Térstatisztikai modellek alkalmazása a biztosításban

Vitéz Ildikó Ibolya
alkalmazott matematikus
Témavezető: Arató Miklós
Valószínűségelméleti és Statisztika Tanszék
Eötvös Lóránd Tudományegyetem

Valószínűségelméleti és Statisztika Tanszék
Eötvös Lóránd Tudományegyetem
Természettudományi Kar

2005. június 15.

Tartalomjegyzék

Ábrák jegyzéke

Köszönetnyilvánítás

Ezúttal mondok köszönetet témavezetőmnek, Arató Miklósnak a munka során nyújtott sok hasznos segítségért és tanácsért.

1. Bevezetés

A biztosítók a különböző termékek biztosítási díját a jövőbeli kifizetések várható értékének megfelelően számolják ki, ezért alapvető céljuk, hogy az egyes szerződők által a jövőben okozott károk számát és nagyságát előrejelezzék. Ahhoz, hogy erre minél pontosabb becslést kapjanak, a múltból rendelkezésre álló adatok alapján egy statisztikai modellt állítanak fel, melyek a szerződőről rendelkezésre álló jellemzők, és az eddig bekövetkezett károk között fennálló összefüggéseket veszik alapul. A már megfigyelt szerződések átlagos kárszáma a legegyszerűbb ilyen modell, amely azonban nem tesz különbséget az ügyfelek között, és így nem adhat igazán jó becslést az egyes szerződések várható kárszámára. Ennél jóval pontosabb becslést kaphatunk, ha figyelembe vesszük a szerződő bizonyos jellemzőit. Nem-élet biztosításban talán a leggyakrabban használt modell az általánosított lineáris modell (Generalized Linear Models, GLM), ami figyelembe tud venni kategorikus változókat (például a gépkocsi típusát), és folytonosakat is (például a gépkocsi korát), amennyiben ez utóbbiakról polinomiális hatást tételezünk fel. Ebben az esetben meg tudjuk adni, hogy milyen típusú eloszlást tételezünk fel a modellezni kívánt változóról, és azt is, hogy az eloszlás paraméterének milyen függvényét akarjuk a többi változó polinomjaként közelíteni. Egyes változóktól (ilyen például a vezető kora, neme, a motor hengerűrtartalma) már előre sejthetjük, hogy összefüggnek a károk számával. Ugyanakkor sok esetben jogosan feltételezhetjük, hogy a térbeli elhelyezkedésnek is szerepe van. Így ezt is figyelembe véve még pontosabb becslést kapunk az egyes szerződők által a jövőben okozott károk számára, ami lehetőséget ad arra, hogy a díjakat még differenciáltabban, a valódi kockázatnak méginkább megfelelően számoljuk ki. Ez a szerződők szempontjából igazságosabb díjakat eredményez, a biztosítónak pedig egy megbízhatóbb előrejelzést ad. Az utóbbi években számos cikk született ebben a témában. A fertőző betegségek terjedésének vizsgálata kapcsán merül fel talán leggyakrabban a térstatisztika fontossága, ugyanakkor más, kevésbé magától értetődő esetekben is érdemes vizsgálni a térbeliség hatását. Ilyen lehet a gépjárműbiztosítás is. Azt, hogy a biztosított autó tulajdonosa mekkora lélekszámú városban, településen él, sok helyen már figyelembe veszik a biztosítási díj számításakor. Azonban más, kevésbé kézzel

fogható tulajdonságai is lehetnek egy-egy régiónak, melyek befolyásolhatják a kockázat mértékét. Dolgozatomban azt vizsgálom, hogy Magyarországon hogyan változik a kárszám régiónként, mekkora az egyes térségek relatív kockázata. Több modellt alkalmazok egy biztosításból származó valós adatsorra, és megvizsgálom az illeszkedés mértékét. A Bayes tétel segítségével a modellben szereplő paraméterekre vonatkozó előzetes elvárásokat is figyelembe tudom venni, ilyen például az egymáshoz közeli területek térbeli hatásának hasonlósága. Az így kapott a posteriori eloszlásokból Markov Lánc Monte Carlo (Markov Chain Monte Carlo, MCMC) módszerrel veszek mintát, aminek segítségével becslést adok az egyes térségek által képviselt relatív kockázatra. Az eddig elterjedt módszerek többsége a térbeli hatást utólag elemzi, miután már egyéb változók figyelembe vételével modellt illesztettek az adatokra (lásd [4]). A hatások vizsgálatának ez a módja azonban nem egészen meggyőző, hiszen keveredik benne a gyakoriságelvű (frequentist) és a Bayes-i megközelítés, továbbá nem veszi figyelembe, hogy a térbeli hatás befolyásolhatja a többi változó hatását. Ezen problémák kiküszöbölésére születtek a [1],[2],[3] cikkek, melyekben a többi változóval együtt kezelik a regionális változót. Az ezen cikkekben leírt modellek kapcsán nem merül fel az előbb említett probléma, hisz csakis a Bayes-i megközelítést alkalmazzák. Gondot jelenthet ugyanakkor az, hogy az itt ismertett eljárásokban tömérdek adatra kell többszázezer (esetenként milliós nagyságrendű) iterációból álló lépéssorozatot alkalmazni, ami nem minden esetben oldható meg. A futásidő csökkentése azonban nem könnyű feladat, ha megfelelő megoldást akarunk adni a fentiekre. Az általam alkalmazott módszerben keveredik ugyan a kétfajta hozzáállás (pusztán a gyakoriságon alapuló, illetve Bayes-i), ugyanakkor tekintetbe veszi, hogy a különböző változók hatása egymástól nem független. További előnye az algoritmusnak, hogy belátható időn belül befejeződik. Az eredmények azt mutatják, hogy azok az értékek, melyeket ezzel az eljárással - a második típusú módszereknél gyorsabban - kaptunk meg, jobban illeszkednek a valódi adatokhoz, mint az első fajta modellekből származó becslések.

2. Modellek

Az alábbiakban paraméteres modelleket mutatok be, melyekben közös vonás, hogy az egyes szerződések kárszámáról feltesszük, hogy Poisson eloszlásúak, a szerződőre jellemző paraméterrel. A cél e paraméterek minél pontosabb meghatározása. Mindazonáltal a paraméterek a priori eloszlásának meghatározásánál egyéb szempontok is szerepet játszanak. Ilyen a regionális hatás esetén az a jogosnak tűnő feltételezés, hogy a szomszédos területek hasonló relatív kockázattal rendelkeznek. Ez a feltevés ráadásul nagyon hasznos lehet olyan esetekben, amikor egy régióra nincs bejelentett kár, hiszen így a szomszédos régiókról rendelkezésre álló adatok segítenek abban, hogy erre a régióra is reális eredményt kapjunk. Az ugyanis, hogy egy régióban az adott évben nem történt káresemény, még nem jelenti azt, hogy a tényleges várható kárszám 0 lenne, miközben azok a becslések, amelyek csak a régió tapasztalatát veszik figyelembe, 0-val becsülnék a kárszámot. A szomszédsági rendszerben viszont egy-egy régió kárszámának becslésekor a szomszédos régiókból származó adatok is számítanak. A Bayes-tétel segítségével az ilyen a priori feltevéseket is figyelembe véve egyszerűen fel tudjuk írni a keresett paraméterek a posteriori eloszlását. Mivel azonban legtöbbször a kapott sűrűségfüggvény nem azonosítható be valamely jól ismert eloszlás sűrűségfüggvényeként, a várható érték meghatározása sokszor analitikusan nem oldható meg. Ezért szimulációs technikát alkalmazunk, és elegendően nagy számú minta esetén a mintaátlagot fogadjuk el várható értékként. A mintavételhez Markov Lánc Monte Carlo módszert használunk.

2.1. A térbeli hatás utólagos vizsgálata

Az ezen fejezetben bemutatásra kerülő módszerek lényege, hogy a regionális hatás figyelmen kívül hagyásával illesztett modellből (például általánosított lineáris modellből) kapott becslést hasonlítjuk össze a valós adatokkal, és ezek ismeretében utólag adunk becslést arra, hogy az egyes régiók mekkora relatív kockázatot képviselnek. [5]-ben található e módszereknek egy részletes bemutatása. A többi változó hatását tehát adottnak tekintjük, és azokat figyelembe véve újra vizsgáljuk az adatokat a térbeliség tekintetében. Legyen $\underline{Y} = (Y_1, ..., Y_m)$ a modellezni kívánt értékek (pél-

dául a kárszámok) vektora, melynek elemei valószínűségi változók, és ahol m a régiók száma. A $\underline{z} = (z_1, ..., z_m)$ jelölje az előzetes modell alapján számolt becsült értékek vektorát, és legyen $\underline{R} = (R_1, ..., R_m)$ az egyes régiók relatív kockázatainak vektora, valószínűségi változó. Y_i eloszlását Poissonnal modellezzük, $z_i R_i$ paraméterrel:

$$[y_i \mid r_i] = e^{-z_i r_i} \frac{(z_i r_i)^{y_i}}{y_i!}$$

A $[y_i \mid r_i]$ jelölés a $P(Y_i = y_i \mid R_i = r_i)$ feltételes valószínűség helyett áll, és a dolgozat további részében is ez a rövidebb alak szerepel majd. Értelemszerűen ahol folytonos valószínűségi változóról van szó, ott a feltételes sűrűségfüggvényt jelöli a fenti alak. Az $\underline{Y}$ elemeinek $\underline{R}$-re való feltételes eloszlása egymástól független, azaz $\left[\underline{y} \mid \underline{r}\right] = \prod_{i=1}^{m} [y_i \mid r_i]$. Az R_i-k helyett ezek logaritmusát, $U_i = log(R_i)$-ket fogjuk közvetlenül becsülni a

$$[y_i \mid u_i] = e^{-z_i exp(u_i)} \frac{(z_i exp\,(u_i))^{y_i}}{y_i!} = exp\,(-z_i exp\,(u_i) + y_i u_i)\, \frac{z_i^{y_i}}{y_i!}$$

kifejezés segítségével.

2.1.1. Korrelálatlan régiók

Ha az egyes régiók relatív kockázatai közt nem tételezünk fel összefüggést, akkor az U_i-k a priori eloszlásai egymástól függetlenek. Ekkor tekintsük az U_i-k feltételes eloszlását normálisnak:

$$\left[u_i \mid u_{j,\, j \neq i},\, \mu,\, \sigma^2\right] \sim \left[u_i \mid \mu,\, \sigma^2\right] \sim N\left(\mu,\, \sigma^2\right).$$

Mivel nincs előzetes információnk a régiók hatásáról, a várható érték és a szórásnégyzet is valószínűségi változók; a μ, illetve a σ^2 hiperparaméterek a priori eloszlásai: $\mu \sim N\,(a,\, b^2)$, $1/\sigma^2 \sim Gamma(c,\, d)$. A szórásnégyzet eloszlását azért választottuk inverzgammának, mert így annak az a posteriori eloszlása is inverzgamma lesz. A Bayes-tétel alapján felírható az $\underline{U}$ paraméter, illetve a μ és a σ^2 hiperparaméterek együttes a posteriori eloszlása:

$$\left[\underline{u},\, \sigma^2,\, \mu \mid \underline{y}\right] \propto \prod_{i=1}^{m} \left[y_i \mid u_i\right] \left[u_i \mid \mu,\, \sigma^2\right] \left[\mu\right] \left[\sigma^2\right],$$

továbbá az U_i a posteriori eloszlása, mely valójában független az u_j értékektől:

$$\left[u_i \mid u_j,\, \underline{y},\, \mu,\, \sigma^2\right] \propto exp\left(-z_i exp\left(u_i\right) + y_i u_i - \frac{\left(u_i - \mu\right)^2}{2\sigma^2}\right),$$

és az alábbi összefüggéssel kifejezhető a hiperparamétereké is:

$$\left[\sigma^2 \mid \underline{u},\, \underline{y},\, \mu\right] \propto \left[\underline{u} \mid \mu,\, \sigma^2\right] \left[\sigma^2\right], \tag{1}$$

$$\left[\mu \mid \underline{u},\, \underline{y},\, \sigma^2\right] \propto \left[\underline{u} \mid \mu,\, \sigma^2\right] \left[\mu\right], \tag{2}$$

melyben az U_i-k a priori függetlensége miatt az $\left[\underline{u} \mid \mu,\, \sigma^2\right] = \prod_{i=1}^{m} \left[u_i \mid \mu,\, \sigma^2\right]$. Továbbá mivel az $\underline{Y}$-nak az $\underline{U}$-ra való feltételes eloszlása független a hiperparaméterektől, azok a posteriori eloszlásában nem szerepel az $\left[\underline{y} \mid \underline{u}\right]$ kifejezés.

