UNIVERSITAT DE BARCELONA

FACULTAT DE MEDICINA – DEPARTAMENT DE SALUT PÚBLICA

TESI DOCTORAL

APORTACIONS ALS MÈTODES ESTADÍSTICS PER A MODELAR DADES AGREGADES AMB CORRELACIÓ ESPACIAL

Rosa Mari Abellana Sangrà

ÍNDEX

AGRAÏMENTS

Aquesta aventura va començar un cop em vaig llicenciar l'any 1997, aleshores vaig tenir l'oportunitat de treballar a la Unitat de Bioestadística del departament de Salut Pública com a becària i més tard com a ajudant. Des d'aquell moment va anar creixent el meu interès per realitzar el doctorat, fins que l'any 1999 els de la Unitat d'Assessorament Suport i Prevenció de l'Hospital Clínic de Barcelona, van embarcar-se conjuntament amb el Dr. Ascaso en crear una plataforma GIS d'una zona de Manhiça i en analitzar la distribució geogràfica de la malària en aquesta àrea. En aquells moments van comptar amb mi, i els hi agreixo la confiança dipositada, perquè no m'imaginava que aquest camp desconegut acabaria essent la base de la meva tesi.

Al llarg dels cursos de doctorat i la realització d'aquesta tesis, el Josep Lluís sempre m'ha donat suport moral, ajuda, i idees per continuar treballant i per poder crear aquesta tesi. Vull agrair-li molt especialment tot el que ha fet per mi, doncs sempre ha tingut temps quan ell ja tenia feina suficient en la seva tesi i projectes.

També un agraïment al Lluís perquè quan ha estat necessari, davant del seu ordinador però d'esquenes al personal, m'ha donat el seu suport, així com els seus comentaris i correccions. De veritat gràcies per tot.

Pel seus comentaris i recomanacions vull donar les gràcies al Dr. MacNab del *Centre for Health Evaluation Research, Bristish Columbia for Children's and Women's Health* de Canadà, a la Dra. Ugarte del departament d'Estadística de la Universitat Pública de Navarra, i al Dr. Ferrándiz del departament d'Estadística i Investigació Operativa de la Universitat de València.

A tots aquells del departament de Salut Pública i del departament d'Estadística de la Universitat de Barcelona que d'alguna manera han col·laborat en la realització d'aquesta tesi.

A la Geòrgia i al Jaume, companys del departament, pel seu suport moral.

Al servei de Llengua Catalana de la Universitat de Barcelona perquè tot i ser catalana de naixement i haver adquirit el nivell C, ells m'han corregit les faltes ortogràfiques i expressions errònies que se m'han anat colant al llarg del document.

Al Dr. Tresseres i a la Conxa Castells del departament de Sanitat i Seguretat Social de la Generalitat de Catalunya, per haver-me facilitat l'accés a les dades de diabetis de tipus I a Catalunya.

Vull agrair als meus pares, a les meves germanes i al meu cunyat el suport que m'han aportat en tota la meva tesi. Al Xavi i al petit Manel, que han aguantat els meus nervis i han estat al meu costat en tot moment.

Als meus amics d'Agramunt i Reus, que gràcies a la seva insistència m'han fet sortir de festa, i han aconseguit donar-me forces per continuar treballant.

INTRODUCCIÓ

1. Introducció

El conjunt de procediments per analitzar dades que són recollides tenint en compte la seva localització geogràfica és coneix com *estadística espacial*, la qual té com a objectiu modelar les dades incorporant els possibles patrons geogràfics. Dins d'aquest marc, les dades es poden classificar segons com es defineix la localització (georeferència) associada a cada unitat d'observació (Cressie, 1993), i són classificades en dades geoestadístiques, patrons puntuals i dades reticulades.

Les dades geoestadístiques es caracteritzen perquè cada localització geogràfica es realitza mitjançant un punt i aquest té associat mesures d'una o més variables d'interès. Aquestes dades provenen d'un procés espacial continu, és a dir, la variable d'interès que es mesura varia contínuament a través de la regió d'estudi i per tant hi haurà un nombre infinit de localitzacions. L'objectiu de treballar amb aquest tipus de dades és modelar i predir la distribució espacial de la variable d'estudi al llarg de tota la zona d'interès a partir de les localitzacions mostrejades. Alguns exemples de dades geoestadístiques són concentracions d'ozó, temperatura o concentracions de minerals del subsòl mesurades en uns punts geogràfics.

En els patrons puntuals també es treballa amb localitzacions puntuals que tenen associades mesures, però es distingeixen de les dades geoestadístiques en el fet que aquestes localitzacions són també una variable aleatòria i aleshores tant les localitzacions com la variable d'interès tenen propietats estocàstiques. Es poden dividir en *spatial point process* i *marked point process*. Els *spatial point process* es caracteritzen pel fet que la variable que es mesura és una variable degenerada, és a dir, ens indica tan sols la presència d'un esdeveniment, com, per exemple, les localitzacions dels arbres en un bosc o les localitzacions de casos d'una malaltia. En canvi, en un *marked point process* la variable aleatòria és no degenerada, com, per exemple, el diàmetre dels arbres localitzats en un bosc.

Per finalitzar, les dades reticulades són observacions associades a les regions en què es divideix una àrea en concret. Aquestes poden ser regulars com els píxels d'una imatge o irregulars com els municipis d'una àrea determinada. Així doncs, són dades reticulades el nombre de malalts en cada província d'un país o el nombre d'accidents per cada segment d'una carretera.

En ciències de la salut es treballa amb dades reticulades quan es pretén estudiar el risc de patir o morir d'una determinada malaltia en una àrea. L'objectiu d'aquests estudis és analitzar la variabilitat geogràfica de les taxes d'una malaltia. Les dades que cal analitzar consisteixen en agregacions de la variable dicotòmica presència o absència d'un esdeveniment, en funció de les regions geogràfiques en què s'ha dividit l'àrea d'estudi. La definició d'aquestes regions sovint és de caràcter administratiu o per motius censals. Aleshores, en aquests estudis, la variable d'interès, Y, serà un recompte el qual habitualment s'assumeix que es distribueix segons una Binomial o una Poisson. A més a més, essent la mesura de risc més utilitzada la taxa d'incidència acumulada o mortalitat, aquest recompte correspondrà al nombre de casos nous o al nombre de morts, referits aquests a una població de risc durant un període determinat. L'estimació del risc de patir o morir d'una determinada malaltia s'utilitza per obtenir mapes informatius de la distribució geogràfica del risc. Aquests mapes permeten representar la informació d'una manera visual i identificar patrons que podrien ser omesos mitjançant una

presentació en forma tabular. A més, aquesta representació de la malaltia mitjançant un mapa té importants funcions, com ara formular i validar hipòtesis sobre la malaltia, identificar possibles factors de riscos i detectar regions on la incidència o la mortalitat de la malaltia és alta o baixa.

Atès que l'objectiu és estudiar la variabilitat geogràfica de la taxa, és pertinent eliminar possibles diferències entre regions provocades per variables confusores. Un dels possibles mètodes per controlar per aquestes variables és l'estandardització i, en particular, l'estandardització indirecta. Normalment, s'utilitza l'estandardització indirecta perquè no es coneixen les taxes específiques de la variable per la qual es vol estandarditzar, o perquè existeixen regions on el nombre de persones de risc és insuficient per obtenir unes taxes específiques representatives. Per tant, per dur a terme aquesta estandardització es necessita tenir unes taxes específiques per a cada estrat de la variable per la qual es vol estandarditzar. Aquestes es poden obtenir internament utilitzant les taxes específiques pel total de la regió estudiada, o externament obtenint-les d'una població de referència.

S'ha de fer notar que quan es treballa amb dades estandarditzades estem assumint que el risc en funció dels possibles estrats de la variable per la qual estem estandarditzant és el mateix per a totes les regions, és a dir, que es considera que no hi ha interacció entre les àrees i aquesta variable.

Quan s'utilitza l'estandardització indirecta habitualment es fa servir la raó estandarditzada de mortalitat o morbilitat (SMR) com a mesura de risc. L'SMR s'obté del quocient del nombre de casos observats i el nombre de casos esperats obtinguts a partir de l'estandardització indirecta. De manera que una SMR superior a 1 indica que hi ha un risc superior en la regió d'estudi que en la població de referència, si l'SMR és inferior a 1, hi haurà menys risc, i si és igual a 1, el risc és el mateix.

El problema de treballar amb SMR és que les àrees amb poca població de risc, normalment rurals, tendiran a presentar SMR extremes perquè el nombre de casos esperats serà petit. D'altra banda, l'expressió de l'error estàndard és:

$$\text{e.s.}\left(\hat{\text{SMR}}\right) \approx \frac{\sqrt{\text{var}(\text{Y})}}{\text{E}},$$

i per tant és inversament proporcional al nombre de casos esperats (E). Aquest fet provoca que SMR extremes en zones poc poblades no siguin significativament diferents d'1 perquè les seves estimacions presentaran molta variabilitat. En conseqüència, es poden tenir dues SMR de la mateixa magnitud però que una sigui significativa i l'altra no per la diferència del nombre de casos esperats. Aquest problema també es pot observar quan es treballa amb les taxes brutes per la variabilitat de la població de risc de les àrees estudiades, és a dir, les zones rurals tendiran a presentar taxes extremes perquè la seva població de risc és petita. Per tant, l'increment de la taxa per la presència d'un nou cas és més elevat en les poblacions petites, mentre que en les poblacions grans la variació de la taxa serà mínima. Paral·lelament, les taxes també seran inestables atès que el seu error estàndard és inversament proporcional al nombre de persones de risc. Tot això afecta les estimacions brutes i condueix a una mesura errònia de l'heterogeneïtat real de les taxes i de l'SMR que repercuteix en la interpretació de la distribució geogràfica del risc al llarg de l'àrea.

Un mètode que permet recollir l'heterogeneïtat real del risc és el modelatge a partir de tècniques de regressió. Aquestes tècniques permeten crear models complexos que poden representar la realitat, les quals anirem desenvolupant al llarg d'aquesta introducció. A més a més, el modelatge ens permet fer una suavització dels valors extrems del risc. Les tècniques de regressió considerades són els models lineals generalitzats i els models lineals generalitzats mixtos.

2. Models lineals generalitzats

2.1. Introducció

Un primer model que es podria considerar per modelar els recomptes de cada àrea és el model de regressió lineal clàssic

$$Y = \alpha + X\beta + \varepsilon,$$

on Y és el vector dels recomptes corresponents a cada regió, α és la mitjana general, X és la matriu de disseny de les possibles variables explicatives que tant poden ser variables confusores per les quals es vol controlar les estimacions com factors de risc. β és el vector de paràmetres que relaciona el vector de variables explicatives amb les respostes, i finalment, ε és el vector dels errors aleatoris.

Atès que les dades són recomptes que es distribueixen sota una distribució Binomial o de Poisson, el model lineal clàssic no serà adequat perquè en primer lloc es violarà l'assumpció de normalitat dels errors, ε, i perquè habitualment la relació entre la variable resposta i les variables explicatives no serà lineal.

Nelder i Wedderburn (1972) van definir els models lineals generalitzats (GLM) de manera que permeten que:

1. la distribució de probabilitat de la variable resposta, i per tant dels errors aleatoris, sigui qualsevol de la família exponencial.
2. la funció que relaciona l'esperança de la variable resposta, $E(Y)$, i les covariables sigui una funció no lineal. Per exemple, la funció lògit, $\log\left(\frac{E(Y)}{1-E(Y)}\right)$, o la funció logaritme, $\log(E[Y])$.

Seguint la parametrització dels models lineals generalitzats definida per McCullagh i Nelder (1989) per a dades independents, la funció de distribució de la variable resposta Y es resumeix de la manera següent:

$$f_Y(Y;\theta,\phi) = \exp\left\{\frac{(Y\cdot\theta - b(\theta))}{a(\phi)} + c(Y,\phi)\right\},$$

on θ és conegut com el paràmetre canònic, i ϕ, com el paràmetre de dispersió. Quan ϕ és desconegut, es considera un paràmetre de soroll (*nuisance* parameter*)*. La funció $a(\phi)$ habitualment s'expressa com ϕ/ω, on ω són pesos coneguts que poden variar d'observació en observació. La funció $b(\cdot)$ és la *cumulant function*, i $c(\cdot)$ és una funció arbitrària que depèn de les dades, de ϕ i ω.

A partir d'aquesta parametrització, l'esperança i la variància de la variable *Y* es defineixen de la manera següent:

$$\mathrm{E}(\mathrm{Y}) = \frac{\partial \mathrm{b}(\theta)}{\partial \theta} \quad \text{i} \quad \mathrm{Var}(\mathrm{Y}) = \mathrm{a}(\phi) \cdot \frac{\partial^2 \mathrm{b}(\theta)}{\partial^2 \theta}.$$

La funció de log-versemblança es resumeix així:

$$\ln \mathrm{L}(\mathrm{Y};\theta) = \frac{\theta \cdot \mathrm{Y} - \mathrm{b}(\theta)}{\mathrm{a}(\phi)} + \mathrm{c}(\mathrm{Y}, \phi_0).$$

En la taula 1 es mostren les funcions anteriors en el cas que les respostes $y_1\ y_2 ... y_n$ segueixin una distribució binomial de paràmetres n_i i π_i, o una distribució de Poisson de paràmetre μ_i.

Taula 1
Components de la parametrització dels models lineals generalitzats per dades Binomials i de Poisson

Distribució	θ_i	$a(\phi)=\phi$	$b(\theta_i)$	$c(y;\phi)$
Binomial(n_i,π_i)	$\log\left(\frac{\pi_i}{1-\pi_i}\right)$	1	$n_i \cdot \log(1+e^{\theta_i})$	$\log\binom{n_i}{y_i}$
Poisson(μ_i)	$\log(\mu_i)$	1	e^{θ_i}	$-\log(y_i!)$

θ_i : paràmetre canònic; $a(\phi)=\phi$: paràmetre dispersió; $b(\theta_i)$: *cumulant function*; $c(y;\phi)$: una funció arbitrària.

Així la funció de probabilitat i la funció de log-versemblança associada a dades binomials és:

$$f_{Y_i}(y_i;\theta_i,\phi_0) = \exp\left\{\left(y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i)\right) + \log\binom{n_i}{y_i}\right\} = \binom{n_i}{y_i} \cdot \pi_i^{y_i} \cdot (1-\pi_i)^{n-y_i}$$

$$\ln \mathrm{L}(\mathrm{Y};\pi) = \sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i},$$

on $\pi=(\pi_1,\pi_2,....\pi_n)$.
I per al cas de la Poisson és:

$$f_{Y_i}(y_i;\theta,\phi) = \exp\left\{\frac{\left(y_i \cdot \log(\mu_i) - e^{\log(\mu_i)}\right)}{1} - \log(y_i!)\right\} = \frac{\exp(-\mu_i)\cdot \mu_i^{y_i}}{y_i!}$$

$$\ln \mathrm{L}(\mathrm{Y};\mu) = \sum_{i=1}^{n} \frac{\log(\mu_i)\cdot y_i - e^{\log(\mu_i)}}{1} = \sum_{i=1}^{n} y_i \cdot \log(\mu_i) - \mu_i - \log(y_i!),$$

on $\mu = (\mu_1,\ldots,\mu_n)$.

En resum, els models lineals generalitzats es defineixen mitjançant tres components:

1. La variable resposta *Y*, que es distribueix sota una distribució de la família exponencial.
2. Un vector de paràmetres desconeguts β de dimensió *q* i una matriu de disseny de les variables explicatives, *X*, de dimensió $n \times q$.

3. Una funció d'enllaç, monòtona i diferenciable que defineix la relació entre l'esperança de la variable resposta, $E(Y_i) = \mu_i$, i les variables explicatives. Aquesta funció s'expressa com:

$$g(\mu_i) = \eta_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_q x_{iq},$$

on $i = 1,....n$.

Matricialment, l'equació es defineix com:

$$g(\mu) = \eta = \beta \cdot X,$$

on $\mu = (\mu_1 \ldots \mu_n)^T$ i $\beta = (\beta_1 \ldots \beta_q)^T$. A més a més, quan la funció d'enllaç coincideix amb el paràmetre canònic, θ, aquesta funció s'anomena *funció d'enllaç canònica*. Per exemple, amb dades binomials de paràmetres n_i i π_i, la funció d'enllaç canònica, $g(\mu_i)$, és la funció lògit

$$\log\left(\mu_i \big/ n_i - \mu_i\right).$$

Per tant, com que l'esperança d'una variable binomial és

$$\mu_i = n_i \cdot \pi_i,$$

la funció d'enllaç se simplifica en

$$\log\left(\pi_i \big/ 1 - \pi_i\right).$$

En conseqüència:

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_{i2} + ... + \beta_q x_{iq})}{1 + \exp(\beta_1 + \beta_2 x_{i2} + ... + \beta_q x_{iq})}.$$

La regressió entre la variable resposta i variables explicatives utilitzant aquesta funció canònica es coneix com a *regressió logística*.

En l'altre cas, és a dir quan es realitza l'assumpció que la variable resposta es distribueix segons una $\text{Poisson}(\mu_i)$, la funció d'enllaç canònica és la funció logarítmica, $\log(\mu_i)$ i, per tant,

$$\mu_i = \exp(\beta_1 + \beta_2 x_{i2} + ... + \beta_q x_{iq}).$$

En aquesta situació, la regressió rep el nom de *regressió de Poisson*.

2.2. Estimació de màxima versemblança

L'estimació dels paràmetres de regressió $\beta_j, j = 1,\ldots q$, la podem obtenir mitjançant el mètode de màxima versemblança. A partir d'aquest mètode, l'estimació dels paràmetres de regressió s'aconsegueix igualant a 0 la derivada de la funció de log-versemblança respecte als paràmetres que s'han d'estimar:

$$U_j = \frac{\partial \ln L(\theta; Y)}{\partial \beta_j} = 0.$$

La funció de log-versemblança s'ha expressat en l'apartat anterior com:

$$\ln L(\theta, Y) = \sum_i^n \frac{\theta_i \cdot y_i - b(\theta_i)}{a(\phi)} + c(y, \phi) = \sum_{i=1}^n l_i.$$

Per tant, la derivada d'aquesta funció en relació amb els paràmetres la podem definir:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j}.$$

Aquesta derivada es pot solucionar mitjançant les relacions parcials que existeixen entre l_i i β_i, amb el qual:

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \cdot \frac{a(\phi)}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}.$$

Per tant, el sistema d'equacions per resoldre serà:

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)} \cdot x_{ij} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0.$$

En tractar-se d'un sistema d'equacions no lineals, caldrà utilitzar un algoritme d'optimització numèrica com l'algoritme de *Fisher-Scoring*. En la iteració *r* d'aquest procediment les estimacions dels paràmetres de regressió s'actualitzen mitjançant:

$$\beta^{(r)} = \beta^{(r-1)} + I^{-1}\left(\beta^{(r-1)}\right) \cdot U\left(\beta^{(r-1)}\right),$$

on $I\left(\beta^{(r-1)}\right)$ és la matriu d'informació de Fisher avaluada en $\beta^{(r-1)}$, la qual es calcula mitjançant:

$$I(\beta) = -E\left[\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} \right] = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu}{\partial \eta_i} \right)^2.$$

L'expressió matricial de $I(\beta)$ és $X^T W X$, on *W* ve definida per:

$$\text{diag}\left\{ \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right\}.$$

Si l'equació $\beta^{(r)} = \beta^{(r-1)} + I^{-1}\left(\beta^{(r-1)}\right) \cdot U\left(\beta^{(r-1)}\right)$ la multipliquem en ambdós costats per $I\left(\beta^{(r-1)}\right)$ s'obtindrà:

$$I\left(\beta^{(r-1)}\right) \cdot \beta^{(r)} = I\left(\beta^{(r-1)}\right) \cdot \beta^{(r-1)} + U\left(\beta^{(r-1)}\right).$$

Per tant, la part dreta d'aquesta equació serà:

$$X^T \cdot W\left(\beta^{(r-1)}\right) \cdot Z\left(\beta^{(r-1)}\right),$$

on $Z\left(\beta^{(r-1)}\right)$ és un vector de *n* components on cada element és igual a

$$z_i = x_i \beta^{(r-1)} + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i},$$

i z_i és coneguda com a pseudovariable.

Així l'equació iterativa que s'ha de resoldre és:

$$\left(X^T W\left(\beta^{(r-1)}\right) X\right) \cdot \beta^{(r)} = X^T W\left(\beta^{(r-1)}\right) Z\left(\beta^{(r-1)}\right)$$

$$\beta^{(r)} = \left(X^T W\left(\beta^{(r-1)}\right) X\right) \cdot X^T W\left(\beta^{(r-1)}\right) Z\left(\beta^{(r-1)}\right).$$

En la taula 2 s'especifiquen cada una d'aquestes expressions en funció de la distribució que s'assumeixi.

Taula 2

Estimació per màxima versemblança dels GLM en dades Binomials i de Poisson

Distribució	μ_i	$\mathrm{Var}(Y_i)$	w_i	z_i	$\frac{\partial \eta_i}{\partial \mu_i}$
Binomial	$n_i \cdot \pi_i$	$n_i\pi_i(1-\pi_i)$	$n_i\pi_i(1-\pi_i)$	$\sum_{k=1}^{q} x_{ik}\beta_k + \frac{(y_i - n_i\pi_i)}{n_i\pi_i(1-\pi_i)}$	$\frac{1}{n_i\pi_i(1-\pi_i)}$
Poisson	μ_i	μ_i	μ_i	$\sum_{k=1}^{q} x_{ik}\beta_k + \frac{(y_i - \mu_i)}{\mu_i}$	$\frac{1}{\mu_i}$

μ_i i Var(Y): esperança i variància de la variable *Y*; w_i: pesos: z_i: pseudovariable, i η_i : predictor lineal.

El procés d'iteracions esquemàticament es resumeix d'aquesta manera:

1. Donat un valor inicial per $\hat{\beta}^0$ es calcula *W* i *Z*. Normalment per començar el procés els valors inicials de $\hat{\beta}^0$ s'obtenen de considerar $\hat{\mu}_i = y_i$, és a dir, fixant la mitjana al valor observat.
2. Se substitueixen els valors anteriors en l'equació
$$\beta^{(r)} = \left(X^T W\left(\beta^{(r-1)}\right)X\right)\cdot X^T W\left(\beta^{(r-1)}\right)Z\left(\beta^{(r-1)}\right),$$
i s'obtenen uns nous valors de β amb els quals es torna a començar el procés fins que s'arribi a la convergència.

2.3. Problemes en l'anàlisi de les dades

En general, un dels problemes que sorgeixen en l'anàlisi de dades binomials i de poisson és que el paràmetre de dispersió, ϕ, és diferent de la unitat. Aquest fenomen és conegut com a *sobredispersió* quan $\phi > 1$ o *sotsdispersió* si $\phi < 1$.

Aquesta situació succeeix quan la variància empírica de les dades $\mathrm{Var}(Y_i)$ és diferent a la variància teòrica $\mathrm{Var}(\mu_i)$, és a dir, en el cas Binomial que $\mathrm{Var}(Y_i) \neq n_i\pi_i(1-\pi_i)$, i en el cas Poisson que $\mathrm{Var}(Y_i) \neq \mu_i$.

El problema principal tant de la sobredispersió com de la sotsdispersió és l'estimació incorrecta dels errors estàndard de les estimacions, de manera que aquests seran infraestimats en el cas de sobredispersió o sobreestimats en el cas de sotsdispersió.

La sobredispersió o la sotsdispersió en dades binomials i poisson es poden donar en les situacions següents:

1. Quan la mitjana de les dades varia segons una variable i aquesta no és inclosa en el model (Shoukri and Pause, 1998). Per exemple, imaginem que tenim *k clusters* de mida n_i, i = 1,...,k, i que el nombre de respostes positives a cada *cluster* Y_1, Y_2,....Y_k, segueixen una binomial de paràmetres n_i i p_i. Aleshores el nombre total de respostes sense tenir en compte les agrupacions serà $Y=Y_1+Y_2+....+Y_k$, i el total d'individus serà $m = \sum_{i=1}^{k} n_i$. L'esperança i la variabilitat de la probabilitat p_i es poden expressar com $E(Y \mid p_i) = m \cdot p_i$ i $\mathrm{Var}(Y \mid p_i) = m \cdot p_i (1 - p_i)$. Aleshores l'esperança i la variància del nombre total de respostes vénen donades per

$$E(Y) = E\left[E\left(Y \mid p_i\right)\right] = m \cdot E\left(p_i\right) = \pi,$$

i

$$\begin{aligned} Var(Y) &= E\left[Var\left(Y \mid p_i\right)\right] + Var\left[E\left(Y \mid p_i\right)\right] = m \cdot E\left[p_i\left(1-p_i\right)\right] + m^2 \cdot Var\left[p_i\right] = \\ &= m \cdot \pi \cdot \left(1-\pi\right) \cdot \left[1 + \left(m-1\right)\sigma^2\right], \end{aligned}$$

on σ^2 expressa la sobredispersió, atès que si $\sigma^2 = 0$ la variància teòrica i l'empírica coincidiran. Com que la variància només pot prendre valors positius, el fet de no incorporar variables rellevants en el model sempre provocarà una situació de sobredispersió.

2. Si s'incompleix l'assumpció d'independència de les dades (Shoukri i Pause, 1998). La presència de dades correlacionades apareix en estudis de *clusters* en els quals les dades d'un mateix *cluster* comparteixen característiques comunes que condueixen a l'existència de correlació entre si. Són exemples d'aquests estudis les mesures repetides que es poden obtenir d'un individu en diversos períodes de temps (estudis longitudinals) o agregacions de les dades en *clusters* com ara ciutats, escoles o estatus social.

   Per exemple, en el cas binomial si definim $Y = \sum_{j=1}^{n} Y_j$ on Y_j són variables binàries amb probabilitat π i $(1-\pi)$. Aleshores, $E(Y) = n\pi$ i $Var(Y) = n\pi(1-\pi)$. Si l'assumpció d'independència no és manté, és a dir, la correlació$(Y_l, Y_k)=\rho$ on $l \neq k$, la variància de les dades esdevé:

$$Var(Y) = Var\left(\sum_{j=1}^{n} Y_j\right) = \sum_{j=1}^{n} \pi \cdot \left(1-\pi\right) + \sum_{l \neq k}^{n} \rho \cdot \pi \cdot \left(1-\pi\right) = n \cdot \pi \cdot \left(1-\pi\right) \cdot \left[1 + \left(n-1\right) \cdot \rho\right]$$

   Si $\rho > 0$, la variància de Y és més gran que la que esperaríem si les respostes fossin independents, i provocaria sobredispersió. En canvi, si $\rho < 0$, la variància serà inferior a l'esperada, aleshores ens trobarem davant d'una situació amb sotsdispersió.

Un cas particular de la violació de la independència de les dades és l'existència de correlació espacial en estudis de la variació geogràfica del risc al llarg d'una àrea. La correlació espacial considera que el risc d'una regió pot estar influenciada pel risc de les regions veïnes i, en conseqüència, aquestes regions tendiran a presentar riscos similars. L'existència d'aquesta correlació en dades de malalties no infeccioses mesurades en regions d'una àrea geogràfica és produïda pel fet que àrees pròximes comparteixin factors de risc que no es troben en àrees més llunyanes. Normalment, aquests factors de risc s'associen a factors mediambientals, socials, culturals i fins i tot a característiques genètiques o a hàbits alimentaris que no estan mesurats. En conclusió, si el risc depèn d'aquestes covariables no mesurades es produirà la presència d'una correlació de les dades a escala regional. Un exemple d'aquest tipus de correlació és l'estudi del càncer de llavis a Escòcia realitzat per Breslow i Clayton (1993). En aquest tipus de càncer un dels principals factors de risc és l'exposició solar, factor que està associat a un component geogràfic.

Finalment, la presència de correlació espacial ens indica l'existència d'unes agrupacions geogràfiques més grans que la mida de les regions estudiades (Bernardinelli *et al.*, 1995*a*). En aquest cas la correlació serà positiva per definició i per tant no tenir-la en compte generarà sobredispersió. Per tant, en l'anàlisi de riscos amb correlació espacial

que es desenvoluparan en aquesta tesi només es considerarà la situació de sobredispersió.

Una de les possibles solucions per tenir en compte la sobredispersió de les dades és incorporar efectes aleatoris en el model. Aquest tipus de models són una extensió dels models lineals generalitzats i reben el nom de models lineals generalitzats mixtos (GLMM).

3. Model lineal generalitzat mixt

3.1. Definició

El model lineal generalitzat mixt (McCulloch i Searle, 2001) incorpora efectes aleatoris, els quals es poden entendre com a variables latents que intenten capturar els efectes de variables no mesurades o desconegudes, que provoquen que la variabilitat de les dades sigui més gran que la variabilitat que es recull sota un model lineal generalitzat. Per tant, ens seran útils per controlar la sobredispersió.

Definim el vector Y_i compost per les n_i observacions realitzades en el *cluster* i-*èsim*, i b_i és el vector d'efectes aleatoris de dimensió $p \leq q$, on q és el nombre de covariables. En el cas d'estudis geogràfics amb dades agregades per regions, en ser la unitat d'observació la mateixa regió, el vector Y_i conté una sola observació ($n_i = 1$); per tant, es pot simplificar la nomenclatura mitjançant el vector Y de dimensió igual al nombre de regions, N, i compost per l'observació de cada regió, y_i.

S'assumeix que els y_i condicionats als efectes aleatoris són independents amb $\mathrm{E}(y_i \mid b_i) = \mu_i^b$ i $\mathrm{Var}(y_i \mid b_i) = \mathrm{Var}(\mu_i^b)$ i que es distribueixen sota una distribució de la família exponencial. Aleshores el model expressat matricialment és:

$$g(\mu^b) = X\beta + Zb,$$

on $\mu^b = (\mu_1^b, \ldots, \mu_N^b)$ és el vector de mitjanes, X i Z són les matrius de disseny pels efectes fixos i aleatoris, β i b són els vectors dels coeficients de regressió dels efectes fixos i aleatoris respectivament, i $g(\cdot)$ és la funció d'enllaç canònica. Per acabar de definir el model, és necessari assumir una distribució pels efectes aleatoris que denotarem mitjançant $f_b(b \mid \Sigma(\varpi))$. Una de les distribucions més utilitzades perquè ens permet incorporar fàcilment estructures de correlació és la distribució Normal (Clayton i Kaldor, 1987). Sota aquesta assumpció els efectes aleatoris es distribuiran sota una distribució normal multivariant de mitjana 0 i matriu de variàncies i covariàncies $\Sigma(\varpi)$, on ω és el vector dels components de la variància. Aquests components de la variància són els que ens permeten recollir la sobredispersió observada en les dades.

Aleshores la funció de versemblança quedarà definida per l'expressió següent:

$$L(Y;\theta,\phi,\Sigma(\varpi)) = \int \prod_{i=1}^{N} f_Y(Y;\varpi,\phi)\cdot f(b,\Sigma(\theta))\,d(b) =$$

$$= \int \prod_{i=1}^{N} \exp\left\{\frac{(y_i\cdot\theta_i - b(\theta_i))}{a(\phi)} + c(y,\phi)\right\}\cdot f(b,\Sigma(\varpi))\,d(b) =$$

$$= \int \prod_{i=1}^{N} \exp\left\{\frac{(y_i\cdot\theta_i - b(\theta_i))}{a(\phi)} + c(y,\phi)\right\}\cdot \frac{1}{\sqrt{(2\pi)^N \Sigma(\varpi)}} \exp\left\{-\left(1/2\right)\cdot b\cdot\Sigma^{-}(\varpi)\cdot b^T\right\} db.$$

En el cas que els efectes aleatoris no siguin independents, la funció de versemblança no es pot avaluar d'una manera tancada. Per solucionar aquest problema, s'han proposat diferents solucions les quals s'explicaran i es desenvoluparan més endavant.

3.2. Parametrització de la matriu de variàncies i covariàncies dels efectes aleatoris en estudis geogràfics reticulats

La parametrització de la matriu de variàncies i covariàncies dels efectes aleatoris és un aspecte fonamental en l'ajustament de models en els estudis geogràfics reticulats, atès que segons com s'efectuï aquesta parametrització es controlaran diferents tipus de sobredispersió, concretament l'estructurada i la no estructurada. La sobredispersió estructurada és la provocada per la presència de correlació espacial entre regions. En canvi, la no estructurada ens indica que en una regió es produeix una de les situacions següents: la mitjana de les dades varia segons una variable que no està inclosa en el model i la no-independència dels individus que componen una regió.

Així, aquesta matriu es determina en funció de quina sobredispersió es vol controlar de manera que s'especifiquen tres models: el d'heterogeneïtat, l'autoregressiu condicional (CAR) intrínsec i l'autoregressiu condicional (CAR) no intrínsec.

El model d'heterogeneïtat ens permet recollir la sobredispersió no estructurada. La matriu es definirà com $\Sigma(\sigma_H^2) = \sigma_H^2 I_N$, on I_N és la matriu identitat de dimensió N. Per tant, sota aquest model els efectes aleatoris són independents i idènticament distribuïts on el paràmetre de la variància σ_H^2 controla aquesta sobredispersió no estructurada.

A partir del model CAR intrínsec, es controlarà la sobredispersió deguda a la correlació espacial entre el risc de les regions veïnes. Aquest model fou introduït per Besag (1974) en l'àmbit de procés d'imatges i Clayton i Kaldor (1987) l'aplicaren en aquest context. La matriu es defineix com $\Sigma(\sigma_S^2) = \sigma_S^2 Q^{-1}$, on Q és una matriu quadrada de dimensió N (nombre de regions), la qual es determina mitjançant l'estructura de veïnatge de les regions. En el nostre cas es consideren dues regions com a veïnes si comparteixen els límits geogràfics. Aleshores, els elements de la diagonal de la matriu de veïnatge són el nombre de veïns de cada regió, i els elements de fora de la diagonal prendran el valor –1 si les dues regions són veïnes, o el valor 0 en el cas contrari.

Sota el model CAR intrínsec, la distribució condicionada dels efectes aleatoris és una distribució Normal de paràmetres:

$$b_i \mid b_{j\neq i} \sim N\left(\overline{b}_i, \sqrt{\frac{\sigma_s^2}{n_i}}\right),$$

on:

$\overline{b}_i = \frac{\sum_{j\in\partial_i} b_i}{n_i}$;

∂_i és el conjunt de regions veïnes i n_i, el nombre de veïns de la i-*èsima* regió. Per tant, $\overline{b_i}$ és la mitjana dels efectes aleatoris de les regions adjacents. Aquest model es pot entendre com una suavització del risc a l'àmbit local, atès que una regió concreta tendirà a tenir un risc similar a la de les regions veïnes.

La variància dels efectes aleatoris, $\frac{\sigma_s^2}{n_i}$, és inversament proporcional al nombre de regions adjacents de la regió i-*èsima*; per tant, com més veïns tingui una regió més estable serà l'efecte aleatori.

El model CAR no intrínsec introduït per Besag, York i Mollié (1991) és una conjunció dels dos models anteriors perquè incorpora dos components de la variància: un per modelar la sobredispersió no estructurada i l'altre per modelar sobredispersió estructurada. Aleshores la matriu es defineix com $\Sigma\left(\sigma_s^2, \sigma_H^2\right) = \sigma_S^2 Q_N^{-1} + \sigma_H^2 I_N$.

En aquest cas, la distribució condicional dels efectes aleatoris serà:

$$E\left[b_i \mid b_j \; i \neq j\right] = \overline{b}_i = \frac{\sum_{j\in\partial_i} b_i}{n_i}$$

$$\mathrm{Var}\left(b_i \mid b_j \; i \neq j\right) = \sigma_H^2 + \frac{\sigma_S^2}{n_i}.$$

Per tant, l'esperança és la mateixa que en el model CAR intrínsec, però la variància ara està definida mitjançant els dos components de la variància.

S'ha proposat una nova parametrització del model CAR no intrínsec (Leroux *et al.*, 1999) que també incloïa dos components, un per a la sobredispersió, σ, i un altre per mesurar el pes entre la sobredispersió estructurada i la no estructurada, λ , que anomenarem *paràmetre de dependència*. Aleshores la matriu de variàncies i covariàncies dels efectes aleatoris s'especifica com

$$\Sigma\left(\sigma^2, \lambda\right) = \sigma^2 R^- = \sigma^2\left[\lambda Q_N + (1-\lambda) I_N\right]^-.$$

El paràmetre λ pot prendre valors entre 0 i 1, on una $\lambda = 1$ implica que $\Sigma\left(\sigma^2, \lambda = 1\right) = \sigma^2 R^- = \sigma^2 Q_N^{-1}$. I, per tant, ens trobaríem amb el model CAR intrínsec, és a dir, tota la sobredispersió és deguda a correlació espacial. En canvi, un valor de λ igual a 0, implicaria $\Sigma\left(\sigma^2, \lambda = 0\right) = \sigma^2 R^- = \sigma^2 I_N$, la qual ens condueix a un model d'heterogeneïtat, i per tant tota la sobredispersió és no estructurada espacialment.

A partir d'aquesta parametrització, la mitjana i la variància de la distribució condicional dels efectes aleatoris són:

$$E(b_i \mid b_j \; i \neq j) = \frac{\lambda n_i}{1-\lambda+\lambda n_i} \frac{\sum_{j \in \partial_i} b_i}{n_i} = \frac{\lambda n_i}{1-\lambda+\lambda n_i} \frac{\sum_{j \in \partial_i} b_i}{n_i}$$

$$Var(b_i \mid b_j \; i \neq j) = \frac{1-\lambda}{1-\lambda+\lambda n_i}\sigma^2 + \frac{\lambda n_i}{1-\lambda+\lambda n_i}\frac{\sigma^2}{n_i} = \frac{\sigma^2}{1-\lambda+\lambda n_i}.$$

Tal com es pot observar, la mitjana i la variància condicionals depenen del nombre de veïns i també del paràmetre de dependència.

3.3. Estimació dels models lineals generalitzats mixtos

Com s'ha mencionat en l'apartat 3.1, l'estimació en els models lineals generalitzats mixtos no és una qüestió trivial, perquè la derivació de la versemblança sovint no es pot aconseguir d'una manera analítica. Han estat proposats diferents mètodes per estimar els paràmetres, els quals es poden dividir en mètodes bayesians i freqüentistes segons la perspectiva en què estan basats.

La utilització de les tècniques bayesianes va ser introduïda per Clayton i Kaldor (1987) i més tard va ser desenvolupada per Besag *et al.* (1991), Clayton i Bernardinelli (1992) i Bernardinelli i Montomili (1992). Aquestes tècniques es basen en l'estimació de la distribució de probabilitat posterior dels paràmetres, obtinguda de combinar la versemblança de les dades amb les creences *a priori* dels paràmetres desconeguts.

Dintre dels mètodes freqüentistes, una primera solució seria la de màxima versemblança. Però, com ja s'ha dit, l'optimització de la versemblança quan els efectes aleatoris no són independents pot ser intractable. Els últims anys han aparegut diferents mètodes que permeten obtenir estimacions dels paràmetres, així Breslow i Clayton (1993) han proposat dues aproximacions basades en la quasiversemblança: la quasiversemblança penalitzada (PQL) i la quasiversemblança marginal (MQL). D'altra banda, McCullogh (1997) ha implementat la realització de l'estimació dels paràmetres mitjançant màxima versemblança a partir de tècniques de simulació com: *Monte Carlo Expectation-Maximization* (MCEM), *Monte Carlo Newton Raphson* (MCNR) i simulació de màxima versemblança (SML). També s'han efectuat aproximacions mitjançant la *Conditional Least Squares* (Yasui i Lele, 1997) i l'algoritme *Expectation-Maximization* (EM) (Militino *et al.*, 2001).

4. Objectius

L'objectiu general d'aquesta tesi és aportar nous coneixements per millorar el modelatge dels riscos amb dades agregades geogràficament en retícules. Els objectius específics són:

1. Avaluar i comparar els procediments d'estimació quasiversemblança penalitzada i *fully bayesian* utilitzats en el modelatge de dades agregades per regions geogràfiques.

2. Millorar la convergència del model CAR no intrínsec quan s'estima mitjançant la quasiversemblança penalitzada.
3. Estudiar el comportament de diverses proves per testar la independència espacial en estudis geogràfics i proposar un estadístic per prendre decisions.

5. Estructura de la tesi

En el capítol I es comparen dos mètodes d'estimació, el *fully bayesian* i la quasiversemblança penalitzada, mitjançant un estudi de simulació.

En el capítol II s'aborda el problema de la convergència dels models quan són estimats per quasiversemblança penalitzada, i es proposa una nova parametrització de la matriu de variàncies i covariàncies dels efectes aleatoris que millora la convergència.

En el capítol III s'avalua el comportament de diferents proves per contrastar la hipòtesi d'independència espacial de les dades, i es proposa una nova mesura útil per prendre decisions sobre aquesta hipòtesi.

En el capítol IV es presenta una il·lustració de les aportacions dels capítols anteriors mitjançant l'anàlisi de les dades de diabetis de tipus I a Catalunya.

Finalment, en el capítol V es resumeixen les principals conclusions d'aquesta tesi.

CAPÍTOL I: MODELACIÓ DE LA SOBREDISPERSIÓ EN ELS ESTUDIS GEOGRÀFICS DEL RISC D'UNA MALALTIA: ESTIMACIÓ DE LA QUASIVERSEMBLANÇA PENALITZADA ENFRONT DE L'ESTADÍSTICA BAYESIANA (*FULLY BAYESIAN*) UTILITZANT LA TÈCNICA DE MOSTRATGE DE GIBBS

1.1. Introducció

L'existència de diferents mètodes d'estimació dels paràmetres d'un model lineal generalitzat mixt implica que abans de començar el modelatge s'hagi de seleccionar un d'aquests mètodes. Per fer aquesta selecció és pertinent conèixer els avantatges i els inconvenients d'aquestes tècniques. Per això, l'objectiu d'aquest capítol és comparar dues de les tècniques més emprades en la bibliografia quan s'estudia la variabilitat geogràfica del risc d'una malaltia al llarg d'una àrea, les tècniques conegudes com a *fully bayesian* i PQL, la quasiversemblança penalitzada.

