Területi faktorok modellezése a nem-élet biztosítás díjkalkulációjában

Diplomamunka

Írta: Szabó Róbert

Alkalmazott matematikus szak

Témavezetők:

Pásztor Gábor, vezető aktuárius
Allianz Hungária Biztosító Zrt.
és
Pröhle Tamás, egyetemi tanársegéd
Valószínűségelméleti és Statisztika Tanszék
Eötvös Loránd Tudományegyetem, Természettudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar
2009

Tartalomjegyzék

1. Bevezetés

Egy biztosító számára nagyon fontos, hogy minél pontosabban meg tudja határozni jövőbeli kifizetéseinek mértékét. Másképpen fogalmazva, azt szeretné tudni, hogy valójában az egyes szerződések során mekkora kockázatot vállal át a biztosítottaktól. A tényleges kockázat felmérése statisztikai módszerekkel végezhető el. Elsődlegesen az a feladatunk, hogy meghatározzuk, melyek azok a tényezők, amelyek befolyásolják az egyes szerződések kockázatát. Ezen tényezőket a továbbiakban faktoroknak nevezzük. A biztosítás jellegétől függően más és más faktorok jöhetnek szóba. Például a gépjármű-felelősségbiztosításban (GFB) jellemző faktorok: az életkor, a gépjármű tulajdonságai, lakhely, stb.... Ezen befolyásoló tényezők meghatározása főképpen intuitív meggondolásokon alapszik. Például nem meglepő az a feltevés, hogy egy 20 éves fiatal nagyobb kockázattal bír személygépkocsivezetőként, mint egy 40 éves férfi. Ezen feltételezéseink helytállását statisztikai módszerekkel támaszthatjuk alá vagy esetleg vethetjük el.

Következő feladatunk annak eldöntése, hogy a különböző faktorok kategóriái mekkora súllyal befolyásolják a kockázatot. Például mekkora súllyal járul hozzá a kockázathoz az, ha a gépjárművezető 30 éves? Ennek eldöntésére is fellelhetők statisztikai módszerek. Az egyik legkedveltebb ezek közül az *általánosított lineáris modell (GLM)*, amely minden faktor minden kategóriájához hozzárendel egy súlyt. Ezek után járható út az, ha egy faktor azon osztályait összevonjuk, amelyek súlya nem tér el jelentősen egymástól. Ezzel a megközelítéssel, a különböző faktorok díjosztályokat alakítanak ki. Egy lehetséges díjosztály például azon gépjárművezetők, akik 20 és 25 év közöttiek, piros színű gépjárművel rendelkeznek és Budapesten laknak.

Az jól érzékelhető, hogy minél részletesebben skálázzuk a faktorokat annál nagyobb lesz a díjosztályok száma. Azonban minél jobban szétdaraboljuk a portfoliót, annál kevesebb statisztikai adat marad az adott osztály kiértékeléséhez, így annál bizonytalanabb lesz a kapott eredmény is. Ezzel ellentétben, ha csupán néhány osztályra daraboljuk szét az állományt abból a megfontolásból, hogy elegendően sok adatunk legyen a statisztikai kiértékeléshez, akkor könnyen előfordulhat, hogy nagyon különböző kockázatokat sorolunk azonos osztályba. Ekkor a konkurencia esetlegesen részletesebb szegmentálással kicsemegézheti az adott osztályunkból a kedvezőbb kockázatú szerződéseket, és a nálunk maradt portfolió rosszabb lesz, mint azt eredetileg kalkuláltuk. Az osztályok számának, részletezettségének meghatározásakor e kettősség között kell megtalálni a megfelelő kompromisszumot.

Jelenleg a legtöbb magyarországi biztosító által a gépjármű-biztosításban használt faktorok: az életkor, a lakóhely és a gépjármű hengerűrtartalma. Ezen kívül speci-

fikusan egyes biztosítók figyelembe veszik még a gépjármű életkorát, a vezetői engedély korát és a gépjármű gyártmányát (pl: BMW, Opel, stb...). A biztosítók díjosztályainak lakóhely szerinti kialakítása részletesség szempontjából nagyon eltérő. Akad olyan biztosító, amely a lakóhelyeket csupán 4 osztályra bontja: Budapest, Budapest vonzáskörzete, megyeszékhelyek és egyéb. Valamint olyan biztosító is van, amely majdnem teljes részletességgel, minden településhez meghatározza a neki megfelelő kockázatot. Azonban egyik esetben sem sorolják a lakóhelyeket 10-15-nél több osztályba.