2.1.2. Korrelált régiók

Mint már korábban is írtuk, sok esetben jogosan feltételezzük, hogy az egymáshoz közel fekvő helyek hasonló relatív kockázattal rendelkeznek. Definiáljunk tehát egy szomszédsági viszonyt a régiók között. Az egyik lehetőség szerint azokat a régiókat nevezzük szomszédosnak, melyeknek van közös határa. Tekinthetünk azonban szomszédosnak két térséget aszerint is hogy, központjaik egymástól mekkora távolságra fekszenek. Ha magyarországi adatokat vizsgálva ezt a távolságot 35 km-nek vesszük, akkor mindenkinek lesz szomszédja, átlagosan 5.5, ami a modell szempontjából ideális, hiszen ha az egyik régióból nincsen adat (például mert nem történt káresemény, vagy mert hiányos az adatsorunk), akkor is van miből kiindulnunk, hiszen információt jelent a szomszédos területekről rendelkezésre álló adat is. Esetünkben azért is jobb ez utóbbi szomszédsági rendszer, mert így a szomszédok száma nem mutat nagy szórást, azaz az egyes térségekre ez az érték hasonló, márpedig ez az alábbi-

akban ismertetett modellben alapvető fontosságú. Jelölje δ_i azon indexek halmazát, amely sorszámú régiók szomszédosak az i. régióval. Ekkor alkalmazható a Markov mező modell, amelyben az U_i-k a priori eloszlását, egy σ^2 hiperparaméter mellett a következőnek választjuk:

$$\left[u_i \mid u_{j,\, j\neq i},\, \sigma^2\right] \sim N(\frac{1}{\mid \delta_i \mid} \sum_{j\in\delta_i} u_j,\ \frac{\sigma^2}{\mid \delta_i \mid}). \tag{3}$$

Az U_i feltételes eloszlása tehát ez esetben is normális, a várható értéke azonban függ a szomszédos U_j-k értékétől; azok átlaga, szórása pedig a szomszédai számával fordítottan arányos. Ekkor az U_i-k együttes eloszlása a következőképpen írható fel:

$$\left[\underline{u} \mid \sigma^2\right] \propto \frac{1}{\sigma^2} exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{m} \sum_{j\in\delta_i,\, j<i} (u_i - u_j)^2\right\}.$$

Az $\underline{Y}$-ról feltettük, hogy Poisson eloszlású, így felírható az U_i-k a posteriori eloszlása:

$$\left[u_i \mid u_{j,\, j\neq i},\, \underline{y},\, \sigma^2\right] \propto exp\left\{-z_i exp\,(u_i) + y_i u_i - \frac{\mid \delta_i \mid (u_i - \bar{u}_i)^2}{2\sigma^2}\right\}$$

ahol $\bar{u_i} = \frac{1}{|\delta_i|} \sum_{j\in\delta_i} u_j$. A σ^2 azt mutatja, hogy milyen mértékű hasonlatosságot tételezünk fel a szomszédos régiók között. Ha a σ^2-t 0-nak választjuk, akkor a priori minden U_i azonos lesz. A $\sigma^2 = \infty$ választással az előző esetet kapjuk, vagyis ekkor az egyes régiók hatása egymástól független. Érdemes azonban σ^2-t is valószínűségi változónak tekinteni, így az algoritmus erre is becslést ad majd, azaz a rendelkezésünkre álló adatok segítségével megkapjuk az optimális σ^2 értéket, ahelyett, hogy mi előre meghatároznánk az értékét. A σ^2 a priori eloszlását most is érdemes inverzgammának választani, elsősorban a számolások megkönnyítése végett. A hiperparaméter a posteriori eloszlását a

$$\left[\sigma^2 \mid \underline{u},\, \underline{y}\right] \propto \left[\underline{u} \mid \sigma^2\right] \left[\sigma^2\right] \tag{4}$$

kifejezés adja meg.

2.1.3. Korrelált és korrelálatlan regionális hatás

A különböző régiók hatásának egymáshoz való viszonyát tekintve egy újabb modellt kapunk, ha az előző két módszert egybeolvasztjuk. Ezt a következő módon tehetjük meg. Ahelyett, hogy választanánk az előbbi két lehetőség közül, bontsuk fel az eddig U_i-val jelölt logaritmusát a térbeli hatásnak két, a priori egymástól független valószínűségi változó összegére: $U_i = V_i + W_i$, ahol a V_i-k eloszlása az imént ismertett Markov mező modell szerinti, míg a W_i-k egymástól független, μ várható értékű, λ^2 szórásnégyzetű normális eloszlású valószínűségi változók. Vezessük be az $\underline{\alpha}$ valószínűségi változót, mely jelölje az aktuális modellben szereplő hiperparaméterek vektorát, azaz jelen esetben: $\underline{\alpha} = (\mu, \lambda^2, \sigma^2)$. Felhasználva, hogy a $\underline{V}$, illetve a $\underline{W}$ a priori függetlenek:

$$[\underline{v}, \underline{w} \mid \underline{\alpha}] = \left[\underline{v}, \underline{w} \mid \mu, \lambda^2, \sigma^2\right] \propto \left[\underline{v} \mid \sigma^2\right] \left[\underline{w} \mid \mu, \lambda^2\right],$$

a paraméterek a posteriori együttes eloszlása a következő:

$$\left[\underline{v}, \underline{w}, \sigma^2, \mu, \lambda^2 \mid \underline{y}\right] \propto \prod_{i=1}^{m} [y_i \mid v_i, w_i] \left[\underline{v} \mid \sigma^2\right] \left[\underline{w} \mid \mu, \lambda^2\right] \left[\sigma^2\right] \left[\lambda^2\right] [\mu].$$

Írjuk fel a $\underline{V}$, illetve a $\underline{W}$ vektor elemeinek a posteriori eloszlását:

$$\left[v_i \mid v_{j,\, j\neq i}, \underline{w}, \underline{y}, \underline{\alpha}\right] \propto exp\left(-z_i exp\left(v_i + w_i\right) + y_i v_i - \frac{\mid \delta_i \mid (v_i - \bar{v}_i)^2}{2\sigma^2}\right),$$

$$\left[w_i \mid w_{j,\, j\neq i}, \underline{v}, \, , \underline{y}, \underline{\alpha}\right] \propto exp\left(-z_i exp\left(v_i + w_i\right) + y_i w_i - \frac{(w_i - \mu)^2}{2\lambda^2}\right),$$

ahol $\bar{v_i} = \frac{1}{|\delta_i|} \sum_{j\in\delta_i} v_j$, továbbá a hiperparaméterekét:

$$\left[\sigma^2 \mid \underline{v}, \underline{w}, \underline{y}, \lambda^2, \mu\right] \propto \left[\underline{v} \mid \sigma^2\right] \left[\sigma^2\right], \tag{5}$$

$$\left[\lambda^2 \mid \underline{v}, \underline{w}, \underline{y}, \sigma^2, \mu\right] \propto \left[\underline{w} \mid \mu, \lambda^2\right] \left[\lambda^2\right], \tag{6}$$

$$\left[\mu \mid \underline{v}, \underline{w}, \underline{y}, \sigma^2, \lambda^2\right] \propto \left[\underline{w} \mid \mu, \lambda^2\right] [\mu]. \tag{7}$$

Így tehát nem kell előre eldöntenünk, hogy vajon a szomszédos régiók kockázatai összefüggenek-e, és ha igen, milyen mértékben, hanem az kiolvasható az eredményből.

E módszerek egy már előzetes modellre építve vizsgálják az egyes régiók relatív kockázatait. Ezzel a megközelítéssel kapcsolatban azonban már említettünk két problémát; a kétféle statisztikai hozzáállás keveredését, illetve azt, hogy a térbeli hatás nem tud visszahatni a többi hatásra. Ezek feloldására születtek a következő módszerek.

2.2. Együttes modellillesztés

Ez esetben is az általánosított lineáris modell felépítéséből indulunk ki, azonban az itt bemutatott modellek egyszerre veszik figyelembe az összes hatást, méghozzá úgy, hogy a regionális hatást beillesztik az említett modellbe (lásd [3]). Egy nem-élet biztosításban az n szerződés közül az i. kárszámát jelöljük Y_i-vel. A kárszámok eloszlását most is Poissonnak tekintjük, szerződésenként különböző paraméterrel: $e_i\theta_i$, ahol e_i a szerződésben töltött idő (általában napokban mérve), θ_i pedig az i. szerződőre jellemző relatív kockázat. Jelölje $\underline{x}_i$ az i. szerződő egy olyan jellemzőiből álló vektorát, melyekről feltételezzük, hogy összefüggésben vannak az okozott károk számával. Ha a k. változó kategorikus, akkor az $\underline{x}_i$-ben annyi elem tartozik hozzá, ahányféle értéket a változó felvehet; a megfelelő helyen 1 áll, a többi 0 (például, ha az autó típusa a 2. kategóriába tartozik, és 5 féle kategória van, az $\underline{x}_i$ ilyen alakú; (..., 0, 1, 0, 0, 0, ...)). A folytonos változók értelemszerűen 1 helyet foglalnak. Keressük θ_i-t a $\theta_i = e^{\underline{\beta}\underline{x}_i + u_{r_i}}$ alakban, ahol U_{r_i} jelöli a regionális hatást, melyben r_i azon régió sorszámát jelöli, amelybe az i. szerződés tartozik. A $\underline{\beta}\underline{x}_i$ a többi változó hatását jelöli; kategorikus változó esetén (például a fenti k. változó), $\underline{\beta}$-ban öt elem mutatja az 5 féle kategória relatív kockázatát, ha a szóban forgó változó folytonos, akkor a $\underline{\beta}$ megfelelő eleme képviseli a változó hatását, melyről feltesszük, hogy lineáris. Az i. szerződő kárszámát tehát az alábbi alakban keressük:

$$\left[y_i \mid \underline{\beta}, \underline{u}\right] = exp\left(-e^{log(e_i)+\underline{\beta}\underline{x}_i+u_{r_i}}\right)\frac{\left(e^{log(e_i)+\underline{\beta}\underline{x}_i+u_{r_i}}\right)^{y_i}}{y_i!}.$$

Most is háromféle modellt különböztetünk meg az U_i a priori eloszlása szerint.

2.2.1. Korrelálatlan régiók

Az első modellben a különböző régiók hatását egymástól függetlennek tekintjük. Minden j-re u_j normális eloszlású μ várható értékkel - melynek a priori eloszlása normális -, illetve σ^2 szórásnégyzettel, aminek a priori eloszlása inverzgamma. Ha a $\underline{\beta}$-ról nincs előzetes információnk, vagy elvárásunk, akkor eloszlását tekinthetjük a $(-\infty, +\infty)$ intervallumon egyenletesnek, ami ugyan a priori nem egy valódi eloszlás, a posteriori mégis értelmes eredményt ad. Ekkor az U_i-k illetve a $\underline{\beta}$ a posteriori eloszlása a Bayes-tétel alapján felírható fel:

$$\left[u_i \mid u_{j,\, j\neq i},\, \underline{\beta},\, \underline{y},\, \mu,\, \sigma^2\right] \propto \prod_{k=1,\, R_k=i}^{n} exp\left\{-e_k exp\left(\underline{\beta}\underline{x}_k + u_i\right) + y_k u_i - \frac{(u_i-\mu)^2}{2\sigma^2}\right\},$$

$$\left[\underline{\beta} \mid \underline{u},\, \underline{y},\, \mu,\, \sigma^2\right] \propto \prod_{i=1}^{n} exp\left\{-e_i exp\left(\underline{\beta}\underline{x}_i + u_{r_i}\right) + y_i \underline{\beta}\underline{x}_i\right\}.$$

A hiperparaméterek eloszlása az előző fejezetben leírtakéhoz hasonlóan fejezhető ki ezen esetekben is, lásd (1)-(2), (4), (5)-(6)-(7) kifejezés, ezeket a továbbiakban nem részletezzük.

Ha a szomszédos régiókat bizonyos tekintetben összefüggőnek, korreláltnak tekintjük, azt ismét az U_i-k a priori eloszlásában tudjuk jelezni.