En aquest context, es considera que el nombre de casos de l'esdeveniment en estudi a cada regió es distribueix sota una distribució de Poisson de mitjana que varia de regió a regió, i aquesta sobredispersió es recull mitjançant efectes aleatoris. Els models que es consideren per als efectes aleatoris són el d'heterogeneïtat, el CAR intrínsec i el CAR no intrínsec definit per Besag, York i Mollié (1991). Les dues tècniques es comparen a partir d'un exemple amb dades reals i mitjançant un estudi de simulació. Les dades reals corresponen al nombre de casos nous de diabètics de tipus I a Catalunya en el període 1989-1999. Pel que fa a la simulació, s'han generat dades tenint en compte els tres models definits pels efectes aleatoris i considerant diferents nombres de regions i de casos esperats.

En l'apartat 2 del capítol, s'introdueixen els dos mètodes d'estimació considerats pels models lineals generalitzats mixtos. En l'apartat 3 es defineix el model per estudiar el risc geogràfic d'una malaltia i les característiques dels mètodes d'estimació en aquest cas. L'aplicació a dades reals es troba en l'apartat 4 i l'estudi de simulació, en l'apartat 5. Finalment, es discuteixen els resultats i s'exposen les conclusions principals en l'apartat 6.

1.2. Estimació

1.2.1. Estadística bayesiana

1.2.1.1. Introducció a l'estadística bayesiana

En l'estadística clàssica quan es vol fer inferència sobre un paràmetre desconegut, θ, aquest es considera constant. L'estimació d'aquest paràmetre es realitza mitjançant una funció d'una mostra, funció que es coneix com a *estimador* i es considera una variable aleatòria. En canvi, des d'una perspectiva bayesiana (Box i Tiao, 1992), aquest paràmetre desconegut és una variable, i com a tal se li pot assignar una distribució de probabilitat, $f(\theta)$. Aquesta distribució, que es coneix com a distribució inicial o *a priori* del paràmetre, reflecteix les creences o els coneixements que té l'investigador sobre el paràmetre. Aleshores, la inferència bayesiana combina la informació de les dades, *x*, que es recull mitjançant la funció de versemblança, $l(x;\theta)$, amb la distribució inicial del paràmetre. Això implica que el coneixement dels paràmetres es modifica amb les dades mitjançant el teorema de Bayes:

$$f(\theta \mid x) = \frac{f(x \mid \theta)\cdot f(\theta)}{f(x)} \propto f(x \mid \theta)\cdot f(\theta) = l(x;\theta)\cdot f(\theta),$$

i s'obté la distribució *a posteriori* o final del paràmetre, $f(\theta \mid x)$. Aquesta distribució posterior és el resultat final de l'anàlisi bayesiana i tota la informació sobre el paràmetre s'obté d'aquesta. Així podem obtenir estimacions puntuals del paràmetre mitjançant la mitjana, mediana o moda de la distribució posterior, o estimacions per interval mitjançant intervals de probabilitat.

En la majoria de problemes hi ha més d'un paràmetre d'interès, i aleshores cal diferenciar entre la distribució posterior conjunta i la distribució posterior marginal de cada paràmetre. El primer objectiu és trobar la distribució conjunta posterior de tots els paràmetres, i després obtenir la distribució marginal de cadascun dels paràmetres mitjançant la integració de la funció conjunta respecte de la resta de paràmetres. Així, si considerem que θ consisteix en dos paràmetres $\theta = (\theta_1, \theta_2)$, la distribució conjunta posterior serà:

$$f(\theta_1, \theta_2 \mid x) \propto l(x;\theta_1,\theta_2)\cdot f(\theta_1,\theta_2).$$

La distribució marginal posterior per θ_1 s'obté a partir de la integral següent:

$$f(\theta_1 \mid x) = \int f(\theta_1, \theta_2 \mid x) d\theta_2 ,$$

i per θ_2

$$f(\theta_2 \mid x) = \int f(\theta_1, \theta_2 \mid x) d\theta_1 .$$

Per tant, l'estimació dels paràmetres en l'anàlisi bayesiana se centrarà en la distribució marginal posterior per a cada paràmetre d'interès.

Addicionalment és necessari estructurar un model probabilístic per recollir les creences de l'investigador definint la distribució *a priori* dels paràmetres. Si no es té cap informació de partida i es vol ser objectiu, una de les possibilitats és elegir una distribució no informativa. Per exemple, donar la mateixa probabilitat a tots els valors que pot prendre θ utilitzant una distribució uniforme, o considerant una distribució proporcional a una constant, $f(\theta) \propto \text{cte}$. Aquest últim cas ens condueix a una distribució impròpia, $\sum_{\theta=1}^{\infty} f(\theta) = \infty$, i per tant viola l'assumpció que la suma de probabilitats sigui 1. Però com que en l'estadística bayesiana la distribució *a priori* es combina amb la versemblança, el fet que la distribució *a priori* sigui impròpia no és un problema sempre que la distribució posterior sigui correcta.

Finalment, el mètode d'estimació rep la denominació de *fully bayesian* si s'especifica la distribució *a priori* de tots els paràmetres desconeguts.

1.2.1.2. Estimació dels models lineals generalitzats mixtos via *fully bayesian*

En un model lineal generalitzat mixt es relaciona la mitjana condicionada als efectes aleatoris, μ^b, de la variable resposta, Y, amb les variables explicatives mitjançant una funció d'enllaç $g(\cdot)$. El model és:

$$g(\mu^b) = \eta^b = X\beta + Zb ,$$

on X és la matriu de disseny dels efectes fixos, β és el vector de paràmetres dels efectes fixos, Z és la matriu de disseny dels efectes aleatoris i b és un vector de dimensió N que conté els efectes aleatoris que es distribueixen sota una distribució Normal multivariant amb vector mitjana 0 i matriu de covariàncies $\Sigma(\delta)$, essent δ els components de la

variància. Aquesta distribució conjunta dels efectes aleatoris s'expressa mitjançant $f[b \mid \delta]$, i la distribució marginal de cada efecte aleatori es denota com $f[b_i \mid b_j\ j \neq i, \delta]$.

En aquest model els paràmetres desconeguts són β, *b* i δ, i caldrà assignar-los una distribució *a priori* per a cadascun sobre la base de les creences de l'investigador, que designarem com $f[\beta]$, $f[b \mid \delta]$, $f[\delta]$. La distribució conjunta posterior dels paràmetres desconeguts s'obté mitjançant el teorema de Bayes:

$$f[b, \delta, \beta \mid Y] \propto f[Y \mid b, \delta, \beta] \cdot f[b \mid \delta] \cdot f[\delta] \cdot f[\beta],$$

i les distribucions marginals posteriors dels paràmetres estan proporcionades per les integrals següents:

$$f[\beta \mid Y] = \int_{\beta} \int_{b_1} \int_{b_2} \dots \int_{b_N} \int_{\delta} f[b, \delta, \beta \mid Y] d\beta db_1 db_2 \dots db_N d\delta$$

$$f[b_i \mid Y] = \int_{\beta} \int_{b_1} \dots \int_{b_{i-1}} \int_{b_{i+1}} \dots \int_{b_N} \int_{\delta} f[b, \delta, \beta \mid Y] d\beta db_1 db_{i-1} db_{i+1} \dots db_N d\delta$$

$$f[\delta \mid Y] = \int_{\beta} \int_{b_1} \int_{b_2} \dots \int_{b_N} f[b, \delta, \beta \mid Y] d\beta db_1 db_2 \dots db_N .$$

L'estimació puntual dels paràmetres desconeguts s'obtindrà a partir de l'esperança de la distribució marginal posterior, $E(\delta \mid Y)$, $E(\beta \mid Y)$, $E(b_i \mid Y)$.

L'obtenció d'aquestes distribucions sovint comporta resoldre integrals complicades i fins i tot intractables. Per evitar de resoldre analíticament aquestes integrals es poden utilitzar tècniques de simulació com ara cadenes de Markov-Montecarlo (MCMC) (Gilks, *et al.,* 1996). Amb aquest procediment, les integrals s'avaluen mitjançant la integració de Montecarlo, que utilitza els valors simulats de tots els paràmetres desconeguts $\theta = (b, \delta, \beta)$, els quals són generats a partir d'una cadena de Markov, és a dir, que la generació d'un nou valor per un paràmetre es condiciona al valor anterior.

Així, l'obtenció de mostres independents de θ_k a partir d'un procés de Markov compleix que

$$f\left(\theta_k^{(t+1)} \mid \theta_k^{(t)}, \ldots \theta_k^0\right) = f\left(\theta_k^{(t+1)} \mid \theta_k^{(t)}\right),$$

on la seqüència $\theta_k^{(0)}, \theta_k^{(1)} \ldots \theta_k^{(t+1)}$ és una cadena de Markov. Per tant, el mètode MCMC permet generar una mostra de la distribució marginal posterior i es poden calcular les propietats d'interès d'aquesta distribució. L'algoritme MCMC més utilitzat i conegut és el mostratge de Gibbs (Casella i George, 1992), que treballa a partir de les distribucions univariants condicionades.

1.2.1.3. Mostratge de Gibbs

Suposem que tenim *n* paràmetres $\theta_1, \ldots \theta_n$ amb distribucions condicionades

$$f\left(\theta_1 \mid \theta_2, \ldots \theta_n\right),\ f\left(\theta_2 \mid \theta_1, \ldots \theta_n\right), \ldots, f\left(\theta_n \mid \theta_2, \ldots \theta_{n-1}\right),$$

i uns valors inicials per a cada paràmetre $\left[\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)}\right]$. El primer pas del mostratge de Gibbs és generar una nova mostra $\left[\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_n^{(1)}\right]$ a partir de les distribucions condicionades als valors inicials. El procediment comença amb la generació de l'observació aleatòria $\theta_1^{(1)}$ a partir de la distribució condicionada de θ_1 fixant tots els altres paràmetres al valor inicial,

$$\theta_1^{(1)} \sim f\left(\theta_1 \mid \theta_2^{(0)}, \ldots \theta_n^{(0)}\right),$$

després es genera l'observació aleatòria de $\theta_2^{(1)}$ a partir de $f\left(\theta_2 \mid \theta_1^{(0)}, \ldots \theta_n^{(0)}\right)$ i així successivament fins al paràmetre $\theta_n^{(1)}$. Un cop s'han actualitzat tots els paràmetres s'obté una seqüència de Gibbs $\left[\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_n^{(1)}\right]$ i el procés torna a començar, però ara en lloc de condicionar els paràmetres als valors inicials, es condicionen als valors generats en la darrera iteració. El procés es pot repetir fins a obtenir *k* mostres dels paràmetres desconeguts, de manera que es generà la cadena de Gibbs següent:

$$\left[\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)}\right], \left[\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_n^{(1)}\right], \ldots\ldots, \left[\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_n^{(k)}\right].$$

En el model que s'ha definit, les distribucions condicionades de cada paràmetre es poden resumir així:

$$f[\delta \mid b, \beta, Y] \propto f[\delta] \cdot f[b \mid \delta]$$

$$f[b_i \mid b_j \; j \neq i, \delta, \beta, Y] \propto f[b_i \mid b_j \; j \neq i] \cdot l[Y_i; \beta, \delta]$$

$$f[\beta \mid b, \delta, Y \cdot] \propto f[\beta] \cdot \prod_{i=1}^{N} l[Y_i; \beta, \delta].$$

Així doncs, en el cas que ens afecta el procés del Gibbs es resumeix en les etapes següents:

1. Assignació d'uns valors inicials per als paràmetres d'interès: $\delta^{(0)}, b_1^{(0)}, \ldots b_N^{(0)}, \beta^{(0)}$.

2. Obtenció d'un valor nou de $\delta^{(1)}$ a partir de la seva distribució condicionada. Els paràmetres als quals està condicionada són substituïts pels valor inicials:

$$f[\delta \mid b^{(0)}, \beta^{(0)}, Y] \propto f[\delta] \cdot f[b^{(0)} \mid \delta].$$

3. Obtenció d'un valor nou del primer efecte aleatori $b_1^{(1)}$ a partir de:

$$f[b_1 \mid b_j^{(0)} \; j \neq 1, \delta^{(0)}, \beta^{(0)}, Y] \propto f[b_1 \mid b_j^{(0)} \; j \neq 1] \cdot l[Y_i; b^{(0)}, \beta, \delta^{(0)}],$$

i d'igual manera per a la resta d'efectes aleatoris fins a l'efecte aleatori N-*èsim* $b_N^{(1)}$:

$$f[b_N \mid b_j^{(0)} \; j \neq N, \delta^{(0)}, \beta^{(0)}, Y] \propto f[b_N \mid b_j^{(0)} \; j \neq N] \cdot l[Y_i; b^{(0)}, \beta, \delta^{(0)}].$$

4. Finalment, obtenció d'un valor nou de l'efecte fix β a partir de:

$$f[\beta \mid b^{(0)}, \delta^{(0)}, Y \cdot] \propto f[\beta] \cdot \prod_{i=1}^{N} l[Y_i; \beta, b^{(0)}, \delta^{(0)}].$$

El procés torna a començar però substituint els valors dels paràmetres pels obtinguts en aquesta nova mostra, fins que la cadena de mostres generi la distribució posterior.

De vegades, les distribucions condicionades que s'utilitzen no presenten una forma analíticament senzilla. Aleshores, això impedeix que mètodes de generació de nombres aleatoris –com ara el mètode Invers– no es puguin utilitzar, i cal utilitzar mètodes més sofisticats com per exemple el del rebuig adaptatiu (Gilks i Wild, 1992).

En conclusió, mitjançant el mostratge de Gibbs ens passegem per l'espai mostral, de manera que el pas següent que cal donar se sorteja des de la posició que s'ha arribat en la simulació anterior.

Normalment, les mostres obtingudes de les iteracions inicials es descarten perquè les mostres no estiguin condicionades a la tria dels valors inicials, procés conegut com a *escalfament*. Un cop se supera la fase d'*escalfament* es comprova si les cadenes simulades provenen de la distribució posterior, fet que es coneix com a *convergència* a la distribució posterior. Així, si hi ha convergència, les mostres extretes s'utilitzaran per calcular les propietats de la distribució posterior. En cas contrari, es continua generant valors fins que s'arribi a la convergència.

Per comprovar si la mostra simulada o cadena per cada paràmetre convergeix a la distribució posterior es pot utilitzar el test de Geweke (Geweke, 1992). Aquest test divideix la mostra en dues parts, una amb les primeres *x*% iteracions i l'altra amb les últimes *y*% iteracions. Si la cadena és estacionària, la mitjana dels primers valors i la dels últims seran iguals. Per diagnosticar la convergència es calcula l'estadístic *Z* com la diferència entre les mitjanes dividit per l'error estàndard de la diferència. Quan la mida de la mostra tendeix a infinit aquest estadístic segueix una distribució normal tipificada. Es conclou que la cadena ha convergit si el valor de *Z* no excedeix els valors crítics basats en la distribució normal estàndard.

1.2.2. Quasiversemblança penalitzada

1.2.2.1. Definició i estimació

S'assumeix que les dades, Y_i, condicionades als efectes aleatoris són independents, amb $E(Y_i \mid b_i) = \mu_i^b$ i $Var(Y_i \mid b_i) = a(\phi)Var(\mu_i^b)$ i que es distribueixen sota una distribució de la família exponencial. Aleshores, tal com s'ha expressat en l'apartat 1.2.1.1, el model expressat matricialment és:

$$g(\mu^b) = X\beta + Zb,$$

on recordem que $g(\cdot)$ és la funció d'enllaç canònica, X és la matriu de disseny dels efectes fixes, β és el vector de paràmetres dels efectes fixos, Z és la matriu de disseny

dels efectes aleatoris i *b* és el vector de paràmetres dels efectes aleatoris que es distribueixen sota una distribució normal multivariant amb vector mitjana 0 i matriu de covariàncies $\Sigma(\delta)$.

La funció de versemblança d'un model lineal generalitzat mixt, com ja s'ha mencionat en la introducció, sovint presenta una expressió analítica difícil de derivar, i per tant l'estimació de màxima versemblança dels paràmetres del model pot arribar a ser extremament complicada. Breslow i Clayton (1993) van proposar estimar els paràmetres fixos i aleatoris del model substituint la funció de densitat de les dades condicionades als efectes aleatoris $f(Y \mid b)$ per la funció de quasiversemblança. Aleshores, van aproximar la integral que defineix la versemblança mitjançant el mètode de Laplace. Per estimar els components de la variància, δ, van utilitzar la màxima versemblança restringida (REML) dels efectes aleatoris.

Així doncs, el procés parteix de l'expressió de la funció de màxima versemblança d'aquest model:

$$L = \int f[Y \mid b] \cdot f[b] db = \int \exp\{\log(f[Y \mid b]) + \log(f[b])\} db,$$

on $f[Y \mid b]$ és la funció de densitat de probabilitat de les dades condicionada als efectes aleatoris, i $f[b]$ és la funció de densitat de probabilitat dels efectes aleatoris.

Atès que s'assumeix que els efectes aleatoris es distribueixen sota una distribució normal de paràmetres $N(0, \Sigma(\delta))$, el $\log(f[b])$ serà:

$$\log(f[b]) = -\frac{1}{2} b^T \Sigma(\delta) b - \frac{N}{2} \log(2 \cdot \pi) - \frac{1}{2} \log|\Sigma(\delta)|.$$

Substituint aquesta expressió en la funció de versemblança s'obté:

$$L = \int \exp\left\{ \log(f[Y \mid b]) - \frac{1}{2} b^T \Sigma(\delta) b - \frac{N}{2} \log(2 \cdot \pi) - \frac{1}{2} \log|\Sigma(\delta)| \right\} db.$$

El terme $\frac{N}{2} \log(2 \cdot \pi)$ es pot ometre perquè no depèn de cap paràmetre i per tant no intervé en el procés de maximització de la versemblança.

Breslow i Clayton (1993) en comptes de $\log(f[Y \mid b])$ utilitzen la funció de quasiversemblança:

$$Q = \sum_{i=1}^{N} \int_{Y_i}^{\mu_i} \frac{Y_i - u}{\mathrm{var}(u)} du ,$$

que es caracteritza perquè només requereix que es defineixi quina és la funció d'enllaç entre la mitjana i les covariables i quina és la relació entre la mitjana i la variància de les dades.

Substituint $\log(f[Y \mid b])$ per la quasiversemblança s'arriba a:

$$\begin{aligned} PQL &= \int \exp\left\{ \sum_{i=1}^{N} \int_{Y_i}^{\mu_i} \frac{Y_i - u}{\mathrm{var}(u)} du - \frac{1}{2} b^T \Sigma(\delta) b - \frac{1}{2} \log|\Sigma(\delta)| \right\} db = \\ &= \int \exp\left\{ \sum_{i=1}^{N} \int_{Y_i}^{\mu_i} \frac{Y_i - u}{\mathrm{var}(u)} du - \frac{1}{2} b^T \Sigma(\delta) b + \log|\Sigma(\delta)|^{-\frac{1}{2}} \right\} db = \\ &= |\Sigma(\delta)|^{-\frac{1}{2}} \int \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \left[(-2) \int_{Y_i}^{\mu_i} \frac{Y_i - u}{\mathrm{var}(u)} du \right] - \frac{1}{2} b^T \Sigma(\delta) b \right\} db. \end{aligned}$$

No obstant això, per estimar els efectes fixos i aleatoris continua sent necessari resoldre la integral. Breslow i Clayton (1993) la resolen utilitzant l'aproximació de Laplace. Aquest mètode resol integrals del tipus $|\Sigma(\delta)|^{-\frac{1}{2}} \int \exp\{-h(b)\} db$ mitjançant l'expressió següent:

$$|\Sigma(\delta)|^{-\frac{1}{2}} \int \exp\{h(b)\} db \approx -\frac{1}{2} \log(|\Sigma(\delta)|) - \frac{1}{2} \log\left(\left| \frac{\partial^2 h(b)}{\partial^2 b} \right| \right) - \frac{\partial h(b)}{\partial b} .$$

Aplicant aquest mètode a PQL, s'obté la funció següent:

$$-\frac{1}{2} \log| I - Z^T W Z D| - \frac{1}{2} \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - u}{\mathrm{var}(u)} du - \frac{1}{2} b^T \Sigma(\delta)^{-1} b ,$$

on, com en el cas dels GLM, *W* és la matriu diagonal de pesos amb termes a la diagonal

$$w_i = \frac{1}{a_i(\phi) v(\mu_i^b)} \left[\frac{\partial \mu_i^b}{\partial \eta_i} \right]^2 .$$

Si s'ignora el terme $-\frac{1}{2} \log|I - Z^T W Z \Sigma(\delta)|$, tal com van fer Breslow i Clayton (1993), la funció que s'ha de maximitzar correspon a:

$$-\frac{1}{2}\sum_{i=1}^{n}\int_{y_i}^{\mu_i}\frac{y_i-u}{var(u)}du-\frac{1}{2}b^T\Sigma(\delta)^{-1}b\ .$$

Per tant, aquesta funció es compon de la funció de quasiversemblança, *Q*, i d'una funció de penalització, que correspon al terme $-\frac{1}{2}b^T\Sigma(\delta)b$. Per aquest fet aquesta funció rep el nom de *quasiversemblança penalitzada*.

Per poder obtenir l'estimació dels paràmetres, aquesta expressió s'ha de derivar respecte de β i *b*. Així doncs, les equacions per resoldre són:

$$\sum_{i=1}^{n}\frac{(Y_i-\mu_i^b)x_i}{a_i(\phi)var(\mu_i^b)}\left(\frac{\partial\mu_i^b}{\partial\eta_i^b}\right)=0 \qquad \sum_{i=1}^{n}\frac{(Y_i-\mu_i^b)z_i}{a_i(\phi)var(\mu_i^b)}\left(\frac{\partial\mu_i^b}{\partial\eta_i^b}\right)=\Sigma(\delta)^{-1}b\ .$$

Aquestes equacions es resolen mitjançant l'algoritme de Fisher-Scoring, on l'actualització dels paràmetres β a la iteració *r* es realitza mitjançant l'equació següent:

$$\hat{\beta}^{(r)}=\left(X^tV^{-1}(\beta^{(r-1)},\delta^{(r-1)})X\right)^{-1}X^tV^{-1}(\beta^{(r-1)})Y^*,$$

i la dels paràmetres *b* mitjançant:

$$\hat{b}^{(r)}=\Sigma(\delta^{(r-1)})Z^tV^{-1}(\beta^{(r-1)},\delta^{(r-1)})(Y^*-X\hat{\beta}^{(r)}),$$

on

$$V(\beta^{(r-1)},\delta^{(r-1)})=W^{-1}(\beta^{(r-1)})+Z\cdot\Sigma(\delta^{(r-1)})\cdot Z^t,$$

i , a l'igual dels models lineals generalitzats, el vector *Y** es defineix com:

$$Y_i^*=\eta_i^b+(Y_i-\mu_i^b)\frac{\partial\eta_i^b}{\partial\mu_i^b}.$$

La variància de les estimacions dels paràmetres β i *b* s'aproxima mitjançant:

$$Var(\hat{\beta})=\left(X^tV^{-1}X\right)^{-1}\ \text{i}\ Var(\hat{b})=\left(\Sigma(\delta)Z^tV^{-1}X\right)^t\left(X^tVX\right)^{-1}\left(\Sigma(\delta)Z^tV^{-1}X\right).$$

1.2.2.2. Estimació dels components de la variància

L'estimació dels components de la variància δ s'efectua utilitzant la màxima versemblança restringida dels efectes aleatoris (REML) (Patterson i Thompson, 1974). Així, l'equació que cal maximitzar és:

$$\text{REML} \approx -\frac{1}{2}\log|\mathrm{V}| - \frac{1}{2}\log\left|\mathrm{X}^{\mathrm{t}}\mathrm{V}^{-1}\mathrm{X}\right| - \frac{1}{2}\left(\mathrm{Y}^{*} - \mathrm{X}\hat{\beta}\right)^{\mathrm{t}}\mathrm{V}^{-1}\left(\mathrm{Y}^{*} - \mathrm{X}\hat{\beta}\right).$$

Si es deriva la REML respecte de δ s'obtenen les equacions d'estimació dels components de la variància δ:

$$\mathrm{U}_{\delta} = -\frac{1}{2}\left[\left(\mathrm{Y}^{*} - \mathrm{X}\hat{\beta}\right)^{\mathrm{t}}\mathrm{V}^{-1}\frac{\partial \mathrm{V}}{\partial \delta}\mathrm{V}^{-1}\left(\mathrm{Y}^{*} - \mathrm{X}\hat{\beta}\right) - \mathrm{tr}\left(\mathrm{P}\frac{\partial \mathrm{V}}{\partial \delta}\right)\right] = 0\,,$$

on $\mathrm{P} = \mathrm{V}^{-1} - \mathrm{V}^{-1}\mathrm{X}\left(\mathrm{X}^{\mathrm{t}}\mathrm{V}^{-1}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{t}}\mathrm{V}^{-1}$ i k = 1,...K; *K* és el nombre de paràmetres de variància que s'ha d'estimar. Aquestes equacions es poden resoldre mitjançant l'algoritme de Fisher-Scoring, on les estimacions a la iteració *r* s'obtenen mitjançant:

$$\delta^{(\mathrm{r})} = \delta^{(\mathrm{r}-1)} + \mathrm{I}_{\delta}^{-1} \cdot \mathrm{U}_{\delta}\,|_{\delta=\delta^{(\mathrm{r})}}\,,$$

on I_{δ} és la matriu d'informació de Fisher que és composta pels elements següents:

$$\mathrm{I}_{\mathrm{kl}} = -\frac{1}{2}\mathrm{tr}\left(\mathrm{P}\frac{\partial \mathrm{V}}{\partial \delta_{\mathrm{k}}}\mathrm{P}\frac{\partial \mathrm{V}}{\partial \delta_{\mathrm{l}}}\right).$$

En resum, l'estimació per PQL es fa seguint el procés següent:

1. Es defineixen uns valors inicials pels efectes fixos, $\hat{\beta}^{(0)}$, assumint que els efectes aleatoris són igual a 0. Amb aquests valors i uns valors inicials per $\delta^{(0)}$ es calcula *W* i *V* i s'obtenen $\hat{\beta}^{(1)}$ i $\hat{\mathrm{b}}^{(1)}$ a partir de les equacions:

$$\hat{\beta}^{(1)} = \left(\mathrm{X}^{\mathrm{t}}\mathrm{V}^{-1}\left(\beta^{(0)}, \delta^{(0)}\right)\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{t}}\mathrm{V}^{-1}\left(\beta^{(0)}, \delta^{(0)}\right)\mathrm{Y}^{*}$$

$$\hat{\mathrm{b}}^{(1)} = \Sigma\left(\delta^{(0)}\right)\mathrm{Z}^{\mathrm{t}}\mathrm{V}^{-1}\left(\beta^{(0)}, \delta^{(0)}\right)\left(\mathrm{Y}* - \mathrm{X}\hat{\beta}^{(0)}\right).$$

2. S'actualitzen les matrius *W* i *V* i s'estimen els components de la variància amb

$$\delta^{(1)} = \delta^{(0)} + \mathrm{I}_{\delta}^{-1} \cdot \mathrm{U}_{\delta}\,|_{\delta=\delta^{(0)}}\,,$$

i torna a començar el procés fins a obtenir convergència en els paràmetres $\hat{\beta}, \hat{\mathrm{b}}$ i $\hat{\delta}$.

L'estimació de l'error estàndard dels components $\hat{\delta}$ es fa a partir de la inversa de la matriu d'informació de Fisher.

1.3. Estudi geogràfic del risc d'una malaltia al llarg d'una àrea d'estudi

1.3.1. Definició del model

En l'estudi de la variabilitat geogràfica del risc d'una malaltia al llarg d'una àrea, s'assumeix que el nombre de casos de l'esdeveniment d'interès en un període de temps determinat a cada regió en què s'ha subdividit l'àrea d'estudi es distribueix sota una distribució de Poisson. Si la mesura de risc per a cada regió utilitzada és la raó estandarditzada d'incidència o mortalitat, la mitjana de la distribució del nombre de casos serà $\mu = (E_1\psi_1, \ldots E_N\psi_N)$; E_i és el nombre de casos esperats a la regió i-*èsima*, i ψ_i és la mesura de risc de la regió i-*èsima*, amb i = 1...N, en què *N* és el nombre de regions. Aleshores, utilitzant com a funció d'enllaç canònica el logaritme, el model que ens permet incloure variables explicatives i efectes aleatoris serà:

$$g(\mu^b) = \log(\mu^b) = \eta^b = \log(E) + X\beta + Zb,$$

on *X* és la matriu de disseny dels efectes fixos, β és el vector de paràmetres dels efectes fixos, *Z* és la matriu de disseny dels efectes aleatoris que correspon a una matriu identitat i *b* és el vector de paràmetres dels efectes aleatoris que es distribueixen sota una distribució normal multivariant amb vector mitjana 0 i matriu de variàncies i covariàncies $\Sigma(\delta)$.

Segons aquesta definició, el vector de mitjanes és $\mu^b = E \cdot \exp(X\beta + Zb)$, on la i-*èsima* component és $\mu_i^b = \{E_i \exp(x_i^t\beta + z_i^t b)\}$.

1.3.2. Característiques dels mètodes d'estimació

L'estimació per quasiversemblança penalitzada amb la funció logarítmica com a funció d'enllaç es du a terme mitjançant la definició del vector mitjanes, $\mu = E \cdot \exp(X\beta + Zb)$, del vector de dades transformades, $Y^* = X\beta + Zb + \frac{Y - \mu}{\mu}$, de la matriu de pesos, $W = \text{diag}(\mu)$, i de la definició de la matriu de variàncies i covariàncies dels efectes aleatoris, $\Sigma(\delta)$. Les estructures d'aquesta matriu per als tres models són:

Model	$\Sigma(\delta)$
Heterogeneïtat	$\sigma_H^2 I_N$
CAR intrínsec	$\sigma_s^2 Q_N$
CAR no intrínsec	$\sigma_H^2 I_N + \sigma_S^2 Q_N$

Les funcions *score*, U_δ, els elements de la matriu d'informació de Fisher, I_δ, i les equacions iteratives per cada component de la variància dels diferents models considerats es presenten en la taula 1.1.

Taula 1.1. Funcions *score* (U_δ), elements de la matriu d'informació de Fisher (I_δ) i equació iterativa dels components de la variància (Eq.) en els tres models considerats

Heterogeneïtat	V	$W^{-1} + Z\left(\sigma_H^2 \cdot I_N\right)Z^t$
	$\frac{\partial V}{\partial \sigma_H}$	$2 \cdot \sigma \cdot Z \cdot I_N \cdot Z^t$
	U_{σ_H}	$-\frac{1}{2}\left[\left(Y^* - X\beta\right)^t V^{-1}\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)V^{-1}\left(Y^* - Xa\right) - tr\left(P\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)\right)\right]$
	I_{σ_H}	$-\frac{1}{2}tr\left(P\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)P\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)\right)$
	Eq.	$\sigma_H^{(r)} = \sigma_H^{(r)} + I^{-1}_{\sigma_H^{(r-1)}} \cdot U_{\sigma_H} \mid_{\sigma_H = \sigma_H^{(r-1)}}$
CAR intrínsec	V	$W^{-1} + Z\left(\sigma_S^2 \cdot Q_N^{-1}\right)Z^t$
	$\frac{\partial V}{\partial \sigma_S}$	$2 \cdot \sigma_S \cdot Z \cdot Q_N^{-1} \cdot Z^t$
	U_{σ_S}	$-\frac{1}{2}\left[\left(Y^* - Xa\right)^t V^{-1}\left(2 \cdot \sigma_S \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)V^{-1}\left(Y^* - Xa\right) - tr\left(P\left(2 \cdot \sigma_S \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)\right)\right]$
	I_{σ_S}	$-\frac{1}{2}tr\left(P\left(2 \cdot \sigma_S \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)P\left(2 \cdot \sigma_S \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)\right)$
	Eq.	$\sigma_S^{(r)} = \sigma_S^{(r)} + I^{-1}_{\sigma_S^{(r-1)}} \cdot U_{\sigma_S} \mid_{\sigma_S = \sigma_S^{(r-1)}}$
CAR no intrínsec	V	$W^{-1} + Z^T \cdot \left(\sigma_s^2 Q_N^{-1} + \sigma_u^2 I_N\right) \cdot Z$
	$\frac{\partial V}{\partial \sigma_H}$	$2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t$
	$\frac{\partial V}{\partial \sigma_S}$	$2 \cdot \sigma_s \cdot Z \cdot Q_N^{-1} \cdot Z^t$
	U_{σ_H}	$-\frac{1}{2}\left[\left(Y^* - Xa\right)^t V^{-1}\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)V^{-1}\left(Y^* - Xa\right) - tr\left(P\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)\right)\right]$
	U_{σ_S}	$-\frac{1}{2}\left[\left(Y^* - Xa\right)^t V^{-1}\left(2 \cdot \sigma_s \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)V^{-1}\left(Y^* - Xa\right) - tr\left(P\left(2 \cdot \sigma_s \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)\right)\right]$
	I_{σ_H}	$-\frac{1}{2}tr\left(P\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)P\left(2 \cdot \sigma_H \cdot Z \cdot I_N \cdot Z^t\right)\right)$
	I_{σ_S}	$-\frac{1}{2}tr\left(P\left(2 \cdot \sigma_s \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)P\left(2 \cdot \sigma_s \cdot Z \cdot Q_N^{-1} \cdot Z^t\right)\right)$
	Eq.	$\sigma_H^{(r)} = \sigma_H^{(r)} + I^{-1}_{\sigma_H^{(r-1)}} \cdot U_{\sigma_H} \mid_{\sigma_H = \sigma_H^{(r-1)}}$ $\sigma_S^{(r)} = \sigma_S^{(r)} + I^{-1}_{\sigma_S^{(r-1)}} \cdot U_{\sigma_S} \mid_{\sigma_S = \sigma_S^{(r-1)}}$

En l'estimació bayesiana, la funció de màxima versemblança assumint que les dades es distribueixen sota una Poisson és:

$$f(Y \mid E, a, b, \Sigma(\delta)) = \prod_{i=1}^{N} f(Y_i \mid E_i, a, b_i, \Sigma(\delta)) = \prod_{i=1}^{N} \exp\left(-\mu_i^b\right) \cdot \frac{\left(\mu_i^b\right)^{Y_i}}{Y_i!} =$$

$$= \prod_{i=1}^{N} \exp\left\{-E_i \exp\left(x_i^t a + z_i^t b\right)\right\} \cdot \frac{\left\{E_i \exp\left(x_i^t a + z_i^t b\right)\right\}^{Y_i}}{Y_i!}.$$

Aleshores, en aquest model els paràmetres desconeguts són β, *b*, i δ, als quals, com s'ha explicat, cal assignar-los una distribució *a priori*. Així, als paràmetres β, atès que no es parteix de cap creença, se'ls assigna una distribució normal difusa, és a dir, una distribució normal de mitjana 0 i variància 1.000. En canvi, la distribució *a priori* que seguiran els efectes aleatoris, $f[b \mid \Sigma(\delta)]$, estarà en funció del model que s'estigui estimant, és a dir, en el model d'heterogeneïtat els efectes aleatoris seguiran una distribució normal multivariant de paràmetres $NM(0, \sigma_H^2 I_N)$, en el CAR intrínsec la parametrització serà $NM(0, \sigma_S^2 Q^-)$ i el CAR no intrínsec es definirà mitjançant $NM(0, \sigma_H^2 I_N + \sigma_S^2 Q^-)$. Finalment, sols queda definir les distribucions pels components de la variància, $\delta = (\sigma_H^2, \sigma_S^2)^T$, coneguts com a *hiperparàmetres*. Una elecció convenient que permet obtenir una distribució *a posterior* no impròpia és assumir que la inversa de la variància es distribueixi segons una gamma. En les anàlisis realitzades, s'ha assumit que la inversa de la variància es distribuirà $f[\delta^-]$, segons una gamma (0,25, 0,005) .

1.3.3. Aspectes computacionals dels mètodes d'estimació

En l'estimació PQL, els valors inicials que s'han determinat per als components de la variància han estat de 0,5, i per als efectes fixos els que s'han obtingut mitjançant màxima versemblança fixant els efectes aleatoris a 0. A més a més, per accelerar la convergència en l'estimació dels paràmetres, s'ha efectuat una convergència prèvia dels efectes fixos i aleatoris fixant els components de la variància als valors inicials. Aquest procés previ finalitza quan la diferència entre les estimacions de la mitjana de cada regió en dues iteracions consecutives és menor que 0,001. Les estimacions dels paràmetres s'han obtingut a partir de les equacions següents:

$$\hat{\beta}^{(r)} = \left(X^t V^{-1}\left(\beta^{r-1}, \delta^{(r-1)}\right)X\right)^{-1} X^t V^{-1}\left(\beta^{r-1}, \delta^{(r-1)}\right)Y^* \text{ i}$$

$$\hat{b}^{(r)} = \Sigma\left(\delta^{(r)}\right)Z^t V^{-1}\left(\beta^{r-1}, \delta^{(r-1)}\right)\left(Y^* - X\hat{\beta}^{(r)}\right).$$

Un cop s'ha arribat a la convergència, s'han actualitzat els components de la variància mitjançant les equacions de Fisher-Scoring presentades en la taula 1.1. En aquesta actualització, per mantenir l'estabilitat d'aquest algoritme, sobretot en les primeres iteracions, s'ha aplicat la tècnica de Marquardt. D'aquesta manera, quan el valor de $I_\delta^{-1} \cdot U_\delta \mid_{\delta=\delta^{(r-1)}}$ ha estat major que 0,5, aquest s'ha reduït en $\frac{1}{s}$, on *s* s'ha obtingut de $s = \frac{1}{0.5}\left\{I_\delta^{-1} \cdot U_\delta \mid_{\delta=\delta^{(r-1)}}\right\}$.

Els valors dels efectes fixos i aleatoris també s'actualitzen en cada pas amb els components de la variància obtingudes. El procés es repeteix fins que els canvis en les estimacions són menors que 0,001 i amb un màxim de 200 iteracions. Aquest procés d'estimació s'ha implementat mitjançant unes funcions creades amb llenguatge S utilitzant el programa S-plus v. 6, les quals es presenten en l'apèndix 1.

La inferència bayesiana ha estat implementada mitjançant el Gibbs Sampling utilitzant el programa BUGS v. 06 (Spiegelhater *et al.*, 1996). Aquest programa s'ha executat des de l'S-plus mitjançant un fitxer executable d'extensió bat. També s'ha comprovat si la cadena ha convergit mitjançant el test de Geweke (1992) implementat en el programa *Convergence Diagnostic and Output Analysis* CODA (Best *et al*., 1995).

En el procés d'estimació mitjançant l'estadística bayesiana s'ha fet un *escalfament* de 10.000 iteracions. Després d'aquest *escalfament*, s'ha realitzat una cadena amb els 5.000 valors següents i se n'ha comprovat la convergència. Si no ha estat així, s'ha tornat a simular 5.000 valors més amb els quals també s'ha comprovat la convergència a la distribució posterior. Si el procés ha continuat sense convergir, s'ha considerat que el model no convergia definitivament i no s'ha obtingut cap estimació. En el cas de convergència, l'estimació puntual i l'error estàndard dels paràmetres s'ha obtingut a partir de la mitjana i la variància de les 5.000 mostres.

En el BUGS, la programació de la funció prior dels efectes aleatoris del model no intrínsec no està definida i per tant no es pot assignar al vector d'efectes aleatoris una distribució normal multivariant de mitjana 0 i matriu de variàncies i covariàncies $\sigma_H^2 I_N + \sigma_S^2 Q^-$. Aleshores, per programar aquest model amb el BUGS s'han definit dos vectors independents d'efectes aleatoris en lloc d'un sol vector $b = (b_1, \ldots b_N)^T$. Aquests dos vectors corresponen als efectes aleatoris d'heterogeneïtat, $h = (h_1, \ldots h_N)^T$, i als efectes aleatoris del CAR intrínsec, $s = (s_1, \ldots s_N)^T$. Per tant, la distribució prior dels efectes $h = (h_1, \ldots h_N)^T$ és $NM(0, \sigma_H^2 I_N)$, i per a $s = (s_1, \ldots s_N)^T$ és $NM(0, \sigma_S^2 Q^-)$. A més, en aquest cas el model que s'ha d'estimar s'expressa mitjançant $\log(\mu^b) = \log(E) + X\beta + Z(h + s)$, que no deixa de ser una altra expressió del $\log(\mu^b) = \log(E) + X\beta + Zb$ però amb una altra parametrització.

En l'apèndix 1 es mostren les funcions utilitzades per a l'anàlisi *fully bayesian* a partir del programa BUGS.

1.4. Incidència de la diabetis de tipus I a Catalunya

En l'exemple següent s'analitzen les dades d'incidència de la diabetis de tipus I per comarques a Catalunya. Aquestes dades consisteixen en el nombre de diabètics menors de 30 anys als quals s'ha diagnosticat la diabetis entre l'any 1989 i l'any 1998. Pel eliminar l'efecte confusor de les variables edat i sexe, les dades s'estandarditzen mitjançant l'estandardització indirecta interna, a partir de la qual s'obté el nombre de casos esperats en cada comarca. Assumint que el nombre de diabètics observats en cada comarca es distribueix sota una distribució de Poisson de mitjana $\mu_i^b = E_i \times \psi_i$, on ψ_i és el risc relatiu, es fita el model següent:

$$\log(\mu_i^b) = \log(E_i) + \beta_o + b_i,$$

on b_i és l'efecte aleatori de la comarca i-*èsima*. Per tant, sota aquest model, l'estimació del risc de patir la diabetis en la comarca i-*èsima* s'obté de l'expressió

$$\psi_i = \exp(\beta_0 + b_i),$$

on ψ_i es coneix com la raó estandarditzada d'incidència (SMR).

