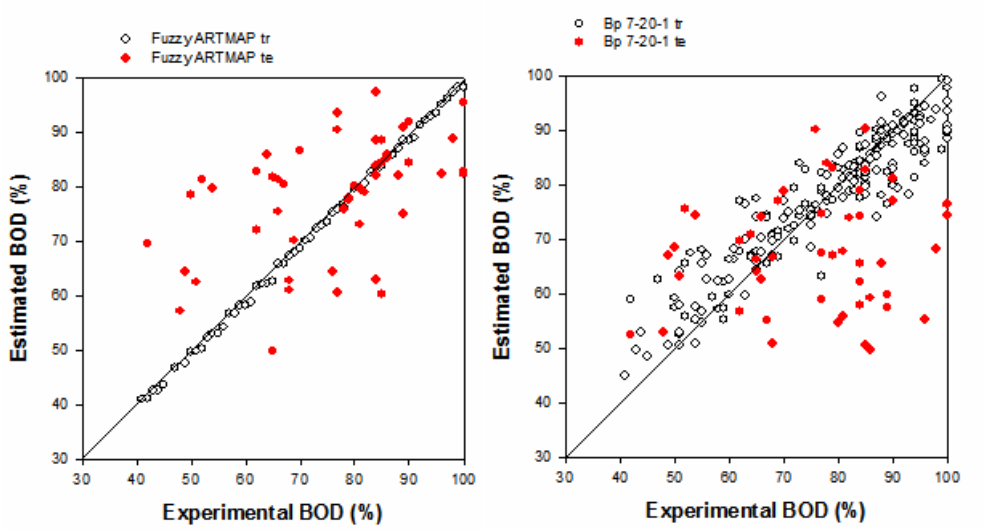


Figure 5.8 QSBR-based neural networks models with descriptors selected by the CFS filter technique. Fuzzy ARTMAP (a); feed-forward network using backpropagation (7-20-1) (b)

Table 5.7. Best subsets of descriptors selected using several feature selection algorithms for the whole range BOD values [1-100%]

Method (number)	Descriptors
CFS (7)	LUMO, kappa 3, ATS Geary1 AlogP98, ATSGeary3 electronegativity, Buffer solubility, Eccentric adjacency index, Vertex degree-distance index
ReliefF (10)	ATS Geary2 AlogP98, ATS Geary2 electronegativity, ATS Geary2 VDW radius, ATS Geary3AlogP98, ATS Geary3 electronegativity, Fraction of 2D-VSAhydrophobic, Fraction of 2D-VSApolar, Fraction of Rotatable bonds, Graph Petitjean, Path/walk4
ANNIGMA (7)	Heat of Formation, HOMO, log K_{ow} (calc.), Log P, LUMO, DeltaChi4path/cluster, Path/walk4

The main conclusion that can be drawn from these experiments is that all models perform similarly and that the error rate for both external and internal validation is consistent in average with a value around 20%.

Additional experiments have been performed to analyze the effect that non-biodegradable chemicals (0% BOD) have in models performance. Chemicals with zero biodegradation have been removed from the dataset in these experiments. Two overlapping ranges of biodegradation have been considered. Values of BOD for non-biodegradable compounds have been assigned to the range [1-60%] while readily biodegradable chemicals span the range [40-100%], with an overlap of chemicals within 40-60% BOD.

Table 5.8. Summary of the absolute mean errors and standard deviations obtained from QSBR models developed for the two overlapping BOD ranges

Method (number)	BOD range %	QSBR model	tr/te instances	All Error (σ)	Train Error (σ)	Test Error (σ)	LOO Error (σ)
CFS (7)	low [1-60]	FAM	140/80	7.21 (13.19)	1.84 (1.64)	16.60 (18.36)	22.73 (18.79)
		feed forward		14.74 (10.72)	14.05 (10.45)	15.96 (11.14)	21.40 (19.78)
	high [40-100]	FAM	159/86	10.29 (10.51)	7.23 (8.95)	16.02 (10.94)	19.55 (12.92)
		feed forward		13.16 (12.21)	12.36 (7.92)	**14.61** (17.71)	17.54 (14.34)
ReliefF (10)	low [1-60]	FAM	131/89	7.72 (13.63)	1.58 (1.52)	16.75 (17.89)	21.37 (10.72)
		feed forward		14.40 (25.20)	4.79 (7.74)	28.54 (33.95)	25.60 (20.67)
	high [40-100]	FAM	151/85	9.67 (12.52)	5.72 (9.53)	16.02 (13.00)	20.42 (10.14)
		feed forward		13.46 (11.88)	9.28 (7.22)	20.17 (14.58)	22.53 (18.45)
ANNIGMA (7)	low [1-60]	FAM	112/108	9.33 (15.05)	1.57 (1.58)	17.37 (18.23)	19.63 (17.21)
		feed forward		12.35 (20.40)	0.78 (1.96)	24.34 (23.72)	26.36 (31.48)
	high [40-100]	FAM	127/118	8.69 (11.25)	2.09 (1.75)	15.81 (12.72)	**16.89** (12.04)
		feed forward		15.16 (21.84)	1.97 (2.45)	29.36 (24.45)	22.8 (16.47)

Since the two data ranges are overlapping the feature selection process has been performed for the whole range of BOD values considered, i.e., excluding zero's.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 5.7 shows the subset of descriptors selected by the methods which yielded better QSBR models. All these methods select a similar number of descriptors. From these data, the generation of the best training/test sets using the SOM procedure and the development of QSBR models based on Fuzzy ARTMAP and feed-forward networks are performed.

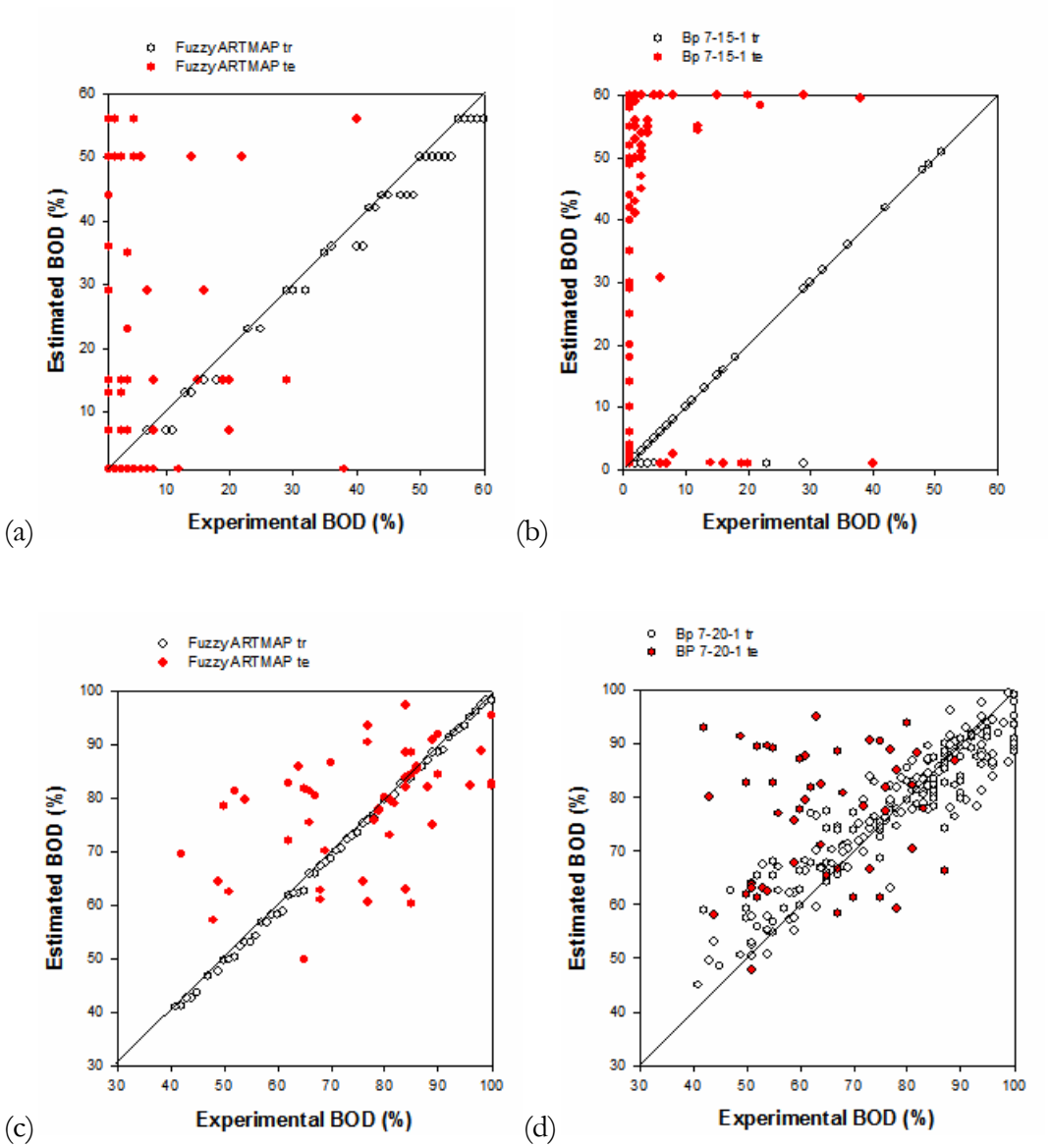


Figure 5.9. FAM and feed-forward based QSBR models developed using the descriptors selected with the CFS filter. Fuzzy ARTMAP for [1-60%] (a); backpropagation with a 7-15-1 architecture for [1-60%] (b); Fuzzy ARTMAP for [40-100%] (c); backpropagation with a 7-20-1 architecture for [40-100%] (d)

Table 5.8 summarizes the absolute mean error and standard deviation for each of the QSBR models developed using the three feature selection methods mentioned above

and the algorithms FAM and feed-forward NN. The best model in terms of external validation error is obtained for the subset of seven molecular descriptors selected by the CFS algorithm. In contrast, from the point of view of internal coherence, the most robust QSBR model corresponds to the subset of seven descriptors selected by ANNIGMA. Within the range of non-degradable chemicals (1-60% BOD) the generalization error is around 16-20% in most cases. In contrast, a slightly better generalization was achieved within the range of 40-100% biodegradation. Figure 5.9 compares the two FAM and feed-forward based QSBR models obtained using the molecular descriptors selected by CFS. As expected, the Fuzzy ARTMAP-based QSBRs yield a less scattered graph. Nevertheless both models yield poor predictions for the low range of biodegradation values (values less than 10% BOD). This is a clear indication that the dynamics of the biodegradation process for these "extremely" low values of BOD isn't well learnt by the networks.

Table 5.8 indicates that in general the performance of Fuzzy ARTMAP is superior to that for feed-forward networks. This applies to both internal and external validation tests.

5.2.1.2 Qualitative Models (SBR)

The results presented in the previous subsection show that the most important sources of error are in the range of low BOD values, specifically values close to or equal to zero. An alternative method to overcome this limitation and that still will be useful for environmental modeling might be to transform the regression problem (quantitative) into a classification problem (qualitative). This means that qualitative SBR models will be used instead of QSBR. As has been mentioned in previous sections there is agreement in the literature about the default threshold value of BOD that best discriminates between ready/non-ready biodegradable chemicals is located around 60% BOD. Furthermore, compounds having BOD values in the range of 45-55% aren't uncommon and constitute a source of uncertainty in biodegradation models because they may be considered either as biodegradable or as non-biodegradable.

The current study considers a value of 50% BOD as a cut-off to discriminate between ready and not ready biodegradation classes. Several experiments were carried out to confront current results with those reported in the literature. The experiments carried out considered the use of several classifiers based on machine learning techniques such as classification trees, M5-Tree, Instance based learners, as well as neural network based classifiers such as Fuzzy ARTMAP and Self organizing maps. In all subsequent experiments, CFS and Relief filters were used to select the input molecular descriptors used in the development of SBR models.

Assessment of classifiers is performed with the metrics discussed in section 5.3. In the following ROC diagrams the positive class is persistent (TP = low biodegradation) and false positives are biodegradable instances predicted as persistent. Again, the SOM procedure was used to select the best partition between training and test sets during the implementation and testing of the SBR models developed.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The CFS filter was used in the first experiment to select the most relevant molecular descriptors. This filter selects a compact subset of four molecular descriptors: ATS Geary1 AlogP98, Balaban index JX, Fraction of Rotatable bonds and Solvation chi4path/cluster. To train the corresponding SBR models, a set of 344 compounds was selected by SOM (202 with %BOD<50% and 142 with %BOD >50), while the remaining 328 compounds were used as test set (237 %B0D<50% and 91 %BOD>50%). Table 5.9 summarizes the results obtained for all SBR models developed.

Table 5.9. Performance of several SBR models with descriptors selected by CFS

ID	SBR Model		Sensitivity	Specificity	Miss classification Error (%)
A	Fuzzy ARTMAP	all	0.82	0.84	16.96
		train	0.98	0.99	1.45
		test	0.73	0.60	30.18
B	M5 Rule	all	0.85	0.84	15.33
		train	1.00	0.98	0.87
		test	0.73	0.62	30.49
C	IBK	all	0.85	0.85	14.73
		train	1.00	1.00	0.00
		test	0.73	0.62	30.18
D	SOM	all	0.88	0.71	17.71
		train	0.91	0.68	18.60
		test	0.86	0.75	**16.77**

Machine learning based methods (M5 and IBk) yield similar performance, showing an approximately 30% of misclassification rate for the test set, with a sensitivity of 0.73 and specificity of 0.62. In contrast, neural network based methods behave quite different. The Fuzzy ARTMAP based SBR was trained to discriminate between ready and not-ready biodegradation. The misclassification error rates obtained for the test set are similar to those for the machine learning algorithms (around 30%), and the same applies to the specificity (0.60) and sensitivity (0.73). SOM based SBR improves the generalization specificity and sensitivity up to 0.75 and 0.86, respectively. The misclassification rate decreases to 17% in this case. The examination of the ROC curve in Figure 5.10 confirms that, in general, all these SBR models with only 4 features are able to discriminate with acceptable accuracy between ready/not-ready biodegradable chemicals. Models A, B and C have similar classification power, while model D (SOM-based SBR) clearly shows an improved performance in this figure.

ISBN:978-84-691-1008-9/DL: T.2230-2007

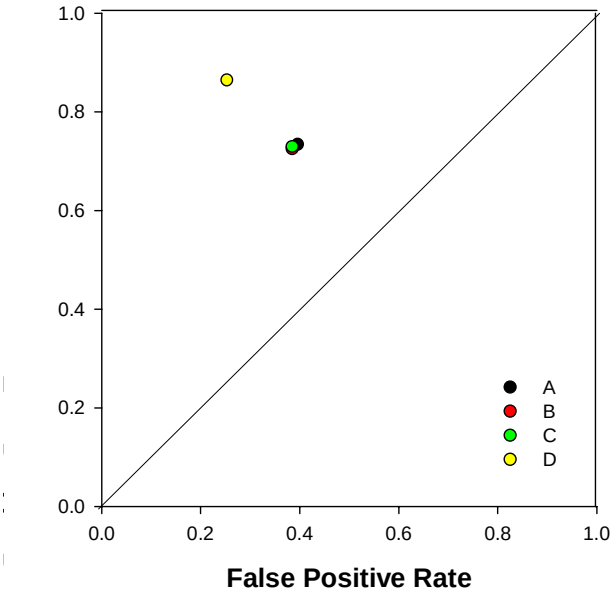


Figure 5.10. ROC plot for the test set of the SBR models developed using the CFS selected descriptors. (A) Fuzzy ARTMAP; (B) M5; (C) IBk; (D) SOM

A similar experiment was conducted using the ReliefF filter as the selection mechanism for the molecular descriptors. In this case, ten autocorrelation indices were selected: ATS Geary1Alog P98, ATS Geary1 E-state, ATS Geary1 electronegativity, ATS Geary1 VDW radius, ATS Geary2 E-state, ATS Geary2 electronegativity, ATS Geary2 mass, ATS Geary2 VDW radius and ATS Geary3 E-state. A training set of 297 compounds was selected (187 %B0D<50% and 110 %BOD>50%). The remaining 375 compounds were included in the test set (252 with %BOD<50% and 123 with %BOD >50).

Table 5.10. Performance of several SBR models with descriptors selected by ReliefF

ID	SBR Model		Sensitivity	Specificity	Miss Classification Error (%)
A	Fuzzy ARTMAP	all	0.84	0.77	18.45
		train	1.00	0.99	0.34
		test	0.72	0.57	32.80
B	M5 Rule	all	0.95	0.91	**6.25**
		train	0.94	0.93	**6.40**
		test	0.96	0.89	**6.13**
C	IBK	all	0.87	0.80	15.48
		train	0.99	0.99	0.67
		test	0.78	0.63	27.20
D	SOM	All	0.91	0.39	27.23
		train	0.87	0.45	28.96
		test	0.94	0.34	25.87

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 5.10 compares the performance of the four algorithms used. The last column indicates the percentage of misclassification error for the entire data, i.e., the training and the test set. Model tree inducer's M5 outperform the rest of classifiers with about 6% of error in each data set. For FAM and IBk model sensitivities are high, around 0.85-0.90, but specificities (not biodegradable) are too low, about 0.6, as in the SBR described before. The performance of SOM is very poor since it is unable to discriminate the not ready biodegradable class (specificity around 0.40). A closer examination of the chemical space of descriptors selected by ReliefF reveals that with autocorrelation descriptors the SOM is unable to discriminate between the two biodegradation classes. Figure 5.11 shows the ROC plot that compares the performances of these four SBR models. The IBk-based SBR is located in the upper left corner, which corresponds to a good performance and a well balanced classifier. On the contrary, the SOM-based SBR is located in the opposite side of the plot, meaning that even though the performance is acceptable, the classifier is unbalanced, i.e., is unable to discriminate one of the classes.

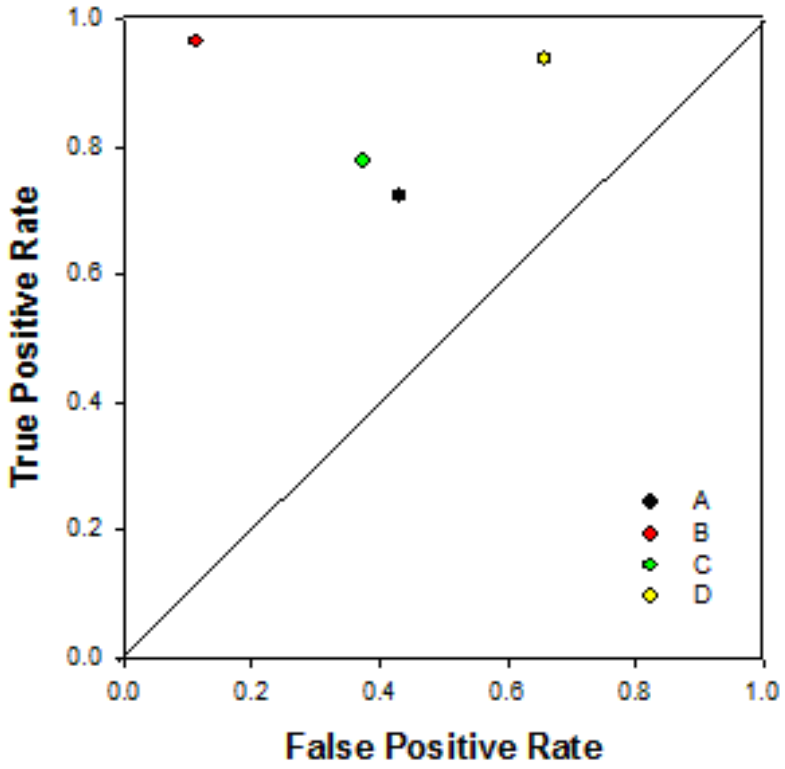


Figure 5.11. ROC plot for the test set of the SBR models developed using the descriptors selected by the ReliefF filter. (A) Fuzzy ARTMAP; (B) M5; (C) IBk; (D) SOM

ISBN:978-84-691-1008-9/DL: T.2230-2007

5.2.2 Models for Biodegradation in Soil

The second case study analyzes the modeling of biodegradation in the soil media. Models capable of estimating the biodegradation kinetics of organic compounds in soils (half lives) would be useful for deciding if an accidental pollutant spill could be naturally degraded or if specific treatment would be needed. An additional application of these models is for limiting the number of biodegradation test needed to characterize existing molecules (which are quite complex and expensive). Soil is traditionally described by the so-called *matrix variables* such as pH, organic matter concentration (from), cationic exchange capacity (CE), and particle size distributions However, matrix information is reported in very few experiments.

The soil database used to derive the biodegradation models for soil consists of 146 chemicals carefully selected from the literature. Biodegradation experiments correspond to different soil experimental conditions, which results in a total data set of 389 patterns. Of those, only a reduced subset of 96 experiments is complete with all the necessary matrix information. This scarcity of data imposes an additional difficulty in the development of accurate biodegradation models for soil.

5.2.2.1 Quantitative Models (QSBR)

A preliminary QSBR model using the complete subset of 96 experiments was built to initially assess the accuracy of the models and avoid uncertainties derived from the use of imputation techniques to recover the missing data. The feature selection model used here was again the CFS filter. Eight features were selected, three representing matrix characteristics (from, humidity and sable) and the rest related to the chemical information (dipole moment, HOMO, ionization potential, positive charged polar SA-MPEOE, and total structure connectivity index). The resulting QSBR model for soil has been assessed by internal cross-validation using a leave-one-out approach (LOO). The external validation is based in the splitting of the dataset into training and test sets with SOM. As a result, 24 chemicals were selected as training and the remaining 72 were used as a test.

Figure 5.12 shows how the eight input variables and the target biodegradation half-life are distributed in the SOM feature space. From the examination of this representation it is clear that the small half-lives of biodegradable compounds (left upper corner in the bottom right c-plane) are related to matrix properties such as moderate-high values of sable content and low humidity. In addition non-biodegradable compounds have a low ionization potential and high HOMO energy values.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

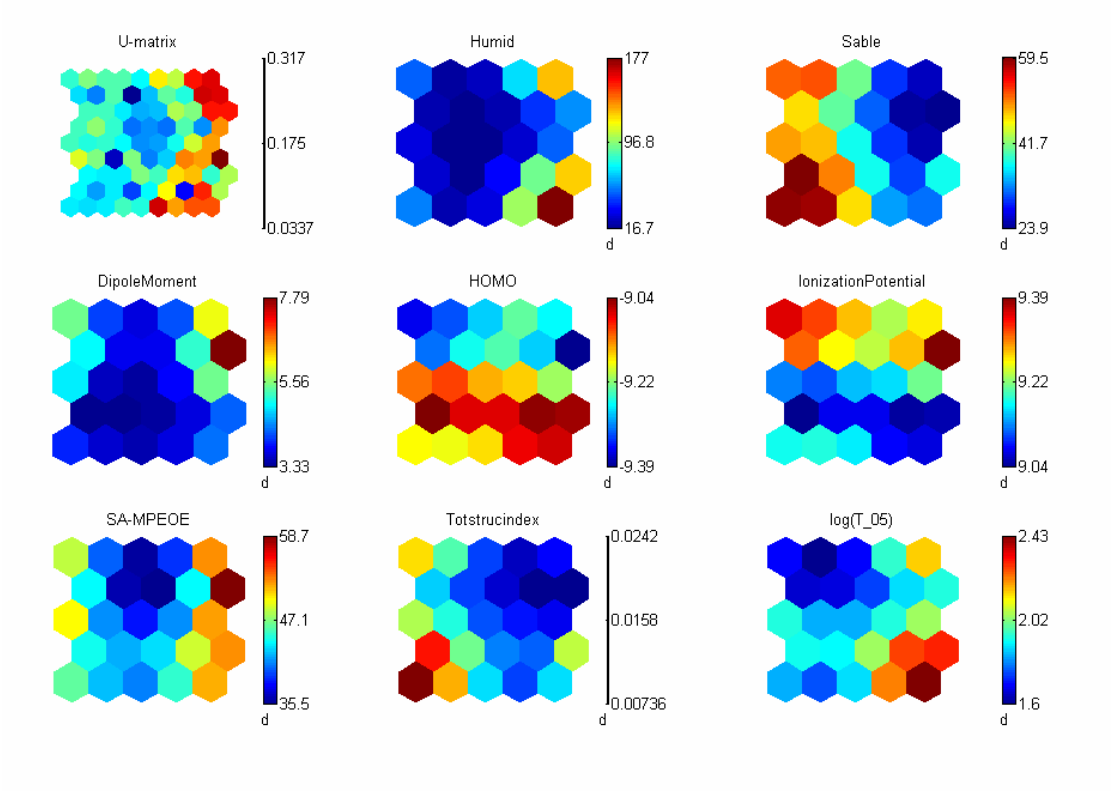


Figure 5.12. Component planes corresponding to the molecular descriptors and soil-matrix information selected by CFS for soil biodegradability half lives data set

The performance of the fuzzy ARTMAP and backpropagation (8-13-1), based QSBR model for half lives is described in Table 5.11 and Figures 5.12a-b.

Table 5.11. Summary of absolute mean errors and standard deviations for Fuzzy ARTMAP and backpropagation neural network QSBR models for soil biodegradation half lives log($T_{1/2}$)

QSBR model	All	Train	Test	LOO
	Error (σ)	Error (σ)	Error (σ)	Error (s)
Feed-forward	0.47 (0.49)	0.10 (0.14)	0.59 (0.50)	0.63 (0.53)
Fuzzy ARTMAP	0.42 (0.36)	0.19 (0.18)	0.50 (0.37)	0.67 (0.49)

It can be observed in Table 5.11 that both models have a similar performance. Both models are robust, with external and internal validation errors of the same magnitude.

ISBN:978-84-691-1008-9/DL: T.2230-2007

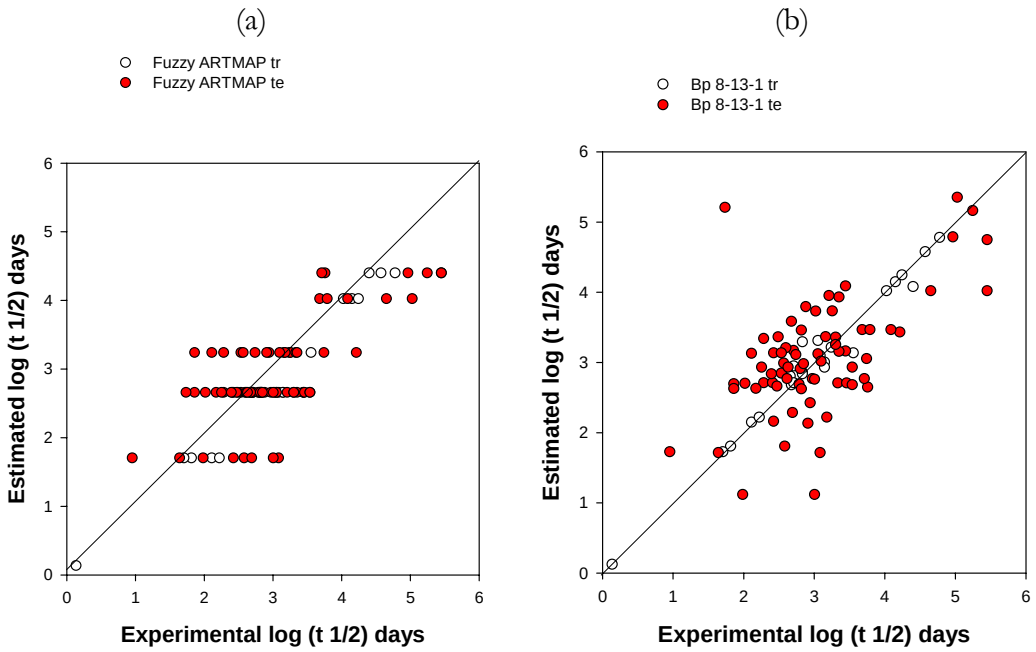


Figure 5.13. Comparison of QSBR models for soil biodegradation. (a) Fuzzy ARTMAP. (b) Backpropagation (8-13-1)

The absolute mean errors in the external validation for the logarithm of the half-life are 0.59 and 0.50 for feed-forward and Fuzzy ARTMAP based models, respectively. Despite the better average performance of Fuzzy ARTMAP, this algorithm learns a series of constant values for a few classes, as shown in Figure 5.13a. This is an indication that there the number of data points is insufficient to develop good experimental models.

There exists an effect of the environment on the biodegradation kinetics of organic compounds, as suggested by Fass et al. (1999). These authors suggest a cut-off value of 120-180 days below which environmental parameters markedly affect biodegradation half lives. A feature selection procedure was performed by splitting the soil biodegradation datasets using and average value of 140 days for the half-life. The feature selection method used was the CFS filter. Table 5.12 lists the variables selected for each half-life range.

Table 5.12 indicates that for half-lives less than 140 days, i.e., for biodegradable compounds, soil-matrix conditions such as humidity, pH and temperature, affect biodegradation rates. Above this value cut-off half-life value biodegradation rates depend solely on molecular structure. This result is in agreement with previous studies published in the literature.

QSBR models have been developed for both biodegradation ranges. The SOM based procedure to generate optimized training/test sets was used again. For the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

range of half-life>140 days 28 compounds were selected as training and 61 as test. For the low range 137 and 164 chemicals were selected as train and test respectively. Both QSBR models were built with Fuzzy ARTMAP, backpropagation, and M5 regression trees. The M5 regression tree models outperform the neural based QSBR models which show similar performances.

Table 5.12. Attributes selected by CFS for each of the half-lives ranges separated by a cut-off value of 140 days

Cut off value (days)	Descriptors
>140	Connectivity Index (order 2, standard), Dipole Moment, Difference chi 5, Positive charged polar SA – MPEOE
<140	Humid, Ph, T, HOMO, Ionization Potential, PNSA1

Table 5.13. Fuzzy ARTMAP, backpropagation, and M5 regression tree QSBR models for soil biodegradation half lives log($T_{1/2}$).

QSBR model	All	Train	Test	Half-life Range (days)
	Error (σ)	Error (σ)	Error (σ)	
Feed-forward	0.31 (0.38)	0.11 (0.17)	0.44 (0.43)	
Fuzzy ARTMAP	0.28 (0.32)	0.17 (0.24)	0.35 (0.34)	>140
M5 Tree	0.16 (0.15)	0.17 (0.16)	0.16 (0.15)	
Feed-forward	0.37 (041)	0.22 (0.23)	0.49 (0.48)	
Fuzzy ARTMAP	0.24 (0.30)	0.09 (0.12)	0.36 (0.34)	≤140
M5 Tree	0.15 (0.16)	0.16 (0.16)	0.14 (0.16)	

5.2.2.2 Qualitative Models (SBR)

Qualitative SBR models for soil biodegradation have also been developed. Dzeroski et al. (1999) define four biodegradation classes as a function of half-life values: fast degradation (half life up to 7 days), moderately fast (one to four weeks), slowly (one to six months) or resistant (more than six months). Taking this as starting point, a two-class problem can be defined as follows. A compound is considered to be degradable if its class is fast or moderate, otherwise it is considered resistant. A cut-off value of 28 days was established to differentiate between these two classes. After labeling data according to these two classes, the CFS filter was used to select the best set of descriptors. Afterwards, two machine learning based method (M5 regression tree and IBk instance based learned) and two neural-based methods (Fuzzy ARTMAP and SOM) were used to build SBR models.

The input variables selected by CFS were HOMO, Ionization Potential, Valence Connectivity Index (order 1), ACGD, Chi 3 path, FNSA2, and HRNCG. The SOM

procedure was applied to generate optimized training and test sets. The training set was formed by 50 and 58 compounds for each class; the remaining 103 and 177 were used as external validation set.

The results are shown in Table 5.14 and Figure 5.14. The IBk instance-based classifier is located in the down side of the ROC graph which means that the classifier behaves as a random guess, i.e., it can be expected to predict half of the positive and half of the negative instances correct. The M5 regresion tree algorithm has a similar behavior and performance. SOM and Fuzzy ARTMAP algorithms are located at the upper side of the ROC graph, yielding better results. The SOM classifier behaves as more conservative compared to fuzzy ARTMAP which classify nearly 80% of chemicals correctly. Table 5.14 shows that FAM gives higher misclassification errors than SOM for the test set. However, the balanced ratio between sensitivity and specificity in FAM yields classifiers with good overall performance.

Table 5.14. Summary of performance of several SBR models

ID	SBR Model		Sensitivity	Specificity	Miss Classification Error (%)
A	SOM	train	0.58	0.79	30.56
		test	0.67	0.94	16.07
B	M5 Rules	train	0.64	0.72	31.48
		test	0.56	0.48	48.93
C	IBk	train	0.59	0.44	50.00
		test	0.62	0.67	35.19
D	Fuzzy ARTMAP	train	0.82	0.73	**23.71**
		test	0.82	0.81	**18.52**

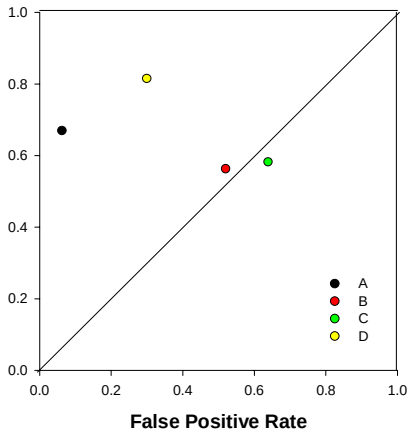


Figure 5.14. ROC curve corresponding to the four SBR models developed for biodegradation in soils. (A) SOM; (B) M5-Rules; (C) IBk; (D) Fuzzy ARTMAP

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

5.3 Conclusions

This chapter has discussed the main components that form the modeling tier. First, an overview of modeling algorithms using machine learning, neural networks, or statistical learning theories has been presented in section 5.1. A new model to construct radial basis functions coupled with a winner-takes-all approach to determine centers and widths of the Gaussian activation function has been introduced.

To illustrate the proposed modeling tier, section 5.2 completes the application of the proposed framework for modeling biodegradation rates of chemicals in different media which was started by the characterization of the chemical space given in chapter 3. Also, the preprocessing techniques proposed in chapter 4 have been applied to select the best set of features and to generate train and test sets adapted to the application domain of the studied data sets. Several feature selection methods such as CFS, ReliefF and ANNIGMA have been applied and compared with the SOM-dissimilarity method proposed in sub-section 4.2.1. The SOM-dissimilarity method yields feature subsets with higher number of variables in most simulations, which lead to models with more predictive power. The development of QSAR models corresponding to biodegradation rates in water has been assessed. The effects of the skewed data distributions detected in the previous EDA have been confirmed in the model development tier. Qualitative SBR models have been developed and tested to distinguish between biodegradable and non-biodegradable chemicals. Feature selection techniques have been applied to determine the influence of the soil matrix in biodegradation models. Results obtained are in agreement with those determined experimentally or previously published in the literature.

5.4 References

AHA, D., KIBLER, D. Instance-based learning algorithms. *Machine learning*, **6**:37-66, 1991.

ANAGNOSTOPOULOS, G.C., GEORGIOPOULOS, M. Category Regions as New Geometrical Concepts in Fuzzy ART and Fuzzy ARTMAP, *Neural Networks*, **15**(10):1205-1221, 2002.

BAUER, E. KOHAVI, R. An Emprical comparison of voting classification algorithms: Bagging, Boosting and variants. *Machine Learning*, **36**:105-139, 1999.

BERTHOLD, M.R., DIAMOND, J. Boosting the performance of RBF networks with dynamic decay adjustments, in *Advances in Neural Information Processing Systems*, G. Tesauro, D.S. Touretzky & T.K. Leen, eds., **7**:521-528, 1995.

BLUME, M., VAN BLERKOM, D. A., ESENER, S. C. Fuzzy artmap modifications for intersecting class distributions. *International Neural Network Society 1996 Annual Meeting*, WCNN'96, 250–255, NJ, USA, 1996.

BREIMAN, L. FRIEDMAN, J. STONE, C. OLSHEN, R. *Classification and Regression Trees*; Chapman & Hall: Boca Raton, FL, 1984.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, P.J. *Classification and Regression Trees*. Wadsworth International Group, 1984.

BROOMHEAD, D. S., LOWE, D. Multivariable functional interpolation and adaptive networks. *Complex Systems*, **2**:321-355,1988.

CARPENTER, G. GROSSBERG, S. ROSEN, D. B. An adaptive resonance algorithm for rapid, stable classification of analog patterns. *Proceedings of the International Joint Conference on Neural Networks (IJCNN-91)*, Piscataway, NJ: IEEE Service Center, II-411-416. Technical Report CAS/CNS-TR- 91-006, Boston, MA: Boston University, 1991.

CARPENTER, G., GROSSBERG, S., REYNOLDS, H. Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, **4**:565-588, 1991.

CARPENTER, G.A., GROSSBERG, S., MARKUZON, N., REYNOLDS, J.H., ROSEN, D.B. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*. **3**:698-713, 1992.

CLEARLY, J. G., TRIGG, L. K* An instance-based learner using an entropic distance measure. *Proc. Int. Conference on Machine learning*, Morgan Kaufmann, 1995.

DZEROSKI, S., BLOCKEEL, H., KOMPARE, B., KRAMER, S., PFAHRINGER, B., VAN LAER, W. Experiments in predicting biodegradability. *Proc. of the Ninth International Workshop on Inductive Logic Programming*, Lecture Notes in Computer Science, **1634**:80-91, 1999.

FASS, S., VOGEL, T.M., VAUDREY, H., BAUD-GRASSET, F., BLOCK, J.C. Prediction of chemicals biodegradation in soils: a tentative of modelling, *Physics and Chemistry of the Earth (B)*, **24**(6):495–499, 1999.

FERRE-GINÉ, J., RALLO, R., ARENAS, A., GIRALT, F. Identification of coherent structures in turbulent shear flows with a fuzzy ARTMAP neural network. *Int. Journal of Neural Systems*, **7**:559-568, 1996.

FISHER, R.A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7:** 179-188, 1936.

FRITZKE, B. Growing cell structure - a self-organizing networks for unsupervised and supervised learning, *Neural Networks*, **7**(9):1441-1460, 1994.

GIRALT, F., ARENAS, A., FERRE-GINÉ, J., RALLO, R, KOPP, G.A. The simulation and Interpretation of free turbulence with a cognitive neural system, *Physics of Fluids*, **12**:1826-1835, 2000.

HAMMETT, L.P. *Physical Organic Chemistry*, McGraw-Hill, 1940.

HAN, J., KAMBER, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Pub., Academic Press, 2001.

HANSCH, C., MALONEY, P.P., FUJITA, T., MUIR, R.M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, **194**:178-180, 1962.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

HOLLAND, J.H. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

HWANG, Y.S., BANG, S.Y. An Efficient method to construct a Radial Basis Function Neural Network Classifier. *Neural Networks*, **10**(8):1495-1503, 1997.

LEE, S., KIL, R.M. Multi-layer feedforward potential function network. *Proc. of the IEEE International Conference of Neural Networks*, San Diego, I, 161-171, 1988.

MCLACHLAN, G.J. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience, 2004.

MOODY, J., DARKEN, C. J. Fast learning in networks of locally-tuned processing units. *Neural computation*, **1**:281-294, 1989.

MUSAVI, M.T., AHMED, W., CHAN, K. H., FARIS, K. B., HUMMELS, D.M. On the training of radial basis function classifiers. *Neural Networks*, **5**:595-603, 1992.

PARK, J., SANDBERG, I.W. Universal approximation using radial basis function networks. Neural Computation, **3**:246-257, 1991.

PLATT, J.C. A resource-allocating network for function interpolation. *Neural Computation*, **3**(2), 213-225, 1991.

POGGIO, T., GIROSI, F. Networks for approximation and learning. *Proc. of the IEEE*, **78**:1481-1497, 1990a.

POGGIO, T., GIROSI, F. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, **247**:978-982,1990b.

POWELL, M. J. D. Radial basis functions for multivariable interpolation: a review, in *Algorithms for approximation*, J. C. Mason & M. G. Cox eds., Oxford, Clarendon Press, 143-167, 1987.

QUINLAN, J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

QUINLAN, J.R. Induction of decision trees. *Machine Learning*, **1**:81-106, 1986.

QUINLAN, J.R. Learning Logical definitions from relations. *Machine Learning*, **5**:239-266, 1990.

RALLO,.R., FERRÉ-GINÉ, J., ARENAS, A., GIRALT, F. Neural Virtual Sensor for the Inferential Prediction of Product Quality from Process Variables. *Computers and Chemical Engineering*, **26**(12):1735-1754, 2002.

RIVAS, V.M., MERELO, J. J., CASTILLO, P. A., ARENAS, M. G., CASTELLANO, J. G. Evolving RBF neural networks for time series forecasting with EvRBF. *Information Sciences*. **165**:207-220, 2004.

SCHÖLKOPF, J., SMOLA, A. *Learning with Kernels*. MIT Press, 2002.

SCHÜÜRMANN, G., APTULA A.O., KÜHNE, R., EBERT, R.W. Stepwise discrimination between Four Modes of Toxic Action of Phenols in the Tetrahymena pyriformis Assay. *Chem. Res.Toxicol.*, **16**:974-987, 2003.

SCHWENKER, F. KESTLER, H. A. PALMM, G. Three learning phases for radial basis function networks. *Neural Networks*. **14**:439-458, 2001.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

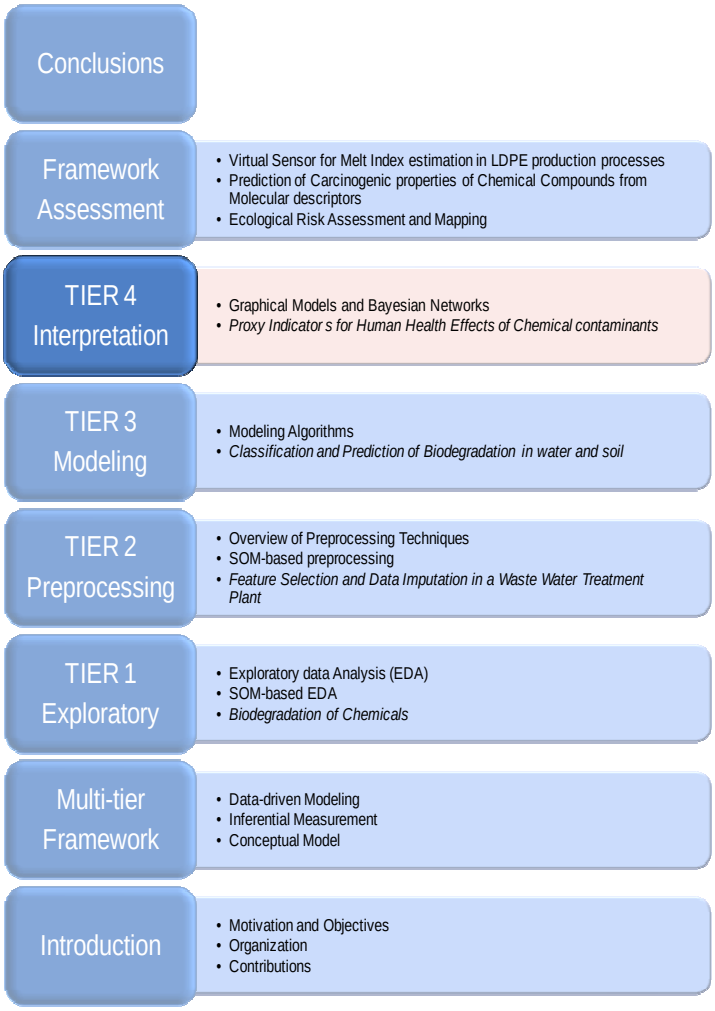
VAPNIK, V. *Statistical Learning Theory*. Wiley, 1998.

YINGWEI, L. SUNDARARAJAN, N. SARATCHANDRAN, P. A sequential learning scheme for function approximation using minimal radial basis function neural networks. *Neural Computation*, **9**:461-478, 1997.

ZADEH, L. Fuzzy sets. *Information and Control*, **8**(3):338-353, 1965.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 6

Tier 4: Interpretation

Chapter six introduces tier four which is the highest level of abstraction in the proposed framework. The interpretation layer is aimed at providing tools to establish and quantify the relationships between the input variables selected to develop the experimental models.

6.1 Graphical Models and Bayesian Networks

The analysis and interpretation of the relationships identified by machine learning algorithms could be performed from two distinct perspectives. First, rule generation systems, such as the method described in section 5.1 to extract rules from decision trees, can be used to extract knowledge from machine learning models. These algorithms use the information encoded within the internal structure of the model to generate clauses that describe, in terms of problem variables, the knowledge extracted from the modeled system. Second, Bayesian dependency models can be used to generate graphical models for the relationships between the input and target variables. In this section a brief overview of these concepts is given. Let us introduce the concept of graphical models using Jordan's definition:

"Graphical models are a marriage between probability theory and graph theory. They provide a natural tool for dealing with two problems that occur throughout applied mathematics and engineering -- uncertainty and complexity -- and in particular they are playing an increasingly important role in the design and analysis of machine learning algorithms." --- Jordan (1998).

In probability theory and statistics, a graphical model (GM) represents dependencies among random variables in a graph where each random variable is a node, and the edges between the nodes represent conditional dependencies. In its simplest case, the network structure of the model is a directed acyclic graph (DAG).

Dependence modeling means finding the model of the probabilistic dependencies of all the variables in a data set. Dependencies can be used to establish cause-effect

ISBN:978-84-691-1008-9/DL: T.2230-2007

relationships among variables. Besides revealing the structure of the domain of the data, dependency models can be used to infer probabilities of any set of variables given another different set of the same variables. Examples of such dependency models are given in Figures 6.1 and 6.2. Figure 6.1 represents the probabilistic dependencies between the Modes of toxic Action (MOA) and molecular structure parameters for a set of organic chemical contaminants. In this figure colors indicate the strength of the dependency (black: strongest dependency; cyan: weak dependency), while arrows indicate related variables. Arrow directions aren't meaningful and are used solely to compute the likelihood of the model using Bayesian inference.

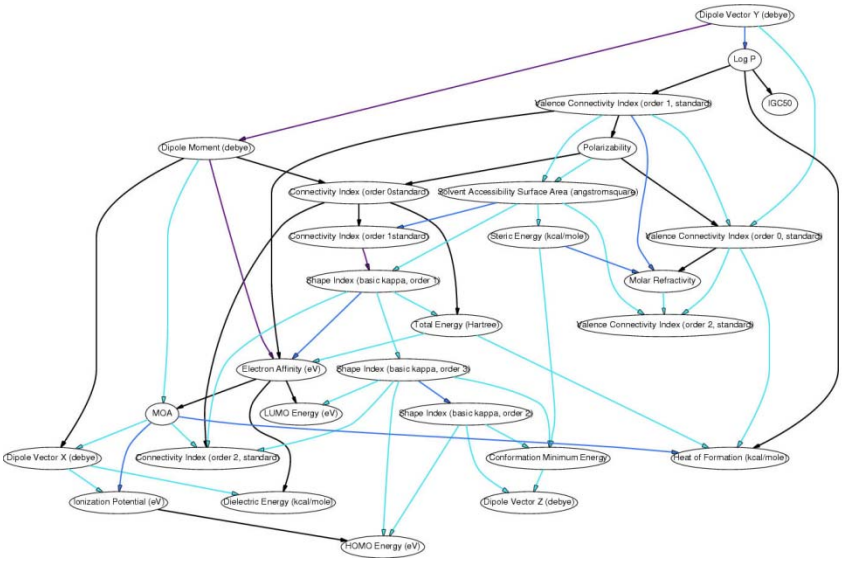


Figure 6.1. Example of a dependency model showing relationships between Modes of Toxic Action and Molecular descriptors for several organic chemicals

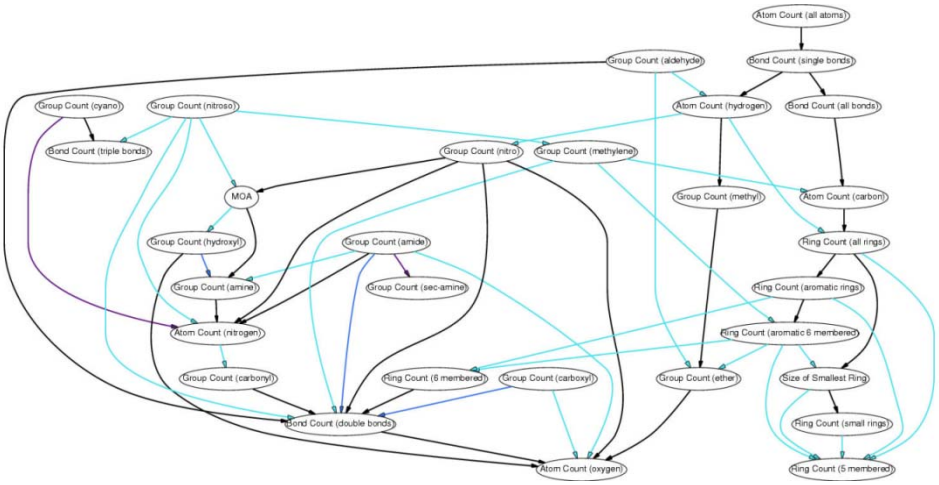


Figure 6.2. Example of a dependency model showing relationships between MOAs and functional group counts for the same set of organic chemicals as in Figure 6.1

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Bayesian reasoning provides a probabilistic approach to data-driven modeling. It is based on the assumption that the variables involved in model development are governed by probability distributions and relationships among these variables can be extracted by reasoning about these probabilities.