2.2.2. Korrelált régiók

A második esetben feltételezzük, hogy egy régióra jellemző U_i függ a vele szomszédos régiókhoz tartozó U_j-któl. Ebben az esetben a $\underline{U}$ eloszlására felírható a már ismertetett Markov mező modell,

$$\left[\underline{u} \mid \sigma^2\right] \propto exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j\in\delta_i,\, j<i}(u_i-u_j)^2\right\},$$

mellyel a paraméterek a posteriori eloszlása:

$$\left[u_i \mid u_{j,\,j\neq i},\, \underline{\beta},\, \underline{y},\, \sigma^2\right] \propto \prod_{k=1,\,R_k=i}^{n} exp\left\{-e_k exp\left(\underline{\beta}\underline{x}_k + u_i\right) + y_k u_i - \frac{\mid \delta_i \mid (u_i - \bar{u}_i)^2}{2\sigma^2}\right\},$$

ahol a $\bar{u}_i$ a korábbi definíció szerint a szomszédos hatások átlaga,

$$\left[\underline{\beta} \mid \underline{u},\, \underline{y},\, \sigma^2\right] \propto \prod_{i=1}^{n} exp\left\{-e_i exp\left(\underline{\beta}\underline{x}_i + u_{r_i}\right) + y_i \underline{\beta}\underline{x}_i\right\}.$$

Végül pedig most is megtehetjük, hogy az előző két esetet egyszerre beillesztjük a modellbe.

2.2.3. Korrelált és korrelálatlan regionális hatás

A harmadik eset tehát az előző kettő egybeolvasztása. Írjuk fel a U-t két független valószínűségi változó összegeként, épp úgy, mint a 2.1.3-as fejezetben: $U_i = V_i + W_i$. Ekkor az a posteriori együttes eloszlásunk:

$$\left[\underline{v},\, \underline{w},\, \underline{\beta},\, \underline{\alpha} \mid \underline{y}\right] \propto exp\left\{\sum_{i=1}^{n}\left[-e_i exp\left(\underline{\beta}\underline{x}_i + v_{r_i} + w_{r_i}\right) + y_i\left(\underline{\beta}\underline{x}_i + v_{r_i} + w_{r_i}\right)\right]\right\}$$

$$\times exp\left\{\sum_{i=1}^{n}\left[-\frac{(w_i - \mu)^2}{2\lambda^2} - \frac{1}{2\sigma^2}\sum_{j\in\delta_i,\,j<i}(v_i - v_j)^2\right]\right\}\left[\sigma^2\right]\left[\lambda^2\right][\mu],$$

melyből a paraméterek a posteriori eloszlása:

$$\left[v_i \mid v_{j,\,j\neq i},\, \underline{w},\, \underline{\beta},\, \underline{y},\, \underline{\alpha}\right] \propto \prod_{k=1,\,R_k=i}^{n} exp\left\{-e_k exp\left(\underline{\beta}\underline{x}_k + v_i + w_i\right) + y_k v_i - \frac{\mid \delta_i \mid (v_i - \bar{v}_i)^2}{2\sigma^2}\right\},$$

$$\left[w_i \mid w_{j,\,j\neq i},\, \underline{v},\, \underline{\beta},\, \underline{y},\, \underline{\alpha}\right] \propto \prod_{k=1,\,R_k=i}^{n} exp\left\{-e_k exp\left(\underline{\beta}\underline{x}_k + v_i + w_i\right) + y_k w_i - \frac{(w_i - \mu)^2}{2\lambda^2}\right\},$$

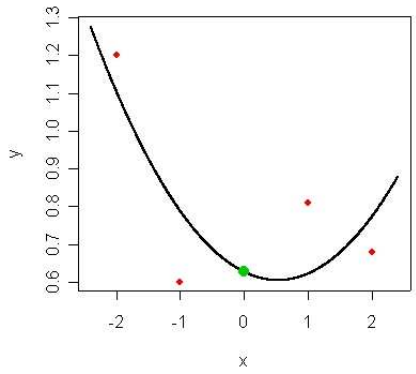
$$\left[\underline{\beta} \mid \underline{v},\, \underline{w},\, \underline{y},\, \underline{\alpha}\right] \propto \prod_{i=1}^{n} exp\left\{-e_i exp\left(\underline{\beta}\underline{x}_i + v_{r_i} + w_{r_i}\right) + y_i \underline{\beta}\underline{x}_i\right\}.$$

2.3. Továbbfejlesztett modellek

A továbbiakban bemutatok még két eljárást, melyek tovább pontosíthatják a kárszámok előrejelzését. Induljunk ki az utolsó modellből; a regionális hatás tehát két, egymástól független összetevőből áll, melyekről azonban most még azt is feltételezzük, hogy kiegyenlítettek abban az értelemben, hogy mind az egymástól független hatások, azaz a fenti jelölést használva a W_i-k, mind az összefüggőek; a V_i-k összege 0. Az [1] cikkben találhatjuk ennek egy gyakorlati megvalósítását, sőt e cikk további újdonsága, hogy más változók hatásának pontosabb becslésével tovább finomít a modellen. Megtehetjük például, hogy nemcsak az egyes régiók térbeli relatív kockázatában tételezünk fel összefüggést, hanem a korcsoportokéban is. Az eddig ismertett módszerekben egy-egy változó (a regionálistól eltekintve) vagy faktorváltozó, ekkor a különböző csoportok egymástól független relatív kockázatot képviselnek, vagy folytonos, amikoris a hatásukat polinommal modelleztük. A következő két modell e két esetet fejleszti tovább. Az első célja, hogy bizonyos kategorikus változók esetében figyelembe vegye azt az előzetes elvárást, hogy az egymáshoz közeli értékekhez tartozó relatív kockázat hasonló. Ez jogos feltételezés például a kor esetében. Ha korcsoportokkal dolgozunk, melyek beosztása elég sűrű, például minden évre külön jut egy, akkor érdemes a következő módszert alkalmazni (lásd [1]). Jelöljük az i. korcsoport relatív kockázatát γ_i-vel. Ezek a priori eloszlása legyen a következő:

$$\gamma_i - 2\gamma_{i-1} + \gamma_{i-2} \sim N(0,\, \sigma_\gamma^2).$$

Azaz a γ_i-ről azt várjuk, hogy a γ_{i-1}-nél annyival lesz több, mint amennyivel a γ_{i-1} több volt a γ_{i-2}-nél. Kicsit szemléletesebben úgy is elképzelhetjük, hogy a γ_i feltételes várható értéke a γ_{i-2}, γ_{i-1}, γ_{i+1}, γ_{i+2}-ek ismerete mellett megkapható a $(-2,\, \gamma_{i-2})$, $(-1,\, \gamma_{i-1})$, $(1,\, \gamma_{i+1})$, $(2,\, \gamma_{i+2})$ pontokra illesztett másodfokú polinom 0-ban felvett értékeként.



1. ábra. Négyzetes illesztés

A másik módszer a folytonos változók hatását próbálja pontosabban közelíteni. A [2] cikkben szintén a kor az a változó, melynek hatását közelebbről vizsgálják a térbeliség mellett. A kort itt folytonos változóként kezelik, és hatását polinom helyett az annál sokkal rugalmasabb spline-nal közelítik. Persze nem csak az életkornak, hanem például az autó gyártási idejének, vagy a hengerűrtartalomnak is feltételezhetjük folytonos függvénnyel leírható, nem feltétlenül polinomiális hatását. Ha η_i-vel jelöljük az i. szerződő Poisson-paraméterének logaritmusát, akkor a modellünket így írhatjuk fel:

$$\eta_i = \tau_0 + \underline{\tau}_1 x_{nem,i} + \underline{\tau}_2 x_{auto,i} + ... + log(e_i) + f_{kor}(x_{kor,i}) + f_{henger}(x_{henger,i}) + ... + f_{reg}(x_{reg,i}).$$

Ahol a nem, autótípus, ... faktorok, így például az $x_{nem,i}$ egy három elemű indikátorvektor; ha a szerződő férfi, akkor a vektor első eleme 1, a többi 0, ha a vezető nő, akkor a második elem értéke 1, ha céges autó akkor a harmadik elemé 1. Ugyanakkor az f_k függvényekben szereplő változók, mint a kor, hengerűrtartalom, ... folytonos változók. A térbeli hatás becslését a folytonos változókéval együtt végezzük majd. Az f_k-kat spline-okkak közelítjük. Ez egy könnyen kezelhető modellhez vezet, ugyanakkor jobb közelítést adhat, mintha polinomokkal dolgoznánk. Legyen $x_{j,min} = \xi_0 < \xi_1 < ... < \xi_{r_j} = x_{j,max}$, a j. változó lehetséges értékeit tartalmazó

intervallum egy felosztása, valamint adjunk meg további l_j (a spline fokszáma) osztópontot a $x_{j,min} = \xi_0$ előtt: ξ_{-l_j}, ξ_{-l_j+1}, ..., ξ_{-1}. Egy ezen a felosztáson definiált spline a következő tulajdonságokkal rendelkezik: egy (ξ_i, ξ_{i+1}) intervallumon l_j-edfokú polinom, az osztópontokban pedig $(l_j - 1)$-szer folytonosan differenciálható. Egy f_j spline felírható $m_j = r_j + l_j$ darab alapspline lineáris kombinációjaként:

$$f_j\left(x_{j,i}\right) = \sum_{k=0}^{m_j} \beta_{j,k} B_{j,k}^{l_j}\left(x_{j,i}\right)$$

ahol a $B_{j,k}^{l_j}$ alapsplineok rekurzió segítségével könnyen számolhatók:

$$B_{j,k}^{0}(x) = \begin{cases} 1 & ha\ \xi_k \leq x < \xi_{k+1} \\ 0 & egy. \end{cases}$$

$$B_{j,k}^{l}(x) = \frac{x - \xi_k}{\xi_{k+l} - \xi_k} B_{j,k}^{l-1}(x) + \frac{\xi_{k+l+1} - x}{\xi_{k+l+1} - \xi_{k+1}} B_{j,k+1}^{l-1}(x) \qquad l \geq 1.$$

Az n szerződő $x_{j,i}$ $(i = 1..n)$ értékeinek f_j-ben felvett értékeit jelölje $f_{j,i} = f_j\left(x_{j,i}\right)$, ezek vektorát - $f_j = (f_{j,1}, f_{j,2}, ..., f_{j,n})$ - mátrixalakban meg tudjuk adni: $f_j = X_j \beta_j$, ahol $\beta_j = \left(\beta_{j,1}, \beta_{j,2}, ..., \beta_{j,m_j}\right)$ a j. változóhoz tartozó spline együtthatóinak becsült értékei, és X_j egy $n \times m_j$-s mátrix, melynek elemei a $B_{j,\cdot}^{l_j}$ alapspline-oknak a szerződések $(i.)$, megfelelő $(j.)$ változóiban felvett értékei: $X_j(i,k) = B_{j,k}^{l_j}\left(x_{i,j}\right)$. Ebbe a jelölésbe a regionális változót egyszerűen tudjuk majd beilleszteni, és együtt kezelni azt a folytonos változókkal. Mivel azt szeretnénk, ha a spline-ok kellően simák lennének, ezért egy adott változó esetén az egymáshoz közel fekvő alapspline-ok együtthatóit hasonlónak szeretnénk tudni. Ezért a következőt feltesszük:

$$\beta_{j,k} - \beta_{j,k-1} \sim N(0, \sigma^2).$$

Azaz elvárjuk, hogy az $k.$, illetve a $(k-1).$ osztópontban induló alapsplineok együtthatói ne térjenek el nagyon egymástól. Vegyük észre, hogy itt is a Markov mezőt alkalmaztuk:

$$\beta_{j,k} \sim N\left(\frac{\beta_{j,k-1} - \beta_{j,k+1}}{2}, \sigma^2\right),$$

ezért a j. spline-hoz tartozó $\beta_{j,k}$-k együttes eloszlása a következő módon írható fel:

$$\left[\underline{\beta}_j \mid \sigma_j^2\right] \propto exp\left(-\frac{1}{2\sigma_j^2}\underline{\beta}_j' K_j \underline{\beta}_j\right), \tag{8}$$

ahol K_j egy $m_j \times m_j$-s mátrix, mely a $K_j = D_j' D_j$ alakban áll elő, ahol D_j az elsőrendű differenciamátrix. A σ_j^2 hiperparaméter a priori eloszlását itt is érdemes inverzgammának választani. Nézzük most, hogy hogyan lehet a térbeli hatás paramétereit is ilyen alakban felírni. Tekintsük a már részletezett Markov mező modellt a β_{reg}-ekre:

$$\left[\beta_{reg,i} \mid \beta_{reg,j},\, i \neq j,\, \sigma^2\right] \sim N\left(\frac{1}{\mid \delta_i \mid}\sum_{j\in\delta_i}\beta_{reg,j},\, \frac{\sigma_{reg}^2}{\mid \delta_i \mid}\right).$$

A regionális hatás felírható az előző mátrixalakban: $f_{reg} = X_{reg}\beta_{reg}$, ha bevezetjük a következő jelöléseket: X_{reg} legyen egy $n \times m$-es incidenciamátrix: az i. sor j. elem 1, ha az i. szerződés a j. régióból való. A 8 alak felírásában a K_{reg} mátrix ez esetben a szomszédsági mátrix, mely az átlós elemeiben az egyes régiók szomszédszámát tartalmazza, míg egy $(i,\, j)$, $i \neq j$ eleme -1, ha az i., és a j. régiók szomszédosak, egyébként 0. Ezzel a jelöléssel egységes lett a modell, és egyszerűen programozható a Markov Lánc Monte Carlo mintavétel is.