Els models considerats pels efectes aleatoris són el CAR intrínsec, el d'heterogeneïtat i el CAR no intrínsec; cadascun s'estima mitjançant les tècniques de la quasiversemblança penalitzada (PQL) i *fully bayesian* (FB) utilitzant el mostratge de Gibbs.

El nombre de comarques en què es divideix el territori de Catalunya és de 41, el nombre de casos de diabètics observats varia entre 1 i 985, i la mitjana del nombre de casos esperats és de 67,59.

En la taula 1.2 es presenten les estimacions amb PQL i FB de l'efecte fix i dels components de la variància conjuntament amb el seu error estàndard per cada model considerat.

Taula 1.2
Estimacions dels paràmetres dels models d'heterogeneïtat, CAR i CAR no intrínsec mitjançant PQL i FB de les dades de diabetis a Catalunya

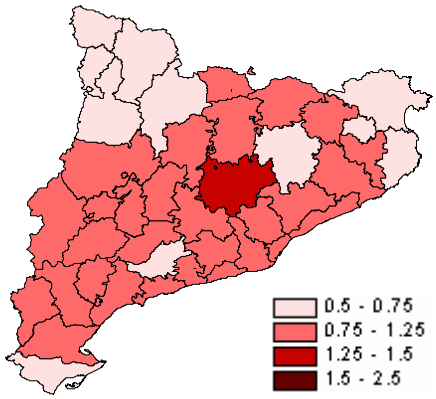
Model		β_0	σ_S	σ_H
CAR intrínsec	PQL	–0,168	0,405	---
		(0,045)	(0,083)	
	FB	–0,176	0,398	---
		(0,049)	(0,093)	
Heterogeneïtat	PQL	–0,134	---	0,232
		(0,051)		(0,046)
	FB	–0,146	---	0,239
		(0,055)		(0,048)
CAR no intrínsec	PQL	–0,150	0,250	0,167
		(0,047)	(0,139)	(0,076)
	FB	–0,185	0,291	0,252
		(0,061)	(0,085)	(0,050)

β_0: coordinada en l'origen; σ_s: desviació típica del component d'espacialitat, i σ_H: desviació típica del component d'heterogeneïtat.

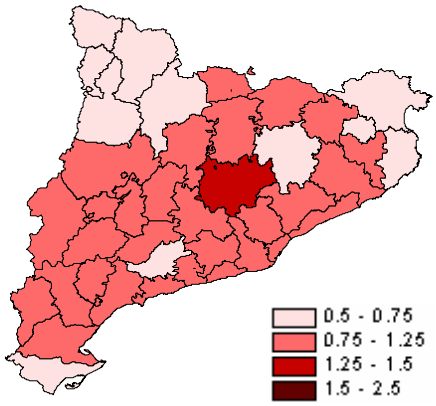
En la taula 1.2 podem observar que tant les estimacions puntuals com la seva variabilitat són molt similars entre ambdós mètodes, fent que les estimacions puntuals de les SMR

també ho siguin. Aquest fet es pot observar en els mapes següents de les SMR, les quals són estimades mitjançant PQL i FB considerant el model CAR no intrínsec.

SMR estimades amb PQL



SMR estimades amb FB



En aquests mapes s'observa que totes les SMR de les comarques es classifiquen en el mateix interval amb les dues tècniques d'estimació. Per tant, es podria pensar que les estimacions fetes tant per PQL com per FB sempre són similars i és indiferent la tècnica que s'utilitzi. Per avaluar aquesta hipòtesi es farà un estudi de simulació en el qual es considerarà un nombre diferent de regions i de casos esperats. Sobre la base d'aquestes combinacions es comparen les estimacions d'ambdues tècniques en termes de biaix i precisió.

1.5. Estudi de simulació

1.5.1 Característiques de l'estudi de simulació

En l'estudi de simulació que es duu a terme s'ha tingut en compte tres àrees compostes per 25, 49 i 100 regions, obtingudes de tres retícules de dimensió 5×5, 7×7 i 10×10, on cada retícula representa una regió. El veïnatge es defineix quan dues retícules comparteixen algun dels seus límits. Per tant, el nombre de veïns de cada regió pot ser 3, 5 o 8. Amb l'objectiu de poder estudiar el comportament dels dos mètodes d'estimació en funció del nombre de casos, s'ha simulat una situació on el nombre de casos és petit, i una altra on el nombre de casos és gran. Així, s'ha generat el nombre de casos esperats a partir, o bé d'una uniforme de paràmetres 1 i 10 o d'una amb paràmetres 10 i 40, respectivament. A més a més, s'ha considerat una variable

explicativa, x_i, la qual ha estat simulada mitjançant una distribució normal amb una mitjana de 0 i una desviació típica de 0,5. Finalment, s'ha fixat la constant del model en 0,1 i el paràmetre de la variable explicativa, en 0,3. Per tant, el vector de paràmetres és $\beta = (0{,}1{,}0{,}3)^T$ i cada fila de la matriu de disseny dels efectes fixos serà $(1, x_i)^T$.

Els efectes aleatoris s'han generat sota una distribució normal multivariant amb mitjana 0, i una matriu de variàncies i covariàncies definida en funció del model seleccionat. Els components de la variància pels models han estat fixades a 1. Així pel model d'heterogeneïtat, $\sigma_H^2 = 1$, pel model CAR intrínsec, $\sigma_S^2 = 1$, i pel model CAR no intrínsec, $\sigma_H^2 = 1$ i $\sigma_S^2 = 1$.

Un cop definits tots els components del model, el nombre de casos observats s'han generat mitjançant una distribució de Poisson de mitjana $E \cdot \exp(X \cdot \beta + Z \cdot b)$.

En resum, les diferents situacions que s'obtenen estan en funció del nombre de regions i del nombre de casos esperats. Per cadascuna d'aquestes situacions s'han simulat 400 sets de dades i s'ha calculat per cada paràmetre estimat $\xi = (\beta, \sigma_H^2, \sigma_S^2)$,

l'esperança,

$$E(\hat{\xi}),$$

el biaix relatiu,

$$RB = \frac{E(\hat{\xi} - \xi)}{\xi},$$

l'error quadràtic mitjà,

$$MSE = E\left[(\hat{\xi} - \xi)^2\right],$$

la desviació estàndard de les estimacions obtingudes,

$$SD = \sqrt{var(\hat{\xi})},$$

i la mitjana dels errors estàndards estimats,

$$MES = E\left[S.E.(\hat{\xi})\right].$$

1.5.2. Resultats

Els resultats de la simulació són descrits separadament per cada model considerat, els quals s'exposen en les taules 1.3, 1.4 i 1.5. Així doncs, la taula 1.3 conté els resultats del CAR intrínsec, la taula 1.4, els del model d'heterogeneïtat, i la taula 1.5, els del model CAR no intrínsec. A més, també es representa gràficament al final del capítol el valor absolut del biaix relatiu i l'error quadràtic mitjà pels paràmetres estimats en les figures 1.1, 1.2 i 1.3.

1.5.2.1. Model intrínsec CAR

Tal com es pot observar en la taula 1.3, el paràmetre β_0 està sobreestimat en l'estimació PQL. Aquest biaix és superior en comparació amb el FB, però tot i així, atès que les estimacions amb PQL presenten menys variabilitat, l'error quadràtic mitjà que s'obté és semblant entre els dos procediments. En relació amb el biaix que es presenta en PQL, es pot observar que aquest decreix quan el nombre de casos esperats creix, però té una lleugera tendència a incrementar quan el nombre de regions creix (figura 1.1*a*).

El paràmetre β_1, en canvi, presenta un biaix menyspreable (figura 1.1*b*) i una estimació de l'error estàndard bastant exacta en els dos tipus d'estimació.

En relació amb la variància dels efectes aleatoris, l'estimació bayesiana està més esbiaixada quan el nombre de casos esperats és petit i quan el nombre de regions és 25 o 49 (figura 1.1*c*). En totes les situacions considerades, l'estimació per PQL presenta una variabilitat inferior que en FB, i aleshores l'error quadràtic mitjà resultant també és menor.

En general, pels dos procediments l'error quadràtic mitjà de tots els paràmetres decreix quan el nombre de regions i el nombre de casos esperats creix (figures 1.1*d*, 1.1*e*, 1.1*f*).

Taula 1.3
Estimació dels paràmetres del CAR intrínsec utilitzant l'aproximació PQL i FB en funció del nombre de regions i casos esperats

Valor real	Nre. regions	Nre. esperats	Esperança		Biaix relatiu x10^2		MSE x10^2		SD		MES	
			PQL	FB	PQL	FB	PQL	FB	PQL	FB	PQL	FB
β_0=0.1	25	Petit	0,131	0,098	30,78	–1,57	0,807	0,875	0,084	0,094	0,082	0,088
		Gran	0,112	0,097	11,85	–2,92	0,172	0,217	0,040	0,047	0,040	0,041
	49	Petit	0,136	0,103	36,55	3,29	0,543	0,495	0,064	0,070	0,063	0,069
		Gran	0,115	0,100	14,58	–0,45	0,114	0,118	0,031	0,034	0,030	0,032
	100	Petit	0,137	0,097	36,8	–2,68	0,330	0,228	0,044	0,048	0,043	0,047
		Gran	0,115	0,098	14,9	–2,04	0,065	0,051	0,021	0,023	0,021	0,021
β_1=0.3	25	Petit	0,296	0,302	–1,31	0,72	2,316	2,450	0,152	0,157	0,140	0,140
		Gran	0,304	0,307	1,50	2,40	1,304	1,360	0,114	0,117	0,113	0,117
	49	Petit	0,303	0,308	0,84	2,58	1,230	1,288	0,111	0,113	0,101	0,102
		Gran	0,299	0,302	–0,31	0,74	0,645	0,662	0,080	0,081	0,079	0,081
	100	Petit	0,298	0,302	–0,56	0,73	0,547	0,572	0,074	0,076	0,074	0,075
		Gran	0,300	0,303	0,13	0,99	0,313	0,323	0,056	0,057	0,056	0,057
σ_s=1	25	Petit	0,935	0,852	–6,48	–14,81	8,105	15,85	0,278	0,370	0,263	0,317
		Gran	0,978	1,000	–2,23	0,01	3,177	3,912	0,177	0,198	0,176	0,196
	49	Petit	0,946	0,915	–5,38	–8,52	4,195	8,363	0,198	0,277	0,192	0,243
		Gran	0,983	1,000	–1,69	–0,04	1,773	2,263	0,132	0,151	0,128	0,141
	100	Petit	0,962	0,963	–3,82	–3,67	1,773	2,779	0,128	0,163	0,134	0,163
		Gran	0,989	1,005	–1,13	0,49	0,700	0,952	0,083	0,098	0,089	0,098

MSE: error quadràtic mitjà; SD: desviació típica de les estimacions, i MES: mitjana dels errors estàndards estimats.

1.5.2.2. Model d'heterogeneïtat

A la taula 1.4 es pot observar que el paràmetre β_0 està sobreestimat en PQL i una mica subestimat en FB. El biaix que es presenta en PQL s'incrementa quan el nombre de regions creix, però és menor amb un nombre de casos esperats gran. En canvi, en FB el biaix decreix quan augmenten tant el nombre de regions com el nombre casos d'esperats (figura 1.2*a*). També s'ha detectat que l'error quadràtic mitjà és similar en les dues tècniques perquè l'estimació mitjançant PQL és més precisa.

En relació amb el paràmetre β_1, la seva estimació està una mica esbiaixada en les dues tècniques. Aquest biaix en FB decreix quan el nombre de regions i de casos esperats s'incrementa, i en canvi en PQL aquest només es redueix quan el nombre de regions s'incrementa (figura 1.2*b*). A més a més, l'error quadràtic mitjà és una mica més gran en FB per la variabilitat de les estimacions.

Pel que fa al paràmetre σ_H, l'estimació a partir de PQL presenta un biaix negatiu, el qual decreix quan el nombre de casos esperats és gran (figura 1.2*c*). L'error quadràtic mitjà és una mica menor en PQL perquè l'estimació d'aquest paràmetre presenta menys variabilitat.

En general, l'error quadràtic mitjà en ambdues tècniques i de tots els paràmetres és redueix quan el nombre de regions i el nombre de casos esperats s'incrementa (figures 1.2*d*, 1.1*e* i 1.1*f*).

Taula 1.4
Estimació dels paràmetres del model d'heterogeneïtat utilitzant l'aproximació PQL i FB en funció del nombre de regions i casos esperats

Valor real	Nre. regions	Nre. esperats	Esperança		Biaix relatiu x10^2		MSE x10^2		SD		MES	
			PQL	FB	PQL	FB	PQL	FB	PQL	FB	PQL	FB
β_0=0.1	25	Petit	0,168	0,077	68,14	–22,67	4,597	4,816	0,204	0,219	0,211	0,236
		Gran	0,113	0,086	12,76	–14,42	3,957	4,185	0,199	0,204	0,202	0,213
	49	Petit	0,182	0,085	81,77	–14,98	2,913	2,638	0,150	0,162	0,152	0,168
		Gran	0,121	0,093	20,58	–6,95	2,054	2,116	0,142	0,146	0,144	0,150
	100	Petit	0,188	0,091	88,28	–9,17	2,019	1,471	0,115	0,121	0,106	0,116
		Gran	0,128	0,101	28,23	0,47	1,117	1,143	0,104	0,107	0,101	0,104
β_1=0.3	25	Petit	0,317	0,325	5,80	8,29	7,537	8,723	0,274	0,295	0,258	0,280
		Gran	0,321	0,321	6,98	7,15	6,926	7,394	0,263	0,271	0,247	0,255
	49	Petit	0,305	0,316	1,63	5,34	3,892	4,446	0,197	0,211	0,185	0,199
		Gran	0,305	0,312	1,81	3,96	3,150	3,307	0,177	0,182	0,178	0,182
	100	Petit	0,293	0,308	–2,31	2,62	1,922	2,170	0,139	0,147	0,139	0,148
		Gran	0,300	0,305	–0,04	1,64	1,678	1,805	0,130	0,134	0,132	0,134
σ_H=1	25	Petit	0,946	1,024	–5,38	2,410	3,107	3,709	0,168	0,191	0,171	0,205
		Gran	0,981	1,020	–1,95	2,044	2,534	2,808	0,158	0,167	0,153	0,169
	49	Petit	0,944	1,015	–5,56	1,475	1,778	1,942	0,121	0,139	0,121	0,142
		Gran	0,978	1,011	–2,16	1,126	1,221	1,320	0,109	0,115	0,109	0,117
	100	Petit	0,945	1,010	–5,46	1,036	1,045	0,981	0,087	0,099	0,083	0,096
		Gran	0,977	1,004	–2,33	0,407	0,554	0,555	0,071	0,074	0,074	0,079

MSE: error quadràtic mitjà; SD: desviació típica de les estimacions, i MES: mitjana dels errors estàndards estimats.

1.5.2.3. Model CAR no intrínsec

L'estimació PQL del paràmetre β_0 en totes les situacions considerades presenta un biaix positiu que disminueix quan es treballa amb un nombre de casos esperats gran (figura 1.3*a*). En canvi, les estimacions amb FB presenten poc biaix excepte quan el nombre de regions és 49. Pel que fa a l'error quadràtic mitjà, es pot observar que aquest és semblant en els dos procediments i que aquest es redueix amb el nombre de regions i de casos esperats (figura 1.3*e*).

L'estimació de β_1 es presenta, tant en PQL com en FB, poc esbiaixada (figura 1.3*b*). El seu error quadràtic mitjà és més gran en FB, però en les dues tècniques disminueix quan el nombre de regions i casos esperats creix (figura 1.3*f*).

Mentre que el procediment PQL clarament infraestima σ_S quan el nombre de regions és 25, en FB s'observa aquest fet quan el nombre de regions és 100. També es pot apreciar que l'error quadràtic mitjà en FB és més gran que en PQL menys en la situació de 25 regions i pocs casos esperats. Addicionalment, el comportament del decreixement d'aquest error és diferent en les dues tècniques, i així, mentre que en FB té lloc amb el nombre de casos esperats, en PQL té lloc amb el nombre de regions (figura 1.3*g*).

El paràmetre σ_H és infraestimat pels dos procediments. A més, aquest biaix en PQL decreix amb el nombre de regions i de casos esperats, i en canvi en FB, només amb el nombre de regions (figura 1.3*d*). També es pot apreciar que l'error quadràtic mitjà en PQL és més petit que en FB i que té una tendència a reduir-se en PQL. Aquest fet s'observa tant quan s'incrementa el nombre de regions com el nombre de casos esperats, en canvi al FB la reducció tan sols es presenta quan s'incrementa amb el nombre de casos esperats (figura 1.3*h*).

Finalment, en general per als tres models, es pot apreciar que la mitjana de l'estimació de l'error estàndard és similar a la desviació típica de les estimacions, és a dir, que l'estimació de l'error estàndard es realitza correctament en els dos procediments. No obstant això, cal remarcar el comportament de l'error estàndard dels components de la variància en el CAR no intrínsec, en el qual sí que hi ha una lleugera diferència entre la mitjana de l'estimació de l'error estàndard i la desviació típica en totes les combinacions excepte en PQL amb 100 regions.

Per finalitzar, en relació amb la convergència dels procediments, es pot apreciar un percentatge de convergència pròxim al 100%, tant en el model CAR intrínsec com en el d'heterogeneïtat i, en canvi, en el CAR no intrínsec aquest percentatge es troba entre el 76,75% i el 83,75% en PQL, i entre el 75,50% i el 85,50% en FB.

Taula 1.5
Estimació dels paràmetres del CAR no intrínsec utilitzant l'aproximació PQL i FB en funció del nombre de regions i casos esperats

Valor real	Nre. regions	Nre. esperats	Esperança		Biaix relatiu x10^2		MSE x10^2		SD		MES	
			PQL	FB	PQL	FB	PQL	FB	PQL	FB	PQL	FB
β_0=0.1	25	Petit	0,193	0,094	93,23	–6,45	5,311	5,961	0,211	0,244	0,203	0,214
		Gran	0,132	0,098	32,12	–1,71	3,858	4,628	0,194	0,215	0,192	0,187
	49	Petit	0,191	0,053	91,22	–47,11	2,996	2,927	0,147	0,165	0,148	0,163
		Gran	0,129	0,076	28,56	–23,53	2,006	2,253	0,139	0,149	0,138	0,131
	100	Petit	0,204	0,100	104,1	0,04	2,151	1,318	0,104	0,115	0,104	0,108
		Gran	0,134	0,096	34,12	–4,41	1,107	1,086	0,100	0,104	0,098	0,095
β_1=0.3	25	Petit	0,274	0,309	–8,74	2,85	7,358	8,796	0,270	0,297	0,276	0,311
		Gran	0,296	0,302	–1,20	0,80	7,197	8,837	0,269	0,298	0,270	0,288
	49	Petit	0,286	0,310	–4,73	3,21	4,124	4,582	0,203	0,214	0,200	0,224
		Gran	0,302	0,302	0,54	0,57	3,984	3,825	0,200	0,196	0,193	0,204
	100	Petit	0,298	0,311	–0,57	3,58	2,221	2,760	0,149	0,166	0,148	0,163
		Gran	0,300	0,308	–0,05	2,69	2,074	2,431	0,144	0,156	0,142	0,149
σ_s=1	25	Petit	0,752	1,003	–24,760	0,253	66,150	47,400	0,776	0,690	0,744	0,855
		Gran	0,826	1,006	–17,440	0,617	62,170	84,590	0,770	0,921	0,809	0,618
	49	Petit	0,944	0,960	–5,636	–3,995	35,390	66,380	0,593	0,815	0,716	0,690
		Gran	0,988	0,931	–1,157	–6,920	33,060	94,950	0,576	0,974	0,674	0,441
	100	Petit	0,951	0,854	–4,898	–14,640	22,210	73,190	0,469	0,844	0,488	0,533
		Gran	0,974	0,800	–2,591	–20,040	20,600	87,630	0,454	0,916	0,464	0,346
σ_H=1	25	Petit	0,885	0,870	–11,480	–12,970	8,871	10,470	0,275	0,297	0,322	0,349
		Gran	0,921	0,876	–7,895	–12,360	7,981	14,290	0,272	0,358	0,299	0,263
	49	Petit	0,896	0,906	–10,420	–9,423	4,921	11,180	0,196	0,321	0,219	0,269
		Gran	0,931	0,898	–6,923	–10,190	3,441	13,690	0,172	0,356	0,196	0,180
	100	Petit	0,908	0,937	–9,170	–6,299	2,953	8,676	0,146	0,288	0,139	0,192
		Gran	0,947	0,945	–5,292	–5,495	2,003	10,730	0,131	0,324	0,125	0,125

MSE: error quadràtic mitjà; SD: desviació típica de les estimacions, i MES: mitjana dels errors estàndards estimats.

1.6. Discussió i conclusions

L'objectiu d'aquest capítol ha estat comparar la quasiversemblança penalitzada *versus* la *fully bayesian* obtinguda a partir del mostratge de Gibbs, quan s'utilitzen per a l'estimació dels models lineals generalitzats mixtos en el cas del modelatge de dades agregades per regions geogràfiques. El comportament de la quasiversemblança penalitzada va ser estudiada per Leroux *et al.* (1999), el qual també va utilitzar diferents nombres de casos esperats però tenint en compte una altra parametrització de la matriu de variàncies i covariàncies dels efectes aleatoris. En aquest treball, s'ha observat un comportament de la quasiversemblança penalitzada semblant a la que va obtenir Leroux *et al.* (1999); és a dir, l'estimació de la coordenada en l'origen també va presentar un biaix positiu que es reduïa quan el nombre de casos esperats era gran. Addicionalment,

s'ha obtingut que aquest biaix té una lleugera tendència a incrementar-se amb el nombre de regions en els models CAR intrínsec i heterogeneïtat.

El coeficient de la variable explicativa ha estat estimat correctament i no presenta els problemes de la coordenada en l'origen. L'estimació mitjançant la quasiversemblança penalitzada de la variància dels efectes aleatoris ha presentat un biaix negatiu com en Leroux *et al.* (1999). Aquest biaix és més pronunciat en el model CAR no intrínsec quan el nombre de regions és petit. A més d'això, la reducció del biaix és més gran quan augmenta el nombre de casos esperats que quan augmenta el nombre de regions. La precisió de les estimacions decreix quan el nombre de casos esperats i el nombre de regions disminueix.

En l'anàlisi de les dades d'incidència de diabetis de tipus I a Catalunya, les estimacions obtingudes amb els dos procediments no presenten unes diferències gaire pronunciades, però quan les dues tècniques són comparades mitjançant l'estudi de simulació s'ha apreciat un comportament diferent en termes de biaix i de precisió.

Quan les dues tècniques són comparades amb l'estudi de simulació, s'ha observat que l'ordenada en l'origen està més esbiaixada amb la quasiversemblança penalitzada. No obstant això, la variabilitat d'aquestes estimacions és menor i això provoca que l'error quadràtic mitjà entre els dos procediments sigui similar.

En relació amb el coeficient de la variable explicativa, l'error quadràtic mitjà en *fully bayesisan* és més gran per la variabilitat de les observacions.

Quant al component de la variància espacial, la quasiversemblança penalitzada proporciona unes estimacions més consistents (menys error quadràtic mitjà) que el *fully bayesian*, excepte en el CAR no intrínsec quan el nombre de casos esperats és petit i el nombre de regions és 25.

Pel que fa al component de la variància d'heterogeneïtat, sempre presenta un error quadratic mitjà superior en *fully bayesian*, i les diferències més grans es localitzen en el model CAR no intrínsec.

En general, amb les dues tècniques, les estimacions dels paràmetres milloren quan el nombre de casos esperats i el nombre de regions s'incrementen, excepte els components de la variància en el CAR no intrínsec estimades mitjançant *fully bayesian.*

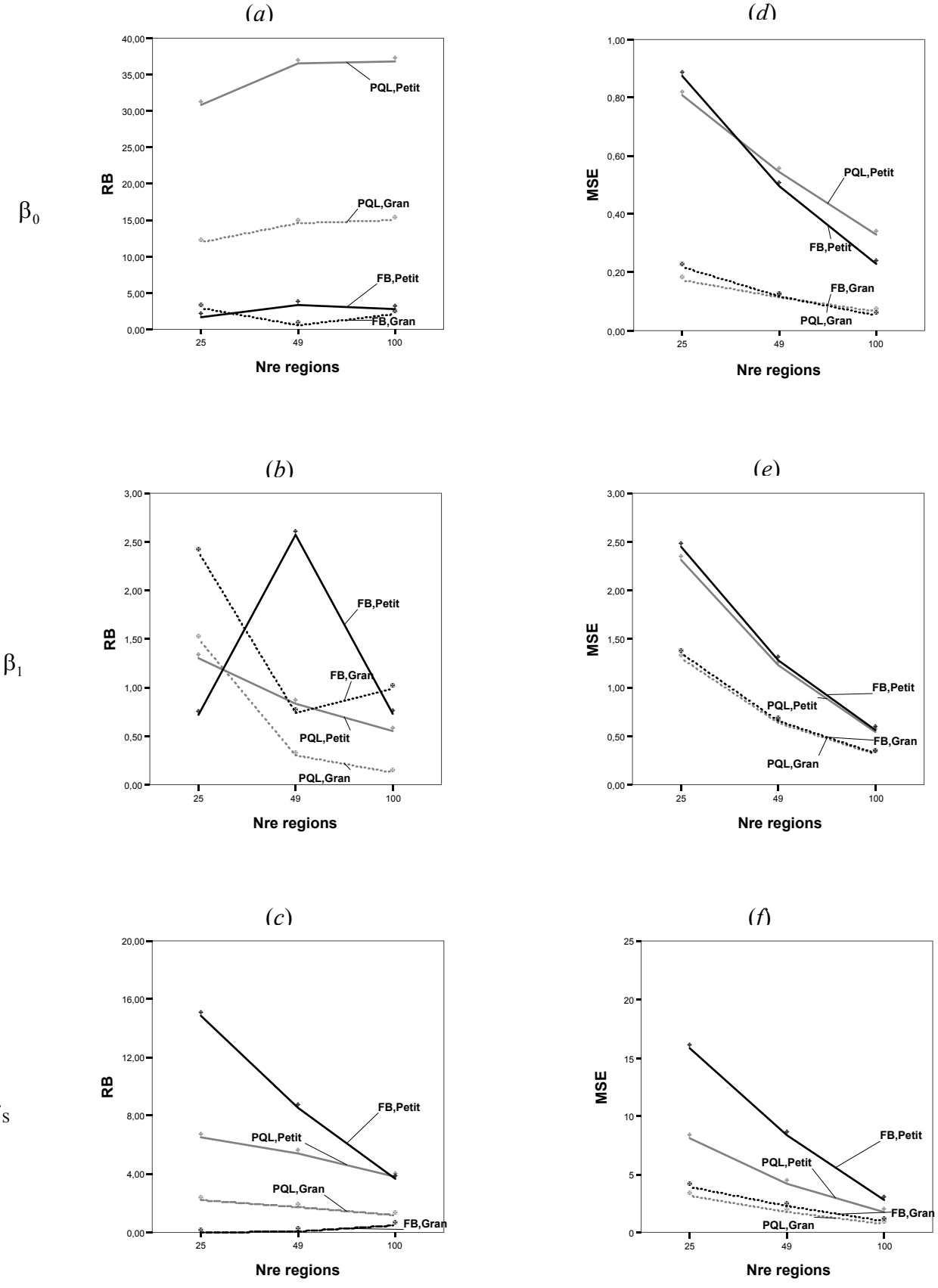
Un fet remarcable és que l'error estàndard dels components de la variància estimades mitjançant la inversa de la matriu d'informació de Fisher i la variància de la distribució marginal posterior dels paràmetres són correctes, excepte en alguns casos en el model CAR no intrínsec.
Finalment, a partir d'aquesta simulació s'ha observat que el model amb més problemes de convergència és el CAR no intrínsec, amb el qual s'obté un percentatge de convergència per sota del 85,5% en *fully bayesian* i del 83, 75% en quasiversemblança penalitzada. A més a més, al llarg de l'estudi de simulació s'ha observat que les estimacions mitjançant la quasiversemblança penalitzada són una mica més esbiaixades però més precises que en *fully bayesian*. Tot i així les diferències de biaix i de variabilitat que es presenten no són gaire pronunciades.

Un dels avantatges de la quasiversemblança penalitzada és la facilitat d'implementació i el baix cost computacional en relació amb el *fully bayesian* utilitzant el mostratge de Gibbs.

En conclusió, l'investigador ha de considerar que quan modela un dels tres models utilitzant l'aproximació quasiversemblança penalitzada, les estimacions que obtindrà de l'ordenada en l'origen seran menys exactes (particularment amb un nombre de casos esperats petit), però les estimacions seran més consistents que en l'aproximació *fully bayesian*. Per tant, la selecció de la tècnica d'estimació dependrà de l'interès de l'investigador a obtenir una estimació més exacta, precisa o consistent. Una possible recomanació podria ser utilitzar l'estimació per FB si es té un nombre petit de casos esperats i de regions. Aquest raonament també es pot basar en el fet que amb pocs casos esperats i poques regions la versemblança és molt inestable, i funciona millor l'estimació FB perquè estabilitza la versemblança mitjançant les funcions de densitat *a priori* dels paràmetres (Bernardinelli *et at*., 1995*b*). En canvi, amb un nombre gran de casos esperats i de regions el procediment PQL seria preferible en termes de consistència.

Figura 1.1

Representació del biaix relatiu (RB) i de l'error quadràtic mig (MSE) dels paràmetres del model CAR intrínsec en funció del nombre de casos esperats, regions i mètode d'estimació



Casos esperats petits ····▲···· Casos esperats grans. El color ■ correspon a PQL i el color ■ correspon a FB.

Figura 1.2

Representació del biaix relatiu (RB) i de l'error quadràtic mitjà (MSE) dels paràmetres del model d'heterogeneïtat en funció del nombre de casos esperats, regions i mètode d'estimació

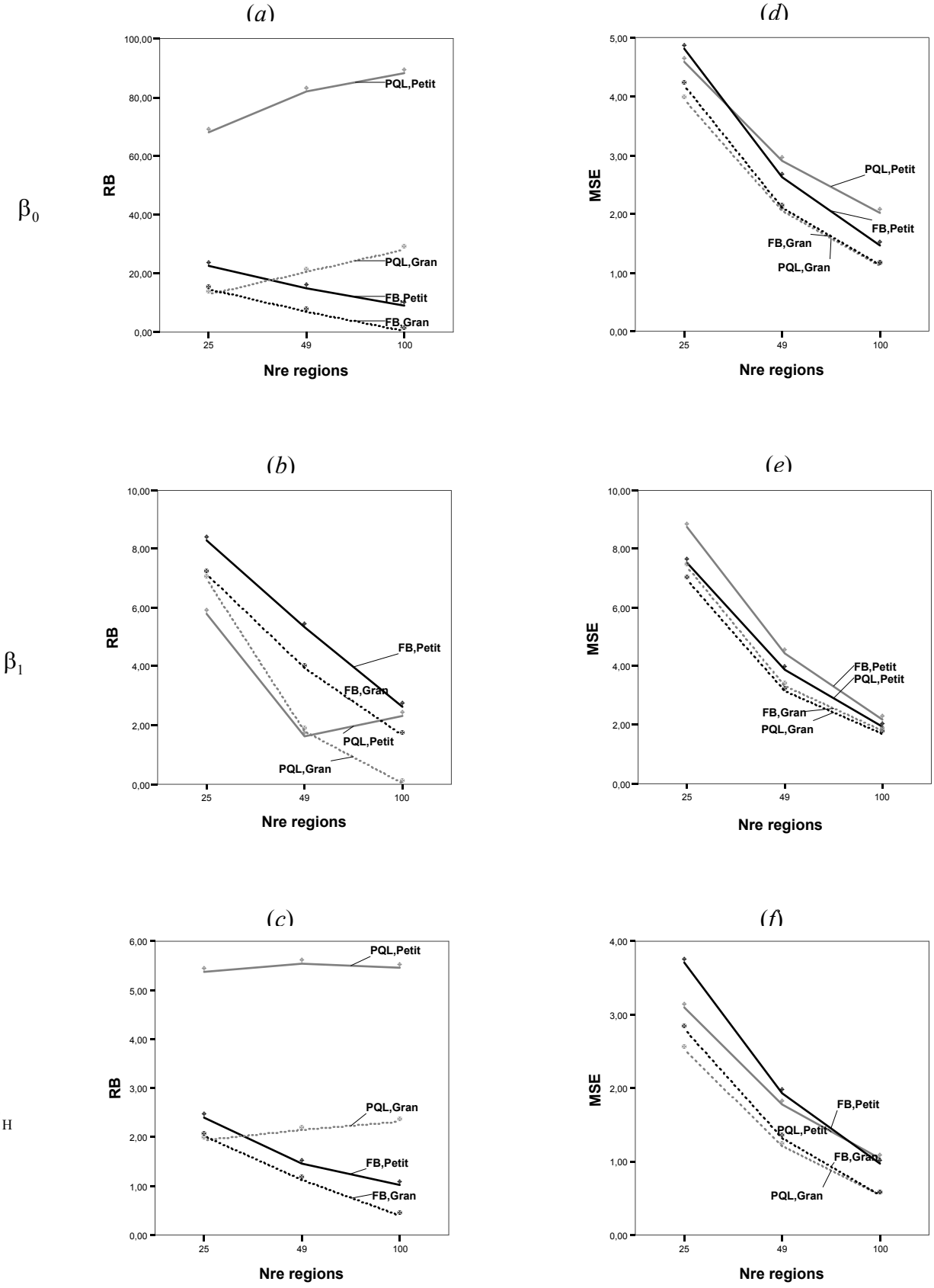
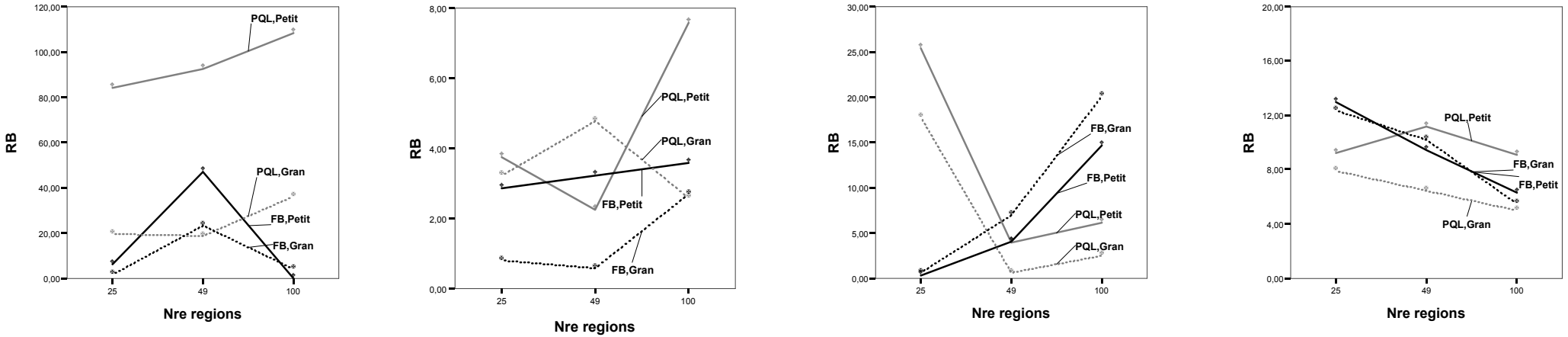


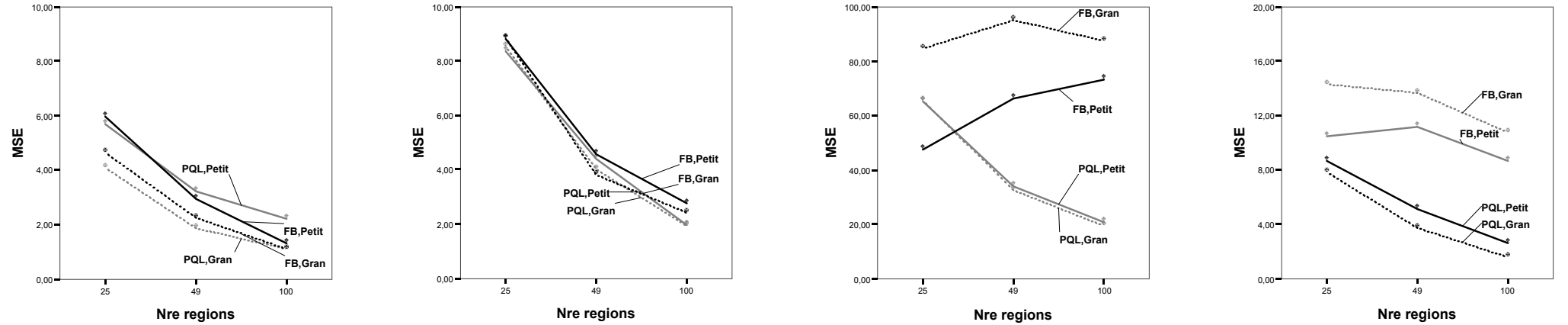
Figura 1.3

Representació del biaix relatiu (RB) i de l'error quadràtic mitjà (MSE) dels paràmetres del model CAR no intrínsec en funció del nombre de casos esperats, regions i mètode d'estimació

β_0 (*a*) β_1 (*b*) σ_S (*c*) σ_H (*d*)



(*e*) (*f*) (*g*) (*h*)



Casos esperats petits Casos esperats grans El color ■ correspon a PQL i el color ■ correspon a FB

CAPÍTOL II: MILLORA DE LA CONVERGÈNCIA EN L'ESTIMACIÓ DELS MODELS CONDICIONALS AUTOREGRESSIUS

2.1. Introducció

En el modelatge de la variabilitat del risc de patir o morir d'una determinada malaltia al llarg d'una àrea d'estudi, com s'ha vist en la introducció, una de les possibles parametritzacions de la matriu de variàncies i covariàncies dels efectes aleatoris és la proposada per Leroux *et al.* (1999). Aquesta parametrització inclou dos components: l'un per controlar la sobredispersió, σ, i l'altre per tenir en compte el pes de la dependència espaial, λ, i es defineix com:

$$\Sigma\left(\sigma^2,\lambda\right)=\sigma^2 R^{-}=\sigma^2\left[\lambda Q_N+(1-\lambda)I_N\right]^{-},$$

on I_N és una matriu identitat i Q_N és la matriu quadrada determinada per l'estructura de veïnatge de les àrees; N és el nombre de regions d'estudi. Sota aquesta parametrització de la matriu, el valor mínim del paràmetre λ és 0 i el màxim és 1. Pel que fa al paràmetre de la sobredispersió, σ, tan sols pot presentar valors positius.

Com ja s'ha mencionat, hi ha diferents mètodes d'estimació d'aquests models, els quals poden estar basats en la versemblança o la quasiversemblança. Sovint, amb aquests procediments, es presenten problemes en l'estimació. Pot produir-se o que l'algoritme de maximització utilitzat no assoleixi la convergència dels paràmetres amb el nivell de tolerància fixat, o que tot i que s'aconsegueixi convergència, l'estimació dels paràmetres no estigui dins del seu domini. Això provoca, en el cas de no-convergència, que no es puguin estimar els paràmetres del model o, en el cas que els paràmetres estiguin fora del seu domini, que les estimacions no siguin de màxima versemblança, la qual cosa condueix a obtenir una matriu de variàncies i covariàncies dels efectes aleatoris que serà incorrecta i, en conseqüència, també ho seran els errors estàndards de les estimacions del paràmetres. Per solucionar aquests problemes en l'estimació, en aquest capítol hom proposa aplicar una transformació a cadascun dels paràmetres que defineixen la matriu de variàncies i covariàncies, λ i σ. Mitjançant les transformacions proposades es garanteix que les estimacions d'aquests paràmetres estiguin dins del seu domini. Amb aquest mateix objectiu, Pinehiro i Bates (1996) també van proposar unes parametritzacions de la matriu de variàncies i covariàncies en el context dels models lineals i no lineals mixtos.

En aquest capítol s'estudia mitjançant una simulació el comportament de l'algoritme que inclou les transformacions, tant en termes de convergència com pel que fa a l'estimació, i s'il·lustra aquesta aproximació a través de l'aplicació a les dades d'incidència del càncer de llavis a Escòcia (Clayton i Kaldor, 1987) i a les dades d'incidència de la diabetis de tipus I de Catalunya.

2.2. Transformació dels paràmetres

En el nostre context, s'assumeix que el nombre de casos observats en cada àrea $Y=\left(Y_1,\ldots,Y_N\right)$ condicionat als efectes aleatoris $b=\left(b_1,\ldots b_N\right)$ es distribueix sota una distribució de Poisson de mitjana $\mu^b=\left(\mu_1^b,\ldots,\mu_N^b\right)$. La funció que enllaça aquestes mitjanes amb el predictor lineal és la logarítmica:

$$\log\left(\mu^b\right)=\log(E)+Xa+Zb,$$

on recordem que X i Z són les matrius de disseny dels efectes fixos i aleatoris, respectivament, a és el vector de coeficients dels efectes fixos, i $E=\left(E_1,\ldots,E_N\right)$ és el vector del nombre de casos esperats calculats a partir de les taxes d'una població

estàndard o de l'àrea en què estem treballant. Els efectes aleatoris *b* és distribueixen sota una distribució normal multivariant amb un vector de mitjanes de 0 i una matriu de variàncies i covariàncies

$$\Sigma(\delta) = \sigma^2 R^{-} = \sigma^2 \left[\lambda Q_N + (1-\lambda) I_N\right]^{-},$$

on $\delta = (\sigma, \lambda)$.