Napjainkban a biztosítási piacon kialakult versenyhelyzet megköveteli a biztosítóktól, hogy a díjaikat differenciáltan, minél inkább a kockázatnak megfelelő mértékben alakítsák ki. Ugyanis ezzel tudják növelni ügyfeleiknek a számát és pontosabban meg tudják határozni jövőbeli kifizetéseik várható értékét. Figyelembe kell venni azt is, hogy ez az ügyfelek számára is igen előnyös, mert így az egyes ügyfelek mindig a leginkább rájuk jellemző kockázatot fizetik meg. Természetesen ez alatt azt értjük, hogy a kevésbé kockázatos ügyfeleket szétválasztjuk a kockázatos ügyfelektől, így az első esetben a díjakat csökkentjük, a második esetben pedig növeljük.

Sajnos nem minden faktor olyan könnyen kezelhető, mint például az életkor. A lakóhely, azaz a területi faktor esetén több probléma is felmerül. Elsőként az, hogy több ezer település található Magyarországon, így a településenkénti bontása a területi faktornak nem kivitelezhető. Ugyanis sok településre nagyon kevés adat áll rendelkezésére a biztosítónak, sőt az is előfordulhat, hogy van olyan település amiről semmilyen információja nincs. Ezért az általánosított lineáris modell által becsült súlyok nagyon nagy ingadozást mutathatnak, eltérhetnek a valóságtól. Nyilvánvalóan egy olyan településhez, amelyről nincsen semmilyen káradatunk, 0 súlyt rendelne, ami egyáltalán nem elfogadható.

Ezen szakdolgozat témája a területi faktorok statisztikai modellezése. A cél olyan statisztikai eljárások ismertetése, amelyek megbízható becslésekkel szolgálnak az egyes területek kockázatáról. Megadják a településeknek a kockázat mértéke szerinti osztályozását, ahol az osztályok száma lehetőség szerint nem túl nagy, és az osztályok káradatai stabilak, azaz évről évre nem mutatnak jelentős ingadozást.

A szakdolgozat két eltérő szemléletű modellezést mutat be három modellen keresztül, de mindkettő esetében élünk azzal a feltételezéssel, hogy a területi faktoron kívüli faktorokra már létezik valahonnan egy megfelelő becslésünk (pl:GLM). Elsődlegesen kiszűrjük a nem területi faktorokat az adatokból. Ezzel megkapjuk minden területre illetve lakóhelyre azt, hogy mekkora a területi faktor által indukált kockázata, mely keverve tartalmazza a területi faktor hatását a véletlen miatti ingadozással. Ez azonban még nem elegendő, hiszen mint említettük előfordulhat

olyan település is, amelyről semmilyen információ nem áll a biztosító rendelkezésére. Képzeljük el magunk elé Magyarország térképét a síkon, minden egyes településhez rendeljük hozzá a neki megfelelő területi faktor mérőszámát. Ekkor egy tűpárnához hasonló képet kapunk. Szemléletesen a feladatunk az, hogy erre egy olyan felületet illesszünk, amely minden település felett a kockázat várható értékét mutatja. Ezen felülettől még azt is megköveteljük, hogy elég sima legyen, mert így módunkban áll kevesebb számú díjosztály kialakítása. A felület simaságát arra a feltételezésre alapozzuk, hogy a területileg közel elhelyezkedő települések kockázata is közel van egymáshoz.

Az első modell a Whittaker féle területi simítást használja. A modell a [3] *Greg Taylor - Geographic Premium Rating By Whittaker Spatial Smoothing* (2001) cikk alapján lett feldolgozva. Nyilvánvalóan az adatok simítása során mindig hibát követhetünk el, hiszen kénytelenek vagyunk a lokálisan nagyon kiugró értékeket levágni, az alacsony értékeket pedig megemelni. Ezen hibák összértéke nő, ha az elvárt felület simaságát növeljük. A probléma megoldására szolgál a Whittaker féle simítás, amely megfelelő egyensúlyt alakít ki a simítás során fellépő hiba és a felületünk simasága között.

A második modell a *credibility elméletet* használja, amely a Bayesi statisztika egy speciális eseteként adódik. A credibility elmélet lehetővé teszi számunkra, hogy egy település kockázatának felmérése során használni tudjuk a szomszédos települések káradatait is. Ez nagyon fontos, mert a területi adataink száma nagyon kevés és sokszor nagy bizonytalansággal bír. Ez az elmélet a biztosítási matematika nagyon sok területén jól alkalmazható, részletes leírása megtalálható [4] *Arató Miklós - Nem-Élet Biztosítási Matematika* (2001) című egyetemi tankönyvben.