A Bayesian Network (BN) encodes the relationships contained in the modeled data. It can be used to describe data as well as to generate new instances of the variables with similar properties as those in the given data. A Bayesian network encodes the probability distributions of a set of variables into a directed acyclic graph (DAG). Arrow directions are meaningful in BN. Informally; an arc between two nodes relates these nodes so that the value of the variable corresponding to the ending node of the arc depends on the value of the variable corresponding to the starting node. Every probability distribution can be defined by a BN (Pearl, 1988). As a result BNs are widely used in problems where uncertainty is handled using probabilities. However, finding the optimal DAG requires searching through all possible network structures, which has proven to be a NP-hard problem (Chickering et al., 1994). Heuristic search algorithms, such as K2 (Cooper and Herskovits, 1992), are commonly applied. This algorithm performs a greedy search that trades off network complexity for accuracy over the training data to find the optimal network structure.

The complexity of the process of determining the network structure can be alleviated by using simpler approaches, such as the Naïve Bayes model. This algorithm applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function $f(x)$ can take on any value from a finite set V. A set of training examples of the target function is provided, and a new instance, described by the tuple of attribute values$\langle a_1, a_2, \dots, a_n \rangle$, is presented,. The learning algorithm is asked to predict the target value (or class) for this new instance.

The Bayesian approach to classify the new instance is to assign the most probable target value given the attributes $\langle a_1, a_2, \dots, a_n \rangle$ by using,

$$v_{MAP} = arg \max_{v_j \in V} P(v_j | a_1, a_{12}, \dots, a_n) \tag{6.1}$$

Eq. (6.1) can be rewritten using the Bayes theorem as,

$$v_{MAP} = arg \max_{v_j \in V} P(a_1, a_{12}, \dots, a_n | v_j) P(v_j) \tag{6.2}$$

so that the two terms in Eq. (6.2) can be estimated from the training data. It is easy to estimate $P(v_j)$ by simply counting the frequency by which each target value v_j occurs in the training data. However, estimating the terms $P(a_1, a_{12}, \dots, a_n | v_j)$ is not feasible unless we have a very large set of training data. The problem is that the number of these terms is equal to the number of possible instances times the number of possible target values. Therefore, we need to observe every instance in the instance space many times in order to obtain reliable estimates.

The Naïve Bayes classifier is based on the simplifying assumption that the attribute values are conditionally independent given the target value. This assumption implies that,

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

$$P(a_1, a_{12}, \ldots, a_n|v_j) = \prod_i P(a_i|v_j) \tag{6.3}$$

The formulation of the Naïve Bayes classifier can be obtained by substituting Eq. (6.3) in Eq. (6.2),

$$v_{NB} = arg \max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j) \tag{6.4}$$

where v_{NB} denotes the output of the classifier. It should be noted that in a naïve Bayes classifier the number of distinct $P(a_i|v_j)$ terms that must be estimated from the training data is the number of distinct attribute values times the number of target values. In naïve Bayes classifiers there is no need to perform an explicit search in the space of the targets to obtain estimates. Instead the response is obtained simply by counting the frequency of various data combinations within the training examples.

A simple naïve Bayes model is presented in Figure 6.3. This example relates the target output value, i.e., the MOA, with a training data set formed by 10 molecular descriptors for the organic chemicals considered above.

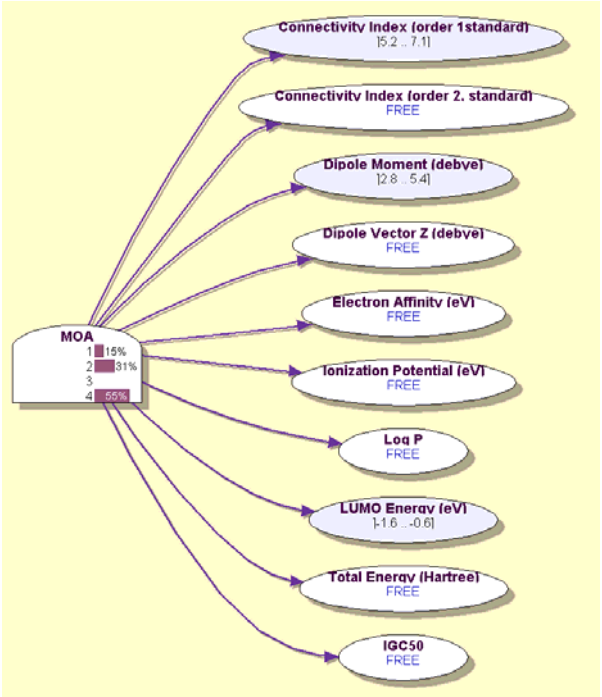


Figure 6.3. Naïve Bayes Model relating MOAs with molecular descriptors for the set of organic pollutants of Figures 6.1 and 6.2

6.2 Proxy indicators for Human Health Effects of Chemical Contaminants

This section presents the results obtained in a knowledge extraction task aimed to illustrate the use of components in tier 4. The main goal of this analysis is, first to

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

characterize the relationships between molecular structure and eco-toxicological effects of chemicals, and second to establish relationships between these environmental effects and their potential effects on human health.

Ecological risk assessment and regulatory standards are typically applied to determine the effects of single stressors on ecosystem components. However, organisms in the environment often experience many stressors simultaneously, including those of a physical, biological, and chemical nature.

It is difficult to obtain useful data sets for human health at the screening level that can help to identify the best subset of chemicals to carry out a detailed risk assessment. The reason is obviously the problem of identifying human toxicity without using expensive and problematic animal testing, including mammals, for very large data sets. Thus, it is a priori interesting to explore the possibility of applying alternative testing methodologies and to assess their level of certainty and validity. Two strategies could be considered for the current exploratory research:

- use existing data for evaluating to what extend relationships can be established between existing data sets for human toxicity and data from both simple eco-toxicological tests and molecular information;
- develop and refine simple human oriented testing procedures.

The current study focuses on the first strategy and uses information on Risk Phrases (RP) since they are the most comprehensive data set for human toxicity at the screening level. Risk phrases have been continuously developed over many years and represent a large accumulated source of expert knowledge. Thus, they will be analyzed using basic molecular descriptors for well investigated chemicals, i.e., for chemicals where the characterization using risk phrases stem from a comprehensive source of information. If it is possible to identify relationships between risk phrases and molecular descriptors it will then be possible to assess risk for many chemicals in a first hand screening procedure. These relationships can be formalized in terms of SAR (Structure Activity Relationship) models describing the relationships between the molecular structure of a chemical and a physicochemical property, environmental fate attribute, or a specific effect (end-point) on human or environmental health. It is important to note that SAR models could be developed not only for single chemicals but for consistent chemical families. The use of these models offers the possibility of performing statistical and trend analysis within the set of property values of each class consistently with the best characterization of each compound in the family. SAR models for environmental fate (biodegradation and abiotic degradation) and toxicity rely heavily on physicochemical properties of the chemical classes considered. Nevertheless, it is well known that SAR models are difficult to develop for certain chemical families. Modeling health effects by using SARs is different from the characterization of other endpoints. This is due to the multiplicity of possible scenarios (acute and chronic exposure, in vitro and in vivo tests, etc.) and the diversity of endpoints (general toxicity, organ specific effects, mutagenicity, etc.). In this situation generic SAR models are either not readily available or not well accepted by the scientific community.

ISBN:978-84-691-1008-9/DL: T.2230-2007

The structure similarity analysis assumes that similar compounds share a common Mode of Action (MOA), which is based on their common chemical reactivity. A method to enhance the accuracy of QSAR models for toxicity is to classify the compounds into sound structural families representing chemical classes, i.e., classes which are expected to have a similar toxicological behavior. However, the correct recognition of these chemical classes is a challenging task. As an alternative, QSAR models can be developed based on the concept of modes of toxic action (MOA), assuming that single models can be produced for compounds with the same mechanism of toxic action.

The main goals of this study can be summarized as:

- Generation of SAR models to predict MOAs
- Generation of QSAR toxicity models for each MOA class
- Extraction of rule sets and relationships of MOA and molecular information with Risk Phrases
- Characterization of certain human health risk scenarios in terms of MOAs and Risk phrases

6.2.1 Estimation of Modes of Toxic Action (MOA) from Molecular Structure

A variety of QSAR models for the MOAs of phenolic compounds have been reported in the literature (Ren, 2002; Schüürmann et al., 2003). Aptula et al. (2002) classified 221 phenolic compounds into four MOA groups: polar narcotics, respiratory uncouplers, pro-electrophiles and soft electrophiles. The structure-based 2D and 3D descriptors, including quantum-chemical calculations, that have been used in the current study are listed in Table 6.1.

The best subset of molecular descriptors to determine the MOA of a chemical is obtained applying the preprocessing approach presented in chapter 4. The algorithms used to select these best set of molecular descriptors are: ReliefF (Kononenko, 1994) and margin-based feature selection filters, such as SIMBA and G-flip (Gilad-Bachrad et al., 2004), the ANNIGMA (Hsu et al., 2002) wrapper, a SOM-dissimilarity filter (Rallo et al., 2005), and an ensemble approach based on majority voting (Hansen and Salamon, 1990). The simulated external validation process splits the whole dataset into two equalized complementary subsets referred as group 1 and group 2. The results are validated with the concordance, κ index, and λ_B parameter statistics (Schüürmann et al., 2003) that measure the agreement between predicted and experimental MOAs while taking into account the effect of random allocation and the predictive power of each classifier.

Two different techniques have been used to develop the classifiers. A deterministic approach which different feature selection techniques combined with a Fuzzy

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

ARTMAP classifier to estimate the MOA has first been used. Secondly, a probabilistic approach that uses dependency models to select relevant variables and Support Vector Classifiers (SVC) to classify the chemicals according to their MOA has been applied. Both approaches have been complemented with ensemble techniques to obtain more accurate MOA estimators by combining single classifiers.

Table 6.1. Pool of molecular descriptors considered

Reference number	Descriptor name	Units	Times selected
1	Toxicity	Mol/L	2
2	logK_{ow}		1
3	pK_a		3
4	N_{Hdon}		1
5	Atom count		1
6	Bond count single		2
7	Bond count double		1
8	Bond count triple		2
9	Conformation Minimum Energy		3
10	χ^0		1
11	χ^1		0
12	χ^2		0
13	μ	*Debye*	4
14	Electron Affinity	eV	3
15	Dielectric Energy	*Kcal/mole*	4
16	Steric Energy	*Kcal/mole*	2
17	Total Energy	*Hartree*	2
18	ΔH_f	*Kcal/mole*	3
19	E_{HOMO}	eV	4
20	Ionization Potential	eV	3
21	log$P$		4
22	E_{LUMO}	eV	2
23	MR		1
24	MW		1
25	κ_{A1}		0
26	κ_{A2}		3
27	κ_{A3}		2
28	Solvent accessibility Surface area	$Å^2$	3
29	χ_v^o		1
30	χ_v^1		3
31	χ_v^2		1

6.2.1.1 Deterministic Approach

Tables 6.1 and 6.2 summarize the feature selection results for MOAs. The most frequently selected molecular descriptors are the dipole moment, the dielectric energy, the E_{HOMO} and $\log P$. The filter methods yield smaller sets of descriptors (between 2 and 6) while the wrapper ANNIGMA selects 9 and 12. For the classification of group 1 compounds using those of group 2 as training, the best results obtained using filter methods are SIMBA, which misclassifies 21 compounds (19.3%), and the SOM-dissimilarity, which misclassifies 24 compounds (22%). For the external validation of group 2 the best results are obtained with SOM (22.3% misclassification). The ANNIGMA wrapper produces the lowest misclassification rates for both group 1 (20,2%) and group 2 (18.8%) validation sets. Table 6.3 summarizes the external validation results for all these feature selection methods as well as for ensembles. The ANNIGMA wrapper yields the best classification results among individual techniques.

Table 6.2. Selection of the best set of molecular descriptors for the classification of MOA

Algorithm	Model I.D.	Indices of Selected descriptors	No. of descriptors
ReliefF	m_1	18-24-28-23-3	5
SIMBA	m_2	17-28	2
G-flip linear	m_3	6-8-9-13-15-21	6
G-flip non linear	m_4	6-8-9-13-15-21	6
G-flip zero	m_5	1-13-15-19-20-26-27	7
ANNIGMA 13H-BSE	m_6	30-14-16-22-9-10-20-27 19-3-18-21	12
ANNIGMA 7H-BSE	m_7	30-31-20-14-27-26-19-18 22	9
SOM-based	m_8	14-1-7-4-3-13-19-15-2-17 21-5-30-29-28	15

Table 6.3 shows that best results are achieved using an ensemble of filters (e_1) and another of filters and wrappers (e_2), the latter being best. The ensemble e_1 misclassifies 14 phenols while in e_2 the misclassification drops to 9 (8.3%) for group 1. For group 2 both ensembles also produce the best results compared with single classifiers, with a misclassification rate of 16.1% and 13.4% for e_1 and e_2, respectively. A detailed study of these misclassification patterns reveals that for group 1 the misclassified polar narcotics are 2-methoxy-4-propenylphenol and 3-acetamidophenol. The last compound had equal voting as polar narcotic and as proelectrophile, and the first MOA was selected. The respiratory uncouplers misclassified in this group 1 are 2,4-dichloro-6-nitrophenol and 2,4,6-trinitrophenol. The misclassified proelectrophiles are bromohydroquinone, 2,3-dimethylhydroquinone, methylhydroquinone and 2-amino-4-(tert)-butylphenol. Finally, there is only one misclassification for the soft electrophile mechanism that corresponds to 2-nitroresorcinol. This compound received the same voting as

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

resporatory uncoupler, pro-electrophile or soft electrophile and the first MOA was again selected.

Table 6.3. Contingency table statistics for the best 8 Fuzzy ARTMAP based classification models for Group 1 (109) and Group 2 (112) compounds using simulated external prediction

Group 1 Train /Group 2 Test

Statistic	m_1	m_2	m_3	m_5	m_7	m_8	e_1	e_2
Concordance	0.71	0.81	0.68	0.76	0.80	0.83	1.00	**1.00**
κ	0.48	0.59	0.35	0.54	0.61	0.65	1.00	**1.00**
λ_B	0.15	0.36	0.06	0.21	0.33	0.42	1.00	**1.00**
	polar narcotics							
Sensitivity	0.72	0.88	0.82	0.84	0.86	0.89	1.00	**1.00**
Predictivity	0.92	0.86	0.84	0.93	0.94	0.94	1.00	**1.00**
	oxidative uncouplers							
Sensitivity	0.78	0.44	0.33	0.44	0.78	0.22	1.00	**1.00**
Predictivity	0.78	0.67	0.19	0.44	1.00	0.67	1.00	**1.00**
	Proelectrophiles							
Sensitivity	0.62	0.54	0.46	0.85	0.69	0.69	1.00	**1.00**
Predictivity	0.33	0.54	0.50	0.46	0.43	0.64	1.00	**1.00**
	soft electrophiles							
Sensitivity	0.64	0.91	0.27	0.36	0.55	1.00	1.00	**1.00**
Predictivity	0.44	0.83	0.43	0.57	0.50	0.55	1.00	**1.00**

Group 2 Train / Group 1 Test

Statistic	m_1	m_2	m_3	m_5	m_7	m_8	e_1	e_2
Concordance	0.75	0.76	0.56	0.71	0.81	0.79	0.88	**0.92**
κ	0.56	0.54	0.24	0.47	0.63	0.59	0.76	**0.83**
λ_B	0.31	0.26	0.09	0.11	0.40	0.34	0.63	**0.74**
	polar narcotics							
Sensitivity	0.75	0.82	0.64	0.75	0.88	0.87	0.94	**0.97**
Predictivity	0.91	0.89	0.80	0.89	0.91	0.88	0.92	**0.95**
	oxidative uncouplers							
Sensitivity	**1.00**	0.67	0.22	0.67	0.67	0.56	0.67	0.67
Predictivity	0.53	**0.67**	0.10	0.33	0.50	0.50	0.60	**0.67**
	Proelectrophiles							
Sensitivity	0.50	0.64	0.43	0.57	0.57	0.64	0.79	**0.79**
predictivity	0.39	0.45	0.30	0.47	0.57	0.64	0.85	**0.85**
	soft electrophiles							
Sensitivity	0.83	0.58	0.50	0.58	0.75	0.67	0.83	**0.92**
Predictivity	0.77	0.58	0.55	0.58	0.82	0.67	0.91	**1.00**

(*) See Table 6.2 for model identification; e_1: ensemble using models developed from filter methods; e_2: ensemble using all models (developed using filters and wrappers).

Compounds corresponding to group 2 exhibit a similar behaviour. The misclassified polar narcotics are phenol, 3,5-diiododsalicylaldehyde, bromovanillin, 2,4,6-tribromophenol and 3-hydroxybenzoicacid. The last two phenols obtained the same

voting as polar narcotic or proelectrophile, and polar narcotic, oxidative uncoupler or proelectrophile, respectively. The misclassified oxidative uncouplers are pentachlorophenol and 2,3,5,6-tetrachlorophenol. In the proelectrophiles there are five misclassified compounds, of which tetrachlorocatechol, trimethylhydroquinone and 1,2,3-trihydroxybenzene are true misclassifications while 4-acetamidophenol and methoxyhydroquinone obtain the same voting as polar narcotics. Finally, the missclassfied soft electrophiles are 2-chloro-4-nitrophenol, 4-nitro-3-(trifluoromethyl)-phenol and 2,6-dibromo-4-nitrophenol. Schüürmann et al.(2003) changed the MOA assignment for two compounds: 2,6-dibromo-4-nitrophenol is now classified as an oxidative uncoupler, and the tetrachlorocatechol has been excluded from the database due to its redox-cycling activity. If we considered these changes the misclassification rate is further reduced. The performance of ensemble e_2 is also superior to that of other methods reported in the literature (Aptula et al., 2002) when the four MOAs are considered.

Among the filter methods for feature selection presented the SOM-based feature selector produces the smallest sets of descriptors to build a MOA classifier with an acceptable accuracy. Wrapper methods produce the best results but its training complexity and size of the descriptor pool is greater than those of filters. The best results are obtained using a simple ensemble of all these classifiers that selects the appropriate MOA type by majority voting. Nevertheless a weighted voting mechanism would be necessary to solve conflicts in the assignation of MOA classes.

6.2.1.2 Probabilistic Approach

In the probabilistic approach toxicity and molecular structure data has been complemented with functional group count information. In these experiments the preprocessing step uses Bayesian dependency models to select relevant features. Following this approach the feature selection is performed by a greedy exploration of the space of possible Naïve Bayes models. The model with highest likelihood is retained and its variables form the selected subset of features. Following this approach, two subsets of 10 molecular descriptors and 12 functional group counts have been selected, respectively. Figure 6.4 shows the set of variables selected from either molecular descriptors or functional group counts. The strength of the dependences detected is depicted in Figure 6.4 using different colors. The most influential variables for models using molecular descriptors are, the ionization potential, the partition coefficient (log P), and the total energy of the molecule. For models developed from functional group count the most influential variables include nitrogen atom counts and the number of aldehyde groups.

Two MOA models have been developed from Support Vector Classifiers (SVC) using RBF kernels by using the variables (descriptors and functional group counts) identified in Figure 6.4. The performance of the resulting classifiers has been assessed using simulated external validation with the same group partitioning used in the assessment of deterministic models described above. Table 6.4 summarizes the performance of these models. The concordances of classifiers developed from molecular descriptors are 83% and 86% for group 1 and group 2, respectively. These models perform in terms of concordance better than all previous single models (m_1

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

to m_8) presented in Table 6.3. However, oxidative uncouplers and proelectrophiles are the most difficult to predict mechanisms of action, as also observed in deterministic models. The SVC models developed from functional group counts also perform better than all single deterministic models, having concordances of 88% and 84% for group1 and group2 respectively. Finally, the performance of a SVC using the 22 variables which result from the combination of both, molecular descriptors and functional group counts is presented in Table 6.4. Although the model is trained using low capacity values (C=10), the concordances obtained are similar to those obtained for previous SVC models and still better than those of single deterministic models

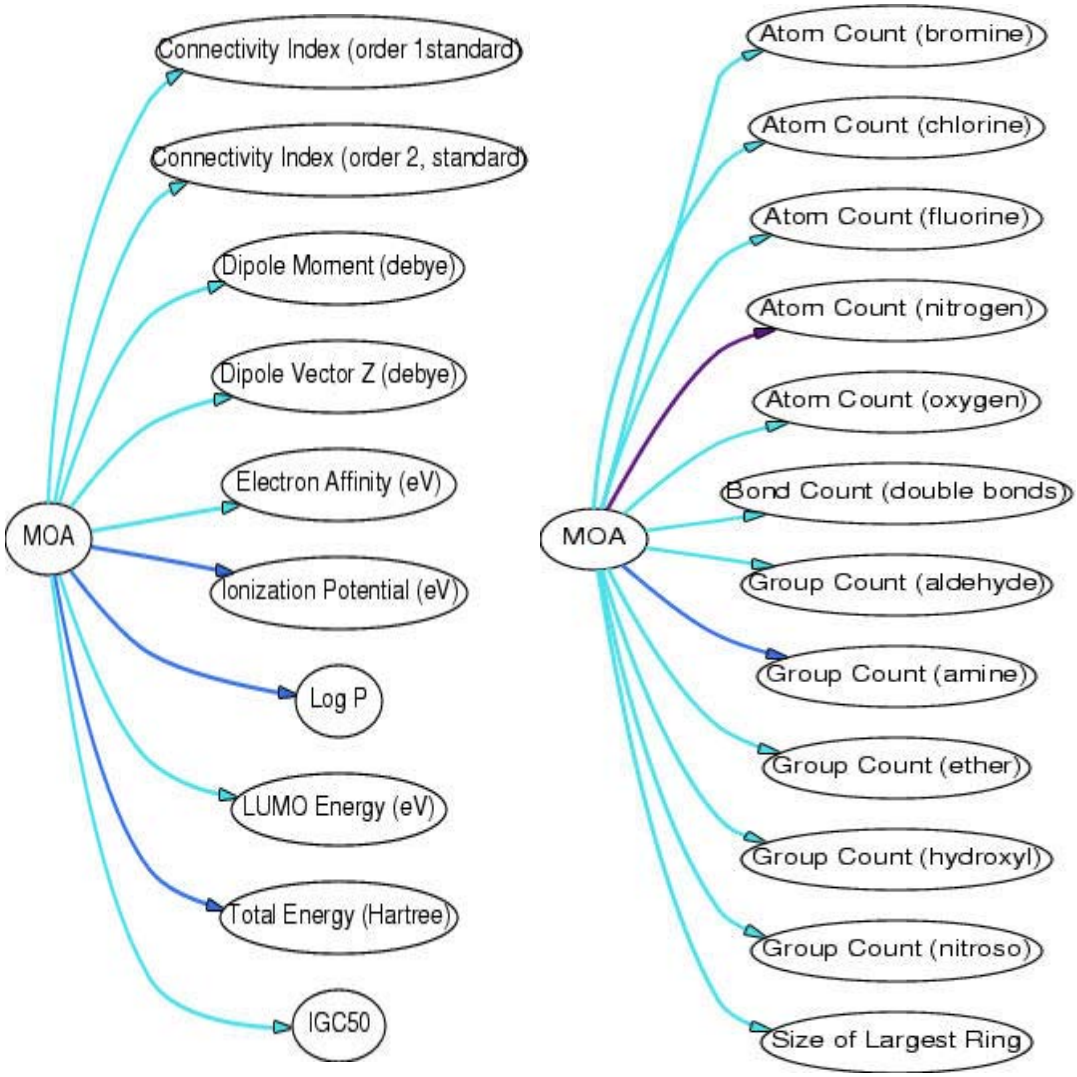


Figure 6.4. Schematic representation of a Naïve Bayes Model to classify MOAs from molecular descriptors (left); Naïve Bayes Model to classify MOAs from functional group counts (right)

ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 6.4. SVM classifiers assessed by simulated external validation and trained using molecular descriptors (a); functional group counts (b); and combination of molecular descriptors and group counts (c)

(a)

Group 1 Train / Group 2 Test	
	All Chemicals
concordance	0.83
κ	0.61
λb	0.45
	polar narcotics
sensitivity	0.97
predictivity	0.86
	oxidative uncouplers
sensitivity	0.56
predictivity	0.56
	proelectrophiles
sensitivity	0.31
predictivity	0.8
	soft electrophiles
sensitivity	0.73
predictivity	0.89

Group 2 Train / Group 1 Test	
	All Chemicals
concordance	0.86
κ	0.69
λb	0.54
	polar narcotics
sensitivity	0.97
predictivity	0.89
	oxidative uncouplers
sensitivity	0.56
predictivity	0.83
	proelectrophiles
sensitivity	0.50
predictivity	0.64
	soft electrophiles
sensitivity	0.75
predictivity	0.82

(b)

	All Chemicals
concordance	0.88
κ	0.75
λb	0.61
	polar narcotics
sensitivity	0.93
predictivity	0.91
	oxidative uncouplers
sensitivity	0.56
predictivity	0.83
	proelectrophiles
sensitivity	0.77
predictivity	0.666667
	soft electrophiles
sensitivity	0.91
predictivity	1.00

	All Chemicals
concordance	0.84
κ	0.65
λb	0.49
	polar narcotics
sensitivity	0.95
predictivity	0.88
	oxidative uncouplers
sensitivity	0.44
predictivity	1.00
	proelectrophiles
sensitivity	0.43
predictivity	0.55
	soft electrophiles
sensitivity	0.92
predictivity	0.79

(c)

	All Chemicals
concordance	0.83
κ	0.63
λb	0.45
	polar narcotics
sensitivity	0.93
predictivity	0.86
	oxidative uncouplers
sensitivity	0.67
predictivity	0.67
	proelectrophiles
sensitivity	0.46
predictivity	0.6
	soft electrophiles
sensitivity	0.73
predictivity	1.00

	All Chemicals
concordance	0.85
κ	0.66
λb	0.51
	polar narcotics
sensitivity	0.97
predictivity	0.86
	oxidative uncouplers
sensitivity	0.56
predictivity	1.00
	proelectrophiles
sensitivity	0.36
predictivity	0.63
	soft electrophiles
sensitivity	0.83
predictivity	0.83

Ensemble methods have been recently gain momentum mainly due to their ability to improve the classification capabilities of single classifiers. The main drawback of this approach resides in the fact that it is very difficult to estimate the uncertainty associated to predictions and to interpret the resulting model. A probabilistic approach based on the use of Bayesian networks has been applied in the current study to integrate and interpret the results of an ensemble of classifiers. This

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

approach integrates the uncertainty of experimental data and the uncertainty of each single model in terms of the conditional probability distribution of the whole ensemble.

Two different kinds of ensembles have been developed using the Bayesian approach. First, four SVC models have been developed to detect each single MOA. These models have a binary output indicating whether each mode of action is detected. The responses of each single MOA detector have been integrated using a Bayesian ensemble. Figure 6.5 depicts the structure of the proposed ensemble approach.

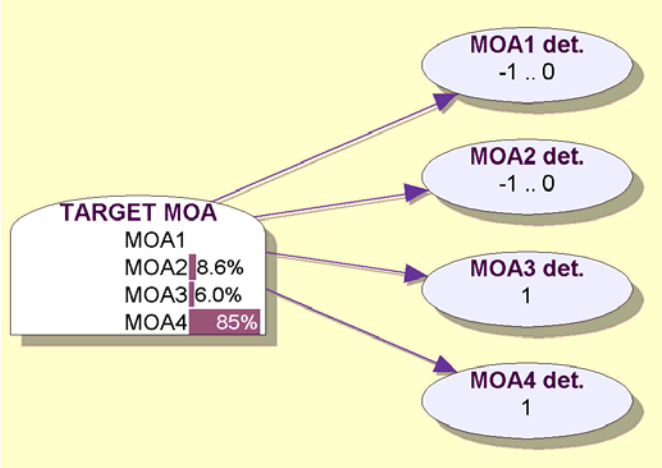


Figure 6.5. Bayesian ensemble of four SVC MOA detectors

The advantage of the approach in Figure 6.5 resides in the fact that the output of the ensemble is able to discriminate between conflicting input information. Figure 6.5 illustrates this with an example showing the simultaneous detection of both proelectrophile (MOA3) and soft-electrophile (MOA4) modes of action. The response of the ensemble uses the learned probability distributions to select the soft-electrophile MOA as the real output with a probability of 85%. The assessment of the performance of the proposed ensemble is performed in terms of internal validation using a leave-one-out (LOO) approach. Table 6.5 presents the results obtained for this model. The concordance of the model is 81%. The best classified MOA corresponds to polar narcosis which is the majority class with a 69.4%. The lowest performance corresponds to the class of oxidative uncouplers which is the MOA with less samples in the data set. The concordance obtained for this model is in average similar to the concordances reported in Table 6.3 for single deterministic models.

The three SVC models developed using different combinations of molecular and functional group information have also been integrated using this ensemble approach. To this end, a new Bayesian ensemble was trained to integrate the responses of the three models. Figure 6.6 illustrates the operation of the proposed ensemble. As in the previous case it can be observed that the probabilistic ensemble is able to deal with contradictory information. In this example the SVC models developed using molecular descriptors and functional group counts predict the mode of action as polar narcosis (MOA1). In contrast, the model developed using

ISBN:978-84-691-1008-9/DL: T.2230-2007

the combination of both molecular descriptors and functional groups predicts the oxidative uncoupler mechanism (MOA2). The integration of this contradictory information in the ensemble results in the prediction of MOA1 and MOA2 with a confidence of 52% and 48% respectively. This gives a clear indication that even though the most probable mode of action is polar narcosis there exists also a strong evidence of MOA2.

Table 6.5. Assessment of the internal performance of the Naïve Bayes classifier using LOO cross-validation

	All Chemicals
concordance	0.81
κ	0.57
λb	0.43
	polar narcotics
sensitivity	0.86
predictivity	0.97
	oxidative uncouplers
sensitivity	0.42
predictivity	0.56
	proelectrophiles
sensitivity	1.00
predictivity	0.15
	soft electrophiles
sensitivity	0.89
predictivity	0.73

The ensemble presented in Figure 6.6 is assessed following an internal validation approach using LOO cross-validation. Table 6.6 indicates that the concordance of the classifier ensemble increases until 84% with respect to the ensemble that integrates single MOA detectors. Also the recognition power for the miss-represented MOA class of oxidative uncoupler increases in both sensitivity and predictability.

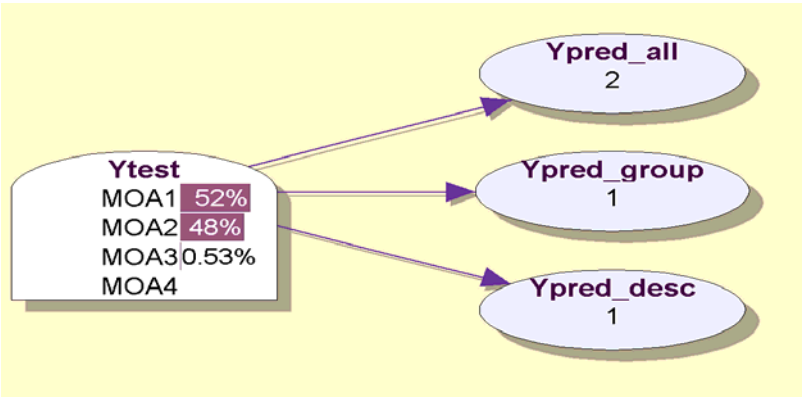


Figure 6.6. Schematic representation of an ensemble of SVM classifiers using only group counts (Ypred_group); only molecular descriptors (Ypred_desc); and all variables (Ypred_all)

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 6.6 indicates that the concordance of the classifier ensemble increased until 84% with respect to the simple Bayes model. Also the recognition power for the oxidative uncoupler MOA increases in both sensitivity and predictability. It should be noted that although the sensitivity for proelectrophile mechanism decreases its predictability increases, resulting globally in best classification performance.

Table 6.6. Assessment of the internal performance of the Naïve Bayes ensemble using LOO cross-validation

	All Chemicals
concordance	0.84
κ	0.66
λb	0.43
	polar narcotics
sensitivity	0.88
predictivity	0.93
	oxidative uncouplers
sensitivity	0.71
predictivity	0.56
	proelectrophiles
sensitivity	0.60
predictivity	0.46
	soft electrophiles
sensitivity	0.91
predictivity	0.91

6.2.2 Development of Toxicity QSAR models

Quantitative structure-activity relationship models (QSAR) have been developed for each mode of action. Two different modeling approaches have been used. First, a QSAR based on Fuzzy ARTMAP, and second a QSAR based on Support Vector Regression (SVR).

The performance and accuracy of the Fuzzy ARTMAP-based QSAR model has been assessed in terms of external validation. Figure 6.7 depicts the results obtained for the prediction of toxicity for group 1 and group 2 MOA compounds using a set of five molecular descriptors selected using the improved SOM-dissimilarity procedure. The absolute mean error for the prediction of toxicity is 0.48 for group 1 and 0.56 for group 2.

SVR-based QSAR models have been developed using either the set of molecular descriptors or the functional group counts selected by a Bayesian network. A single model to predict toxicity for all MOAs and individual models for each MOA have been developed and assessed.

ISBN:978-84-691-1008-9/DL: T.2230-2007

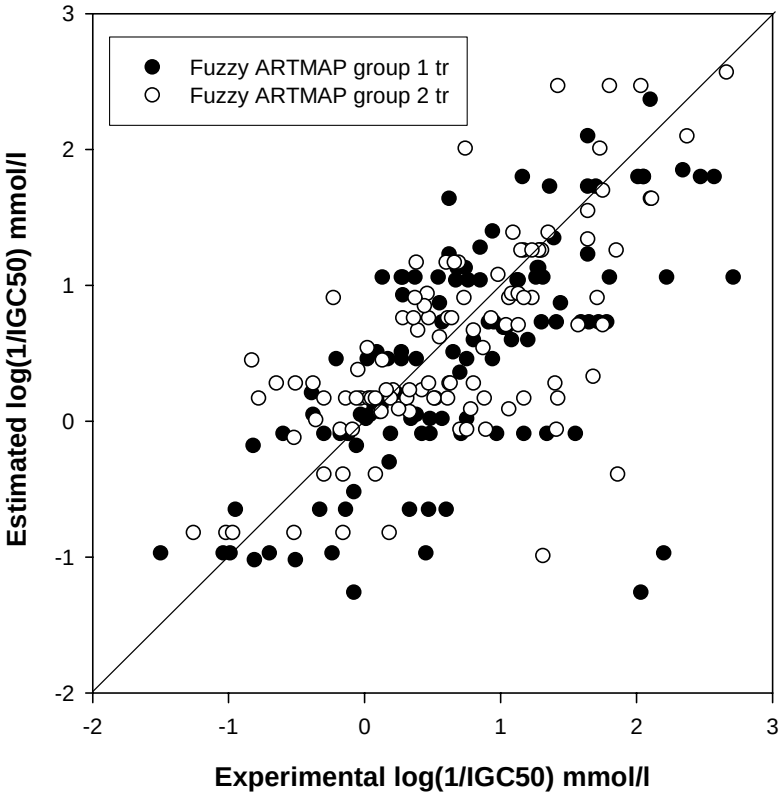


Figure 6.7. Fuzzy ARTMAP based toxicity prediction using simulated external validation techniques (using the SOM-based selection of indices)

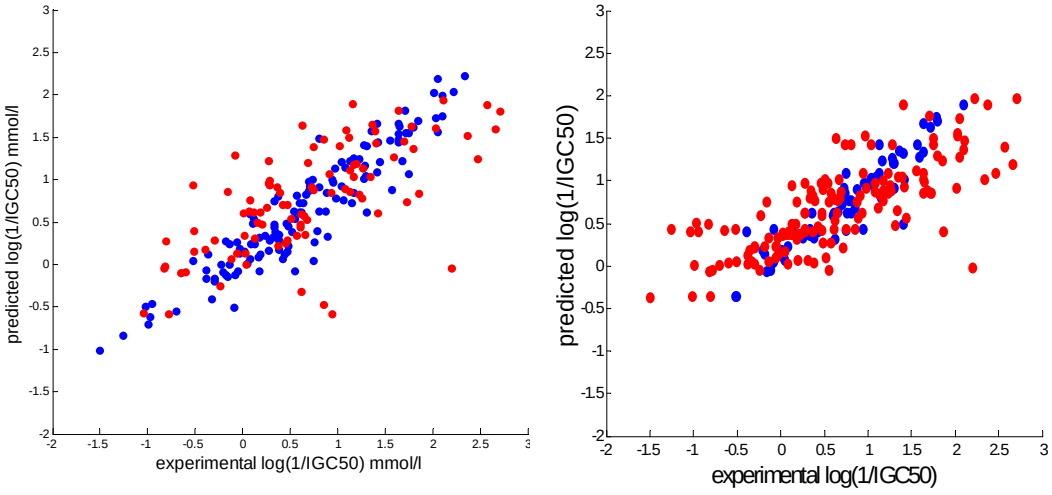


Figure 6.8. Single model for all MOAs using molecular descriptors (left); and functional group counts (right). Training data is in blue and test data in red

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The internal performance using 10-fold cross-validation for the model using molecular descriptors and for the model using functional group counts yields a mean average error of 0.38 and 0.36 respectively, which compare well with the leave-one-out cross-validation errors of 0.36 and 0.35 obtained for these two models, respectively. Finally, the mean average error corresponding to the external validation process is 0.49 for the two models. However, it should be noted that the SOM-based procedure for the generation of the train/test sets selects a training set composed by 126 chemicals for models using molecular descriptors and only 68 chemicals for models using functional group counts. This is an indication that for this data set the information provided by functional group counts is more diverse than the information contained in molecular descriptors. As a consequence, models developed from functional group counts require less training exemplars to learn the structure of the chemical space. Figure 6.8 depicts the results obtained for the external validation of the single models.

Table 6.7. Summary of the mean absolute errors (MAE) obtained for the internal and external validation of QSAR using only molecular descriptors. Train and test sets have been obtained using the SOM-dissimilarity selection technique described in chapter 5

MOA Class (tr/te)	External validation MAE	Internal 10-fold CV MAE	Internal LOO MAE
Polar Narcotics (89/62)	0.31	0.34	0.34
Oxidative Uncouplers (10/8)	0.62	0.52	0.52
Pro electrophiles (17/10)	0.66	0.36	0.47
Soft electrophiles (13/10)	0.45	0.45	0.43

The performance of models developed for each MOA from molecular descriptors is presented in Table 6.7. Internal validations using 10-fold and leave-one-out approaches, and external validation using the SOM-based partitioning of the data set, have been used to evaluate the performance of the QSAR. The best performance is achieved for the majority MOA class (i.e., the best represented class) with an average error consistent with the errors reported for single models. The low performance observed for the rest of models arises from the scarcity of training data.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Figure 6.9 presents the results obtained for the toxicity model for polar narcosis. It can be seen that the performance is similar in both models and that the results obtained are consistent with those reported for the single models for all MOA depicted in Figure 6.8.

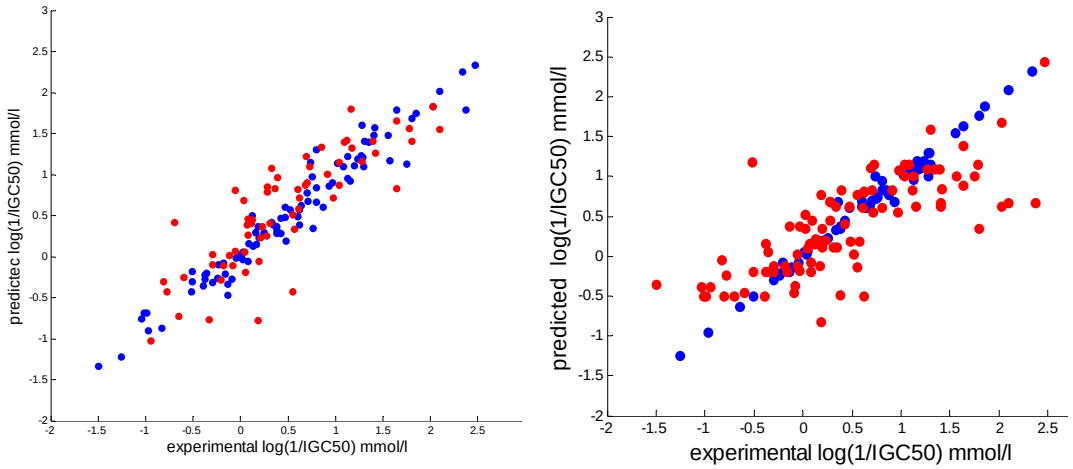


Figure 6.9. Toxicity model for all polar narcosis using (left) molecular descriptors and (right) functional group counts. Training data is in blue and test data in red

From all these sets of experiments it can be concluded that it is possible to generate classification models to estimate the mode of action from chemical structure parameters such as molecular descriptors or functional group counts. Also it has been proven that the use of ensembles increases the classification performance of individual models. Furthermore, incomplete or contradictory information can be dealt with by the use of Bayesian ensembles. The automatic detection of modes of action can be used to build QSAR models to estimate toxicity for each single MOA.

6.2.3 Knowledge Extraction

Once it has been verified that it is suitable to generate classifiers to detect the MOA from molecular structure information, we can attempt to establish the relationships between these modes of action and the potential harmful effects of chemicals in humans. Different criteria have been currently applied to each endpoint to link biological activity or Mode of Action (MOA) with risk phrases (RPs) based on a classification scheme. The hypothesis behind this analysis is that if any relationship exists between ecotoxicological endpoints, MOAs, molecular descriptors and human health related RPs, monitored effects on natural ecosystems could then be used as proxies for risk to human health. For example, an aquatic ecosystem may be used as proxy to assess health effects by oral intake. To the extent that indications of toxicity potencies for different chemical families, across species and endpoints are derived, the resulting relationships may be used as early warning indicators of contributing factors to aggregate and cumulative human health risk scenarios for related routes of exposure. The following data sources have been considered.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Test Substances and Endpoints. The source-specific fields were downloaded from EPAFJM database, which includes LC_{50} (concentration producing lethality in 50% of test animals after 96 hours exposure) in mg/l test results, and Mode of action (MOA) for 394 chemicals. The different MOAs considered are listed in Table 6.8 (Russom et al., 1997).

Table 6.8. Classification Scheme for MOAs considered

MOA	Description
NARCOSIS I	Base line narcosis
NARCOSIS II	Polar Narcosis
NARCOSIS III	Narcosis III primarily observed in esters and some acrilates
NARCOSIS I & II	Identified as both Narcosis I & II
UNCOUPLER	Uncoupler of oxidative phosphorylation
ACHE	Acetylcholinesterase inhibition
BLOCKER	Respiratory blocker/inhibitor
REACTIVE	Electrophile/proelectrophile reactivity
NEUROTOX	Central nervous system seizure/stimulant
NEURODEP	Neurodepressant
UNSURE	MOA could not be determined – insufficient evidence
MIXED	Conflicted evidence

Chemicals which do not interact with specific receptors in an organism show a mode of action called narcosis and the potency of a chemical to induce narcosis is entirely dependent on its hydrophobicity. Narcosis may be described as a process of chemicals interacting directly on the cell membrane, resulting in swelling and interference with the normal structure and functioning of the cell membrane. This type of toxicity is called non-polar narcosis or base-line toxicity. The high toxicity of the remaining chemicals depends on their hydrophobicity and their enhanced toxicity is among other factors connected with hydrogen bond donor acidity. These chemicals act by the mechanism called polar narcosis. The toxicity of reactive chemicals, e.g. electrophiles, results from their reactions with nucleophilic sites of biological molecules. Such chemicals have specific interaction with certain receptor molecules (e.g. organic phosphorous esters which inhibit acetylcholinesterase, ACHE, uncoupler, reactive) in addition to the MOA listed in Table 6.8.

Risk Classification data. The Risk Phrases, which are part of the EU Classification and Labeling system, include 59 phrases to describe different adverse effects. Some phrases describe specific effects, whereas others cover a wide range of effects. A detailed description of the testing methods required to obtain all required information for the risk phrase labeling scheme may be obtained from the EU standardized Testing Methods developed to determine the hazardous properties of chemicals. Risk phrases are listed in Table 6.9.

The first step in the current analysis has been the preliminary extraction of knowledge by the generation of descriptive rules that explain the relationships between human health indicators (risk phrases), ecotoxicological parameters (MOA)

ISBN:978-84-691-1008-9/DL: T.2230-2007

and molecular structure (molecular descriptors). The algorithm used has been the PART (Partial C4.5 Decision Trees). The resulting rule base encodes knowledge that describes the main characteristics of some subsets of chemicals (families). The behavior of these families in terms on their MOAs, RPs and descriptors can be inferred from this rule base.

The second step in the current analysis is aimed at extracting relationships between RPs by using the information provided by the molecular structure of the chemicals involved. The Correlation based Feature selection (CFS) algorithm (Hall, 1998) has been used to outline relevant relationships between risk phrases and chemical structure taking into account chemical group fingerprints (aldehyde, amide, amine, carbonyl, carboxyl, carboxylate, cyano, ether, hydroxyl, methyl, methylene, nitro, nitroso, sec-amine, sulfide, sulfone, sulfoxide, thio), and other geometrical properties (size of the molecule in terms of Ring Count like 5 member, 6 member, 7-12 member, all rings aromatic 5 member, aromatic 6 member, aromatic 7-12 member, aromatic rings, small rings). Finally, self-organizing maps (SOM) have been used to analyze the relationships between RPs and MOAs.

Table 6.9. List of Risk Phrases

R1 Explosive when dry.
R2 Risk of explosion by shock, friction, fire or other source of ignition.
R3 Extreme risk of explosion by shock, friction, fire or other sources of ignition.
R4 Forms very sensitive explosive metallic compounds.
R5 Heating may cause an explosion.
R6 Explosive with or without contact with air.
R7 May cause fire.
R8 Contact with combustible material may cause fire.
R9 Explosive when mixed with combustible material.
R10 Flammable.
R11 Highly flammable.
R12 Extremely flammable.
R13 Extremely flammable liquefied gas
R14 Reacts violently with water.
R15 Contact with water liberates extremely flammable gases.
R16 Explosive when mixed with oxidizing substances.
R17 Spontaneously flammable in air.
R18 In use may form inflammable/explosive vapor-air mixture.
R19 May form explosive peroxides.
R20 Harmful by inhalation.
R21 Harmful in contact with skin.
R22 Harmful if swallowed.
R23 Toxic by inhalation.
R24 Toxic in contact with skin.
R25 Toxic if swallowed.
R26 Very toxic by inhalation.
R27 Very toxic in contact with skin.
R28 Very toxic if swallowed.
R29 Contact with water liberates toxic gas.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

R30 Can become highly flammable in use.
R31 Contact with acids liberates toxic gas.
R32 Contact with acid liberates very toxic gas.
R33 Danger of cumulative effects.
R34 Causes burns.
R35 Causes severe burns.
R36 Irritating to eyes.
R37 Irritating to respiratory system.
R38 Irritating to skin.
R39 Danger of very serious irreversible effects.
R40 Limited evidence of a carcinogenic effect.
R41 Risk of serious damage to the eyes.
R42 May cause sensitization by inhalation.
R43 May cause sensitization by skin contact.
R44 Risk of explosion if heated under confinement.
R45 May cause cancer.
R46 May cause heritable genetic damage.
R47 May cause birth defects
R48 Danger of serious damage to health by prolonged exposure.
R49 May cause cancer by inhalation.
R50 Very toxic to aquatic organisms.
R51 Toxic to aquatic organisms.
R52 Harmful to aquatic organisms.
R53 May cause long-term adverse effects in the aquatic environment.
R54 Toxic to flora.
R55 Toxic to fauna.
R56 Toxic to soil organisms.
R57 Toxic to bees.
R58 May cause long-term adverse effects in the environment.
R59 Dangerous to the ozone layer.
R60 May impair fertility.
R61 May cause harm to the unborn child.
R62 Risk of impaired fertility.
R63 Possible risk of harm to the unborn child.
R64 May cause harm to breastfed babies.
R65 Harmful: may cause lung damage if swallowed.
R66 Repeated exposure may cause skin dryness or cracking.
R67 Vapors may cause drowsiness and dizziness.
R68 Possible risk of irreversible effects.

6.2.3.1 Generation of Descriptive Rules

Two different processes of rule generation have been carried out. Since the rule database contains information about the chemical family of each compound a first attempt to establish relationships between this classification and the MOAs, Risk Phrases and Molecular descriptors has been performed. In a second attempt, the chemical family classifications was left out of the analysis and a second set of rules

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

was obtained by relating the reported MOA for each compound with the RPs and the molecular descriptors.