Si el model es maximitza utilitzant el mètode de la quasiversemblança penalitzada (Breslow i Clayton 1993), els efectes fixos i aleatoris s'estimen a partir de solucionar iterativament,

$$\hat{\beta} = \left(X^t V^{-1} X\right)^{-1} X^t V^{-1} Y^* \qquad \text{i} \qquad \hat{b} = \Sigma(\delta) Z^t V^{-1} \left(Y^* - X\hat{\beta}\right),$$

on

$$V = W^{-1} + Z\Sigma(\delta)Z^t;$$

W és una matriu diagonal amb els termes igual a μ_i^b i *Y** la pseudovariable definida mitjançant

$$Xa + Zb + \left(\frac{Y - \mu^b}{\mu^b}\right).$$

L'estimació de δ s'obté de maximitzar la màxima versemblança restringida dels efectes aleatoris mitjançant l'algoritme de Fisher-Scoring. A partir d'aquest algoritme, en cada pas els components de la matriu de variàncies i covariàncies s'actualitzen mitjançant l'equació:

$$\hat{\delta}^{(r)} = \hat{\delta}^{(r-1)} + I^{-1}\left(\hat{\delta}^{(r-1)}\right) \cdot U\left(\hat{\delta}^{(r-1)}\right),$$

on *r* és el nombre d'iteració, $U(\hat{\delta})$ és el vector *score* i $I(\hat{\delta})$ és la matriu d'informació de Fisher, els quals es defineixen com:

$$U(\hat{\delta}) = \frac{\partial L}{\partial \hat{\delta}} = -\frac{1}{2}\left\{\left(Y^* - X\hat{\beta}\right)^t V^{-1} \frac{\partial V}{\partial \hat{\delta}} V^{-1}\left(Y^* - X\hat{\beta}\right) - \text{tr}\left(P\frac{\partial V}{\partial \hat{\delta}}\right)\right\}$$

$$I(\hat{\delta}) = -E\left[\frac{\partial^2 L}{\partial \hat{\delta}_i \partial \hat{\delta}_j}\right] = \frac{1}{2}\text{tr}\left(P\frac{\partial V}{\partial \hat{\delta}_i} P \frac{\partial V}{\partial \hat{\delta}_j}\right),$$

on $P = V^{-1} - V^{-1}X\left(X^t V^{-1} X\right)^{-1} X^t V^{-1}$.

Per assegurar que l'estimació d'aquests paràmetres estigui dins el seu domini, es proposa aplicar el logaritme a la desviació estàndard dels efectes aleatoris, $\tau_\sigma = \log(\sigma)$ i la funció lògit al pes de la dependència espacial $\tau_\lambda = \log\left(\frac{\lambda}{1-\lambda}\right)$. Per tant, σ només podrà prendre valors positius, i λ, només valors entre 0 i 1. Considerant aquestes transformacions, la matriu de variàncies i covariàncies s'especificarà com:

$$\Sigma(\tau_\sigma, \tau_\lambda) = \exp(2 \cdot \tau_\sigma) \cdot R^{-1},$$

on

$$R = \frac{\exp(\tau_\lambda)}{1 + \exp(\tau_\lambda)} \cdot Q_N + \frac{1}{1 + \exp(\tau_\lambda)} I_N.$$

Així doncs, els paràmetres a estimar seran τ_σ i τ_λ, i els elements de l'equació iterativa de l'algoritme de Fisher-Scoring per a aquests dos components seran:

$$U(\hat{\tau}_\sigma)=\frac{\partial L}{\partial \hat{\tau}_\sigma}=-\frac{1}{2}\left\{\left(Y^*-X\hat{\beta}\right)^t V^{-1}\frac{\partial V}{\partial \hat{\tau}_\sigma}V^{-1}\left(Y^*-X\hat{\beta}\right)-\mathrm{tr}\left(P\frac{\partial V}{\partial \hat{\tau}_\sigma}\right)\right\}$$

$$U(\hat{\tau}_\lambda)=\frac{\partial L}{\partial \hat{\tau}_\lambda}=-\frac{1}{2}\left\{\left(Y^*-X\hat{\beta}\right)^t V^{-1}\frac{\partial V}{\partial \hat{\tau}_\lambda}V^{-1}\left(Y^*-X\hat{\beta}\right)-\mathrm{tr}\left(P\frac{\partial V}{\partial \hat{\tau}_\lambda}\right)\right\}$$

$$I(\hat{\tau}_\sigma,\hat{\tau}_\sigma)=\frac{1}{2}\mathrm{tr}\left(P\frac{\partial V}{\partial \hat{\tau}_\sigma}P\frac{\partial V}{\partial \hat{\tau}_\sigma}\right);$$

$$I(\hat{\tau}_\sigma,\hat{\tau}_\lambda)=\frac{1}{2}\mathrm{tr}\left(P\frac{\partial V}{\partial \hat{\tau}_\sigma}P\frac{\partial V}{\partial \hat{\tau}_\lambda}\right);$$

$$I(\hat{\tau}_\lambda,\hat{\tau}_\lambda)=\frac{1}{2}\mathrm{tr}\left(P\frac{\partial V}{\partial \hat{\tau}_\lambda}P\frac{\partial V}{\partial \hat{\tau}_\lambda}\right),$$

on

$$\frac{\partial V}{\partial \hat{\tau}_\sigma}=\frac{\partial\left(W^{-}+Z'\Sigma(\hat{\delta})Z\right)}{\partial \hat{\tau}_\sigma}=-2\exp(2\cdot\hat{\tau}_\sigma)\cdot R_N^{-}$$

$$\frac{\partial V}{\partial \tau_\lambda}=\frac{\partial\left(W^{-1}+Z'DZ\right)}{\partial \tau_\lambda}=Z'\frac{\partial D}{\partial \tau_\lambda}Z=Z'\frac{\partial\left(\exp(2\cdot\tau_\sigma)R^{-1}\right)}{\partial \tau_\lambda}Z=Z'\cdot\exp(2\cdot\tau_\sigma)\frac{\partial\left(R^{-1}\right)}{\partial \tau_\lambda}Z=$$

$$=-\exp(2\cdot\tau_\sigma)\cdot Z'R^{-1}\frac{\partial(R)}{\partial \tau_\lambda}R^{-1}Z=-\exp(2\cdot\tau_\sigma)Z'R^{-1}\left(\frac{\exp(\tau_\lambda)}{[1+\exp(\tau_\lambda)]^2}\cdot Q-\frac{\exp(\tau_\lambda)}{[1+\exp(\tau_\lambda)]^2}I\right)R^{-1}Z.$$

Un cop s'ha arribat a la convergència, l'estimació de σ i λ s'obtindrà com

$$\hat{\sigma}=\exp(\hat{\tau}_\sigma)\quad\text{i}\quad\hat{\lambda}=\frac{\exp(\hat{\tau}_\lambda)}{1+\exp(\hat{\tau}_\lambda)}.$$

Els errors estàndard de τ_σ i τ_λ es poden aproximar mitjançant la inversa de la matriu d'informació de Fisher, i a partir d'aquests es poden obtenir els dels paràmetres σ i λ utilitzant el mètode delta. Aplicant aquest mètode, els errors estàndards seran:

$$\mathrm{Var}(\hat{\sigma})\approx\left(\frac{\partial\hat{\sigma}}{\partial\hat{\tau}_\sigma}\right)^2\mathrm{Var}(\hat{\tau}_\sigma)=\left(\frac{\partial(\exp(\hat{\tau}_\sigma))}{\partial\hat{\tau}_\sigma}\right)^2\mathrm{Var}(\hat{\tau}_\sigma)=\exp(2\cdot\hat{\tau}_\sigma)\mathrm{Var}(\hat{\tau}_\sigma)$$

$$\mathrm{Var}(\hat{\lambda})\approx\left(\frac{\partial\hat{\lambda}}{\partial\hat{\tau}_\lambda}\right)^2\mathrm{Var}(\hat{\tau}_\lambda)=\left(\frac{\partial\left(\frac{\exp(\hat{\tau}_\lambda)}{1+\exp(\hat{\tau}_\lambda)}\right)}{\partial\hat{\tau}_\sigma}\right)^2\mathrm{Var}(\hat{\tau}_\sigma)=\frac{\exp(2\cdot\hat{\tau}_\lambda)}{(1+\exp(\hat{\tau}_\lambda))^4}\mathrm{Var}(\hat{\tau}_\sigma).$$

2.3. Estudi de simulació

Amb l'objectiu d'examinar el comportament i el percentatge de convergència de les transformacions proposades, s'ha dut a terme l'estudi de simulació següent.
S'han considerat dues àrees dividides en 49 i 225 regions, respectivament, les quals s'han creat mitjançant dues retícules de dimensions 9×9 i 15×15. El nombre de

casos observats ha estat generat a partir d'una distribució de Poisson de mitjana igual a $E \cdot \exp(X \cdot \beta + Z \cdot b)$. S'ha tingut en compte un nombre de casos esperats petit i gran, els quals han estat simulats mitjançant una uniforme de paràmetres (1,10) i (10,40), respectivament. A més, el model inclou una ordenada en l'origen i una variable explicativa amb uns coeficients de regressió fixats a $\beta = (0{,}1, 0{,}3)$. La covariable s'ha generat a partir d'una distribució normal de mitjana 0 i desviació típica de 0,5. Finalment, els efectes aleatoris, *b*, s'han obtingut a partir d'una distribució normal de mitjana 0 i la matriu de variàncies i covariàncies definida en la secció 2. Els valors considerats per λ han estat 0,25, 0,5 i 0,75, i per a σ són 0,25 i 1.

El nombre de mostres simulades per cada combinació de paràmetres, nombre de regions i nombre de casos esperats, ha estat de 400, i el model per cada mostra de dades ha estat fitat tant considerant les transformacions dels paràmetres com sense tenir-les en compte. En relació amb el procediment que no inclou les transformacions dels paràmetres, els models en els quals les estimacions obtingudes han estat fora del seu domini també han estat considerades casos de no-convergència.

Amb l'objectiu de comprovar si les estimacions dels paràmetres són correctes quan emprem les transformacions proposades, hom ha calculat per a cada combinació la mitjana i la desviació estàndard de les estimacions obtingudes. A més a més, també s'ha mesurat el biaix ($BIAIX_{SMR}$) i l'error quadràtic mitjà (MSE_{SMR}) per a la predicció del logaritme de la raó estandarditzada de mortalitat. Les expressions per obtenir aquestes mesures són:

$$BIAIX = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{N}\sum_{i=1}^{N}\log(SMR_{ij}) - \log(\hat{SMR}_{ij})\right] = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{SMR_{ij}}{\hat{SMR}_{ij}}\right)\right]$$

$$MSE = \frac{1}{m}\sum_{j=1}^{m}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{SMR_{ij}}{\hat{SMR}_{ij}}\right)\right)^2 + Var\left(\log\left(\frac{SMR_{ij}}{\hat{SMR}_{ij}}\right)\right)\right],$$

on *m* és el nombre de models que han convergit i *N* és el nombre de regions.

La simulació proposada i l'estimació del model incorporant les transformacions i sense incorporar-les s'han fet mitjançant funcions en llenguatge S (apèndix 2).

En les taules 2.1 i 2.2, es mostren els percentatges de convergència utilitzant ambdós procediments quan $\sigma = 0{,}25$ i $\sigma = 1$, respectivament. En ambdues taules es pot observar que el percentatge de convergència del procediment que incorpora les transformacions és sempre superior al que no les inclou. A més, les diferències entre els dos procediments són més extremes quan el nombre d'àrees és 49 i el nombre de casos esperats és petit.

En el procediment sense transformacions i quan el valor de σ és 0,25, s'obtenen uns percentatges de convergència pitjors que quan σ es fixa a 1. Per a les combinacions on el valor real de σ és 0,25, es pot apreciar que hi ha poca convergència excepte en el cas de 225 àrees i amb el nombre de casos esperats gran. En aquest cas el percentatge de convergència sense considerar les transformacions varia entre 12,5% i 96,5% i, en canvi, amb transformacions, entre 82% i 100%. També es pot observar

que amb un nombre d'àrees de 49 i pocs casos esperats el percentatge no supera el 21%.

Quan el valor real del paràmetre σ val 1, les diferències majors entre els dos procediments es poden trobar quan el nombre d'àrees és 49. En aquest cas, els percentatges de convergència del procediment sense transformacions són menors del 90% i, en canvi, amb la transformació el percentatge més petit observat és del 99,75%. Les diferències més grans entre els dos procediments s'observen en la combinació de 49 àrees, pocs casos esperats i λ igual a 0,75, situació en la qual pel procediment sense transformacions el percentatge de convergència és del 64% mentre que emprant les transformacions el percentatge puja fins al 99,75%.

Taula 2.1

Nombre de simulacions que han convergit utilitzant els procediments amb transformacions i sense, fixant σ a 0,25 i 1 i λ a 0,25, 0,5 i 0,75, amb un nombre de regions de 49 i 225 i casos esperats petits o grans

Real σ	Real λ	Nre. de regions	Casos Esperats	Sense transformació	Amb transformació
0,25	0,25	49	Petit	81 (20,25)	328 (82,00)
			Gran	235 (58,75)	380 (95,00)
		225	Petit	215 (53,75)	340 (85,00)
			Gran	386 (96,50)	400 (100,00)
	0,5	49	Petit	54 (13,50)	324 (81,00)
			Gran	204 (51,00)	336 (84,00)
		225	Petit	200 (50,00)	364 (91,00)
			Gran	384 (96,00)	400 (100,00)
	0,75	49	Petit	50 (12,50)	326 (81,50)
			Gran	148 (37,00)	357 (89,25)
		225	Petit	138 (34,50)	318 (79,50)
			Gran	362 (90,50)	400 (100)

*El percentatge de convergència està entre parèntesis.

Taula 2.2

Nombre de simulacions que han convergit utilitzant els procediments amb transformacions i sense, fixant σ a 1 i λ a 0,25, 0,5 i 0,75, amb un nombre de regions de 49 i 225 i casos esperats petits o grans*

Real σ	Real λ	Nre. de regions	Casos esperats	Sense transformació	Amb transformació
1	0,25	49	Petit	323 (80,75)	399 (99,75)
			Gran	340 (85,00)	400 (100,00)
		225	Petit	365 (91,25)	400 (100,00)
			Gran	356 (89,00)	400 (100,00)
	0,5	49	Petit	291 (72,75)	400 (100,00)
			Gran	336 (84,00)	400 (100,00)
		225	Petit	394 (98,50)	400 (100,00)
			Gran	398 (99,50)	400 (100,00)
	0,75	49	Petit	256 (64,00)	399 (99,75)
			Gran	282 (70,50)	400 (100,00)
		225	Petit	373 (93,25)	400 (100,00)
			Gran	381 (95,25)	400 (100,00)

*El percentatge de convergència està entre parèntesis.

En les taules 2.3 i 2.4, es presenten la mitjana, l'error estàndard de les estimacions i el biaix i l'error quadràtic mitjà de la predicció del log(SMR) quan $\sigma = 0{,}25$ i $\sigma = 1$, respectivament. Malgrat que en algunes combinacions hi ha un lleuger biaix, en general els paràmetres estan estimats correctament. Cal esmentar que en els casos on l'estimació es duu a terme mitjançant la tècnica de la quasiversemblança penalitzada, aquest petit biaix també es presenta (Leroux *et al.*, 1999). En conseqüència, aquest biaix no és atribuïble a la transformació dels paràmetres sinó al mètode d'estimació.

Taula 2.3

Valor mitjà de les estimacions de σ i λ, biaix ($BIAIX_{SMR}$) i error quadràtic mitjà (MSE_{SMR}) de les prediccions de l'SMR utilitzant el procediment amb les transformacions quan s'ha fixat σ a 0,25 i λ a 0,25, 0,5 i 0,75, amb un nombre de regions de 49 i 225 i casos esperats petits i grans*

				Valor mitjà			
Real σ	Real λ	Nre. de regions	Casos esperats	σ	λ	$BIAIX_{SMR}$	MSE_{SMR}
0,25	0,25	49	Petit	0,303 (0,133)	0,446 (0,457)	0,004 (0,057)	0,035 (0,009)
			Gran	0,254 (0,084)	0,345 (0,355)	0,004 (0,027)	0,018 (0,004)
		225	Petit	0,245 (0,080)	0,367 (0,351)	0,008 (0,027)	0,026 (0,003)
			Gran	0,242 (0,048)	0,273 (0,187)	0,007 (0,013)	0,016 (0,002)
	0,5	49	Petit	0,289 (0,135)	0,537 (0,468)	-0,0004 (0,061)	0,029 (0,010)
			Gran	0,240 (0,091)	0,461 (0,392)	0,003 (0,026)	0,015 (0,004)
		225	Petit	0,226 (0,080)	0,443 (0,380)	0,007 (0,027)	0,019 (0,003)
			Gran	0,230 (0,289)	0,466 (0,258)	0,005 (0,013)	0,012 (0,001)
	0,75	49	Petit	0,285 (0,139)	0,548 (0,465)	-0,004 (0,057)	0,025 (0,001)
			Gran	0,247 (0,105)	0,576 (0,416)	0,001 (0,027)	0,013 (0,004)
		225	Petit	0,238 (0,088)	0,606 (0,386)	0,007 (0,028)	0,016 (0,002)
			Gran	0,233 (0,046)	0,662 (0,266)	0,004 (0,013)	0,001 (0,001)

*La desviació estàndard està entre parèntesis.

Taula 2.4

Valor mitjà de les estimacions de σ i λ, biaix ($BIAIX_{SMR}$) i error quadràtic mitjà (MSE_{SMR}) de les prediccions de l'SMR utilitzant el procediment amb les transformacions quan s'ha fixat σ a 1 i λ a 0,25, 0,5 i 0,75, amb un nombre de regions de 49 i 225 i casos esperats petits i grans*

				Valor mitjà			
Real σ	Real λ	Nre. de regions	Casos esperats	σ	λ	$BIAIX_{SMR}$	MSE_{SMR}
1	0,25	49	Petit	0,922 (0,286)	0,323 (0,299)	0,064 (0,067)	0,160 (0,043)
			Gran	0,956 (0,257)	0,282 (0,240)	0,019 (0,031)	0,047 (0,015)
		225	Petit	0,958 (0,152)	0,270 (0,145)	0,062 (0,031)	0,138 (0,016)
			Gran	0,977 (0,132)	0,260 (0,124)	0,020 (0,014)	0,043 (0,006)
	0,5	49	Petit	0,839 (0,256)	0,460 (0,351)	0,051 (0,064)	0,131 (0,037)
			Gran	0,896 (0,218)	0,478 (0,313)	0,017 (0,031)	0,042 (0,014)
		225	Petit	0,935 (0,139)	0,481 (0,199)	0,050 (0,029)	0,112 (0,014)
			Gran	0,966 (0,124)	0,486 (0,170)	0,017 (0,015)	0,038 (0,005)
	0,75	49	Petit	0,797 (0,210)	0,606 (0,355)	0,048 (0,066)	0,118 (0,032)
			Gran	0,830 (0,186)	0,632 (0,329)	0,019 (0,032)	0,041 (0,014)
		225	Petit	0,925 (0,118)	0,696 (0,205)	0,044 (0,029)	0,098 (0,013)
			Gran	0,951 (0,104)	0,703 (0,182)	0,016 (0,014)	0,035 (0,118)

*La desviació estàndard està entre parèntesis.

2.4. Exemples

2.4.1. Dades del càncer de llavis a Escòcia

Les dades de càncer de llavis a Escòcia (Clayton i Kaldor, 1987) consisteixen en el nombre d'homes que presenten aquest càncer en cadascun dels 56 comtats d'Escòcia. El model considerat per modelar el risc del càncer de llavis al llarg dels comtats és:

$$\log\left(\mu_i^b\right) = \log\left(E_i\right) + \beta_o + \beta_1 x_i + b_i,$$

on E_i és el nombre de casos esperats pel comtat i-*èsim*, els quals es calculen a partir de la distribució de la població per edats i sexe, x_i és una covariable que mesura el percentatge de la població que es dedica a l'agricultura, la piscicultura i la silvicultura, i b_i és l'efecte aleatori corresponent al comtat i-*èsim*. En la taula 2.5 es presenten les estimacions dels paràmetres i els seus errors estàndard utilitzant el procediment amb transformacions i sense:

Taula 2.5
Estimacions dels paràmetres del model per les dades de càncer de llavis a Escòcia considerant els procediments amb transformacions i sense dels paràmetres

	Paràmetres			
Procediment	β_0	β_1	σ	λ
Sense transformació	–0,17 (< 0)	0,34 (0,112)	0,618 (0,126)	1,225 (0,035)
Amb transformació	–0,192 (1,160)	0,376 (0,115)	0,645 (0,124)	0,994 (0,122)

β_0: ordenada en l'origen; β_1: coeficient de regressió de la covariable percentatge que es dedica a l'agricultura, la piscicultura i la silvicultura; σ: sobredispersió, i λ: pes de la dependència espacial.

Malgrat que podem obtenir una solució utilitzant els dos procediments, el model sense transformació proporciona una estimació del paràmetre λ superior a 1, per sobre del límit superior per a aquest paràmetre. En conseqüència, aquestes estimacions no són de màxima versemblança i la matriu de variàncies i covariàncies tampoc no és correcta. Això provoca que els errors estàndards no siguin correctament estimats, com per exemple en aquestes dades, on l'error de l'ordenada a l'origen és negatiu.

En canvi, utilitzant el procediment amb transformacions, l'estimació dels components de la matriu de variàncies i covariàncies obtinguts són dins del seu domini; σ és 0,645 i λ és 0,994. Aleshores, s'han pogut obtenir els errors estàndard dels efectes fixos, amb els quals és possible poder testar la significació de la covariable si s'utilitza el test F-Wald (Brown i Prescott, 1999). Aplicant aquest test s'obté un *p*-valor de 0,0019 i, per tant, la relació entre el risc i la covariable és significativa.

2.4.2. Dades de la diabetis de tipus I

En l'exemple següent s'analitzen les dades d'incidència de la diabetis de tipus I a Catalunya. Aquestes dades consisteixen en el nombre d'homes menors de 15 anys als quals els ha estat diagnosticada la diabetis de tipus I entre els anys 1989 i 1998. Aquestes dades s'han agregat per a les 41 comarques de Catalunya i el model considerat és el següent:

$$\log\left(\mu_i^b\right) = \log\left(n_i\right) + \beta_o + b_i,$$

on b_i és l'efecte aleatori de la i-*èsima* comarca i n_i és el nombre de persones en risc per a la mateixa comarca. Atès que la població d'estudi en el període considerat és bastant estable, s'ha utilitzat la població censada l'any 1996 com a població de risc.

En la taula 2.6 es presenten tan sols les estimacions dels paràmetres del model mitjançant el procediment amb les transformacions, ja que sense transformacions no se n'ha obtingut.

Taula 2.6
Estimacions dels paràmetres del model per les dades de diabetis de tipus I a Catalunya considerant els

procediments amb transformacions i sense dels paràmetres

Procediment	Paràmetres		
	β_0	σ	λ
Sense transformació	No convergeix		
Amb transformació	-8,983 (0,012)	0,245 (0,109)	0,834 (0,841)

β_0: ordenada en l'origen; σ: sobredispersió, i λ: pes de la dependència espacial.

Amb el procediment que inclou les transformacions dels paràmetres s'ha obtingut convergència a la iteració 23 i s'ha estimat una sobredispersió (σ) de 0,245 i una dependència espacial (λ) de 0,834.

2.5. Discussió

Amb les reparametritzacions proposades en aquest capítol s'ha intentat assolir l'objectiu de millorar el percentatge de convergència de l'algoritme de maximització quasiversemblança penalitzada quan les taxes d'una malaltia són modelades a partir d'un model lineal generalitzat mixt. Aquesta millora s'ha obtingut aplicant unes transformacions als paràmetres que defineixen la matriu de variàncies i covariàncies dels efectes aleatoris. A partir de l'estudi de simulació s'ha observat que el percentatge de convergència és sempre elevat amb el procediment que inclou les transformacions. Els avantatges en relació amb la convergència es mostren principalment quan el valor de la sobredispersió, el nombre de regions i casos esperats són petits, i en aquests casos és on la falta de convergència del procediment sense transformacions dels paràmetres és més extrema, ja que el percentatge varia entre el 12,5% i el 20,5% i, en canvi, si les transformacions són aplicades, aquests valors són del 81% i del 81,5%.

Aquesta millora en la convergència també és palesa en els exemples. En el primer exemple l'estimació de la lambda està fora del seu domini i per tant les estimacions no són de màxima versemblança i, a més, la matriu de variàncies és incorrecta. Això provoca que l'error estàndard de l'ordenada a l'origen sigui negatiu. En aquesta situació, una solució possible és fixar el paràmetre lambda en el seu valor límit i estimar el model una altra vegada (Leroux, 2000). Però aquesta solució comporta una decisió subjectiva, que quan s'incorporen les transformacions no cal fer perquè l'obtenció de les estimacions dels paràmetres dins del seu domini ens permet efectuar un test estadístic, com, per exemple, el quocient de versemblances, per testar si els paràmetres poden ser fixats en un valor determinat, i en aquest cas es pren la decisió sobre la base dels resultats estadístics. En el segon exemple, els avantatges de la transformació són encara més obvis, perquè amb aquest procediment s'aconsegueix una estimació dels paràmetres que sense transformacions no s'assoleix.

Addicionalment, treballar amb els paràmetres transformats es pot estendre a altres parametritzacions de la matriu de variàncies i covariàncies dels efectes aleatoris,

com, per exemple, en el model del CAR no intrínsec definit per Besag, York i Mollié (1991), on les dues variàncies poden ser transformades mitjançant el logaritme. Malgrat que s'ha utilitzat la quasiversemblança penalitzada per obtenir les estimacions dels paràmetres, aquestes transformacions també es poden aplicar en altres procediments basats en la versemblança.

Per finalitzar, la implementació d'aquestes transformacions en el procés de maximització és senzilla i no comporta més cost computacional.

CAPÍTOL III: MÈTODES PER TESTAR L'ABSÈNCIA D'UN PATRÓ GEOGRÀFIC

3.1. Introducció

L'anàlisi geogràfica del risc d'una malaltia incorporant una estructura de CAR no intrínsec, permet controlar la sobredispersió tant d'origen espacial com no espacial. Sota aquest model, la matriu de variàncies i covariàncies dels efectes aleatoris està composta per dos paràmetres (Leroux *et al.*, 1999), l'un per mesurar la sobredispersió, σ^2 i l'altre, λ, per mesurar el pes de la correlació espacial en la sobredispersió. De manera que, si aquest pes és igual a 1, tota la sobredispersió és deguda a la correlació espacial entre regions, i en canvi, si és 0, hi ha independència espacial.

La independència espacial implica l'absència d'un patró geogràfic en les dades, de manera que tota la sobredipersió està produïda per altres factors d'agregació de les dades. Per tant, un cop que s'han modelat les dades és interessant testar l'existència de dependència espacial. Aquest test es redueix a contrastar si el paràmetre λ és igual a 0, i aleshores el contrast d'hipòtesi associat és:

$$H_0 : \lambda = 0$$
$$H_A : \lambda > 0.$$

En la bibliografia es poden trobar diferents aproximacions per testar els components de la variància dels efectes aleatoris, entre els quals s'ha volgut destacar el test de Wald, el quocient de les versemblances (Brown i Prescott, 1999), l'*score test* (Lin, 1997) i el *bootstrap* paramètric (MacNab i Dean, 2000.). També hi ha mesures de bondat de l'ajust dels models com ara l'*AKAIKE's information criterion* (Vonesh i Chinchilli, 1997) i el coeficient de concordança entre matrius (Vonesh *et al.*, 1996).

L'objectiu d'aquest capítol és estudiar el comportament de les diverses proves per testar independència espacial en termes de l'error de tipus I i de potència associada a cadascuna d'aquestes. A més, es vol avaluar el comportament del coeficient de concordança com a mesura de bondat d'ajust i la seva utilitat per prendre decisions en relació amb la independència espacial. Per tal d'analitzar aquests aspectes es durà a terme un estudi de simulació.

Les proves proposades també estan il·lustrades en dues aplicacions: les dades d'incidència del càncer de llavis a Escòcia (Clayton i Kaldor, 1987) i les dades d'incidència de la diabetis de tipus I de Catalunya.

3.2. Model i proves d'independència

3.2.1. Definició

El model lineal mixt generalitzat introduït en el capítol 1 assumia que el nombre de casos observats en cada àrea $Y = (Y_1, \ldots, Y_N)$ condicionat als efectes aleatoris $b = (b_1, \ldots b_N)$ es distribuïa sota una distribució de Poisson de mitjana $\mu^b = (\mu_1^b, \ldots, \mu_N^b)$. La funció que enllaça aquestes mitjanes amb el predictor lineal és la logarítmica:

$$\log(\mu^b) = \log(E) + Xa + Zb,$$

on *X* i *Z* són les matriu de disseny dels efectes fixos i aleatoris, respectivament, *a* és el vector de coeficients dels efectes fixos, i $E = (E_1, \ldots, E_N)$ és el vector del nombre de casos esperats, els quals són calculats a partir de les taxes d'una població estàndard o les de l'àrea d'estudi en què es troben les regions. Els efectes aleatoris *b* es distribueixen sota una distribució normal multivariant amb un vector de mitjanes de 0 i una matriu de covariàncies

$$\Sigma(\delta) = \sigma^2 R^- = \sigma^2 [\lambda Q_N + (1-\lambda) I_N]^-,$$

on $\delta = (\sigma, \lambda)$.

Sota aquesta definició de la matriu de variàncies i covariàncies dels efectes aleatoris, aquests són considerats independents quan el valor de λ és igual a 0, per tant, la hipòtesi nul·la del contrast d'independència que cal dur a terme correspondrà a $H_0 : \lambda = 0$, i la hipòtesi alterna, a $H_1 : \lambda > 0$.

3.2.2. Proves d'independència espacial

Les proves proposades per dur a terme el contrast d'hipòtesis sobre independència espacial són: l'*score test*, el test de Wald, el quocient de les versemblances, l'*AKAIKE's*

information criterion, el coeficient de concordança entre matrius i el *bootstrap* paramètric.

3.2.2.1. Score test

Lin (1997) proposà un *score test* global per contrastar si tots els components de la variància són igual a zero, i un *score test* individual per fer el contrast de tan sols un d'aquests components. En el contrast d'independència espacial només es fa un contrast sobre el paràmetre de correlació espacial, per tant s'ha d'utilitzar l'*score test* individual per resoldre la hipòtesi d'independència espacial. Aleshores, si es defineix γ com el vector de paràmetres que s'ha d'estimar sota la hipòtesi nul·la (en el nostre cas $\gamma = \sigma$) l'estadístic de contrast es defineix com:

$$z_\lambda = \frac{U_\lambda(\hat{\gamma})}{S_\lambda(\hat{\gamma})^{1/2}},$$

on $U_\lambda(\hat{\gamma})$ és la funció *score* del paràmetre que s'ha de contrastar, λ, avaluada sota la hipòtesi nul·la, és a dir,

$$U_\lambda(\hat{\gamma}) = \left.\frac{\partial l(\gamma;\lambda)}{\partial \lambda}\right|_{\lambda=0,\hat{\gamma}},$$

on $l(\gamma;\lambda)$ és la log-versemblança, i $S_\lambda(\hat{\gamma})$ la *informació eficient* del paràmetre λ, que està definida per:

$$S_\lambda(\hat{\gamma}) = I_{\lambda,\lambda} - I_{\gamma,\lambda} I_{\gamma,\gamma}^{-1} I_{\gamma,\lambda},$$

on $I_{\lambda,\lambda} = E\left(\frac{\partial l(\gamma;\lambda)}{\partial\lambda\partial\lambda}\right)$, $I_{\gamma,\lambda} = E\left(\frac{\partial l(\gamma;\lambda)}{\partial\gamma\partial\lambda}\right)$ i $I_{\gamma,\gamma} = E\left(\frac{\partial l(\gamma;\lambda)}{\partial\gamma\partial\gamma}\right)$.

Sota la hipòtesi nul·la, aquest estadístic de contrast es distribueix sota una distribució normal de mitjana 0 i desviació típica 1.

Si el model es maximitza utilitzant el mètode de la quasiversemblança penalitzada definida per Breslow i Clayton (1993), els components de la variància s'estimen mitjançant el mètode de la màxima versemblança restringida. Així doncs, l'equació que cal maximitzar és:

$$l_{REML} \approx -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|X^t V^{-1} X\right| - \frac{1}{2}\left(Y^* - X\hat{\beta}\right)^t V^{-1}\left(Y^* - X\hat{\beta}\right),$$

on es recorda que *Y** és la pseudovariable amb els elements igual a:

$$y_i^* = \eta_i^b + \left(y_i - \mu_i^b\right)\frac{\partial \eta_i^b}{\partial \mu_i^b} = Xa + Zb + \left(\frac{Y - \mu^b}{\mu^b}\right),$$

$$\text{i } V = W^{-1} + Z\Sigma(\delta)Z^t,$$

on *W* és la matriu diagonal de pesos composta pels termes

$$w_i = \frac{1}{v\left(\mu_i^b\right)}\left[\frac{\partial \mu_i^b}{\partial \eta_i}\right]^2 = \mu_i^b.$$

Llavors, la funció *score* per al paràmetre λ és:

$$U(\lambda) = \frac{\partial l_{REML}}{\partial \lambda} = -\frac{1}{2}\left\{\left(Y^* - X\hat{\beta}\right)^t V^{-1}\frac{\partial V}{\partial \lambda}V^{-1}\left(Y^* - X\hat{\beta}\right) - tr\left(P\frac{\partial V}{\partial \lambda}\right)\right\},$$

on $\frac{\partial V}{\partial \lambda}$ és igual a:

$$\frac{\partial V}{\partial \lambda} = \frac{\partial\left(W^{-1} + Z'DZ\right)}{\partial \lambda} = -\sigma^2 Z'R^{-1}(Q - I)R^{-1}Z.$$

Per tant, la funció *score* és defineix de la manera següent:

$$U(\lambda) = -\frac{1}{2}\left\{\left(Y^* - X\hat{\beta}\right)^t V^{-1}\left(-\sigma^2 Z'R^{-1}(Q - I)R^{-1}Z\right)V^{-1}\left(Y^* - X\hat{\beta}\right) - tr\left(P\left(-\sigma^2 Z'R^{-1}(Q - I)R^{-1}Z\right)\right)\right\}$$

Els elements de la informació eficient es defineixen mitjançant les expressions següents:

$$I_{\lambda,\lambda} = E\left(\frac{\partial l_{REML}(\sigma;\lambda)}{\partial \lambda \partial \lambda}\right) = \frac{1}{2}tr\left(P\frac{\partial V}{\partial \lambda}P\frac{\partial V}{\partial \lambda}\right) = \frac{1}{2}tr\left(P\left(-\sigma^2 Z'R^{-1}(Q - I)R^{-1}Z\right)P\left(-\sigma^2 Z'R^{-1}(Q - I)R^{-1}Z\right)\right)$$

$$I_{\gamma,\lambda} = I_{\sigma,\lambda} = E\left(\frac{\partial l_{REML}(\gamma;\lambda)}{\partial \sigma \partial \lambda}\right) = \frac{1}{2}tr\left(P\frac{\partial V}{\partial \sigma}P\frac{\partial V}{\partial \lambda}\right) = \frac{1}{2}tr\left(P\left(2\sigma R^{-1}\right)P\left(-\sigma^2 Z'R^{-1}(Q - I)R^{-1}Z\right)\right)$$

$$I_{\gamma,\gamma} = I_{\sigma,\sigma} = E\left(\frac{\partial l_{REML}(\sigma;\lambda)}{\partial \sigma \partial \sigma}\right) = \frac{1}{2}tr\left(P\frac{\partial V}{\partial \sigma}P\frac{\partial V}{\partial \sigma}\right) = \frac{1}{2}tr\left(P\left(2\sigma R^{-1}\right)P\left(2\sigma R^{-1}\right)\right),$$

on $P = V^{-1} - V^{-1}X\left(X^tV^{-1}X\right)^{-1}X^tV^{-1}$.

3.2.2.2. Test de Wald

El test de Wald és defineix com el quocient entre l'estimació de λ i el seu error estàndard:

$$Z = \frac{\hat{\lambda}}{e.s\left(\hat{\lambda}\right)}.$$

Sota la hipòtesi nul·la, aquest estadístic de contrast es distribueix sota una distribució normal de mitjana 0 i desviació típica 1.

El comportament d'aquesta prova està subjecte a la distribució de probabilitat de $\hat{\lambda}$. Si aquesta s'aproxima a una distribució normal el test funcionarà correctament, atès que *Z* també s'aproximarà a una distribució normal estàndard.

Habitualment aquesta prova es presenta sota una distribució khi quadrat perquè es generalitza a més d'un paràmetre, però en el cas que tan sols es vulgui avaluar un paràmetre, el resultat de la prova és equivalent tant si es fa sota una distribució normal estàndard o si es fa amb una khi quadrat d'un grau de llibertat.

3.2.2.3. Quocient de les versemblances

El quocient de les versemblances és una prova que s'utilitza per comparar la millora de l'ajust en models imbricats, és a dir, en situacions on un dels models està compost per un subconjunt de paràmetres de l'altre model. En el cas d'independència espacial aquesta prova és aplicable perquè el model sota la hipòtesi nul·la és un model niat respecte al model CAR.

Per tant, aquest test compara la versemblança del model amb tots els components de la variància σ i λ, amb la versemblança del model restringit sota la hipòtesi nul·la. Aquesta raó és defineix com:

$$\mathrm{LRT} = -2 \times \log\left(\frac{\mathrm{L}_2(\beta, \mathrm{b}, \sigma)}{\mathrm{L}_1(\beta, \mathrm{b}, \sigma, \lambda)} \right) = (-2) \times \left[\log(\mathrm{L}_2(\beta, \mathrm{b}, \sigma)) - \log(\mathrm{L}_1(\beta, \mathrm{b}, \sigma, \lambda)) \right],$$

A més sota la hipòtesis nul·la aquest estadístic es distribueix sota una mixtura de dos distribucions chi-quadrat, que es defineix com: $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$.

L'estimació dels components de la variància es fa maximitzant una aproximació de les log-versemblances restringides; per tant, el test es construirà a partir d'aquestes.

La log-versemblança restringida del model CAR correspon a:

$$\log\left(L_{REML}\left(\hat{\beta},\hat{\lambda},\hat{\sigma}\right)\right)\approx -\frac{1}{2}\log\left|V\left(\hat{\beta},\hat{\lambda},\hat{\sigma}\right)\right|-\frac{1}{2}\log\left|X^{t}V\left(\hat{\beta},\hat{\lambda},\hat{\sigma}\right)^{-1}X\right|-\frac{1}{2}\left(Y^{*}-X\hat{\beta}\right)^{t}V\left(\hat{\beta},\hat{\lambda},\hat{\sigma}\right)^{-1}\left(Y^{*}-X\hat{\beta}\right)$$

i del model restringit o sota la hipòtesi nul·la, és a dir amb $\lambda = 0$, es defineix com:

$$\log\left(L_{REML}\left(\hat{\beta}_0,\hat{\sigma}_0\right)\right)\approx -\frac{1}{2}\log\left|V\left(\hat{\beta}_0,\hat{\sigma}_0\right)\right|-\frac{1}{2}\log\left|X^{t}V\left(\hat{\beta}_0,\hat{\sigma}_0\right)^{-1}X\right|-\frac{1}{2}\left(Y^{*}-X\hat{\beta}_0\right)^{t}V\left(\hat{\beta}_0,\hat{\sigma}_0\right)^{-1}\left(Y^{*}-X\hat{\beta}_0\right)$$

on $\left(\hat{\beta}_0,\hat{\sigma}_0\right)$ són les estimacions del model sota la hipòtesi nul·la, per tant, del model d'heterogeneïtat.

3.2.2.4. AKAIKE's information criterion

L'*AKAIKE information criterion* (AIC) és un estadístic que s'utilitza per valorar la bondat d'ajust. L'AIC es defineix com:

$$AIC = \log\left[L_{REML}\left(\hat{\beta},\hat{\lambda},\hat{\sigma}\right)\right]- s^{*},$$

on s^{*} és el nombre de paràmetres estimats.

El model que millor fita les dades observades és el que té un valor més alt d'AIC. La decisió d'elecció del model a partir de l'AIC no es basa en resultats probabilístics; és el mateix investigador qui decideix si l'increment de l'AIC és suficientment gran per seleccionar el model amb més paràmetres. Utilitzant l'AIC d'una forma estricta, s'hauria de triar el model amb un valor més gran d'AIC.

3.2.2.5. Coeficient de concordança

Sota la hipòtesi nul·la, la matriu de variàncies i covariàncies dels efectes aleatoris correspon a:

$$\Sigma_0 = \Sigma\left(\sigma,\lambda = 0\right) = \sigma^{2}I,$$

i sota l'alternativa:

$$\Sigma_1 = \Sigma\left(\sigma,\lambda\right) = \sigma^{2}R^{-} = \sigma^{2}\left[\lambda Q + \left(1-\lambda\right)I\right]^{-},$$

per tant, si $\lambda = 0$, les dues matrius coincideixen.

El coeficient de concordança entre matrius (Vonesh *et al.*, 1996) és una mesura de similitud entre matrius que ens pot ser d'ajuda per valorar si $\Sigma_0 = \Sigma_1$. Però en realitat el

test que es vol fer és si $\lambda = 0$, i per tant això implica que la concordança que s'ha d'avaluar és entre la matriu identitat, *I*, i la matriu R^-.