A harmadik modell a Bayes elméleten alapszik. A modell [5] *M. Boskov és R.J. Verrall - Premium Rating by Geographic Area Using Spatial Models* (1994) cikk alapján került feldolgozásra. A Bayesi statisztika lehetővé teszi számunkra, hogy a területi ismereteinket illetve feltételezéseinket beépíthessük a modellbe az ún. *a priori* eloszláson keresztül. Majd az *a poszteriori* eloszlás meghatározásával és maximalizálásával megkaphatjuk a paramétereink becslését. Ez valamilyen értelemben tekinthető a második modell általánosításának is.

2. Díjosztályok és díjkalkulációs módszereik

A biztosítók átvállalják a biztosítottak kockázatát, illetve kockázatuknak egy részét, biztosítási díj ellenében. A nem-élet biztosítások egyik legnagyobb részét a gépjármű-biztosítás teszi ki. A gépjármű-biztosítások két nagy csoportba sorolhatóak, amelyeket a biztosítási kockázat jellege szerint különböztetjük meg. Az egyik a casco biztosítás, amikor a kockázat az, hogy a gépjármű tulajdonosa saját hibájából kárt okoz a gépjárművében. Casco biztosítás esetén az alapbiztosítás a töréskár, de kiegészíthető még elemi károkra, lopás kárra és rongálásra is. A másik nagy csoport a kötelező gépjármű-felelősségbiztosítás, ebben az esetben a kockázat a biztosított gépjárművezető másnak okozott kárát jelenti. A biztosításnak ez a fajtája törvényileg előírt, minden, a forgalomban résztvevő gépjármű esetén. A továbbiakban ebben a dolgozatban biztosítás alatt mindig gépjármű-biztosításra gondolunk, a feltételezéseket és eredményeket ezen biztosítási területre kiterjedően fogalmazzuk meg. Azonban az ismertetésre kerülő módszerek más biztosítási területeken is alkalmazhatóak.

2.1. Biztosítói díjosztályok és faktorok

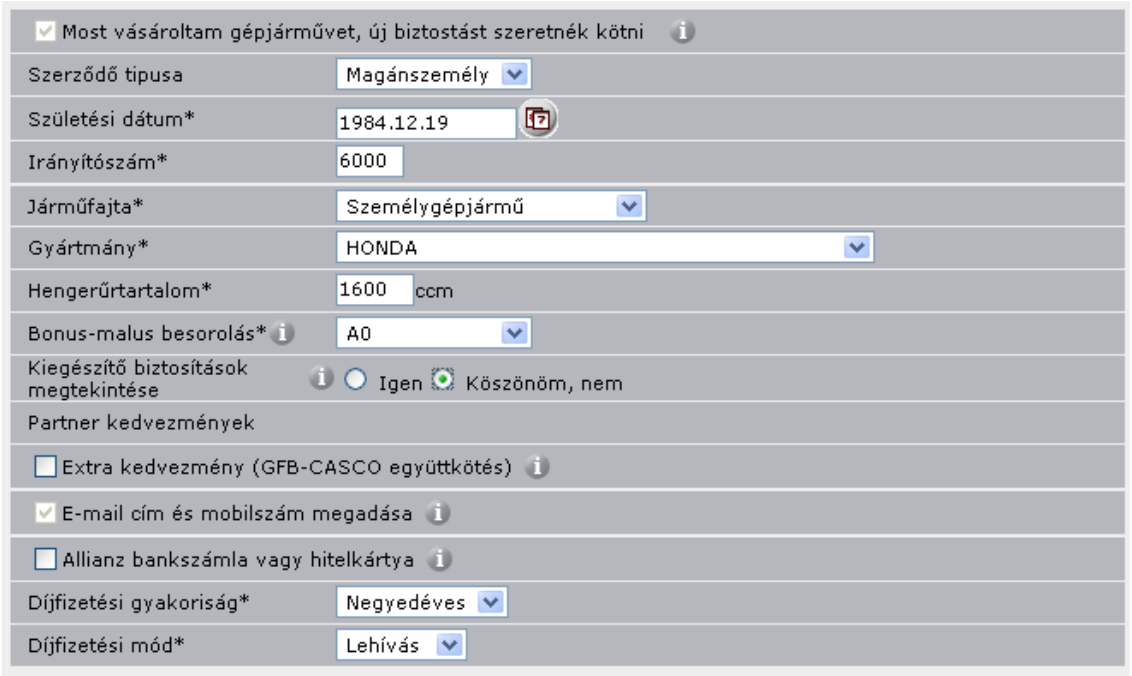
A biztosító célja, hogy minden biztosított számára az átvállalt kockázatnak megfelelő díjat szabja ki. Itt természetesen a tiszta díjról beszélünk, amely nem tartalmazza a biztosító járulékos költségeit, a várt profitját, ugyanakkor a biztosított esetleges kedvezményeit sem. A biztosítási matematikában a kockázatot egy nemnegatív valószínűségi változóval írjuk le. A kockázat két legfontosabb mérőszáma a kárgyakoriság és az átlagkár. A dolgozatban leginkább a kárgyakoriság becslésével foglalkozunk, de az eljárások általánossága, sokszor az átlagkárra is kiterjed, vagy könnyedén átfogalmazható. Tiszta díj alatt csupán a kockázat pénzbeli ellenértékét értjük. A díj számítása során leggyakrabban alkalmazott díjelv a *várható érték elv*, ebben az esetben a díjat $\lambda \geq 0$ paraméter mellett az