Láthatjuk tehát, hogy sokféle módszer létezik a feladat megoldására, és többnyire ezek azonos megfontolásokra épülnek; az azonos típusú, valamilyen szempont szerint egymáshoz közel álló változókhoz tartozó paraméterek közötti feltételezett összefüggéseket a Markov mező modellel valósítják meg, és mivel az esetek többségében az a posteriori eloszlás analitikusan kezelhetetlen, ezért a becslésekhez Monte Carlo módszer segítségével vesznek mintát.

3. Markov Lánc Monte Carlo mintavétel

Ahhoz, hogy az előző fejezetben kapott a posteriori eloszlásokból mintát tudjunk venni, a sűrűségfüggvény bonyolultsága miatt Markov Lánc Monte Carlo szimulációs technikát kell alkalmazni, melynek leírását, és elméleti hátterét megtaláljuk a [6], illetve [4] könyvekben. Mivel a paraméterek esetében 1 komponensű Metropolis-Hastings algoritmust használok, valamint a σ^2 hiperparaméterből Gibbs-lépések segítségével veszek mintákat, ezért e fejezet célja elsősorban ezek rövid ismertetése.

A számítógépek könnyedén szolgáltatnak véletlen mintát bizonyos jól ismert eloszlásokból, mint például a normális, egyenletes, béta, gamma ... A Markov Lánc Monte Carlo módszer célja, hogy ezek felhasználásával olyan bonyolultabb eloszlásokból is tudjunk véletlen értékeket generálni, melyeknek ismert a sűrűségfüggvénye (konstans szorzó erejéig).

3.1. Elméleti háttér

Ahhoz, hogy egy f sűrűségfüggvényű eloszlásból mintát vegyünk, nem szükséges közvetlenül az f-ből végezni a szimulációt. Az alapötlet az, hogy használjunk egy ergodikus (pozitív visszatérő és irreducibilis) Markov-láncot, melynek stacionárius eloszlását az f adja meg. Ha egy ilyen Markov-láncot figyelünk, elegendő idő elteltével azt tapasztaljuk, hogy a lánc elemeinek eloszlása közel f szerinti. A Markov-lánc ezen tulajdonságát kihasználva olyan eloszlásokból is tudunk véletlen mintát venni, melyekből analitikusan lehetetlen.

Diszkrét idejű, folytonos állapotterű Markov-lánc

Definíció: Legyen X_n diszkrét idejű folyamat, $\mathcal{X}$ mérhető halmaz az állapotterünk, és $\mathcal{B}(\mathcal{X})$ egy, az $\mathcal{X}$-en végesen generált σ-algebra. Ekkor egy $K = K(x, A)$ leképezést, ahol $x \in X$, $A \in \mathcal{B}(\mathcal{X})$, *átmeneti valószínűségmag*nak nevezünk, ha

(i) $\forall x \in X$ esetén $K(x, .)$ valószínűségi mérték

(ii) $\forall A \in \mathcal{B}(X)$ esetén $K(., A)$ x-ben mérhető és nemnegatív.

Folytonos állapottér esetén a feltételes *átmeneti sűrűségfüggvény*t jelöljük k-val,

$k(x,\, x')$-re teljesül a következő:

$$P(X_{n+1} \in A \mid X_n = x) = K(x,\, A) = \int_A k(x,\, x')\, dx'.$$

Definíció: X_n legyen sztochasztikus folyamat egy $\mathcal{X}$ állapottéren. Ez *Markov-lánc*, ha

$$P(X_{n+1} \in A \mid X_0 = x_0,\, ...,\, X_n = x_n) = P(X_{n+1} \in A \mid X_n = x_n)$$

$$= \int_A k(x_n,\, x')\, dx'$$

A Markov-lánc *homogén*, ha

$$P(X_{n+1} \in A \mid X_n = x) = K(x,\, A)$$

független n-től minden x, A mellett.

Definíció: π σ-véges mérték *invariáns* az átmeneti valószínűségmagra, ha $\forall B \in \mathcal{B}(\mathcal{X})$-re

$$\pi(B) = \int_{\mathcal{X}} K(x,\, B)\, \pi(dx).$$

Megj.: Ha a π egyben valószínűségi mérték is, akkor *stacionárius eloszlásnak* nevezik, mivel ha az $X_0 \sim \pi$, akkor $X_n \sim \pi$, bármely n-re.

Definíció: Legyen adott egy ψ mérték. Az X_n lánc *ψ-irreducibilis*, ha $\forall A \in \mathcal{B}(\mathcal{X})$, melyre $\psi(A) > 0$, létezik olyan pozitív n, hogy $K^n(x,\, A) > 0$ minden $\mathcal{X}$-beli x esetén.

Definíció: Legyen az X_n A-ban való *megjelenéseinek száma*:

$$\eta_A = \sum_{n=1}^{\infty} \chi_{(X_n \in A)},$$

mely azt mutatja meg, hogy hányszor vett fel az X_n A-beli értéket.

Definíció: Az A halmaz *Harris-visszatérő*, ha $P(\eta_A = \infty \mid x_1 = x) = 1$ az A minden x elemére. Az X_n lánc *Harris-visszatérő*, ha létezik olyan ψ mérték, melyre

az X_n irreducibilis, és minden A-ra, melyre $\psi(A) > 0$, az A halmaz Harris-visszatérő.

Tétel: Legyen $S_n(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. Ekkor ha az X_n lánc Harris-visszatérő, akkor $g \in L^1(\pi)$ esetén igaz a következő:

$$\lim_{n\to\infty} S_n = \int_{\mathcal{X}} g(x)\,\pi(dx).$$

A tétel alapján tehát ahhoz, hogy egy f sűrűségfüggvényű eloszlás valamilyen g függvényének a várható értékét megkapjuk, generálnunk kell egy Markov-láncot, mely kielégíti a fenti feltételt, és mely stacionárius eloszlásának sűrűségfüggvénye éppen f. Az ebben a fejezetben tárgyalt módszer lényege, hogy egy tetszőleges x_0 kezdőértékből kiindulva a megfelelő átmeneti valószínűségmag segítségével egy olyan Markov-láncot generálunk, amely eloszlásban konvergál az f által meghatározott eloszláshoz. Így a $g = id$ választással a fenti tétel alapján elegendő a Markov-lánc kellő hosszúságú "beégetési idő" után megfigyelt elemeinek átlagát venni ahhoz, hogy becslést kapjunk a keresett várható értékre.

Definíció: Egy f sűrűségfüggvényű eloszlásból való mintavételt célzó *Markov Lánc Monte Carlo módszer* alatt értünk minden olyan módszert, mely egy olyan ergodikus Markov-láncot szolgáltat, melynek stacionárius eloszlása f sűrűségfüggvényű.

3.2. Gyakorlati megvalósítás

A feladatunk tehát, hogy olyan Markov-láncot állítsunk elő, mely teljesíti a tétel feltételeit, és stacionárius eloszlását az f adja meg. Az átmeneti valószínűségmag meghatározásnál erre a két dologra kell ügyelnünk. A következő pontban e probléma megoldására láthatunk egy módszert.

3.2.1. Metropolis-Hastings algoritmus

Ebben az algoritmusban az f függvény mellett választanunk kell egy $q(z \mid x)$ feltételes eloszlást (proposal eloszlás[1]), amelyre teljesül, hogy $q(. \mid x)$-ből könnyen tudunk mintát venni. Jelölje $x[t]$ a Markov-lánc t. elemét. Az algoritmus a következő:

Legyen adott $x[t]$

1. Generáljunk egy $z \sim q(. \mid x[t])$ értéket, ezt javasoljuk $x[t+1]$-nek.
2. Legyen

$$x[t+1] = \begin{cases} z & \alpha \; vlszggel \\ x[t] & (1-\alpha) \; vlszggel \end{cases}$$

ahol

$$\alpha = \alpha(x[t], z) = min\left\{\frac{f(z)}{f(x[t])}\frac{q(x[t] \mid z)}{q(z \mid x[t])}, 1\right\}.$$

A $q(. \mid x[t])$ eloszlásból származó javasolt értéket tehát csak egy bizonyos valószínűséggel fogadjuk el, egyébként a lánc előző elemét választjuk újra.

E két lépést addig ismételjük felváltva, amíg a kapott $x[t]$-k eloszlása közel f-szerinti nem lesz. Ezután tovább folytatva az algoritmust már úgy tekinthetjük, mintha tényleg az f-ből vettünk volna mintákat. Ha a q-t szimmetrikusnak választjuk abban az értelemben, hogy $q(z \mid x) = q(x \mid z)$, azzal jelentősen meggyorsítjuk a számolást, hiszen ebben az esetben az $\alpha = min\left\{\frac{f(z)}{f(x[t])}, 1\right\}$. Sokszor ezért érdemes normális eloszlást választani, hiszen ebből egyszerű mintát venni, és szimmetrikus.

Könnyen beláthatjuk, hogy ez a módszer a kívánt eloszlásból szolgáltat mintákat. Az α definíciója alapján

$$f(x[t])\, q(x[t+1] \mid x[t])\, \alpha(x[t], x[t+1]) = \tag{9}$$

$$= f(x[t+1])\, q(x[t] \mid x[t+1])\, \alpha(x[t+1], x[t]),$$

[1]Sajnos a szakirodalomban nem találtam a proposal eloszlásnak magyar megfelelőjét, így a dolgozatban a javaslati eloszlás megnevezést használom majd.

hiszen a bal oldalt kifejtve kapjuk a következő kifejezést

$$f\left(x\left[t\right]\right)q\left(x\left[t+1\right]\mid x\left[t\right]\right)min\left\{\frac{f(x\left[t+1\right])}{f(x\left[t\right])}\frac{q(x\left[t\right]\mid x\left[t+1\right])}{q(x\left[t+1\right]\mid x\left[t\right])},\,1\right\}=$$

$$min\left\{f(x\left[t+1\right])q(x\left[t\right]\mid x\left[t+1\right]),\,f\left(x\left[t\right]\right)q\left(x\left[t+1\right]\mid x\left[t\right]\right)\right\},$$

ami $x\left[t\right]$, $x\left[t+1\right]$-ben szimmetrikus, így a jobb oldal is ilyen alakra hozható. Írjuk fel az $x\left[t+1\right]$-nek $x\left[t\right]$-re való feltételes átmenetvalószínűségét:

$$p\left(x\left[t+1\right]\mid x\left[t\right]\right)=q\left(x\left[t+1\right]\mid x\left[t\right]\right)\alpha\left(x\left[t\right],\,x\left[t+1\right]\right)+$$

$$\mathcal{I}_{x[t+1]=x[t]}\left[1-\int q\left(y\mid x\left[t\right]\right)\alpha\left(x\left[t\right],\,y\right)\,dy\right].$$

Az egyenlőséget $f\left(x\left[t\right]\right)$-vel beszorozva, az (9) egyenlőség felhasználásával kapjuk, hogy

$$f\left(x\left[t\right]\right)p\left(x\left[t+1\right]\mid x\left[t\right]\right)=f\left(x\left[t+1\right]\right)p\left(x\left[t\right]\mid x\left[t+1\right]\right).$$

Így ha feltesszük, hogy az $x\left[t\right]$ az f sűrűségfüggvényű eloszlásból származik, akkor az $x\left[t+1\right]$ eloszlására kapjuk:

$$\int f\left(x\left[t\right]\right)p\left(x\left[t+1\right]\mid x\left[t\right]\right)\,dx\left[t\right]=f\left(x\left[t+1\right]\right).$$

Ezzel beláttuk, hogy az f sűrűségfüggvényű eloszlás stacionárius eloszlása a fent definiált Markov-láncnak. Belátható továbbá, hogy a Metropolis-Hastings algoritmusból kapott lánc f-irreducibilis, és Harris-visszatérő, tehát teljesíti a tétel feltételeit. Az eljárás helyességének részletes bizonyítása a ([6]) könyvben található.