El coeficient de concordança entre les matrius V_1 i V_2 s'ha definit com:

$$r = 1 - \frac{\|d - i\|^2}{\|d\|^2 + \|i\|^2},$$

on $d = \text{vech}\left(D = V_2^{-½} V_1 V_2^{-½}\right)$ i $= \text{vech}(\text{Identitat})$, i on *vech* és la funció que vectoritza el triangle superior de la matriu $D = V_2^{-½} V_1 V_2^{-½}$ i de la matriu identitat, respectivament, i $\|x\|$ és la norma euclidiana del vector *x*. En el nostre cas, V_2 correspon a la matriu Identitat i V_1 a la R^-. Per tant, $V_2^{-½} V_1 V_2^{-½}$ és igual a $I^{-½} R^- I^{-½}$, i simplificant la matriu *D* és R^-. Si s'escriuen els elements de la matriu *D* i *I* amb els subíndexs matricials, aquests elements corresponen a:

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1j} & \cdots & d_{1N} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ d_{i1} & d_{i2} & \cdots & d_{ij} & \cdots & d_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nj} & \cdots & d_{NN} \end{pmatrix} \quad I = \begin{pmatrix} i_{11} & i_{12} & \cdots & i_{1j} & \cdots & i_{1N} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ i_{i1} & i_{i2} & \cdots & i_{ij} & \cdots & i_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ i_{N1} & i_{N2} & \cdots & i_{Nj} & \cdots & i_{NN} \end{pmatrix},$$

on en la matriu *I* els elements de la diagonal, i_{ij} amb $i = j$, són 1, i fora de la diagonal, i_{ij} amb $i \neq j$, són 0.

L'element $\|i\|^2$ del coeficient de concordança és igual al nombre de regions, atès que és la suma al quadrat dels elements del triangle superior de la matriu identitat, $\|i\|^2 = \sum_{i \geq j}^{k} i_{ij}^2$.

A partir d'aquesta notació el coeficient de concordança s'expressa:

$$r_c = 1 - \frac{\sum_{i \geq j}^{K} (d_{ij} - i_{ij})^2}{\sum_{i \geq j}^{k} d_{ij}^2 - N},$$

on *k* és el nombre d'elements resultants de la vectorització del triangle superior de les matrius, és a dir, $k = \frac{N((N-1)}{2}$.

El coeficient prendrà el valor 1 quan *d* sigui igual a *i*, i això es produeix quan $R^{-} = I$ i el valor de 0 quan els elements de *d* siguin ortogonals als elements de *i*.

Per tant, s'espera que el valor del coeficient obtingut sigui proper a 1 quan s'estima λ d'unes dades on hi ha independència espacial. En canvi, el coeficient de concordança prendrà valors pròxims a 0 en situacions de dependència espacial.

Atès que una de les matrius és la identitat, l'expressió d'aquest coeficient es pot simplificar així:

$$r_c = \frac{\sum_{i\geq j}^{k} d_{ij}^2 + N - \sum_{i\geq j}^{k} (d_{ij} - i_{ij})^2}{\sum_{i\geq j}^{k} d_{ij}^2 + N} = \frac{\sum_{i\geq j}^{k} d_{ij}^2 + N - \sum_{i\geq j}^{k} d_{ij}^2 - N + 2\sum_{i\geq j}^{k} d_{ij} i_{ij}}{\sum_{i\geq j}^{k} d_{ij}^2 + N} = \frac{2 \times tr(D)}{\sum_{i\geq j}^{k} d_{ij}^2 + N} = \frac{2 \times \sum_{i=1}^{N} d_{ii}}{\sum_{i\geq j}^{k} d_{ij}^2 + N}.$$

Un dels avantatges d'aquest coeficient és que és una mesura no paramètrica, en el sentit que no requereix l'especificació d'una funció de versemblança.

3.2.2.6. Bootstrap paramètric

MacNab i Dean (2000) proposen un altre procediment per testar la independència espacial: el *bootstrap* paramètric. Aquest procediment es basa en la generació de la distribució de la λ sota la hipòtesi nul·la per definir la regió de rebuig. Aleshores es comprova si l'estimació de λ obtinguda de les dades cau en la zona de rebuig i es pren una decisió.

La distribució sota la hipòtesi nul·la de λ es genera simulant *m* mostres mitjançant la funció de probabilitat de les dades sota $\lambda = 0$, que es coneixen com les *mostres bootstrap*. D'aquesta manera, si s'especifica la distribució de les dades *Y* sota la hipòtesi nul·la com $F_0(\hat{\beta}_0, \hat{\sigma}_0^2)$, es generen *m* mostres d'aquesta distribució:

$$Y^{(1)} \sim F_0(\hat{\beta}_0, \hat{\sigma}_0^2),\ Y^{(2)} \sim F_0(\hat{\beta}_0, \hat{\sigma}_0^2) \ldots Y^{(m)} \sim F_0(\hat{\beta}_0, \hat{\sigma}_0^2).$$

En el nostre cas la distribució de les dades, $F_{0,}$ és una mixtura de la funció de probabilitat dels efectes aleatoris i de les dades condicionades als efectes aleatoris, i aquestes són una distribució normal i una de Poisson, respectivament. Per tant, una mostra *bootstrap* s'obté primer generant una mostra d'efectes aleatoris amb els quals es determinen les mitjanes de cada regió, i a partir d'aquestes mitjanes se simulen les dades. Les distribucions utilitzades es parametritzen mitjançant les estimacions obtingudes i fixant $\lambda = 0$.

En cada mostra *bootstrap* simulada s'estima el paràmetre λ, i s'aconsegueixen un conjunt de *m* estimacions: $\hat{\lambda}_1, \hat{\lambda}_2, \ldots \hat{\lambda}_m$.

El valor crític que defineix la regió de rebuig associada al contrast d'independència, que cal recordar que és unilateral, i si es treballa amb un error de tipus I dels 5% correspon al percentil 95% de la mostra d'estimacions *bootstrap* $\hat{\lambda}_1, \hat{\lambda}_2, \ldots \hat{\lambda}_m$, es denota com $\hat{\lambda}_{(95\%)}$. Aleshores, si el valor estimat de $\hat{\lambda}$ cau a la zona de rebuig, la decisió serà rebutjar la hipòtesi nul·la.

També es pot estimar el *p*-valor o grau de significació del test de la manera següent:

$$p = \frac{1 + \sum_{i=1}^{m} \left(\hat{\lambda}_i \geq \hat{\lambda} \right)}{m + 1}.$$

3.3. Simulació

3.3.1. Definició

Per poder comprovar el comportament de les diverses proves explicades per contrastar la independència espacial, s'ha efectuat un estudi de simulació. En aquest estudi les proves es comparen en termes d'error de tipus I i de potència. El procediment de generació dels 400 conjunts de dades simulades és el mateix que s'ha utilitzat en la simulació del capítol dos. Els valors considerats per als components de la variància dels efectes aleatoris han estat 0 i 0,25 per a λ, i 0,09, 0,25, 0,5 i 1 per a σ^2, i el nombre de regions s'ha fixat a 49 i 100.

Quan les dades se simulen sota el model de $\lambda = 0$, la característica de les proves que serà avaluada és l'error de tipus I. En canvi, quan se simulen sota una λ de 0,25, s'estudia la potència. No s'han considerat valors més alts per a λ perquè la potència augmentarà a mesura que la distància entre la hipòtesi nul·la i alterna sigui més gran.

El valor nominal de la probabilitat d'error de tipus I s'ha fixat al 5%. En la simulació s'ha estudiat el comportament del test de Wald, del quocient de les versemblances, de la diferència d'AIC, de l'*score test* i del coeficient de concordança.

3.3.2. Resultats

Els resultats referents a l'error de tipus I i la potència es presenten en les taules 3.1 i 3.2:

Taula 3.1

Percentatge de rebuig de la hipòtesi d'independència espacial obtingut per les proves quocient de les versemblances (LRT), *score test*, test de Wald i diferència d'AIC quan $\lambda = 0$, i en funció del nombre de regions, de casos esperats i de valors de sobredispersió

Valor real	Nre. de regions	Nre. de casos esperats	LRT	*Score test*	Test de Wald	Diferència d'AIC
$\sigma^2 = 1$	49	Petit	4,58	4,58	0,00	8,14
		Gran	4,52	5,03	0,00	7,54
	100	Petit	3,81	4,82	0,25	9,39
		Gran	5,53	5,28	0,25	7,04
$\sigma^2 = 0{,}5$	49	Petit	4,25	4,25	0,00	7,00
		Gran	3,50	4,25	0,00	5,75
	100	Petit	5,50	5,25	0,25	9,00
		Gran	4,50	4,50	0,00	6,75
$\sigma^2 = 0{,}25$	49	Petit	2,77	4,50	0,00	6,55
		Gran	3,50	3,75	0,00	5,50
	100	Petit	5,75	8,00	0,75	11,00
		Gran	4,25	4,25	0,00	7,00
$\sigma^2 = 0{,}09$	49	Petit	2,47	5,37	0,00	4,40
		Gran	4,25	4,75	0,00	5,75
	100	Petit	4,95	6,77	2,86	7,29
		Gran	4,50	6,25	0,25	7,25

Taula 3.2

Percentatge de rebuig de la hipòtesi d'independència espacial obtingut per les proves quocient de les versemblances (LRT), *score test*, test de Wald i diferència d'AIC quan $\lambda = 0{,}25$, i en funció del nombre de regions, de casos esperats i valors de sobredispersió

Valor real	Nre. de regions	Nre. de casos esperats	LRT	*Score test*	Test de Wald	Dif. d'AIC
$\sigma^2 = 1$	49	Petit	38,60	42,50	3,51	44,11
		Gran	47,25	49,00	2,75	55,50
	100	Petit	55,75	58,50	11,00	61,75
		Gran	73,50	76,00	21,00	80,75
$\sigma^2 = 0{,}5$	49	Petit	20,55	26,50	2,51	28,82
		Gran	44,00	48,75	2,25	53,00
	100	Petit	44,00	50,50	9,00	51,75
		Gran	69,75	71,00	17,00	74,75
$\sigma^2 = 0{,}25$	49	Petit	21,24	23,23	1,34	27,69
		Gran	36,75	41,25	1,75	46,25
	100	Petit	27,34	33,25	9,37	34,68
		Gran	64,00	64,25	12,25	68,75
$\sigma^2 = 0{,}09$	49	Petit	6,90	12,98	0,00	12,07
		Gran	26,97	28,00	3,82	32,82
	100	Petit	13,25	20,59	15,66	17,77
		Gran	40,75	44,00	9,75	49,75

Respecte a la probabilitat d'error de tipus I, en la taula 3.1 podem observar que l'*score test* en les combinacions on la sobredispersió és superior o igual a 0,5, el test funciona bastant bé atès que l'error de tipus I estimat és pròxim al valor nominal. En canvi, quan la sobredispersió és més petita, els resultats són més variables, i s'observa que amb una σ^2 de 0,25 el percentatge de rebuig varia entre el 3,75% i el 8%, i amb una σ^2 de 0,09, entre el 4,75% i el 6,77%.

Els resultats obtinguts amb l'LRT en general són més variables que amb l'*score*, i en algunes combinacions on el nombre de regions és 49 s'obté un error de tipus I inferior al 3,5%. De totes formes es pot apreciar que aquest error millora quan s'incrementa el nombre de regions.

En relació amb el test de Wald, en totes les combinacions realitzades se subestima clarament el valor nominal, i s'obté un valor màxim d'error de tipus I del 2,86%.

L'AIC s'ha utilitzat de manera que la decisió de rebutjar la hipòtesi nul·la s'ha dut a terme quan la diferència entre els valors d'AIC dels dos models era inferior a 0. Amb aquest punt de decisió s'ha obtingut un percentatge de rebuig entre el 4,4% i l'11%.

En relació amb les potències de les proves (taula 3.2), podem observar un comportament comú: un increment de la potència quan augmenta tant el valor de la sobredispersió com del nombre de regions i de casos esperats. El test amb uns resultats pitjors és el test de Wald, en canvi l'AIC és el que presenta un nombre de rebuigs més elevat. Es fa notar que, entre l'*score test* i l'LRT, el primer presenta uns valors de potències una mica superiors.

Pel que fa al coeficient de concordança, a les taules 3.3 i 3.4 es mostren la mitjana, i la desviació estàndard d'aquest coeficient quan $\lambda = 0$ i $\lambda = 0{,}25$ respectivament. A les figures 3.1.*a*, 3.1.*b*, 3.2.*a* i 3.2.*b* (presentades al final del capítol) es representen els gràfics de caixes dels coeficients per a les combinacions de $\lambda = 0$ i $\lambda = 0{,}25$ en funció dels valors de sobredispersió, de nombre de regions i de casos esperats, respectivament.

Taula 3.3

Mitjana i desviació típica del coeficient de concordança quan $\lambda = 0$, i en funció dels valors de correlació espacial i sobredispersió i del nombre de regions i de casos esperats

	Valor real	Nre. de regions	Nre. de casos esperats	Mitjana	Desviació típica
λ=0	$\sigma^2 = 1$	49	Petit	0,9539	0,0070
			Gran	0,9666	0,0041
		100	Petit	0,9728	0,0033
			Gran	0,9785	0,0017
	$\sigma^2 = 0{,}5$	49	Petit	0,9461	0,0099
			Gran	0,9646	0,0046
		100	Petit	0,9659	0,0048
			Gran	0,9781	0,0019
	$\sigma^2 = 0{,}25$	49	Petit	0,9272	0,0213
			Gran	0,9614	0,0051
		100	Petit	0,9472	0,0097
			Gran	0,9752	0,0022
	$\sigma^2 = 0{,}09$	49	Petit	0,8492	0,0806
			Gran	0,9478	0,0084
		100	Petit	0,8912	0,0385
			Gran	0,9633	0,0051

Taula 3.4

Mitjana i desviació típica del coeficient de concordança quan $\lambda = 0,25$, i en funció dels valors de correlació espacial i sobredispersió i del nombre de regions i de casos esperats

	Valor real	Nre. de regions	Nre. de casos esperats	Mitjana	Desviació típica
λ=0,25	$\sigma^2 = 1$	49	Petit	0,7580	0,0539
			Gran	0,7858	0,0308
		100	Petit	0,7720	0,0221
			Gran	0,7688	0,0183
	$\sigma^2 = 0.5$	49	Petit	0,7661	0,0616
			Gran	0,7802	0,0344
		100	Petit	0,7580	0,0286
			Gran	0,7695	0,0188
	$\sigma^2 = 0,25$	49	Petit	0,6893	0,1135
			Gran	0,7847	0,0355
		100	Petit	0,7595	0,0494
			Gran	0,7720	0,0210
	$\sigma^2 = 0,09$	49	Petit	0,6478	0,1723
			Gran	0,7280	0,0780
		100	Petit	0,6818	0,1139
			Gran	0.7611	0.0388

Els resultats presentats a les taules 3.3 i 3.4 mostren que quan hi ha independència espacial la mitjana dels coeficients de concordança són, en general, superiors a 0,9. En canvi, quan $\lambda = 0,25$, la mitjana dels coeficients de concordança entre les dues matrius és inferior al 0,8.

En general, es pot observar que el coeficient de concordança augmenta i la seva variabilitat disminueix quan augmenta el nombre de regions. En relació amb la variabilitat, també s'observa que són més variables les estimacions del coeficient de concordança quan λ val 0,25 que quan val 0. Aquest comportament també es pot apreciar en els *box-plots* de les figures 3.1 i 3.2.

Segons els resultats, s'ha observat que quan hi ha independència espacial el coeficient de concordança tendeix a ser superior a 0,9, i, en canvi, quan hi ha dependència la tendència és a ser inferior a 0,8. Aquest fet permet plantejar-se un valor per decidir l'existència o no d'independència espacial. Aquest valor es podria ubicar dins de l'interval de 0,8 a 0,9. Per estudiar-ho s'ha calculat per a cada combinació els

percentatges de valors del coeficient de concordança que són superiors a 0,8, 0,85 i 0,9. Aquests resultats es presenten en la taula 3.5.

Taula 3.5

Percentatge de valors del coeficient de concordança superiors al 0,8, 0,85 i 0,9 en funció dels valors de correlació espacial i sobredispersió i del nombre de regions i de casos esperats

				Percentatge		
	Valor real	Nre. de regions	Nre. de casos esperats	≥ 0,8	≥ 0,85	≥ 0,9
λ=0	$\sigma^2 = 1$	49	Petit	92,50	87,75	83,25
			Gran	95,00	91,75	89,25
		100	Petit	98,00	95,75	89,25
			Gran	99,50	97,75	93,75
	$\sigma^2 = 0{,}5$	49	Petit	89,75	85,25	79,75
			Gran	96,00	92,75	86,50
		100	Petit	95,75	93,75	89,75
			Gran	99,00	96,50	93,50
	$\sigma^2 = 0{,}25$	49	Petit	85,89	80,35	77,08
			Gran	94,50	91,50	84,25
		100	Petit	91,00	87,50	80,75
			Gran	99,00	96,50	91,75
	$\sigma^2 = 0{,}09$	49	Petit	76,65	74,18	70,05
			Gran	90,75	85,25	80,25
		100	Petit	78,65	75,78	71,88
			Gran	95,25	92,75	86,75
λ=0,25	$\sigma^2 = 1$	49	Petit	47,37	42,11	37,84
			Gran	50,50	41,50	29,50
		100	Petit	44,00	35,00	24,25
			Gran	44,00	31,00	18,50
	$\sigma^2 = 0.5$	49	Petit	52,63	47,12	39,35
			Gran	49,00	39,50	30,50
		100	Petit	41,75	34,75	24,00
			Gran	41,50	31,50	22,75
	$\sigma^2 = 0{,}25$	49	Petit	47,04	43,01	37,63
			Gran	50,25	43,25	34,00
		100	Petit	47,09	40,76	33,16
			Gran	41,75	32,25	25,25
	$\sigma^2 = 0{,}09$	49	Petit	52,87	50,86	49,14
			Gran	48,09	41,73	34,10
		100	Petit	43,37	41,27	40,06
			Gran	45,75	37,25	29,75

A la taula 3.5 es pot observar que quan hi ha independència espacial el percentatge de valors del coeficient de concordança amb un valor superior al 0,8 es mou entre el

76,65% i el 99,5%, quan val 0,85 és entre el 74,18% i el 97,75%, i quan val 0,9 és entre el 70,5% i el 93,75%; aquests percentatges s'entenen com a proporcions de decisions correctes quan $\lambda = 0$.

En canvi, quan hi ha dependència espacial s'obté que el percentatge de valors amb un coeficient de concordança superior al 0,8 és inferior al 50%. i aquest disminueix fins a valors inferiors al 40% en el cas d'un valor de referència del 0,9. En aquest cas aquests percentatges es corresponen amb decisions incorrectes ja que ens trobem en la situació $\lambda = 0{,}25$.

3.4. Aplicacions

3.4.1. Dades del càncer de llavis a Escòcia

En les dades de càncer de llavis a Escòcia (Clayton i Kaldor, 1987) que han estat analitzades en el segon capítol, es modela la raó estandarditzada de mortalitat masculina produïda pel càncer de llavis en cadascun dels 56 comtats d'Escòcia. El model considerat per modelar el risc del càncer de llavis en tots els comtats és:

$$\log\left(\mu_i^b\right) = \log\left(E_i\right) + \beta_o + \beta_1 x_i + b_i,$$

on E_i és el nombre de casos esperats pel comtat i*th*, els quals són calculats a partir de la distribució de la població per edats i sexe, x_i és una covariable que mesura el percentatge de població que es dedica a l'agricultura, la piscicultura i la silvicultura, i b_i és l'efecte aleatori del i*th* comtat.

En la taula 3.6 es presenten les estimacions dels paràmetres i els seus errors estàndard del model considerant una estructura per als efectes aleatoris CAR no intrínsec i d'heterogeneïtat. L'estimació s'ha dut a terme mitjançant el procediment de quasiversemblança penalitzada, i s'hi han incorporat les transformacions per als components de la variància proposades al capítol 2.

Taula 3.6
Estimacions dels paràmetres del model d'heterogeneïtat i CAR no intrínsec per a les dades de càncer de llavis a Escòcia*

	Paràmetres			
Model	β_0	β_1	σ	λ
Heterogeneïtat	-0,441 (0,157)	0,679 (0,141)	0,596 (0,082)	--
CAR no intrínsec	-0,192 (1,160)	0,376 (0,115)	0,645 (0,124)	0,994 (0,122)

β_0: ordenada en l'origen; β_1: coeficient de regressió de la covariable percentatge que es dedica a l'agricultura, la piscicultura i la silvicultura; σ: sobredispersió, i λ: pes de la dependència espacial.
*L'error estàndard s'ha posat entre parèntesis.

S'ha estimat una sobredispersió (σ^2)de 0,36, i el pes del component espacial és de 0,994. A partir d'aquest modelatge es mostra el comportament de les proves considerades per testar la independència espacial.

En la taula 3.7 es presenta el resultat dels estadístics de contrast i els *p*-valors associats. En aquesta aplicació també s'ha incorporat el procediment del *bootstrap* paramètric realitzat a partir de 1.000 remostres.

Taula 3.7
Estadístic de contrast i *p*-valor associat al test de Wald, *score test*, quocient de versemblances (LRT), diferència d'AIC i *bootstrap* paramètric i coeficient de concordança

Test	Estadístic de contrast	*p*-valor
Wald	8,15	<0,001
Score	3,562	<0,001
LRT	26,46	<0,001
Diferència d'AIC	-12,23	
Coeficient de concordança	0,015	
Bootstrap paramètric		0,001

El temps utilitzat per efectuar el procediment del *bootstrap* paramètric a partir de 1.000 mostres (mitjançant un Intel Pentium IV, a 2 Ghz i 512 Mb de RAM), ha estat al voltant de cinc hores. Per tant, es pot observar que aquest procediment és molt costós computacionalment.

En relació amb els resultats obtinguts, podem observar que totes les proves utilitzades han rebutjat la hipòtesi d'independència espacial. En la prova d'AIC s'ha obtingut un resultat negatiu, per tant, el model que incorpora la correlació espacial fita millor les dades, i aleshores es pot decidir rebutjar l'existència d'independència espacial. La concordança entre les matrius dels efectes aleatoris dels dos models considerats és notablement baixa i propera a 0, fet que referma la decisió de rebutjar la hipòtesi d'independència.

3.4.2. Dades de la diabetis de tipus I

En l'exemple següent s'analitzen les dades d'incidència de la diabetis de tipus I a Catalunya. Aquestes dades consisteixen en el nombre d'individus als quals s'ha diagnosticat la diabetis entre l'any 1989 i l'any 1998. Aquestes dades han estat agregades per a les 41 comarques de Catalunya i el model considerat és el següent:

$$\log\left(\mu_{ij}^{b}\right) = \log\left(n_{ij}\right) + \beta_{o} + \beta_{1}\text{Gènere}_{ij} + \beta_{2}\text{Edat}_{ij} + b_{i},$$

on b_i és l'efecte aleatori de la i-èsima comarca i n_i és el nombre de persones de risc per a la mateixa comarca. S'han incorporat al model les covariables dicotòmiques gènere i edat dels individus. La covariable edat consta de dos nivells referents a si la diabetis s'havia diagnosticat abans o després dels 14 anys. En relació amb aquestes variables es considera la categoria basal, les dones i els majors de 14 anys.

En la taula 3.8 es presenten les estimacions dels paràmetres conjuntament amb el seu error estàndard del model CAR no intrínsec i del d'heterogeneïtat.

Taula 3.8
Estimacions dels paràmetres del model d'heterogeneïtat i CAR no intrínsec de les dades de la diabetis de tipus I a Catalunya*

	Paràmetres				
Model	β_0	β_1	β_2	σ	λ
Heterogeneïtat	-9,474	0,273	0,353	0,237	--
	(0,059)	(0,039)	(0,038)	(0,047)	
CAR no intrínsec	-9,491	0,272	0,354	0,351	0,44
	(0,088)	(0,039)	(0,038)	(0,01)	(0,464)

β_0: ordenada en l'origen; β_1: coeficient de la variable gènere; β_2: coeficient de la variable edat; σ: sobredispersió, i λ: pes de la dependència espacial.

*L'error estàndard s'ha posat entre parèntesis.

En la taula 3.9 es presenten els resultats de les proves per testar la independència espacial.

Taula 3.9
Estadístic de contrast i *p*-valor associat al test de Wald, *score test*, quocient de versemblances (LRT), diferència d'AIC i *bootstrap* paramètric i coeficient de concordança

Test	Estadístic de contrast	*p*-valor
Wald	0,952	0,171
Score	0,600	0,274
LRT	1,856	0,086
Diferència d'AIC	1,928	
Coeficient de concordança	0,784	
Bootstrap paramètric		0,182

En la taula es mostra que totes les proves presenten un grau de significació superior al 5% i una diferència del valor AIC positiva, per tant la decisió serà no rebutjar la hipòtesi d'independència espacial. La situació en què ens trobem és la d'una àrea amb 41 regions, amb un nombre gran de casos esperats i un valor de sobredispersió de $\sigma^2 = 0{,}12$. En l'estudi de simulació s'ha generat una situació similar (N = 49,

$\sigma^2 = 0{,}09$), en la qual les potències de l'LRT, l'*score* i l'AIC es trobaven al voltant del 30%. Per tant, podríem suposar que ens trobem en una situació de manca de potència per detectar dependència espacial. Aquí és on pot ser útil el coeficient de concordança que ha donat un valor de 0,8. En la mateixa combinació de la simulació, més del 50% dels coeficients es trobaven per sota de 0,8 i gairebé el 60% per sota de 0,85 quan $\lambda = 0{,}25$. En canvi, si $\lambda = 0$, gairebé el 80% dels coeficients eren superiors a 0,8. Aleshores, basant-nos en el coeficient de concordança, la decisió podria ser la de dependència espacial, i es rebutjaria el resultat de les altres proves perquè aquesta és una situació de potencia baixa.

A més, cal remarcar que en aquest cas per calcular el *p*-valor del *bootstrap* paramètric van ser necessàries aproximadament 12 hores amb un processador AMD a 900 MHz i 384 MB de RAM.

3.5. Discussió i conclusions

L'objectiu principal d'aquest capítol ha estat estudiar el comportament de diverses proves proposades per testar independència espacial en estudis geogràfics de malalties. La realització d'aquest contrast és important atès que l'existència de correlació espacial ens indica que certes regions comparteixen factors de risc geogràfics desconeguts que provoquen que aquestes regions tinguin uns riscos similars. En conseqüència, l'existència de dependència espacial ens permet identificar conjunts de regions amb més o menys risc i dur a terme polítiques sanitàries més eficients.

Les proves que s'han considerat són: el test de Wald, l'*score test*, el quocient de les versemblances i l'*AKAIKE's information criterion* (AIC), les quals han estat avaluades en termes d'error de tipus I i de potència. També s'ha descrit el comportament del coeficient de concordança entre matrius.

De totes les proves aplicades, el test de Wald, l'*score test* i el coeficient de concordança presenten l'avantatge que per poder-los calcular únicament cal fitar un model, en canvi l'AIC i el quocient de les versemblances requereixen avaluar el model amb $\lambda = 0$ i el model on s'estima aquest paràmetre.

Els resultats obtinguts ens han mostrat que el test de Wald no té un bon comportament ni en relació amb l'error de tipus I ni amb la potència; concretament infraestima l'error de tipus I i no té capacitat per detectar l'existència de correlació espacial. Aquests problemes poden ser deguts al fet que falla l'assumpció de normalitat de l'estimació del paràmetre λ, tal com es pot observar en els *QQ-plots* de normalitat presentats en les figures 3.3.*a*, 3.3.*b*, 3.4.*a* i 3.4.*b* (presentades al final del capítol).

En relació amb el quocient de les versemblances i l's*core test* s'ha observat que el comportament d'aquestes dues proves són similars, però l'avantatge de l'*score test* és que presenta uns resultats més estables en totes les combinacions de sobredispersió, nombre de regions i nombre de casos esperats. Pel que fa a l'AIC s'ha observat que presenta un error de tipus I al voltant del 10%, però un bon comportament en termes de potència.

En vista dels resultats l'*score test* és la prova que presenta un equilibri més bo entre l'error de prendre la decisió d'assumir dependència quan realment hi ha independència espacial i l'error d'assumir independència quan no n'hi ha.

En aquest capítol també s'ha estudiat descriptivament el comportament del coeficient de concordança entre matrius i, a més, s'han analitzat possibles valors de referència d'aquest coeficient. Sobre la base dels resultats, es poden establir dos criteris: un amb els punts 0,8 i 0,9, i l'altre amb el punt 0,85. Així doncs, el primer consistiria a donar suport a la decisió de dependència quan el valor és inferior a 0,8, d'independència quan és superior a 0,9, i d'indecisió quan el valor es troba en l'interval de 0,8 i 0,9. En canvi, amb el segon criteri, valors superiors a 0,85 fonamenten l'elecció d'independència, i valors inferiors a aquest límit porten a la decisió de dependència. El primer criteri es podria utilitzar en les situacions on el nombre de regions és petit, atès que les estimacions del coeficient de concordança són més variables. En canvi, quan el nombre de regions és superior, el punt de tall de 0,85 es presenta força estable en relació amb el percentatge d'error, tant en rebutjar una hipòtesi d'independència com de dependència. La virtut principal que ha presentat el coeficient de concordança respecte de la resta de proves és precisament aquesta estabilitat en el percentatge d'errors, independentment del nombre de regions, variància de la sobredispersió i nombre de casos esperats.

Aquest fet pot ser degut a la naturalesa no paramètrica d'aquest coeficient, atès que, a diferència de la resta de proves, no fa assumpcions distribucionals sobre l'estadístic de contrasts ni treballa directament sobre la versemblança de les dades. Això s'ha posat de manifest en l'exemple de les dades de diabetis a Catalunya, ja que es tractava d'una situació on el nombre de regions i el valor de la sobredispersió definien una situació en què la resta de proves es caracteritzaven per una manca de potència i, per tant, per una manca de fiabilitat en no rebutjar la hipòtesi nul·la.

Una altra prova que s'ha proposat (Vonesh *et al.*, 1996) per avaluar la bondat d'ajustament dels models lineals generalitzats mixtos quan s'ha assumit una estructura de variàncies i covariàncies, és el test *pseudo likelihood ratio*. Aquest test s'ha utilitzat àmpliament en l'àmbit d'equacions estructurals (Bentler I Bonett, 1980) per decidir entre models niats. No obstant això, aquest test es basa en l'obtenció d'una matriu de covariàncies empírica i en l'aproximació a una distribució de Wishart de les matrius de covariàncies proposades en cada model. Malauradament en el nostre cas no és possible obtenir una matriu empírica de covariàncies perquè no disposem de mesures repetides, és a dir, a cada regió disposem d'una mesura. Addicionalment, l'aproximació de les matrius a la distribució de Wishart requereix que el nombre de mesures repetides sigui superior al nombre d'efectes aleatoris, cosa que en aquest cas tampoc no es compleix.

En resum, s'ha vist que quan el nombre de regions, el nombre de casos esperats i la sobredispersió són grans, l'LRT i l'*score test* funcionen força bé tant en termes d'error de tipus I com de potència, encara que és lleugerament preferible l'*score test*. Ara bé, quan aquestes condicions no es donen i s'arriba a l'extrem d'un nombre de regions, de casos esperats i de sobredispersió petits, aquestes proves no són gaire fiables en termes de potència, i en aquest cas el coeficient de concordança és més útil per prendre decisions.

Figura 3.1.*a*

Diagrames de caixes del coeficient de concordança per $\lambda = 0$ i $\sigma^2 = 1$ o $\sigma^2 = 0{,}5$ i nombre de regions (Nre.): 49 o 100, i de casos esperats (E.): petit o gran

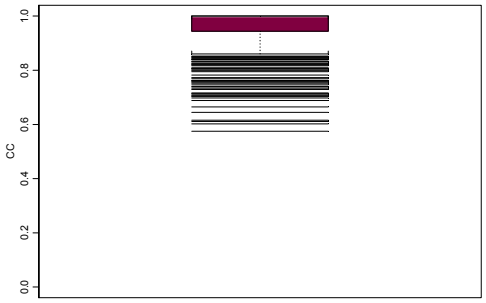
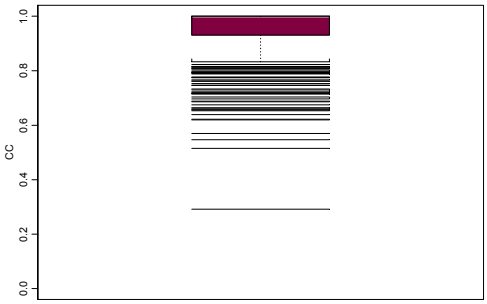
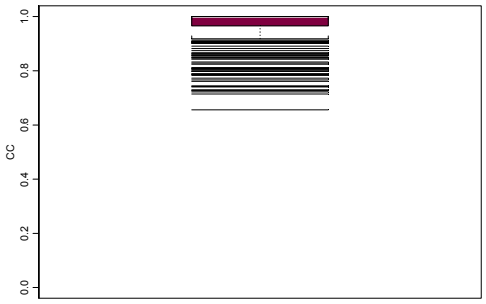
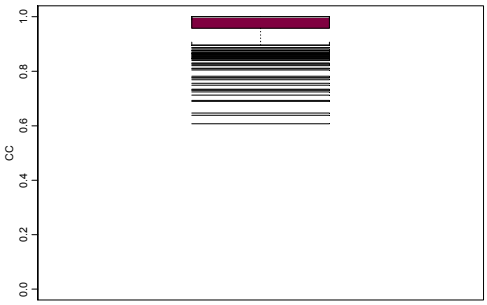
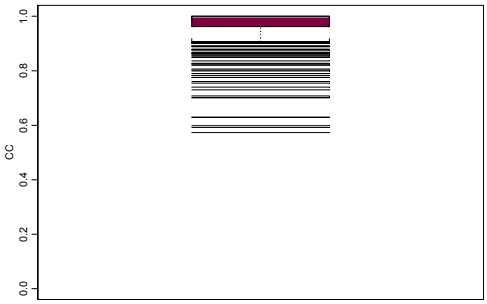
	$\lambda = 0$ i $\sigma^2 = 1$	$\lambda = 0$ i $\sigma^2 = 0{,}5$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.1.*b*

Diagrames de caixes del coeficient de concordança amb $\lambda = 0$ i $\sigma^2 = 0.25$ o $\sigma^2 = 0.09$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

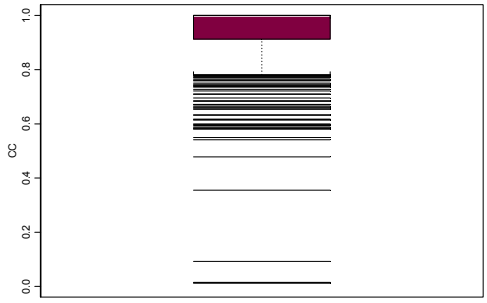
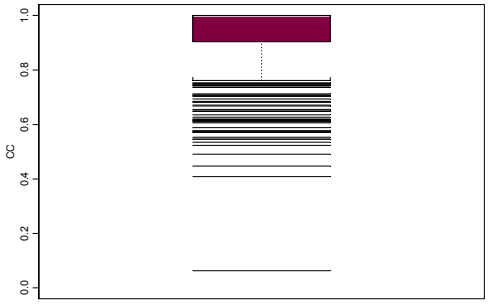
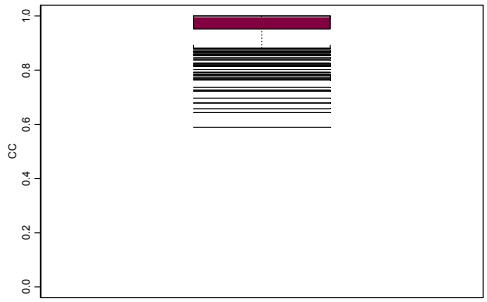
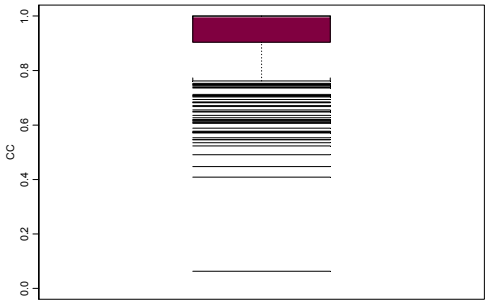
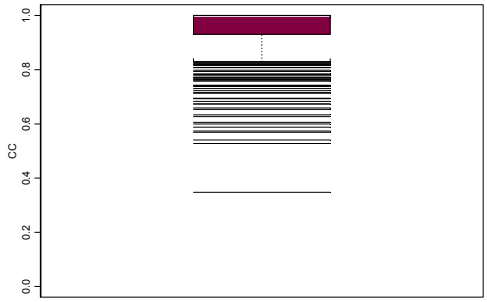
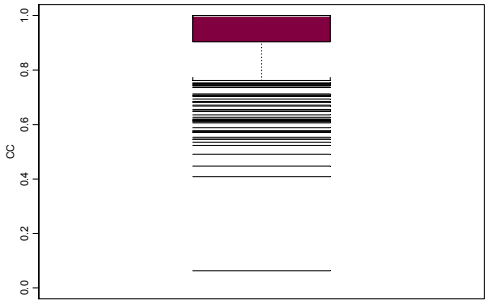
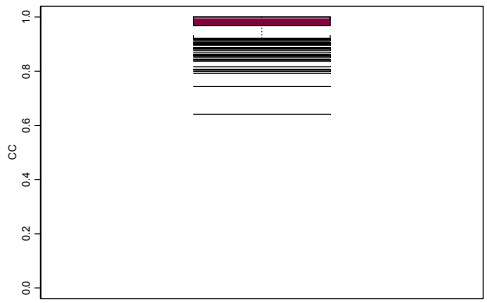
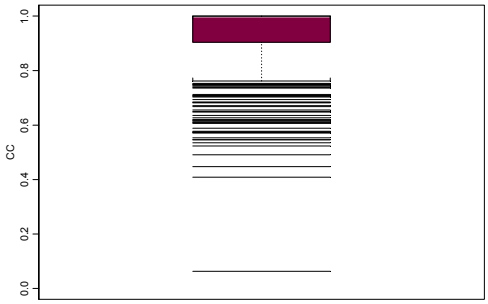
	$\lambda = 0$ i $\sigma^2 = 0.25$	$\lambda = 0$ i $\sigma^2 = 0.09$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.2.*a*

Diagrames de caixes del coeficient de concordança amb $\lambda = 0.25$ i $\sigma^2 = 1$ o $\sigma^2 = 0.5$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

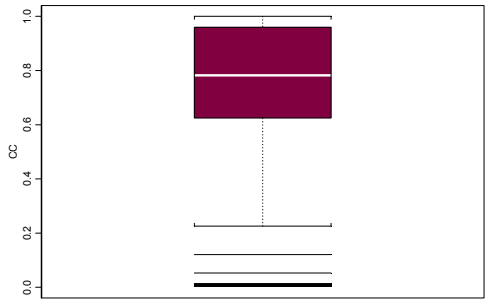
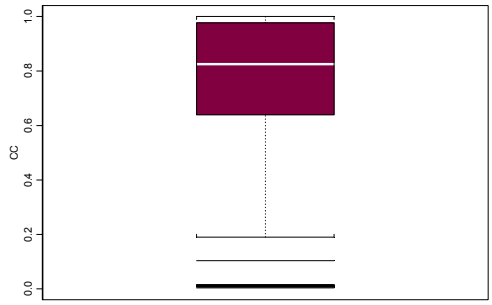
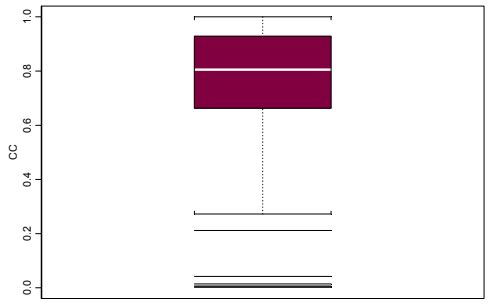
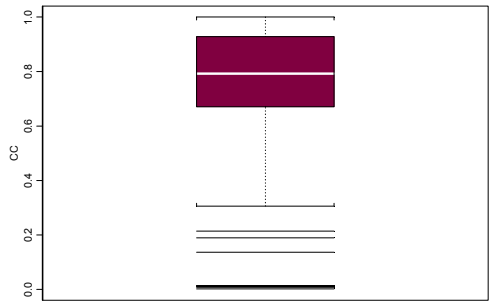
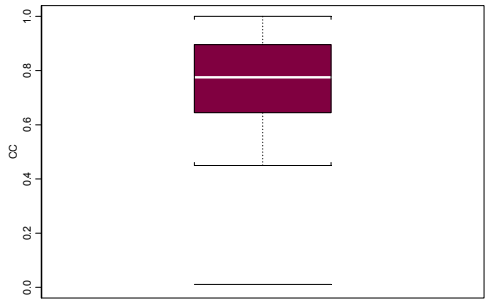
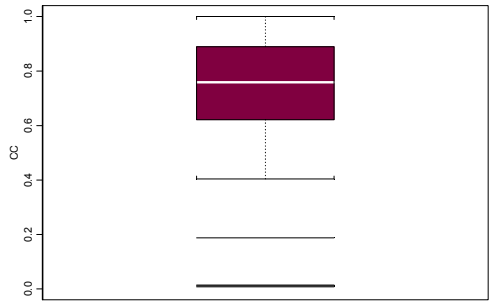
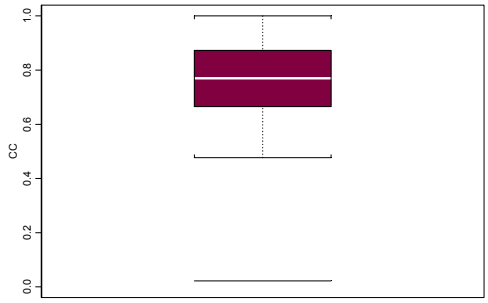
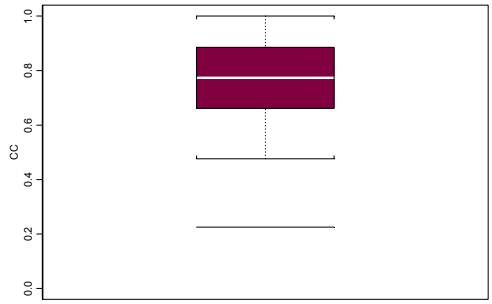
	$\lambda = 0.25$ i $\sigma^2 = 1$	$\lambda = 0.25$ i $\sigma^2 = 0.5$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.2.*b*