Rules describing relationships between Risk Phrases, Molecular Descriptors, and MOAs in relation to Chemical Families. The idea of rule-based programming is to represent a domain expert's knowledge in the form of rules. A rule-based system consists mainly of three components: facts, rules and an engine that acts upon them. Rules represent knowledge and facts represent data. A rule-based system solves problems by applying rules on facts, i.e., matching facts with rules' if clauses. A rule consists of two parts: conditions (if clauses) and actions. The action part of a rule might assert new facts that trigger other rules. Table 6.10 summarizes the rules that relate MOAs, RPs and chemical families for the chemicals considered. We consider the presence (value 1) or absence (value 0) of functional groups or the weight of these functional groups within the molecule. The algorithm identifies threshold values and branches at these points are generated by > or < than the threshold value. The presence or absence of elements in the risk phrases are respectively identified by (Yes) or (No).

The dataset used in Table 6.10 includes 44 chemicals having LC_{50} <= 1 mg/l (R50 compounds), 63 compounds with LC_{50} between 1 and 10 mg/l (R51 compounds), and 66 compounds with LC_{50} values between 10 and 100 mg/l (R52 compounds). Chemical families are not described by the lethal fish exposure classifications as given by the R50, R51 and R52 concentration ranges, i.e., the rules listed in Table 6.10 include no RP = YES classifications explanatory for clusters of chemicals according to chemical families.

Table 6.10. Rule set generated using the PART algorithm to relate molecular information. MOA, RP and chemical families

Rule ID	Rule Description
1	GroupCount(aldehyde) > 0: **Aldehydes**
2	GroupCount(amide) > 0 AND R52 = no AND GroupCount(methyl) > 0 AND R43 = no AND AtomCount(chlorine) <= 1: **Carbamates**
3	GroupCount(sec-amine) > 0 AND GroupCount(amide) > 0 AND R52 = no: **Anilides_and_Ureas**
4	GroupCount(sec-amine) > 0 AND GroupCount(amide) <= 1 AND RingCount(5membered) <= 0 AND AtomCount(nitrogen) <= 1 AND DipoleVectorZ <= 0.153: **Secondary_aromatic_amines**
5	BondCount(doublebonds) <= 0 AND AtomCount(nitrogen) > 0 AND GroupCount(cyano) <= 0 AND GroupCount(amine) > 0 AND RingCount(6membered) <= 0: **Primary_aliphatic_amines**
6	GroupCount(carbonyl) > 0 AND MOA = NARCOSIS_I AND GroupCount(methyl) > 0: **Basic_Ketones**
7	GroupCount(amine) > 0 AND R34 = no AND LOG(LC50) <= 2.225309 AND GroupCount(hydroxyl) <= 0: **Primary_aromatic_amines**
8	BondCount(doublebonds) <= 0 AND AtomCount(nitrogen) > 0 AND

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
	GroupCount(cyano) <= 0 AND AtomCount(nitrogen) <= 1 AND GroupCount(sec-amine) <= 0: **Tertiary_aliphatic_amines**
9	BondCount(doublebonds) <= 0 AND AtomCount(nitrogen) > 0 AND R43 = yes: **Piperazines**
10	GroupCount(hydroxyl) > 0 AND BondCount(doublebonds) <= 2 AND GroupCount(methylene) > 0 AND R43 = yes: **Acrylates**
11	BondCount(doublebonds) <= 0 AND AtomCount(nitrogen) > 0 AND GroupCount(sec-amine) <= 0: **Nitriles**
12	GroupCount(hydroxyl) > 0 AND BondCount(doublebonds) <= 2 AND AtomCount(oxygen) <= 1 AND BondCount(doublebonds) <= 0 AND BondCount(triplebonds) <= 0: **Basic_Alcohols**
13	GroupCount(hydroxyl) > 0 AND SizeofLargestRing <= 0 AND R41 = no AND AtomCount(oxygen) > 1: **Diols**
14	GroupCount(hydroxyl) > 0 AND SizeofLargestRing <= 0 AND R41 = no AND BondCount(doublebonds) <= 0: **Alkyne_Alcohols**
15	GroupCount(hydroxyl) > 0 AND LOG(LC50) > 2.103804 AND R41 = no AND ElectronAffinity > -0.175: **Pyridines**
16	GroupCount(hydroxyl) > 0 AND BondCount(triplebonds) > 0: **Nitriles**
17	GroupCount(hydroxyl) > 0 AND SizeofLargestRing <= 0 AND R41 = no: **Alkene_Alcohols**
18	GroupCount(hydroxyl) > 0 AND SizeofLargestRing > 0 AND AtomCount(chlorine) <= 0 AND RingCount(smallrings) <= 2 AND AtomCount(bromine) <= 1 AND DipoleMoment > 0.79 AND BondCount(doublebonds) <= 6 AND R51 = no AND ConformationMinimumEnergy <= 18.259: **Phenols**
19	AtomCount(oxygen) <= 0 AND AtomCount(nitrogen) > 0 AND R23 = no AND R34 = no AND R52 = no AND RingCount(5membered) <= 0 AND GroupCount(sec-amine) <= 0: **Pyridines**
20	AtomCount(oxygen) <= 0 AND AtomCount(nitrogen) <= 0 AND RingCount(aromaticrings) > 0 AND AtomCount(sulphur) <= 0 AND DipoleMoment <= 0.735: **Benzenes**
21	AtomCount(oxygen) <= 0 AND AtomCount(nitrogen) <= 0 AND AtomCount(sulphur) > 0 AND AtomCount(sulphur) > 1: **Disulfides**
22	AtomCount(oxygen) <= 0 AND AtomCount(nitrogen) <= 0 AND RingCount(aromaticrings) <= 0 AND BondCount(doublebonds) <= 0 AND R65 = no AND BondCount(allbonds) <= 25: **Saturated_Hydrocarbons**
23	AtomCount(oxygen) <= 0 AND RingCount(aromatic5membered) > 0 AND AtomCount(sulphur) <= 0: **5_Membered_ring_aromatics**
24	AtomCount(oxygen) <= 0 AND R53 = no AND

ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
	GroupCount(sec-amine) > 0: **Secondary_aliphatic_amines**
25	GroupCount(hydroxyl) > 0 AND GroupCount(ether) <= 0 AND SizeofLargestRing > 0 AND RingCount(aromatic6membered) > 0 AND LOG(LC50) <= 1.659916 AND BondCount(doublebonds) <= 6 AND R23 = no AND RingCount(smallrings) <= 0: **Chlorinated_Phenols** (17.05/1.05)
26	SolventAccessibilitySurfaceArea > 249.119 AND GroupCount(amide) <= 1 AND R40 = no AND R51 = no AND ValenceConnectivityIndex(order1standard) > 7.991 AND DipoleVectorZ > -1.721: **Other_pesticides**
27	GroupCount(ether) > 0 AND RingCount(6membered) <= 1 AND R50 = no AND AtomCount(nitrogen) <= 0 AND GroupCount(methylene) <= 0 AND AtomCount(hydrogen) <= 9: **Cyclic_Ethers**
28	RingCount(aromatic6membered) > 0 AND GroupCount(ether) > 0 AND R50 = no AND RingCount(aromatic6membered) > 1: **Diphenyl_Ethers**
29	AtomCount(oxygen) > 1 AND GroupCount(methylene) <= 0 AND MOA = NARCOSIS_III: **Basic_Esters**
30	AtomCount(nitrogen) > 0 AND GroupCount(sec-amine) <= 0 AND RingCount(aromatic5membered) > 0: **Multiple_heteroatom_compounds**
31	RingCount(aromatic6membered) > 0 AND GroupCount(hydroxyl) > 0 AND HeatofFormation > -37.692: **Phenols**
32	RingCount(aromatic6membered) > 0 AND GroupCount(carbonyl) > 0 AND AtomCount(oxygen) <= 2: **Pyridines**
33	RingCount(aromatic6membered) > 0 AND ConnectivityIndex(order0standard) > 12.629 AND R22 = no: **Phthalates**
34	RingCount(aromaticrings) > 0 AND ElectronAffinity <= -0.192 AND R21 = no AND AtomCount(nitrogen) > 0: **Tertiary_aromatic_amines**
35	RingCount(aromaticrings) > 0 AND LOG(LC50) <= 1.463893 AND GroupCount(carbonyl) <= 0 AND RingCount(5membered) <= 0 AND AtomCount(carbon) <= 9 AND DipoleVectorZ <= 0: **Chlorinated_Benzenes**
36	AtomCount(nitrogen) > 0 AND GroupCount(sec-amine) <= 0 AND AtomCount(fluorine) <= 1 AND R21 = no AND GroupCount(carbonyl) <= 0 AND GroupCount(ether) <= 0 AND R22 = yes: **Amides**
37	AtomCount(oxygen) <= 1 AND AtomCount(nitrogen) > 0 AND MOA = NARCOSIS_I: **Nitriles**
38	AtomCount(oxygen) <= 1 AND GroupCount(ether) > 0: **Basic_Ethers**
39	GroupCount(carbonyl) > 0 AND R10 = no: **Basic_Ketones**
40	AtomCount(oxygen) <= 1 AND R53 = no AND R21 = no:

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
	Cyclic_Ketones
41	AtomCount(oxygen) <= 1 AND R34 = no AND AtomCount(oxygen) <= 0 AND BondCount(doublebonds) > 0 AND DipoleVectorZ > -0.001: **Unsaturated_Hydrocarbons**
42	AtomCount(oxygen) <= 1 AND R21 = no AND BondCount(doublebonds) <= 1: **Alkanes**
43	GroupCount(sec-amine) > 0: **Barbitals**
44	RingCount(aromatic6membered) > 0 AND R21 = no AND GroupCount(hydroxyl) <= 0 AND ConnectivityIndex(order2standard) <= 5.167: **Benzenes**
45	AtomCount(nitrogen) > 0 AND R21 = no: **Pyridines**
46	AtomCount(nitrogen) <= 0 AND AtomCount(phosphorus) <= 0 AND GroupCount(carboxyl) <= 0 AND RingCount(smallrings) <= 1 AND AtomCount(oxygen) > 1 AND GroupCount(methylene) > 0: **Acrylates**
47	R21 = no AND AtomCount(phosphorus) <= 0 AND GroupCount(carboxyl) <= 0 AND RingCount(smallrings) <= 1 AND AtomCount(oxygen) > 1 AND AtomCount(hydrogen) > 10: **Basic_Esters**
48	R21 = no AND GroupCount(hydroxyl) <= 0 AND AtomCount(phosphorus) <= 0 AND AtomCount(oxygen) > 1: **Carboxylic_Acids**
49	R21 = yes: **Tertiary_aliphatic_amines** (3.0/2.0)
50	GroupCount(hydroxyl) <= 0 AND AtomCount(allatoms) <= 36: **Alkenes**
51	R51 = no: **Phosphorus_compounds**
52	: **Chlorinated_Phenols**

Table 6.10 indicates that risk phrases related to the aquatic environment, R50, R51 and R52, are present in many rules (2, 3, 18, 19, 24, 26, 27, 28, 40 and 51). This is consistent with the fact that the dataset analyzed measures the ecotoxicological effects (LC_{50}) in aquatic environments for the *fathead minnow*. It can also be observed in Table 6.10 that most rules refer to "inactive" risk phrases, i.e., RP = NO *(LC_{50} values above 100 mg/l)*, which means that the chemical families referred to in the rules are characterized by not being dangerous to the aquatic environment according to the RPs. Of a total of 394 heterogeneous chemicals, 173 have LC_{50} below 100 mg/l and still grouping into chemical families does not result in any active RP classifications according to toxicity potential towards the aquatic ecosystem. No co-occurrences of LC_{50} and R50, R51 and R52, respectively, are observed in the rules according to chemical families. Based on this data set, the similar structure - similar activity hypothesis is rejected, as expected, when using chemical families as descriptors of aquatic toxicity classification according to R-phrases. MOAs appear in fewer rules (6, 29 and 37) in Table 6.8. Rules 6 and 37 relate the groups of ketones and nitriles with baseline narcosis mode of action (narcosis I). Rule 29 links esters with the narcosis III mode of action which is characteristic of esters. The ecotoxicity indicator, LC_{50}, appears in some rules (7, 15, 25 and 35); for instance rule 25

describes chlorinated phenols as chemicals with a relatively high toxicity, $\log(LC_{50}) \leq 1.66$, and without toxic effects by inhalation. Active risk phrases are only covered in Table 6.10 by rules 9, 10, 36 and 49. Rules 9 and 10 correspond to sensitization by skin contact and are used to describe piperazines and acrylates. Rule 36 defines the amides group as not harmful in contact with skin but harmful if swallowed. Finally, rule 49 defines tertiary aliphatic amines as harmful in contact with skin. Both secondary and tertiary aliphatic and aromatic amines, piperazines and acrylates have been reported to be involved in occupational asthma; the mechanism is unclear, and R37 do not appear in any of the rules in Table 6.10 (Chan-Yeung and Malo, 1994). Piperazine scored 50 and 100% positive skin sensibility test and bronchial reactions, respectively. It can be concluded from Table 6.10, that the rules referring to chemical families describe ecotoxicity and human health related patterns separately.

The following conclusions can be drawn on the use of chemical family groupings for the identification of similarity patterns between fish toxicity, risk phrases and MOAs:

- No significant occurrence of common clustering between human health effects (risk phrases) and modes of action was found from the classification of compounds into chemical families.
- The fact that no rules include both the risk phrases R50, R51 and/or R52 and LC_{50} range rules, strongly indicates that these endpoints are not significantly related to chemical families.
- Four rules have been identified to capture important chemical structural alerts related to skin allergy and to harmful effects by oral intake.

Relationships between Risk Phrases, Molecular Descriptors and MOAs. A second rule base has been created to characterize homogeneous families of compounds in terms of their modes of action. These rules, which relate the MOAs with risk phrases and molecular descriptors, are listed in Table 6.11.

Table 6.11. Rule set generated with PART to relate RPs, molecular information and MOAs

Rule ID	Rule Description
1	R28 = yes AND GroupCount(nitro) <= 1: **ACHE**
2	GroupCount(nitro) > 1 AND DipoleVectorZ > 0: **UNCOUPLER**
3	GroupCount(methylene) > 0 AND GroupCount(sec-amine) = 0 AND GroupCount(ether) = 0: **REACTIVE**
4	GroupCount(aldehyde) > 0 AND RingCount(6membered) <= 1 AND DipoleVectorZ > -0.051: **REACTIVE**
5	GroupCount(sec-amine) > 0 AND GroupCount(amide) <= 1 AND ConnectivityIndex(order1standard) <= 6.77 AND RingCount(5membered) <= 0 AND BondCount(singlebonds) > 14 AND R48 = no AND AtomCount(nitrogen) <= 1 AND GroupCount(hydroxyl) = 0 AND DipoleVectorZ <= 0.309: **NARCOSIS_I**
6	GroupCount(sec-amine) > 0 AND GroupCount(amide) <= 1 AND

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
	R48 = no AND RingCount(5membered) = 0 AND MW <= 193.288 AND BondCount(singlebonds) > 14: **UNSURE**
7	GroupCount(sec-amine) > 0 AND GroupCount(amide) <= 1 AND AtomCount(carbon) <= 10 AND DielectricEnergy <= -0.214: **NARCOSIS_I**
8	GroupCount(sec-amine) > 0 AND RingCount(aromatic6membered) = 0: **NEURODEP**
9	GroupCount(sec-amine) > 0 AND AtomCount(carbon) > 9: **ACHE**
10	ValenceConnectivityIndex(order1standard) > 8.535 AND AtomCount(phosphorus) > 0: **ACHE**
11	ConnectivityIndex(order2standard) > 9.697 AND MW > 336.3: **NEUROTOX**
12	R24 = no AND R66 = yes AND HOMOEnergy > -11.138: **NARCOSIS_I**
13	R24 = no AND R66 = yes: **NARCOSIS_III**
14	R24 = no AND GroupCount(amine) = 0 AND SolventAccessibilitySurfaceArea <= 190.642 AND BondCount(allbonds) > 22 AND AtomCount(nitrogen) > 0: **UNSURE**
15	R24 = no AND GroupCount(amine) = 0 AND R21 = yes AND ShapeIndex(basickappaorder3) <= 3.811 AND R43 = no AND GroupCount(methyl) = 0 AND DipoleVectorZ > -0.001: **NARCOSIS_II**
16	GroupCount(nitro) > 1 AND R33 = no: **REACTIVE**
17	R24 = no AND GroupCount(amine) = 0 AND RingCount(aromatic5membered) > 0 AND DipoleVectorX <= 0.092: **NARCOSIS_I**
18	R24 = no AND RingCount(aromatic5membered) = 0 AND GroupCount(amine) = 0 AND R11 = yes: **NARCOSIS_I**
19	ElectronAffinity <= -1.965 AND GroupCount(amine) > 0 AND R10 = no AND ConformationMinimumEnergy > 2.618: **UNSURE**
20	RingCount(aromatic5membered) = 0 AND GroupCount(cyano) > 0 AND R23 = no AND DipoleVectorX <= 2.082: **NARCOSIS_I**
21	GroupCount(cyano) > 0 AND R22 = no: **BLOCK**
22	RingCount(aromatic5membered) > 0: **REACTIVE**
23	AtomCount(nitrogen) > 1 AND R23 = yes AND LOG(LC50) <= 1.245513: **REACTIVE**
24	AtomCount(nitrogen) > 1 AND ConnectivityIndex(order0standard) <= 9.845 AND ShapeIndex(basickappaorder1) > 4.84 AND R22 = yes AND ConformationMinimumEnergy > -25.763 AND GroupCount(amine) >

ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
	0: **NARCOSIS_II**
25	R24 = no AND SolventAccessibilitySurfaceArea<= 185.395 AND R25 = no AND RingCount(aromatic6membered) = 0 AND AtomCount(nitrogen) <= 1 AND R40 = no AND RingCount(smallrings) = 0 AND R22 = no: **NARCOSIS_I**
26	MolarRefractivity > 42.167 AND R26 = no AND AtomCount(carbon) > 17 AND AtomCount(carbon) > 18: **UNSURE**
27	MolarRefractivity > 42.167 AND ElectronAffinity <= 0.012 AND R40 = no AND BondCount(doublebonds) > 2 AND R43 = yes AND RingCount(allrings) <= 1 AND DipoleVectorX > -0.698: **NARCOSIS_I**
28	MolarRefractivity > 42.167 AND ElectronAffinity <= 0.012 AND R43 = no AND RingCount(5membered) <= 1 AND R51 = yes: **NARCOSIS_I**
29	MolarRefractivity <= 42.167 AND GroupCount(sec-amine) = 0 AND R60 = no AND BondCount(triplebonds) > 0 AND AtomCount(nitrogen) <= 1: **REACTIVE**
30	MolarRefractivity <= 42.167 AND R60 = no AND GroupCount(sec-amine) = 0 AND RingCount(aromatic6membered) = 0 AND RingCount(allrings) <= 2 AND AtomCount(nitrogen) = 0 AND R48 = no AND AtomCount(bromine) <= 1 AND BondCount(doublebonds) = 0: **NARCOSIS_I**
31	MolarRefractivity > 42.167 AND ElectronAffinity > 0.012 AND R37 = no AND R33 = no AND GroupCount(methylene) <= 0 AND RingCount(aromatic6membered) <= 2 AND LOG(LC50) <= -0.552842 AND ShapeIndex(basickappaorder2) > 3.6: **REACTIVE**
32	MolarRefractivity > 42.167 AND ElectronAffinity > 0.012 AND R37 = no AND R33 = no AND HeatofFormation > -50.772 AND AtomCount(hydrogen) > 1 AND GroupCount(amine) <= 0 AND AtomCount(nitrogen) <= 1 AND AtomCount(sulphur) <= 1 AND ElectronAffinity <= 1.304: **NARCOSIS_I**
33	ShapeIndex(basickappaorder1) <= 8.1 AND GroupCount(sec-amine) <= 0 AND LOG(LC50) <= 0.463893 AND AtomCount(carbon) <= 6: **REACTIVE**
34	MolarRefractivity <= 42.167 AND R60 = no AND GroupCount(sec-amine) <= 0 AND ShapeIndex(basickappaorder1) <= 4.84: **MIXED**
35	MolarRefractivity <= 42.167 AND R60 = no AND GroupCount(sec-amine) <= 0 AND RingCount(6membered) = 0 AND DielectricEnergy > -0.279: **NARCOSIS_I**
36	SizeofLargestRing <= 0 AND GroupCount(ether) <= 0 AND R50 = no AND AtomCount(carbon) > 5 AND ValenceConnectivityIndex(order2standard) <= 3.027: **UNSURE**

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
37	AtomCount(chlorine) <= 3 AND SizeofSmallestRing <= 3 AND AtomCount(oxygen) <= 2 AND DielectricEnergy > -0.284: **NARCOSIS_I**
38	AtomCount(chlorine) <= 3 AND SizeofSmallestRing <= 3 AND AtomCount(oxygen) > 2: **NARCOSIS_III**
39	AtomCount(chlorine) <= 3 AND GroupCount(nitro) <= 1 AND RingCount(allrings) > 0 AND SolventAccessibilitySurfaceArea <= 188.268 AND GroupCount(sec-amine) <= 0 AND R41 = no AND AtomCount(fluorine) > 0 AND AtomCount(hydrogen) <= 5: **NARCOSIS_I**
40	AtomCount(chlorine) <= 3 AND AtomCount(fluorine) > 0: **REACTIVE**
41	RingCount(allrings) > 0 AND TotalEnergy(Hartree) > -86.137 AND RingCount(smallrings) = 0 AND RingCount(6membered) <= 1 AND RingCount(aromatic6membered) > 0 AND R41 = no AND BondCount(triplebonds) = 0 AND GroupCount(aldehyde) = 0 AND R33 = no AND GroupCount(nitro) = 0 AND DielectricEnergy > -0.143 AND AtomCount(chlorine) = 0: **NARCOSIS_I**
42	SizeofLargestRing > 0 AND TotalEnergy(Hartree) > -86.137 AND RingCount(smallrings) = 0 AND RingCount(6membered) <= 1 AND RingCount(aromatic6membered) > 0 AND LOG(LC50) > 1.245513 AND R33 = no AND BondCount(doublebonds) <= 3 AND DipoleVectorZ <= 0 AND R24 = no AND DipoleMoment > 1.885: **NARCOSIS_II**
43	SizeofLargestRing > 0 AND TotalEnergy(Hartree) > -86.137 AND RingCount(smallrings) = 0 AND RingCount(6membered) <= 1 AND RingCount(aromatic6membered) > 0 AND LOG(LC50) <= 1.245513 AND AtomCount(nitrogen) = 0 AND DipoleMoment > 1.156: **MIXED**
44	RingCount(allrings) > 0 AND HeatofFormation <= -55.531 AND DipoleVectorY <= 0.673 AND GroupCount(methyl) <= 1 AND DipoleVectorX <= 0.335: **NARCOSIS_III**
45	RingCount(allrings) = 0: **REACTIVE**
46	AtomCount(chlorine) > 3: **UNCOUPLER**
47	AtomCount(oxygen) > 3 AND R33 = no AND R22 = yes: **NARCOSIS_III**
48	GroupCount(nitro) <= 1 AND AtomCount(bromine) <= 2 AND DielectricEnergy <= -0.214 AND GroupCount(hydroxyl) > 0 AND R22 = yes: **NARCOSIS_I**
49	GroupCount(nitro) <= 1 AND AtomCount(bromine) <= 2 AND GroupCount(hydroxyl) > 0 AND RingCount(allrings) <= 1: **NARCOSIS_II**
50	GroupCount(hydroxyl) > 0 AND HOMOEnergy <= -8.896: **UNCOUPLER**

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Rule ID	Rule Description
51	GroupCount(hydroxyl) > 0: **NARCOSIS_I_and_II**
52	R33 = yes AND LOG(LC50) <= 1.841985: **NARCOSIS_I**
53	R33 = no AND R24 = yes AND AtomCount(hydrogen) > 5: **NARCOSIS_II**
54	R33 = no AND HOMOEnergy <= -8.475 AND AtomCount(bromine) <= 0 AND AtomCount(oxygen) <= 2 AND DipoleVectorZ <= 0.662 AND DipoleVectorY <= 1.652 AND R43 = no: **NARCOSIS_I**
55	R23 = no AND DipoleVectorY > 0.673 AND R22 = yes: **UNSURE**
56	ElectronAffinity <= -0.001 AND R52 = no: **NARCOSIS_I_and_II**
57	AtomCount(bromine) = 0 AND R51 = no AND ConnectivityIndex(order2standard) > 3.975 AND R22 = yes: **NARCOSIS_I**
58	TotalEnergy(Hartree) <= -69.481 AND R52 = no AND ElectronAffinity <= 1.475 AND DipoleVectorY <= 0.545: **NARCOSIS_I**
59	ShapeIndex(basickappaorder1) > 7.111 AND ElectronAffinity <= 1.475: **MIXED**
60	AtomCount(carbon) <= 7: **NARCOSIS_II**
61	: **REACTIVE**

Rules in Table 6.11 indicate that the Acetylcholinesterase inhibition mode of action (ACHE) is described by rules 1, 9 and 10. Specifically rule 1 relates high toxicity by swallowing with this MOA. In relation to rule 10 and phosphorus (phosphorus > 0), it is well-known that organophosphate pesticides binds covalently to the active site of cholinesterase. Since phosphorylated cholinesterase is stable, this results in an inhibition of the enzymatic activity and acetylcholine poisoning. Concerning rule nine, secondary amine are a bit trickier since they may be related to pharmacological undesired effects. The baseline narcosis is mainly related to RPs 11, 22, 43, 51 and 66, which respectively correspond to highly flammable, harmful if swallowed, sensitization by skin contact, toxicity to aquatic organisms and skin dryness or cracking by repeated exposure. The polar narcosis mode of action is related to harmful effects in contact with skin or by swallowing (RPs 21 and 22). The MOA narcosis type III is also related to skin dryness effects by repeated exposure. Finally, electrophile or pro-electrophile modes of action are mainly related to toxicity by inhalation exposure RPs.

We may conclude that the occurrence of active risk phrases is more consistent when grouped according to fish toxicity MOAs compared to grouping by chemical families. The simultaneous identification of RPs and molecular descriptor classification in the MOA rule base is also higher. Fish MOAs show Based in the fish toxicity dataset some tendency for relational patterns with acute toxicity risk phrases.

Analysis of relationships between Risk-Phrases. Table 6.12 summarizes the results obtained using the CFS method to detect the RPs that mostly contribute to characterize diverse human health scenarios. Ecotoxicity parameters such as MOAs

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

and LC_{50} have also been considered in the analysis. The idea is to see if what are called *exposure RPs* contain some correlation with other RPs. Individual RPs are the correlated features with each exposure RP, and the common features are represented by the intersection of these features.

Table 6.12. Relevant relationships between RPs for different health scenarios

Human Health scenario	Exposure RPs	Individual Relationships	Common Relationships
Toxic Level (inhalation; skin contact; swallow)	R23, R24, R25	R33, R39, R48	
Harmful Level (inhalation; skin contact; swallow)	R20, R21, R22	R10, R48, R59, R60, R34, R35, R41, R42, R11, R24, R25, R26, R27, R28, R37, R51, R52, R59, R63, R67	R48, R59, R60
Respiratory route: **Inhalation** (harmful; toxic; very toxic)	R20, R23, R26	R10, R21, R48, R59, R60, R24, R25, R33, R39, R28	R48
Dermal route: **Skin contact** (harmful; toxic; very toxic)	R21, R24, R27	R20, R34, R35, R41, R42, R60, R23, R25, R28	
Oral route: **Swallow** (harmful; toxic; very toxic)	R22, R25, R28	R11, R21, R24, R26, R27, R37, R51, R52, R59, R63, R67, R23	R24, R26
Skin allergies (exposure)	R24, R27	R23, R25, R28	
Skin allergies (endpoints)	R38, R43	R36, R37, R41, R48, R65, R68, R11, R20, R25, R27, R33, R37, R42, R49, R61, R67	R37, R48
Asthma/respiratory diseases (exposure)	R23, R26	R24, R25, R33, R39, R48	
Asthma/respiratory diseases (endpoints)	R37, R42	R10, R19, R36, R63, R10 R11, R34	
Cancer (endpoints)	45, 46, 49	R12, R60, R62, R68, R12 R20, R24, R25, R26, R27, R28, R44, R51, R53, R61	
Ecotoxicity (LC50)		R22, R38, R40, R62	

The main relationships among RPs that can be extracted from the CFS analysis summarized in Table 6.10 are: (i) Risk phrases describing toxicity effects by different routes (first and second row) are mainly related to RPs referred to cumulative irreversible and long term effects; (ii) the analysis of a single exposure scenario such as inhalation reveals a close relation with skin and oral intake evidences and with long term effects. In addition, there exists a relation with other endpoints such as danger for the ozone layer and fertility related problems. Similar relationships can be extracted for dermal and oral intake; (iii) relationships between specific endpoints, such as for instance those related with respiratory diseases (R37, R42), indicate that this kind of alert is present when characteristics like flammable or explosive (RPs 10, 11 and 19), and consequences such as eyes irritation (RP36), burns (RP34), and harmful effects to the unborn child (RP63), are also present; (iv) a correlation with

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

RPs for the carcinogenicity related criteria referring to respiratory and oral exposure routes is found, as expected. Relationships with long term effects for the aquatic environment are also observed. Concerning human health endpoints relationships with infertility, harm to the unborn child, and possible risk of irreversible effects are observed; (v) ecotoxicity is mainly related to oral exposure and the effects are skin irritation, evidence of carcinogenic effects and risk of impaired fertility.

6.2.3.2 Analysis of Relationships using SOM-based EDA

A combined analysis of the relationships between risk-phrases and MOA has been carried out with binary encoding. An 11-dimensional binary vector has been appended to the input data to characterize the modes of action of each chemical. From this encoding, relationships between risk-phrases and MOAs can be established and visualized by using self-organizing maps. Figure 6.10 depicts the distribution of each single variable in the output space spanned by the SOM.

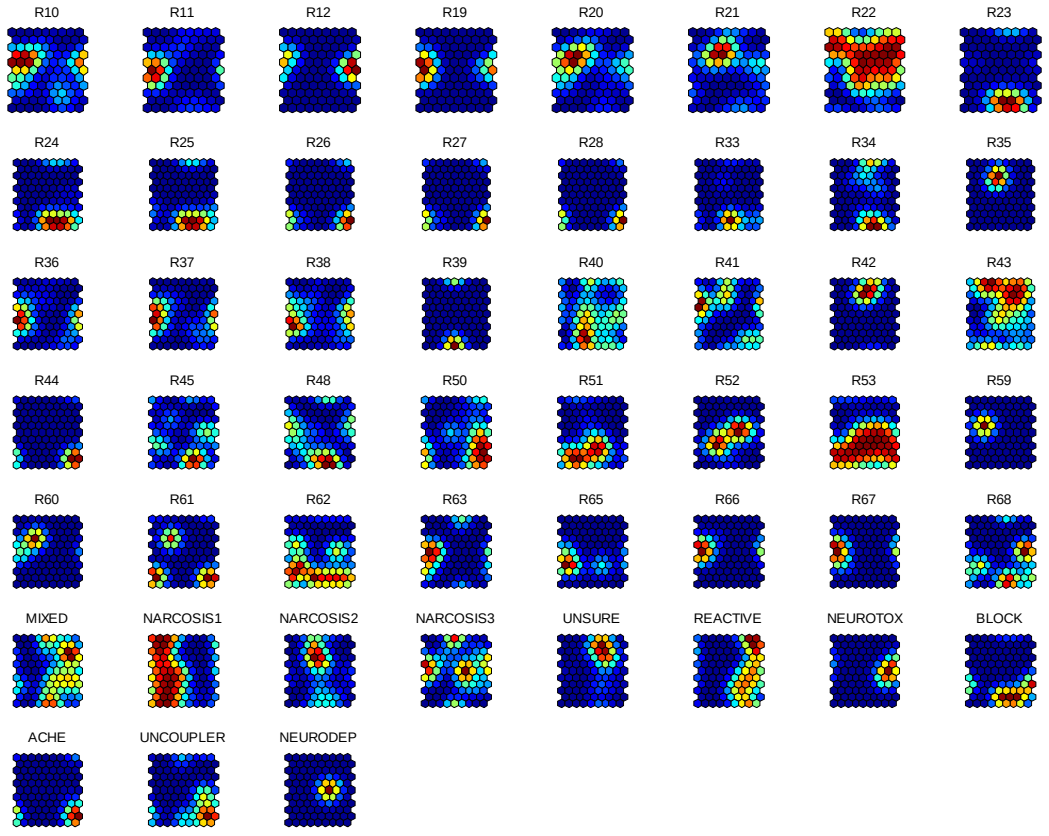


Figure 6.10. C-planes representing the complete set of risk phrases and MOA types for studied substances

The distribution of compounds having similar behavior in terms of RP and MOA in Figure 6.10 can be used to characterize homogeneous clusters of chemicals (families of chemicals). It can be seen for instance that R23, R24 and R25 group chemicals that are located in the lower section of the map (red spot). From the interpretation of these risk phrases it can be seen that their group contains chemicals that are toxic

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

either by inhalation, contact or swallow. From the examination of the c-planes corresponding to the MOAs it can be seen that chemicals with R23, R24 or R25 have a similar space distribution with BLOCK mechanisms. The comparison between c-planes has been carried out by using a clustering approach also based on self organization.

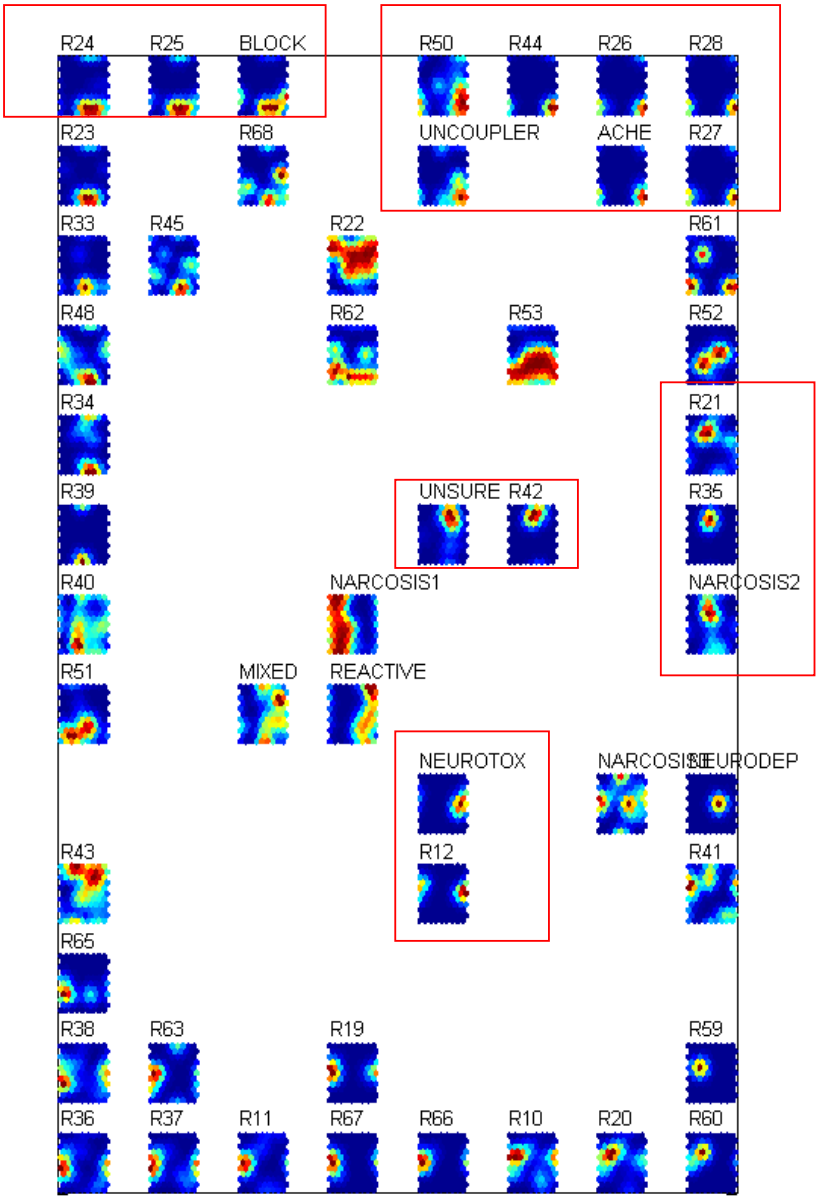


Figure 6.11. Similarities between c-planes corresponding to Risk-Phrases and MOAs

Figure 6.11 depicts clustering between input variables. From this clustering it can be seen that R23, R24, R25 and R68 are mainly related with the BLOCKER mode of action which is mainly associated to respiratory alterations. This is compatible with

the effect of R23 exposure which accounts for toxicity by inhalation. Figure 6.11 also illustrates that R50 and R44 are closely related with the uncoupler of the oxidative phosphorylation modes of action. R26, R27 and R28 are related with ACHE inhibition and account for very toxic substances, either by inhalation, contact or swallowing. Polar narcosis (narcosis II) is related to chemicals with R25 active, which corresponds to substances that can cause severe burns. Finally, neurotoxic activity is related to R12, which accounts for extremely flammable substances.

6.2.3.3 Analysis of Risk Scenarios

An alternative analysis of the relationships between RPs and MOAs can be performed by defining generic risk scenarios and visualizing the distribution of the compounds that belong to the scenario, either by contributing to exposure or endpoint RPs or both. Figure 6.12 represents the projection of the skin allergies scenario over the SOM space. The exposure RPs related to this scenario cover compounds that are located in the lower right area of the map. Inspection of c-planes plotted in Figure 6.10 reveals that this area is mainly related to REACTIVE, BLOCKER and UNCOUPLER modes of action.

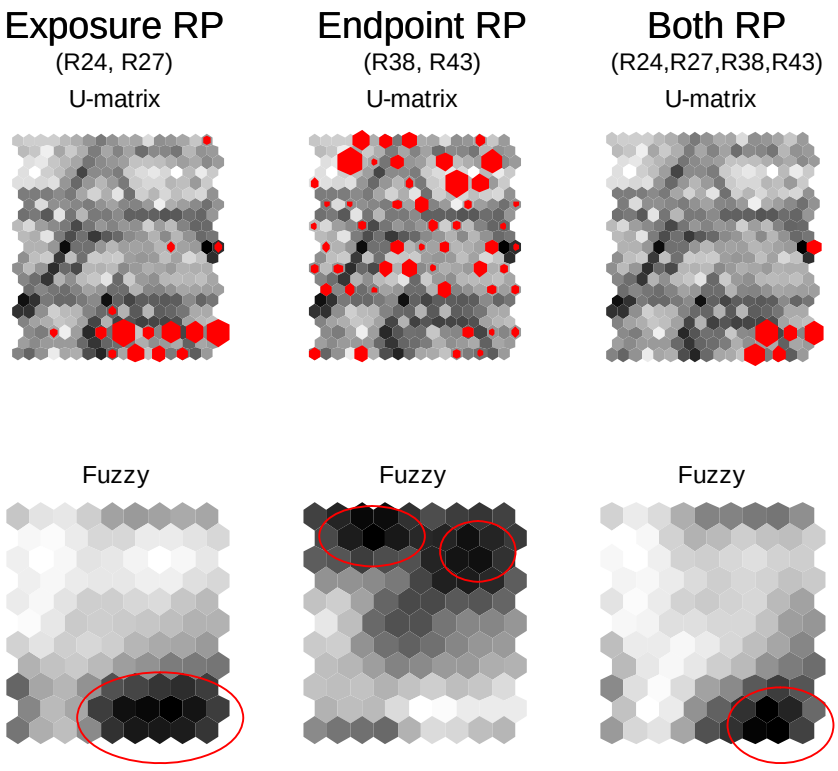


Figure 6.12. Skin allergies scenario

Risk phrases related to endpoints are located in the opposite side of the map. These RPs are mainly related to UNSURE and NARCOSIS mechanisms. Finally, chemicals having simultaneously both types of RPs active are located in the lower right area and are related mainly to REACTIVE and BLOCKER mechanisms.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

A similar analysis can be performed for other scenarios. Figure 6.13 shows the asthma/respiratory diseases and cancer by inhalation scenarios. In these scenarios the exposure route is the same (respiratory). The main area of activity for exposure is located in the lower center-right area of the map and corresponds to chemicals having REACTIVE, BLOCK and UNCOUPLER modes of action.

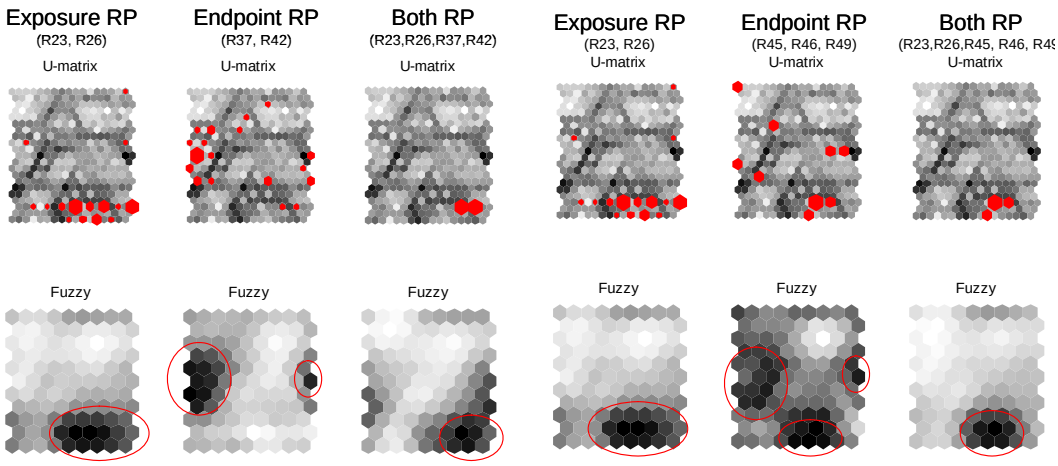


Figure 6.13. Respiratory intake related scenarios: asthma/respiratory diseases (left); cancer by inhalation (right)

Endpoints have also a similar structure in both scenarios. For asthma, the activity area is located in the left side of the map and corresponds to chemicals with baseline narcosis (narcosis type I) and narcosis type III modes of action, as seen in Figure 6.12. For cancer by inhalation, the projection of RPs describing the endpoints shows an additional activity area in the same location as the exposure RPs. The simultaneous activation of both groups of RPs affects in different ways the two scenarios. In the case of asthma and respiratory diseases the main activity is located in the lower-right area which corresponds to chemicals having NEUROTOX, BLOCKER, ACHE or UNCOUPLER modes of action. For the cancer by inhalation scenario the activation area is concentrated in the lower area of the map and mainly corresponds to chemical families with MIXED, REACTIVE or BLOCKER modes of action.

6.2.4 Summary

The characterization of chemical families based on modes of action for the fathead minnow derived from acute LC_{50} toxicity data has been carried out. The study has also examined the effects of chemicals on human health and their relationship with risk-phrases and molecular structure. A complete screening by means of knowledge extraction methodologies has been performed, including the application of rule generation using the PART algorithm, the extraction of relevant relationships using correlation based feature selection techniques, and the classification of chemicals using self-organizing maps. The information extracted from each single method is

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

compatible and completes the information extracted from the other techniques. This can be illustrated by the examination of the acetylcholinesterase inhibitor (ACHE) mode of action which is described by rules 1, 9 and 10 in Table 6.11. Rule 1 relates this mechanism to chemicals with high toxicity by swallowing (R28). This affirmation is confirmed by the examination of the clustering of the self-organized maps shown in Figure 6.12. This figure indicates that ACHE is very close to R28. The re-examination of rules 1, 9 and 10 in Tables 6.10 and 6.11 completes the information about this mode of action indicating the possible presence of a group nitro, a secondary amine or phosphorous atoms. It also indicates that it occurs when having more than 9 carbon atoms and a valence connectivity index of order 1 greater than 8.5.

The analysis of asthma / respiratory diseases scenario shows that this mode of action is also compatible with the simultaneous activation of its exposure and endpoint RPs. In fact, Figure 6.12 shows that ACHE is very close to chemicals that are very toxic by inhalation. Rules 14, 24 and 53 in Table 6.11 reveal that the polar narcosis (NARCOSIS II) mode of action is related to harmful effects either in contact with skin (R21, R24) or by swallowing (R22). This is confirmed in Figure 6.12 where the c-plane corresponding to polar narcosis is close to R21. Furthermore, from this representation a relationship of this MOA with R35 (possibility of severe burns) is observed. The relationship of R21 and R22 with R35 is confirmed by the CFS algorithm in Table 6.12 (harmful level). The detailed examination of rules 15 and 24 add additional information about this MOA. From rule 15 it can be seen that the effects on skin of chemicals having this mode of action is not severe because they aren't toxic in contact with skin and don't cause sensitization by skin contact. This affirmation seems to contradict rule 53 which states that polar narcotics are toxic in contact with skin. In fact both affirmations can be compatible; the detailed examination of c-planes in figure 6.11 reveal that the lighter activity area (light blue) located in the lower region of the plane is compatible with the activation of R24. It should be noted that rule 53 covers only 4 of 35 polar narcotics indicating a possible group of chemicals having a slightly different behavior.

The uncoupler of oxidative phosphorylation (UNCOUPLER) mode of action is mainly related to R50 which indicates high toxicity to aquatic organisms and to R44 which states the risk of explosion under confinement. Also Figure 6.12 indicates that uncoupler activity is related to high toxicity (R26, R27 and R28) for different exposure routes and is also related to acetylcholinesterase inhibition MOA. Examination of Table 6.11 shows that chemicals having uncoupler mode of action are grouped into three families described by rules 2, 46 and 50. None of these rules contains information related to risk phrases; instead the description is only based on molecular properties (presence of group nitro and hydroxyl, presence of chlorine atoms, component Z of the dipole vector, and HOMO energy less than -9 eV). This MOA is also present in the skin allergies and asthma scenarios and is related to the simultaneous activation of exposure and endpoint RPs. Similar relationships can be extracted for the rest of MOAs for this dataset.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

6.3 Conclusions

This chapter has introduced a probabilistic approach to the interpretation of information contained in models. In the first section, a short overview of dependency modeling and Bayesian networks was given. The relationships between Modes of Action, Molecular information and human health have been analyzed in the second part of this chapter. Classification models to detect the MOA have been implemented and tested using diverse approaches such as SOM, SVM and Bayesian ensemble techniques. In addition quantitative models of toxicity have been developed and assessed for each single MOA.

Current analysis has uncovered the well-known complexity of overlapping mechanisms of toxicity, chemical classes and toxic effects. All three endpoint scenarios considered are related to MOAs reactive chemicals. Local acute effects are partially related to the same MOAs, while the effects are different, e.g., inhalation of irritant gas or skin contact with a corrosive agent. The analysis supports the idea that single chemicals use multiple mechanisms of action depending on whether the effect relates to local or to systemic toxicity.

6.4 References

APTULA A. O., NETZEVA T. I., VALKOVA I. V., CRONIN M. T. D., SCHULTZ T. W. D., KÜHNE R. , SCHÜÜRMANN G. Multivariate Discrimination between Modes of Toxic Action of Phenols. *Quant. Struct.-Act. Relat.* **21**:12-22, 2002.

CHAN-YEUNG M., MALO J.L. Aetiological agents in occupational asthma. *Eur. Respir. J.* **7**:346-371, 1994.

CHICKERING, D.M., GEIGER, D., HECKERMAN, D.E. Learning Bayesian Networks is NP-Hard, *Microsoft Research Technical Report*, MSR-TR-94-17, 1994.

COOPER, G.F., HERSKOVITS, E. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, **9**:309-347, 1992.

COWELL, R.G., DAWID, A.P., LAURITZEN, S.L., SPIEGELHALTER, D.J. *Probabilistic Networks and Expert Systems*. Springer-Verlag. 1999.

F. V. JENSEN. *Bayesian Networks and Decision Graphs*. Springer. 2001.

GILAD-BACHRAD, R. NAVOT, A. TISHBY, N. Margin Based Feature Selection – Theory and Algorithms. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

HANSEN, L.K., SALAMON, P. Neural network ensembles, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **12**(10): 993-1001, 1990.

HSU, CH., HUANG, H., SCHUSCHEL, D. The ANNIGMA-wrapper approach to fast feature selection for neural nets. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, **32**:207-212, 2002.

JORDAN, M.I. *Learning in Graphical Models*. MIT Press. 1998.

ISBN:978-84-691-1008-9/DL: T.2230-2007

KONONENKO, I. Estimating Attributes: Analysis and Extensions of Relief. *Proc. European Conference on Machine Learning*, 171-182, 1994.

LAURITZEN, S. *Graphical Models*, Oxford. 1996.

PEARL, J *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 1988.

PEARL, J. *Causality*. Cambridge. 2000.

REN, S. Determining the mechanisms of Toxic Action of Phenols to Tetrahymena pyriformis. *Environ Toxicol.* **17**(2),119-127, 2002.