Ha az eloszlás, amiből mintákat akarunk venni többdimenziós, az eljárás akkor is alkalmazható. Ha a dimenziószám nem túl nagy, akkor a leírt algoritmus jól működik, de nagy dimenziószám esetén érdemes kisebb blokkokra vágni a valószínűségi változó vektorát. Speciálisan megtehetjük, hogy a valószínűségi változó minden egyes eleméből külön-külön veszünk mintát. Legyen $\underline{X}=(X_1,\,...,\,X_k)$ a valószínűségi változó, és vezessük be a következő jelölést: $\underline{X}_{-j}=(X_1,\,...,\,X_{j-1},\,X_{j+1},\,...,X_k)$, jelölje továbbá $x_i\left[t+1\right]$ az X_i-re a $(t+1)$-edik iterációban kapott minta értékét, és

$\underline{x}_{-j}[t+1] = (x_1[t+1], ..., x_{j-1}[t+1], x_{j+1}[t], ..., x_k[t])$. A $(t+1)$-edik itrációban a j. elemre a következő módon kaphatunk mintát: a j. elem javaslati eloszlása legyen $q_j(z \mid x_j[t], \underline{x}_{-j}[t+1])$, melyből vett véletlen szám elfogadásának valószínűsége:

$$\alpha = min\left\{\frac{f_j(z \mid \underline{x}_{-j}[t+1])}{f_j(x_j[t] \mid \underline{x}_{-j}[t+1])}\frac{q_j(x_j[t] \mid z, \underline{x}_{-j}[t+1])}{q_j(z \mid x_j[t], \underline{x}_{-j}[t+1])}, 1\right\},$$

ahol az f_j a marginális sűrűségfüggvény.

3.2.2. Gibbs lépések

A Gibbs-féle eljárás gyakorlatilag az 1-komponensű Metropolis-Hastings módszer egy speciális esete. Akkor tudjuk alkalmazni, ha az (a posteriori eloszlás) marginálisa egy olyan ismert eloszlás, amiből könnyen tudunk véletlen számot generálni. Ekkor a javaslati eloszlás maga a marginális eloszlás:
$q_j(z \mid x_j[t], \underline{x}_{-j}[t+1]) = f_j(z \mid \underline{x}_{-j}[t+1])$. A Gibbs lépésekben így minden iterációban elfogadjuk a proposal eloszlásból származó z-t, mint újabb mintát.

A feladat megoldásakor mindkét módszerre szükség lesz.

4. Egy gépkocsibiztosításból származó adatsor

Térjünk rá a dolgozatban vizsgált adatok jellemzésére. A gépjárműbiztosításból származó adatsorunk tartalmazza a 323 808 szerződés kockázatban töltött időtartamát, a bejelentett károk számát, a vezető korcsoportját, nemét, az autó hengerűrtartalmát, azt, hogy mikor kötötték a szerződést, a helység lélekszámát, a települést, valamint a régiót. Tartalmazza továbbá a bonusfaktort, aminek figyelembe vételével módosítottam a biztosításban töltött időt, és nem ezt tekintettem e_i-nek, hanem ennek a bonusfaktorral beszorzott értékét, hiszen a bonusfaktor egy, az előzetes tapasztalatokra alapozott szorzótényező, mely az egyes (már ismert) vezetők relatív kockázatát jelzi. Az eredeti adatok közt szerepel még a vezető születési éve is, így nemcsak a korcsoportok állnak rendelkezésre, hanem a kor maga is. Ez azért hasznos, mert így mód nyílik arra, hogy ellenőrizzük, valóban jó-e a korcsoportok beosztása. Az adatok elemzésénél kiderült, hogy érdemes a 25 év alattiak kategóráját ketté bontani egy 20 évesnél fiatalabbak, illetve egy 20 és 25 év közöttiek kategóriájára, ugyanis e két csoportban az okozott károk számát tekintve jelentős eltérés figyelhető meg. A lineáris modell illesztésekor a 25 év alatti korcsoportra 382-t kaptam várható kárszámnak, és valóban, ennyi eset történt. Ugyanakkor ha megnézzük, hogy hány eset tartozik a 20 év alatti korcsoportba, illetve mennyi a 20 és 25 év közöttiekébe, akkor azt tapasztaljuk, hogy míg az utóbbiban a modellből kapott 347 jóval több, mint a valóságban bekövetkezett 327, addig a fiatalabbaknál éppen fordítva a modell alábecsüli a kárszámot: 35-t ad, míg valójában 55 volt. Hasonló a helyzet az 50 éven felüliek esetében is. Itt a 70 év alatti vezetőket - akiknél a kapott 2359-nél jóval kevesebb, 2209 káresemény történt - érdemes elválasztani a 70 éven felüliektől, ahol viszont a bekövetkezett 437 kárt jócskán alábecsülve a modell 359-t adott. Így kaptam az eredetileg 6 féle csoport helyett 8-at. Nem szerint 3 kategória van, a harmadik a valamely cég nevén lévő autókat jelöli. 5 autótípust különböztetünk meg a hengerűrtartalom szerint, továbbáb a település lélekszámát tekintve 10 kategória van. Aszerint, hogy a szerződés mennyire új, 3 eset lehetséges, 1-es jelzi az 1998, vagy azelőtt kötött szerződéseket, 2-es az 1999-ben, illetve 3-as a 2000-ben kötötteket. A régiók száma 168, a településeké 3307. Bár az eredeti adatsorban volt még sok más

jellemzője a szerződéseknek, hogy ha túl sok adattal akarnánk egyszerre dolgozni, az személyi számítógép mellett nagyon lelassítja a számolásokat. Ezért ahhoz, hogy eldöntsem, mely változókat veszem bele a modellbe, előzetesen egy általánosított lineáris modellt illesztettem az adatokra R programcsomag segítségével. Így maradtak az igazán szignifikáns változók, melyeket az előbbiekben felsoroltam.

A modell tehát, amit ebben a konkrét esetben alkalmaztam, a következő. A fenti jelölést használva tekintsük Y_i-t, az i. szerződő kárszámát Poisson eloszlásúnak, $e_i\lambda_i$ együtthatóval (e_i az i. szerződés biztosításban töltött napjainak a bonusfaktorral módosított száma). Jelölje ν_i az i. Poisson-paraméter logaritmusát; $\nu_i = log(e_i\lambda_i) = log(e_i) + log(\lambda_i)$. A ν_i értékét a következő alakban keressük:

$$\nu_i = \beta_0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i) + u_{r_i}$$

ahol $\underline{\beta}^{kor}$ egy 8 elemű vektor, benne az egyes korcsoportok relatív kockázatát mutató együtthatókkal, k_i jelzi, hogy a i. szerződő hanyadik korcsoportba tartozik. Hasonlóan értelmezhető a többi jelölés is, az sz a szerződéskötés időpontjára vonatkozik. Az első 5 változóhoz tartozó $\underline{\beta}$, valamint az β_0 a priori eloszlását egyenletesnek tekintem a $(-\infty, +\infty)$ intervallumon, a régiók hatásáról viszont feltételezem, hogy nem függetlenek. Így a Markov mező modellt alkalmazom, a szomszédsági viszonyt a régióközéppontok egymástól való távolság határozza meg. Mivel a szomszédok közti hasonlóság mértékétől nincs plusz információ, ezért a σ^2 paramétert is valószínűségi változónak tekintem; a, b paraméterű inverzgamma eloszlással. Így annak az a posteriori eloszlása is inverzgamma eloszlású. Felteszem továbbá, hogy az Y_i-k a paraméterek ismerete mellett feltételes függetlenek, és a , β^0, $\underline{\beta}^{kor}$, $\underline{\beta}^{nem}$, $\underline{\beta}^{auto}$, $\underline{\beta}^{sz}$, $\underline{\beta}^{pop}$, $\underline{U}$ a priori eloszlásaik sem függenek egymástól. A keresett paraméterek, és a hiperparaméter együttes a posteriori eloszlása tehát:

$$[\underline{\beta}, \sigma^2, \underline{u} \mid \underline{y}] \propto \prod_{i=1}^{n} [y_i \mid \underline{\beta}_i, \underline{u}_i][\underline{u}_i \mid \sigma^2][\sigma^2].$$

Ahol a $\underline{\beta}_i$ az i. szerződés változóinak megfelelő β értékek vektora. Ebbe behelyette-

sítve a fent ismertetett a priori eloszlásokat a következőt kapjuk:

$$\prod_{i=1}^{n} exp(-exp(\beta^0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i) + u_{r_i}))$$

$$\times exp(y_i(\beta^0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i) + u_{r_i}))$$

$$\times (\sigma^2)^{-\frac{168}{2}} exp(-\frac{1}{2\sigma^2}\underline{u}' K \underline{u})(\sigma^2)^{-(a-1)} exp(-b\sigma^{-2}).$$

A K egy 168×168-as mátrix, melynek elemei: $k_{r,r} = \mid \delta_r \mid$, $k_{r,s} = -1$, ha az r. és az s. régiók szomszédosak, egyébként $k_{r,s} = 0$. Az ebből a sűrűségfüggvényből kapott a posteriori eloszlások:

$$[\beta_j^{kor} \mid \underline{\beta}^k, \, _{k \neq kor}, \, \underline{u}, \, \underline{y}] \propto \prod_{i=1, k_i=j}^{n} exp(-exp(\beta^0 + \beta_j^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i) + u_{r_i}))$$

$$\times exp(y_i \beta_j^{kor}),$$

hasonlóan felírható a többi β-ra is.

$$[u_j \mid u_k, \, _{k \neq j}, \, \underline{\beta}, \, \sigma^2, \, \underline{y}] \propto \prod_{i=1, r_i=j}^{n} exp(-exp(\beta^0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i) + u_j)$$

$$\times exp(y_i u_j) \times exp(-\frac{1}{2\sigma^2}\underline{u}' K \underline{u}),$$

$$[\sigma^2 \mid \underline{\beta}, \, \underline{u}, \, \underline{y}] \propto (\sigma^2)^{-\frac{168}{2}} exp(-\frac{1}{2\sigma^2}\underline{u}' K \underline{u})(\sigma^2)^{-(a-1)} exp(-b\sigma^{-2}) =$$

$$= (\sigma^{-2})^{(\frac{168}{2}+a-1)} exp((-\sigma^{-2})(\frac{1}{2}\underline{u}' K \underline{u} + b)).$$

Ez utóbbi a posteriori eloszlás a σ^{-2}-ra nézve éppen a $Gamma(a + \frac{168}{2}, \, b + \frac{1}{2}\underline{u}' K \underline{u})$ sűrűségfüggvénye. Az a posteriori eloszlásokból mintát véve becslést kapunk a keresett paraméterek várható értékére.

5. Mintavétel a példában

A modellben öt kategorikus változó, egy konstans, és a térbeli elhelyezkedés szerepel. A kategorikus változók hatásának becsléséhez sorban 1, 8, 3, 5, 3, 10 darab β paraméterre van szükség. További 168 β kell a térbeli hatáshoz; minden régióhoz egy-egy. Továbbá valószínűségi változó a regionális hatásban az összefüggés mértékét meghatározó σ^2 hiperparaméter is. Így összesen 199 paraméterre keresünk becslést az y_i (szerződésenkénti kárszámok), e_i (szerződésben töltött napok bonusfaktorral módosított száma), valamint a többi, a modellben használt változó (kor, nem, ...) ismerete mellett. Mivel a σ^{-2} a posteriori eloszlása beazonosítható, *Gamma*, ezért abból Gibbs lépések alkalmazásával könnyű mintát venni: minden iterációban az aktuális $\underline{u}$ mellett a $Gamma(a+\frac{168}{2},\ b+\frac{1}{2}\underline{u}'K\underline{u})$-ból kell egy véletlen értéket generálni. A többi paraméter a posteriori eloszlása azonban bonyolult, ezért a mintavételhez Metropolis-Hastings algoritmust használunk, méghozzá a gyors konvergencia érdekében 1-komponensűt. A 3.2.1. fejezetben leírt algoritmust alkalmazzuk. A második fejezetben leírtak alapján világos, hogy érdemes az összes változót egyszerre figyelembe venni a pontosabb eredmény elérése érdekében. Ugyanakkor más szempontok is fontosak egy-egy ilyen típusú feladat megoldásakor. Lényeges kérdés például a futásidő. A 2.2.2. fejezetben leírt módszer beprogramozásakort azt tapasztaltam, hogy ekkora mennyiségű adat mellett túl sok időt vesz igénybe a program futása. Ugyanakkor logikus az ilyen modellek célkitűzése, azaz hogy ne olyan legyen a modellillesztés, hogy az először kor, nem, ... változókra illesztett modellt utólag finomítjuk a regionális hatás figyelembe vételével, hanem a regionális hatás is befolyással lehessen a többi változó hatásának vizsgálatára. Ezért a következő modellt alkalmaztam az adatsorra:

1. Először illesszünk általánosított lineáris modellt (GLM) az R programcsomag segítségével:

   $$glm(y \sim offset(log(e)) + kor + nem + auto + sz + pop,\ family = Poisson),$$

   ahol az e, kor, nem, ... változók vektorok, melyek az egyes szerződések adott

változóban felvett értékeit tartalmazzák, míg az y-ban a kárszámok szerepelnek. Az Y_i tehát Poisson eloszlású, a paramétere helyett az R automatikusan annak a logaritmusát modellezi a felsorolt változók elsőfokú polinomjaként. Az $offset(log(e_i))$ azt biztosítja, hogy a modellben a $log(e_i)$ együtthatója 1 legyen. Erre azért van szükség, mert azt feltételezzük, hogy a kárszám egyenes arányban áll a biztosításban töltött napok számával. Így kapunk a β-kra egy első közelítést.