Diagrames de caixes del coeficient de concordança amb $\lambda = 0.25$ i $\sigma^2 = 0.25$ o $\sigma^2 = 0.09$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

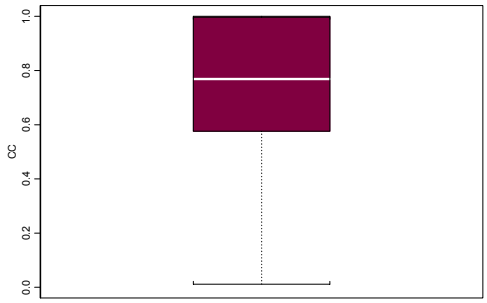
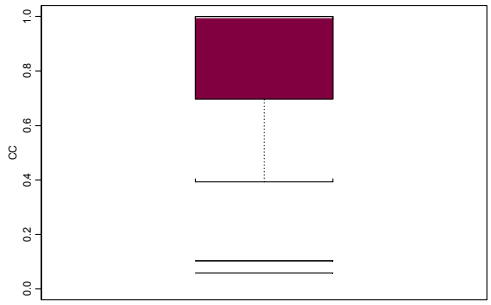
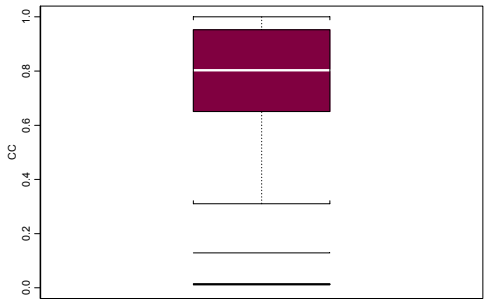
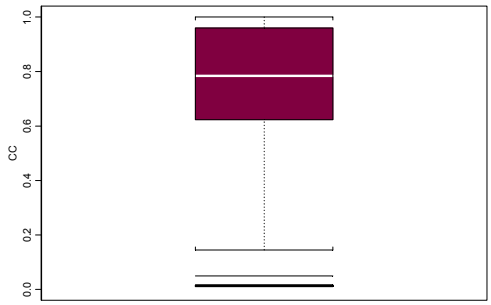
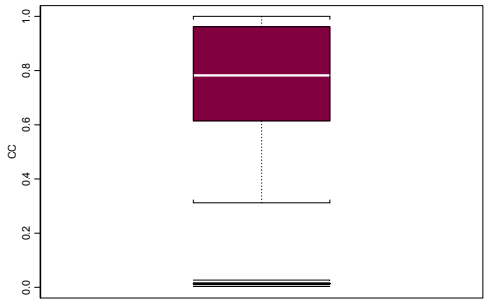
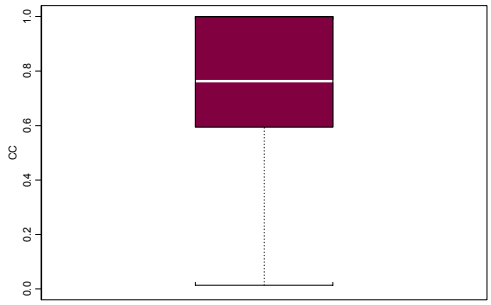
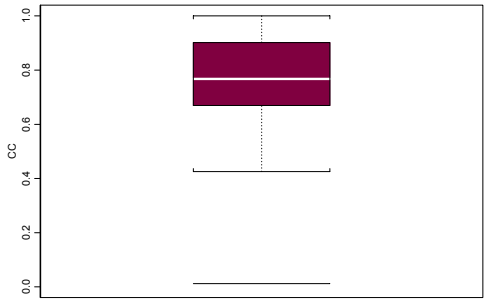
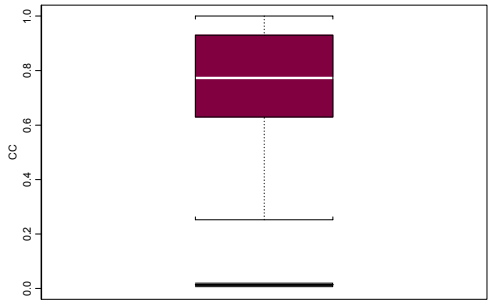
	$\lambda = 0.25$ i $\sigma^2 = 0.25$	$\lambda = 0.25$ i $\sigma^2 = 0.09$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.3.*a*

QQ-plot de normalitat per $\lambda = 0$ i $\sigma^2 = 1$ o $\sigma^2 = 0.5$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

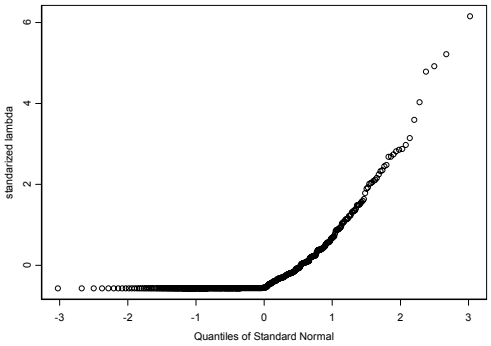
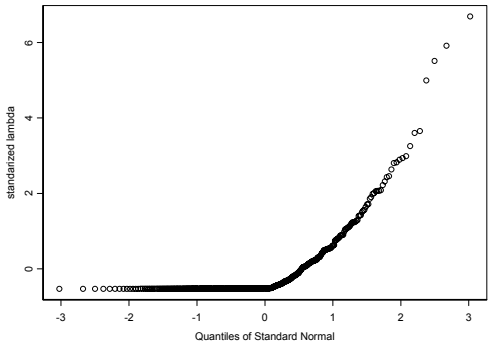
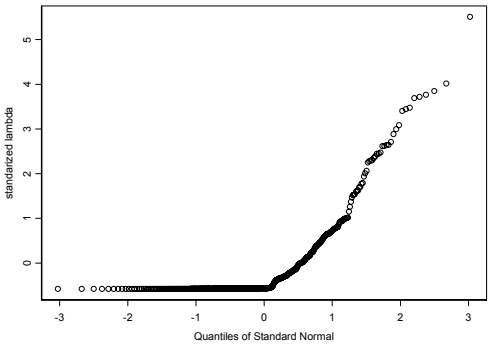
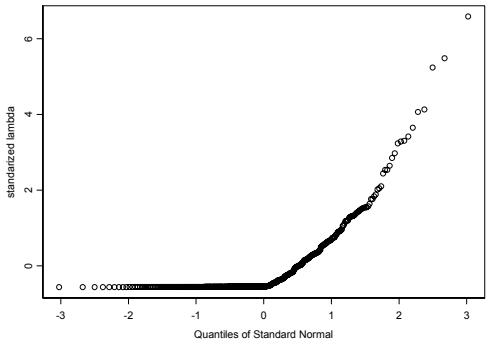
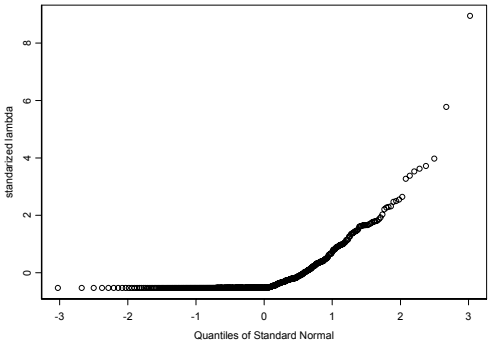
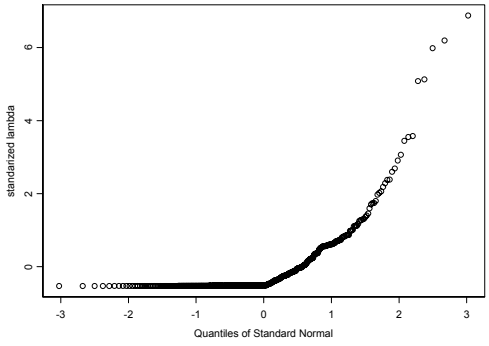
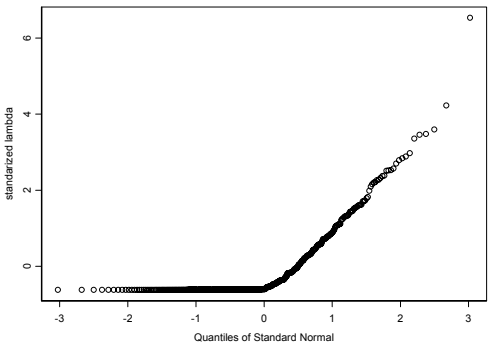
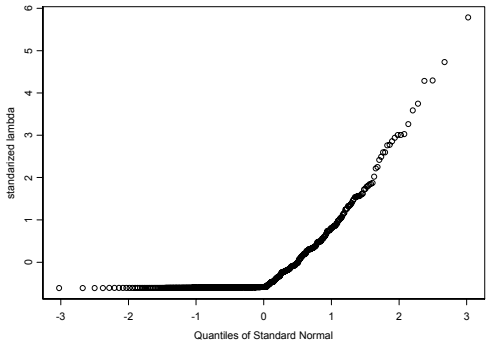
	$\lambda = 0$ i $\sigma^2 = 1$	$\lambda = 0$ i $\sigma^2 = 0.5$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.3.*b*

QQ-plot de normalitat per $\lambda = 0$ i $\sigma^2 = 0.25$ o $\sigma^2 = 0.09$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

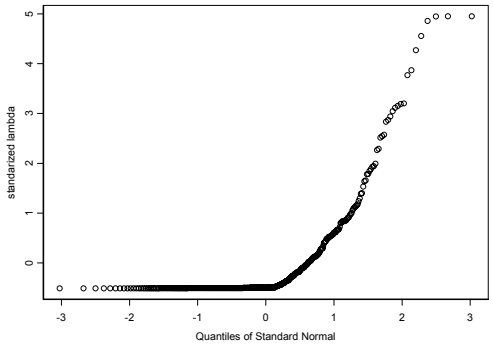
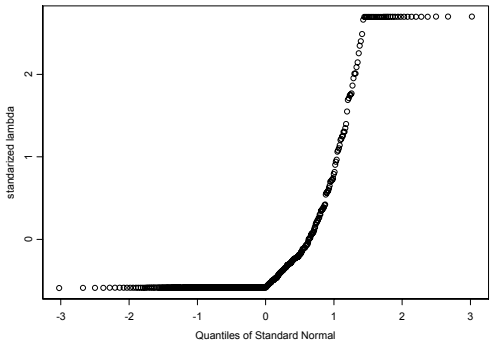
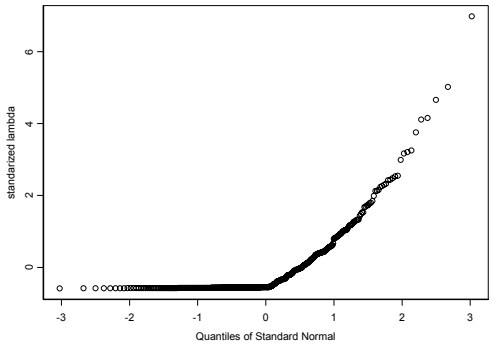
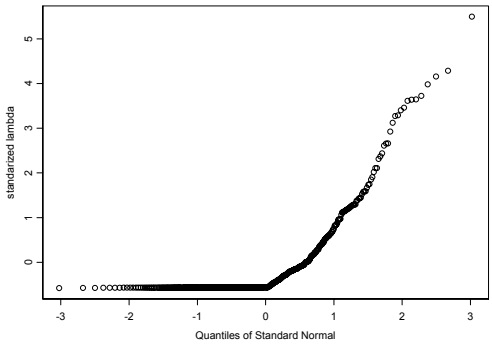
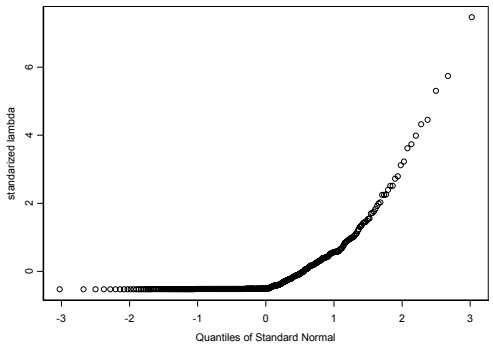
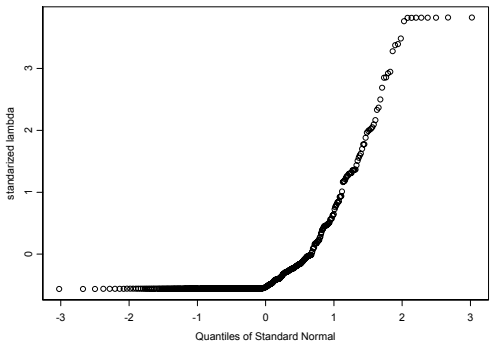
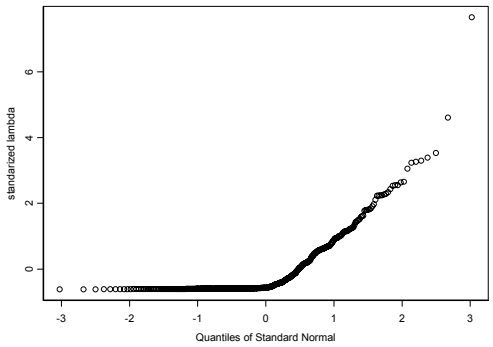
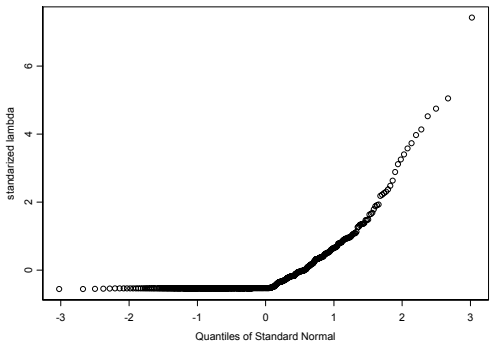
	$\lambda = 0$ i $\sigma^2 = 0.25$	$\lambda = 0$ i $\sigma^2 = 0.09$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.4.*a*

QQ-plot de normalitat per $\lambda = 0.25$ i $\sigma^2 = 1$ o $\sigma^2 = 0.5$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

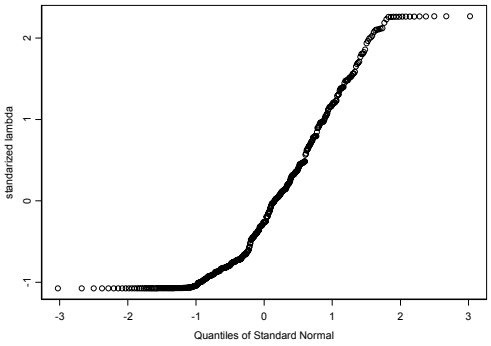
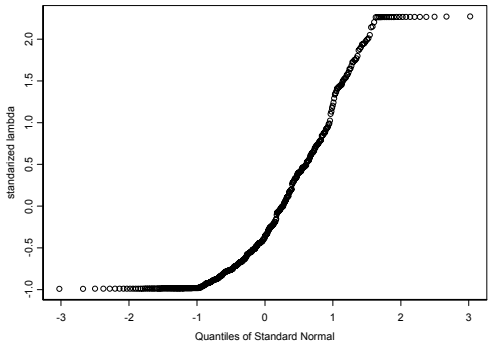
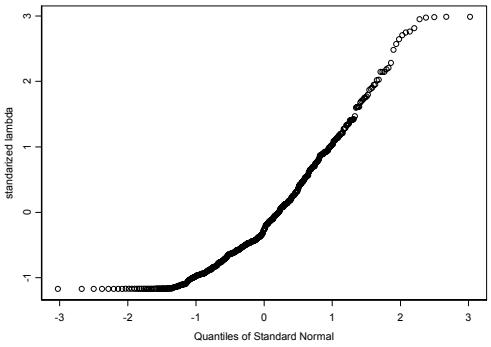
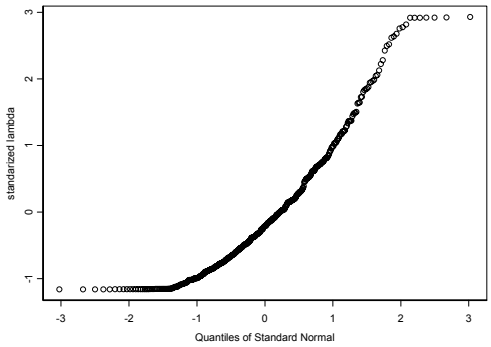
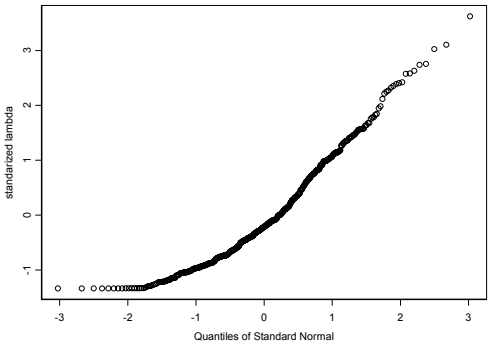
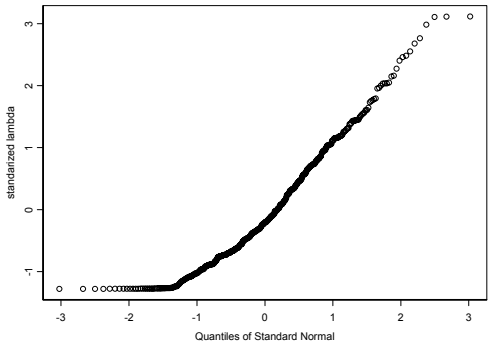
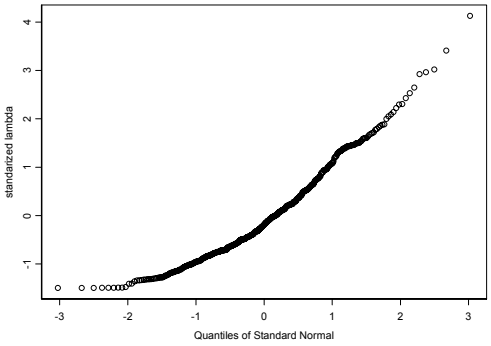
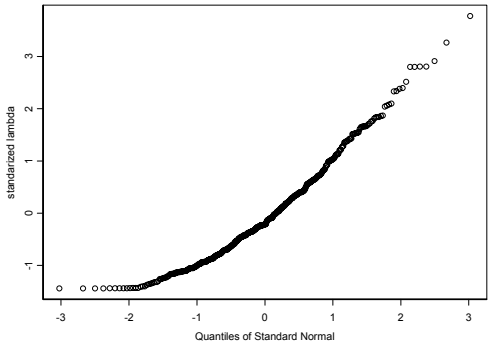
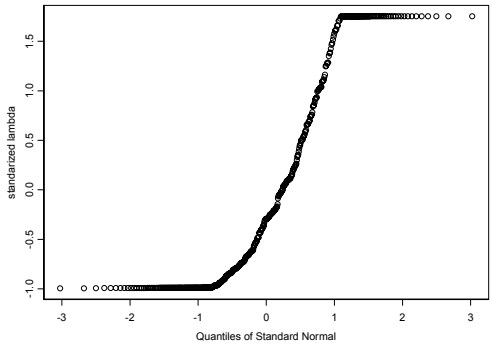
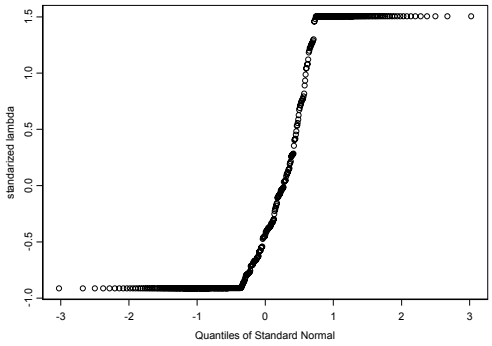
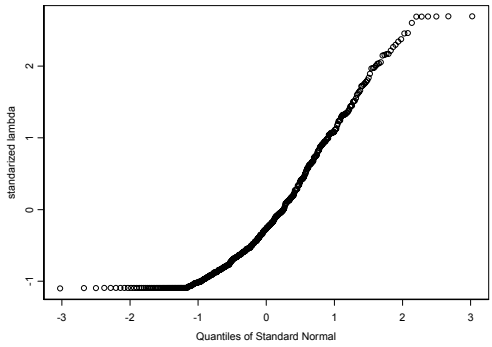
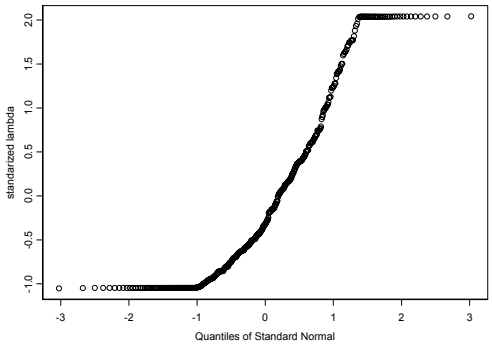
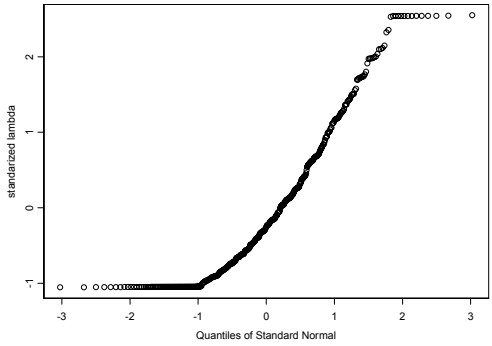
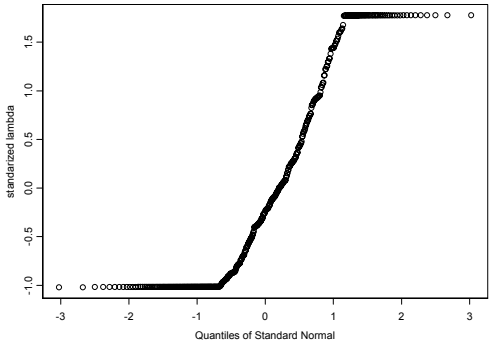
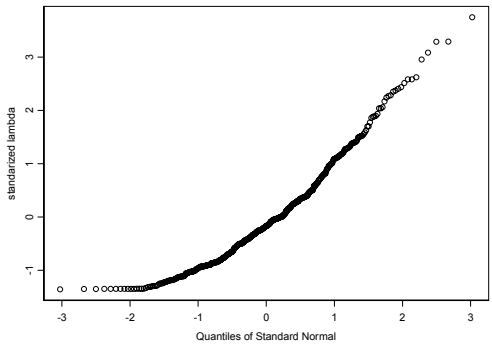
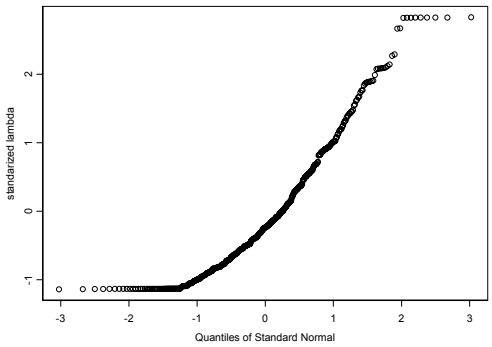
	$\lambda = 0.25$ i $\sigma^2 = 1$	$\lambda = 0.25$ i $\sigma^2 = 0.5$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

Figura 3.4.*b*

QQ-plot de normalitat per $\lambda = 0.25$ i $\sigma^2 = 0.25$ o $\sigma^2 = 0.09$ i nombre de regions (Nre.): 49 o 100 i de casos esperats (E.): petit o gran

	$\lambda = 0$ i $\sigma^2 = 0.25$	$\lambda = 0$ i $\sigma^2 = 0.09$
Nre.: 49 E.: Petit		
Nre.: 49 E.: Gran		
Nre.: 100 E.: Petit		
Nre.: 100 E.: Gran		

CAPÍTOL IV: ANÀLISI GEOGRÀFICA DE LA INCIDÈNCIA DE LA DIABETIS TIPUS I A CATALUNYA EN EL PERÍODE 1989-1998

4.1. Introducció

En aquest capítol es durà a terme una anàlisi geogràfica del risc de la diabetis de tipus I a Catalunya aplicant la metodologia que s'ha anat definint al llarg de la tesi.

La diabetis de tipus I és una de les malalties cròniques d'aparició a l'edat infantil i juvenil que més morbilitat comporta a les poblacions dels països occidentals (Krans *et al*., 1992). Es tracta d'una malaltia que és la causant principal de ceguesa, d'insuficiència renal terminal i d'amputacions d'extremitats inferiors en aquests països, a més, és un dels factors de risc cardiovascular més rellevants. Tot plegat fa que la diabetis de tipus I sigui la causa d'un deteriorament molt important de la qualitat de vida de les persones que la pateixen i d'un cost sanitari elevat.

En diferents estudis s'ha demostrat l'existència d'una distribució geogràfica dels nous casos de diabetis de tipus I (Songini *et al*., 1998; Rytkönen *et al*., 2001; Green *et al*., 2003). Així doncs, l'objectiu d'aquest capítol és analitzar si els casos de diabetis de tipus I a Catalunya també estan associats a un patró geogràfic. Així mateix, també es vol comprovar si el risc és homogeni respecte del gènere i de l'edat dels individus.

4.2. Material i mètodes

Els individus d'estudi són els casos declarats i confirmats en el registre de diabetis de tipus I a Catalunya entre els anys 1989 i 1998. La població de risc menor de 30 anys s'ha obtingut del padró poblacional de l'any 1996. Aquesta població es pot considerar representativa de tot el període, atès que Catalunya a la dècada dels noranta tenia un creixement de la població proper a 0.

Les variables recollides pels individus diagnosticats com a diabètics són el lloc de residència, l'edat i el gènere. Las unitats geogràfiques estudiades són les 41 comarques en què es divideix Catalunya, i per a cadascuna d'aquestes unitats s'obté el nombre de casos de diabetis de tipus I.

La variable d'estudi per a cada comarca en funció del gènere i l'edat és la taxa d'incidència de la malaltia.

El mètode d'estimació utilitzat per modelar les taxes d'incidència de la malaltia ha estat la quasiversemblança penalitzada utilitzant les transformacions dels components de la variància presentades en el capítol 2. Per testar l'homogeneïtat de la taxa respecte del gènere i de l'edat s'ha emprat el test F de Wald. L'avaluació de la hipòtesi d'independència espacial s'ha fet mitjançant el quocient de les versemblances, l'*score test* i el coeficient de concordança.

Finalment, es representen les raons estandarditzades de morbilitat (SMR) brutes i les que s'han obtingut a partir dels models mitjançant els mapes que s'han creat amb el programa Arc View 3.2. Les SMR brutes es calculen dividint la taxa bruta de cada gènere, edat i comarca per la taxa bruta global de Catalunya, i les SMR estimades pel model s'obtenen dividint les taxes estimades per l'estimació de la taxa mitjana.

4.3. Resultats

En les taules 4.1 i 4.2 es presenten les freqüències de casos de diabetis i les taxes brutes d'incidència per gènere i edat del diagnòstic. En total, s'han observat 2.771 casos, que representen una taxa d'incidència d'11,8 casos de diabetis per cada 100.000 habitants i any. En la figura 4.1 es representen el nombre de casos per any d'estudi, on es pot comprovar que els casos de diabetis han variat poc d'any a any.

Taula 4.1
Nombre de casos de diabetis per edat i gènere

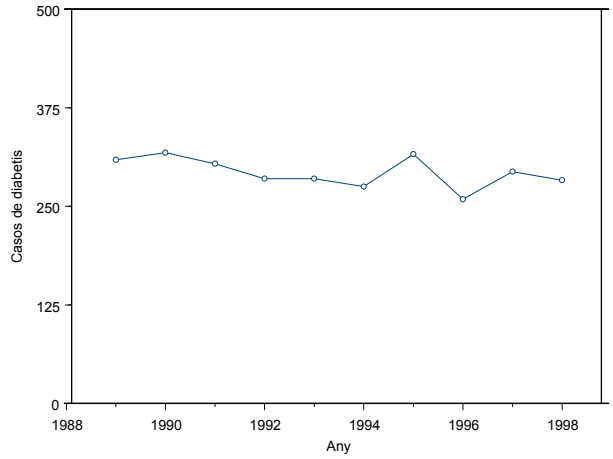
	Edat		
Gènere	≤ 14 anys	> 14 anys	Total
Homes	653	948	1.601 (57,80%)
Dones	631	539	1.170 (42,20%)
	1.284 (46,30%)	1.487 (53,70%)	2.771

Taula 4.2

Taxes d'incidència brutes de diabetis per edat i gènere per cada 100.000 habitants i any

	Edat		
Gènere	≤ 14 anys	> 14 anys	Total
Homes	14,33	12,83	13,41
Dones	14,44	7,53	10,15
	14,39	10,22	11,80

Figura 4.1
Nombre de casos de diabètics per any d'estudi



En la figura 4.2 es mostra la distribució per tota la població de les SMR brutes per a cadascuna de les comarques estudiades. Aquestes SMR s'han classificat en les quatre categories següents: taxes clarament inferiors a la taxa mitjana, que serien les SMR amb un valor inferior a 0,75; taxes al voltant de la mitjana però inferiors a aquesta, categoria definida per SMR entre 0,75 i 1; taxes al voltant de la mitjana però superiors a aquesta, definides per SMR entre 1 i 1,25, i finalment taxes notablement superiors a la mitjana, definides per SMR superiors a 1,25.

Es pot observar que la majoria de comarques que presenten una SMR inferior a la mitjana estan ubicades al nord i al sud de Catalunya. A més, solament hi ha quatre comarques amb una SMR superior a 1,25, que són el Segrià, les Garrigues, el Bages i el Vallès Oriental.

En les figures 4.3, 4.4, 4.5 i 4.6 es presenten les SMR brutes de les comarques en funció del gènere i de l'edat. A partir d'aquestes representacions es pot observar que el comportament de les SMR en totes les comarques és diferent en funció d'aquestes variables, per la qual cosa es podria plantejar l'existència de patrons geogràfics diferents segons les combinacions de gènere i edat.

Figura 4.2

SMR brutes de diabetis de tipus I de les comarques catalanes

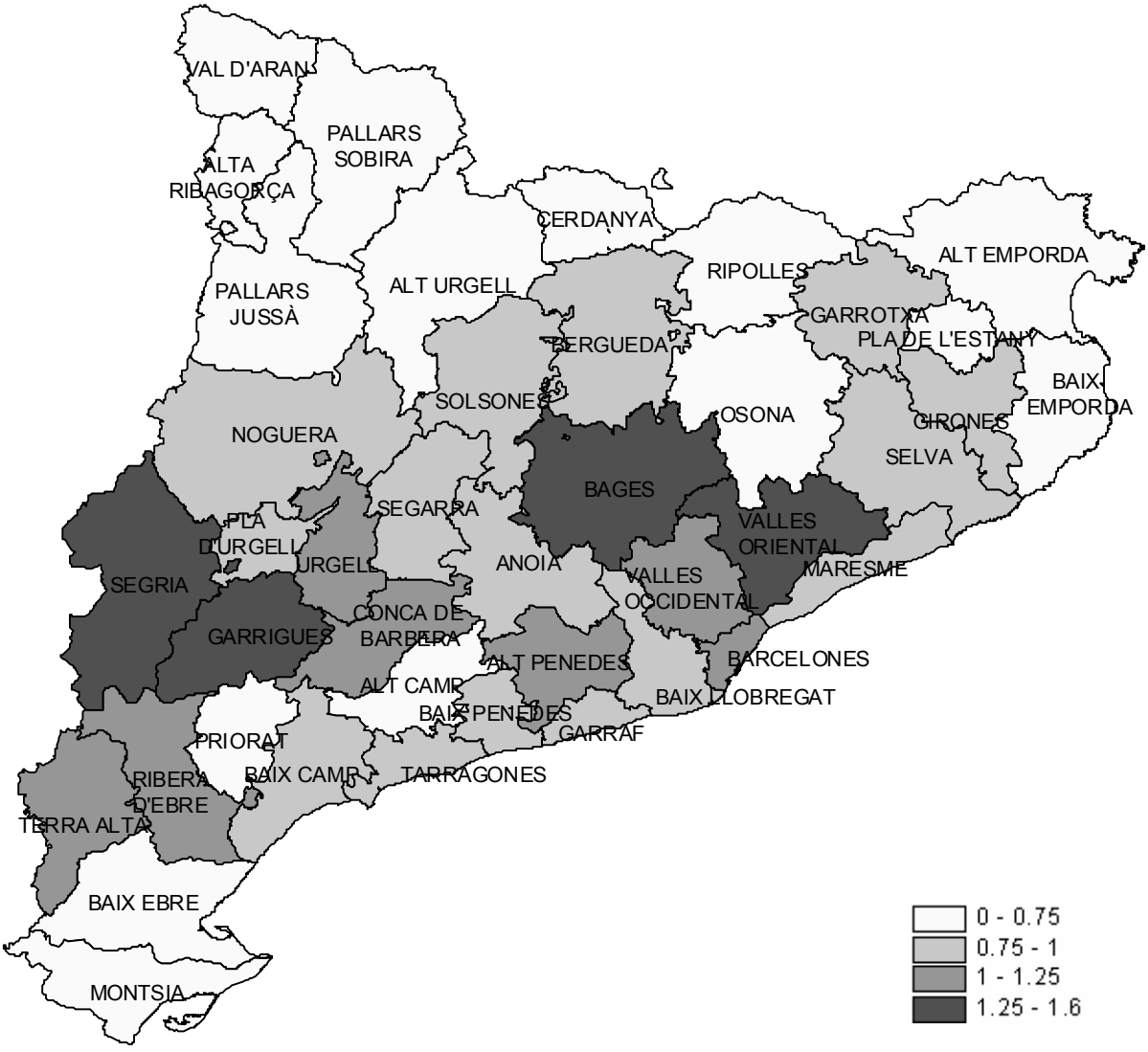


Figura 4.3
SMR brutes de diabetis de tipus I de les comarques catalanes. Dones de ≤ 14 anys

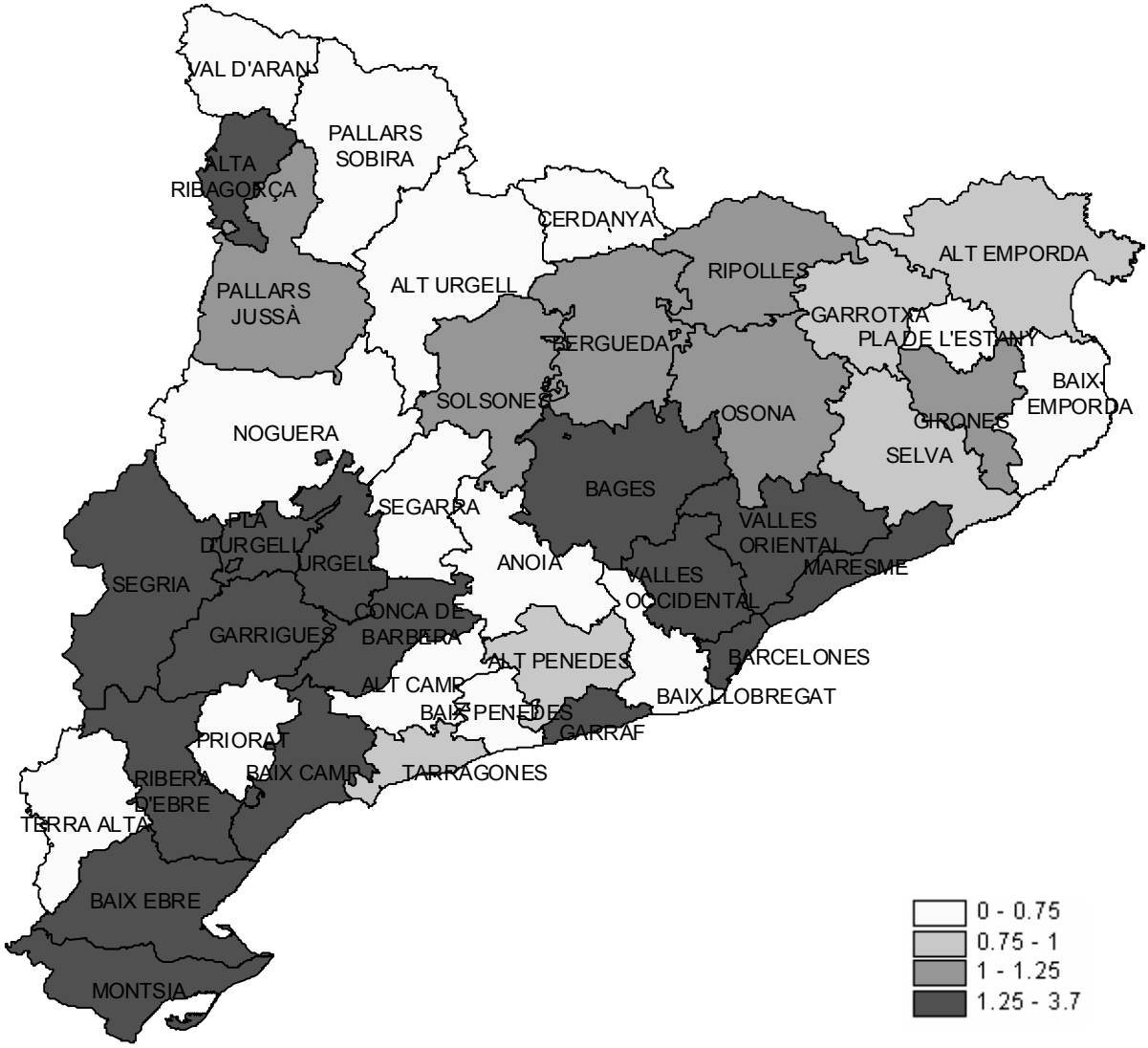


Figura 4.4
SMR brutes de diabetis de tipus I de les comarques catalanes. Dones > 14 anys