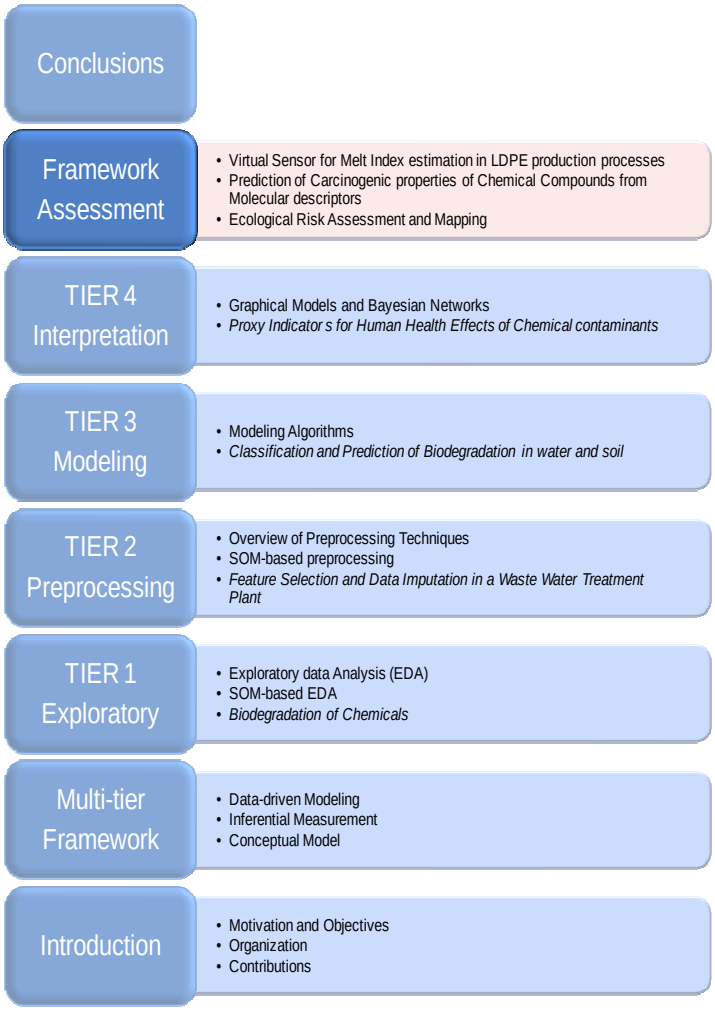
RUSSOM,C.L., BRADBURY, S.P., BRODERIUS, S.J. Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (Pimephales Promelas). *Environmental Toxicology and Chemistry*, **16**(5):948-967, 1997.

SCHÜÜRMANN G., APTULA A. O., KÜHNE R., EBERT, R-U. Stepwise Discrimination between four modes of Toxic action of Phenols in the *Tetrahymena pyriformis* Assay. *Chem. Res. Toxicol.*, **16**:974-987, 2003.

STEWARD, J. J. P. Optimization of parameters for semiempirical methods. I Methods; II. Applications. *J. Comput. Chem.* **10**:209-220;221–264, 1989.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 7

Framework Assessment

In this last chapter the framework proposed in chapter 2 is assessed in three different application domains of engineering and scientific interest. In the first case study a virtual (software) sensor system is developed for an industrial process using real plant data. The second case study deals with data-driven modeling to develop a QSAR model to predict carcinogenicity for aromatic chemicals with nitrogen containing substituent. Finally, the third case study applies the proposed framework to risk assessment in environmental modeling problems.

7.1. Virtual Sensor for Melt Index estimation in LDPE production processes

The application of the proposed framework to inferential measurement is assessed by the development of a virtual sensor to estimate on-line the melt index (MI), which is a quality indicator used in low density polyethylene (LDPE) production plants, from 25 process variables (pressures, flow rates, temperatures of the cooling/heating streams of the reactor, etc.) measured continuously. The characterization and prediction of the time-variation of this index have been approached using the different tiers of the proposed framework.

7.1.1 Problem statement

The polymerization of ethylene to produce Low Density Polyethylene (LDPE) is usually carried out in tubular reactors 800-1500 m long and 20-50 mm in diameter at pressures of 600-3000 atmospheres (Chan et al., 1993; Lines and Hartlen, 1993). The quality of the polymer produced is determined essentially by the Melt Index (MI), which is measured by the flow rate of polymer through a die. The on-line measurement of this quantity is difficult and requires close human intervention because the extrusion die often fouls and blocks. As a result, the MI is evaluated in most plants off-line with an analytical procedure that takes between 2 to 4 hours to

complete in the laboratory, leaving the process without any real-time quality indicator during this period. Consequently, a model for estimating the MI on-line would be very useful both as an on-line sensor and as a forecasting system. In addition, it would allow the supervision of the overall process and to avoid any mismatch of product quality during product grade transitions, i.e., changes in the properties of the polymer produced. However, a model derived from first principles, capable to continuously predict accurate MI values for any LDPE process in real time, is still non-existent. Instead, some production plants use data based, linear and non-linear correlation models to overcome this drawback.

The aim of the current case study is to implement a virtual sensor to predict the MI of LDPE from the state of a process plant. The virtual sensor will behave as a black-box model that relates the MI of produced LDPE to other process variables measured at the time when the ethylene monomer is fed into the plant, i.e., accounting for the plant residence time or a lag-time τ. It is important to remark that the current approach is not a time-series analysis, since it is time independent. The functionality between the output and the input variables is of the form $MI(t+\tau)=\Phi[v_1(t),...,v_n(t)]$, where each $v_i(t)$ is any of the process variables listed in Table 1, excluding MI, measured at time t. Nevertheless, since field measurements are available as a time series the current neural sensor can also be developed and operated according to production time-sequences and cycles.

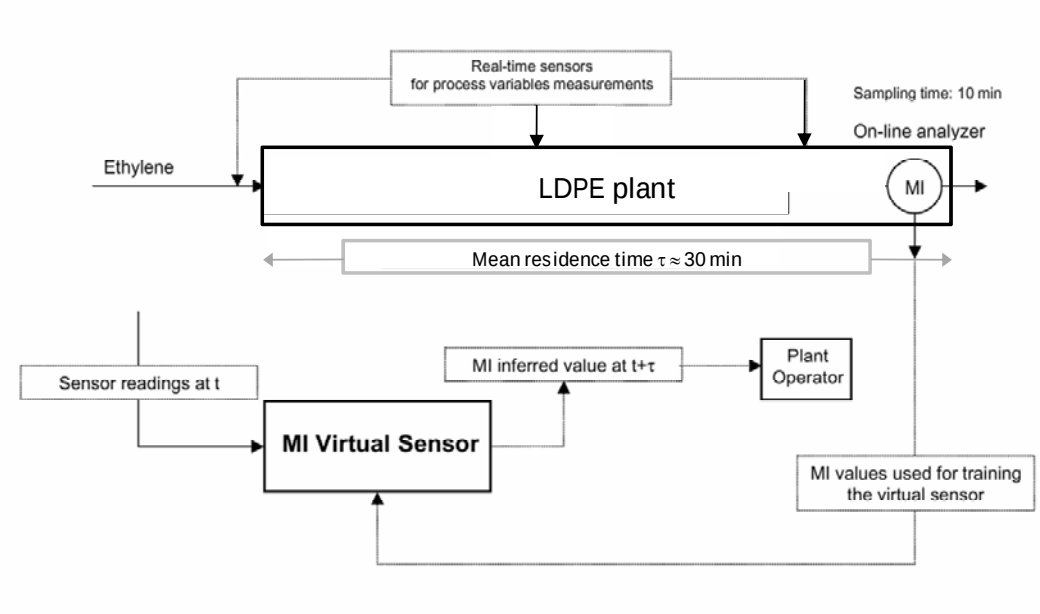


Figure 7.1. LDPE plant diagram with typical time scales

Figure 7.1 summarizes the time scales of the different elements of a LDPE production plant. The MI data used in this analysis correspond to time intervals of 10 minutes. The virtual sensor has to anticipate the MI values from process variables approximately measured when the production cycle begins. The duration of cycles is determined by the mean residence time inside the plant, including pre and post-processing units ($\tau \approx 30$ minutes). This choice is independently confirmed by the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

spectral analysis of plant data for all grades of LDPE produced, which shows an underlying periodicity around 1.1 residence time τ units.

The MI values were determined every 10 minutes with on-line sensors that were calibrated by off-line laboratory determinations. The error associated with these on-line MI measurements is ±2%. Changes in MI correspond to changes in the physical or chemical characteristics of the desired product (grade transitions) and to variations in process conditions. In the following subsections each of the tiers proposed in the framework are used to develop, assess and implement a virtual sensor to infer LDPE MI from process variables.

7.1.2 Tier 1: SOM-based EDA

The time records of the 25 process variables that were used for training and testing the MI virtual sensor are listed in Table 7.1. These data were measured in several LDPE production plants by field instruments and sent to the control computer of the plant for their processing. This computer receives voltage signals that are converted into fix point numeric data within a certain range. Subsequently, these values are used for controlling the production process and are stored in historical logs for later processing.

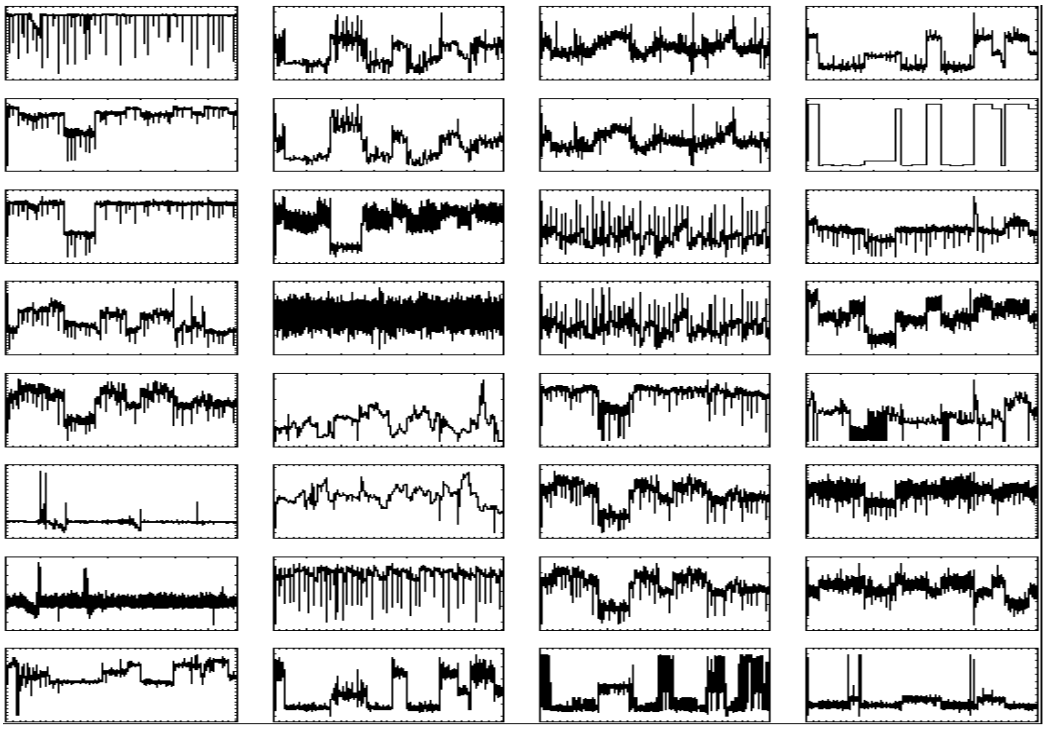


Figure 7.2. Raw data records as received from real-time field sensors

The run sequence plots of these variables for one of the plants analyzed are shown in figure 7.2. Further details of data and process plant information are subject to confidentiality agreements signed with several chemical manufacturers. Nevertheless, it can be stated that the MI virtual sensor has been built with consistent and coherent field data. The graphs in Figure 7.2 indicate that field data are noisy due to

the measurement system and mechanical effects of the process. This figure also indicates that grade transitions produce different scale effects on process variables.

Table 7.1. Process variables and correlations with the Melt Index for LDPE grades

Variable Name	Units	|R| with MI
compressor throughput	Tm/h	0.044
concentration 1	%	**0.527**
concentration 2	%	0.249
concentration 3	%	0.168
concentration 4	%	0.018
density	g/cm^3	0.183
extrusion power consumption	A	**0.583**
extrusion speed	rpm	0.056
flow rate 1	kg/h	0.021
flow rate 2	kg/h	0.042
flow rate 3	kg/h	**0.595**
flow rate 4	kg/h	**0.626**
level	%	0.126
Melt Index	g/10min	1.000
pressure	Kg/cm^2	0.052
temperature 1	ºC	0.305
temperature 2	ºC	0.023
temperature 3	ºC	**0.522**
temperature 4	ºC	0.115
temperature 5	ºC	0.122
temperature 6	ºC	0.136
temperature 7	ºC	0.428
temperature 8	ºC	0.446
temperature 9	ºC	0.112
volumetric flow rate 1	l/h	0.324
volumetric flow rate 2	l/h	**0.518**

The process variables selected are sufficient to capture the dynamics of LDPE plants analyzed for the six different quality grades (MI) considered, which have been grouped into three families of final LDPE products for convenience. Table 7.1 also

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

includes the absolute value of their correlation of process variables with MI. Data were filtered to discard abnormal situations and to improve the quality of the inference system. The input and output variables were normalized with respect to their maximum operation values for all the LDPE grades considered, so that $\forall MI, v_i \in [0,1]$.

Variables with correlations higher than 0.5 are indicated in bold. It can be observed that the most influential variable, labeled as *flowrate 4*, corresponds to a flow rate measure in the reaction section. Flow rates to the reactor are closely related to the polymer being produced and, thus, to MI. Similar explanations can be given for the other highly correlated variables. The relationships observed from the correlation analysis in Table 7.1 can be enhanced by SOM-based EDA. To this end, a SOM formed by 204 units was used to explore this data set. The component planes and the slices of the distance matrix (U-matrix) corresponding to each variable are given in Figure 7.3. It can be seen that the distribution of data on the component plane corresponding to the extrusion speed is closely related to the distribution observed for MI. Also there exist a clear relationship between the extrusion speed and its power consumption; however this is not detected by the corresponding correlation index (0.05). *Temperature 3* is inversely related to MI and the areas where high values of this variable are observed correspond to low MI values.

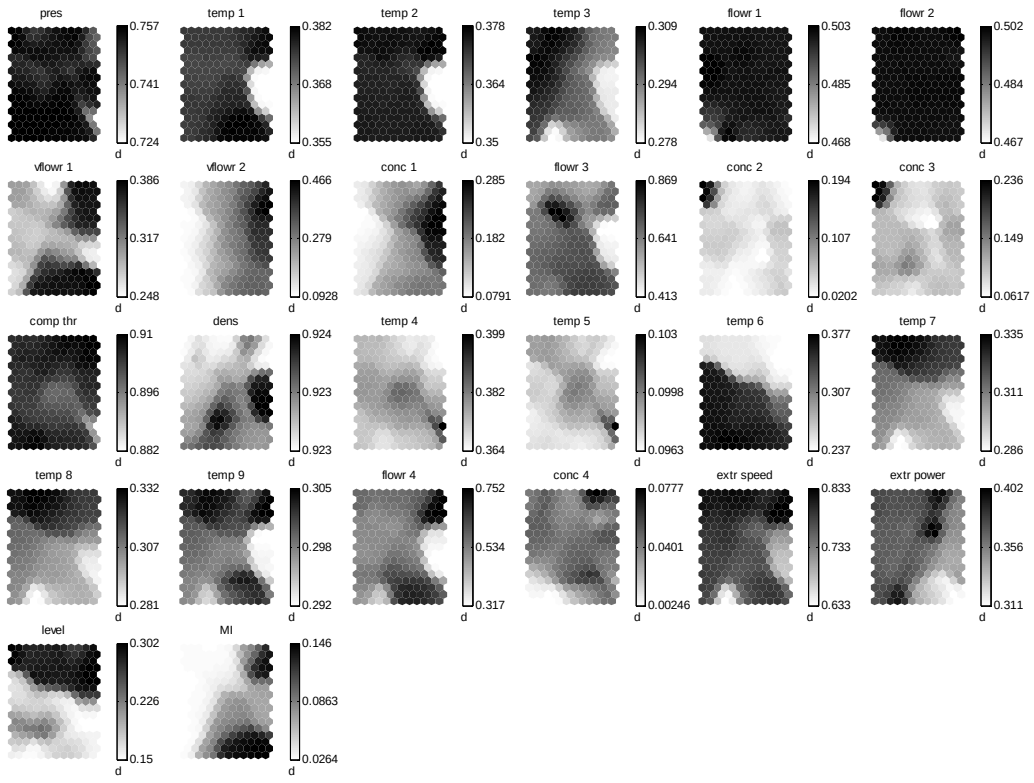


Figure 7.3. C-planes corresponding to each measured process variable

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

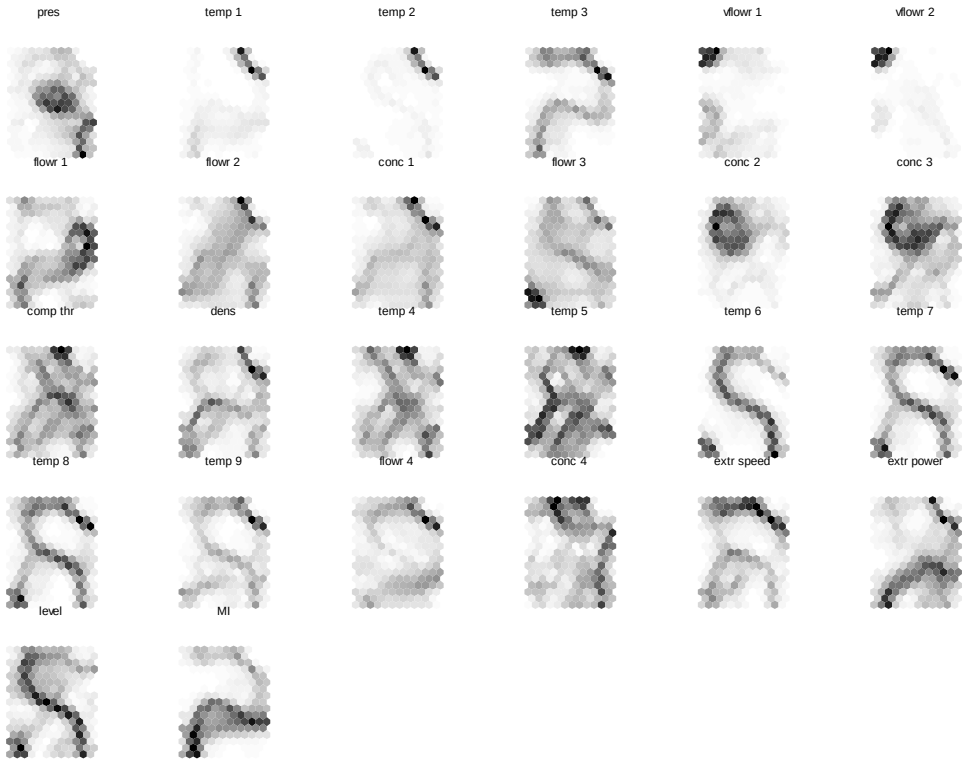


Figure 7.4. Slices of the U-matrix corresponding to each measured process variable

Figure 7.5 shows a scatter-plot of density vs. melt-index. The three clouds of points in this graph indicate the presence of at least three families of MI values which correspond to different LDPE product grades. The six LDPE product grades cluster into three families according to their average MI values and polymer densities. Each of these families contains two grades. Also each cluster has a different population density, with the highest one corresponding to the lowest value of MI.

This partitioning is also suggested by the U-matrix slices in Figure 7.4. The slice corresponding to MI clearly outlines three different regions in the U-matrix; the bottom-right, the middle-left, and the top-right region. Each of these areas corresponds to a coherent range of melt-index. It can be observed that other variables such as the extrusion speed and *temperature 3* contribute to the partitioning of the SOM space in a similar way.

The complete U-matrix is computed and labeled in Figure 7.6 by majority voting with the product grade identifier to confirm the presence of these clusters of data. This figure corroborates the separation in three families as previously suggested by the scatter-plot in Figure 7.5. Data points corresponding to low MI values are mainly grouped in the upper portion of the map in Figure 7.6. In this area, grades A and B are distributed in the middle region while grades C and D span a small portion of the top-right corner. Data points corresponding higher MI values (grades E and F) group in the lower part of the map. A close examination of Figure 7.6 suggest that there exists two different sub-families for grades A and B which are located in the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

right and left sides and mainly stem from the effects induced by temperatures in the clustering structure.

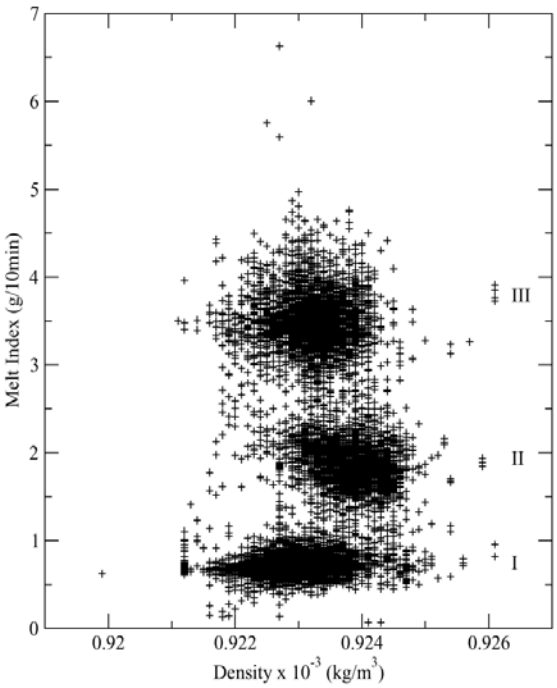


Figure 7.5. Families of LDPE identified by the variation of MI with polymer density

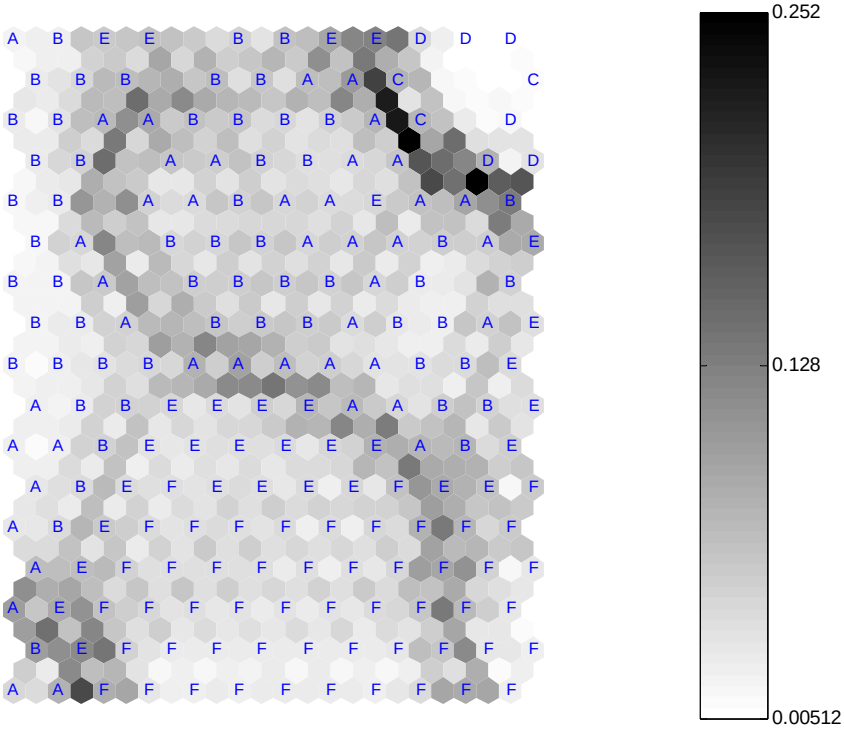


Figure 7.6. Distribution of polymer families identified by SOM

ISBN:978-84-691-1008-9/DL: T.2230-2007

7.1.3 Tier 2: Preprocessing

The c-planes resulting from the SOM are used at this level to select the best set of features for training the MI models, and to generate optimized train/test sets suitable for the external validation of the models.

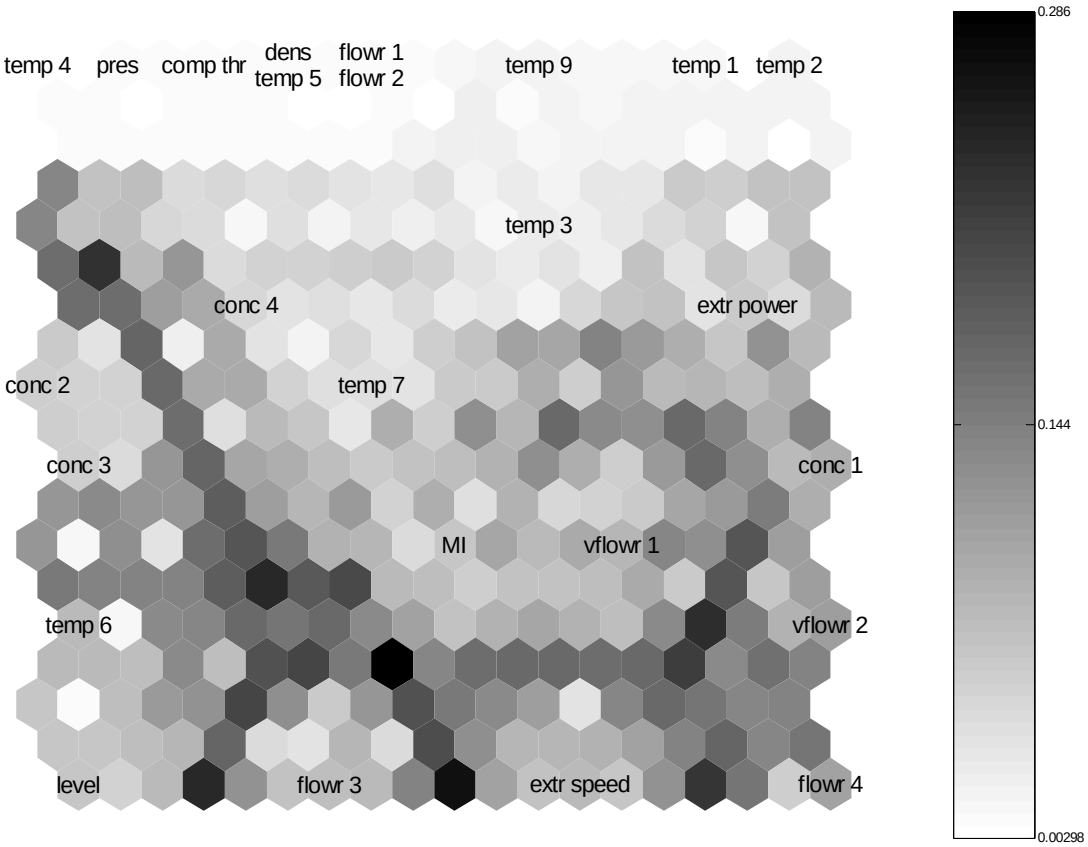


Figure 7.7. Clustering of the c-planes resulting from the SOM in Figures 7.3, 7.4 and 7.6

Figure 7.7 shows the U-matrix corresponding to the clustering of c-planes obtained from native data. Several groups of similar variables can be detected by visual inspection; for instance, *concentrations 2* and *3*. However, it is difficult to determine the position of class borders in some cases. The process based in the minimization of the Davies-Bouldin index is applied to detect an optimal partition that maximizes both, intra-cluster coherence and inter-cluster separation. The optimal partition obtained is formed by five clusters and is depicted in Figure 7.8. It is interesting to note that variables which by visual inspection of Figure 7.7 would be allocated to different classes, such as (*temperature6,level*) and (*flowrate3*), are grouped together by the the minimization of the Davies-Bouldin index. This proves the suitability of the proposed K-means clustering strategy to detect optimal cluster boundaries. The *volumetric flow rate 1* is the only variable which is grouped with the target, despite not being highly correlated with MI ($|R|=0.32$). Note that the span of this variable over the SOM space is similar to the distribution of MI values. Most variables (14) are

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

classified as belonging to the same cluster. The remaining clusters are more compact and contain fewer variables (2 to 4).

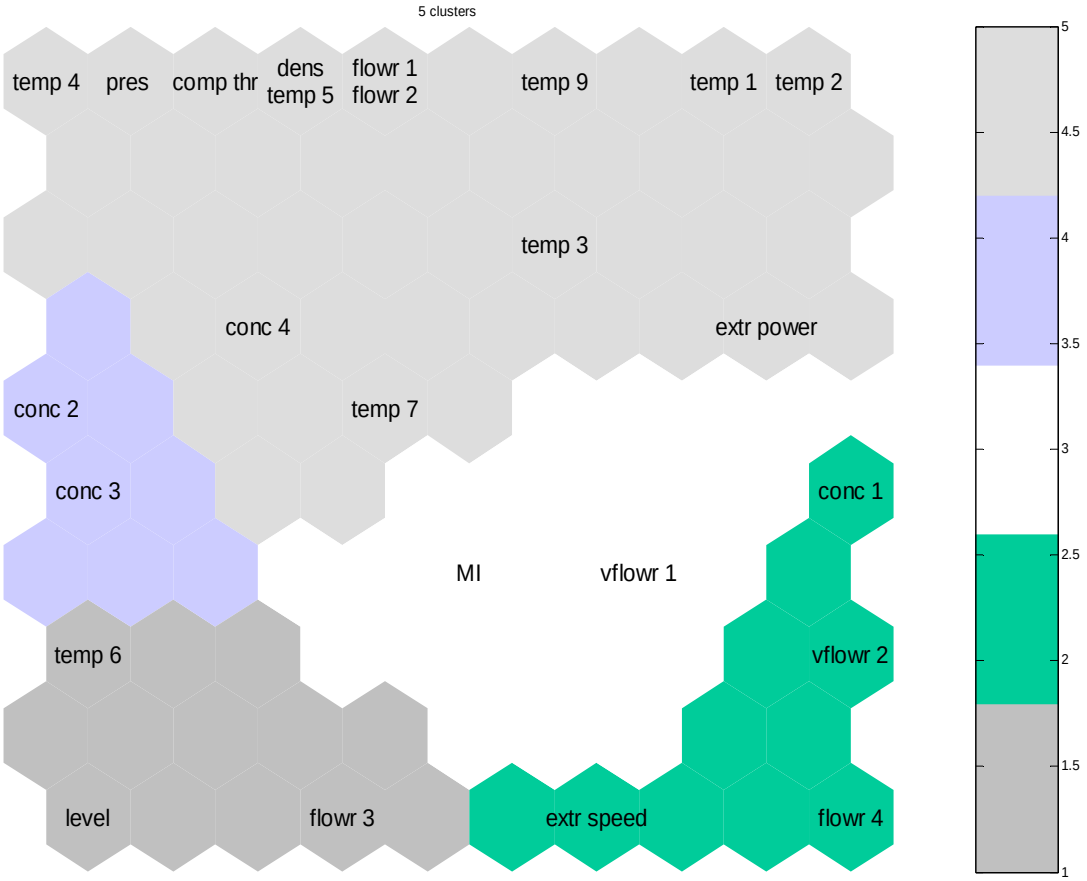


Figure 7.8. K-mean clustering with the minimization of Davies-Bouldin index of the SOM presented in Figure 7.7

The clusters in Figure 7.8 are used as the seed to carry out the SOM-dissimilarity feature selection procedure presented in subsection 4.2.1. The reduction in the size of the training set resulting from feature selection also decreases the CPU time needed to train the virtual sensor. The sets of ordered features for each model are summarized in Tables 7.2 and 7.3. Note that the 25 input process variables, ordered in Tables 7.2 and 7.3 by correlation, do not include the product grade label, even for the composite model situation, since this information is added after the best set of reduced variables has been selected from the ordered complete set. Figure 7.9 depicts the average dissimilarity curves corresponding to the feature selection procedure performed for each single grade.

It is interesting to note that, except for grade C, the *volumetric flow rate 1*, which belongs to the same class as MI, is always selected by the SOM method. Also, *density* is included in the selected subset of variables, except for grade E. The variables related to the extrusion process, either speed or power, are always selected. In

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

contrast, flow rates 3 and 4, which have the highest correlation with MI, are only simultaneously selected for the composite data set.

Table 7.2. Variables selected by the SOM-dissimilarity method for low MI grades (A-B, and C-D)

GRADE A (11 vars.)	GRADE B (15 vars.)	GRADE C (13 vars.)	GRADE D (14 vars.)
Temperature 3	Temperature 3	Temperature 2	Flow Rate 1
Density	Density	Temperature 9	Volumetric Flow Rate 2
Temperature 9	Concentration 1	Volumetric Flow Rate 2	Temperature 2
Concentration 1	Volumetric Flow Rate 1	Extrusion Power	Concentration 3
Temperature 8	Temperature 9	Temperature 1	Temperature 9
Extrusion Speed	Extrusion Power	Concentration 1	Temperature 1
Temperature 7	Temperature 8	Extrusion Speed	Volumetric Flow Rate 1
Concentration 2	Temperature 7	Temperature 3	Concentration 2
Concentration 3	Concentration 4	Density	Extrusion Speed
Volumetric Flow Rate 1	Volumetric Flow Rate 2	Concentration 3	Flow Rate 4
Temperature 4	Extrusion Speed	Flow Rate 3	Flow Rate 2
Compressor throughput	Flow Rate 2	Temperature 6	Density
Extrusion Power	Temperature 1	Concentration 2	Concentration 1
Level	Flow Rate 1	Level	Concentration 4
Pressure	Flow Rate 3	Concentration 4	Pressure
Flow Rate 2	Temperature 5	Flow Rate 4	Flow Rate 3
Temperature 6	Level	Flow Rate 2	Level
Temperature 1	Compressor throughput	Flow Rate 1	Temperature 8
Temperature 5	Temperature 2	Temperature 4	Temperature 6
Flow Rate 1	Concentration 3	Volumetric Flow Rate 1	Temperature 3
Temperature 2	Concentration 2	Pressure	Extrusion Power
Concentration 4	Temperature 4	Temperature 8	Temperature 7
Volumetric Flow Rate 2	Pressure	Temperature 7	Temperature 5
Flow Rate 3	Flow Rate 4	Compressor throughput	Compressor throughput
Flow Rate 4	Temperature 6	Temperature 5	Temperature 4

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 7.3. Variables selected by the SOM-dissimilarity method for high MI grades (E-F) and for all grades together

GRADE E (12 vars.)	GRADE F (12 vars)	ALL GRADES (17 vars.)
Level	Extrusion Power	Flow Rate 4
Temperature 6	Volumetric Flow Rate 1	Flow Rate 3
Flow Rate 4	Temperature 9	Extrusion Power
Temperature 7	Temperature 5	Concentration 1
Temperature 8	Temperature 4	Temperature 3
Volumetric Flow Rate 1	Compressor throughput	Volumetric Flow Rate 2
Concentration 1	Temperature 7	Temperature 8
Extrusion Speed	Level	Temperature 7
Volumetric Flow Rate 2	Temperature 8	Volumetric Flow Rate 1
Temperature 1	Temperature 3	Temperature 1
Compressor throughput	Concentration 2	Concentration 2
Flow Rate 3	Density	Density
Temperature 9	Flow Rate 3	Concentration 3
Concentration 2	Flow Rate 2	Temperature 6
Temperature 4	Temperature 6	Level
Density	Concentration 1	Temperature 5
Extrusion Power	Flow Rate 1	Temperature 4
Concentration 4	Flow Rate 4	Temperature 9
Concentration 3	Concentration 4	Extrusion Speed
Temperature 5	Extrusion Speed	Pressure
Temperature 2	Temperature 1	Compressor throughput
Flow Rate 2	Temperature 2	FlowRate2
Temperature 3	Pressure	Temperature2
Flow Rate 1	Concentration 3	Flow Rate 1
Pressure	VolumetricFlow Rate 2	Concentration 4

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

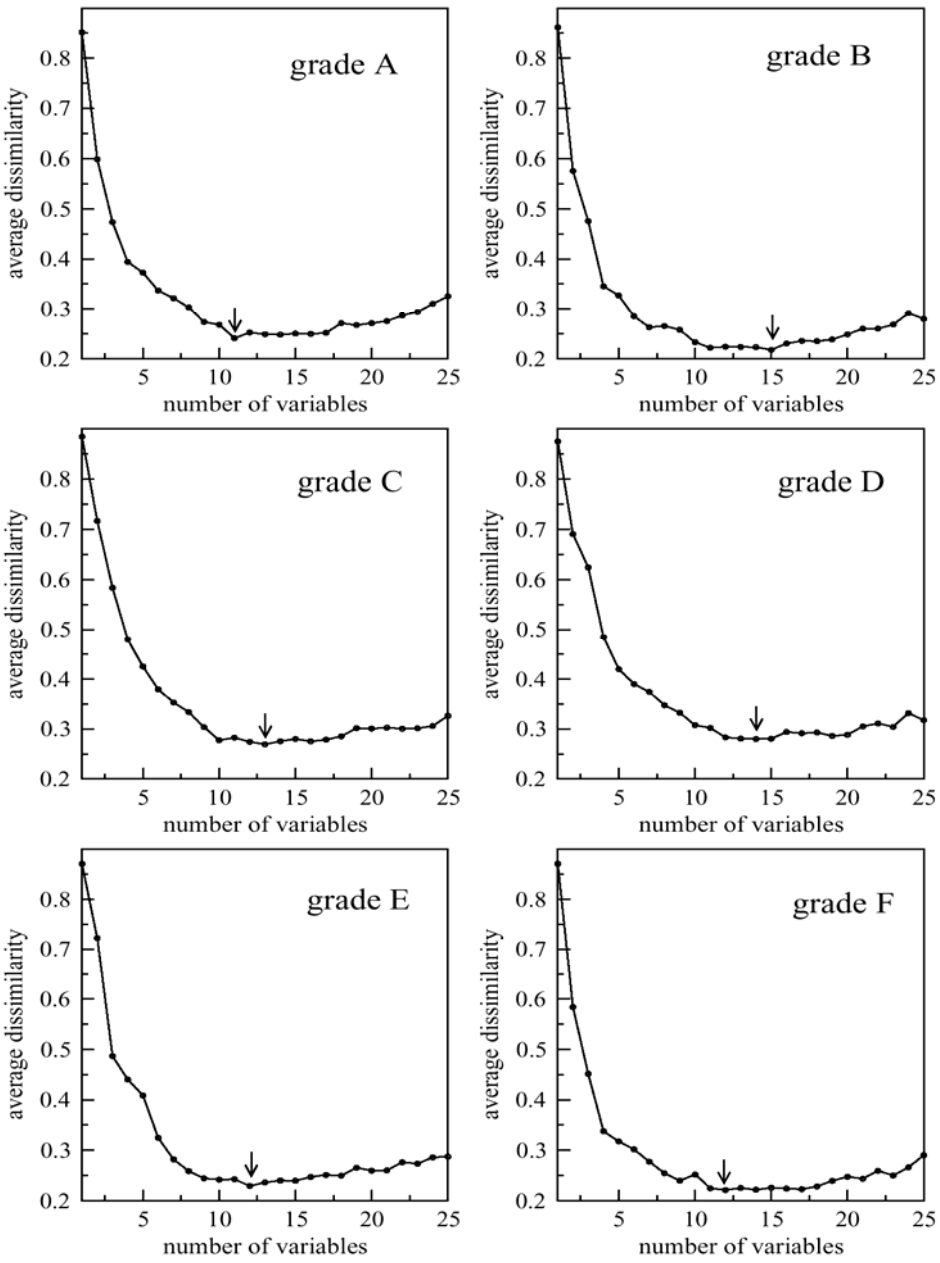


Figure 7.9. Variation of the SOM average dissimilarity measure with the addition of variables selected for each LDPE grade

The proper selection of training and testing datasets is fundamental to develop reliable and robust models, as discussed in chapter 4. Two different strategies have been used to select the partition of examples into train and test sets. The first selection procedure that was applied aimed at preserving the time-series structure of the recorded plant data for each type of LDPE analyzed. This sequential procedure consisted in separating from the time records of all variables the last 100 patterns for testing. This facilitated the representation of data and the evaluation as if sensors were operating under real plant conditions, i.e., predicting MI sequentially in time.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 7.4. Characteristics of the training and test data sets

TRAINING SET FOR THE COMPLETE PROCESS INPUT INFORMATION								
Product Families	**Grade**	**# patterns**	**MI avg**		**# different MI values**		**# single MI values**	
			Sequential	**Pre-classification**	**Sequential**	**Pre-classification**	**Sequential**	**Pre-classification**
I	A	3043	0.74	0.74	321 (11%)	321 (11%)	99 (3%)	99 (3%)
	B	3823	0.73	0.72	284 (7%)	283 (7%)	82 (2%)	81 (2%)
II	C	640	1.76	1.75	250 (39%)	277 (43%)	104 (16%)	115 (18%)
	D	957	1.74	1.73	261 (27%)	273 (29%)	113 (12%)	115 (12%)
III	E	1445	2.00	2.00	371 (26%)	383 (27%)	150 (10%)	152 (11%)
	F	4295	3.46	3.45	721 (17%)	730 (17%)	243 (6%)	241 (6%)
I+II+III	A - F	14203	1.80	1.80	1389 (10%)	1406 (10%)	292 (2%)	287 (2%)
TEST SET FOR THE COMPLETE PROCESS INPUT INFORMATION								
Product Families	**Grade**	**# patterns (test/train%)**	**MI avg**		**# different MI values**		**# single MI values**	
			Sequential	**Pre-classification**	**Sequential**	**Pre-classification**	**Sequential**	**Pre-classification**
I	A	100 (3.3%)	0.73	0.72	69	68	46	50
	B	100 (2.6%)	0.70	0.75	70	65	50	43
II	C	100 (15.6%)	1.68	1.73	76	88	57	78
	D	100 (10.4%)	1.73	1.72	80	74	62	53
III	E	100 (6.9%)	1.94	1.99	84	84	68	71
	F	100 (2.3%)	3.46	3.49	71	82	46	67
I+II+III	A - F	600 (4.2%)	1.70	1.43	367 (61%)	228 (38%)	221 (37%)	102 (17%)
TRAINING SET FOR THE REDUCED PROCESS INPUT INFORMATION								
Product Families	**Grade**	**# patterns**	**MI avg**		**# different MI values**		**# single MI values**	
			Sequential	**Pre-classification**	**Sequential**	**Pre-classification**	**Sequential**	**Pre-classification**
I	A	3043	0.74	0.72	321 (11%)	305 (10%)	99 (3%)	85 (3%)
	B	3823	0.73	0.72	284 (7%)	280 (7%)	82 (2%)	79 (2%)
II	C	640	1.76	1.72	250 (39%)	266 (41%)	104 (16%)	113 (18%)
	D	957	1.74	1.72	261 (27%)	258 (27%)	113 (12%)	104 (11%)
III	E	1445	2.00	2.00	371 (26%)	378 (26%)	150 (10%)	154 (11%)
	F	4295	3.46	3.50	721 (17%)	558 (13%)	243 (6%)	111 (2%)
I+II+III	A - F	14203	1.80	1.80	1389 (10%)	1358 (9%)	292 (2%)	284 (2%)
TEST SET FOR THE REDUCED PROCESS INPUT INFORMATION								
Product Families	**Grade**	**# patterns (test/train%)**	**MI avg**		**# different MI values**		**# single MI values**	
			Sequential	**Pre-classification**	**Sequential**	**Pre-classification**	**Sequential**	**Pre-classification**
I	A	100 (3.3%)	0.73	0.72	69	52	46	34
	B	100 (2.6%)	0.70	0.72	70	65	50	43
II	C	100 (15.6%)	1.68	1.77	76	69	57	47
	D	100 (10.4%)	1.73	1.72	80	46	62	23
III	E	100 (6.9%)	1.94	2.00	84	77	68	58
	F	100 (2.3%)	3.46	3.50	71	63	46	38
I+II+III	A - F	600 (4.2%)	1.70	1.25	367 (61%)	271 (45%)	221 (37%)	160 (27%)

The SOM-based selection procedure outlined in subsection 4.2.2 was applied afterwards to generate optimized training and test sets. The main characteristics of these training and test sets are included in Table 7.4.

7.1.4 Tier 3: Modeling

The three types of neural architectures introduced in chapter 5 were used to develop the current virtual sensor for LDPE MI. A fourth linear model was also developed to have a reference for evaluation purposes. This linear model is formulated in terms of normalized variables as,

$$MI = a_o + a_1 v_1 + \cdots + a_N v_N + a_{N+1} v_{N+1} \quad (7.1)$$

In this Eq. (7.1) $a_j \forall$ j:0,…,N+1 are adjustable coefficients and $v_i \forall$ i:1,…,N are each of the N=25 normalized process variables listed in Table 7.1 and measured simultaneously at the beginning of the production cycle. The last variable v_{N+1} in Eq. (7.1) correspond to a label identifying product grades. This label was only used in the composite model and its normalized value identifies both product quality and grade. Results showed that high or low values of these coefficients did not necessarily correspond with high or low correlations of variables with MI.

Two types of virtual sensors were developed for each neural system considered: Single models for every product grade and composite models for all LDPE grades jointly. Each kind of model was trained and tested with data sets defined according to the sequential and pre-classification procedures described in previous sections, so that the performance of each sensor could be assessed as independently as possible of the limited number of training patterns available. Furthermore, the original (complete) and reduced sets of variables have been used in order to investigate the sensitivity of the input pattern dimension in sensor's performance. On the other hand, the composite sensors, which are applicable to all grades simultaneously, should provide a good estimate of the benefits derived from the classification capabilities of all the neural systems considered. Table 7.4 includes the most relevant information concerning the training and test sets used in the current study to develop sensors from the complete input data set of 25 process variables (upper part of Table 7.4) and from the reduced sets selected by the sensitivity analysis using SOMs (lower part of Table 7.4). From the point of view of MI, the two sets of data used for training and testing each kind of neural sensor, characterized by the different input dimension, had similar characteristics, i.e., approximately equal average MI values, number of different MI values, and number of single MI values in the data records. The parameters used to develop (train and test) the three current neural sensor models are summarized in Table 7.5.

Table 7.5 summarizes the characteristics of the data sets used to build and test the different virtual sensors developed for three different families of LDPE identified in Figure 3 according to their different MI values. Each of these families includes two product grades that are produced under different process conditions (dynamics). Single models for each product grade as well as composite ones, valid for the ensemble of all LDPE families, have been developed to test and compare the performance of the different sensor currently under consideration. The

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

aforementioned data preprocessing techniques aimed at reducing the number of variables and at optimizing the composition of the training set were applied in all cases. The results obtained for all virtual sensors considered are summarized in Tables 7.6 and 7.7.

Table 7.5. Training parameters for the three neural models developed.

	Fuzzy ARTMAP	Clustering Average	DynaRBF
Single Grade Model All variables	ρ=0.995	d_{max} =0.05 α_0=0.1 (Eq. 5.11)	d_{max}=0.05 α_0=0.1 (Eq. 5.11) β=0.1 (Eq. 5.17) σ_0=1.0 (Eq. 5.14)
Composite Model All variables	ρ=0.9995		
Single Grade Model Reduced variables			
Composite Model Reduced variables			

It should be noted that the average characteristics of the training and test sets listed in Table 7.4 differ slightly for single grade models and significantly for composite ones, as illustrated by the average MI values listed for each case. Differences are even more noticeable when the pre-classification technique for data selection is applied, and between models built with either all input variables or with the most relevant input features only. This should be kept in mind when comparing relative errors between different models, particularly for composite models. In this case the average MI values of the sparse test sets will neither coincide with that of the training set nor with that of any of the six grades considered individually, and will differ significantly when both pre-classification and variable reduction techniques were applied to develop the composite models. For example, the average MI values for both the sequential and pre-classification training sets used to develop composite models with the reduced set of input variables is equal to 1.80 in Table 7.4, while that for the corresponding test sets are 1.7 and 1.25, respectively. Thus, only absolute errors are reported in Tables 7.6 and 7.7 for the composite models.

7.1.4.1 Models with the complete set of variables

Table 7.6 summarizes the performance of all neural sensor models developed with the complete set of input process variables after training with both the sequential and the pre-classified sets of data given in the upper part of Table 7.4, which also includes the test sets used for these models. The absolute and relative mean errors listed in Table 7.6 for single grade models and sequential training (upper part) indicate that all neural sensors, especially DynaRBF, function better than the linear model on the overall, with absolute mean errors for the linear model ranging from 0.080 to 0.510 respectively for I(B) and II(C), and those for the neural sensors from 0.057 to 0.460. The standard deviations listed also in Table 7.6 are of the same order of magnitude as the absolute mean errors for all models with sequential training.

ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 7.6. Absolute and relative mean errors and standard deviations for the test sets predicted by all sensor models built with all process variables after training with both the sequential and pre-classified sets of patterns

	Family	Grade	LINEAR MODEL Abs (%) Std. Dev.	CLUSTERING AVERAGE Abs (%) Std. Dev.	Fuzzy ARTMAP Abs (%) Std. Dev	DynaRBF Abs (%) Std. Dev.
Sequential	I	A	0.100 (13.7%) 0.099	0.072 (9.9%) 0.066	0.069 (9.4%) 0.052	0.066 (9.0%) 0.053
		B	0.080 (11.4%) 0.063	0.063 (9.0%) 0.065	0.073 (10.4%) 0.061	0.057 (8.1%) 0.050
	II	C	0.510 (30.3%) 0.411	0.438 (26.1%) 0.378	0.460 (27.4%) 0.385	0.456 (27.1%) 0.384
		D	0.148 (8.5%) 0.126	0.190 (10.9%) 0.132	0.163 (9.4%) 0.118	0.152 (8.8%) 0.106
	III	E	0.323 (16.6%) 0.303	0.295 (15.2%) 0.281	0.302 (15.6%) 0.276	0.286 (14.7%) 0.291
		F	0.393 (11.3%) 0.272	0.383 (11.1%) 0.262	0.358 (10.3%) 0.260	0.339 (9.8%) 0.234
	I+II+III	A – F	0.295 0.301	0.157 0.160	0.163 0.152	0.151 0.148
	Family	Grade	LINEAR MODEL Abs (%) Std. Dev.	CLUSTERING AVERAGE Abs (%) Std. Dev.	Fuzzy ARTMAP Abs (%) Std. Dev.	DynaRBF Abs (%) Std. Dev.
Pre-classified	I	A	0.107 (14.9%) 0.165	0.057 (7.9%) 0.050	0.049 (6.8%) 0.045	0.048 (6.7%) 0.067
		B	0.113 (15.1%) 0.243	0.051 (6.8%) 0.053	0.038 (5.1%) 0.027	0.041 (5.5%) 0.065
	II	C	0.195 (11.3%) 0.202	0.147 (8.5%) 0.113	0.131 (7.6%) 0.116	0.149 (8.6%) 0.166
		D	0.132 (7.8%) 0.159	0.063 (3.7%) 0.065	0.058 (3.4%) 0.060	0.059 (3.4%) 0.064
	III	E	0.164 (8.2%) 0.165	0.064 (3.2%) 0.066	0.076 (3.8%) 0.077	0.056 (2.8%) 0.049
		F	0.213 (6.1%) 0.165	0.149 (4.3%) 0.151	0.165 (4.7%) 0.174	0.146 (4.2%) 0.151
	I+II+III	A – F	0.182 0.178	0.121 0.114	0.118 0.120	0.125 0.115

The better performance of neural sensors is clearer in family I, where all models yield consistently good predictions. Increases of approximately 40% and 20% in

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

prediction accuracy are obtained for grades I(A) and I(B), respectively, with the neural models compared to the linear model. Deficient training is the cause for the anomalous high errors of grades II(C) and III(E) and for the slightly better performance of the linear model for grade II(D) since they disappear when the pre-classified set of data was used for training, as shown in the lower part of Table 7.6. Note that the small number of 640 training patterns available for training grade II(C) (see Table 7.4) aggravates this situation. In addition, the training set has the largest test to train ratio (15.6%) and is the most heterogeneous one, as indicated by its high percentage of different and single MI values in Table 7.4. The fact that an increase of 50% in the number of training patterns between II(C) and II(D) reduces errors by approximately a factor of three, i.e., to the average levels in Table 7.6, reinforces the above arguments but also indicates that this family of LDPE was produced under distinct non-linear process dynamics.

To illustrate the genesis of the average errors given in Table 7.6, i.e., the detailed performance of the sensors built with sequential training, Figure 7.10 shows the measured and predicted time-records for grade I(A). The linear model is unable to follow the time-sequence of the measured MI, while clustering average sensor and particularly the DynaRBF model perform reasonably well, consistently with the average errors in Table 7.6. The performance of fuzzy ARTMAP, while being reasonable on the average (see Table 7.6) it yields a step-like response around the average MI. This is again a clear indication of the insufficient training information provided by the sequential set. The fact that this effect is more evident for the usually highly performing fuzzy ARTMAP algorithm in difficult classification problems arises from its need for increased training for periodic signals, as illustrated in the benchmark reported by Carpenter et al. (1992) and in tests carried out during the course of the current study. The plant data for all grades have an underlying periodicity around 1.1 residence time τ units.

The performance of the single grade models developed with all variables and trained with the pre-classified set of MI values is summarized in the lower part of Table 7.6. The average performance of all virtual sensors for single grades trained using this pattern pre-classification selection procedure is significantly better in terms of both mean errors and standard deviations for all product grades than that reported in the upper part of Table 7.6 for sequential training. This effect is most significant for the grades with less number of patterns in Table 7.4, i.e., grades II(C), II(D) and III(E), where errors in the predictions decrease by more than a factor of three with the change of training sets. It is also more noticeable in the fuzzy ARTMAP sensor for its especial sensitivity to training in systems with periodical behavior, as explained before. For instance, the absolute mean error of predictions obtained for grade II(C) with the DynaRBF model drops from 0.456 (27.1%) to 0.149 (8.6%) and from 0.460 (27.4%) to 0.131 (7.6%) for fuzzy ARTMAP. The same applies to grade III(E) and for the other grades with errors decreasing between 30% and 60%.

The better performance of the single grade neural sensors with respect to the linear model is more evident and consistent when training is carried out with the pre-classified sets of data. The average relative errors for the former models range from 2.8% to 8.6% while for the latter they range from 6.1% to 15.1%. This tendency is even clearer in terms of standard deviations. The average performance of DynaRBF

and fuzzy ARTMAP are similar, with an overall relative error of 5.2%, followed by clustering average with 5.7%, and the linear model with 10.6%. The classification capabilities of fuzzy ARTMAP can be inferred by the lower standard deviations of predictions that are obtained when the more appropriate pre-classified set of data is used to train the networks. The results for the three neural sensors are remarkable considering the ±2% error associated with the on-line MI measurements.

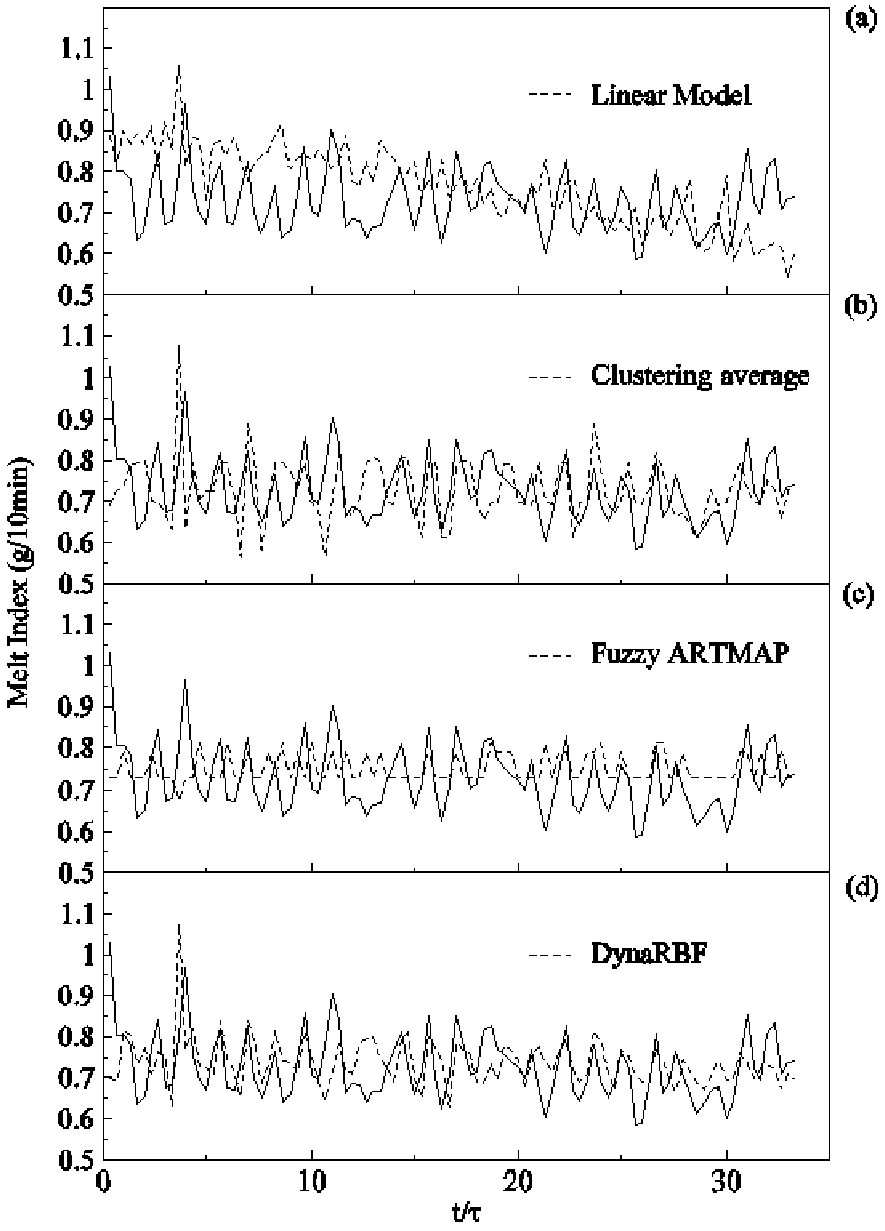


Figure 7.10. Comparison between measured and predicted MI for grade I(A) using single neural models with sequential training and all available variables. (a) Linear; (b) clustering average; (c) fuzzy ARTMAP; (d) DynaRBF

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

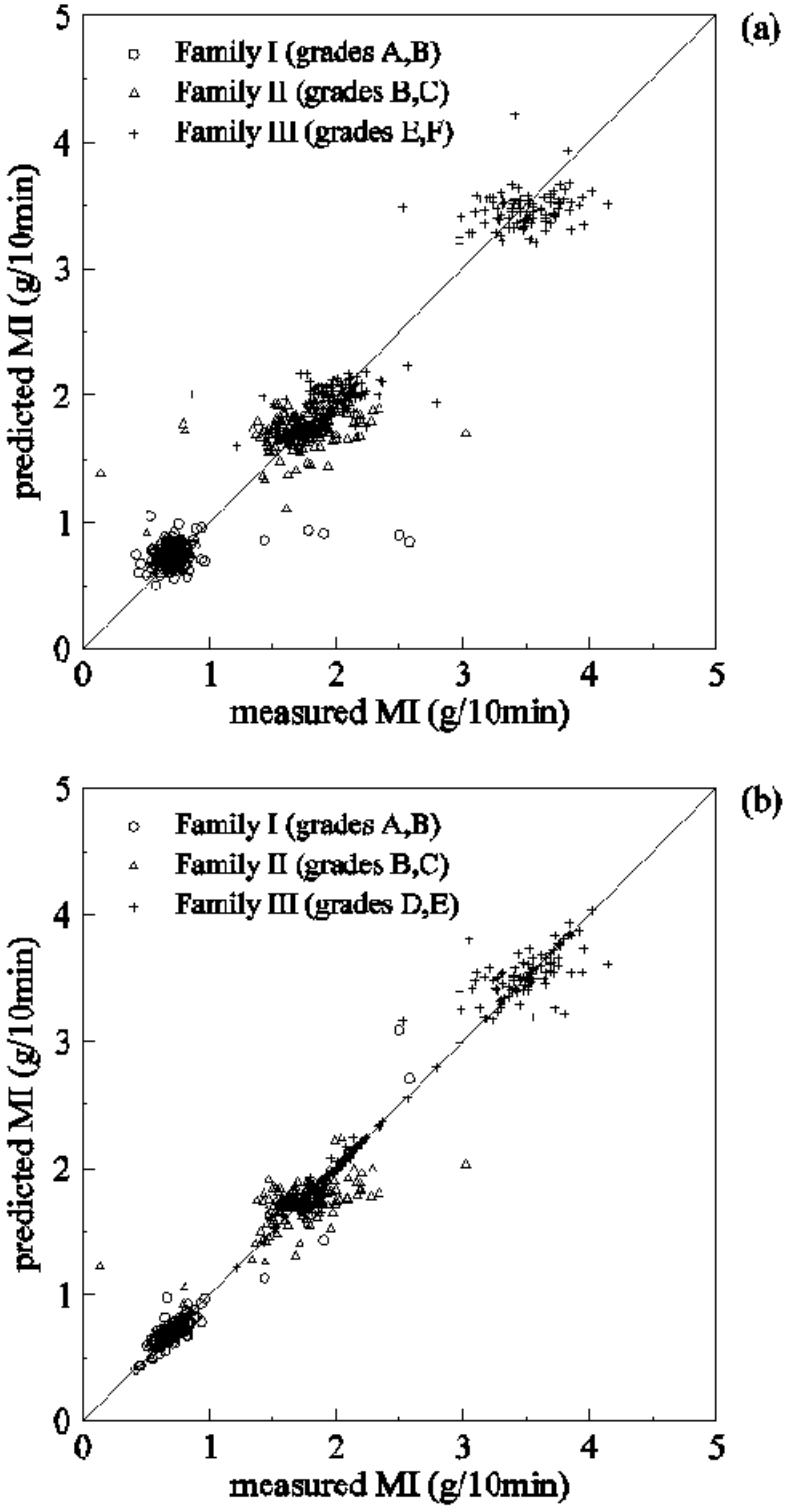


Figure 7.11. Comparison between measured MI time-records and predictions obtained by using: (a) Single grade linear model; (b) single grade DynaRBF model trained with the pre-classified data set for the three families of LDPE grades studied

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

A detailed comparison of performances between the linear model and the DynaRBF sensor is shown in Figure 7.11 for the three product families. Under the training scheme with pre-classified set of patterns detailed results can only be evaluated in terms of deviations of predicted MI with respect to measured values since training patterns are not presented to the neural models according to the time-sequence of plant measurements. Figure 7.11 confirms the improved performance obtained with the DynaRBF model during testing. This figure also shows that there is an overlap in the MI distribution between product grades II(C) and III(E). This could be one of the reasons why product grade II(C) exhibits the maximum errors in Table 7.6.

The potential of the currently proposed neural sensing technology should become more evident when attempting to predict the behavior of the ensemble of LDPE grades simultaneously with only one sensor, i.e., with a composite sensor model. As a consequence, the four models have also been tested with the more difficult problem of forecasting the quality of the three LDPE families (I+II+III) simultaneously, i.e., forecasting grade transitions. Table 7.6 also summarizes the composite model results obtained with both training sets formed by the 14,203 patterns indicated in Table 7.4 and using the 25 process variables listed in Table 7.1 complemented by the normalized label to identify the six grades. The composite models were tested with the remaining 600 patterns, so that the test to train ratio was kept comparable to that for single grade models.

Table 7.6 indicates that the absolute mean errors for the three composite neural sensor models with sequential training are approximately equal to 0.157 compared to 0.295 for the composite linear correlation model. This expected good performance of the neural systems, also reflected by the respective standard deviations of 0.153 and 0.301, is examined in Figure 7.12 for a production cycle including the three families. The behavior of the three neural sensors is similar over time according to the time sequence of the plots of the errors (Figures 7.12b-d) with respect to the measured MI (Figure 7.12a). Predictions made over patterns corresponding to family I (grades A and B) show the lowest fluctuations in error while those for family III are the highest, as was also the case for the single grade sensors in sequential training in Table 7.6. This behavior is partially due to deficient training as discussed previously. Note that the MI variation shown in Figure 7.12a also illustrates the differences in process dynamics between product families. The use of the most suitable training set obtained by pre-classification in the four composite models reduces absolute errors and standard deviations of predictions from 0.157 and 0.153 to 0.121 and 0.116, respectively, for all virtual sensor models and from 0.295 and 0.301 to 0.182 and 0.178 for the linear correlation model, as shown also in Table 7.6. In this case of composite models the performance of neural sensors is also better.

7.1.4.2 Models with the reduced set of variables

All single and the composite sensor models developed for the reduced set of process variables selected by dissimilarity of SOMs have been trained and tested by using the data sets given in the lower part of Table 7.3. The reduction in the number of input variables has the advantages of (i) dealing with a lower dimensional problem, (ii) cutting-back any noise that could contaminate the measurements of the discarded

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

variables, and (iii) avoiding variables that could provide conflicting information with respect to more correlated variables in the relation to the target MI. Figure 7.9 shows the variation of the average cumulative dissimilarity between self-organizing maps that has been calculated for each grade by successive insertion of the variables listed in Tables 7.3 and 7.4. The variables located beyond the minimum dissimilarity point in the plots of Figures 7.9 or below the separation line in Tables 7.3 and 7.4 do not contribute with any additional relevant information to explain the qualitative and quantitative behavior of MI and can be discarded from the input data set of process plant information.

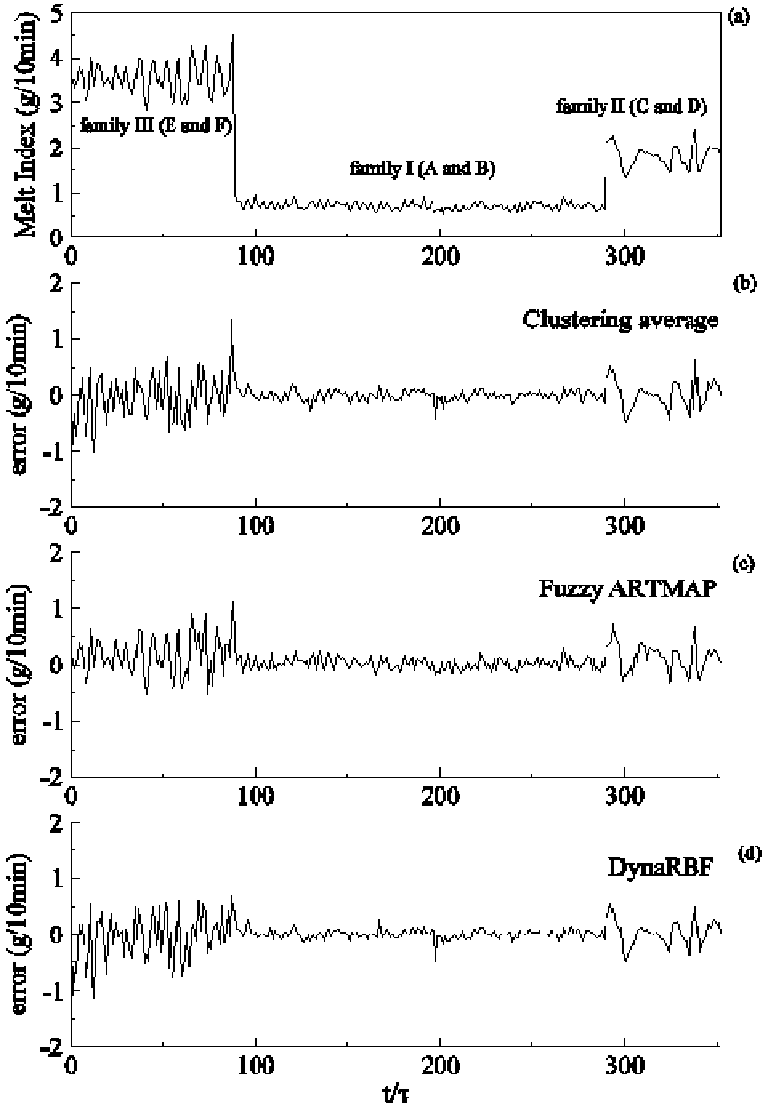


Figure 7.12. Time-records of measured MI and of the errors of values predicted by the composite neural sensor models applicable simultaneously to all LDPE families considered and trained with the sequential set of data; measured MI (a); Clustering average (b); fuzzy ARTMAP (c); DynaRBF (d)

ISBN:978-84-691-1008-9/DL: T.2230-2007

It should be noted that the same variables but in a different order would have been selected if an ordering criteria based on both topological and correlation information (Espinosa et al., 2001) would have been adopted. The application of this selection procedure allowed an approximate 50% reduction in the number of input data needed to develop all single sensor models. Table 6 summarizes the performance of all single grade models using the reduced set of input variables and training with sequential and pre-classified data.

Table 7.7. Absolute and relative mean errors and standard deviations for the test sets predicted by all sensor models built with the reduced set of process variables after training with both the sequential and the pre-classified set of patterns

	Family	**Grade**	**LINEAR MODEL** Abs (%) Std. Dev.	**CLUSTERING AVERAGE** Abs (%) Std. Dev.	**Fuzzy ARTMAP** Abs (%) Std. Dev.	**DynaRBF** Abs (%) Std. Dev.
Sequential	I	A	0.083 (11.4%) 0.066	0.064 (8.8%) 0.051	0.066 (9.0%) 0.061	0.059 (8.1%) 0.055
		B	0.070 (10.0%) 0.055	0.058 (8.3%) 0.052	0.070 (10.0%) 0.057	0.057 (8.1%) 0.049
	II	C	0.432 (25.7%) 0.378	0.380 (22.6%) 0.322	0.368 (21.9%) 0.296	0.338 (20.1%) 0.275
		D	0.153 (8.8%) 0.135	0.166 (9.6%) 0.126	0.185 (10.7%) 0.147	0.175 (10.1%) 0.135
	III	E	0.291 (15.0%) 0.294	0.224 (11.5%) 0..224	0.217 (11.2%) 0.166	0.219 (11.3%) 0.202
		F	0.376 (10.9%) 0.493	0.374 (10.8%) 0.537	0.397 (11.5%) 0.320	0.360 (10.4%) 0.507
	I+II+III	A – F	0.281 0.325	0.247 0.338	0.235 0.318	0.237 0.343
	Family	**Grade**	**LINEAR MODEL** Abs (%) Std. Dev.	**CLUSTERING AVERAGE** Abs (%) Std. Dev.	**Fuzzy ARTMAP** Abs (%) Std. Dev.	**DynaRBF** Abs (%) Std. Dev.
Pre-classified	I	A	0.055 (7.6%) 0.047	0.037 (5.1%) 0.034	0.038 (5.3%) 0.037	0.031 (4.3%) 0.034
		B	0.068 (9.4%) 0.061	0.042 (5.8%) 0.038	0.048 (6.7%) 0.050	0.039 (5.4%) 0.035
	II	C	0.113 (6.4%) 0.132	0.073 (4.1%) 0.082	0.088 (5.0%) 0.098	0.048 (2.7%) 0.051
		D	0.043 (2.5%) 0.062	0.038 (2.2%) 0.048	0.046 (2.7%) 0.046	0.020 (1.2%) 0.022
	III	E	0.113 (5.6%) 0.168	0.097 (4.8%) 0.200	0.083 (4.1%) 0.120	0.097 (4.8%) 0.143
		F	0.055 (1.6%) 0.064	0.066 (1.9%) 0.069	0.091 (2.6%) 0.132	0.050 (1.4%) 0.041
	I+II+III	A – F	0.232 0.286	0.078 0.166	0.095 0.167	0.081 0.162

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The effects of variable reduction when the single grade sensor models were trained with the set of patterns selected sequentially are summarized in upper part of Table 7.7. Comparison with the corresponding results in Table 7.6 shows that the performance of all single grade models, linear and neural, is maintained or improves slightly despite the reduction in the information provided to the virtual sensor. For example, the single models for grade I(B), all built with the first 15 variables listed in Table 7.3 according to the minimum dissimilarity in Figure 7.9, yield predictions with an average absolute error and standard deviation of 0.062 and 0.053, respectively, for sequential training in Table 7.7, which is slightly better than the corresponding 0.064 and 0.059 for all 25 variables in Table 7.6. The same holds for the more difficult to predict grade II(C), where the 26.9% average relative error obtained for all variables with sequential sets (Table 7.6) drops to about 21.5% (Table 7.7) for the reduced set of 13 input variable selected from Figure 7.9 and Table 7.4. In the case of grade III (E) the reduction in relative error is from 15.2% for all variables to 11.3% using 12 variables. Similar behaviors are observed for the other grades, both in terms of mean errors and standard deviations.

The effects of variable reduction when the single grade sensor models were trained with the best set of patterns selected by the pre-classification procedure can be observed in the lower part of Table 7.7. The errors of predictions drop on the average from approximately 5% in Table 5 to 4% in Table 7.7 for the neural sensors, and from 10% to 5.5% for the linear model. The reduction in input variables also causes a decrease in standard deviations. The same tendency of improvement or comparable performance is observed for each individual sensor and grade. DynaRBF is confirmed in Table 7.7 as the best single grade neural sensor with an overall absolute error and standard deviation of 0.048 (3.3%) and 0.054 for the six grades A-F, followed by clustering average with 0.059 (4%) and 0.079, and fuzzy ARTMAP with 0.066 (4.4%) and 0.081. In all cases, the mean errors of predictions are comparable to the ±2% error associated with MI on-line measurements. As discussed before, the unexpected poorer performance of the powerful ARTMAP classifier is due to insufficient training considering the underlying periodicity of the inferred target MI variable. The analyses of detailed input pattern classification and MI forecasting during testing support the average results in Table 7.7, but have not been included here for brevity. Comparison between the upper and lower parts of Table 7.7 confirms again the adequacy of the pre-classification approach to select data for training all sensors.

This significant improvement in the performance of properly trained single grade virtual sensors caused by the adequate reduction in input information indicates that redundancy of information may be disadvantageous when dealing with field variables contaminated by measurement errors. Also, the inclusion of variables with conflicting or contradicting information in relation to that contributed by other variables with higher correlations with the target MI could result in a detrimental effect. Note that beyond the minimum points in the plots of Figure 7.9 dissimilarity changes very slowly with the addition of more variables indicating that their inclusion or exclusion will not affect information but could contribute to the addition or subtraction of noise and/or of conflicting information with respect to their effect on MI.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The effects of the reduction of variables in the composite models are also included in Table 7.7 and the most relevant results highlighted in Figure 7.13. The arrow in Figure 7.13a indicates that the minimum dissimilarity is reached in this case when the first 17 variables are considered from the ordered list in the last column of Table 7.4. It is also clear from this figure that noise and experimental errors make the choice of the minimum dissimilarity more difficult. All composite models have been developed with an input formed by these 17 variables plus the normalized product grade label to facilitate product identification.

The performance of the linear composite model improves slightly with variable reduction when sequentially trained, with absolute errors decreasing from 0.295 in Tables 7.6 to 0.281 in Table 7.7, but worsens when the best pre-classification set of data is used at the learning stage, increasing from 0.182 in Table 7.6 to 0.232 in Table 7.7. It should be noted that the simple additive nature of this model (Eq. 7.1) could more easily cancel the noise effects and/or handle conflicting information and, thus, gain from the redundant information brought by any extra variable when appropriately trained; the lowest error of 0.182 and standard deviation of 0.178 correspond to the composite linear correlation built with all variables and trained with the pre-classified set of patterns (Table 7.6).

The functioning of the composite neural sensor models trained sequentially worsens significantly with the reduction of input variables; relative errors increase on the overall from 9% in Table 5 to 14% in Table 7.7, while standard deviations approximately double. For the more appropriate training set selected by pre-classification performance remains unchanged or improves slightly with variable reduction in terms of errors but the dispersion of predictions increase. These results indicate the difficulties encountered by the sensors to properly classify certain input patterns during testing and to generate adequate outputs for MI for the large variety of process dynamics that occur in the LDPE plant. The best composite neural sensors are dynaRBF and clustering average, which yield the lowest absolute errors of about 0.080 for the case of pre-classified training (lower part of Table 7.7). The corresponding error for fuzzy ARTMAP is an acceptable 0.095 considering the periodicity effects discussed before. When the inadequate sequential training is applied in this case of less input information (reduced variables) the mean absolute error of predictions triples (upper part of Table 7.7), which is the highest in Tables 7.6 and 7.7 for neural composite models. This is consistent with the fact that any model deficiency should become more evident when dealing with the more difficult problem of predicting simultaneously all grades with reduced input information. These results are illustrated and corroborated in detail in Figures 7.13b and 7.13c where the measured and predicted MI values corresponding to the test set of patterns are compared for the four composite models trained with pre-classified data. The superior performance of the neural models compared to the linear model is clear in these figures. The origin of the observed deviations is similar to that illustrated in Figure 7.12 for the complete set of variables and sequential training. Again, the lowest deviations correspond to family I.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

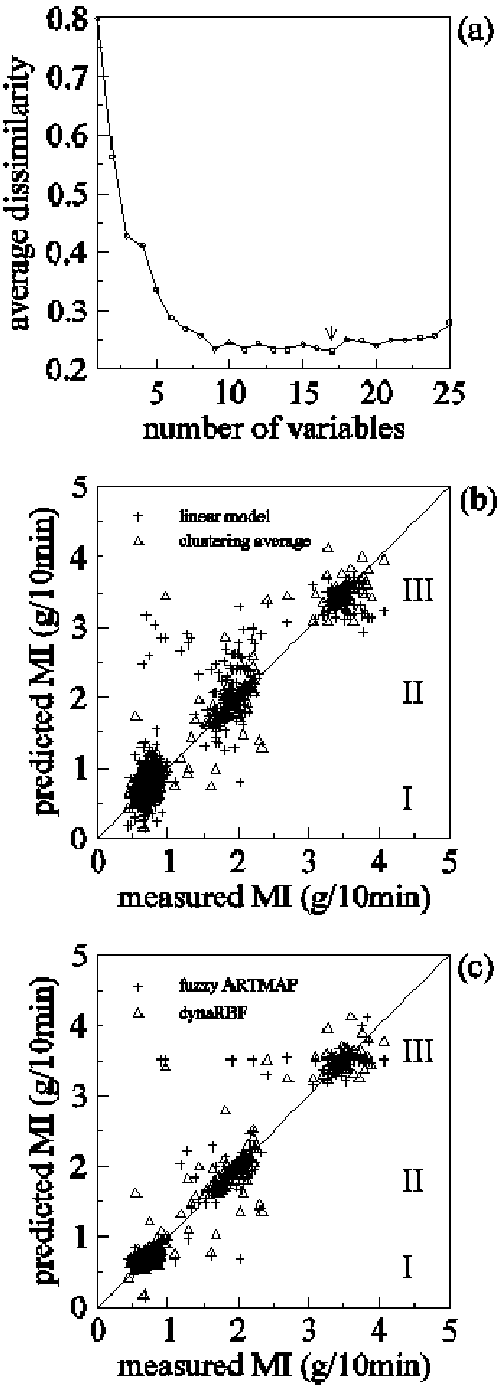


Figure 7.13. Results obtained for the composite models with all sensors trained using the best set of pre-classified data and the reduced set of input variables. Variation of average dissimilarity between self-organizing maps with the inclusion of ordered variables (a); measured vs. predicted MI for the linear and clustering average models (b); measured vs. predicted MI for the fuzzy ARTMAP and DynaRBF models (c)

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

7.1.5 Integration of the Virtual Sensor in a production Plant

Figure 7.14 depicts the implementation of the virtual sensor systems within the control flow sheet of an LDPE plant. It can receive real-time readings of process variables as well as feedback signals of downstream on-line analyzers; both sets of data were used for training (and later adapting) the virtual sensor. Once trained, this virtual device uses only real time measurements of the selected process variables made by process sensors at any time to infer the value of the product target property when leaving the reactor. The output can be redirected as information to the plant operator or to the control system to maintain optimal plant operation for a given product quality.

A successful model for estimating the MI on-line was developed in previous sections. The main drawback of this approach resides in the fact that the failure of a single sensor makes the virtual sensor unusable. This is a major issue for the integration of the virtual sensor systems in a real processing plant. The data imputation scheme developed in section 4.2.3, is integrated into the sensor system to address this issue.

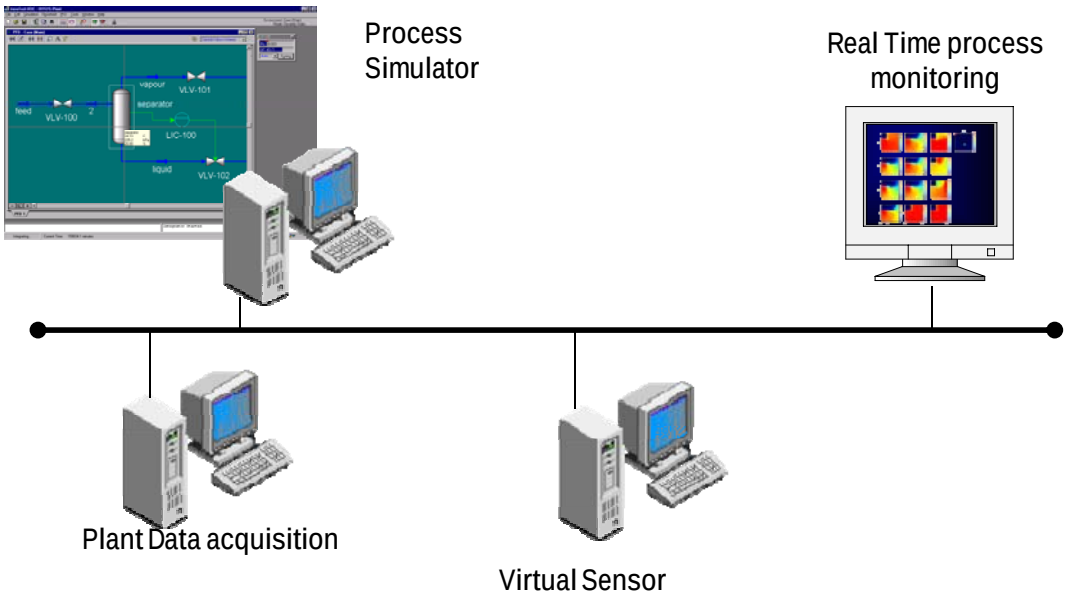


Figure 7.14. Integration of the virtual sensor in the control scheme of a real LDPE production plant

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

7.1.5.1 Missing Data Reconstruction

An extended pool of the LDPE process information, formed by the 148 process variables (pressures, flow rates, temperatures of the cooling/heating streams of the reactor, etc.) closely related to the polymerization process and sampled at intervals of 10 minutes was used to implement the SOM-based multiple imputation system described in section 4.2.3. For simplicity, only one grade of LDPE, containing 5548 input patterns, is discussed here. The data used to train the system included all the 148 variables, with both complete and incomplete patterns. Table 7.8 summarizes the results obtained using the imputation system under different conditions and using diverse imputation models. Three main cases have been analyzed. First, the effect of random failures in a temperature sensor, second the effects of failures in a flow meter, and finally random failures in any of the 148 sensors.

Temperature and flow rate have been chosen because of their distinct distribution of data. Temperature data are smoothly distributed and the amount of unique values is low (4.3%) compared with data corresponding to flow rate, were the amount of unique values raises till 56%. This situation is specially challenging for a prototype-based classifier such is the SOM. In all these situations the models compared had been: mean substitution, single imputation (SI) based on the best matching unit (*bmu)*, SI based on a neighborhood of the *bmu*, multiple imputation (MI) using bagging and MI using maps of different size. Also, in each case the effect of the amount of missing data has been studied, varying the missing data ratio from 10% to 70% of the available patterns. Details of the methodology used to develop the single and multiple imputation systems are given in chapter 4.

Inspection of Table 7.8 leads to several observations: (i) Regardless of the methods used to construct the individual single imputation systems the absolute mean error of the aggregated response of an ensemble of maps is always lower than that of each individual system; (ii) all imputation systems maintain an stable behavior with the increase of missing data, independently of the imputation technique used. The imputation systems based on the SOM are very robust with respect to the amount of missing data. The performance of these systems (measured as the absolute mean error) does not degrade in a significant way when the amount of missing data increases from 10% up to 70%. In all situations, the SOM imputation yield an absolute mean error (AME) lower than that obtained with mean-substitution; (iii) all SOM imputation methods tend to overestimate the value of the mean of the imputed data, while those based solely in mean-substitution are more precise but cannot reproduce well the variance of data. For the temperature sensor, the best imputation model presents a maximum relative error in the mean estimation of 0.6% for the dataset with 10% of missing data. On the other hand, for the flow rate meter, the maximum relative error in the mean for the best model is 0.40% which corresponds to the dataset with 70% of missing data.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 7.8. Comparison of the imputation models in the LDPE plant for two simulated failures. Absolute mean errors and statistical properties of the imputed dataset (mean, median and variance) for three different levels of missingness

	Imputation model	%missing data	AME	μ	median	var
Temperature Sensor Failure μ=0.516 median=0.510 var=0.031 unique=240 (4.3%)	mean	10%	0.136	0.515	-	-
		30%	0.141	0.516	-	-
		70%	0.142	0.515	-	-
	SI_{bmu}	10%	0.101	0.512	0.516	0.005
		30%	0.106	0.513	0.517	0.005
		70%	0.106	0.515	0.520	0.005
	$SI_{neigh=3}$	10%	0.108	0.511	0.512	0.003
		30%	0.114	0.512	0.512	0.003
		70%	0.115	0.512	0.515	0.003
	$MI^{bagging}_{bmu}$	10%	0.080	0.522	0.514	0.017
		30%	0.080	0.516	0.507	0.017
		70%	0.074	0.518	0.511	0.018
	$MI^{bagging}_{neigh=3}$	10%	0.080	0.520	0.509	0.017
		30%	0.081	0.515	0.510	0.017
		70%	0.077	0.517	0.511	0.017
	MI^{dim}_{bmu}	10%	**0.055**	0.519	0.514	0.019
		30%	**0.055**	0.516	0.504	0.020
		70%	**0.053**	0.516	0.500	0.021
	$MI^{dim}_{neigh=3}$	10%	0.057	0.518	0.509	0.018
		30%	0.057	0.515	0.505	0.019
		70%	0.054	0.516	0.497	0.020
Flow rate Measurement System Failure μ=0.540 median=0.555 var=0.024 unique=3084 (56%)	Mean	10%	0.117	0.540	-	-
		30%	0.124	0.540	-	-
		70%	0.125	0.544	-	-
	SI_{bmu}	10%	0.109	0.542	0.537	0.001
		30%	0.115	0.541	0.537	0.001
		70%	0.117	0.544	0.534	0.001
	$SI_{neigh=3}$	10%	0.114	0.541	0.536	0.000
		30%	0.120	0.540	0.535	0.000
		70%	0.121	0.543	0.539	0.000
	$MI^{bagging}_{bmu}$	10%	0.096	0.545	0.562	0.005
		30%	0.094	0.543	0.565	0.006
		70%	0.095	0.545	0.564	0.005
	$MI^{bagging}_{neigh=3}$	10%	0.097	0.545	0.557	0.005
		30%	0.095	0.544	0.562	0.005
		70%	0.096	0.544	0.561	0.005
	MI^{dim}_{bmu}	10%	**0.078**	0.539	0.564	0.007
		30%	**0.079**	0.541	0.567	0.008
		70%	**0.079**	0.542	0.564	0.008
	$MI^{dim}_{neigh=3}$	10%	0.079	0.541	0.564	0.007
		30%	0.081	0.542	0.565	0.007
		70%	0.081	0.543	0.564	0.007

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

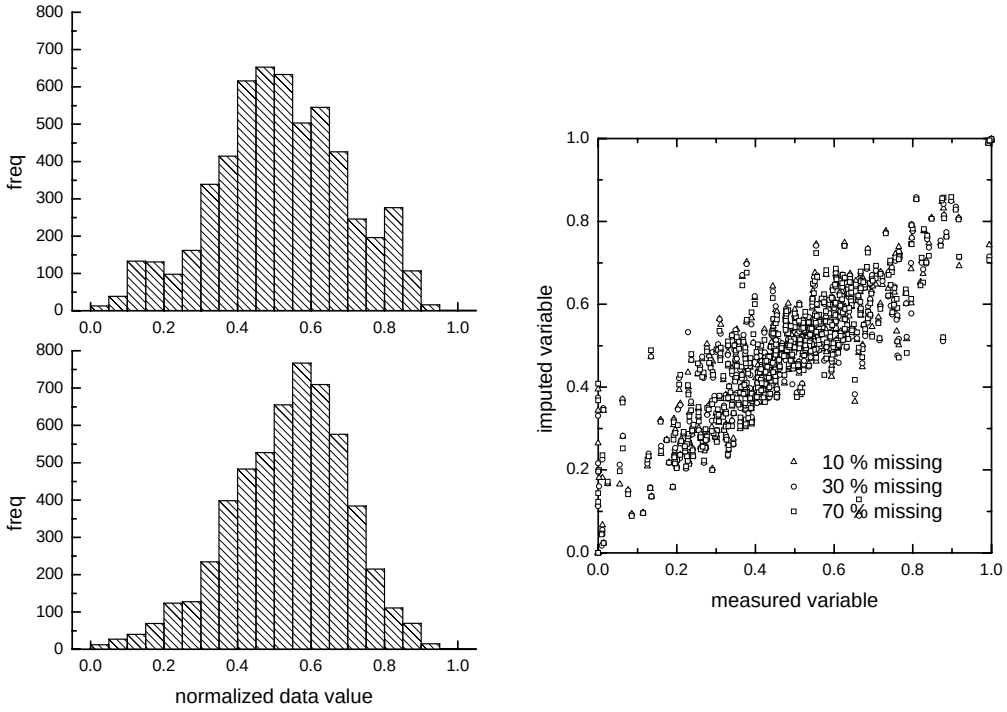


Figure 7.15. (Left) Frequency histograms for the original LDPE data corresponding to the temperature sensor. (upper) and the flow rate (lower). (Right) Measured and imputed variables corresponding to a set of random failures using the best multiple imputation model

The variance of a dataset is more difficult to reproduce by means of imputation techniques. For the temperature sensor, the model that reproduces best the variance yields a relative error of 32.2%. The situation is worse for the flow rate, with a relative error around 66.6%. This high value can be explained mainly because the data corresponding to flow rate is more diverse (56% of unique values) than the data for the temperature sensor (4.3%). This high amount of unique values cannot be handled conveniently by prototype-based imputation systems such is the SOM; (iv) All current imputation systems exhibit lower errors for the imputation of temperatures. That constitutes a clear indication of the fact that depending on the statistical properties of data being modeled, imputation systems based on the use of prototype-based classifiers will have low performance.

Table 7.9 summarizes the results obtained in the later case, where the misbehaving sensor is chosen randomly. This situation is equivalent to considering all the sensors in the process plant with the same uniform probability distribution of failure. In this situation, the results observed confirm that SOM-based imputation systems outperform mean-substitution techniques.

Table 7.9. Comparison of all the imputation models for a random sensor failure in the LDPE plant. Absolute mean errors for three different levels of missingness

	Imputation Model	% missing data	AME
Random sensor Failure	Mean	10%	0.116
		30%	0.119
		70%	0.119
	SI_{bmu}	10%	0.099
		30%	0.101
		70%	0.100
	$SI_{neigh=3}$	10%	0.103
		30%	0.105
		70%	0.105
	$MI^{bagging}{}_{bmu}$	10%	0.081
		30%	0.077
		70%	0.073
	$MI^{bagging}{}_{neigh=3}$	10%	0.082
		30%	0.078
		70%	0.075
	$MI^{dim}{}_{bmu}$	10%	0.068
		30%	0.067
		70%	0.065
	$MI^{dim}{}_{neigh=3}$	10%	0.069
		30%	0.068
		70%	0.066

Figure 1 shows that in all the cases the behavior of the imputation system remains stable independently of the amount of missing data since incomplete patterns were also used during training. Best results are also obtained when using the multiple imputation system implemented using maps of different sizes.

To corroborate the quality of the imputation process a hypothesis test to compare the means of the imputed data with the original dataset was also performed. In all the cases, at a significance level of 0.05 this test revealed that there are no statistically significant differences between the means of both samples. Unfortunately significant differences are observed for the variances.

7.1.5.2 Assessment under real Operating Conditions

The virtual sensor provided with the missing data recovering module, has been integrated within the process control scheme of a LDPE production plant. Figure 7.16 depicts the performance of the system. It can be observed that the virtual sensor produces more stable inferential measurements than the correlation approach. The high deviations observed for the correlation model are produced by

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

sensor faults. In this case the approach followed by the correlation system is to use the previous sensor value.

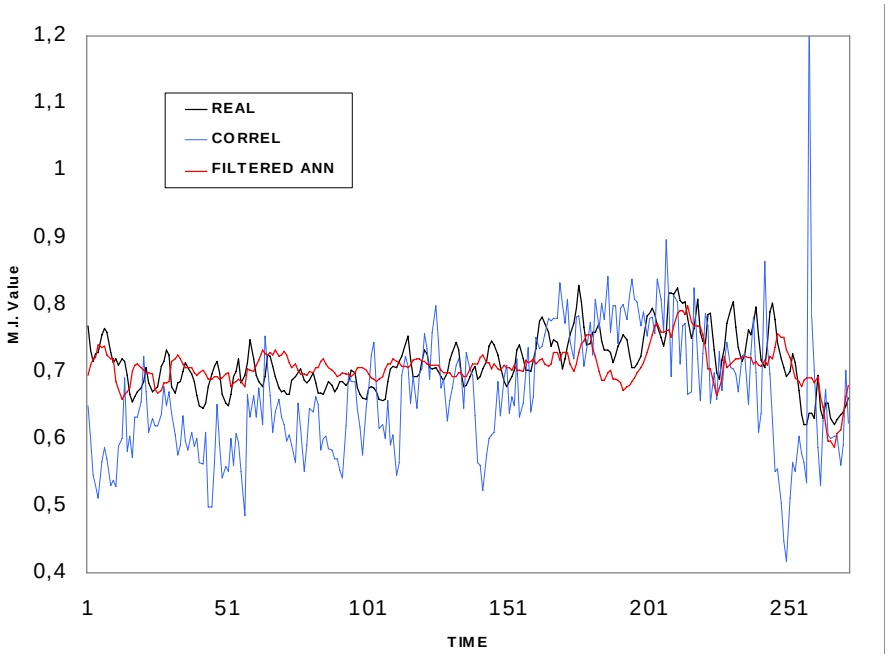


Figure 7.16. Data recorded from the virtual sensor operating in real-time conditions using the Fuzzy ARTMAP algorithm. Comparison with real MI data from laboratory and the estimations given by the correlation used in the process plant

7.2. Prediction of Carcinogenic Properties of Chemical Compounds from Molecular Descriptors

The application of the proposed framework to data-driven modeling is assessed here by the development of a Quantitative Structure-Activity Relationship (QSAR) model to estimate carcinogenicity of chemicals. The development of this QSAR model has been approached using the tiers of the proposed framework.

7.2.1 Problem statement

In recent years there has been an increasing research interest on quantitative-structure-activity relationship (QSAR) models for the prediction of biological activities of chemical compounds. Applications range from drug design (Senese and Hopfinger, 2003; Douali et al., 2003) to the assessment of toxicity of chemical pollutants (Espinosa et al., 2002; Mazzatorta et al., 2003a; 2003b). It is widely accepted that the biological activity of a chemical molecule is associated with its steric, electronic and lipophilic properties that are derived from molecular sites of action, usually receptors or enzyme active sites, with complex three-dimensional shapes. The intricate relationships between biological activity and molecular structure requires a careful selection of the number and type of descriptors necessary to accurately characterize the highly non-linear relationships among the molecular

descriptors and the biological activity of the molecule (Randic, 1975; Kier and Hall, 1976; Carbó et al., 1980).

Various novel approaches based in the use of artificial intelligence have been reported in the literature for the prediction of biological activities of organic compounds (Grauel et al., 1999; Ludwig et al., 1999) based on chemical structure information. Furthermore, the applications of novel computational learning methods to model QSAR are rapidly expanding (Amat et al., 1990; Vendrame et al., 1999; Espinosa et al., 2002). A key issue in the modeling process is the identification of a set of descriptors that are suitable from both model predictability and activity mechanisms points of view since a selection based on a fundamental understanding of the mechanism of action would be desirable. However, the present state-of-the-art with respect to mechanistic understanding of biological activity, does not lend itself to the identification of a descriptor set which is appropriate for a large and diverse group of compounds. As a consequence, it is a major challenge to find an appropriate set of descriptors for mapping the strong non-linear dependencies between the chemical structure and the activity of the chemical compound. Moreover, this set of descriptors should not contain irrelevant information that would act as uncontrolled error sources, i.e., noise.

A possible approach for the selection of the optimal set of descriptors to build accurate QSAR models is to rely on data mining and computational intelligence techniques (Burden et al., 2000; Izrailev and Agrafiotis, 2001; Yasri and Hartsough, 2001). Among the numerous methodologies available to extract relevant information (molecular descriptors) those that provide the clearest cause-effect relationships should be preferred. Thus, classification algorithms are a suitable option since they can cluster compounds according to structure and properties or activities by using the minimal set of molecular information. Self-organizing feature maps (Kohonen, 1990; Vesanto, 1999) (SOM) have been successfully applied to classify molecules into three-dimensional self-organized maps. The analysis of the shape and surface properties of these maps has provided valuable information about the biological activity of the molecules involved (Gasteiger et al., 1994; Anzali et al. 1998; Espinosa et al., 2002). The classification and visualization capabilities of SOM have been applied to select the most suitable set of molecular descriptors to predict target variables, while assuring the consistency of chemical classification. Similarity measures (Espinosa et al., 2002) were used to compare the resulting SOMs and remove redundant or irrelevant descriptors from the initial pool of information. This analysis is compatible with other statistical techniques such as correlation analysis, stepwise regression, and principal component analysis.