2. Ezután ezen paramétereket fixnek tekintve vizsgáljuk a regionális hatást, a 2.1.2. fejezetben leírt módon. Itt tehát már Markov Lánc Monte Carlo algoritmust kell használni. Az i. iteráció j. lépésében az u_j-re akarunk egy újabb mintát generálni, ezt jelöljük $u_j\,[i]$-vel. Ehhez kiindulásnak vegyük az $u_j[i-1]$-nek egy normális eloszlásból vett véletlen számmal való módosítását. A javaslati eloszlás tehát $z \sim N(u_j\,[i-1]\,, 0.001)$. A következő lépésben számoljuk ki az a posteriori eloszlásfüggvény értékét az $u_j = z$, illetve az $u_j = u_j\,[i-1]$ pontokban, az előzőleg kiszámolt β értékek mellett. Ha az első érték nagyobb, elfogadjuk $u_j\,[i]$-nek a z-t, ha a második érték a nagyobb, akkor a kapott két szám arányának megfelelő valószínűséggel fogadjuk el a z-t, vagy hagyjuk meg az $u_j\,[i-1]$-t az új mintának. Az $u_j\,[i]$ aktuális értékekkel a σ^2-ből kell még mintát vennünk, ezt pedig az R-ben egy egyszerű utasítással megtehetjük: $\sigma^{-2} < -rgamma(a + \frac{168}{2},\, b + \frac{1}{2}\underline{u}'K\underline{u})$. 10 000 lefutás után azt tapasztaltam, hogy az így generált Markov-lánc tagjai már a kívánt eloszlásból vannak, ezért újabb 90 000 iteráció átlagaként becslést kaptam az u_j-kre, illetve a σ^2-re.

Az eljárásunk további részében e két lépést ismételjük. Térjünk vissza tehát a β paraméterekre. Ezekre újabb becslést kapunk, ha az y_i, e_i, illetve a változók értékein kívül figyelembe vesszük a regionális hatásra számolt értékeket. Ha a β-kat továbbra is egyszerre kezeljük ahelyett, hogy minden egyes paraméterre szimulációs lépéseket futtatnánk, gyorsabban eredményt kapunk. Illesszünk tehát újra általánosított lineáris modellt az 1. pontban felsorolt változókra, azzal a változtatással, hogy az e_i-k

helyett azoknak a megfelelő regionális hatással módosított értékeit használjuk:

$$\nu_i = \beta_0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + (log(e_i) + u_{r_i}) =$$

$$\nu_i = \beta_0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i \times exp(u_{r_i})).$$

Az így módosított $e_i' = e_i \cdot exp(u_{r_i})$ értékekkel számoljuk újra az általánosított lineáris modellt.

A két lépést felváltva ismételve elérjük, hogy az egyes paraméterekre kapott becslések visszahathatnak a többi paraméter becsült értékére. Az algoritmus akkor ér véget, amikor a paraméterekre kapott becslések már nem mutatnak változást.

6. Az eredmények értékelése

6.1. A hatásokra kapott becslések

A szerződéseknek olyan tulajdonságai alapján közelítettem a kárszámot, melyek szignifikáns voltáról előzőleg lineáris illesztéssel meggyőződtem. Így kaptam, hogy a várható kárszám becslésekor figyelembe kell venni a kort, a nemet, a kocsitípust, a szerződéskötés idejét, a populációt (lélekszám), és persze a vizsgálatunk középpontjában álló régiókat.

$$\nu_i = \beta_0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop} + log(e_i) + u_{r_i}.$$

Először ismertetem azon változók hatását, melyeket e komplexebb modellben is lineáris modellillesztéssel vizsgáltuk, ezeket jelöltük β-val, majd rátérek a régiók hatására kapott becslések ismertetésére. A lineáris modellillesztésnél egy adott változó esetén az első kategória relatív kockázata mindig 0, és ehhez viszonyítva kapjuk a többi kategória kockázatát. Mivel a paraméterezésünk szerint az i. szerződő kárszámának Poisson-paramétere a $\lambda_i e_i = exp(\nu_i) = exp(\beta_0 + \beta_{k_i}^{kor} + \beta_{n_i}^{nem} + \beta_{a_i}^{auto} + \beta_{s_i}^{sz} + \beta_{p_i}^{pop}) \cdot e_i \cdot exp(u_{r_i})$ alakban írható fel, ezért érdemes a kapott β-k exponenseit ábrázolni, hiszen ezek azt mutatják, hogy ha tekintünk két szerződést, melyek egy (nem a régiót jelző) változó - legyen ez a kor - kivételével azonosak, és az első például az 1. korcsoportba tartozik, második az m.-be, akkor a második szerződő várható kárszáma $exp\left(\beta_m^{kor}\right)$-szerese az elsőének.

6.1.1. A konstans

A konstans $\beta_0 = -8.32267$, mely minden szerződésre jellemző. Ha eltekintünk a regionális hatástól, akkor ennek az exponense (≈ 0.0002429) egy olyan szerződő várható kárszáma, melynek a többi jellemzője mind az 1. kategóriába tartozik, és egy napig volt szerződésben (így a kárszám is egy napra vonatkozik). Egy évre egy ilyen tulajdonságokkal rendelkező, 1-es bonusfaktorú szerződő várható kárszáma 0.0884.

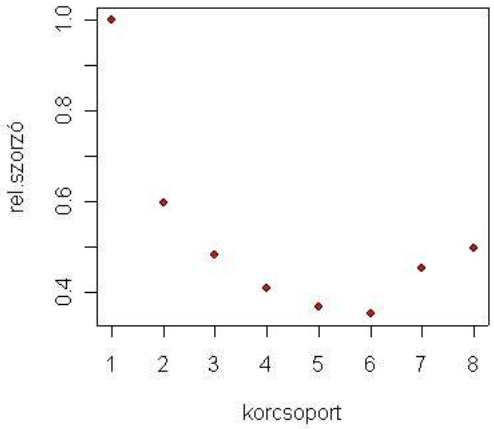
6.1.2. A kor hatása

A β^{kor} vektor elemei az lenti táblázat középső sorában szerepelnek, az alsó sorban ezek exponense látható, mely tehát azt fejezi ki, hogy egy az adott korcsoportba tartozó szerződőnek hányszoros a várható kárszáma egy más tekintetben azonos, de az első korcsoportba tartozó szerződőhöz képest.

korcsop.	1	2	3	4	5	6	7	8
kor	<20	$20\leq<25$	$25\leq<30$	$30\leq<35$	$35\leq<50$	$50\leq<70$	$70+$	cég
rel. kock.	0	-0.51	-0.73	-0.89	-0.99	-1.04	-0.79	-0.7
rel. szorzó	1	0.6	0.48	0.4	0.37	0.35	0.45	0.5

Az 1. korcsoportba a 20 évnél fiatalabb vezetők tartoznak, míg a másodikba a 20 és 24 év közöttiek. Ez utóbbi csoport relatív szorzótényezője 0.6, azaz egy ide tartozó vezető átlagosan 3/5 annyi balesetet okoz, mint egy húsz év alatti sofőr. Ez az érték azt mutatja, hogy a két csoport relatív kockázata nagyon eltérő, azonban a kapott arányt óvatosan kell fogadnunk, mert a húsz évnél fiatalabb szerződők csoportjában mindössze 22-en voltak az adatsorban. Kicsit kevésbé látványos, de egyértelmű különbség van az 6. illetve 7. korcsoportok relatív kockázata között, a 70 éven felüliek relatív szorzója közel 1.3-szorosa az 51 és 70 év közöttieknek.

Ábrázoljuk a szorzótényezőket!



2. ábra. A kor hatása

Látható, hogy a korral eleinte csökken a balesetek számának várható értéke, de a 70 éven felüli vezetőknél ez a tendencia megfordul, és nő a baleset kockázata. A csoportbontásunknak köszönhetően az is világossá válik, hogy az 51-70-es korcsoport relatív kockázata az addigi csökkenő tendenciát folytatja, és csak a 70 éven felülieknél fordul meg a folyamat iránya. A nyolcadik a céges autók csoportja. Egy ebbe a kategóriába tartozó autót kor szerint egy meglehetősen széles réteg használ, általában 25 és 70 közöttiek, így érdemes a vezetők kora szerinti kockázatokkal összevetni a céges autó kockázatát. Azt tapasztaljuk, hogy gyakorlatilag minden korosztály kezében veszélyesebbé válik az autó, ha céges, hiszen a relatív kockázat itt minden, 25 évnél idősebb, saját autót vezető magánszemélyénél nagyobb.

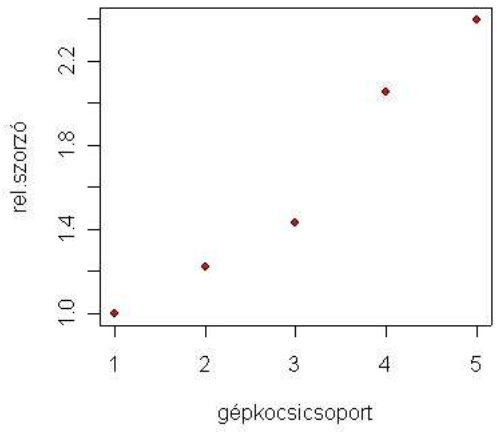
6.1.3. A gépkocsitípus (auto) hatása

Az autó típusának is könnyen magyarázható, és jelentős szerepe van a kárszámok várható értékének alakulásában. A kategóriák sorszámának növekedése a hengerűrtartalom növekedését jelzi. 850 köbcenti alattiak kerültek az első osztályba, az ennél

nagyobb, de 1150 ccm-nél kisebb űrtartalmúak, a másodikba, az 1150 és 1500 ccm közöttiek a harmadikba. A két utolsó csoport az 1501 és 2000, illetve a 2001 és 3000 köbcenti közötti hengerűrtartalmú járműveket tartalmazza.

autótípus	1	2	3	4	5
rel. kockázat	0	0.2	0.36	0.72	0.87
rel. szorzó	1	1.22	1.43	2.05	2.39

A táblázatból jól látszik, hogy az hengerűrtartalom növekedésével nő a kockázat, tehát minél nagyobb, erősebb egy autó, annál nagyobb a várható kárszám.