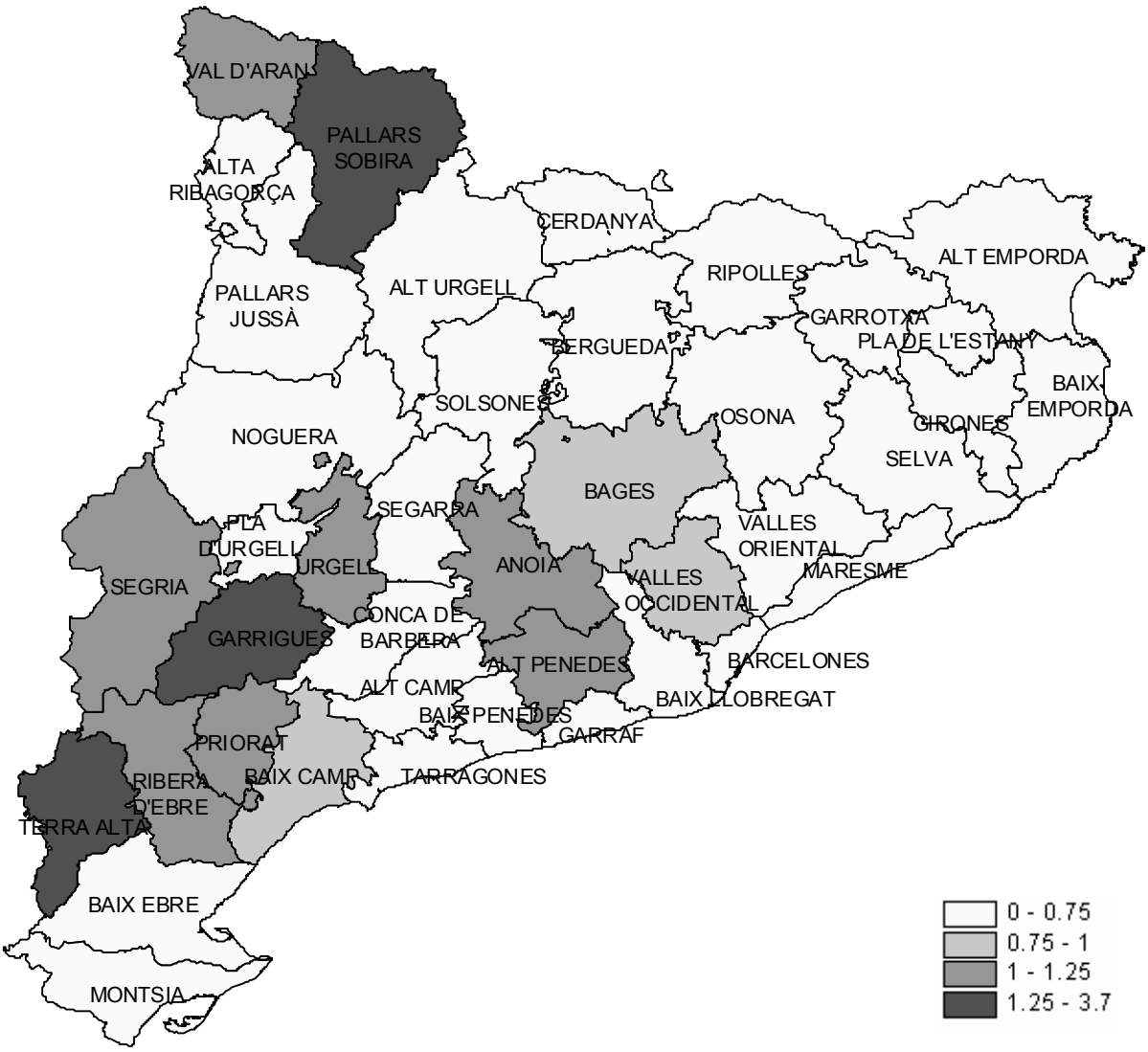


Figura 4.5

SMR brutes de diabetis de tipus I de les comarques catalanes. Homes de ≤ 14 anys

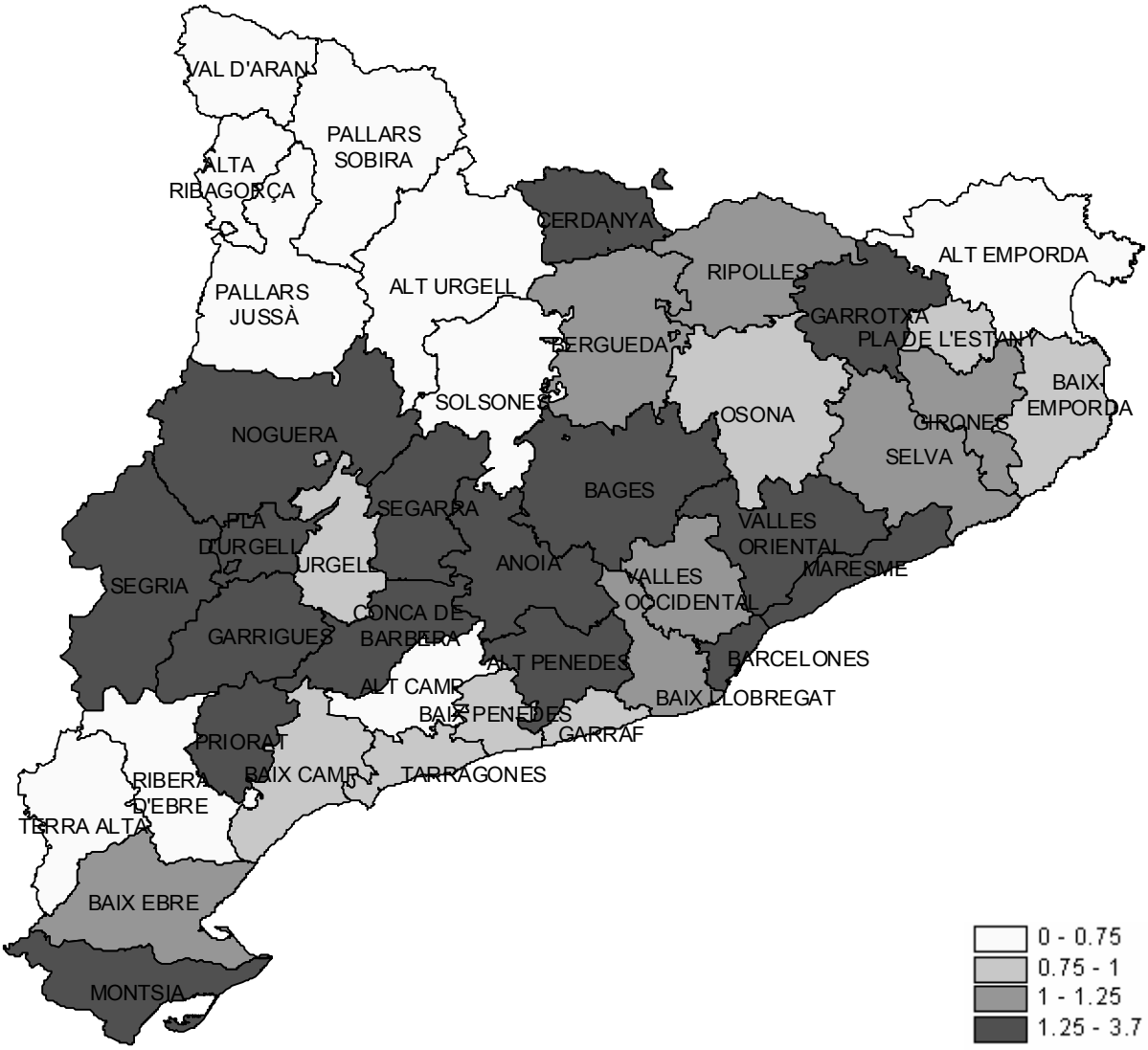
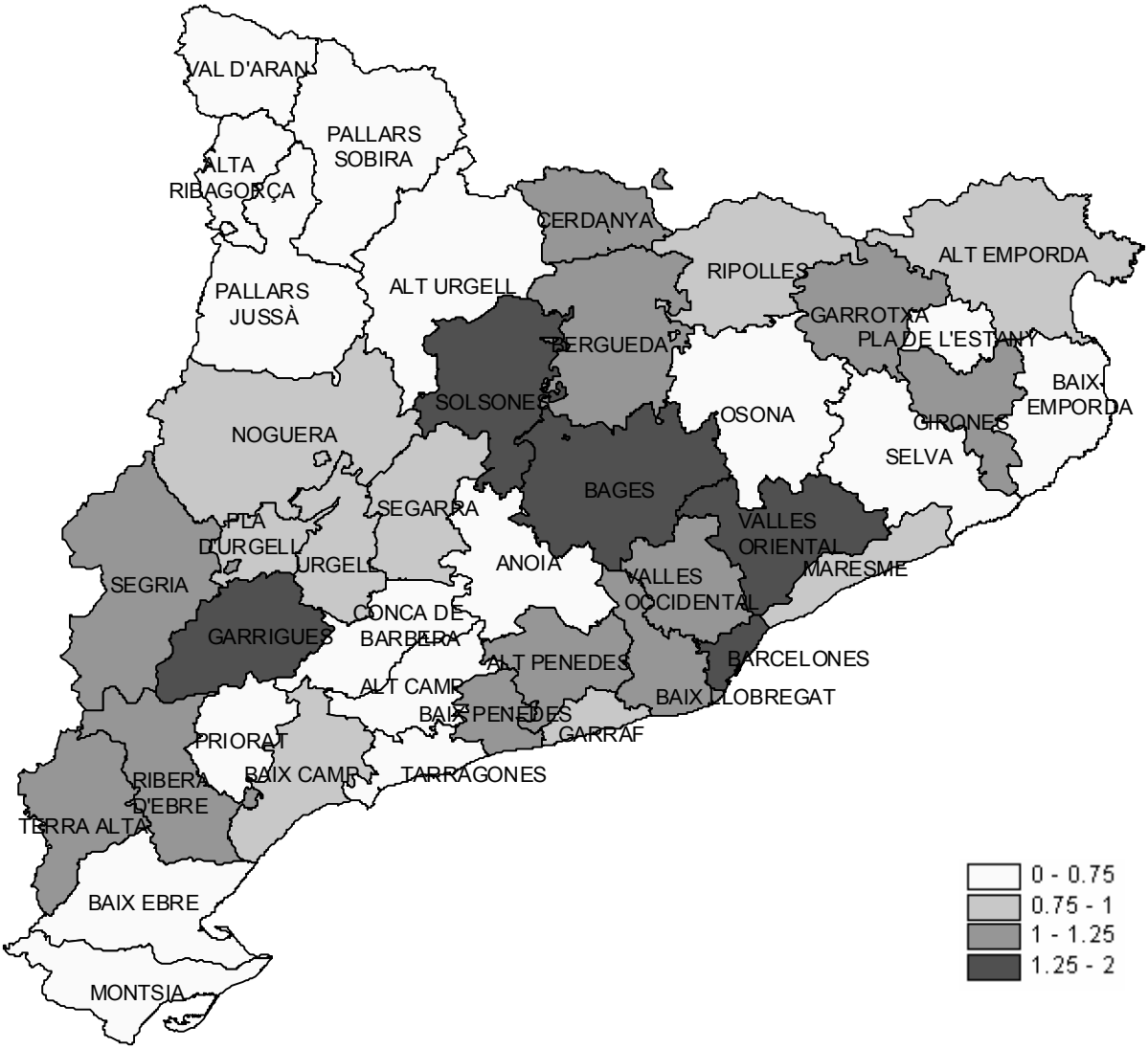


Figura 4.6
SMR brutes de diabetis de tipus I de les comarques catalanes. Homes > 14 anys



Per analitzar les taxes d'incidència de la diabetis de tipus I a Catalunya, en primer lloc es fita un model lineal generalitzat amb el gènere i l'edat com a covariables i el logaritme com a funció d'enllaç. Es procedeix d'aquesta manera perquè si en fitar el model lineal generalitzat no s'observa sobredispersió significarà que no caldrà buscar fonts de sobredispersió ni fitar el model lineal generalitzat mixt. Les covariables s'han categoritzat considerant com a categories basals les dones i els majors de 14 anys. Les estimacions d'aquest model es presenten en la taula 4.3.

Taula 4.3
Resultats del model lineal generalitzat. Estimacions de les covariables i del paràmetre de sobredispersió*

Intercept	Gènere	Edat	Gènere × Edat	Sobredispersió
–9,037	0,1313	0,1905	–0,1349	2,018
(0,019)	(0,019)	(0,019)	(0,019)	

*L'error estàndard de les estimacions es presenta entre parèntesis.

El paràmetre de sobredispersió és de 2,018, per tant ens trobem en una situació de sobredispersió on la variabilitat de les dades és el doble que la variabilitat assumida pel model. Aleshores, és pertinent fer les estimacions amb un model que tingui en compte aquesta sobredispersió. Així es fita un model lineal generalitzat mixt incorporant-hi l'efecte aleatori *comarca* mitjançant els models CAR no intrínsec i d'heterogeneïtat (taules 4.4 i 4.5). Per decidir entre els dos models s'efectuen les proves d'independència espacial, els resultats de les quals es mostren en la taula 4.6.

Taula 4.4
Resultats del model lineal generalitzat mixt amb l'efecte aleatori *comarca* amb estructura de CAR no intrínsec

	Paràmetres						
	Intercept	Gènere	Edat	Gènere × Edat	σ	λ	Log REML
Estimació	-9,643	0,525	0,658	-0,533	0,338	0,407	-157,2
Error estàndard	0,088	0,054	0,059	0,078	0,010	0,46	
Valor P test F-Wald		<0,001	<0,001	<0,001			

σ : desviació típica de l'efecte *comarca*; λ : pes espacial, i Log REML fa referència al logaritme de la versemblança restringida dels efectes aleatoris.

Taula 4.5
Resultats del model lineal generalitzat mixt amb l'efecte aleatori *comarca* amb estructura d'heterogeneïtat

	Paràmetres					
	Intercept	Gènere	Edat	Gènere × Edat	σ	Log REML
Estimació	-9,627	0,526	0,657	-0,534	0,232	-157,9
Error estàndard	0,065	0,054	0,059	0,078	0,046	
Valor P test F-Wald		<0,001	<0,001	<0,001		

σ: desviació típica de l'efecte comarca, i Log REML: fa referència al logaritme de la versemblança restringida dels efectes aleatoris.

Taula 4.6
Resultats de les proves d'independència espacial pel model amb l'efecte aleatori *comarca*

	Proves d'independència espacial		
	Score test	Quocient de versemblances	Coeficient de concordança
Estadístic	0,508	1,4	0,7964
P-valor	0,305	0,237	

En tots dos models el tests F de Wald pels efectes fixos són significatius (taules 4.4 i 4.5); per tant, el comportament de la taxa global de Catalunya és diferent segons les combinacions de gènere i edat.

Pel que fa al component espacial, les proves del quocient de versemblances i l'*score test* no rebutgen la hipòtesi nul·la d'independència espacial. No obstant això, cal dir que ens trobem davant d'una situació de 41 regions, de 67,59 casos esperats i d'una variància dels efectes aleatoris al voltant de 0,10 ($0{,}338^2 = 0{,}114$). En el capítol III es van analitzar les potències de les proves d'independència en una combinació semblant a la que es dóna aquí, amb 49 regions, un nombre de casos esperats gran i una variància de 0,09, i es va comprovar que les potències eren baixes. Per tant, ens podríem trobar en un cas de manca de potència per rebutjar la hipòtesi d'independència espacial. D'altra banda, el valor del coeficient de concordança està per sota de 0,85 i també de 0,8, per la qual cosa la decisió seria de dependència espacial. En resum, ens trobem davant d'una situació on és complicat decidir quin model triem.

Un aspecte que cal no passar per alt és el fet que en els mapes de les SMR brutes per gènere i edat de diagnòstic (figures 4.3, 4.4, 4.5 i 4.6) el comportament espacial era força diferent. Per tant, es podria plantejar la hipòtesi de si la distribució geogràfica de la diabetis a Catalunya és diferent segons el gènere i l'edat, és a dir, si hi ha interacció entre el gènere, l'edat i l'efecte aleatori *comarca*.

En la taula 4.7 es presenten les estimacions del model lineal generalitzat mixt amb l'efecte aleatori amb estructura de CAR, interaccionant amb el gènere i l'edat i el coeficient de concordança entre les matrius de correlacions dels efectes aleatoris i la matriu identitat. En la taula 4.8 es mostren els resultats de les proves d'independència; cal dir que en aquest cas, atès que s'està testant més d'un component de la variància, s'utilitza l'*score test* generalitzat (Lin, 1997). El quocient de versemblances està comparant el model amb l'efecte aleatori *comarca* amb estructura de CAR no intrínsec respecte del model on l'efecte aleatori *comarca* amb una estructura de CAR intrínsec interacciona amb el gènere i l'edat.

Taula 4.7

Resultats del model lineal generalitzat mixt amb l'efecte aleatori *comarca* amb estructura de CAR no intrínsec i interaccionant amb el gènere i l'edat

Estrat		σ	λ	Coeficient de concordança
Dones	Estimació	0,251	0,103	0,954
Edat ≤ 14 anys	Error estàndard	0,168	0,512	
Dones	Estimació	0,586	0,996	0,014
Edat > 14 anys	Error estàndard	0,155	0,293	
Homes	Estimació	0,245	0,8336	0,628
Edat ≤ 14 anys	Error estàndard	0,109	0,841	
Homes	Estimació	0,371	0,002	1,000
Edat > 14 anys	Error estàndard	0,150	0,145	
Log REML	–146,1			

σ : desviació típica de l'efecte *comarca*; λ : pes espacial, i Log REML fa referència al logaritme de la versemblança restringida dels efectes aleatoris.

Taula 4.8
Resultats de les proves d'independència espaial. El quocient de versemblances es realitza entre el model CAR intrínsec amb un efecte aleatori *comarca* i el model CAR intrínsec on l'efecte aleatori *comarca* interacciona amb el gènere, l'edat i la seva interacció

	Proves d'independència espacial	
	Score test	Quocient de versemblances
Estadístic	10,133	22,2
P-valor	0,0191	0,0011

Segons les proves d'independència espacial ara es rebutjaria la hipòtesi nul·la, és a dir, l'ajustament del model s'ha millorat amb la interacció entre l'efecte aleatori *comarca* amb el gènere i l'edat.

L'heterogeneïtat dels patrons espacials segons el gènere i l'edat es pot comprovar observant que les estimacions puntuals de les correlacions espacials són molt diferents, i ens trobem amb dues correlacions molt elevades, amb coeficients de concordança baixos, i amb dues correlacions baixes on els coeficients de concordança són clarament superiors a 0,9. Basant-nos en aquests resultats, es podria argumentar si en els estrats on el component espacial és baix hi ha en realitat independència espacial. Per verificar aquesta hipòtesi es fa de nou l'estimació fixant un component espacial de 0 per als estrats "Dones ≤ 14 anys" i "Homes > 14 anys". Els resultats d'aquest model es mostren en la taula 4.9. En la taula 4.10 es mostren els resultats de l'*score test* i del quocient de versemblances. L'*score test* testa si els dos components espacials considerats són 0, mentre que el quocient de versemblances compara el model amb un component espacial per cada estrat respecte del model amb dos components espacials, un per l'estrat "Dones diagnosticades després dels 14 anys" i "Homes diagnosticats abans dels 14 anys".

Taula 4.9

Resultats del model lineal generalitzat mixt amb l'efecte aleatori *comarca* amb estructura de CAR no intrínsec. Només es considera el component espacial per als estrats "Dones > 14 anys" i "Homes ≤ 14 anys"

Estrat		σ	λ
Dones	Estimació	0,214	---
Edat ≤ 14 anys	Error estàndard	0,073	---
Dones	Estimació	0,586	0,996
Edat > 14 anys	Error estàndard	0,155	0,293
Homes	Estimació	0,245	0,8336
Edat ≤ 14 anys	Error estàndard	0,109	0,841
Homes	Estimació	0,4087	---
Edat > 14 anys	Error estàndard	0,084	---
Log REML	–146,6		

σ: desviació típica de l'efecte *comarca*; λ: pes espacial, i Log REML fa referència al logaritme de la versemblança restringida dels efectes aleatoris.

Taula 4.10

Resultats de les proves *score test* i quocient de versemblances. L'*score test* comprova la hipòtesi de si els components espacials dels estrats "Dones > 14 anys" i "Homes ≤ 14 anys" són igual a 0. El quocient de versemblances compara el model amb un component espacial per a cada estrat respecte del model amb dos components espacials

	Score test	Quocient de versemblances
Estadístic	4,036	1
P-valor	0,0084	0,9098

El resultat de l'*score test* és significatiu a un nivell d'error de tipus I del 5%, per la qual cosa es rebutja la hipòtesi d'independència espacial, és a dir, es rebutja que els dos components espacials considerats siguin 0. D'altra banda, el quocient de versemblances indica que el fet que els quatre estrats tinguin component espacial no millora l'ajustament respecte de la situació que només dos estrats tinguin aquest component.

Per tant, el model definitiu és compost pels efectes fixos gènere, edat i la seva interacció, amb una estructura per a l'efecte aleatori *comarca* definida en la taula 4.9.

En resum, hi ha un component espacial en les dones amb edat superior a 14 anys i en els homes amb una edat inferior o igual als 14 anys. En canvi, en els altres dos estrats tan sols hi ha un component de sobredispersió no estructurat.

En la taula 4.11 es mostren els riscos relatius respecte als efectes fixos, prenent com a categoria de referència les dones amb una edat de diagnòstic superior als 14 anys.

Taula 4.11
Riscos relatius segons l'edat i el sexe

	Edat en el diagnòstic	
	≤14 anys	>14 anys
Homes	1,83	1,44
Dones	1,94	1

Com es pot observar en la taula 4.11, l'estrat que té un risc superior de patir diabetis de tipus I són les dones amb menys de 14 anys, seguit dels homes amb menys de 14 anys. La interacció es manifesta en el fet que la diferència de risc respecte de l'edat no és la mateixa en els homes que en les dones.

En les figures 4.7, 4.8, 4.9 i 4.10 es representen els mapes de les SMR estimades amb el model definitiu.

Figura 4.7

SMR per comarca estimades en les dones amb una edat ≤ 14 anys

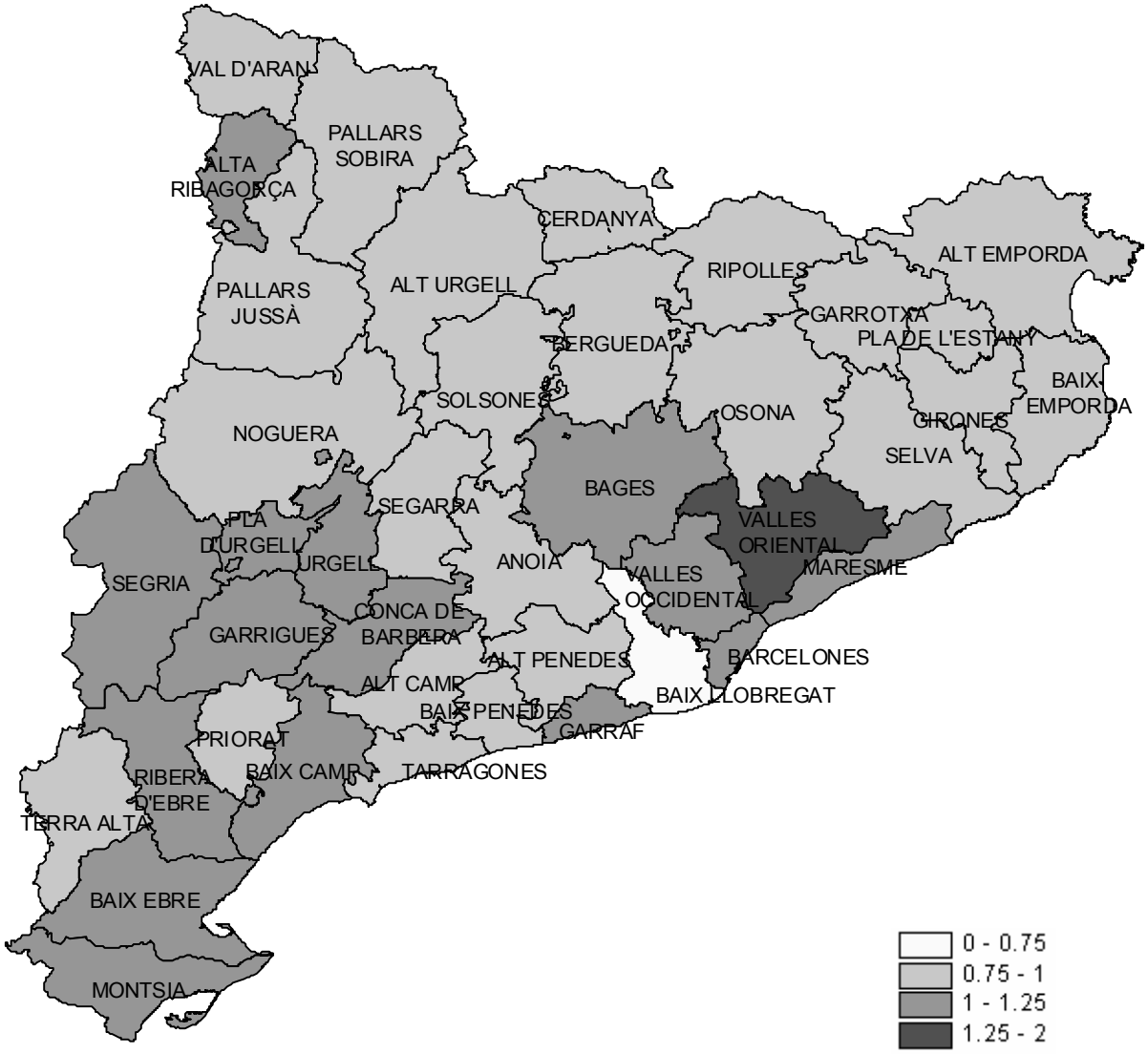


Figura 4.8

SMR per comarca estimades en les dones amb una edat > 14 anys

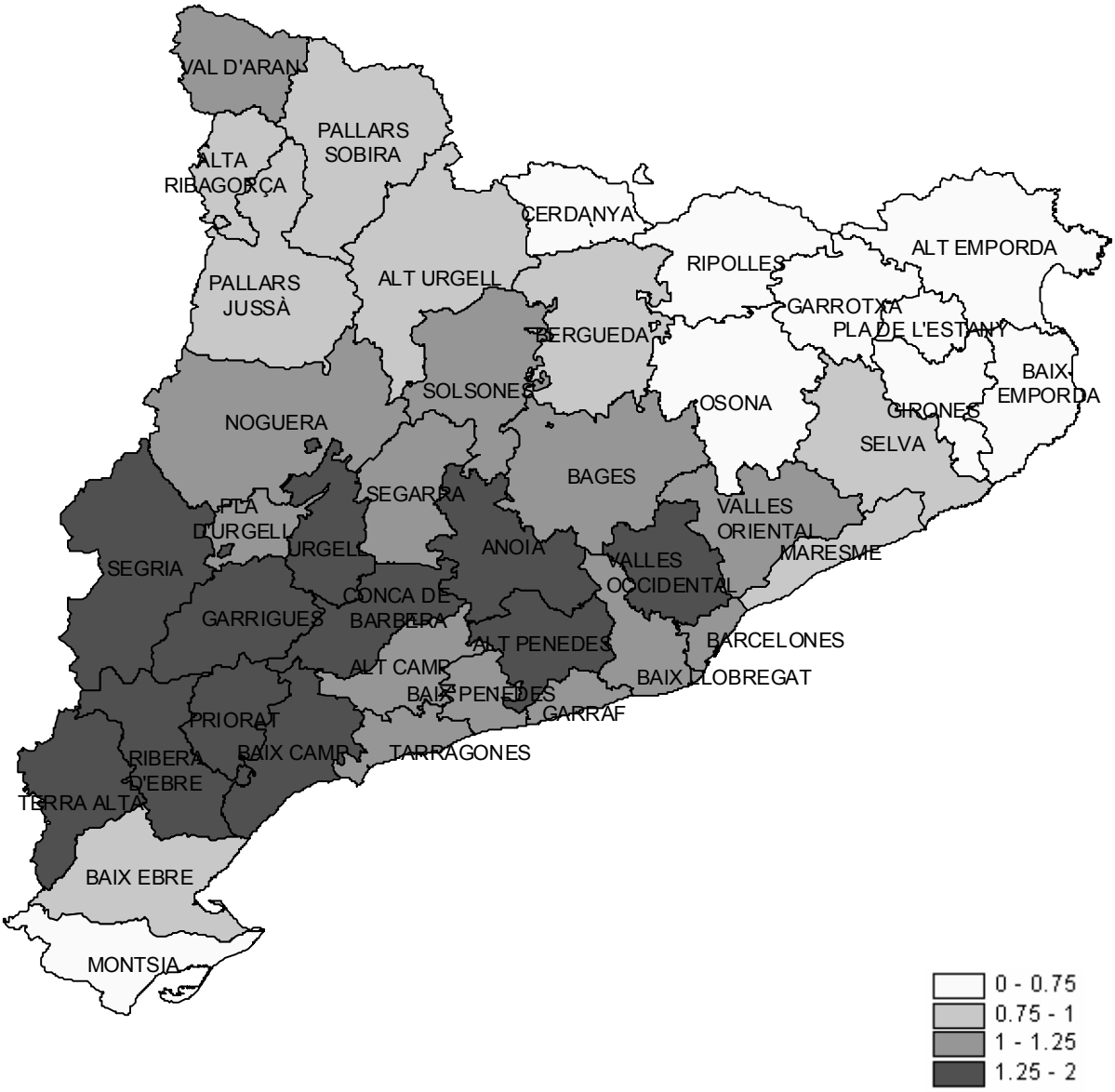
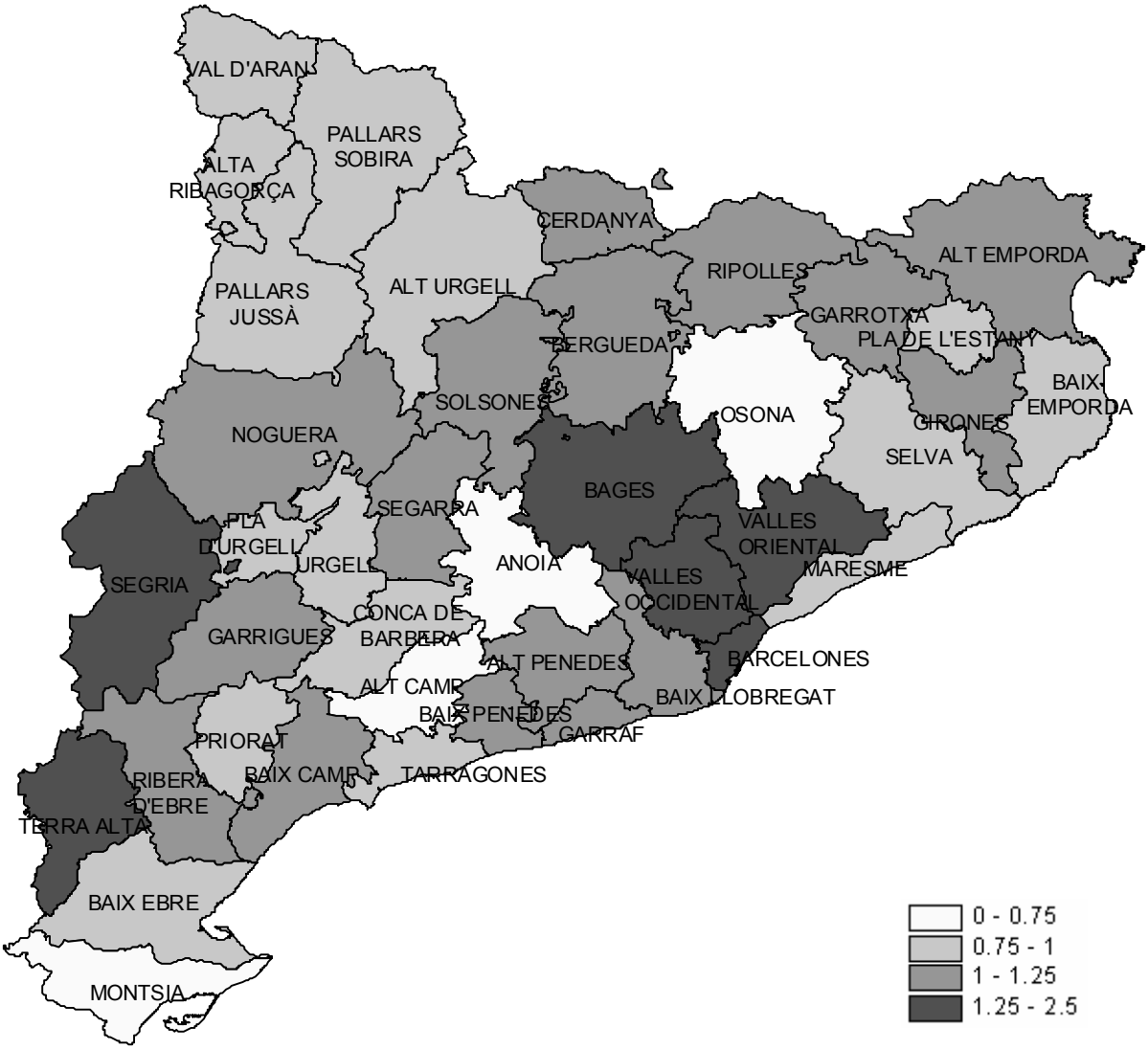


Figura 4.9

SMR per comarca estimades en els homes amb una edat ≤ 14 anys

Figura 4.10

SMR per comarca estimades en els homes amb una edat > 14 anys



En general, s'observa que les SMR en les dones amb edats superiors a 14 anys i en els homes amb menys de 14 anys hi ha una zona central on el risc és superior a la mitjana, i a mesura que ens allunyem d'aquesta zona en direcció nord i sud el risc va minvant. No obstant això, la variació geogràfica és més evident en el grup de les dones perquè la variància de l'efecte aleatori és superior que en el grup dels homes, fet que permet identificar més fàcilment comarques amb riscos més extrems.

Respecte als estrats sense components espacials també es posa de manifest la qüestió de la variància de l'efecte aleatori, però és el grup dels homes el que presenta una dispersió superior, de manera que les SMR presenten més variabilitat al llarg del mapa.

4.4. Conclusions

En aquest últim capítol s'ha presentat l'anàlisi geogràfica de la incidència de la diabetis de tipus I a Catalunya. En aquesta anàlisi s'han utilitzat alguns dels resultats més importants presentats en aquesta tesi, com ara l'estimació dels paràmetres mitjançant la tècnica de la quasiversemblança penalitzada, la utilització de les transformacions dels components de la variància per millorar la convergència dels models i l'aplicació de proves per testar la independència espacial.

Segons els resultats presentats, s'ha observat que les taxes d'incidència de la diabetis de tipus I són diferents segons el gènere i l'edat. Els individus amb una edat inferior als 14 anys presenten un risc superior, però la diferència de risc respecte de l'edat és superior en les dones que en els homes.

A més, també s'ha obtingut que la distribució d'aquestes taxes al llarg del territori català era diferent en funció d'aquestes variables explicatives. Així, les taxes que presenten un patró geogràfic espacial han estat les de les dones amb edat superiors a 14 anys i les dels homes amb edat inferior o igual a 14 anys. La distribució que s'ha observat correspon a una incidència més elevada a les comarques centrals, amb una gradació a mesura que ens allunyem d'aquestes comarques. Aquest patró s'aprecia més clarament en el grup de les dones atès que hi ha més variabilitat de l'efecte aleatori.

CAPÍTOL V: RESUM I CONCLUSIONS

L'objectiu d'aquesta tesi ha estat l'estudi i la millora de les tècniques emprades en el modelatge de riscos amb correlació espacial. Amb aquest tipus de dades, els models lineals generalitzats no es poden utilitzar perquè aquests models assumeixen independència entre les dades. La violació d'aquesta assumpció provoca l'aparició del fenomen anomenat *sobredispersió*, és a dir, les dades presenten una variabilitat superior que la variabilitat assumida pel model. A més, aquesta sobredispersió també pot ser deguda a l'absència de factors de variables explicatives. La conseqüència principal de la presència de sobredispersió és que els errors estàndards de les estimacions són incorrectes. Així doncs, per poder modelar aquests riscos tenint en compte aquesta sobredispersió, s'han introduït els models lineals generalitzats mixtos. Aquests models es classifiquen en funció del tipus de dependència que es modela i se'n defineixen tres tipus: el model autoregressiu condicional intrínsec (CAR intrínsec), on s'assumeix que la dependència de les dades és deguda totalment a la correlació espacial; el model d'heterogeneïtat, on la dependència modelada és no estructurada i, per tant, s'assumeix l'absència de correlació espacial, i finalment l'autoregressiu condicional no intrínsec (CAR no intrínsec), on es modela tant la dependència estructurada espacialment com la d'origen no espacial. Els models lineals generalitzats mixtos, a l'igual dels models lineals generalitzats, requereixen l'especificació de la funció que enllaça la mitjana de les dades amb les variables explicatives i la distribució de l'error, entenent-se l'error com les dades condicionades a les variables explicatives. En el marc de la modelització de riscos, atès que les dades són recomptes, habitualment s'assumeix que la distribució de l'error és la de Poisson, i que la funció d'enllaç és la logarítmica. A més s'ha d'especificar la distribució dels efectes aleatoris que típicament és fixa a una distribució normal. Un cop s'han determinat aquestes especificacions la dificultat és dur a terme les estimacions. Una possibilitat seria fer les estimacions per màxima versemblança, definint-la com la integral de la combinació de la distribució de les dades condicionades als efectes aleatoris (Poisson) i la mateixa distribució dels efectes aleatoris (Normal). Malauradament aquest procediment requereix la resolució d'integrals múltiples, que sovint són complicades de resoldre analíticament. S'han dut a terme altres aproximacions per poder estimar aquests models, entre les quals s'ha de destacar la quasiversemblança penalitzada (Breslow i Clayton, 1993). Aquest procediment substitueix la versemblança de les dades condicionada als efectes aleatoris per la

quasiversemblança (McCullagh i Nelder, 1989) i resol la integral mitjançant l'aproximació de Laplace.

Un altre procediment que permet obtenir estimacions per models lineals generalitzats mixtos és la *fully bayesian* a partir de les cadenes de Markov-Montecarlo (MCMC). Aquesta metodologia es basa en la definició de distribucions a priori pels paràmetres d'interès i en el fet de trobar la distribució posterior d'aquests paràmetres; l'estimació dels paràmetres es duu a terme mitjançant alguna característica (mitjana, mediana, moda) de la distribució posterior. Trobar la distribució posterior també requereix resoldre una integral que sovint és intractable, problema que se salva mitjançant mètodes de simulació que generen mostres o cadenes dels paràmetres que provenen d'aquesta distribució. Com es duu a terme aquesta simulació o generació de cadenes acaba definint el procediment; així, ens podem trobar els mètodes *Gibbs sampling* o *Metropolis-Hasting*, el primer dels quals és el que s'ha utilitzat en aquesta tesi.

En el capítol I s'ha comparat el comportament dels procediments d'estimació dels paràmetres de la quasiversemblança penalitzada i el *fully bayesian* via *Gibbs sampling* considerant funcions *a priori* no informatives mitjançant una simulació. Un dels resultats més remarcables ha estat que la quasiversemblança penalitzada s'ha presentat lleugerament més esbiaixada que la *fully bayesian*, sobretot pel que fa a l'ordenada en l'origen, però més eficient, fet que ha comportat una millor consistència de les estimacions obtingudes amb el procediment de la quasiversemblança penalitzada. En realitat, les diferències han estat petites, cosa que fa difícil de poder recomanar una tècnica o una altra, però si es volgués fer una recomanació d'aquesta mena podria ser la següent: en situacions amb poques regions i pocs casos esperats, es recomana la *fully bayesian*; en canvi, si la situació és de moltes regions i molts casos esperats s'obtenen resultats similars en ambdues tècniques, però és més avantatjós el procediment de la quasiversemblança penalitzada, perquè presenta més precisió en les estimacions. Aquest resultat es pot fonamentar també teòricament si ens fixem en l'esperit de com es du a terme l'estimació bayesiana, que es basa a combinar la versemblança de les dades amb les distribucions a priori. Per tant, el pes de la versemblança en l'estimació dependrà sobretot de la quantitat d'informació que conté la versemblança, és a dir, de la quantitat de dades. Una situació amb poques regions i pocs casos esperats és una situació amb poques dades, i aquesta manca de força de la versemblança és corregida per les

distribucions a priori. En canvi, quan la situació és de força dades, la versemblança té poder suficient per si sola per obtenir bones estimacions.

D'altra banda, les dificultats principals del procediment de quasiversemblança penalitzada són que convergeixi en una solució, ja que es tracta d'un procés iteratiu, o que les estimacions es trobin dins del seu domini. La conseqüència de la manca de convergència és obvia: no es tenen estimacions. En canvi, el fet que les estimacions es trobin fora del seu domini, per exemple una variància negativa, provoca que les estimacions dels errors estàndards siguin totalment incorrectes, i poden arribar a ser negatius, és a dir, la convergència s'ha trobat en una solució no vàlida. En el capítol II s'ha presentat una parametrització de la matriu de variàncies i covariàncies dels efectes aleatoris que millora notablement el percentatge de convergència i garanteix que les estimacions obtingudes es trobin dins del seu domini.

L'anàlisi geogràfica de riscos intenta obtenir estimacions corregides pel fet que les dades no siguin independents per la presència de correlació espacial, però també és interessant per trobar pautes o agrupacions geogràfiques en funció dels riscos. En aquest sentit una hipòtesi fonamental que tot investigador s'ha de plantejar és la hipòtesi d'independència espacial, perquè aquesta independència implicarà l'absència de patrons espacials. Òbviament, això no significarà que les dades siguin independents, però la causa de la sobredispersió no serà d'origen espacial. En el capítol III s'han avaluat diferents proves per contrastar aquesta hipòtesi, com ara el test de Wald, el test del quocient de les versemblances, l'*score test* i l'*AKAIKE Information Criterion.* També s'ha volgut estudiar el comportament del coeficient de concordança entre matrius. En resum, el test de Wald ha presentat un mal comportament tant en termes d'error de tipus I com de potència a causa de la gran dificultat de rebutjar la hipòtesi nul·la. Aquesta dificultat s'ha raonat que podria basar-se en la mala aproximació del paràmetre del pes espacial a una distribució normal, condició necessària perquè el test de Wald es comporti correctament. El test del quocient de les versemblances i l'*score test* han tingut un comportament similar, però s'ha observat que l'*score test* és més estable en relació amb la proporció d'error de tipus I i de potència quan les situacions de nombre de regions, de casos esperats i de variància de sobredispersió han estat modificades. Cal destacar, però, que en algunes situacions extremes, com ara un nombre petit de regions i de casos esperats i una sobredispersió petita, les potències han estat força baixes,

qüestió que haurien de tenir en compte els investigadors quan utilitzin aquestes proves. D'altra banda, l'*AKAIKE Information Criterion* s'ha utilitzat com un criteri no basat en probabilitat, és a dir, el model que té un AIC millor és el model elegit. Sota aquest criteri s'ha observat un percentatge de decisions incorrectes (al voltant del 10%) en situacions d'independència espacial. El coeficient de concordança s'ha presentat com una mesura força útil per prendre decisions sobre independència espacial, sobretot per la seva estabilitat respecte a les altres proves quan les condicions de nombre de regions, nombre de casos esperats i sobredispersió varien. Aquest fet és degut principalment a la naturalesa no paramètrica d'aquest coeficient, ja que el que fa és mesurar la distància entre la matriu de correlacions dels efectes aleatoris i la matriu identitat, i reescala aquesta distància en l'interval [0,1]. S'ha trobat que una concordança superior a 0,85 dóna suport a la decisió d'independència espacial.

Finalment, en el capítol IV s'ha il·lustrat la metodologia presentada al llarg de la tesi mitjançant una aplicació sobre el risc de diabetis de tipus I a Catalunya. El procediment d'estimació utilitzat ha estat la quasiversemblança penalitzada, atès que, en treballar amb un nombre de casos esperats gran, aquest procediment és més consistent que el *fully bayesian*. Per assegurar la convergència de l'algoritme de maximització s'han considerat les transformacions proposades pels components de la variància, i finalment s'ha utilitzat l'*score test*, el quocient de les versemblances i el coeficient de concordança per testar l'absència d'un patró geogràfic. En aquest exemple s'ha posat de manifest que el risc de diabetis de tipus I és diferent respecte del gènere i l'edat, de manera que tant en els homes com en les dones presenten més risc els individus amb una edat inferior o igual als 14 anys. També s'ha observat que els patrons geogràfics al llarg de les comarques han estat diferents en funció del gènere i de l'edat del diagnòstic.

En resum, les conclusions principals de la tesi són:

— Els models lineals generalitzats no permeten modelar la sobredispersió espacial en els estudis de la variabilitat del risc al llarg d'una àrea determinada.