$$(1 + \lambda)E(X) \tag{2.1}$$

formulával számoljuk, ahol $X \geq 0$ jelöli a kockázatot leíró valószínűségi változó. $\lambda = 0$ esetén *nettó várható érték elvről* beszélünk.

A díjkalkulációban az egyik legfontosabb momentum a biztosítottak kockázatának, azaz X-nek a lehető legpontosabb meghatározása, leírása. A kockázatot a biztosítók a saját kártapasztalatuk alapján mérik fel, a kockázatra hatással lévő faktorok beazonosításával és a kockázatra ható mértéküknek becslésével. Annak ellenére, hogy például az életkor egy folytonos változó, ezeket a változókat kategorikus

változóként kezeljük, ahol a kategóriákon belüli értékek esetén nincs szignifikáns eltérés a kockázatra mért hatásban. Egy dimenziós faktorok kategorizálására természetesen léteznek statisztikai eljárások (pl.: polinom illesztés, GLM segítségével). Így a faktorok díjosztályokat alakítanak ki a különböző kategóriáik szerint. A kategóriák helyességét a biztosítási piac is megköveteli. Ugyanis, ha egy biztosító nem differenciálja megfelelően valamely változó egy kategóriáját, akkor lehetőséget ad a konkurens biztosítóknak arra, hogy ezt a kategóriát pontosabban differenciálva, a kisebb kockázatú ügyfeleket átcsábítsák. Így ugyanazon díj mellett csak a kockázatosabb ügyfeleket tartja meg, akiknek az együttes kockázata már nagyobb, mint amivel a biztosító a díjat számolta.



1. ábra. Online kalkulátor

A díjosztályok díjai minden biztosító esetén publikusak, azonban azt nem tudjuk, hogy egy biztosító menyire tekinti kockázatosnak például a 20 éves gépjárművezetőket. Hiszen egyes díjosztályok díjkalkulációjában több faktor hatása is figyelembe van véve, így nem könnyű feladat elkülöníteni a kor faktor hatását a többi faktortól. A feladatot az is nehezíti, hogy a biztosító egy díjosztály díját önkényesen is megemelheti, például a profit érdekében. Ez többnyire, akkor fordulhat elő, ha a konkurens biztosítók ugyanezt a díjosztályt magasabb kockázatúnak tartják, és így magasabb díjat számítanak rá. A biztosítók honlapjain többnyire fellelhetők díjkalkulátorok, az Allianz biztosító honlapján található kalkulátort láthatjuk az 1.ábrán.

A biztosítók jogszabály alapján minden év október 30-án kötelesek 2 országos napilapban közzétenni a GFB tarifáikat. 2008-ban ez a *Népszava* és a *Magyar Hírlap* volt. Egy ilyen díjtáblának egy részletét mutatja az *1.táblázat*, amely a 2009-es évi díjakat tartalmazza, az OTP Garancia Biztosító Zrt. vonatkozásában. A táblázatból leolvasható, hogy az OTP biztosító, a hengerűrtartalom, az életkor és a lakóhely hatásának tekintetében számolja ki a díjat. Továbbá az is látszik, hogy az életkor 4 kategóriára, a hengerűrtartalom pedig 5 kategóriára bomlik fel. A táblázat nem mutatja, de a lakóhelyeket 4 területi csoportba sorolja a biztosító, irányítószám alapján. Így 80 díjosztály keletkezik. A díjosztály díját, a végső díj számításához meg kell még szorozni a bonus-malus besorolásnak megfelelő díjszorzóval, illetve egyéb korrekciós szorzókkal (pl.: kedvezmények).