3. ábra. Az autótípus hatása

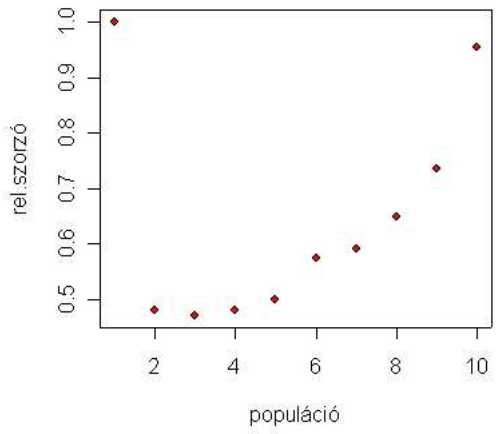
6.1.4. A lélekszám (pop) hatása

A település lélekszámát tekintve tíz csoportot határoztunk meg. Az 1. kategória Budapestet jelöli, hiszen több, mint két millió lakosával a főváros Magyarország legsűrűbben lakott városa. Mivel a második legnagyobb város Debrecen a maga háromszáz ezer lakosával messze elmarad mögötte, így jogos Budapestet önálló kategóriának tekinteni. A második kategóriába tartoznak a legkisebb, 500-nál kevesebb

lakosú falvak, a harmadikba az 501 és 1000 közöttiek, etc. A tizedik csoport a legnagyobb, százezernél több lakóval rendelkező városoké. Az ezekre kapott becslések a következők:

pop	1	2	3	4	5	6	7	8	9	10
rel. kock.	0	-0.73	-0.75	-0.73	-0.7	-0.55	-0.52	-0.43	-0.3	-0.04
rel. szorzó	1	0.48	0.47	0.48	0.49	0.57	0.59	0.65	0.74	0.95

Mint ahogy azt talán vártuk is, a legnagyobb relatív kockázatot Budapest képviseli. A második, harmadik, és negyedik csoportok között nincsen nagy különbség, ezeknek a kétezernél kevesebb lakossal rendelkező településeknek a legkisebb a kockázata. Az ötödik, és hatodik csoport a 2001 és 5000, illetve az 5001 és 10000 közötti lakossal rendelkező városok kockázata hasonló, az előzőeknél nagyobb. A hatodik kategóriától kezdve a lakosság számával együtt monoton nő a kockázat is. Olyannyira, hogy a legnagyobb városok majdnem utólérik a Budapestet a veszélyesség tekintetében.



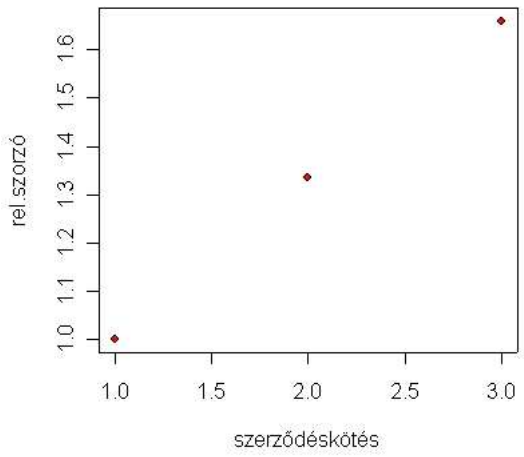
4. ábra. A lélekszám hatása

Ez a tendencia józan ésszel is logikusnak tűnik, hiszen minél több ember él egy városban, annál több a gépkocsi, esetleg nagyobb a népsűrűség is. Ez pedig az egyre több baleset irányában hat.

6.1.5. A nem és a szerződés korának (sz) hatása

Vizsgáltuk a nem hatását is, ahol azt tapasztaltuk, hogy a nők relatív kockázata 0.06, azaz egy női vezető várható kárszáma 1.06-szor nagyobb, mint egy a modell szempontjából ugyanolyan tulajdonságokkal rendelkező férfié. Nem szerint a harmadik kategória a céges autóké volt, de annak a hatását, hogy egy autó magánszemélyhez tartozik-e, vagy céghez már becsültük a korcsoportok vizsgálatánál, hiszen ezek az autók ott is egy külön kategóriát alkottak, így erre most nem is kaphattunk újabb becslést.

A szerződés korának hatása a következőképp alakult: $\beta^{sz} = (0, 0.29, 0.5)$, azaz a szorzótényezők sorban (1, 1.28, 1.65).

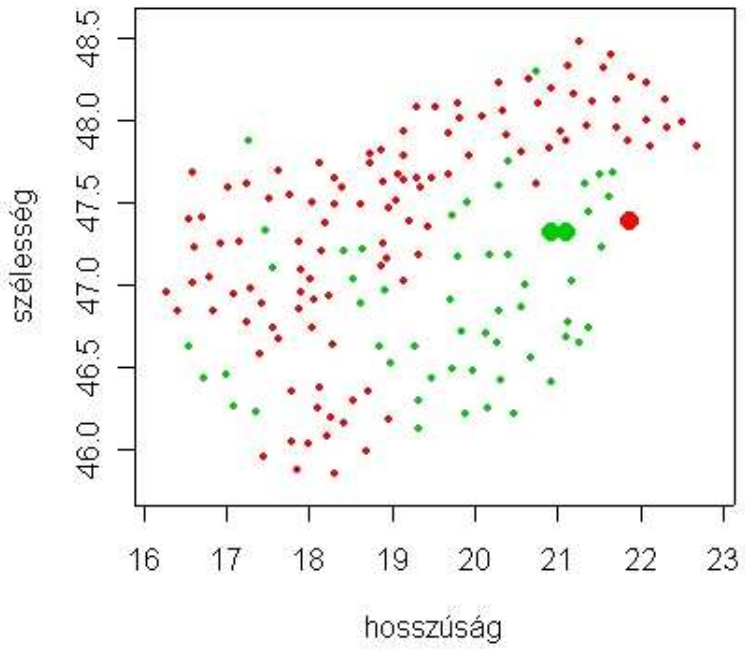


5. ábra. A szerződéskötés idejének hatása

Legkisebb azoknak a kockázata, akik még 1998-ban, vagy azelőtt kötötték a szerződésüket. 1.3-szoros a kockázata az 1999-ben, és több, mint másfélszeres, a 2000-ben szerződötteknek. Az eredmények tehát azt mutatják, hogy a szerződések első pár évében csökken a várható kárszám.

6.1.6. A regionális hatás

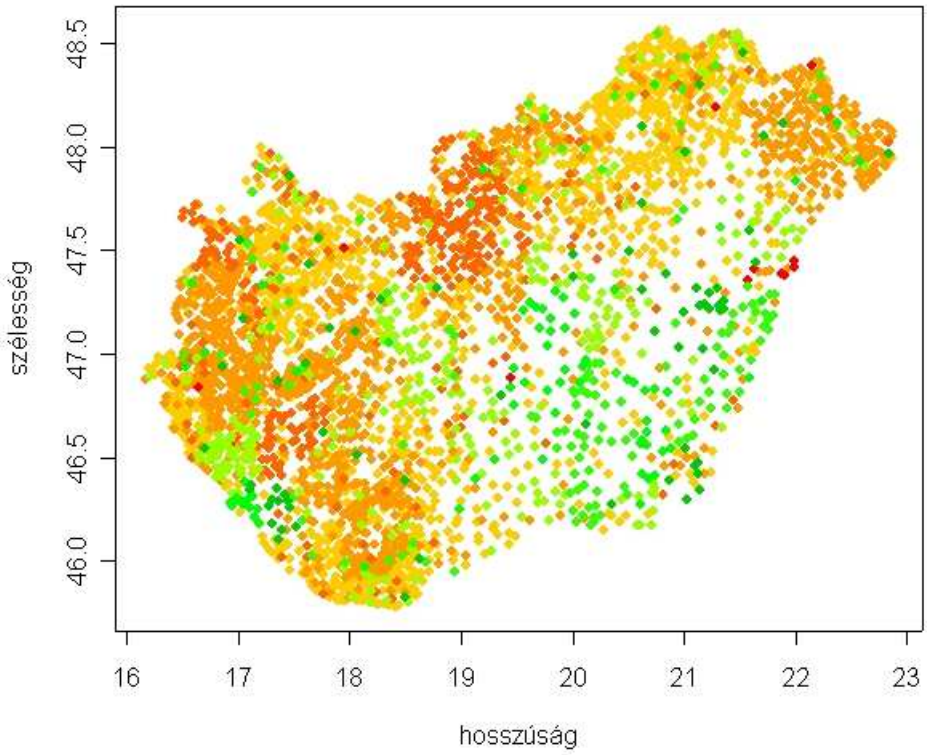
A fenti eredményeket nyolc iteráció után kaptuk. Egy iterációban a második lépés a regionális hatás becslése. Általánosságban elmondhatjuk, hogy ezek az értékek kisebb abszolút értékűek lettek, mint az előzőek, tehát ez a fajta hatás nem olyan jelentős, mint például az, hogy a vezető milyen korú. A legnagyobb 0.399 már kiugróan magas értéknek számít, többnyire 0.1 körüli, vagy annál kisebb abszolút értékű értékeket kaptunk. Bár e változó esetében nem jelöltünk ki egy települést sem 0 kockázatúnak (azaz viszonyítási alapnak), mint ahogy az az előbbiekben automatikusan adódott, a kapott u_i kockázatok átlaga közel 0; 0.003, így mégiscsak van értelme a 0-hoz hasonlítani az u_i-ket. A következő térképen láthatjuk, hogy milyen eredményt kaptunk az egyes régiókra. Piros szín jelöli a pozitív kockázatú régiókat, ezeken belül az egyetlen kiugró érték a keleti határon lévő Derecskéé, zöld színűek a negatív kockázatú, azaz biztonságosabb régiók, ezeken belül legkisebb Karcag, illetve Püspökladány kockázata ≈ -0.27.



6. ábra. A regionális hatás

Nem meglepő, hogy a nagyvárosokra általában nem kaptunk kiugróan nagy pozitív értékeket, hiszen a lélekszám vizsgálatával a régiók hatásának a lakosság számával magyarázható részét már leválasztottuk az adatokról. Marad tehát pusztán a térbeli elhelyezkedés hatása. Az ábrán jól látszik, hogy észak-kelet, dél-nyugat irányban meg tudunk húzni egy vonalat, mely kettévágja az országot, és kevés kivételtől eltekintve szétválasztja a pozitív, illetve a negatív kockázatú helyeket. Az északi, nyugati részen a pozitív kockázat a jellemző, míg a dél-keleti, főként alföldi városok többnyire negatív u_i-vel rendelkeznek. Ennek egyik oka talán a földrajzi viszonyokban kereshető, hiszen az északi táj jellemzően hegyvidékes, ami veszélyesebb, míg az ország dél-keleti része sík. Az alföldi városokban ráadásul elterjedt a kerékpár használata a városon belüli közlekedésre, azaz arányaiban kevesebb gépjármű van a

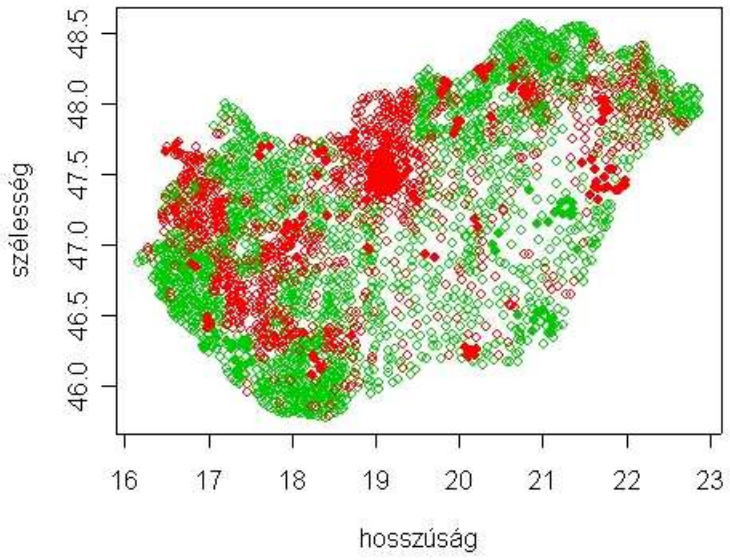
városok útjain, ami természetesen csökkenti az autóbalesetek valószínűségét. Ábrázoljuk a regionális hatásra kapott értékeket településenként a szemléletesség kedvéért.