— Els models lineals generalitzats mixtos permeten controlar la sobredispersió deguda a les regions mitjançant efectes aleatoris, i són necessaris per obtenir unes estimacions correctes dels paràmetres i del seu error estàndard.

— Els mètodes d'estimació dels paràmetres mitjançant la quasiversemblança penalitzada i la *fully bayesian* via *Gibbs sampling* no presenten unes diferències gaire pronunciades en relació amb el biaix i la precisió de les estimacions.

— El procediment de la quasiversemblança penalitzada presenta estimacions més consistents quan el nombre de regions i de casos esperats és gran.

— El mètode de la *fully bayesian* via *Gibbs sampling* proporciona estimacions més consistents quan el nombre de regions i casos esperats és petit.

— Les transformacions proposades pels components de la variància milloren notablement la convergència del mètode d'estimació de la quasiversemblança penalitzada.

— L'*score test* és la prova que presenta un equilibri millor entre l'error tipus I i la potència per testar la independència espacial. No obstant això, s'ha apreciat una potència baixa en les situacions de poques regions i pocs casos esperats.

— El coeficient de concordança entre matrius es presenta també com una mesura útil per prendre decisions sobre la hipòtesi d'independència espacial, especialment quan es treballa amb poques regions i pocs casos esperats.

APÈNDIX :FUNCIONS EN LLENGUATGE S

Apèndix 1: Funcions en llenguatge S utilitzades en el capítol I

1.1. Funció per calcular la matriu de variàncies i covariàncies total, V:

```
calculV<-function(W,D,Z){                    #D: matriu d'efectes aleatoris  W: matriu
    de pesos
    V<-ginverse(W)+(Z%*%D%*%t(Z))
    return(V)
}
```

1. 2. Funció per calcular l'estimació dels efectes fixes

```
est.a<-function(X,Z,V,Y,n){                  #Estimació efectes fixes
    aux1<-t(X)%*%ginverse(V)%*%X
    aux2<-t(X)%*%ginverse(V)%*%(Y)
    a<-ginverse(aux1)%*%aux2
    return(a)
}
```

1. 3. Funció per calcular l'estimació dels efectes aleatoris

```
est.b<-function(X,Z,V,D,Y,a){              #Estimació efectes aleatoris
    aux3<-D%*%t(Z)%*%ginverse(V)
    aux4<-Y-(X%*%a)
    b<-aux3%*%aux4
    return(b)
}
```

1.4. Funció per calcular el vector de les dades transformades

```
y.trans<-function(Y,X,Z,n,a,b,mu){  #No cap posar offset
    eta<-(X%*%a)+(Z%*%b)
    yt<-eta+((Y-mu)/mu)
    return(yt)
}
```

1.5. Funció per generar una mostra de valors simulats amb una estructura de la matriu de variàncies i covariàncies dels efectes aleatoris definida per Besag, York i Mollié 1991

```
simulacio.dades <- function(X.r, Z.r, m, n, sigmas.r, sigmah.r, alpha.r, alpha1.r, Q)
{
    #  sigmas.r    # Desviació  real de l'efecte aleatori espaial
    #  sigmah.r   # Desviació real de l'efecte aleatori heterogeneitat
    #    alpha.r   # Efecte fixe, intercept
    #  alpha1.r   # Covariable fixa
    #    m     # Nombre de regions
```

```
    #    n     # Vector valors esperats per cada zona
    #    Q     # Disseny Veins
    fixed.r <- c(alpha.r, alpha1.r)
    D.r <- (sigmas.r^2) * ginverse(Q) + (sigmah.r^2) * diag(m)
    #Matriu de covariàncies dels efectes aleatoris
    if(sigmah.r == 0) b.r <- as.vector(rmultnorm(1, rep(0, m), D.r)) else
        b.r <- as.vector(rmvnorm(1, rep(0, m), D.r))
    mu.r <- n * exp(X.r %*% fixed.r + Z.r %*% b.r)
    Y.r <- rpois(m, mu.r)
    return(Y.r)
}
```

1.6. Funció que estima els paràmetres del model mitjançant la Quasi Versemblança Penalitzada sense considerar les transformacions dels paràmetres

```
carnon.Leroux<-function(Y.r,b.r,X.r,Z.r,m,n,Q,n.iter.pql,cate){
##############################################################################
       Estimació PQL d'un model de Poisson amb l'estructura efectes aleatoris#
#      proposat per Leroux:                                                  #
##############################################################################
    a.r<-array(c(0.1,0.3),dim=c(2,1))
    SMR.actual<-X.r%*%a.r+Z.r%*%b.r
    sigma<-0.5               #Assignació valors inicials i primera estimació dels
    paràmetres equació
    lambda<-0.5
    prova <- data.frame(Y.r, X.r[, 2], n)
    a <- array(glm(Y.r ~ X2 + offset(log(n)), data = prova, family = poisson) $coefficients,
    dim = c(2, 1))
    a.inicials <- a
    mu <- n * exp(X.r %*% a)
#   a<-rep(0,ncol(X.r))
    b<-array(0,m)
#   mu<-n
    R<-lambda*Q+(1-lambda)*diag(m)
    D<-(sigma^2)*ginverse(R)
    W<-diag(cate*m)
    diag(W)<-mu
    V<-calculV(W,D,Z.r)
    YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
    for (i in 1:n.iter.pql){
            print(c("Iter",i))
            a<-est.a(X.r,Z.r,V,YT,n)
            b<-est.b(X.r,Z.r,V,D,YT,a)
            mu.new<-n*exp(X.r%*%a+Z.r%*%b)
            criteri.mu<-0
            for (j in 1:m){
                dif.mu<-mu[j]-mu.new[j]
                if  (dif.mu<=0.001) criteri.mu<-criteri.mu+1
            }
```

```
            mu<-mu.new
            if (is.infinite(sum(mu))){
                print("Les mitjanes son infinit")
                a<-a.inicials
                mu<-n*exp(X.r%*%a)
                b<-array(0,m)
                YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
                res<-YT-(X.r%*%a)
                W<-diag(cate*m)
                diag(W)<-mu
                V<-calculV(W,D,Z.r)
            break}
            YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
            res<-YT-(X.r%*%a)
            #Reestimació matrius de variàncies
            W<-diag(cate*m)
            diag(W)<-mu
            V<-calculV(W,D,Z.r)
            if (criteri.mu==m) break
    }
#Fisher Scoring
status<-1
status1<-0
    for (i in 1:n.iter.pql){
        print(c("Iter_FS",i))
        print(det(V))
        if (det(V)>100) g.V<-solve(V)
        if (det(V)<=100) g.V<-ginverse(V)
        P<-g.V-(g.V%*%X.r%*%(ginverse(t(X.r)%*%g.V%*%X.r))%*%t(X.r)%*%g.V)
        der.s<-2*sigma*(Z.r%*%ginverse(R)%*%t(Z.r))
        der.l<-(-1)*(sigma^2)*Z.r%*%ginverse(R)%*%(Q-diag(m))%*%ginverse(R)%*%t(Z.r)
        der.ss<-2*Z.r%*%ginverse(R)%*%t(Z.r)
        der.sl<-(-2)*sigma*Z.r%*%ginverse(R)%*%(Q-diag(m))%*%ginverse(R)%*%t(Z.r)
        der.ll<-2*(sigma^2)*Z.r%*%ginverse(R)%*%(Q-diag(m))%*%ginverse(R)%*%(Q-
    diag(m))%*%ginverse(R)%*%t(Z.r)
        U.s<-((0.5)*t(YT)%*%P%*%der.s%*%P%*%YT)-(0.5*sum(diag(P%*%der.s)))
        U.l<-((0.5)*t(YT)%*%P%*%der.l%*%P%*%YT)-(0.5*sum(diag(P%*%der.l)))
    #   U.s<-((0.5)*t(res)%*%g.V%*%der.s%*%g.V%*%res)-(0.5*sum(diag(P%*%der.s)))
    #   U.l<-((0.5)*t(res)%*%g.V%*%der.l%*%g.V%*%res)-(0.5*sum(diag(P%*%der.l)))
        U<-c(U.s,U.l)
        I<-array(,c(2,2))
        I[1,1]<-(0.5)*sum(diag(P%*%der.s%*%P%*%der.s))
        I[2,1]<-(0.5)*sum(diag(P%*%der.s%*%P%*%der.l))
        I[1,2]<-(0.5)*sum(diag(P%*%der.l%*%P%*%der.s))
        I[2,2]<-(0.5)*sum(diag(P%*%der.l%*%P%*%der.l))
        sigma.new<-sigma+(ginverse(I)%*%U)[1]
        lambda.new<-lambda+(ginverse(I)%*%U)[2]
        mat.inf<-ginverse(I)%*%U
#       print(mat.inf)
        if (is.na(mat.inf[1,1])||(is.na(mat.inf[2,1]))){status<-0;break}
        if (is.infinite(mat.inf[1,1])||(is.infinite(mat.inf[2,1]))){status<-0;break}
```

```
        if ((abs(mat.inf[1])>0.5) || (abs(mat.inf[2])>0.5)) {
#           print("Marquardt technique is used")
            step<-max(ceiling(abs(mat.inf*4)))
            sigma.new<-sigma+((1/step)*ginverse(I)%*%U)[1]
            lambda.new<-lambda+((1/step)*ginverse(I)%*%U)[2]
#           print(c("step",step))
        }
        dif.sigma<-sigma.new-sigma
        dif.lambda<-lambda.new-lambda
    #   print(c("dif.sigma",dif.sigma,"dif.lambda",dif.lambda))
        lambda<-lambda.new
        sigma<-sigma.new
#       print(c("sigma",sigma))
        print(c("lambda",lambda))
    #Update de matrius R, D i V
        R<-lambda*Q+(1-lambda)*diag(m)
        D<-(sigma^2)*ginverse(R)
        V<-calculV(W,D,Z.r)
    #Update a i b
        a.new<-est.a(X.r,Z.r,V,YT,n)
        b.new<-est.b(X.r,Z.r,V,D,YT,a)
        mu<-n*exp(X.r%*%a.new+Z.r%*%b.new)
        if (is.infinite(sum(mu))){status<-0;break}
        criteri.b<-0
        for (j in 1:m){
            dif.b<-b[j]-b.new[j]
            if  (dif.b<=0.001) criteri.b<-criteri.b+1
        }
        dif.a<-a.new-a
        a<-a.new
        b<-b.new
        YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
        if (is.na(sum(YT))){status<-0;break}
        res<-YT-X.r%*%a
    #Update W i V
        W<-diag(cate*m)
        diag(W)<-mu
        V<-calculV(W,D,Z.r)
        if         ((abs(dif.sigma)         <=0.001)         &&         (abs(dif.lambda)
    <=0.001)&&(abs(dif.a[1,1])<=0.001)&&(abs(dif.a[2,1])<=0.001)&&(criteri.b==m))
    break
        if(i==n.iter.pql) {status<-0; break}
        }
    if ((lambda>1)|| (lambda<0)) status1<-1
    if (sigma<0) status1<-1
    residu<-YT-(X.r%*%a)-(Z.r%*%b)
    SMR.pred<-(X.r%*%a)+(Z.r%*%b)
    SMR.error<-SMR.pred-SMR.actual
    m.SMR.error<-mean(SMR.error)
    var.SMR.error<-var(SMR.error)
    inv.I<-ginverse(I)
```

```
    inv.V<-ginverse(V)
    err.s<-diag(inv.I)
    err.fe<-ginverse(t(X.r)%*%inv.V%*%X.r)
    estimacio<-
    list(a,b,lambda,sigma,mu,residu,err.s,err.fe,status,status1,m.SMR.error,var.SMR.e
    rror)
    names(estimacio)<-
    c("fixed.effects","random.effects","lambda","sigma","mu","res","error.s","error.f
    e","status","status1","m.SMR.error","var.SMR.error")
    return(estimacio)
}
```

1.7. Funció que crea el fitxer que inclou les dades simulades i la definició de la matriu de veïns necessaris per estimar mitjançant el programa BUGS.

```
c.f.data <- function(var1, var2, var3, var4, var5, m, file1)
{
    var1 <- paste(var1,  , sep = "", collapse = ",")
    #Creació del vector d'observats
    var2 <- paste(var2,  , sep = "", collapse = ",")
    #Creació del vector d'esperats
    llista <- c("list(O=c(", var1, "),", "E=c(", var2, "),", "map=c(")
    write(llista, file = file1)
    pos.in <- 1
    #Creació del vector de veins acumulats
    pos.fin <- 0
    off <- array(0, (m + 1))
    for(i in 1:(m - 1)) {
    pos.in <- pos.fin + 1
    pos.fin <- pos.fin + var4[i]
    a2 <- paste(var3[pos.in:pos.fin], ",", sep = "", collapse = ""
        )
    #Creació de la matriu map
    write(a2, file = file1, append = T)
    off[i + 1] <- off[i] + var4[i]
    }
    pos.in <- pos.fin + 1
    #Última línea feta apart per problemes amb les comes
    pos.fin <- pos.fin + var4[m]
    a2 <- paste(var3[pos.in:pos.fin], collapse = ",")
    write(a2, file = file1, append = T)
    off[m + 1] <- off[m] + var4[m]
    #Creació del vector del nombre de veins acumulat
    aux <- paste(off,  , sep = "", collapse = ",")
    aux1 <- paste(var5,  , sep = "", collapse = ",")
    write(c("),", "x=c(", aux1), file = file1, append = T)
    write(c("),", "off=c(", aux, "))"), file = file1, append = T)
}
```

1.8. Funció que genera la matriu de veïns utilitzada en l'estimació mitjançant la quasi versemblança penalitzada

```
crea.quadricula <- function(lon, lat, m)
{
    v1 <- dist(cbind(lon, lat))
    aux <- 0
    v2 <- 0
    v3 <- 0
    ini.vec <- 1
    fin.vec <- m - 1
    for(i in 1:(m - 1)) {
    v2[ini.vec:fin.vec] <- i
    v3[ini.vec:fin.vec] <- c((i + 1):m)
    ini.vec <- ini.vec + (m - i)
    fin.vec <- fin.vec + (m - i) - 1
    }
    veins <- matrix(0, m, m)
    combi <- data.frame(cbind(v2, v3, v1))
    for(i in 1:((m * (m - 1))/2)) {
    if(combi$v1[i] < 2)
        veins[combi$v2[i], combi$v3[i]] <- -1
    if(combi$v1[i] < 2)
        veins[combi$v3[i], combi$v2[i]] <- -1
    }
    for(i in 1:m) {
    aux[i] <- sum(veins[i,  ] == -1)
    }
    diag(veins) <- aux
    return(veins)
}
```

1.9. Funcions que creen els vectors per definir la matriu de veïns necessaris per l'estimació dels model mitjançant el programa BUGS.

```
assigna <-function(coor, mat)
{
    i <- coor[1]
    j <- coor[2]
    mat[i, j] <- 1
    mat[j, i] <- 1
    return(mat)
}

veins.b <-function(lon, lat, m)
{
    v1 <- dist(cbind(lon, lat))
    #calcula les distancies
    aux <- 0
    v2 <- 0
```

```
    v3 <- 0
    mat <- matrix(0, m, m)
    V <- 0
    num <- array(, m)
    ini.vec <- 1
    fin.vec <- m - 1
    for(i in 1:(m - 1)) {
    #Crea totes les possibles parelles
    v2[ini.vec:fin.vec] <- i
    v3[ini.vec:fin.vec] <- c((i + 1):m)
    ini.vec <- ini.vec + (m - i)
    fin.vec <- fin.vec + (m - i) - 1
    }
    parelles <- matrix(0, m, m)
    parelles <- cbind(v2, v3, v1)
    p.veins <- parelles[parelles[, 3] < 2, 1:2]
    #Agafa les  parelles que estan a una distancia menor de 2
    n <- nrow(p.veins)
    # Nombre de parelles
    for(i in 1:n) {
    vec <- p.veins[i,  ]
    mat <- assigna(vec, mat)
    }
    for(i in 1:nrow(mat)) {
    aux1 <- which(mat[i,  ] == 1)
    #Ens diu quins són veins
    num[i] <- sum(mat[i,  ])
    #Nombre de veins per cada zona
    V <- c(V, aux1)
    }
    V <- V[2:((2 * n) + 1)]
    return(list(V, num))
}
```

Apèndix 2: Funcions en llenguatge S utilitzades en el capítol II

2.1. Funció per generar una mostra de valors simulats amb una estructura de la matriu de variàncies i covariàncies dels efectes aleatoris definida per Leroux *et al.* 1999.

```
simulacio.dades.lambda< function(X.r,Z.r,m,n,sigma.r,lambda.r,alpha.r,alpha1.r,Q){
#    sigma.r           # Desviació  real de l'efecte aleatori espaial
#    lambda.r          # Lambda real
#    alpha.r           # Efecte fixe, intercept
#    alpha1.r          # Covariable fixa
#    m                 # Nombre de regions
#    n                 # Vector valors esperats per cada zona
#    Q                 # Disseny Veins
       fixed.r<-c(alpha.r,alpha1.r)
```

```
        R<-lambda.r*Q+(1-lambda.r)*diag(m)
        D.r<-(sigma.r^2)*ginverse(R)  #Matriu de covariàncies dels efectes aleatoris
        if  (lambda.r==1)    b.r<-as.vector(rmultnorm(1,  rep(0,m),D.r))  else  b.r<-
        as.vector(rmvnorm(1,rep(0,m),D.r))
        mu.r<-n*exp(X.r%*%fixed.r+Z.r%*%b.r)
        Y.r<-rpois(m, mu.r)
        return(Y.r)
}
```

2.2. Funció que estima els paràmetres del model mitjançant la Quasi Versemblança Penalitzada considerant les transformacions dels paràmetres

```
m.carnon.Leroux.t<-function(Y.r,b.r,X.r,Z.r,m,n,Q,n.iter.pql,cate){
##############################################################################
###     Estimació PQL d'un model de Poisson amb l'estructura efectes aleatoris#
#       proposat per Leroux:                                                 #
##############################################################################
    a.r<-array(c(0.1,0.3),dim=c(2,1))
    SMR.actual<-X.r%*%a.r+Z.r%*%b.r
    tau.s<-log(0.5)
    sigma<-exp(tau.s)            #Assignació valors inicials i primera estimació dels
    paràmetres equació
    tau.l<-log(0.5/(1-0.5))
    lambda<-exp(tau.l)/(1+exp(tau.l))
    prova <- data.frame(Y.r, X.r[, 2], n)
    a   <-   array(glm(Y.r   ~   X2   +   offset(log(n)),   data   =   prova,   family
    =poisson)$coefficients, dim = c(2, 1))
    a.inicials <- a
    mu <- n * exp(X.r %*% a)
    b<-array(0,m)
    R<-lambda*Q+(1-lambda)*diag(m)
    D<-(sigma^2)*ginverse(R)
    W<-diag(cate*m)
    diag(W)<-mu
    V<-calculV(W,D,Z.r)
    YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
    for (i in 1:n.iter.pql){
            print(c("Iter_mu",i))
            a<-est.a(X.r,Z.r,V,YT,n)
            b<-est.b(X.r,Z.r,V,D,YT,a)
            mu.new<-n*exp(X.r%*%a+Z.r%*%b)
            criteri.mu<-0
            for (j in 1:m){
                dif.mu<-mu[j]-mu.new[j]
                if  (dif.mu<=0.001) criteri.mu<-criteri.mu+1
            }
            mu<-mu.new
                if (is.infinite(sum(mu))){
                print("Les mitjanes son infinit")
                a<-a.inicials
                mu<-n*exp(X.r%*%a)
```

```
                b<-array(0,m)
                YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
                res<-YT-(X.r%*%a)
                W<-diag(cate*m)
                diag(W)<-mu
                V<-calculV(W,D,Z.r)
                break}
            YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
            res<-YT-(X.r%*%a)
            #Reestimació matrius de variàncies
            W<-diag(cate*m)
            diag(W)<-mu
            V<-calculV(W,D,Z.r)
            if (criteri.mu==m) break
    }
#Fisher Scoring i actualització parametres
    r.fs<-
    F.S.S.L(mu,V,X.r,Z.r,R,Q,W,tau.s,tau.l,sigma,lambda,m,YT,Y.r,a,b,n,n.iter.pql,cat
    e)
    residu<-r.fs$YT-(X.r%*%r.fs$a)-(Z.r%*%r.fs$b)
    SMR.pred<-(X.r%*%r.fs$a)+(Z.r%*%r.fs$b)
    SMR.error<-SMR.pred-SMR.actual
    m.SMR.error<-mean(SMR.error)
    var.SMR.error<-var(SMR.error)
    inv.I<-ginverse(r.fs$I)
    inv.V<-ginverse(r.fs$V)
    err.t<-inv.I
    err.fe<-ginverse(t(X.r)%*%inv.V%*%X.r)
    #Càlcul log versemblança restringida
    l.REML<-(-0.5)*(log(det(r.fs$V))+log(det(t(X.r)%*%inv.V%*%X.r))+(t(r.fs$YT-
    (X.r%*%r.fs$a))%*%inv.V%*%(r.fs$YT-(X.r%*%r.fs$a))))
    estimacio<-
    list(r.fs$a,r.fs$b,r.fs$lambda,r.fs$sigma,r.fs$tau.s,r.fs$tau.l,r.fs$mu,residu,er
    r.t,err.fe,r.fs$status,r.fs$s.L,r.fs$I,m.SMR.error,var.SMR.error,l.REML)
    names(estimacio)<-
    c("fixed.effects","random.effects","lambda","sigma","tau.sigma","tau.lambda","mu"
    ,"res","error.tau","error.fe","status","s.L","I","m.SMR.error","var.SMR.error","l
    .REML")
    return(estimacio)
}
##############################################################################
    #Funció Fisher scoring per les components de la variància i actualització efectes
    #fixes i aleatoris.
##############################################################################
F.S.S.L<-
    function(mu,V,X.r,Z.r,R,Q,W,tau.s,tau.l,sigma,lambda,m,YT,Y.r,a,b,n,n.iter.pql,ca
    te){
        for (i in 1:n.iter.pql){
        print(c("Iter_FS",i))
        #Update de matrius R, D i V    #Només serveix per quan fixem la lambda a 1
        R<-lambda*Q+(1-lambda)*diag(m)
```

```
        D<-(sigma^2)*ginverse(R)
        V<-calculV(W,D,Z.r)
        g.V<-ginverse(V)
        P<-g.V-(g.V%*%X.r%*%(ginverse(t(X.r)%*%g.V%*%X.r))%*%t(X.r)%*%g.V)
        der.ts<-2*exp(2*tau.s)*(Z.r%*%ginverse(R)%*%t(Z.r))
#       der.l<-exp(tau.l)/((1+exp(tau.l))^2)
        der.l<-lambda*(1-lambda)
        aux1<-der.l*Q-(der.l*diag(m))
        der.tl<-(-1)*(exp(2*tau.s))*Z.r%*%ginverse(R)%*%(aux1)%*%ginverse(R)%*%t(Z.r)
        U.s<-((0.5)*t(YT)%*%P%*%der.ts%*%P%*%YT)-(0.5*sum(diag(P%*%der.ts)))
        U.l<-((0.5)*t(YT)%*%P%*%der.tl%*%P%*%YT)-(0.5*sum(diag(P%*%der.tl)))
    #   U.s<-((0.5)*t(res)%*%g.V%*%der.ts%*%g.V%*%res)-(0.5*sum(diag(P%*%der.ts)))
    #   U.l<-((0.5)*t(res)%*%g.V%*%der.tl%*%g.V%*%res)-(0.5*sum(diag(P%*%der.tl)))
        U<-c(U.s,U.l)
        I<-array(,c(2,2))
        I[1,1]<-(0.5)*sum(diag(P%*%der.ts%*%P%*%der.ts))
        I[2,1]<-(0.5)*sum(diag(P%*%der.ts%*%P%*%der.tl))
        I[1,2]<-(0.5)*sum(diag(P%*%der.tl%*%P%*%der.ts))
        I[2,2]<-(0.5)*sum(diag(P%*%der.tl%*%P%*%der.tl))
        tau.s.new<-tau.s+(ginverse(I)%*%U)[1]
        tau.l.new<-tau.l+(ginverse(I)%*%U)[2]
        mat.inf<-ginverse(I)%*%U
#       print(mat.inf)
        if (is.na(mat.inf[1,1])||(is.na(mat.inf[2,1]))){status<-0;break}
        if (is.infinite(mat.inf[1,1])||(is.infinite(mat.inf[2,1]))){status<-0;break}
            if ((abs(mat.inf[1])>1) || (abs(mat.inf[2])>1)) {
#           print("Marquardt technique is used")
            step<-max(ceiling(abs(mat.inf*4)))
            tau.s.new<-tau.s+((1/step)*ginverse(I)%*%U)[1]
            tau.l.new<-tau.l+((1/step)*ginverse(I)%*%U)[2]
#           print(c("step",step))
        }
        dif.tau.s<-tau.s.new-tau.s
        dif.tau.l<-tau.l.new-tau.l
        if (is.na(mat.inf[1,1])||(is.na(mat.inf[2,1]))){status<-0;break}
#       print(c("dif.tau s",dif.tau.s,"dif.tau l",dif.tau.l))
        tau.s<-tau.s.new
        tau.l<-tau.l.new
        sigma.new<-exp(tau.s)
        if (tau.l>709) lambda.new<-1
        if ((tau.l<710)&&(tau.l>=(-380)))lambda.new<-exp(tau.l)/(1+exp(tau.l))
        if (tau.l<(-380)) lambda.new<-0    #Aquesta condició es per que si hi ha un
    valor de lambda molt proper a 0 0.*10-183 després no pot fer la ginverse
        dif.sigma<-sigma.new-sigma
        dif.lambda<-lambda.new-lambda
#       print(c("dif.lambda",dif.lambda,"dif.sigma",dif.sigma))
#       print(c("lambda",lambda,"sigma",sigma))
        sigma<-sigma.new
        lambda<-lambda.new
    #Update de matrius R, D i V
        R<-lambda*Q+(1-lambda)*diag(m)
```

```
        D<-(sigma^2)*ginverse(R)
        V<-calculV(W,D,Z.r)
    #Update a i b
        a.new<-est.a(X.r,Z.r,V,YT,n)
        b.new<-est.b(X.r,Z.r,V,D,YT,a)
        mu<-n*exp(X.r%*%a.new+Z.r%*%b.new)
        if (is.infinite(sum(mu))){status<-0;break}
        criteri.b<-0
        for (j in 1:m){
            dif.b<-b[j]-b.new[j]
            if  (dif.b<=0.001) criteri.b<-criteri.b+1
        }
        dif.a<-a.new-a
        a<-a.new
        b<-b.new
#       print(c("dif.a",dif.a,"criteri.b",criteri.b))
        YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
        res<-YT-X.r%*%a
    #Update W i V
        W<-diag(cate*m)
        diag(W)<-mu
        V<-calculV(W,D,Z.r)
        if((abs(dif.lambda)<=0.001)&&(abs(dif.sigma)<=0.001)&&
        (abs(dif.a[1,1])<=0.001)&&(abs(dif.a[2,1])<=0.001)&&(criteri.b==m))
        {status<-1; break}
        if(i==n.iter.pql) {status<-0; break}
        }
    f.sc<-list(a,b,lambda,sigma,tau.s,tau.l,mu,V,I,YT,R,W,status)
    names(f.sc)<-
    c("a","b","lambda","sigma","tau.s","tau.l","mu","V","I","YT","R","W","status")
    return(f.sc)
    }
```

Apèndix 3: Funcions en llenguatge S utilitzades en el capítol III

3.1. Funció que estima els paràmetres del model mitjançant la Quasi Versemblança Penalitzada considerant les transformacions dels paràmetres però fixant el valor del paràmetre del pes de la dependència (λ)

```
lamb.Leroux.t<-function(Y.r,b.r,X.r,Z.r,m,n,Q,n.iter.pql,lambda,cate){
##############################################################################
#   Estimació PQL d'un model de Poisson amb efectes aleatoris:                #
#               Si lambda=1 és un model CAR intrinsic.                        #
#               Si lambda=0 és un model amb un efecte aleatorio (white noise).  #
##############################################################################
#Assignació valors inicials i primera estimació dels paràmetres equació
    a.r<-array(c(0.1,0.3),dim=c(2,1))
    SMR.actual<-X.r%*%a.r+Z.r%*%b.r
    tau.s<-log(0.5)
```

```
sigma<-exp(tau.s)
prova <- data.frame(Y.r, X.r[, 2], n)
a <- array(glm(Y.r ~ X2 + offset(log(n)), data = prova, family
=poisson)$coefficients, dim = c(2, 1))
a.inicials <- a
b<-array(0,m)
mu<-n
R<-lambda*Q+(1-lambda)*diag(m)
D<-(sigma^2)*ginverse(R)
W<-diag(cate*m)
diag(W)<-mu
V<-calculV(W,D,Z.r)
YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
for (i in 1:n.iter.pql){
        print(c("Iter",i))
        a<-est.a(X.r,Z.r,V,YT,n)
        b<-est.b(X.r,Z.r,V,D,YT,a)
        mu.new<-n*exp(X.r%*%a+Z.r%*%b)
        criteri.mu<-0
        for (j in 1:m){
            dif.mu<-mu[j]-mu.new[j]
            if  (dif.mu<=0.001) criteri.mu<-criteri.mu+1
        }
        mu<-mu.new
            if (is.infinite(sum(mu))){
            print("Les mitjanes son infinit")
            a<-a.inicials
            mu<-n*exp(X.r%*%a)
            b<-array(0,m)
            YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
            res<-YT-(X.r%*%a)
            W<-diag(cate*m)
            diag(W)<-mu
            V<-calculV(W,D,Z.r)
            break}
        YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
        res<-YT-(X.r%*%a)
        #Reestimació matrius de variàncies
        W<-diag(cate*m)
        diag(W)<-mu
        V<-calculV(W,D,Z.r)
        if (criteri.mu==m) break
}
        if (criteri.mu<m){  #Si finalitza l'algoritme i no ha trobat solució,
restablir els valors inicials
            a<-a.inicials
            mu<-n*exp(X.r%*%a)
            b<-array(0,m)
            YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
            res<-YT-(X.r%*%a)
            W<-diag(cate*m)
```

```
                diag(W)<-mu
                V<-calculV(W,D,Z.r)
            }
#Fisher Scoring
status<-1
    for (i in 1:n.iter.pql){
        print(c("Iter_FS",i))
        g.V<-ginverse(V)
        P<-g.V-(g.V%*%X.r%*%(ginverse(t(X.r)%*%g.V%*%X.r))%*%t(X.r)%*%g.V)
        der.ts<-2*exp(2*tau.s)*(Z.r%*%ginverse(R)%*%t(Z.r))
        U<-((0.5)*t(YT)%*%P%*%der.ts%*%P%*%YT)-(0.5*sum(diag(P%*%der.ts)))
        U<-c(U)
        I<-(0.5)*sum(diag(P%*%der.ts%*%P%*%der.ts))
        tau.s.new<-tau.s+(ginverse(I)%*%U)
        mat.inf<-ginverse(I)%*%U
        if (is.na(mat.inf)){status<-0;break}
        if (is.infinite(mat.inf)){status<-0;break}
        if (abs(mat.inf)>1) {
            print("Marquardt technique is used")
            step<-(ceiling(abs(mat.inf*4)))
            sigma.new<-sigma+((1/step)*ginverse(I)%*%U)
            print(c("step",step))
        }
        dif.tau.s<-tau.s.new-tau.s
        print(c("dif.tau s",dif.tau.s,))
        tau.s<-c(tau.s.new)
        sigma.new<-c(exp(tau.s))
        dif.sigma<-sigma.new-sigma
#       print(c("sigma",sigma))
        sigma<-sigma.new
    #Update de matrius R, D i V
        R<-lambda*Q+(1-lambda)*diag(m)
        D<-(sigma^2)*ginverse(R)
        V<-calculV(W,D,Z.r)
    #Update a i b
        a.new<-est.a(X.r,Z.r,V,YT,n)
        b.new<-est.b(X.r,Z.r,V,D,YT,a)
        mu<-n*exp(X.r%*%a.new+Z.r%*%b.new)
        if (is.infinite(sum(mu))){status<-0;break}
        criteri.b<-0
        for (j in 1:m){
            dif.b<-b[j]-b.new[j]
            if  (dif.b<=0.001) criteri.b<-criteri.b+1
        }
        dif.a<-a.new-a
        a<-a.new
        b<-b.new
#       print(c("dif.a",dif.a,"criteri.b",criteri.b))
        YT<-y.trans(Y.r,X.r,Z.r,n,a,b,mu)
        res<-YT-X.r%*%a
    #Update W i V
```

```
        W<-diag(cate*m)
        diag(W)<-mu
        V<-calculV(W,D,Z.r)
        if ((abs(dif.sigma)
    <=0.001)&&(abs(dif.a[1,1])<=0.001)&&(abs(dif.a[2,1])<=0.001)&&(criteri.b==m))
    {status<-1; break}
        if(i==n.iter.pql) {status<-0; break}
        }
    residus<-YT-(X.r%*%a)
    pred<-log(n)+(X.r%*%a)+(Z.r%*%b)
    SMR.pred<-(X.r%*%a)+(Z.r%*%b)
    SMR.error<-SMR.pred-SMR.actual
    m.SMR.error<-mean(SMR.error)
    var.SMR.error<-var(SMR.error)
    inv.I<-ginverse(I)
    inv.V<-ginverse(V)
    error.t<-inv.I
    err.fe<-ginverse(t(X.r)%*%inv.V%*%X.r)
#Càlcul log versemblança restringida
    aux1<-log(det(V))+log(det(t(X.r)%*%inv.V%*%X.r))
    aux2<-(t(YT-(X.r%*%a)))%*%inv.V%*%(YT-(X.r%*%a))
    l.REML<-(-0.5)*(aux1+aux2)
    dev.Poisson(Y.r,mu)
    gg<-calcul.quasi.likelihood(Y.r,mu,b,sigma,lambda,Q,m,1,Z.r,2,49,2)
    qL<-gg$Quasilikelihood
#Càlcul score test
#   g.V<-ginverse(V)
    P<-g.V-(g.V%*%X.r%*%(ginverse(t(X.r)%*%inv.V%*%X.r))%*%t(X.r)%*%g.V)
    der.s<-2*sigma*(Z.r%*%ginverse(R)%*%t(Z.r))
    der.l<-(-1)*(sigma^2)*Z.r%*%ginverse(R)%*%(Q-diag(m))%*%ginverse(R)%*%t(Z.r)
    U.s<-((0.5)*t(YT)%*%P%*%der.s%*%P%*%YT)-(0.5*sum(diag(P%*%der.s)))
    U.l<-((0.5)*t(YT)%*%P%*%der.l%*%P%*%YT)-(0.5*sum(diag(P%*%der.l)))
    U<-c(U.s,U.l)
    G<-array(,c(2,2))
    G[1,1]<-(0.5)*sum(diag(P%*%der.s%*%P%*%der.s))
    G[2,1]<-(0.5)*sum(diag(P%*%der.s%*%P%*%der.l))
    G[1,2]<-(0.5)*sum(diag(P%*%der.l%*%P%*%der.s))
    G[2,2]<-(0.5)*sum(diag(P%*%der.l%*%P%*%der.l))
    score.test<-U.l/((G[2,2]-G[2,1]*(G[1,1]^(-1))*G[2,1])^(0.5))
    estimacio<-
    list(a,b,lambda,sigma,tau.s,mu,residus,pred,error.t,err.fe,status,m.SMR.error,var
    .SMR.error,score.test,l.REML,qL)
    names(estimacio)<-
    c("fixed.effects","random.effects","lambda","sigma","tau.sigma","mu","res","pred"
    ,"error.tau.s","error.fe","status","m.SMR.error","var.SMR.error","score.test","l.
    REML","q.L")
    return(estimacio)
}
```

BIBLIOGRAFIA

Bentler P. M. and Bonett D.G. (**1980**). *Significance Tests and Goodness of Fit in the Analysis of Covariance Structures*. Psychological Bulletin. **88**: 588-606.

Bernardinelli L., and Montomoli C. (**1992)**. *Empirical Bayes versus Fully Bayesian analysis of geographical variation in disease risks*. Statistics in Medicine. **11**: 983-1007.

Bernardinelli,L., Clayton,D., Pascutto,C., Montomoli,C., Ghislandi,M., and Songini,M.. (**1995**a). *Bayesian analysis of space-time variation in disease risk*. Statistics in Medicine. **14**: 2433-2443.

Bernardinelli L., Clayton D. and Montomoli C. (**1995b**). *Bayesian estimates of disease maps: How important are priors?* Statistics in Medicine. **14**: 2411-2431.

Besag J. (**1974**). *Spatial interaction and the statistical analysis of lattice systems*. Journal of the Royal Statistical Society, Series B. **36**:192-236.

Besag J., York J. and Mollié A.(**1991**). Bayesian image restoration, with applications in spatial statistics. Annals of the Institute of Statistical Mathematics. **43**: 1-59.

Best, N.G., Cowles, M.K. and Vines, S.K. (**1995**). CODA: Convergence Diagnosis and Output Analysis Software for Gibbs sampling output, Version 0.3. MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs/documentation/coda04/cdaman04.html

Box G. and Tiao G.C. (**1992**) *Bayesian Inference in Statistical Analysis*. Wiley Classics Library.Canada.

Breslow N.E. and Clayton D.G.(**1993**). *Approximate Inference in Generalized Linear Mixed Models*, Journal of the American Statistical Association. **88**: 9-25.

Brown H. and Prescott R. (**1999**). Applied Mixed Models in Medicine. John Wiley & Sons (Eds). England.

Casella G., and George E.I. (**1992**). *Explaining the Gibbs Sampler.* American Statistical Association. **46**: 167-174.

Clayton D. and Kaldor J. (**1987)**. *Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping*. Biometrics, **43**: 671-681

Clayton D.and Bernardinelli L. (**1992**). *Bayesian methods for mapping disease risks*. In small Area Studies in Geographical and Environmental Epidemiology (P.Elliot, J. Cuzich, et al) Oxford University Press. 205-20.

Cressie N.A. (**1993**). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. Canada.

Geweke, J. (**1992**). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bayesian Statistics 4, (ed. Bernardo, J. M., Berger, J.O., Dawid, A. P. and. Smith, A. F. M). Oxford, UK: Clarendon Press,.

Gilks W.R. and Wild P. (**1992**). Adaptive Rejection Sampling for Gibbs Sampling. Applied Statistics. **41**: 337-348.

Gilks W.R, Richardson S. and Spiegelhalter D.J. (**1996**). *Markov Chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman & Hall. London:

Green C., Hoppa R.D., Young T.K., Blanchard J.F.(**2003**). *Geographical analysis of diabetes prevalence in an urban area.* Social Science and Medicine. **57**: 551-560.

Krans H.M.J., Porta M., Keen H.(**1992**). *Diabetes care and research in Europe: the St Vincent Declaration action programme*. Giornale italiano de Diabetologia; **12** (suppl. 2).

Leroux, B.G, Lei X. and Breslow N. (**1999**). *Estimation of disease rates in small areas: A new mixed model for spatial dependence*. In Statistical models in epidemiology, the Environment, and Clinical Trials. (Halloran M.E, Berry D editors). Springer.

Leroux B.G. (**2000**). *Modeling spatial disease rates using maximum likelihood.* Statistics in Medicine. **19**: 2321-2332.

Lin X. (**1997**). *Variance component testing in generalised linear models with random effects*. Biometrika. **84**: 309-326.

MacNab Y.C. and Dean C.B. (**2000**). *Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models*. Statistics in Medicine, **19**: 2421-2435.

McCullagh P. and Nelder J.A. (**1989**). *Generalized Linear Models* (2nd ed.): Chapman and Hall. London

McCulloch C.E. (**1997**). *Maximum likelihood algorithms for generalized linear mixed models*. Journal of the American Statistical Association. **92**: 162-170.

McCulloch C. E. and Searle S.R. (**2001**). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. Canada.

Militino A.F., Ugarte M.D. and Dean C.B. (**2001**). *The use of mixture models for identifying high risks in disease mapping.* Statistics in Medicine. **20**: 2035-2049.

Nelder J.A. and Wedderburn R.W.M. (**1972**). *Generalized Linear Models* Journal Royal Statistical Society. **135**: 370-384.

Patterson, H.D. and Thompson, R. (**1974**). *Recovery of Interblocks Information When Block Sizes Are Unequal*, Biometrics. **44**: 1033-1048

Pinheiro J.C. and Bates D.M. (**1996**). Unconstrained parameterizations for variance-covariance matrices. Statistics and Computing. **6**: 289-296.

Rytkönen M., Ranta J., Tuomilehto J., Karvonen M., for the SPAT Study Group and the Finnish Childhood Diabetes Registry Group. (**2001**). *Bayesian analysis of geographical variation in the incidence of Type I diabetes in Finland.* Diabetologia. **44**: 37-44

Shoukri MM and Pause CA (**1998**). *Statistical Methods for Health Sciences.* CRC Press. United States of America

Songini M., Bernardinelli L., Clayton D., Montomoli C., Pascutto C., Gislandi M., Fadda D., Bottazzo G.F., and the Sardinian IDDM Study Groups. (**1998**). *The sardinian IDDM study:1. Epidemiology and geographical distribution of IDDM in Sardinia during 1989 to 1994.* Diabetologia **41**: 221-227.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (**1996**). BUGS: Bayesian inference Using Gibbs Sampling, Version 0.6. MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs.

Vonesh E.F., Chinchilli V.M., and K Pu.(**1996**). Goodness of Fit in Generalized Nonlinear Mixed-Effects Models. Biometrics, **52**: 572-587.

Vonesh E.F. and Chinchilli V.M. (**1997**). *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker Corporation. New York.

Yasui,Y. and Lele,S. (**1997**). *A Regression Method for Spatial Disease Rates - An Estimating Function-Approach.* Journal of the American Staitstical Association. **92**: 21-32