A variety of QSAR models have been reported in the literature employing different algorithms to map chemical structure information (chemical descriptors) to the carcinogenicity index (Vendrame et al., 1999; Gini et al., 1999; Mazzatorta et al., 2003a,b). Training algorithms vary from statistical multi-regression methods to intelligent soft computing techniques such as neural networks (Gini et al., 1999; Bahler et al., 2000). The performance of conventional neural network-based QSAR models can be improved by using ensembles of individual models generated with different architectures and/or different sampling strategies of the training set (Rallo et al., 2005). An alternative to this performance enhancing approach is the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

application of advanced cognitive classifiers, such as fuzzy ARTMAP. The advantages of fuzzy ARTMAP over regression or group-contribution based models, as well as those based on feed-forward neural networks, to build QSPR models for physicochemical properties of organic compounds have been reported and discussed elsewhere (Espinosa et al., 2000; Espinosa et al., 2001).

The aim of the current work is to develop an improved QSAR model for the prediction of the carcinogenicity potency (CP=log MW*1000/TD_{50}) of aromatic compounds with nitrogen containing substituents with a fuzzy ARTMAP neural network combined with a SOM-based methodology for the selection of descriptors. This approach is particularly appropriate not only because the data set used in the present analysis (Gini et al., 1999) consists of only 104 compounds but also to minimize false positive/negative identification of carcinogens. The chemical space is characterized by descriptors previously considered by Gini (1999) to be adequate to predict CP for this small set of chemicals. Additional quantum and topological molecular similarity descriptors are also considered. The effects of data scarcity are damped by introducing prototype vectors that represented the center of mass of each cluster found in the SOM classification process. The performance of the current approach is assessed by using (i) alternative feature selection algorithms, (ii) different training/test sets, and (iii) prediction methods.

7.2.2 Carcinogenic Activity Data and Molecular Information

The TD_{50} data set reported by Gini et al. (1999) consisted of 104 aromatic compounds containing nitrogen substituent. This homogeneous data set is well documented in the Carcinogenic Potency Database (Gold et al., 1984). The TD_{50} activity values for carcinogenicity represent the continuous dose in mg/kg that when administered to mouse will cause tumors in a half of the exposed population within the standard life span of two years. The list of chemicals and their corresponding experimental activity values are provided in Table 7.10. The carcinogenic potency $CP = \log(MW*1000/TD_{50})$, where MW is the molecular weight, is used thereafter as the target variable for the development and evaluation of QSAR models for carcinogenicity.

Table 7.10. Experimental and predicted carcinogenic potency CP (mg/kg) for the 104 aromatic compounds considered in the current study. Training and test chemicals are respectively identified by tr and te in the last column

No	Name	CAS	Exper.	Gini (1999)	Fuzzy ARTMAP	LOO	set
1	(n-6)-(methylnitroso)adenine	---	0.6665	0.4923	0.6665	0.5838	tr
2	(n-6)-methyladenine	443-72-1	0.0000	0.4462	0.0000	0.0000	tr
3	1,5-naphthalenediamine	2243-62-1	0.5838	0.5713	0.5838	0.6535	tr
4	1-(1-naphthyl)-2-thiourea	86-88-4	0.8274	0.6979	0.8264	0.7018	tr
5	1-amino-2-methylanthraquinone	82-28-0	0.5516	0.6981	0.4609	0.4630	te
6	1-[(5-nitrofurfurylidene)amino]hydantoin	67-20-9	0.4588	0.4831	0.4568	0.4276	tr
7	2,2',5,5'-tetrachlorobenzidine	15721-02-5	0.5963	0.6738	0.5963	0.6194	tr
8	2,2,2-trifluoro-n-[4-(5-nitro-2-furyl)-2-	42011-48-3	0.7321	0.6992	0.7321	0.6194	tr

ISBN:978-84-691-1008-9/DL: T.2230-2007

No	Name	CAS	Exper.	Gini (1999)	Fuzzy ARTMAP	LOO	set
	thiazolyl]acetamide						
9	2,4,5-trimethylaniline	137-17-7	0.7129	0.6384	0.7129	0.6498	tr
10	2,4,6-trimethylaniline.HCl	6334-11-8	0.6498	0.6310	0.6498	0.7129	tr
11	2,4-diaminoanisole sulfate	39156-41-7	0.4965	0.4567	0.4965	0.5963	tr
12	2,4-diaminotoluene.2HCl	636-23-7	0.5643	0.5146	0.5643	0.4458	tr
13	2,4-dimethoxyaniline.HCl	54150-69-5	0.4257	0.4197	0.4233	0.4952	tr
14	2,4-dinitrophenol	51-28-5	0.0000	-0.0145	0.0000	0.0000	te
15	2,4-dinitrotoluene	121-14-2	0.0000	0.3873	0.0000	0.0000	te
16	2,4-xylidine.HCl	21436-96-4	0.6608	0.5765	0.6363	0.4458	te
17	2,5-xylidine.HCl	51786-53-9	0.4458	0.5227	0.4451	0.6600	tr
18	2,6-dichloro-p-phenylenediamine	609-20-1	0.4405	0.4430	0.4384	0.6942	tr
19	2-acetylaminofluorene	53-96-3	0.7563	0.7638	0.7563	0.6665	tr
20	2-amino-4-(5-nitro-2-furyl)thiazole	38514-71-5	0.7243	0.6966	0.7018	0.7018	te
21	2-amino-4-(p-nitrophenyl)thiazole	2104-09-8	0.7133	0.6690	0.7129	0.7018	tr
22	2-amino-4-nitrophenol	99-57-0	0.4384	0.4929	0.4384	0.3238	tr
23	2-amino-5-nitrophenol	121-88-0	0.3238	0.3026	0.3238	0.4384	tr
24	2-amino-5-nitrothiazole	121-66-4	0.0000	-0.0862	0.0000	0.4233	tr
25	2-aminoanthraquinone	117-79-3	0.4630	0.6501	0.4609	0.5516	tr
26	2-aminodiphenylene oxide	3693-22-9	0.7344	0.7324	0.7321	0.7086	tr
27	2-biphenylamine.HCl	2185-92-4	0.4241	0.3075	0.4233	0.3807	tr
28	2-chloro-p-phenylenediamine sulfate	61702-44-1	0.4001	0.4022	0.4001	0.3807	tr
29	2-hydrazino-4-(5-nitro-2-furyl)thiazole	26049-68-3	0.6857	0.6391	0.7018	0.7018	te
30	2-hydrazino-4-(p-aminophenyl)thiazole	26049-71-8	0.7018	0.6003	0.7018	0.8264	tr
31	2-hydrazino-4-(p-nitrophenyl)thiazole	26049-70-7	0.7134	0.6021	0.7321	0.7321	te
32	2-methyl-1-nitroanthraquinone	129-15-7	0.8404	0.7969	0.8404	0.5516	tr
33	2-naphthylamine	91-59-8	0.6557	0.6456	0.6535	0.5838	tr
34	2-nitro-p-phenylenediamine	5307-14-2	0.4532	0.2208	0.4514	0.4514	tr
35	2-sec-butyl-4,6-dinitrophenol	88-85-7	0.8360	0.8256	0.836	0.8264	tr
36	3,3'-dimethoxybenzidine-4,4'-diisocyanate	91-93-0	0.2791	0.4109	0.2791	0.2717	tr
37	3-(3,4-dichlorophenyl)-1,1-dimethylurea	330-54-1	0.4788	0.6364	0.4769	0.4588	tr
38	3-chloro-p-toluidine	95-74-9	0.3807	0.3849	0.3807	0.3995	tr
39	3-nitro-p-acetophenetide	1777-84-0	0.3995	0.4186	0.3984	0.3995	te
40	4'-fluoro-4-aminodiphenyl	324-93-6	0.8306	0.5675	0.8306	0.7086	tr
41	4,4'-methylene-bis(2-chloroaniline).2HCl	64049-29-2	0.6141	0.6300	0.6178	0.6194	te
42	4,4'-methylenebis(n,n-dimethyl)benzenamine	101-61-1	0.5456	0.5662	0.5450	0.5922	tr
43	4,4'-methylenedianiline.2HCl	13552-44-8	0.6760	0.6068	0.6760	0.5922	tr
44	4,4'-oxydianiline	101-80-4	0.6680	0.5299	0.6665	0.5024	tr
45	4-amino-2-nitrophenol	119-34-6	0.0000	0.2578	0.0000	0.4514	tr
46	4-aminodiphenyl	92-67-1	0.8312	0.6604	0.8306	0.7086	tr
47	4-chloro-m-phenylenediamine	5131-60-2	0.4088	0.3924	0.4082	0.4233	tr
48	4-chloro-o-phenylenediamine	95-83-0	0.4233	0.4286	0.4233	0.4088	tr
49	4-chloro-o-toluidine.HCl	3165-93-3	0.6942	0.6084	0.6942	0.4088	tr
50	4-nitro-o-phenylenediamine	99-56-9	0.0000	0.3094	0.0000	0.0000	te
51	4-nitroanthranilic acid	619-17-0	0.2882	0.3812	0.2859	0.0000	tr
52	5-nitro-2-furaldehyde semicarbazone	59-87-0	0.6600	0.4923	0.6600	0.2859	tr
53	5-nitro-o-anisidine	99-59-2	0.4276	0.4530	0.4276	0.4384	tr
54	5-nitroacenaphthene	602-87-9	0.6194	0.6508	0.6178	0.7321	tr
55	Acetaminophen	103-90-2	0.0000	0.3215	0.0000	0.5100	tr
56	AF-2	3688-53-7	0.5922	0.5664	0.5922	0.5344	tr

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

No	Name	CAS	Exper.	Gini (1999)	Fuzzy ARTMAP	LOO	set
57	Aniline.HCl	142-04-1	0.2679	0.2523	0.2679	0.4276	tr
58	Anthranilic acid	118-92-3	0.1737	0.1693	0.1737	0.4952	tr
59	Atrazine	1912-24-9	0.6881	0.6902	0.7277	0.7277	te
60	Azobenzene	103-33-3	0.7571	0.7360	0.7563	0.7321	tr
61	Benzidine.2HCl	531-85-1	0.7086	0.6738	0.6703	0.6535	te
62	c.i. disperse yellow 3	2832-40-8	0.4769	0.4644	0.4769	0.3995	tr
63	Chloramben	133-90-4	0.3473	0.2602	0.3473	0.4384	tr
64	Chlorambucil	305-03-3	1.0000	0.9094	1.0000	0.9803	tr
65	Cinnamyl anthranilate	87-29-6	0.4017	0.4539	0.4001	0.4769	tr
66	D & c red no. 9	5160-02-1	0.4336	0.4384	0.4333	0.4876	tr
67	Dacarbazine	4342-03-4	0.8653	0.5674	0.8653	0.7277	tr
68	Dapsone	80-08-0	0.0000	0.4293	0.0000	0.0000	tr
69	fd & c red no. 4	4548-53-2	0.2512	0.2209	0.2512	0.2717	tr
70	fd & c yellow no. 6	2783-94-0	0.2717	0.2126	0.2512	0.2512	te
71	Fluometuron	2164-17-2	0.5344	0.4913	0.5344	0.5922	tr
72	Formic acid 2-[4-(5-nitro-2-furyl)-2-thiazolyl]hydrazide	3570-75-0	0.7277	0.6196	0.7277	0.7321	tr
73	Furosemide	54-31-9	0.4876	0.5560	0.4876	0.4844	tr
74	Hydrochlorothiazide	58-93-5	0.4514	0.5654	0.4514	0.4769	tr
75	m-cresidine	102-50-1	0.5100	0.5057	0.5963	0.5963	te
76	m-phenylenediamine.2HCl	541-69-5	0.4844	0.4144	0.4844	0.3807	tr
77	m-toluidine.HCl	638-03-9	0.3831	0.3642	0.3807	0.4276	tr
78	Melamine	108-78-1	0.3475	0.4286	0.3473	0.4333	tr
79	Melphalan	148-82-3	0.9803	1.0032	1.0000	1.0000	te
80	Methotrexate	59-05-2	0.6450	0.4927	0.6450	0.6498	tr
81	Metronidazole	443-48-1	0.4927	0.4924	0.4927	0.4927	tr
82	Mexacarbate	315-18-4	0.8264	0.8305	0.8264	0.3995	tr
83	n-(1-naphthyl)ethylenediamine.2HCl	1465-25-4	0.0000	0.2226	0.0000	0.4233	te
84	n-nitrosodiphenylamine	86-30-6	0.4952	0.4837	0.4927	0.5024	tr
85	n-phenyl-p-phenylenediamine.HCl	2198-59-6	0.0000	0.4836	0.0000	0.0000	tr
86	n-[4-(5-nitro-2-furyl)-2-thiazolyl]formamide	24554-26-5	0.7325	0.7051	0.7129	0.7129	te
87	n-[5-(5-nitro-2-furyl)-1,3,4-thiadiazol-2-yl]acetamide	2578-75-8	0.7440	0.5990	0.7440	0.7321	tr
88	Nithiazide	139-94-6	0.4609	0.4735	0.4609	0.4769	tr
89	Nitrofen	1836-75-5	0.6198	0.5780	0.6178	0.6141	tr
90	o-aminoazotoluene	97-56-3	0.5936	0.5913	0.5922	0.4703	tr
91	o-anisidine.HCl	134-29-2	0.4162	0.4160	0.4162	0.0000	tr
92	o-phenylenediamine.2HCl	615-28-1	0.4333	0.4379	0.4333	0.3807	tr
93	o-toluidine.HCl	636-21-5	0.4296	0.4531	0.3807	0.3807	te
94	p-anisidine.HCl	20265-97-8	0.0000	0.3522	0.0000	0.4162	tr
95	p-chloroaniline	106-47-8	0.3917	0.3951	0.3909	0.6942	tr
96	p-cresidine	120-71-8	0.5986	0.5234	0.5963	0.5100	tr
97	p-isopropoxydiphenylamine	101-73-5	0.4703	0.4558	0.4703	0.3995	tr
98	p-nitrosodiphenylamine	156-10-5	0.5024	0.6000	0.6665	0.6665	te
99	p-phenylenediamine.2HCl	624-18-0	0.3813	0.3249	0.3807	0.4844	tr
100	Pentachloronitrobenzene	82-68-8	0.6161	0.6816	0.6161	0.5963	tr
101	Phenacetin	62-44-2	0.2859	0.3255	0.2859	0.3807	tr
102	Phenylhydrazine	100-63-0	0.0000	0.0428	0.0000	0.0000	tr
103	Proflavine.HCl hemihydrate	952-23-8	0.6535	0.6667	0.6535	0.6760	tr
104	Pyrimethamine	58-14-0	0.5199	0.6243	0.5199	0.6141	tr

The topological and geometrical 2D and 3D molecular descriptors used in the present study are listed in Table 7.11. They include the descriptors previously

selected by Gini et al. (1999) as well as molecular quantum similarity (MQS) measures (Carbo-Dorca and Besalu, 1998).

Table 7.11. Symbols and definitions of molecular descriptors

Symbol	Definition	Description	References
		2-D	
${}^{m}_{t}\chi$ - ${}^{m}_{t}\chi^{v}$	Molecular conectivity and valence connectivity indices, order m=0-4	Measures structural features such as size, branching, unsaturation, heteroatom content and cycles	Randic M. 1984 Kier B. et al., 1986 Hall L. et al., 1991
κ	Kappa index or molecular shape index	Counts one-bond, two-bond and three-bond fragments, relatively to fragment counts in some reference structures	Kier L. 1989 Hall L. et al., 1991
N	Sum of atomic number	Number of protons	Mackay D., et al., 1993 Lyman W., 1990
*Randic	Randic Index	Quantify the molecular connectivity	Randic M., 1975 Randic M., 1984
*Balaban	Balaban Index	Measures the average distance sum connectivity	Balaban A., 1982 Balaban A., 1988
		3-D from semi-empirical PM3	
μ	Dipole moment	Product of the charge on a molecule and the distance between two charges of equal magnitude with opposite sign	Stewart J., 1989 McWeeny R. 1999
*ΔH_f	Heat of formation	Amount of energy per kilomole of substance required to form the molecule in standard conditions	Stewart J., 1989 McWeeny R. 1999
AP	Average molecular polarizability	Measures the response of the electron density distribution to a static electric field	Stewart J., 1989 McWeeny R. 1999
TE	Total energy	Total electronic energy of the molecule	Stewart J., 1989 McWeeny R. 1999
HOMO	Homo index	Highest occupied molecular orbital	Stewart J., 1989 McWeeny R. 1999
LUMO	Lumo index	Lowest unoccupied molecular orbital	Stewart J., 1989 McWeeny R. 1999
W^{3D}	Wiener index3D	Number of bonds in the shortest path connecting the every pair of atoms	Wiener, H., 1947 Bondanov B., et al., 1989
Ove, Cou, Kin	Molecular quantum similarity indices of Overlap, Coulomb and Kinetic operators	Molecular Quantum Similarity Measures (MQSM), which allow quantitative comparison between molecular electronic density distributions	Carbó-Dorca R., et al., 1998 Amat L. et al., 1997

* Taken from Gini et al. (1999)

The MQS descriptors establish quantitative similarity measures between molecular structures by means of the projection of their density functions. A MQS is defined as the volume integral of the product of two density functions weighted by a non-differential hermitic operator,

$$MQS_{ij} = \int \psi_i(\mathbf{r}) \Pi \psi_j^*(\mathbf{r}) d\mathbf{r} \qquad (7.2)$$

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

in which MQS_{ij} represents the molecular quantum similarity between molecules *i* and *j*, ψ_i and ψ_j are the corresponding density functions and Π is a non-differential hermitic operator.

Even though different MQS measures can be defined by the operator in Eq. (7.2), the most widely used are the overlap-like operator (Ove), the coulomb-like operator (Cou), and the kinetic-like operator (Kin). The calculation of these integrals is highly dependent on the relative orientation of the molecules and therefore the structures should be aligned to maximize this measure. When analyzing a set of compounds, all pairs of MQS are collected in a matrix, where the elements of the diagonal represent the self-similarity values. These elements can be used as descriptors in QSAR studies as suggested in a number of previous studies (Amat and Carbo-Dorca, 1997;Carbo-Dorca and Besalu, 1998; Amat et al., 1998).

Although the number of descriptors contained in the matrix can be very extensive for large data sets, feature selection algorithms can be applied to select the most adequate data from each matrix. In the present work, the SOM-based feature selection algorithm was used to select the best set of elements from the Overlap, Coulomb and Kinetic MSQ matrices to be used as input descriptors to the fuzzy ARTMAP based QSAR model. This procedure resulted in the selection of the overlap projection of the whole data set with compounds 50 and 101 (see Table 7.10), corresponding to 4-nitro-o-phenylenediamine and Phenacetin. Similarly, compounds 101 and 61, corresponding to 4-nitro-o-phenylenediamine and Benzidine.2HCl, were selected for the Coulomb MQS matrix. Finally, compounds 50 and 71, corresponding to 4-nitro-o-phenylenediamine and Mexacarbate were selected for the Kinetic MQS matrix. Through the above selection process the following MQS descriptors were considered: the three self-similarities (Ove, Cou, Kin) and the cross similarities of each compound with the above reference compounds (Ove_{50}, Ove_{101}, Cou_{61}, Cou_{101}, Kin_{50}, Kin_{71}).

7.2.3 Tier 2: Data Preprocessing

The methodology described in section 2.1 was applied to select the most suitable set of molecular descriptors for the prediction of CP. This approach was assessed by comparing current results with other well-known feature selection techniques such as Correlation-based Feature Selection-CFS (Hall, 2000) and the ReliefF method (Kononenko, 1994).

7.2.3.1 Selection of Molecular Descriptors

Table 7.12 summarizes the sets of molecular descriptors selected by using each of the above three methodologies, as well as those thoroughly selected by Gini et al (1999) with Principal Component Analysis (PCA).

In the current methodology SOM-dissimilarity approach the SOM was trained with input vectors formed by all descriptors and the target index, CP. The comparison of the resulting c-planes given in Figure 7.17 identifies redundant descriptors accounting for the same type of information.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 7.12. Comparison of indices selected by PCA (Gini et al., 1999) with those selected by using SOM, CFS and ReliefF

Molecular Descriptor	SOM	CFS	ReliefF	PCA
MW	●			●
Volume	●			
Size1	●			
Size2			●	
Size3	●			
Randic	●			
Balaban	●		●	●
$^{v}\chi^{1}$			●	
$^{v}\chi^{2}$	●		●	
$^{v}\chi^{3}$			●	●
$^{v}\chi^{4}$			●	
Flex	●	●	●	●
K_{A2}			●	
K_{A3}			●	
κ_1	●			
κ_2	●			
κ_3	●		●	
Electrotopological sum				●
Geometrical Indices				
Ellipsoidal volume				●
MOM1	●			
MOM3				●
Quantum Indices				
Heat	●			
MQS-Cou	●			
MQS-Kin	●	●		
HOMO			●	●
LUMO			●	●
$MQS\text{-}Kin_{71}$		●		
MQS-Ove	●	●		
Polarizability	●	●		●
Total Energy	●			
Dipole Moment				●
Physicochemical Parameters				
logD(pH=2)				●
logD(pH=7.4)	●	●		
logD(pH=10)	●	●	●	●
Total number of molecular descriptors selected	**20**	**7**	**13**	**13**

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

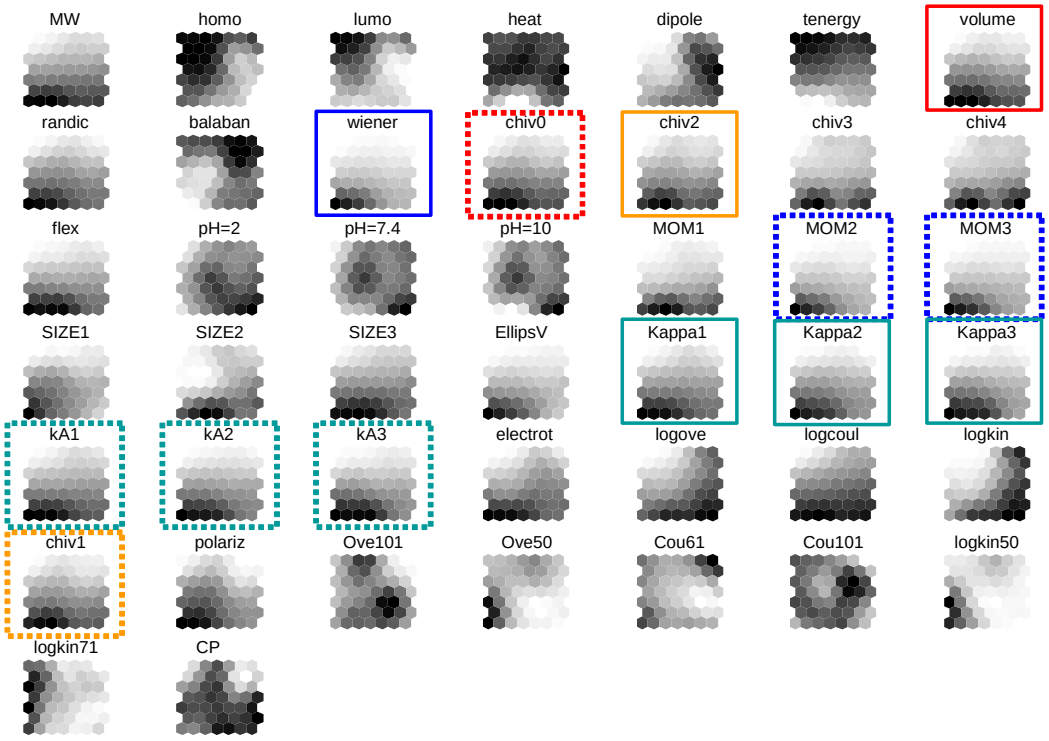


Figure 7.17. C-planes for the complete set of descriptors and carcinogenic potencies. Redundant descriptors are framed

Figure 7.17 shows that descriptors {volume, χ^0} have very similar c-planes and the same applies to {Wiener,MOM2,MOM3}, {Kappa1,k_{A1}}, {Kappa2, k_{A2}}, {Kappa3, k_{A3}} and {χ^2, χ^3}. The redundant descriptor with the highest correlation with the target property (solid line color frame in Figure 7.17) was kept while the others were discarded.

The next step in data preprocessing consisted in the classification of descriptors according to their c-planes after excluding redundant information. A new SOM was created and c-planes were extracted and classified using the K-means algorithm. The optimal number of clusters was obtained following an iterative procedure to minimize the Davies - Bouldin index (Davies and Bouldin, 1979). The consistency of the twelve classes obtained was checked by means of the covariance between all variables. It was observed that while some classes were formed solely by quantum-chemical descriptors, others included both topological and quantum information but with highest correlations for the quantum descriptors.

Table 7.13 shows the combination of these individual classes into four clusters. It should be noted that this grouping implies that only descriptors within the same cluster would contribute with the same type of information to the QSAR model for CP.

Table 7.13. Clusters of the descriptors c-planes obtained by minimizing the Davies - Bouldin index

Clusters from the Davies-Bouldin Index	Classes detected by SOM	Molecular descriptors
Cluster I	Class 1	Balaban Cou_{101}
	Class 2	Heat logD(pH=7.4) logD(pH=10)
	Class 3	logD(pH=2)
	Class 4	Ove_{101}
Cluster II	Class 5	Homo Lumo Tenergy Ove_{50} Cou_{61}
	Class 6	$logKin_{50}$ $logKin_{71}$
Cluster III	Class 7	Dipole LogKin
	Class 8	Chiv3 Chiv4 Size2 logOve
Cluster IV	Class 9	Wiener Size1 EllipVolume
	Class 10	MOM1 Electrot LogCoul
	Class 11	MW Chiv2 Flex Size3 Kappa1
	Class 12	Volume Randic Kappa2 Kappa3 Polariz

A representative index from each cluster having a correlation value with the target variable higher than the average correlation for the complete set of descriptors was then selected to form the initial set of descriptors. The selected indices were polarizability and log D (pH=10), which respectively belong to clusters IV and I in Table 7.13. The most suitable set of descriptors was completed by adding sequentially more descriptors in order of decreasing absolute covariance of their c-planes with that of the target variable CP, independently of the cluster to which each

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

c-plane belonged to. The ordered addition of indices was finalized when no more relevant information resulted from the inclusion of more descriptors. This is indicated by a minimum (or stabilization) in the average dissimilarity curve obtained by comparing different maps formed by these sets of descriptors, as depicted in Figure 7.18. This figure shows how dissimilarity diminishes when 18 additional descriptor [Coulomb, flex, volume, Randic Kappa2, MW, Kappa1, SIZE3, $\chi(2)$, Ove, Kappa3, SIZE1, tenergy, MOM1, log D (pH=7.4), Kin, ΔHf, Balaban] ordered by their c-plane correlation with CP are added to the two initially selected ones [polarizability and log D(pH=10)].

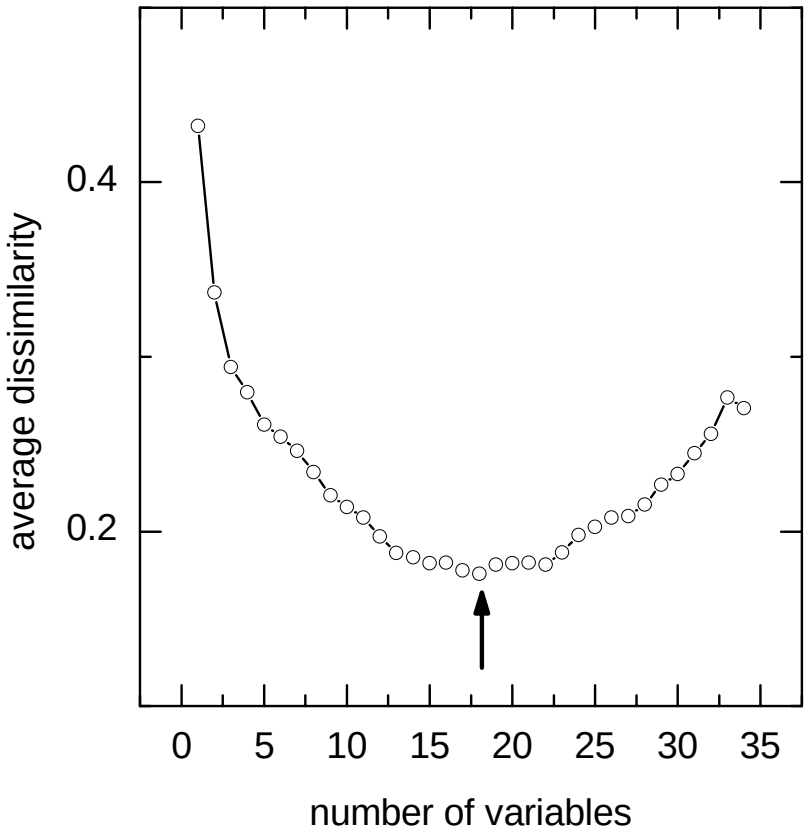


Figure 7.18. Selection of 20 indices by using the SOM-dissinilarity approach

Examination of Table 7.12 shows that the SOM-dissimilarity methodology selects the largest set of 20 molecular features, which approximately doubles the number of those selected by the other methods considered. It should be noted that the current cluster-based selection procedure can detect small changes in the classification of chemicals caused by the consideration of more descriptors in the vectors presented to the SOM. The current approach selects more descriptors because they contribute to improving the compactness and separation of chemical classes in the SOM. All feature extraction methodologies included in Table 7.13 select *flex* and *logD(pH=10)*. *Polarizability* is also selected in all cases except for ReliefF.

7.2.3.2 Generation of Optimized Train/Test Sets

To build the QSAR model with the above most suitable set of 20 descriptors and also with those given in Table 7.13, it is necessary to split the entire data set into the training and test sets. This was accomplished by using one ART module of fuzzy ARTMAP. Data were first classified and only the compounds belonging to classes with two or more elements were considered candidates to test the QSAR. This assures uniformity of the test group. The selection of the best training set is crucial to develop accurate QSAR models from sets of data that are very limited in size. Table 7.10 identifies the compounds used for training (tr) and testing (te) the neural system. The results obtained with the trained fuzzy ARTMAP architecture as well as with other techniques are presented and discussed in the next subsection.

7.2.4 Tier 3: QSAR Modeling

Several models have been calculated to compare the proposed integrated neural network approach with previous QSARs. (Gini et al., 1999; Rallo et al., 2005) by using the different sets of descriptors selected with current and other well-known machine learning algorithms. The results of the two groups of calculations A and B carried out are respectively summarized in Tables 7.14 and 7.15. Group A includes a series of experiments comparing the performance of different neural architectures for the prediction of the CP, with and without considering the enrichment of the original data set of chemicals with the SOM prototypes representative of the clusters that constitute the map. Group B includes the results obtained by using the most suitable set of descriptors selected by the SOM dissimilarity methodology. The description of the different experiments contained in each group is as follows:

Group A: The set of descriptors proposed by Gini et al. (1999) [MW, Homo, Lumo, Balaban, μ, AP, χ^3, flex, logD (pH=2), logD (pH=10), TE, MOM3, and ellipsoidal volume] were used to build the QSAR models.

A.1: Results obtained using the backpropagation architecture 13-6-1 reported by Gini et al. (1999): (a) Original results reported by these authors; (b) the complete data set split into training and testing sets, keeping 85% of the data set for training and 15 % for testing. The purpose is to reproduce the published results with the present backpropagation algorithm. Nevertheless, since these authors do not report the training and testing chemicals used in their QSAR model it was necessary in the current work to apply fuzzy ART to generate those sets (85% and 15%, respectively). However, comparisons are made in terms of all data points (see for example Figure 7.19).

A.2: Results obtained using an optimized backpropagation architecture 13-17-1 and enriching the data set with SOM prototypes.

A.3: Results obtained using a fuzzy ARMAP neural network. (a) The complete data set split into training (85%) and testing (15%) with fuzzy ART. (b) Training with both the whole data set and SOM prototypes. The vigilance parameter ρ was decreased from 0.999 to 0.995 to enhance the generalization capabilities of the model.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Table 7.14. Group A calculations with all models obtained with the original set of descriptors reported by Gini et al. (1999). Effect of backpropagation optimization, pre-classification of data in training-test sets, and of data enrichment with the inclusion of SOM prototypes

	Data Set	Number of samples	AME	σ
Original results reported by Gini	all data	104	0.0856	0.1053
Backpropagation (13-6-1) using pre-classified training/test set partition	train	85	0.0623	0.0551
	test	19	0.2790	0.2376
Optimized backpropagation (13-17-1) using pre-classified training/test sets and including SOM prototypes	train	110	0.0193	0.0240
	test	19	0.2375	0.2354
Fuzzy ARTMAP using pre-classified training/test sets (ρ=0.999)	train	85	0.0004	0.0008
	test	19	0.1260	0.1455
Fuzzy ARTMAP using pre-classified training/test sets and including SOM prototypes (ρ=0.995)	train	110	0.0426	0.0906
	test	19	0.0504	0.0707

Group B: Results obtained by generating the QSAR models with the most suitable set of 20 descriptors selected by the SOM dissimilarity procedure. [Polarizability, log D(pH=10), Coulomb, flex, volume, Randic κ_2, MW, κ1, SIZE3, $^{v}\chi^2$, Ove, κ_3, SIZE1, TE, MOM1, log D (pH=7.4), Kin, ΔHf, Balaban.].

B.1: Results obtained using a MLR model by splitting the whole data set into training (85%) and testing (15%) with fuzzy ART.

B.2: Results obtained using an optimized backpropagation architecture 20-30-1 and splitting the whole data set into training (85%) and testing (15%) with fuzzy ART.

B.3: Results obtained using a fuzzy ARMAP neural network. (a) Complete data set split into training (85%) and testing (15%) with fuzzy ART. (b) Training with both the entire data set and SOM prototypes. The vigilance parameter ρ was lowered from 0.999 to 0.995 to increase the generalization capabilities of the model.

Table 7.14 (group A models) shows that the performance of the Gini et al. (1999) QSAR model (13-6-1 backpropagation) is remarkable with an absolute mean error (AME) of 0.0856 for the entire data set. When the data set was split into 85 chemicals for training and 19 for testing the AME for the 19 test compounds slightly decreases from 0.2790 for the backpropagation (13-6-1) architecture used previously(Gini et al. 1999) to 0.2375 when both SOM prototypes, which act as virtual compounds, and a (13-17-1) architecture are considered. Figure 7.19 illustrates the slight improvement attained by considering SOM prototypes with the backpropagation architecture. These errors further decreases to 0.1260 when the fuzzy ARTMAP algorithm is used. The vigilance parameter for fuzzy ARTMAP was

ISBN:978-84-691-1008-9/DL: T.2230-2007

decreased to 0.995 when working with the training set enriched with SOM prototypes to allow for better generalization of the model. Despite the penalty in error over the training set, which increases to AME=0.0426, the prediction error over the test set further decreases to AME=0.0504 when SOM prototypes are considered, i.e., decreases another 50% compared with the same architecture without the inclusion of prototypes.

Table 7.15. Group B calculations with all models obtained with the set of descriptors selected by the SOM-dissimilarity procedure, CFS and ReliefF

			SOM	**CFS**	**ReliefF**
	Data Set	**Number of samples**	**AME (σ)**	**AME (σ)**	**AME (σ)**
Multilinear regression model using pre-classified training/test sets	Train	85	0.1354 (0.1190)	0.1748 (0.1304)	0.1526 (0.1280)
	Test	19	0.2375 (0.1447)	0.2187 (0.1470)	0.1835 (0.1723)
Optimized backpropagation[a] using pre-classified training/test sets	Train	85	0.0284 (0.0287)	0.0793 (0.0709)	0.0682 (0.0521)
	Test	19	0.2514 (0.1346)	0.2090 (0.1905)	0.1580 (0.1275)
Fuzzy ARTMAP using pre-classified training/test sets (ρ=0.999)	Train	85	0.0004 (0.0008)	0.0004 (0.0007)	0.0004 (0.0007)
	Test	19	0.0389 (0.0367)	0.1595 (0.1652)	0.1529 (0.1257)
Fuzzy ARTMAP using pre-classified training/test sets and including SOM prototypes (ρ=0.995)	Train	134	0.0006 (0.0009)	-	-
	Test	19	0.0323 (0.0416)	-	-
Ensemble of backpropagation networks (20-30-1) trained using bagging Size of ensemble=3 Averaged 50 times	Train	185	0.0887 (0.1005)	-	-
	Test	19	0.1170 (0.1299)	-	-

[a]The optimized backpropagation architectures are 20-30-1, 7-10-1 and 13-20-1, for SOM, CFS and ReliefF, respectively

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

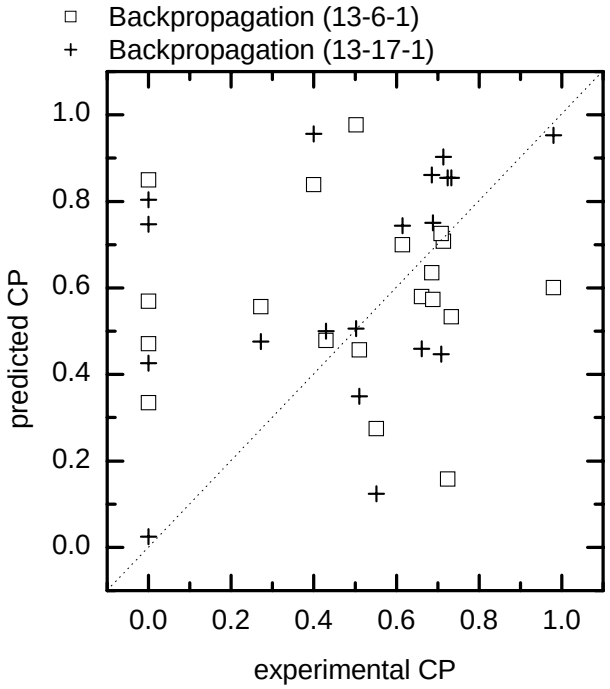


Figure 7.19. Comparison between the optimized backpropagation (13-17-1) including SOM prototypes for CP with the architecture (13-6-1) proposed by Gini et al. (1999)

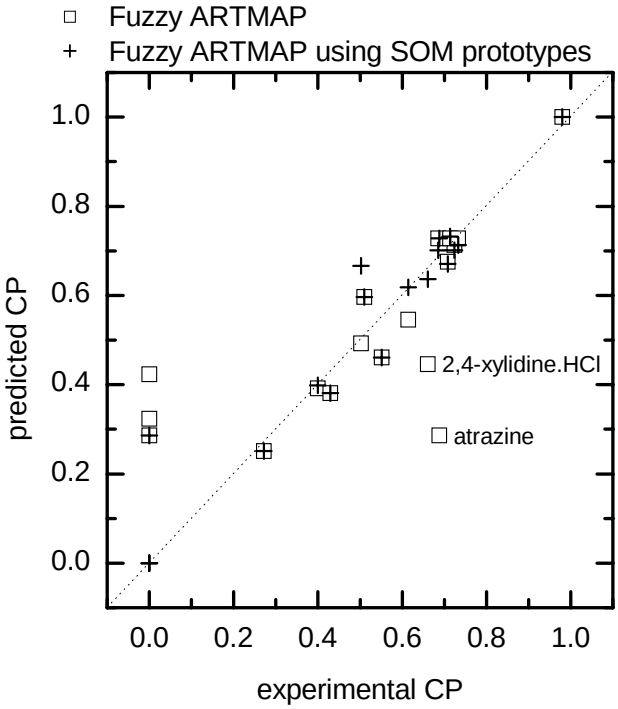


Figure 7.20. Fuzzy ARTMAP-based QSAR for CP with the descriptors proposed by Gini et al. (1999) and the inclusion of SOM prototypes into the training set

Figure 7.20 illustrates the favorable enrichment effect that is attained by adding 25 SOM prototypes (map with a 5 x 5 hexagonal grid) into the training set of 85 chemicals when building the QSAR model with a classification algorithm such as fuzzy ARTMAP. These results show that the adoption of Fuzzy ARTMAP to build QSAR models and the inclusion of SOM prototypes into the training set improve absolute mean errors significantly with respect to backpropagation-based models. Figure 7.20 indicates that the most important source of error arises from the misclassification of non-carcinogenic compounds. Another significant source of error, which is reduced by the inclusion of SOM prototypes in the fuzzy ARTMAP model, corresponds to atrazine and 2,4-xylidine.HCL. The inclusion of SOM prototypes, however, does not affect the non-carcinogenic group of chemicals, indicating that the pool of descriptors used might be insufficient to map the chemical space.

The best fuzzy ARTMAP-based model for CP outperforms previous QSARs based on conventional neural network algorithms reported in the literature (Gini et al., 1999). Figure 7.21 depicts this improvement under equivalent conditions, i.e., complete dataset without SOM prototypes.

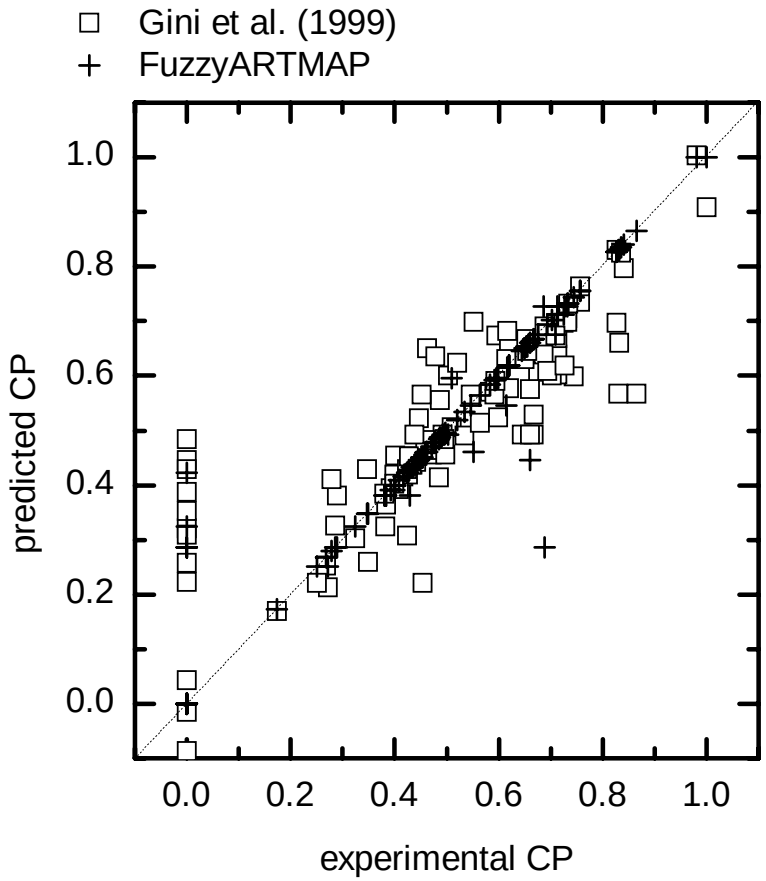


Figure 7.21. Comparison between the CP predicted by Gini et al. (1999) with a backpropagation neural network and by the current Fuzzy ARTMAP algorithm trained with the same set of indices for the complete dataset

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The effects of improving the approach to select the most suitable set of molecular descriptors to build the QSAR model (group B calculations) are summarized in Table 7.15. The application of the SOM-dissimilarity feature extraction approach to select the best set of descriptors (see table 7.12) instead of PCA, does not improve significantly the backpropagation-based models, as observed when comparing results in Tables 7.14 and 7.15. Multilinear regression (MLR) models perform equally well as backpropagation models. The most noticeable gain is observed in Table 7.15 when using classifiers both in the selection procedure (SOM) and in the QSAR model (fuzzy ARTMAP). When fuzzy ARTMAP-based models are respectively trained without and with SOM prototypes, errors decrease from AME=0.1260 and 0.0505 in Table 7.14 for the Gini descriptors (PCA) to AME=0.0389 and 0.0323 for the current descriptors selected with SOM-dissimilarity measures. The corresponding standard deviations of σ=0.0367 and 0.0416 for these two SOM-fuzzy ARTMAP QSARs are also the lowest of all models. The excellent performance of these two models is depicted in Figure 7.22. Comparison of this figure with Figures 7.19-7.21 illustrates the improvement accomplished with the integration of SOM with fuzzy ARTMAP to develop QSAR models.

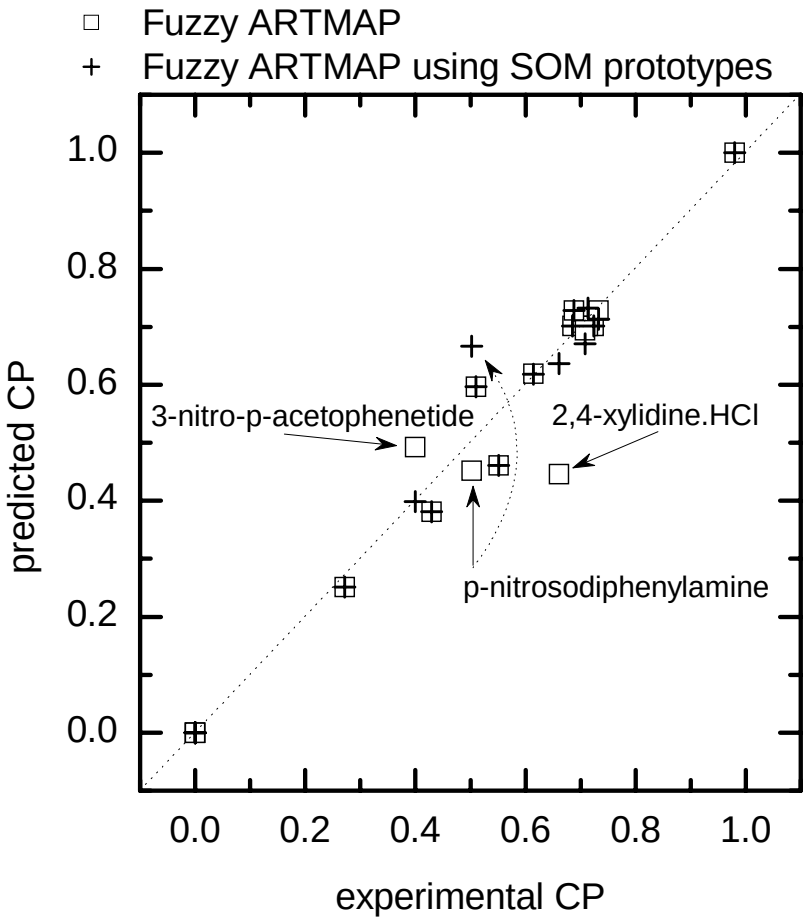


Figure 7.22. Performance of the best fuzzy ARTMAP-based QSAR models for CP with molecular descriptors selected by SOM-dissimilarity measures

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Figure 7.22 illustrates that the main source of error for the best QSAR model (integrated SOM-fuzzy ARTMAP approach) corresponds to p-nitrosodiphenylamine, 3-nitro-p-acetophenetide and 2,4-xylidine.HCl. The addition of SOM prototypes increases the accuracy of the Fuzzy ARTMAP predicted CP for 3-nitro-p-acetophenetide and 2,4-xylidine.HCl, but not for p-nitrosodiphenylamine. In this case all non-carcinogenic chemicals are classified correctly. Since the above results indicate the importance of selecting an adequate set of molecular descriptors to establish QSAR for CP, other feature selection algorithms, such as the CFS and the ReliefF algorithms, where also applied in the current work. Table 7.15 also present the results obtained with these two feature extraction algorithms in combination with MLR, backpropagation, and Fuzzy ARTMAP-based models.

The descriptors selected by each technique are listed in Table 7.12. The CFS algorithm, which extracts the smallest set of 7 descriptors, yields comparable or slightly better results than PCA descriptors for backpropagation models (Table 7.14) and than SOM-dissimilarity measures (Table 7.15) for both MLR and backpropagation models. In the case of fuzzy ARTMAP models, the molecular information extracted by CFS is unable to classify well all compounds and errors are five times larger than for the best fuzzy ARTMAP-based model, as shown in table 7.15. The ReliefF algorithm, which extracts 13 descriptors, shows a trend similar to CFS, i.e., errors five times larger than for the best fuzzy ARTMAP-based model. Table 7.15 indicate that these two feature extraction algorithms (CFS and ReliefF) yield better models when applied in conjunction with MLR and backpropagation neural networks, while the current classification-based SOM-dissimilarity approach yields best results with fuzzy ARTMAP. The differences between CFS, ReliefF and SOM-dissimilarity measures when coupled with fuzzy ARTMAP-based models can be explained based on the operational characteristics of each feature selection algorithm. While CFS is mainly correlation based and ReliefF uses a probabilistic approach, the SOM-dissimilarity measures takes into account not only correlation measures with the target property but also structural relationships between descriptors, as well as their distribution over the whole feature space. It can be surmised that SOM explores the chemical information space detecting all descriptors that are relevant for mapping the target property CP.