7. ábra. A regionális hatás településenként

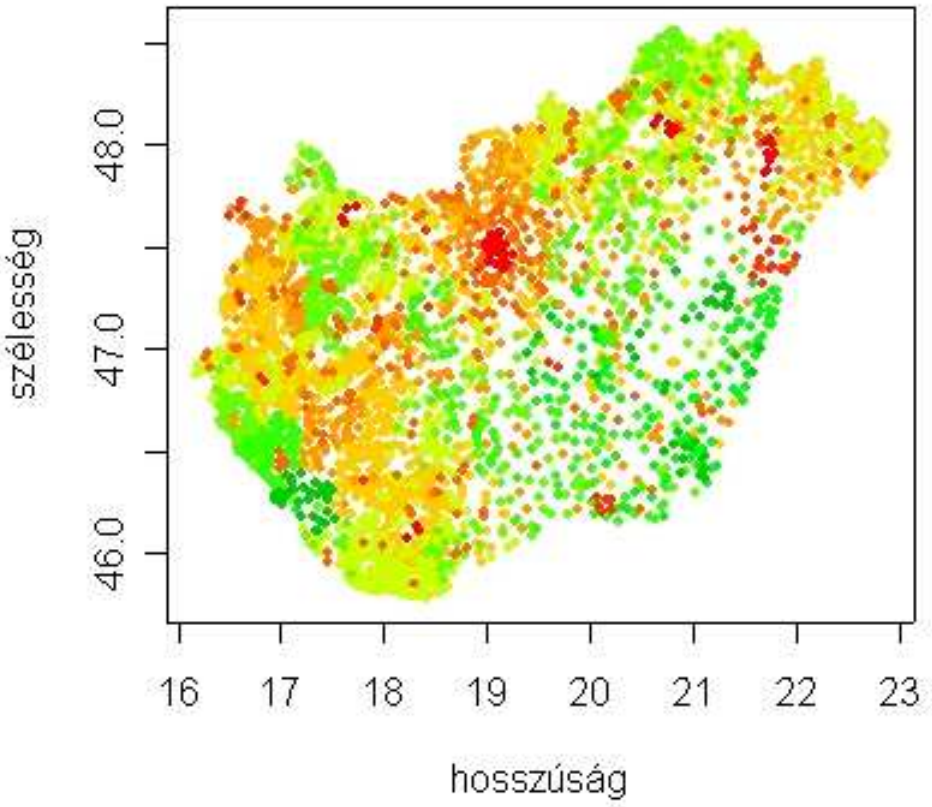
Az ábrán továbbra is a zöld szín árnyalatai jelzik a negatív, piros a pozitív kockázatú régiók településeit, drapp színűek a 0 közeli, átlagos relatív kockázatú települések. A sötét színek a szélsőséges értékeket mutatják. Az alföldi térség itt is jól láthatóan elkülönül az ország többi részétől, azonban az árnyalatoknak köszönhetően az is megfigyelhető, hogy nem minden hegyvidéki térség kockázatosabb az átlagosnál, például észak-kelet Magyarország hegyeiben átlagosnak mondható a területi hatás.

A fenti két térképen az u_i értékeket ábrázoltam, ugyanakkor a térbeli hatás vizsgálatánál érdemes megnézni, hogy mit kapunk, ha a földrajzi elhelyezkedésre kapott becsléshez hozzávesszük a lélekszám hatását is. Ha ugyanis egy város vagy falu relatív kockázatáról beszélünk, általában nem választjuk szét, hogy mekkora az a hatás, amit a lélekszámnak, és mekkora, amit a földrajzi elhelyezkedésnek köszönhetünk, hanem annak minden jellemzőjével együtt vagyunk kíváncsiak az adott helység veszélyességének mértékére. Ezért településenként kiszámoltam a két jellemző együttes hatását, és az így kapott relatív kockázatokat ábrázoltam:



8. ábra. A regiók és a lélekszám hatása

A piros itt is az átlagoshoz képesti pozitív, a zöld a negatív eltérést jelöli. Az előzőleg megfigyelt elkülönülése a térségeknek most már nem tapasztalható, hiszen a két hatás közül a populáció a domináns. Több színárnyalatot használva pontosabb képet kapunk:



9. ábra. A régiók és a lélekszám hatása 2.

Ahogy várható, több nagyváros, és környéke is kirajzolódik, piros színnel. Legjelentősebb természetesen Budapest, de felismerhető Miskolc, Nyíregyháza, Győr, Debrecen is. A kevésbé lakott területek általában (sárgás)zöldek. Ha nem is olyan egyértelműen, de azért itt is megfigyelhető, hogy a dél-keleti országrészre kisebb értékeket kaptunk, a zöld szín dominál.

A kapott értékeket térképen ábrázolva értelmes eredményt kaptunk, ám ahhoz, hogy a módszer helyességét ellenőrizni tudjuk, vizsgálni kell a kárszámra kapott

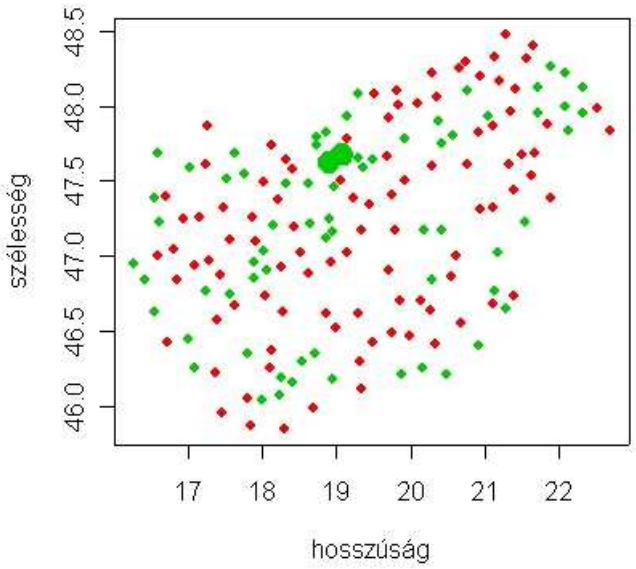
becslésnek a valódi értékektől vett eltérését.

6.2. A modellek jóságának mérése

A modell illesztésénél az elsődleges cél az volt, hogy minél jobb becslést adjunk az egyes szerződések kárszámára. Ugyanakkor a Markov mezőt abból a meggyőződésből választottuk, hogy a szomszédos területek bizonyos hasonlóságot mutatnak. Így azonban az egyes térségek paraméterének becslésekor figyelembe vettük a szomszédos régiókat is, ezért tökéletes illeszkedést nem várhatunk. Nehéz tehát eldönteni, hogy hogyan lehetne mérni egy modell jóságát. Nincs erre egységes módszer a szakirodalomban sem. Egy lehetséges ellenőrzési módszer, ha minden régióra kiszámoljuk a becslés lenormált hibáját:

$$h_i = \frac{\widehat{y_i} - y_i}{\sqrt{\widehat{y_i}}}$$

Ahol az $\hat{y}_i$ az i. szerződés várható kárszámára, azaz az i. Poisson-paraméterre kapott becslés; $\lambda_i e_i = exp(\nu_i)$. Számoljuk ki ezen értékeket, és ábrázoljuk térképen. Így láthatjuk, hogy mely régiók esetén kaptunk alul- illetve felülbecslést.



10. ábra. A regionális hatás becslési hibája

Piros pontok állnak azon régióközéppontok helyén, ahol a becsült érték nagyobb, mint a bekövetkezett kárszám. A pontok mérete a hiba nagyságát mutatja, látható, hogy nagy mértékű felülbecslés nem történt. A zöld az alulbecslés színe, ez esetben Budapesten és Dunakeszin a többi helyhez képest modellünk jelentősen alulbecsülte a várható kárszámot. Az ok a Markov mező modellben keresendő, hiszen ennek lényege éppen az, hogy egy város kockázatát a szomszédosak átlagától nem engedi nagyon eltérni. Így érthető, ha egy nagy várost csupa kisebb kockázattal rendelkező település vesz körül, akkor az csökkentőleg hat a város modell által számolt relatív kockázatára.

Ha egyetlen számban akarjuk megadni az illszkedés mértékét, akkor vehetjük e számok négyzetösszegét:

$$H = \sum_{i=1}^{m} h_i^2,$$

minél kisebb a H, annál pontosabb a közelítésünk. Ha azonban több régióban is kevés volt a károk száma (esetleg 0), érdemes definiálni egy $H*$-t is, melyben ezeket a h_i^2-eket kihagyjuk az összesítésből, hiszen bár a Markov mező alkalmazásának egy pozitív következménye, hogy olyan régiók esetén is értelmes becslést kaphatunk, ahol az adott évben nem történt kár, ezen régiókra nem várható el, hogy az illeszkedés (pl. 0 valódi kárszám, nem 0 becsült érték) jó legyen. Legyen tehát a $H*$ azon régiók hibáinak négyzetösszege, ahol a becsült kárszám legalább 10 volt:

$$H* = \sum_{i=1,\, \hat{y}_i > 10}^{m} h_i^2$$

A H, illetve a $H*$ értékek segítségével tehát össze tudjuk hasonlítani a különböző módszerekkel kapott becsléseket. A következő két táblázatban ezt láthatjuk. Az első táblában minden térség hibáját beleszámítottam a H-ba, a másodikban a $H*$ értékek szerepelnek, melyek számolásakor csak azokat a régiókat vettem figyelembe, melyekben a becsült kárszám legalább 10 volt. A táblázatok első sora azt mutatja, hogy hány iterációt alkalmaztam a modellben. Az első oszlopban a regionális hatás figyelmen kívül hagyásával számolt általánosított lineáris becslés hibája található, a másodikban azon modell hibája van, melyben csak egy utólagos illesztést végeztünk a regionális hatás vizsgálatára, míg a harmadik oszlopban a 8 iterációval elért eredmény látható.

Iterációk száma	0	1	8
H	354.48	272.08	231.01

Láthatjuk, hogy egyrészt a regionális hatást mindenképpen érdemes figyelembe venni, hiszen a négyzetes hibát nagymértékben javítani tudtuk ezzel. Ugyanakkor az is kiolvasható a táblázatból, hogy ha egy utólagos illesztés helyett több lépésben végezzük el a becslést, további javulást érhetünk el az illeszkedés tekintetében.

Iterációk száma	0	1	8
H*	324.09	242.85	207.89

A $H*$ értékek is megerősítik abbéli meggyőződésünket, hogy érdemes több lépésben megismételni az általánosított lineáris modellillesztésből, majd a regionális hatás becsléséből álló módszert.

7. Konklúzió

Dolgozatomban egy gépkocsibiztosításból származó adatsor elemzése kapcsán különböző módszereket mutattam be. Az elsődleges cél a régiók relatív kockázatának becslése volt, de a számolások során a szerződőnek több más, a kárszámot befolyásoló tulajdonságát is figyelembe vettem. A szakirodalomban legelterjedtebb modellek bemutatása során ismertettem olyan módszereket is, melyek a nem-regionális változók hatásának becslését finomítják. Mivel azonban a vizsgálat középpontjában a földrajzi elhelyezkedésnek a kárszámra gyakorolt hatása állt, ezért a többi változó hatására a legegyszerűbb modell szerint adtam becslést. A térbeli hatás utólagos becslésével kapott modell illesztése után egy új módszert vizsgáltam, melyben több iterációban felváltva alkalmaztam a következő két lépést; először általánosított lineáris modell segítségével becslést adtam a nem-regionális változók hatására, majd Markov Lánc Monte Carlo módszerrel becsültem a regionális hatást. Az illeszkedések mértékének összehasonlítását a négyzetes hibák segítségével végeztem el. E hiba mellett ténylegesen jobb illeszkedést kaptam a több iterációt alkalmazó módszerrel, mint az egyszerivel. Bár valószínű, hogy a teljesen Bayes-i megközelítéssel illesztett (azaz minden változót egyszerre figyelembe vevő) modellek pontosabb eredményt adnak, ezzel az eljárással gyorsabban kaphatunk eredményt, hiszen leválasztva azokat a változókat, melyekről jogosan tételezünk fel lineáris, vagy faktorhatást, azokra egyszerű általánosított lineáris becslést alkalmazhatunk. A továbbiakban érdemes lenne a kárnagyságra is vizsgálni a módszer hatékonyságát, esetleg más hibadefiníció mellett. Mivel az itt alkalmazott módszer hatékonyságának nincs elméleti bizonyítéka, ezért további vizsgálatok szükségesek azt megállapítandó, hogy milyen esetekben várhatunk el jó közelítést, illetve pontos előrejelzést ettől a módszertől.

Hivatkozások

[1] Arató, N. M., I. L. Dryden, C.C. Taylor (2004). Hierarchical Bayesian modelling of spatial age-dependent mortality.

[2] Denuit, M., S. Lang (2004). Non-life rate-making with Bayesian GAMs.

[3] Dimakos, X. K., A. Frigessi di Rattalma (2002). Bayesian premium rating with latent structure. Scand. Actuarial J., pp. 162-184

[4] Gilks, W. R., S. Richardson & D. J. Spiegelhalter, Markov Chain Monte Carlo in practice, Chapman and Hall , London

[5] Mollié, A. (1996). Bayesian mapping of disease, in W. R. Gilks, S. Richardson & D. J. Spiegelhalter, Markov Chain Monte Carlo in practice, Chapman and Hall, London, pp. 359-376.

[6] Robert, C. P., G. Casella (1999), Monte Carlo Statistical Methods, Springer-Verlag, London