1.területi csoport					
Kor/cm^3	**850-ig**	**851-1150**	**1151-1500**	**1501-2000**	**2001-től**
<25	96960	129600	160560	270000	340080
25-29	65760	87840	104400	162480	237120
30-34	48960	68880	80880	123120	175200
35<	43680	58080	72240	108960	130320

1. táblázat. OTP Biztosító díjtáblája (részlet)

A díjkalkuláció kapcsán több kérdés is felmerülhet.

1. Milyen részletezettséget használjunk egy bizonyos faktor esetén?
   Ha kicsi a részletezettség mértéke, akkor a valódi kategóriák elmosódnak, ha nagy akkor pedig nagyon megnő a díjosztályaink száma.

2. Ha megvannak a kategóriák, akkor hogyan lehet hozzájuk, a valóságnak leginkább megfelelő súlyokat meghatározni?

3. Ha ismerjük egy biztosító díjtábláját (pl.:1. táblázat), akkor abból, hogyan lehet meghatározni a kategóriáknak megfelelő szorzókat?
   Ez a kérdés leginkább az újonnan megalakuló és kicsi biztosítókat érdekelheti a legjobban, hiszen semmilyen vagy kicsi kártapasztalattal rendelkeznek, ezért célszerű lenne számukra, egy nagy biztosító kártapasztalatán alapuló eredmény használata.

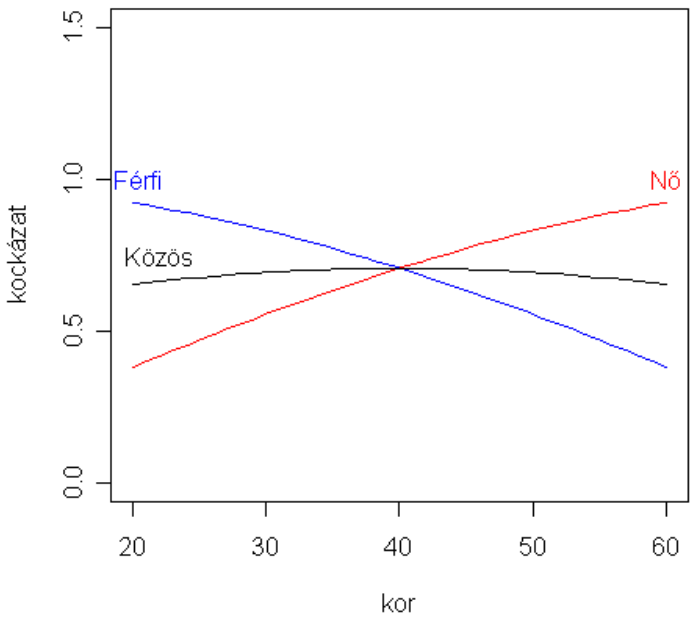
A faktorok kategóriáihoz rendelt súly becslésére több eljárást is kidolgoztak, amelyeket a következő fejezetekben röviden ismertetünk. Ezek közül a leggyakrabban használt módszert, az általánosított lineáris modellt (GLM), fontossága miatt,

nagyobb részletezettséggel tárgyaljuk. Az általánosított lineáris modellt követően tárgyalt módszerek, kezdeti módszerek. A gyakorlatban csak akkor használják, ha a GLM valamilyen oknál fogva nem használható, vagy gyors kezdeti eredmény előállítása a cél, amellyel tovább lehet lépni.

2.2. Általánosított lineáris modell (GLM)

Az *általánosított lineáris modell*, angol szakirodalomban Generalized Linear Model (GLM), számos területen alkalmazott és jól bevált statisztikai modell. A világon sok ország biztosítói használják portfoliójuk statisztikai elemzéséhez. Kezdetben a faktorok hatásának vizsgálatát külön-külön végezték el. Így azonban nem tudták kezelni a korreláló faktorokat.

Korreláló faktorokra példa személygépjármű esetén, a hengerűrtartalom és a gépjármű teljesítménye közötti összefüggés. A faktorokat külön-külön elemezve, arra jutottak, hogy a nagy hengerűrtartalmú és teljesítményű gépjárművek jelentősen veszélyesebbek. A két faktor korreláltsága miatt mindkét faktor ugyanazt a hatást magyarázza, ha nem vigyázunk előfordulhat, hogy ezt kétszer vesszük figyelembe.