A leave-one-out (LOO) cross validation analysis of the Fuzzy ARTMAP model was performed. The results obtained for each compound are summarized in Table 7.10. The absolute mean error obtained in this case is 0.099, which is lower than the errors reported for all MLR and backpropagation architectures. When compared to the errors of the best Fuzzy ARTMAP models given in Table 7.15 and Figure 7.22, with split training and test sets, i.e., avoiding the problem of isomers, it is three times higher, indicating the potential of the fuzzy ARTMAP algorithm to build QSAR models under very demanding situations (scarce dataset and very similar compounds) and with good generalization capabilities, i.e., without over fitting (Hawkins, 2004. Results in Table 7.10 indicate that LOO cross validation doesn't provide a good error estimate for the problem under consideration. This is so because the current dataset is scarce and contains either very similar chemicals or classes with single-chemicals. Gini et al. (1999) reported that p-anisidine.HCL (CP=0) and o-anisidine.HCl (CP=0.4162) are structurally equivalent compounds but

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

with conflicting carcinogenic properties. The LOO procedure predicts in Table 7.10 a CP = 0 for the o-isomer instead of CP = 0.4162 when the p-isomer was included for training and vice versa, it predicts a CP = 0.4162 for the p-isomer when the o-isomer was included in the training set. A similar situation occurs with isomers 2,4-xylidine.HCl (CP=0.6608) and 2,5-xylidine.HCl (CP=0.4458), and with 4-chloro-m-phenylenediamine and 4-chloro-o-phenylenediamine, which are all carcinogenic.

Rallo et al. (2005) applied ensembles of backpropagation networks to enhance the predictive capabilities of feedforward neural algorithms for carcinogenicity (CP). This approach increased the performance of single models. Notwithstanding, the current fuzzy ARTMAP model outperforms the results obtained using the ensembles of conventional neural network models, as shown in Table 7.15. The best ensemble QSAR reported by these authors, which was generated with a combination of three backpropagation networks with a 20-30-1 architecture and trained using a bagging approach, yielded predictions with an absolute mean error of 0.117 compared to 0.032 for the current fuzzy ARTMAP model, as indicated in Table 7.15.

The smallest errors for the test set reported in Tables 7.14 and 7.15 for the cognitive fuzzy ARTMAP classifier can be attributed to the capability of this algorithm to classify chemicals as either carcinogenic or non-carcinogenic when sufficient chemical information is provided and the test set is representative of the training set and excludes isomers. In the original paper of Gini et al. (1999), 12 compounds (9 non-carcinogenic and 3 carcinogenic) with prediction errors higher than 0.2 were considered outliers of the model and, thus, neglected. In the current approach none of the information from the original dataset was discarded. The prediction of whether a compound is potentially carcinogenic or not is a main task embedded in the QSAR model. The intelligent design of the training and test sets using fuzzy ART includes most of these compounds into the training set and confirms that most of them are unique, i.e., constitute a single chemical class. The additional information contributed by SOM prototypes mitigates the scarcity of data. It is worth noting that a significant contribution to the CP prediction error corresponds to 2,4-dinitrotoluene, for which a CP = 0.299 is predicted instead of zero. This chemical was also in the class of chemicals with zero carcinogenicity potency for which the model of Gini et al. (1999) yields the highest errors, including negative CP values. The above observations suggest that more toxicity data are needed to account for such outliers. Nevertheless, it is instructive to briefly review the behavior of the current models with respect to the prediction of these non carcinogenic (CP=0) compounds: the number of non carcinogenic compounds in the original data set is 12, from which 4 compounds are used in all test sets. These are: 2-4 dinitrophenol, 2-4 dinitrotoluene, 4-nitro-o-phenylenediamine and n-(1-naphthyl)ethylenediamine.2HCl.

Predictions of carcinogenicity CP values for the above four non-carcinogenic compounds (in the order written above) resulting from the different models using indices reported by Gini et al. (1999) are respectively as follows: (a) for the A.1 (backpropagation 13-6-1) model they are 0.3345, 0.8497, 0.4788 and 0.5691; (b) for the A.2 (backpropagation 13-17-1) model they are 0.7471, 0.0250, 0.4262 and 0.8041; (c) for the A.3 (fuzzy ARTMAP) model they are 0.2859, 0.2859, 0.3238, and

0.4233; (d) for the fuzzy ARTMAP model with SOM prototypes the CP values are 0.2859, 0.2859, 0.000, and 0.000. It is important to note that the use of a classifier such as fuzzy ARTMAP is the reason for the correct prediction of two of the four non-carcinogenic chemicals. The inclusion of prototypes enables increased accuracy of the classification process. The behavior is equivalent for models using the set of descriptors selected using SOM-dissimilarity measures. In this latter case, however, the fuzzy ARTMAP model with the most suitable set of indices and including SOM prototypes classified correctly all the compounds into the non-carcinogenic class. From the above observations it can be concluded that the correct prediction of non-carcinogenic compounds is likely to be associated with the architecture of the predictive model to a greater degree than with the specific set of molecular descriptors used. Overall, the fuzzy ARTMAP classifier was found to be more reliable than backpropagation neural network based QSARs for all cases studied with the current data set for carcinogenicity.

7.3. Ecological Risk Assessment and Mapping

In this subsection a novel environmental application for the components of the proposed framework is presented. The Self-Organizing Map (SOM) algorithm is used at different levels of the Ecological Risk Assessment Framework (ERA) to generate probabilistic risk maps from incomplete field data measurements. The proposed methodology is assessed in a groundwater pollution scenario.

7.3.1 Motivation and Overview

Our society is increasingly aware of ecological issues including climate change, acid deposition, biological diversity, and the ecological impacts of xenobiotic compounds such as pesticides and toxic chemicals. Ecological risk assessment can help identifying these environmental problems, establish priorities, and provide a scientific basis for regulatory actions.

A successful assessment of these ecological problems requires the use of intelligent tools able to visualize and extract causal relationships among stressors and their effects. The main challenges related to the application of these techniques are the highly non-linear nature of the relationships between the stressors and their effects as well as the lack of reliable data. Machine learning algorithms and intelligent data-mining have proven to be very successful in the management of high dimensional datasets in the presence of uncertainty. The SOM algorithm (Kohonen, 1990) has demonstrated to be an appropriate tool for the classification and visualization of complex data. A detailed description of the self-organizing map algorithm and its properties is provided in Chapter 3.

In the following subsections the applications of the SOM in different areas of ecological risk assessment are evaluated. A set of new methods based on SOM to perform (i) data exploration and analysis, (ii) data imputation, (iii) spatial interpolation, and (iv) risk mapping are described and tested.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

7.3.2 Exposure and Risk Assessment

The intrinsic classification and visualization capabilities of SOM can be exploited to implement a novel approach to assess and map ecological risk. In this context, several applications can be considered for the SOM within the ERA framework. The following subsections briefly outline some of these application areas.

7.3.2.1 TIER 1: SOM based EDA for scenario selection

The estimation of exposure and risk often involves the assessment of a large number of different environmental parameters and system configurations. The selection of the most appropriate scenarios requires the use of classifiers capable of grouping similar configurations of input variables into coherent classes to characterize a concrete exposure and the corresponding risk scenarios. SOM provides a number of visualization methods which are useful for data analysis (see Chapter 3 for a detailed discussion on this topic). The comparison of the component planes (c-planes) of each input variable provides a direct indication of their relationships and information content, as shown in Figure 7.23. The characterization of environmental risk scenarios for Catalunya is used to assess the use of SOM-based EDA for ERA. Catalunya is divided into 41 administrative counties for which a set of 22 geographic, environmental, human activity and ecological indicators have been collected and classified using SOM.

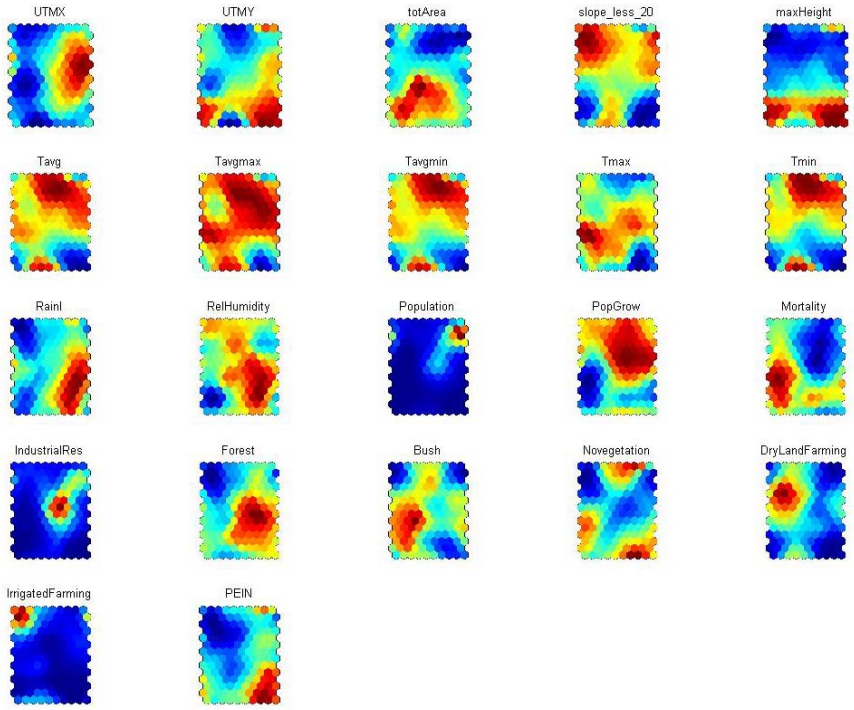


Figure 7.23. C-planes resulting after performing the SOM-based EDA for the 41 counties of Catalunya as characterized by 22 geographic, environmental, human activity and ecological indicators

ISBN:978-84-691-1008-9/DL: T.2230-2007

Figure 7.23 depicts the clustering of 41 counties of Catalunya using a multidimensional set of parameters including UTM coordinates, elevation, temperature, rainfall, humidity, population, land-use, urban waste, and forest areas. At a preliminary screening level, the SOM is used to extract relations between these indicators. The visual inspection of c-planes presented in Figure 7.23 quickly reveals relationships such as the inverse dependency between Population Grow (*PopGrow*) and Mortality index. Less evident and more complex relationships also arise from this analysis; for instance, it can be seen that counties with a substantial percent of area dedicated to farming (either dry or irrigated) are grouped in the upper-left corner of the map (see *DryLandFarming* and *IrrigatedFarming* c-planes in Figure 7.23). This area corresponds to flat lands (counties with large areas with a slope less to 20%; *slope_less_20*), with moderate temperatures (*Tavg*) and with low population (*Population*).

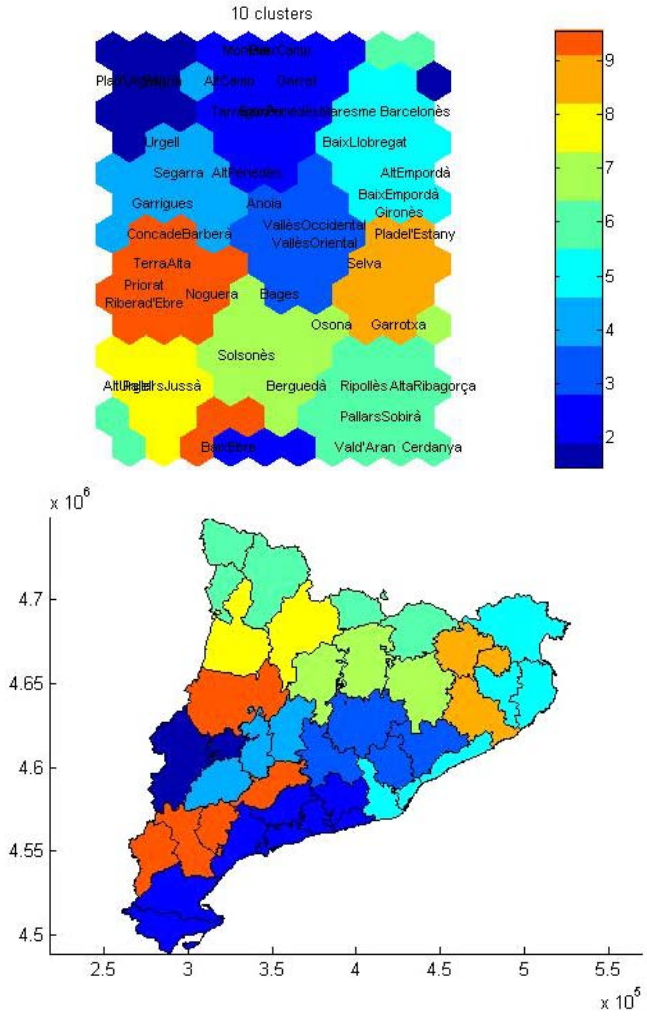


Figure 7.24. Classification of the 41 Catalan counties according to the SOM prototypes in Figure 7.23

The prototype vectors of each SOM unit in Figure 7.23 completely describe representative scenarios for all counties having this node as their best matching unit

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

(BMU). To obtain more compact classes, the prototype vectors can be further grouped into clusters using the minimization of the Davies-Bouldin index previously described in Chapter 3. Figure 7.24 depicts the clustering of the SOM c-planes shown in figure 7.23 using the K-means algorithm with a k value optimized to obtain good cluster partitions. Each cluster groups counties that can be considered as multiple instances of the same basic risk scenario. Using this strategy, the complexity involved in risk assessment can be reduced, and the results obtained can be extrapolated to other scenarios belonging to the same class (same BMU). In the case study corresponding to Catalunya, the optimal clustering in Figure 7.24 results in 10 classes. Thus, the complexity of the exposure and risk analyses is reduced from 41 to 10 distinct scenarios. It is important to note that the inclusion of the geographical UTM coordinates produces clusters in which spatial continuity is preserved.

7.3.2.2 TIER 2: Recovery of missing environmental data

Environmental datasets are prone to be incomplete. In addition to the very limited number of data stations usually considered, sensor failures and manual data recording procedures are the main sources of data scarcity. SOM is an efficient methodology to infer the values of missing data based on the classification of existing measures (Rallo et al., 2005). The map captures the probability distribution of the dataset and therefore it can be used as an imputation mechanism to infer any missing value. The quantization error of the map (first term in Eqn. 3.17) indicates the reliability of imputed values. In this subsection the SOM-based data imputation method proposed in Chapter 4 is applied to the groundwater pollution scenario described in Figure 7.25.

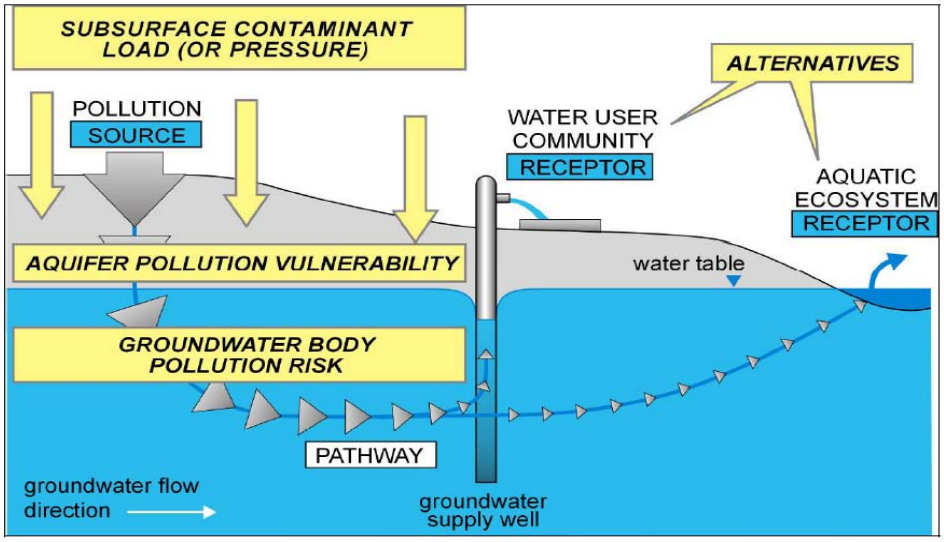


Figure 7.25. Basic elements of a groundwater pollution scenario

The Catalan Water Agency (ACA) databases provided groundwater information for fourteen cations and anions (Al, K, Fe, Mn, Si, Se, Ba, Cu, Pb, Zn, Mg, NO_3, NO_2, SO_4) as well as piezometrics data for 559 measurement stations distributed over the

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

complete geography as shown in Figure. 7.26. In this study, nitrate pollution is expressed as mg/l of NO_3. Data in 15% (99) of all available measurement stations (559) have missing values for nitrate concentrations. Table 7.16 lists the main statistical information for the measured pollutant concentrations data.

Table 7.16. Statistics for measured and reconstructed nitrate concentrations using different map sizes

	Measured as mg/l NO_3	**Recovered SOM 476 units**	**Recovered SOM 117 units**	**Recovered SOM 32 units**
mean	39.65	38.53	39.40	39.42
dev std.	36.68	34.52	34.19	33.99
median	27.90	27.77	33.32	32.53
max.	173.75	173.75	173.75	173.75
min.	0.3	0.3	0.3	0.3
SOM q_{error}	-	0.008	0.022	0.049
SOM t_{error}	-	0.079	0.075	0.075

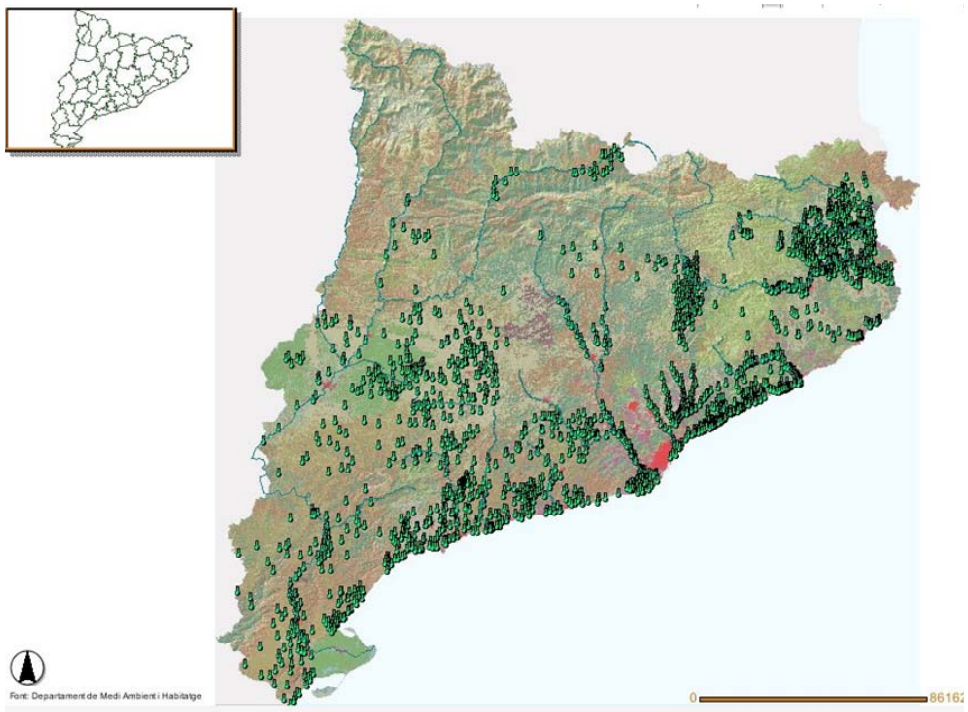


Figure 7.26. Location of groundwater measurement stations in Catalunya

The reconstruction of missing data has been performed using the SOM-based imputation method outlined in Chapter 3. The input data set consisted in 460 input points comprising the UTM_x , UTM_y and NO_3 concentration at each groundwater measurement station presented in Figure 7.26. Table 7.16 includes the main statistics corresponding to the reconstructed values using different map sizes. The best imputation approach is obtained with a SOM of 476 units. The relative errors in the estimation of the mean, standard deviation and median are 3%, 5.8%, and 0.4%, respectively. It should be noted that all three maps have similar topological error

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

(t_{error}) rates. However the lowest quantization error (q_{error}) is obtained for the largest map indicating a more coherent reconstruction of missing data.

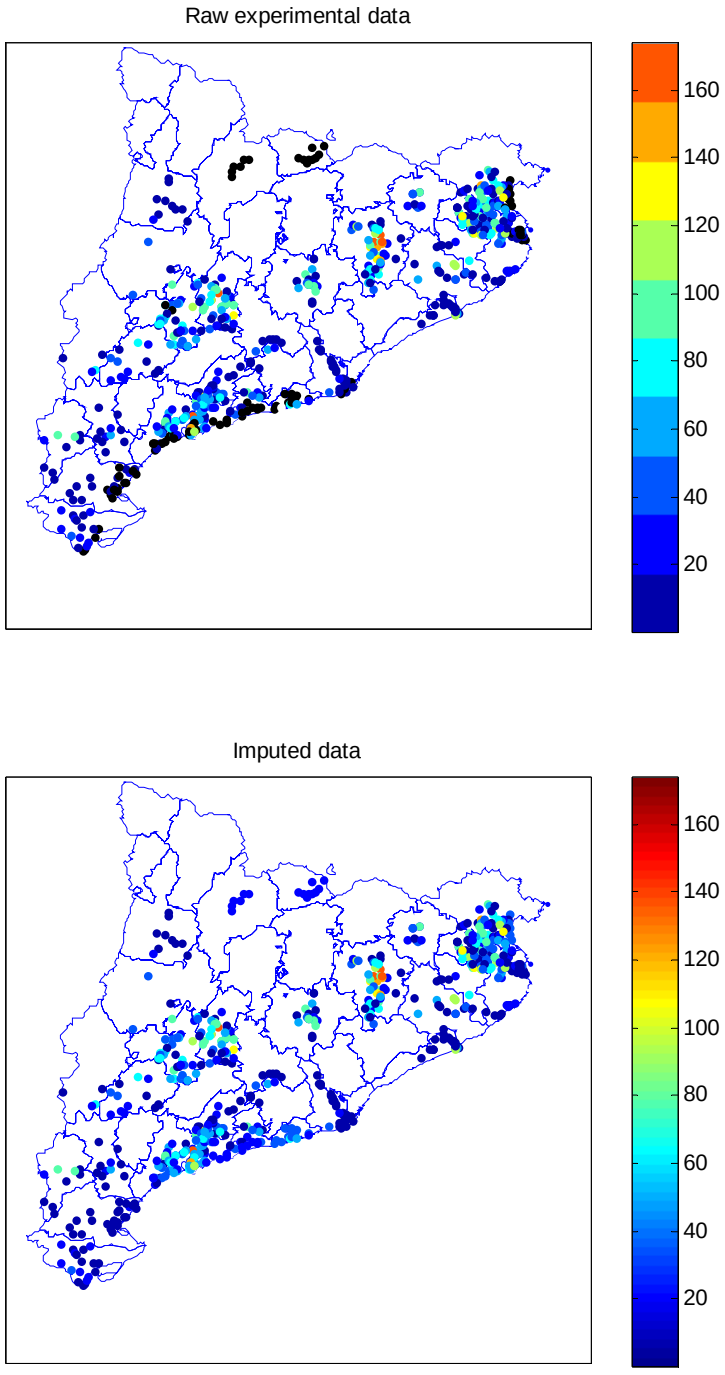


Figure 7.27. Nitrate groundwater pollution scenario. Data measurement stations with missing data are depicted in black in the top map. Reconstructed data after the imputation process are depicted in the bottom map

Figure 7.27 depicts groundwater quality data corresponding to the values of nitrate concentration (mg/l NO_3) in 559 measurement stations across Catalunya. Measurement stations with missing nitrate data are identified with black dots. After training a SOM with these data, missing values have been recovered by assigning the value of their corresponding SOM prototype. The SOM is trained using geo-referenced data, i.e., UTM coordinates, and as a result the SOM topology preservation properties also preserve the spatial contiguity of data. It can be observed that the recovered values are consistent with measures of neighbor locations.

The empirical cumulative density function has been computed for both the measured and the reconstructed data to assess the distribution of reconstructed data. It can be observed in Figure 7.28 that the imputation slightly underestimates NO_3 concentrations for values below 30 mg/l. In the range 50-100 mg/l the behavior is the inverse with the imputation model slightly overestimating values. It can be also perceived that histograms for both distribution models are quite similar. Thus, the SOM-based imputation method provides reliable estimations for missing data in environmental monitoring scenarios.

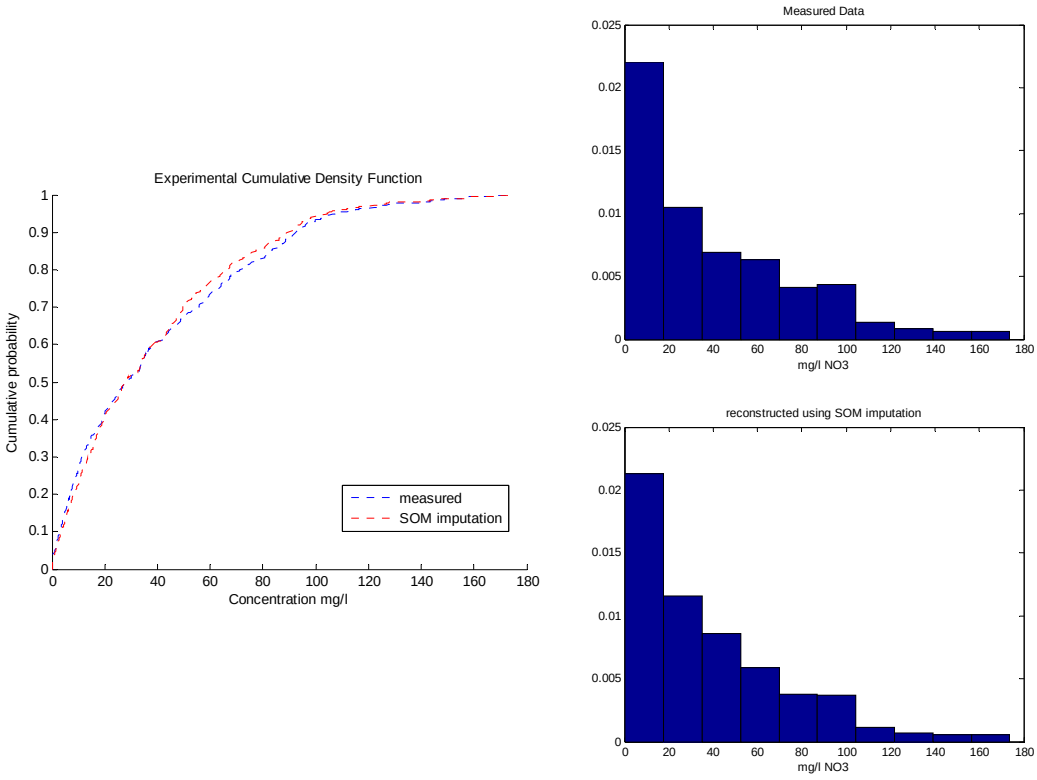


Figure 7.28. Comparison of cumulative distribution function (left) and histograms for measured (upper) and reconstructed (lower) nitrate concentration using SOM based imputation

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

7.3.2.3 TIER 3: ECD Modeling through SOM-based interpolation

Once the measurement stations field is completely filled with data, exposure concentration distribution (ECD) models can be developed taking into account the spatial continuity of data. Geostatistic methods provide tools to deal with these discrete measures over continuous geographic regions. These methods are based on the assumption that measurements lying closer together tend to be more alike than those farther apart. The exact nature of this pattern varies from data set to data set because each one has its own unique function of variability and distances between data points. This fundamental geographic principle is called spatial autocorrelation and can be examined by means of variogram analysis (Wackernagel, 2003).

Three functions can be used in geostatistics to describe the spatial correlation of observations: the correlogram, the covariance, and the semivariogram. The last is usually simply called variogram. The variogram is the key function in geostatistics since it is used to fit a model of the spatial correlation of the observed phenomenon. Semivariance is a measure of the degree of spatial dependence between samples. A smaller distance yields a smaller semivariance and a larger distance results in a larger semivariance. The plot of the semivariances as a function of distance from a point is referred to as a semivariogram and can be calculated as,

$$\gamma(h) = \frac{1}{2N_p(h)} \sum_{i=1}^{N_p(h)} (Z(xi) - Z(x_i + h))^2 \qquad (7.3)$$

where $N_p(h)$ is total number of pairs at distance h and $Z(x_i)$ is the experimental value at each location x_i. When a variogram is used to describe the correlation of different variables it is called *cross-variogram*. If the variogram is constructed from binary variables or class labels it is then called *indicator variogram*.

Kriging (Williams, 1998) is the estimation procedure used in geostatistics to determine unknown values by using known ones and a semivariogram. The procedures involved in kriging incorporate measures of error and uncertainty when determining estimations. Based on the semivariogram used, optimal weights are assigned to known values to calculate unknown ones. Since the variogram changes with distance, the weights depend on the known sample distribution.

The SOM algorithm is able to build a model of the data distribution during the learning phase and, thus, it can be used as an interpolation algorithm (Göppert and Rosenstiel, 1997; Sarzeaud and Stephan, 2000). Methods to approach geospatial interpolation using the SOM have also been pointed out by Correia (2005). The use of SOM as a spatial interpolation tool requires the inclusion of georeferenced data into the data set and the modification of the SOM learning rule to maintain the geospatial topology of data.

To use the SOM as a spatial interpolation method requires a trade-off between the topology preservation of the input space and the spatial continuity of data. This can be introduced in the form of a set of weights affecting distance calculations. The proposed modification to perform geographic interpolation is to give higher weights to spatial coordinates in the calculation of the Euclidean distance to find the best

ISBN:978-84-691-1008-9/DL: T.2230-2007

matching unit (BMU). Once the BMU is found the adaptation is performed in the usual way, affecting all components of the input data vector.

The proposed procedure to perform the SOM-based spatial interpolation for environmental data can be stated as follows:

1) Train a SOM with available data without missing data;

2) recover the missing value and complete all field data by using the SOM-based imputation;

3) train a new SOM using the reconstructed field of measurements;

4) generate a geo-referenced grid at the desired resolution;

5) For all points in the grid

- compute the BMU using the spatial weighted Euclidean distance

- Assign the prototype value of the BMU to the interpolated value of the target variable

A spatial interpolation was performed using the nitrate data set reconstructed by SOM imputation to illustrate the procedure described above. A regular grid of 8010 nodes spanned in the geographical domain was used to obtain interpolated nitrate concentrations at each grid point.

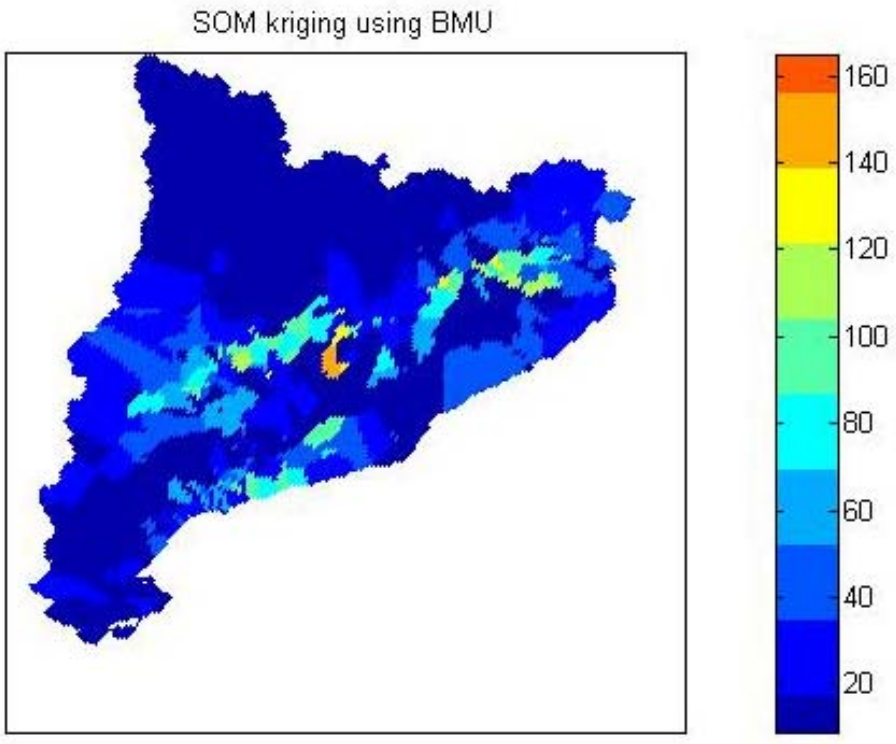


Figure 7.29. SOM based kriging for nitrate concentrations using non-weighted distances and BMUs

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Figure 7.29 shows the resulting nitrate distribution corresponding to the SOM interpolation using the Euclidean distance and assigning the same weight to all input variables (UTM_x, UTM_y and $[NO_3]$). Interpolation was performed in a regular grid with a resolution of 8010 nodes. The comparison of nitrate values in Figure 7.29 with the reconstructed nitrate measurements in Figure 7.27 indicates that the SOM interpolation is able to capture the spatial distribution of data. However, the structure of the interpolated areas appears tessellated as the result of the SOM topology preservation process, and the resulting map overestimates nitrate concentrations in many areas. To obtain smoother maps, the average of the 4 BMUs is used as the interpolated value for nitrate concentration. Figure 7.30 depicts the resulting map. It can be observed that lower estimates for nitrate concentration are obtained as a result of the averaging process. Although the concentration distributions over the map are smoother still remains the discontinuous effect caused by the influence of the value of nitrates in distance calculations.

Figure 7.31 depicts the interpolated concentration map which results after considering only UTM coordinates in distance calculations. The resulting nitrate concentrations are lower than those shown in Figure 7.29 since in that case the effects of nitrate values did not contribute to topology preservation

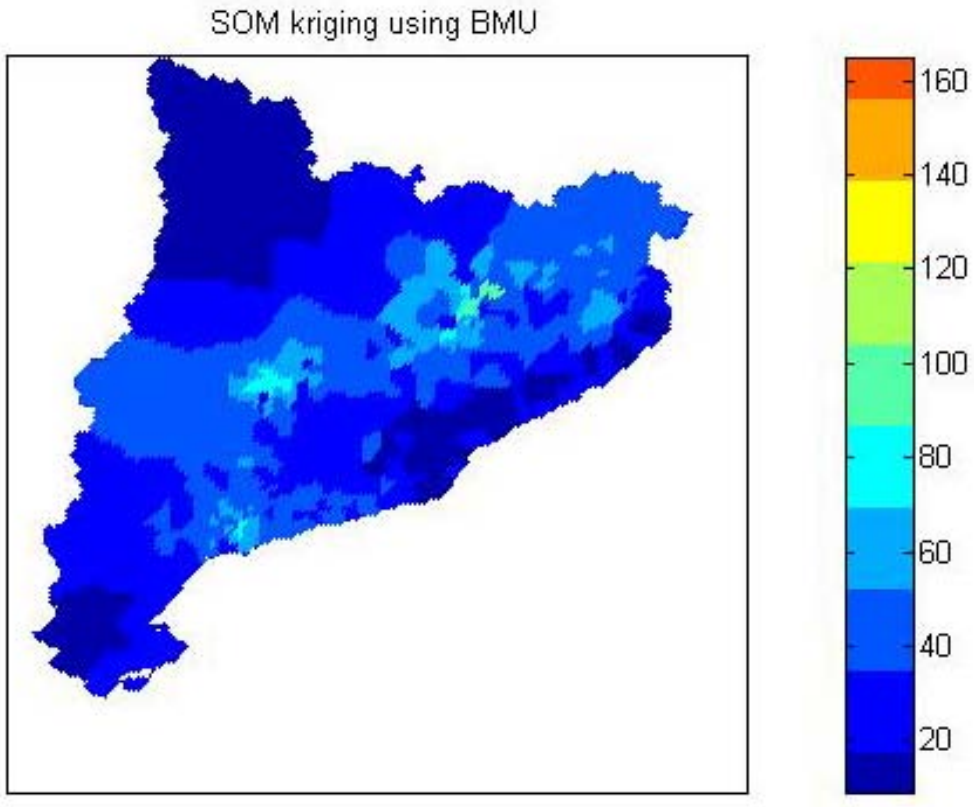


Figure 7.30. SOM based kriging for nitrate concentrations using non-weighted distances and the average of the four BMUs

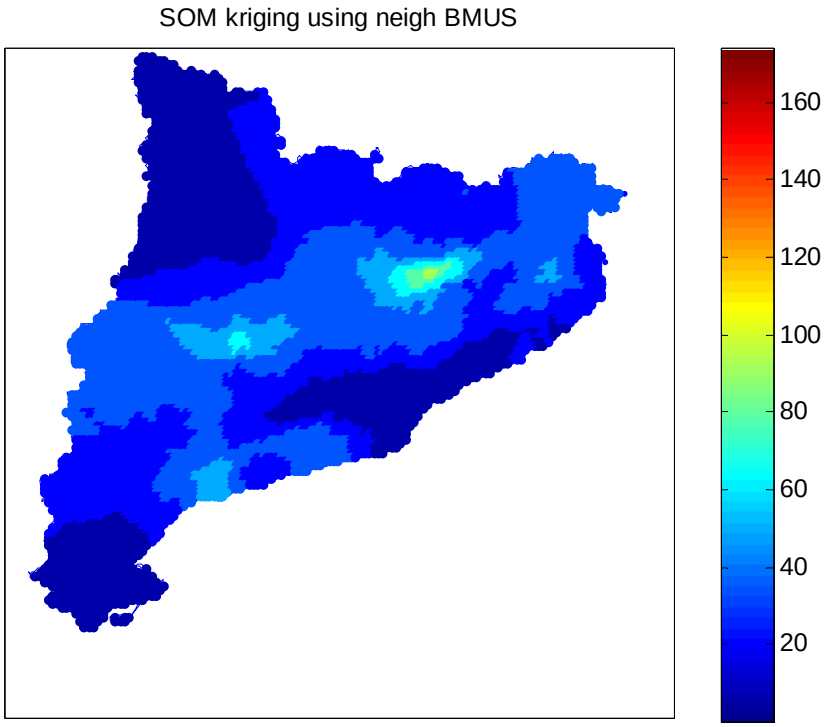


Figure 7.31. SOM based kriging for nitrate concentrations using weighted distances and the BMU

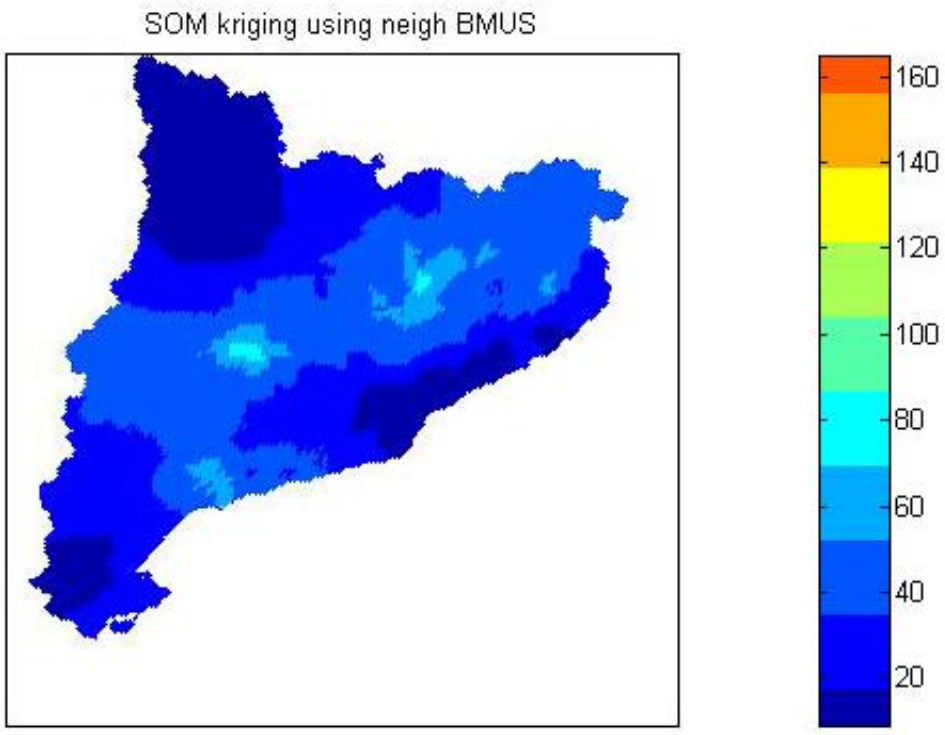


Figure 7.32. SOM based kriging for nitrate concentrations using weighted distances and the average of the four best matching units

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

The same smoothing procedure using the four BMUs applied in Figure 7.30 is applied in Figure 7.32 to the map but by using weighted distances. The resulting map is smoother than previous ones.

To assess the SOM-based spatial interpolation procedure, the above maps can be compared with the nitrate vulnerability map presented in Figure 7.33. This vulnerability map was obtained by Catalan governmental authorities from both, natural and anthropogenic factors influencing the occurrence of high nitrate concentrations in groundwater, such as population density, nitrogen fertilizer loading, groundwater recharge, soil protective capacity, vadose zone hydraulic conductivity, groundwater depth, and saturated zone hydraulic conductivity.

The comparison of figures 7.32 and 7.33 indicate that the three nitrate hotspots observed in Figure 7.32 match with the high vulnerability areas identified in Figure 7.33. The high vulnerability zone located in the coastal area nearby Barcelona is mainly due to the effect of high population density. This hotspot is not reproduced in the interpolated map since there are few groundwater measurement stations in this area and its nitrate concentration values are very low. The vulnerability map presented in Figure 7.33 should be considered only as an indication of areas susceptible of nitrate pollution and not as actual pollution spots.

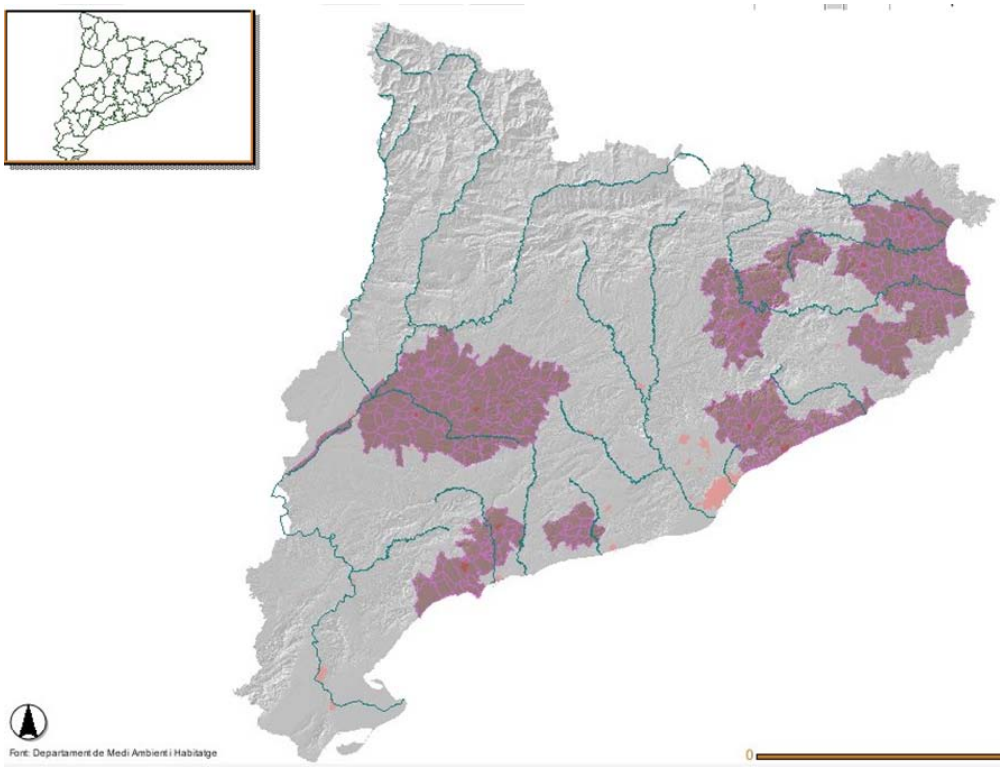


Figure 7.33. Nitrate vulnerability areas provided by Catalan government models

The same approach could be used to obtain an interpolation method similar to co-kriging in which additional variables are included in the interpolation process as

ISBN:978-84-691-1008-9/DL: T.2230-2007

constraints. The hydrogeological units presented in Figure 7.34 can be used to illustrate the inclusion of additional variables in the SOM spatial interpolation approach.

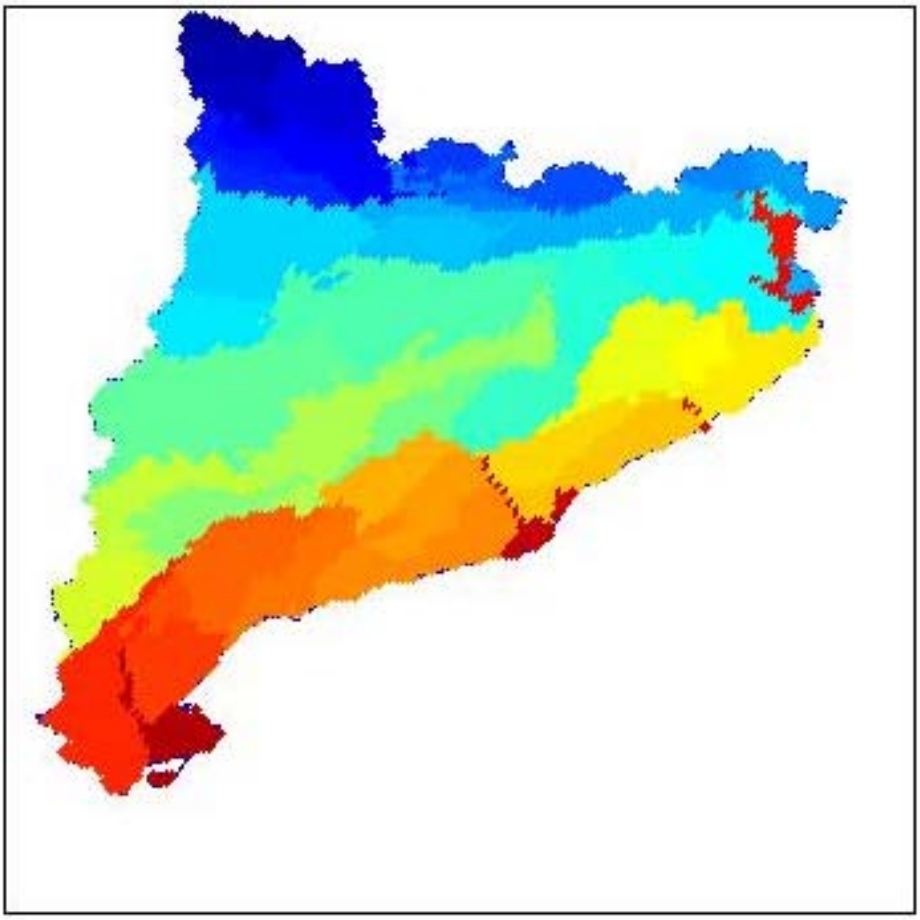


Figure 7.34. Hydrogeological areas in Catalunya

The effects of the inclusion of constraints in the SOM interpolation model is controled by the weighting assigned for distance calculations.The tuning of these weights produces more realistic maps. In the current study the tuning process is performed using a manual search. The development of optimal weighting strategies would lead to more accurate spatial interpolation schemes.

In groundwater pollution modeling an important issue to be considered is not only the spatial continuity but also the independence of aquifers (not mixing). Figure 7.34 depicts the borders of the main aquifers in the hydrogeological map of Catalunya. Nitrate values cannot be interpolated only in terms of spatial distances, i.e., independently of the span of aquifers.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

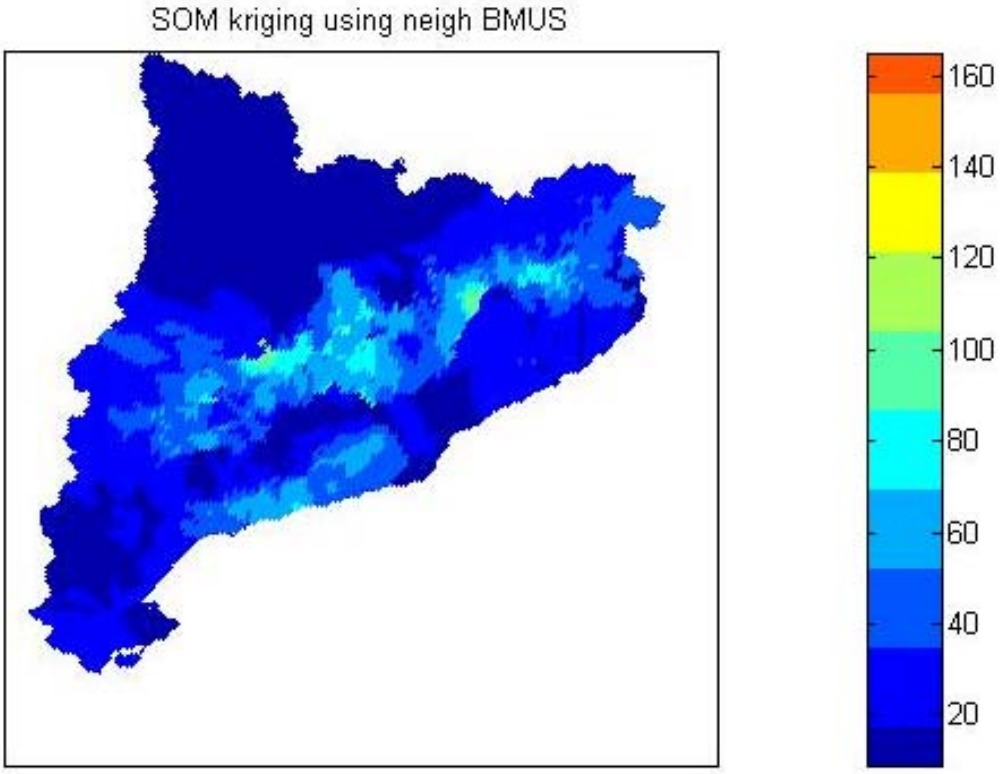


Figure 7.35. SOM based co-kriging of nitrate concentrations using hydrogeological units, weighted distances, and the average of the four BMUs

It can be observed in Figure 7.35 that the inclusion of aquifer boundaries yields nitrate distributions which are compatible with the distribution of their corresponding hydrogeological areas. Thus, it is possible to derive reliable models for the exposure concentration distributions (ECD) of a pollutant from SOM interpolated values.

The SOM-based spatial interpolation scheme yields consistent results when applied to different exposure scenarios, even in cases where data are scarce. Figure 7.36 shows the results obtained using the SOM-based method to estimate the distribution of suspended particles in air. In this case the scarcity of experimental data available makes the estimation less reliable in some areas, as indicated in Figure 7.37. The average quantization error is used as an indicator of data quality. Points located in the north of Catalunya (the Pyrenees), i.e., located far away from any measurement station, exhibit higher estimation errors. As a consequence, the exposure cumulative function obtained is skewed to low particle concentration values due to the effect of points located far away where concentration estimates tend to be close to zero.

ISBN:978-84-691-1008-9/DL: T.2230-2007

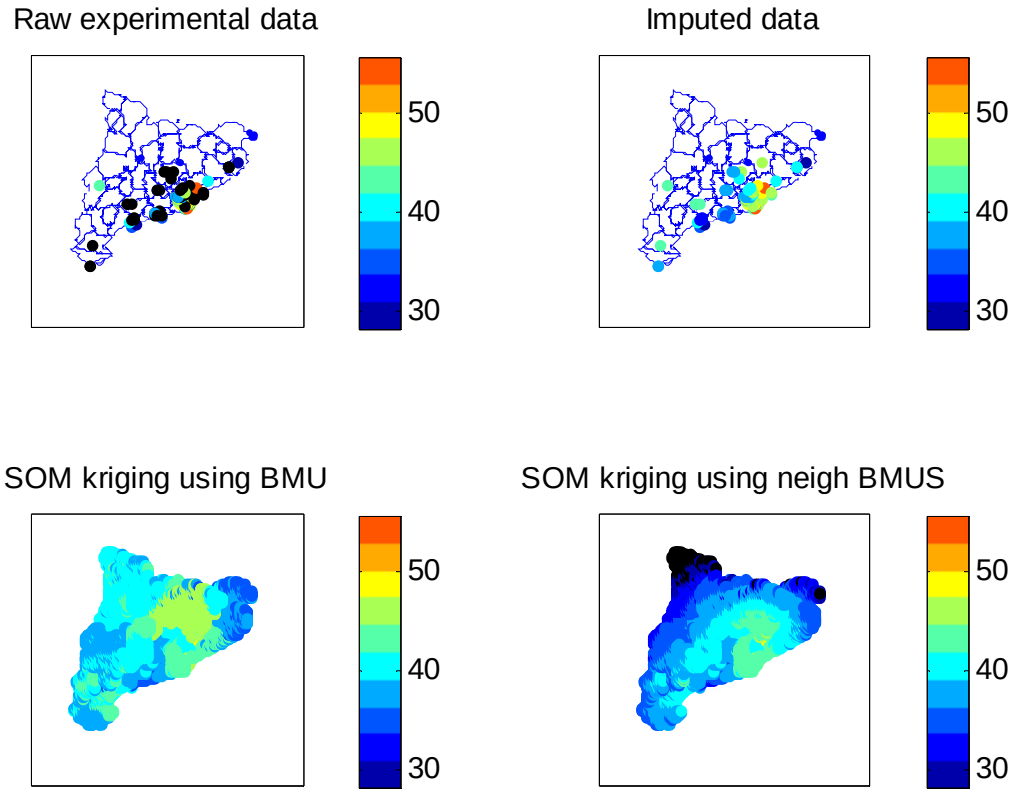


Figure 7.36. SOM-based data imputation process and geographic interpolation for suspended particles (PM) including average wind direction and speed as constraints in the spatial interpolation

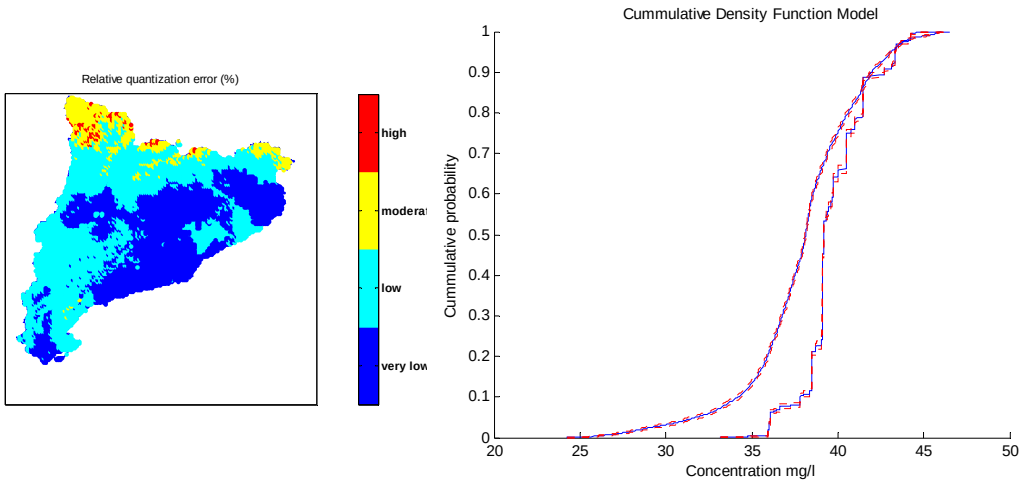


Figure 7.37. Errors for the distribution of suspended particles (PM) in air. Estimation of the imputation error using the average quantization error (left); estimation of the Cumulative Exposure distribution function for measured and interpolated PM values (right)

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

7.3.2.4 TIER 4: Probabilistic Risk Analysis

Risk characterization is the final phase in ecological risk assessment. Risk characterization:

- integrates analyses from the exposure and ecological effects characterization;
- describes uncertainties, assumptions, and strengths and limitations of analyses; and
- summarizes the overall risk information used by regulators to make risk management decisions.

Risk characterization has two major components: risk estimation and risk description. Risk estimation compares exposure and effects data, considers integrated exposure and effects data in the context of Levels of Concern (LOCs), and states the potential for risk. The risk description interprets risks based on the assessment of endpoints. The lines of evidence supporting or refuting risk estimates are evaluated to interpret the risk in terms of the following factors:

- Adequacy and quality of data;
- degree and type of uncertainty;
- relationship between evidence and risk assessment questions.

Risk characterization must be transparent, clear, consistent, and reasonable to be useful for regulatory purposes. Once the risk characterization is finalized, it may be used as the basis for producing fact sheets, press releases, technical briefings, and other communication products.

Risk assessment can be performed using either deterministic or probabilistic approaches. In the deterministic approach, a risk quotient (RQ) is calculated by dividing a point estimate of exposure by a point estimate of effects. This ratio is a simple, screening-level estimate that identifies high- or low-risk situations. Chemical risk quotients can be calculated from ecological effects data, use data, fate and transport data, and estimates of exposure to the pollutant. In the deterministic method, the estimated environmental concentration (EEC) is compared to an effect level indicator, such as LC50 (the concentration of a pollutant where 50% of the exposed organisms die.) or some regulatory threshold.

In the probabilistic approach, risk assessment incorporates probabilistic tools and methods to predict the magnitude of the expected impact of a contaminant as well as the uncertainty and variability involved in these estimates. A probabilistic risk assessment yields a distribution or range of values instead of a fixed one. Since the results of the refined risk assessment indicate the range of possible environmental impacts and which ones are most likely to occur, they provide a better basis for decision-making.

The interpolation capabilities of the SOM are used in risk assessment to generate risk maps using the RQ approach. Figure 7.38 depicts the RQ of exceeding the Spanish regulatory threshold for nitrates in groundwater (50 mg/l). EEC values have been obtained from a SOM-based interpolation using hydrogeological constraints in a regular grid of 8010 points for Catalunya. The two main hot spots detected in this

ISBN:978-84-691-1008-9/DL: T.2230-2007

Figure 7.38 correspond to areas with high livestock farming activities, which are responsible for nitrate groundwater pollution. Again, it can be noted that the "high risk" spots are consistent with the nitrate vulnerability map previously presented in Figure 7.33.

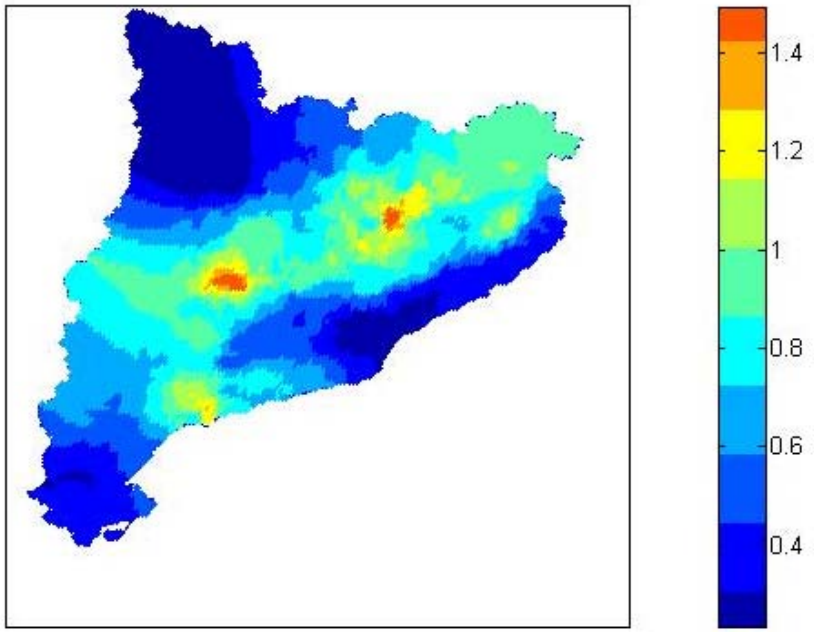


Figure 7.38. EEC and regulatory threshold risk quotient (RQ) for nitrate concentration in a groundwater pollution scenario in Catalunya

SOM can also be regarded as a probability estimation method (Lampinen and Kostiainen, 2000). The main advantages resulting from the association of a generative model (probability density) to a mapping method such as the SOM are:

- The density model enables de computation of the likelihood of any data presented to the model;
- the density model facilitates the quantitative analysis of the clustering process, for instance, by computing conditional densities to test the visually found hypothesis;
- Bayesian inference can be used to quantify established relationships.

Following this approach, the deterministic risk determination provided by the RQ method is enhanced by integrating a probabilistic framework within the SOM. The basic idea is to quantify the probability of each SOM unit of being the BMU for an input sample of data. However, the direct use of the Euclidean distance to perform this estimation poses two main problems:

- New sample data could be far from the closest SOM prototype (novelty detection);
- there may exist several (almost) equidistant prototypes.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

To avoid the last problem, a function of the distance between the prototype vector and the measured data can be used as the response of the SOM unit. A suitable functional form for this response (Alhoniemi et al., 1999) is,

$$r_i = \frac{1}{1 + \|\mathbf{x} - \mathbf{m}_i\|^2}, i = 1, \dots, N \tag{7.4}$$

where $\boldsymbol{m_i}$ is the prototype vector of unit $\boldsymbol{i}$, $\boldsymbol{x}$ is the new data sample and $\boldsymbol{N}$ is the number of units in the SOM. The function (7.2) has some beneficial properties such as $r_i \in [0,1]$, and it can be considered as a fuzzy indicator of the data sample hit for the map unit. However, the probabilistic interpretation response of Eq. (7.2) is not clear.

Alternatively, the closeness of a new multidimensional data sample to each prototype vector $\boldsymbol{i}$ can be defined using Bayes' theorem as the conditional probability of unit $\boldsymbol{i}$ given the data sample $\boldsymbol{x}$ as,

$$r_i = P(i|\mathbf{x}) = \frac{p(\mathbf{x}|i)P(i)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|i)P(i)}{\sum_i p(\mathbf{x}|i)P(i)} \tag{7.5}$$

where $\boldsymbol{i}$ is the SOM node, $\boldsymbol{P(i)}$ is the prior probability of node, $\mathbf{p(x)}$ is the prior probability of the data sample, and $\boldsymbol{p(x/i)}$ and $\boldsymbol{P(i/x)}$ conditional probabilities. To estimate the posterior probability $\boldsymbol{P(i/x)}$, both $\boldsymbol{p(x)}$ and and $\boldsymbol{p(x/i)}$ are needed. A Gaussian mixture model (Titterington et al., 1985) can be used to estimate these probabilities.

The modeled data is assumed to be generated by a set of Gaussian distributions in Gaussian mixture models (GMM). The parameters of the Gaussians, i.e., centers and widths, are usually determined using Expectation-Maximization (EM) algorithms (Dempster et al., 1977). The initial values for these parameters are usually computed using the K-means algorithm. Following this approach a Gaussian kernel density function can be adjusted to each SOM unit (Alhoniemi et al., 1999). The SOM prototype vector is used as a center for the Gaussian kernel and the variance is estimated by multiplying the distances between data samples and SOM prototypes with a neighborhood kernel defined as the final neighborhood radius multiplied by the average distance among SOM prototypes. The prior probability for a map unit $\boldsymbol{i}$ is defined by the ratio,

$$P(i) = \frac{\#(x_n \in \mathbf{m}_i)}{\#x_n}, n = 1, \dots, M \tag{7.6}$$

where $\boldsymbol{m_i}$ is a SOM prototype vector and $\boldsymbol{M}$ is the number of data samples. The conditional probability $\boldsymbol{p(x/i)}$ is estimated from the Gaussian kernel density function constructed for each SOM unit using the prior probability and the covariance matrix of the kernel width. i.e., the variance of each SOM node. Once the prior and conditional probabilities have been estimated the posterior probability $\boldsymbol{P(i/x)}$ is computed by using Eq. (7.3) and a probability estimate for each data value being generated by each SOM unit is obtained.

As a result, probability risk estimates for a data sample can be computed as the sum of the individual probabilities corresponding to each SOM unit which exceed some reference threshold value,

$$\text{risk}(\epsilon) = \sum_{i} P(i|\mathbf{x}), \qquad \text{for all i with } m_i^{target} > \epsilon \tag{7.7}$$

where ϵ is the threshold value and m_i^{target} is the value of the target property in the prototype vector of unit i.

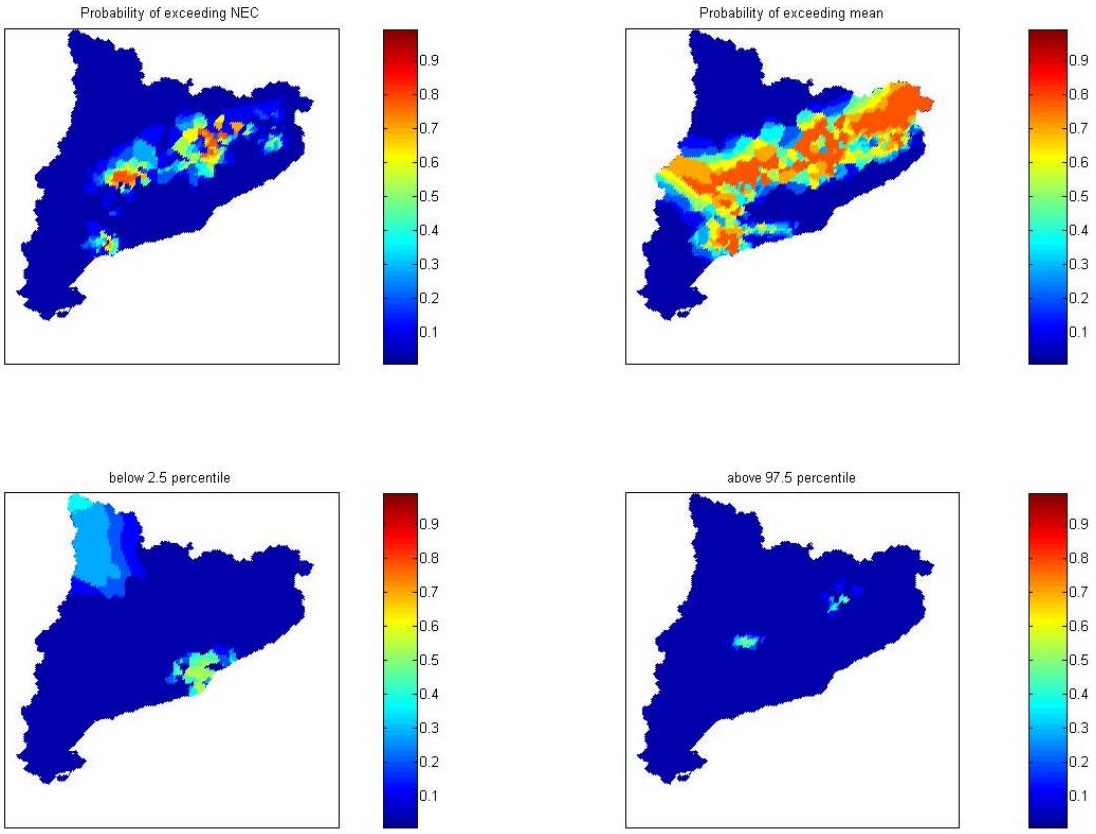


Figure 7.39. Comparison of Risk Maps for exceeding diverse nitrate concentration thresholds obtained with the integrated GMM-SOM technique

The application of the proposed procedure is illustrated by computing the risk estimates corresponding to the nitrate values obtained at each grid point. Different probabilistic risk maps for the nitrate pollution scenario are presented in Figure 7.39. Each data point is presented to the GMM model and the probability of exceeding the threshold control value is computed by addition of the posterior probabilities of corresponding to the sample point for units whose prototype vector has a value for the target value higher than the threshold value used (see Eq. 7.5). The probabilistic risk corresponding to nitrate concentrations higher than the regulatory No-effect Concentration (NEC) of 50 mg/l is coherent with the high vulnerability areas presented in Figure 7.33. Additional maps have been generated to represent the probability of a sample point of being above the measured average nitrate

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

concentration, indicating potential vulnerable areas. Maps for the probability of having low nitrate concentrations (below the 2.5 percentile) or high concentrations (above the 97.5 percentile) have also been computed indicating areas with low and high incidence of nitrate pollution.

The proposed integrated GMM-SOM method has been extended to deal with cumulative effects of different stressors. In this approach multiple stressors are introduced as regular variables during the generation of the SOM. The resulting maps will approximate the joint probability distributions of these stressors.

The inclusion of multiple effects is illustrated in a groundwater quality assessment scenario. The salinity of the aquifer is one of the most ecological adverse effects for groundwater in the Mediterranean coastal areas. The main stressors identified as responsible of the salinity effect are the concentrations of Cl, SO_4, Ca and Mg. The c-planes for these four concentrations together with the electric conductivities that used to train a SOM using the current GMM-SOM methodology are plotted in Figure 7.40. After adjusting a GMM to the resulting map, the risk map of being simultaneously above the average concentration for the four stressors considered was generated. Regions with high probability of exceeding the mean values of each indicator can be considered as more susceptible to salinity effects.

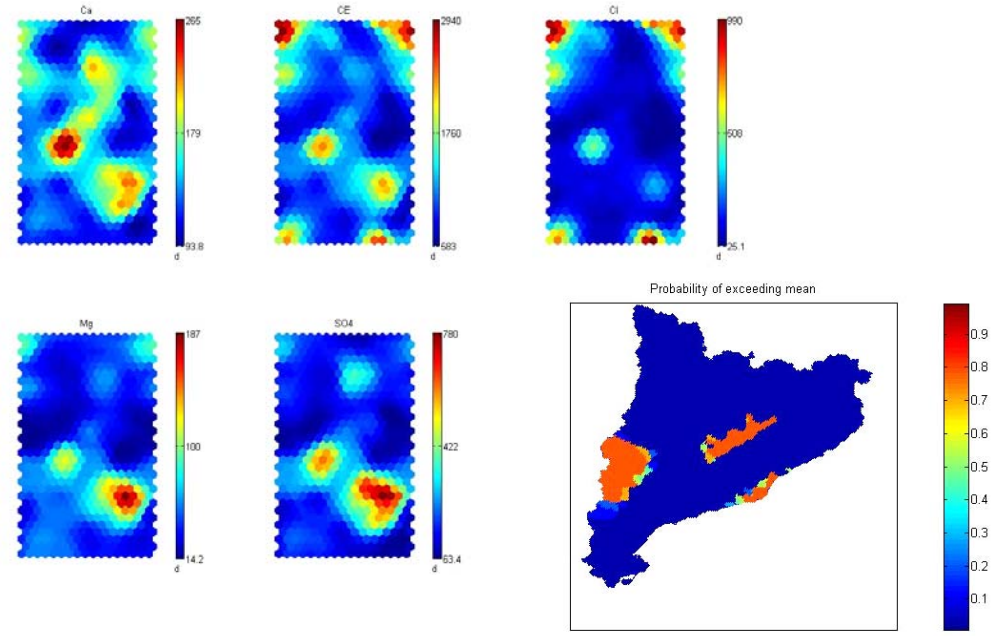


Figure 7.40. Joint effect of multiple stressors. Component planes for each input variable (stressor) and risk map corresponding to the salinity of Catalan aquifers

The application of the SOM in a groundwater pollution / quality scenario has been used to demonstrate the application of the different tiers of the proposed framework in an environmental modeling problem. The SOM provides a basic mechanism to perform imputation and spatial interpolation due to its intrinsic point density modeling properties. In addition the use of Gaussian Mixture models on top of SOM has been introduced as a probabilistic approach to risk assessment.

7.4 Conclusions

Chapter 7 has presented the results obtained when the components of the proposed framework are applied to different engineering domains. A neural network-based methodology to design and build virtual sensors to infer product quality from process variables has been developed and tested in section 7.1. Three neural systems, together with a linear model, have been used to build different virtual sensor models. A predictive fuzzy ARTMAP algorithm is one of the architectures considered. The other two architectures, identified as clustering average and DynaRBF, have been built by the combination of a dynamic unsupervised clustering layer with two different supervised mapping procedures to implement the desired outputs. This hybrid approach facilitates the use of more elaborate learning algorithms for the supervised layer without affecting the underlying infrastructure based on the dynamic unsupervised clustering.
As a proof-of-concept of the generic virtual sensor model, three types of neural models have been developed to infer the Melt Index (MI) of LDPE so that the accuracy of on-line correlation based techniques that are commonly used in industry is increased. Both single sensors for each LDPE grade and a composite model for all grades simultaneously have been implemented and testedto estimate MI. The three neural models for single grades, with all process variables measured at the beginning of the production cycle considered as input, predict MI values with relative mean errors and standard deviations of approximately 5% when appropriately trained with pre-classified patterns, compared with the average errors and standard deviations of approximately 10% and 15%, respectively, obtained with linear correlation models. A reduction in the number of variables up to 50% by dissimilarity measures of SOM decreases theses errors for single grade neural and linear models to approximately 4% and 5.5%, respectively, with comparable decreases in standard deviations. This reduction, which sets the accuracy standards of virtual sensors close to the ±2% experimental error for on-line MI measurements, could be explained in terms of noise reduction and elimination of conflicting input information with respect to the target MI.

DynaRBF yields slightly more accurate predictions of MI than fuzzy ARTMAP and clustering average for single grade sensors. It is well known that fuzzy ARTMAP needs extra amount of training information when the problem under study possess some underlying periodicity, as is the case in the current on-line time-variation of MI. The results obtained both for single grade and composite models indicate that all a neural implementation provide prediction reliability and accuracy. The proposed virtual sensors are capable of learning the relationships between process variables measured at the beginning of the production cycle and the quality of the final product. Their superior performance does not require any readjustment of parameters during production cycles.

The three neural sensors perform similarly for composite models. Sensors built with the reduced set of input process variables and trained by pre-classification yield the best predictions. The out-performance of neural sensors with respect to linear correlations is even more evident in this case of composite models since neural sensors are capable of quickly adapting to new operating conditions of the plant,

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

including grade transitions. Nevertheless, the effect of the reduction of input variables increases the standard deviation indicating that better training is needed.

A data imputation module has been integrated with the virtual sensor to provide input variable estimations in case of sensor failures. Single and multiple imputation approaches have been implemented and assessed. The main conclusions that can be drawn from these analysis are: (i) Multiple imputation methods are more stable and outperform those based on single imputation; (ii) the use of an ensemble of SOM maps of different sizes in multiple imputation systems yields better results that techniques based on bagging; (iii) the results obtained using only the components of the prototype of the *BMU* in SOM-based imputation are better than those obtained using the average with neighboring units. Finally, it has been demonstrated that the virtual sensor produces stable and reliable inferential measures of the target quality index once implemented in a real production process.

In section 7.2, a data-driven model to predict carcinogenicity of organic compounds from their molecular structure has been developed and tested. The performance of the fuzzy ARTMAP-based QSAR model for the carcinogenic potency (CP) of 104 aromatic compounds with nitrogen-contained substituent demonstrated the ability of the proposed integrated approach to accurately predict carcinogenicity as measured by CP. The inclusion of prototype vectors derived from the clusters in the SOM together with adequate molecular information, extracted from quantum-chemical and topological parameters, yields sufficient molecular information to characterize the carcinogenicity of this set of chemicals when they are not a single-chemical class compound as determined by fuzzy ART. The performance of the current methodology has shown to be superior when fuzzy ARTMAP is used since it captures the relevant non-linear relationships present in the data set and reduces the false negative/positive identification of carcinogens.

Finally, section 7.3 has presented an environmental modeling scenario in which some of the proposed framework components have been applied. The exploratory capabilities of tier 1 have been used to classify geographical areas according to their ecological and geophysical properties and to define useful scenarios for exposure and risk assessment. Data imputation in tier 2 has been used to mimic kriging methods to infer missing environmental data by including geophysical constraints in the estimates. The SOM training algorithm has been modified to take into account geographical continuity. Exposure modeling distributions for some groundwater pollutants have been generated using SOM interpolation. Finally, probabilistic risk analysis has been performed by adjusting a Gaussian Mixture on top of the SOM maps.

7.5 References

AMAT, L.; CARBÓ-DORCA, R. Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations. J. of Compt. Chem. 1997, 18, 2023-2039.

AMAT, L.; CARBÓ-DORCA, R.; PONEC, R. Simple linear QSAR models based on quantum similarity measures. J. Med. Chem. 1990, 42, 5169-5180.

AMAT, L.; ROBERT, D.; BESALÚ, E.; CARBÓ-DORCA, R. Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study". J. Chem. Inf. Compt. Sci. 1998, 38, 624-631.

ANZALI, S.; GASTEIGER, J.; HOLZGRABE, U.; POLANSKI, J.; SADOWSKI, J.; TECKENTRUP, A.; WAGENER, M. The Use of Self-Organizing Neural Networks in Drug Design. In 3D QSAR in Drug Design, 2, H. Kubinyi, G. Folkers, Y. C. Martin Ed., Kluwer/ESCOM, Dordrecht, NL. 1998, 273-299.

BAHLER, D.; STONE, B.; WELLINGTON, C.; BRISTOL, D.W. Symbolic, neural, and Bayesian machine learning models for predicting carcinogenocity of chemical compounds. J. Chem. Inf. Compt. Sci. 2000, 39, 906-914.

BURDEN, F.R.; FORD, M.G.; WHITLEY, D.C.; WINKLER, D.A. Use of Automatic Relevance Determination in QSAR Studies using Bayesian Neural Networks. J. Chem. Inf. Comput. Sci. 2000, 40, 1423-1430.

CARBÓ, R.; ARNAU, J.; LEYDA, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. Int J. Quantum Chem. 1980, 17, 1185-1191.

CARBÓ-DORCA, R.; BESALÚ, E. A general survey of molecular quantum similarity. J. of Mol. Struct. 1998, 451, 11-23.

CARPENTER, G.A.; GROSSBERG, S.; MARKUZON, N.; REYNOLDS, J.H.; ROSEN, D.B. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Networks. 1992, 3, 698 -713.

CARPENTER, G.A.; GROSSBERG, S.; REYNOLDS, J.H. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks, 1991, 4, 565-588.

CARPENTER, G.A.; GROSSBERG, S.; ROSEN, D.B. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks. 1991, 4, 759-771.

CHAN, W.M., GLOOR, P.E., HAMIELEC, A.E. A kinetic model for Olefin polymerization in high-pressure autoclave reactors. *AIChE Journal*, **1**(39):111, 1993.

CORREIA, F. *Exploratory geospatial data analysis using self-organizing maps*, Ph.D. thesis, Universidade Nova de Lisboa, 2005.

DAVIES, D.L.; BOULDIN, D.W. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1 (2), 224-227.

DEMPSTER, A.P., LAIRD, N.M., RUBIN, D.B. Maximum Likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, B*, **39**(1):1-38, 1977.

DOUALI, L. VILLEMIN, D. AND CHERQAOUI, D. Neural Networks: Accurate Nonlinear QSAR Model for HEPT Derivatives. J. Chem. Inf. Comput. Sci. 2003, 43, 4, 1200-1207.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

ESPINOSA, G.; ARENAS, A.; GIRALT, F. Integrated SOM-fuzzy ARTMAP Neural System for the Evaluation of Toxicity. J. Chem. Inf. Compt. Sci. 2002, 42, 2, 343-359

ESPINOSA, G.; YAFFE, D.; ARENAS, A.; COHEN, Y.; GIRALT, F. A fuzzy ARTMAP based Quantitative Structure-Property Relationships (QSPRs) for predicting physical properties of organic compounds, Ind. Eng. Chem. Res. 2001, 40, 2757-2766.

ESPINOSA, G.; YAFFE, D.; COHEN, Y.; ARENAS, A.; GIRALT, F. Neural network based Quantitative Structural Property Relations (QSPRs) for predicting boiling points of aliphatic hydrocarbons. J. Chem. Inf. Comput. Sci. 2000, 40, 859-879.

GASTEIGER, J.; LI, X.; USCHOLD, A. The beauty of molecular surfaces as revealed by Self-organizing neural networks. J. Mol. Graphics. 1994, 12, 90-97.

GINI, G.; LORENZINI, M.; BENFENATI, E.; GRASSO, P.; BRUSCHI, M. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, based on Molecular Descriptors Using Artificial Neural Networks. J. Chem. Inf. Compt. Sci. 1999, 39, 1076-1080.

GINI, G.; LORENZINI, M.; VITTORE, A.; BENFENATI, E.; GRASSO, P. Some Results for the Prediction of Carcinogenicity Using Hybrid Systems. In Predictive Toxicology of Chemicals: experiences and Impact of AI Tools; AAAI 1999 Spring Symposium Series; Gini, G.C., Katritzky, A.R., Eds.; AAAI Press: Menlo Park, CA, 1999, 74-77.

GOLD, L.S.; SAWYER, C.B.; MAGAW, R.; BACKMAN, G.M.; DE VECIANA, M.; LEVINSON, R.; HOOPER, N.K.; HAVENDER, W.R.; BERNSTEIN, L.; PETO, R.; PIKE, M.C.; AMES, B.N. A Carcinogenic Potency Database of the standardized results of animal bioassays. Environmental Health Perspectives. 58: 9-319 (1984).

GÖPPERT, J., ROSENSTIEL, W. The continuous interpolating Self-organizing Map. *Neural Processing Letters*, **5**:185-192, 1997.

GRAUEL, A.; LUDWIG, L.A.; RENNERS, I.; BERK, F. Computational Intelligence and Predictive Toxicology. Proc. AAAI Spring Symposium on Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools, Technical Report SS-99-01, 116-118, AAAI Press: Menlo Park (USA), 1999.

HALL, M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Proc. 17th International Conf. on Machine Learning., 2000, 359-366. Morgan Kaufmann, San Francisco, CA.

HAWKINS, D.M. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 2004, 44, 1-12.

IZRAILEV, S.; AGRAFIOTIS, D.A. Novel Method for Building Regression Tree Models for QSAR based on Artificial Ant Colony Systems. J. Chem. Inf. Comput. Sci. 2001, 41, 176-180.

KASKI, S.; LAGUS, K. Comparing Self-organizing maps. In Proc. of ICANN'96. 1996, 809-816.

KIER, L.; HALL, L. Molecular Connectivity in Chemistry and Drug Research. Academic Press: New York, 1976.

KOHONEN, T. The Self-Organizing Map. Proc. IEEE. 1990, 78, 1443-1464.

KONONENKO, I. Estimating Attributes: Analysis and Extensions of Relief. European Conference on Machine Learning, 1994, 171-182.

LINES, B., HARTLEN, D. Polyethylene reactor modeling and control design. *Hydrocarbon Processing*, 119, 1993.

LUDWIG, L.A.; GRAUEL, A.; RENNERS, I. Quantitative Structure-Activity Relationships and Computational Intelligence. Proc. European Symposium on Intelligent Techniques, ESIT '99 (on CD-ROM), ELITE Foundation, Aachen, 1999.

MAZZATORTA, P.; BENFENATI, E.; NEAGU, C. D.; GINI, G. Tuning Neural and Fuzzy-Neural Networks for Toxicity Modeling. J. Chem. Inf. Comput. Sci. 2003, 43, 513-518

MAZZATORTA, P.; VRAČRO, M.; JEZIERSKA, A.; BENFENATI, E. Modeling Toxicity by Using Supervised Kohonen Neural Networks. J. Chem. Inf. Comput. Sci. 2003, 43, 485-492.

RALLO, R.; ESPINOSA, G.; GIRALT, F. Using an ensemble of neural based QSARs for the prediction of toxicological properties of chemical contaminants. Trans. IChemE, Part B. Process Safety and Environmental Protection. 2005, 83, B4, 387-392.

RALLO, R.; FERRE-GINÉ, J.; ARENAS, A.; GIRALT, F. Forecasting product quality in industrial processes with virtual sensors. Sensor Technology, Proceedings of the AIChE Annual Meeting, Indianapolis, IN. 2002, 127-136.

RANDIC, M. On the characterization of molecular branching. J. of Am. Chem. Soc. 1975, 97, 6609-6615.

SARZEAUD, O. , STEPHAN, Y. Data interpolation using kohonen networks. In *Proceedings of the International Joint Conference on Neural Networks*, **6**:197-202, 2000.

SENESE, C.L.; HOPFINGER, A.J. Receptor-independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease. J. Chem. Inf. Comput. Sci. 2003, 43, 4, 1297-1307.

TITTERINGTON, D., SMITH, A., MAKOV, U. *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons, 1985.

VENDRAME, R.; BRAGA, R.S.; TAKAHATA, Y.; GALVAO, D.S. Structure-activity relationship studies of carcinogenic activity of polycyclic aromatic hydrocarbons using calculated molecular descriptors with principal component analysis and neural network methods. J. Chem. Inf. Comput. Sci. 1999, 39, 1094-1104.

VESANTO, J. SOM-Based Data Visualization Methods. Intelligent Data Analysis. 1999, 6, 11

VESANTO, J.; ALHONIEMI, E. Clustering of the Self Organizing Map. IEEE Transactions on Neural Networks, 2000, 11 (3), 586-600.

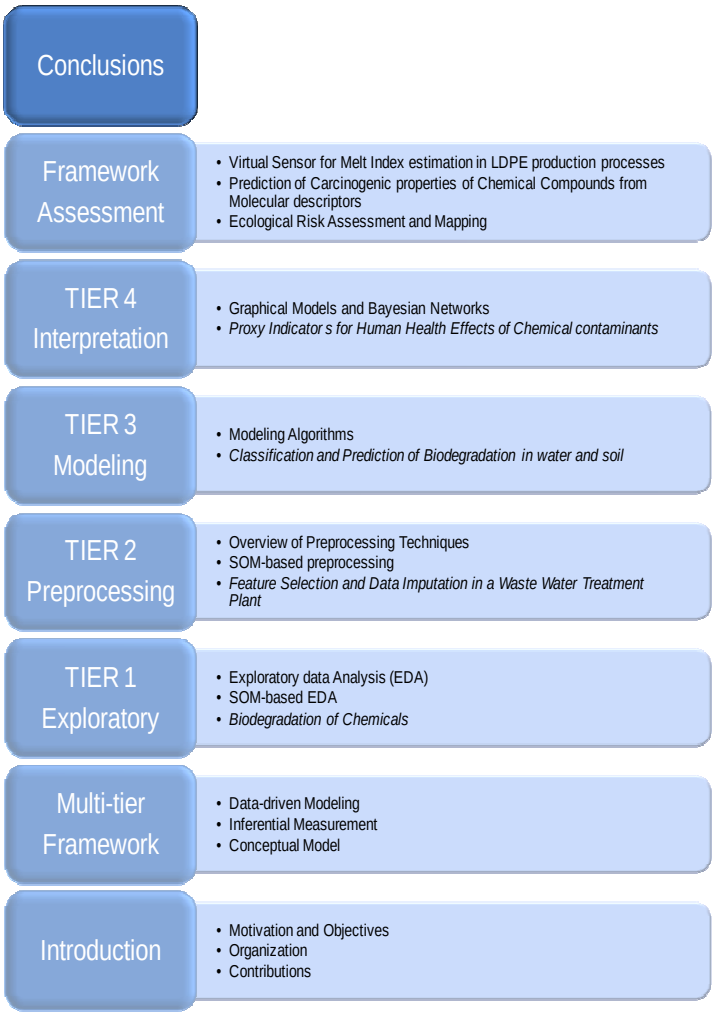
WACKERNAGEL, H. *Multivariate Geostatistics*, Springer-Verlag, 2003.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

WILLIAMS, C. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond, in M. I. Jordan (Ed.) *Learning in graphical models*. MIT Press, 599-612, 1998.

YASRI, A.; HARTSOUGH, D. Toward an Optimal Procedure for Variable Selection and QSAR Model Building. J. Chem. Inf. Comput. Sci. 2001, 41, 1218-1227.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007



UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Chapter 8

Conclusions, Perspectives and Future Work

Modeling natural systems has always been a challenging task. To develop accurate models of physical systems it is necessary to have a good understanding of the laws that drive their inner dynamics. However, most natural systems are complex and it is difficult or even impossible to find a rigorous mathematical formulation describing its behavior. Data-driven modeling constitutes a suitable approach to obtain reliable models in situations where information about the system can be only be achieved by observation.

The goal of this thesis has been to study the development of inferential measurement systems and data-driven models from both chemical engineering and computer science perspectives. The architecture of a multi-tier framework has been proposed to integrate all the exploratory, preprocessing, modeling and interpretation components involved in the development of data-driven models. Each individual tier of the model as well as the complete framework have been assessed in several real data modeling scenarios.

The Self-Organizing Map (SOM) has been introduced, at the exploratory level, as an alternative to the classical statistical exploratory data analysis (EDA) techniques. It has been demonstrated that the clustering capabilities of SOM are able to find and extract complex relationships from data. The exploration of the chemical space corresponding to biodegradation in water, which is complex to model, has been successfully carried out with the SOM-based EDA. Two data sets containing physicochemical properties and molecular descriptors have been used to assess the EDA capabilities of the SOM. It has been shown that SOM component planes provide precise indications about the relationships between variables and permit the detection of redundancy. The internal organization of the data space has been analyzed using distance matrices such as the U-matrix. It has been shown that the clustering SOM reference vectors coarsens the U-matrix classes and permits the extraction of similar chemicals in terms of the variables describing their chemical

space. The effects of data transformations (normalization and scaling) as well as different parameters concerning both the SOM algorithm and the input data have also been analyzed.

It has also been demonstrated that the SOM constitutes a valid alternative to perform most of the tasks needed at the preprocessing level. A new method for feature selection using dissimilarity measures has been developed and assessed. The SOM has been used to detect redundant information by developing a redundancy index which takes into account the correlation between variables and their representation over the map (c-planes and U-matrix). In addition, the SOM clustering capabilities have been applied to optimally select the training and test examples. This methodology ensures a proper data distribution within the model application domain. The use of SOM as both a single and a multiple data imputation system has also been studied.

The proposed preprocessing methodology has been applied to the development of virtual sensors to infer quality indicators for the effluent waters in a Waste Water Treatment Plant. Data reconstructed using single imputation techniques have been evaluated by developing an inferential model for the biological oxygen demand (BOD) of effluent water. The model has been developed using Radial basis Functions (RBF). Multiple imputations have also been assessed by developing SOM ensembles in which diversity has been introduced by topology changes and bagging. Results obtained in the two cases indicate that the SOM constitutes a valid approach to deal with data sets containing an important amount of missing information. This approach solves an important issue in inferential and data-driven modeling since a single sensor failure makes the model useless. The modeling process has been studied in terms of different algorithmic approaches: machine learning, neural networks, and statistical learning theories. A new model to construct radial basis functions coupled with a winner-takes-all approach to determine centers and widths of the Gaussian activation function has been introduced and evaluated. It has been demonstrated that virtual sensors using the proposed method outperform other neural systems.

The development of a data-driven model for a complex and biologically dependent property such as the biodegradation rate of chemicals in different media has been used to assess the modeling part of the proposed framework. In this example, the exploratory and preprocessing tiers of the framework have also been successfully used to characterize the chemical space of biodegradation rates, to select the best set of features, and to generate train and test sets adapted to the application domain of the studied data sets. Several feature selection methods such as CFS, ReliefF and ANNIGMA have been applied and compared with the proposed SOM-dissimilarity method. The current SOM-dissimilarity approach yields feature subsets with higher number of variables, leading to models with enhanced predictive power.

Quantitative and Qualitative Structure-Biodegradation Relationship (QSBR and SBR) models have been developed and assessed using the current approach. The effects of the skewed data distributions detected in previous EDA have been addressed in the model development tier by using different partition ranges of biodegradation rates. QSBR models have been developed for each range. However

ISBN:978-84-691-1008-9/DL: T.2230-2007

prediction errors remained still high for persistent chemicals. The approach used to solve this issue was to integrate qualitative SBR models to distinguish between biodegradable and non-biodegradable chemicals. The inclusion of these classifiers within the modeling scheme improved the estimation of biodegradation in the low range of biodegradation rates. In addition, models for biodegradation in soil have also been developed. Feature selection techniques have been applied to determine the influence of the soil matrix in these models. Results obtained are in agreement with those determined experimentally or previously published in the literature.

The framework also provided tools to facilitate the interpretation of models at the highest level of abstraction. The use of probabilistic algorithms to complement the deterministic approach provided by neural models has been addressed and assessed. Graphical dependency models and simple Bayesian networks have been introduced to provide a probabilistic point of view of relationships between model variables. These methodologies have been screened in the development of models to characterize the Modes of Toxic Action (MOA) of different chemicals. The extraction of descriptive rules to relate these MOAs with the molecular structure of chemicals has been performed. The MOAs related to aquatic toxicity of chemicals have been used as a proxy to extrapolate their effects in human health using a set of descriptive sentences known as Risk Phrases (RP). Preliminary results obtained using this approach indicate that the integration of this heterogeneous information would provide adequate tools to assess the toxicity of chemicals through the establishment of *a cross-spices* relationships. However, research still in progress is needed to confirm the validity and applicability of these models.

The components of the complete framework have been assessed in three different modeling scenarios. First, a virtual sensor for the on-line estimation of quality indicators in a chemical process plant has been designed and assessed in the context of inferential measurement. A neural network-based methodology to design and build virtual sensors to infer product quality from process variables has been developed and tested. Three neural systems, together with a linear model, have been used to build different virtual sensor models. The neural systems used included a predictive fuzzy ARTMAP algorithm and the combination of the proposed dynamic unsupervised clustering layer with two different supervised mapping procedures to implement the desired outputs. This hybrid approach allowed the adoption of more elaborate learning algorithms for the supervised layer without affecting the underlying infrastructure based on the dynamic unsupervised clustering.

As a proof-of-concept of the generic virtual sensor model, three types of neural models have been developed to infer the Melt Index (MI) of Low Density Polyethylene (LDPE) so that the accuracy of on-line correlation based techniques that are commonly used in industry is increased. Both single sensors for each LDPE grade and a composite model for all grades simultaneously have been implemented and tested. The three neural models for single grades, with all process variables measured at the beginning of the production cycle considered as input, predict MI with acceptable accuracy when appropriately trained with pre-classified patterns. The application of the SOM-dissimilarity feature selection method lead to smaller data sets which further decrease the prediction error rates for single grade neural and linear models. This reduction, which sets the accuracy standards of virtual sensors

close to the experimental error for on-line MI measurements, could be explained in terms of noise reduction and elimination of conflicting input information with respect to the target variable MI. It has been demonstrated that the proposed DynaRBF approach yields slightly more accurate predictions of MI than fuzzy ARTMAP and clustering average for single grade sensors. It is well known that fuzzy ARTMAP needs extra amount of training information when the problem under study possess some underlying periodicity, as is the case of the current on-line time-variation of MI. The results obtained both for single grade and composite models indicate that all neural implementations provide prediction reliability and accuracy. The proposed virtual sensors have been capable of learning the relationships between process variables measured at the beginning of the production cycle and the quality of the final product. Their superior performance does not require any readjustment of parameters during production cycles.

Composite models including all LDPE grades have been developed to assess the suitability of implementing a single model for all grades. The three neural sensors performed similarly for composite models. Sensors built with the reduced set of input process variables and trained by pre-classification yielded the best predictions. The out-performance of neural sensors with respect to linear correlations is even more evident in this case of composite models since neural sensors are capable of quickly adapting to new operating conditions of the plant, including grade transitions. Nevertheless, the effect of the reduction of input variables increased the standard deviation indicating that better training is needed.

A data imputation module has been integrated with the virtual sensor to provide input variable estimations in case of sensor failures. Single and multiple imputation approaches have been implemented and assessed. The main conclusions that can be drawn from this analysis are: (i) Multiple imputation methods are more stable and outperform those based on single imputation; (ii) the use of an ensemble of SOM maps of different sizes in multiple imputation systems yields better results that techniques based on bagging; (iii) the results obtained using only the components of the prototype of the best matching unit in SOM-based imputation are better than those obtained using the average with neighboring units. Finally, it has been demonstrated that the virtual sensor produces stable and reliable inferential measures of the target quality index once implemented in a real production process.

Second, in the context of data-driven modeling, the proposed framework has been used to predict carcinogenicity of organic compounds from its molecular structure. The performance of the fuzzy ARTMAP-based QSAR model for carcinogenic potency (CP) of 104 aromatic compounds with nitrogen-containing substituent has demonstrated the ability of the proposed integrated approach to accurately predict carcinogenicity as measured by CP. The inclusion of prototype vectors derived from the clusters in the SOM together with the selection of adequate molecular information, extracted from quantum-chemical and topological parameters, has yielded sufficient molecular information to characterize the carcinogenicity of this small set of data. The performance of the current methodology has shown to be superior when a classifier such as fuzzy ARTMAP is used since it captures the relevant non-linear relationships present in the data set and reduces the false negative/positive identification of carcinogens.

ISBN:978-84-691-1008-9/DL: T.2230-2007

Third, an environmental modeling scenario related to groundwater pollution has been selected to apply and assess some framework components. The exploratory capabilities of tier 1 have been used to classify geographical areas according to its ecological and geophysical properties to define useful scenarios for exposure and risk assessment. A new approach based in data imputation methods has been used to mimic the behavior of kriging methods to infer missing environmental data by including geophysical constraints in the estimates. The proposed SOM-based kriging allowed the easy inclusion of constraints in the imputation model resulting in more realistic estimates. Also it has been demonstrated that the inherent point density modeling produced by the SOM algorithm generates estimates without any prior assumption about the shape and properties of the underlying data distribution. A quantization error measure has been used as an indicator of the uncertainty of estimates provided by SOM. The SOM training algorithm has been adapted to take into account geographical continuity by using different weights for each input vector component in distance calculations. Exposure concentration distributions (ECD) for some groundwater pollutants, such as nitrates, have been generated using SOM interpolation. These inferred ECD using the interpolated data are in close agreement with those obtained from experimental data. Finally, a new probabilistic risk analysis methodology based on SOM has been proposed and assessed in the nitrate groundwater pollution scenario. Gaussian Mixture models adjusted in top of SOM reference vectors have been used to provide probability estimates of exceeding certain regulatory threshold for nitrate pollution.

The complete assessment of the framework proposed in this thesis shows that data-driven modeling, when combined properly with all the necessary components to explore and preprocess the data, provides a convenient and accurate methodology to develop experimental models for real-world applications. However, the interpretability of the models is still the main drawback of this approach. This work has only outlined the use of machine learning models such as rule generation algorithms and Bayesian dependency models, to explain relationships between variables. Further research is needed to enhance the proposed framework at its highest abstraction level.

UNIVERSITAT ROVIRA I VIRIGILI
MULTI-TIER FRAMEWORK FOR THE INFERENTIAL MEASUREMENT AND DATA-DRIVEN MODELING
Robert Rallo Moya
ISBN:978-84-691-1008-9/DL: T.2230-2007

Universitat Rovira I Virgili, Tarragona.