2. ábra. A nem és kor faktorok kölcsönhatása (fiktív példa)

Egy másik fontos példa, amikor az egyik faktor hatása kioltja a másik faktor hatását. Ilyen például a nem és az életkor változók kapcsolata. A faktorok önálló vizsgálata azt mutatta, hogy a férfiak nagyjából ugyanannyira veszélyesek, mint a nők. Azonban kimutatható, hogy a fiatal férfiak veszélyesebbek, mint a fiatal nők és az idős nők veszélyesebbek, mint az idős férfiak. Tehát nem elegendő a nemek faktorát külön elemezni, ugyanis a kor faktor ellentétes hatással van a férfiakra és a nőkre, ezért a hatások kioltják egymást. Ezt a kölcsönhatást láthatjuk a 2. ábrán.

Az előbbi példák mutatják, hogy szükség volt egy olyan módszerre, ami képes együtt is kezelni a faktorokat és megfelelően mérni azoknak hatását a kockázatra. Erre alkalmasnak bizonyult az általánosított lineáris modell. Amelyben az első típusú korreláltságot az egyik faktorok kihagyásával oldják meg, a másodikat, pedig a faktorok kombinálásával.

Lineáris modell

Az általánosított lineáris modell megértéséhez először a lineáris modellt (LM) tárgyaljuk. Amint azt az elnevezés is mutatja, a lineáris modell az általánosított lineáris modell speciális eseteként adódik. A lineáris modell:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{2.2}$$

alakú, ahol

- $\underline{Y}$: a megfigyelések vektora
- $\underline{\beta}$: a modell paramétervektora
- X : a megfigyeléseket magyarázó változók mátrixa
- $\underline{\varepsilon}$: a hibatag, normális eloszlású valószínűségi változó 0 várható értékkel és $I\sigma^2$ szórással

A megfigyelések vektorára ($\underline{Y}$), úgy tekintünk, mint egy valószínűségi változó realizációira, ezt Y-al jelöljük. A modellezés során az alapgondolat az, hogy ez a valószínűségi változó

$$Y = E(Y) + \varepsilon \tag{2.3}$$

formában írható fel és a $\mu = E(Y)$ előáll a magyarázó változók lineáris kombinációjaként.

A megfigyeléseket magyarázó változók mátrixa (X) kategorikus változók estén egy 0-1 mátrix. A mátrix oszlopai a faktorok kategóriáit, sorai pedig, a díjosztályokat reprezentálják. Az érthetőség kedvéért a mátrix felírását bemutatjuk egy példán.

Példa

Tegyük fel, hogy az alábbi táblázat, valamely biztosító kárszámainak megfigyeléseit tartalmazza a gépjármű színe és az életkor faktorok tekintetében.

Kor/Szín	**feketen**	**piros**
18-30 közötti	100	200
31-50 közötti	400	300
50 feletti	100	600

Az életkor és a szín három illetve kettő kategóriája összesen hat díjosztályt határoz meg. Ezekben a díjosztályokban vannak a megfigyeléseink. Ekkor a modell a következőképpen írható fel:

$$\begin{bmatrix} 100 \\ 200 \\ 400 \\ 300 \\ 100 \\ 600 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{21} \\ \beta_{22} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix}$$

Itt a mátrix első három oszlopa az életkor három kategóriáját reprezentálja, az utolsó kettő pedig a színeknek megfelelő rész.

A modell paraméterei ($\underline{\beta}$), azok a paraméterek, amiket meg szeretnénk becsülni. Az előző példában β_{13}, az 50 év feletti gépjárművezetők hatását méri, β_{22}, pedig a piros gépjárművekét. Így kiszámolható az 50 év feletti piros gépjárművel rendelkező vezetők kockázata várható értékének becslése a $\beta_{13} + \beta_{22}$ formula segítségével. A paraméterbecsléseket a hibatagok négyzetösszegének minimalizálásával számíthatjuk ki.

Sajnos ez a modell nem mindig alkalmas a faktorok hatásának mérésére. Egyrészt nagyon erős megkötés számunkra a hibatag normalitása. Másrészt $E(Y)$ sem áll elő mindig a magyarázó változók lineáris kombinációjaként, de előfordulhat, hogy $E(Y)$ valamely függvénye már igen. Ez ad okot az általánosított lineáris modell bevezetésére.

Exponenciális eloszláscsalád

Az átalánosított lineáris modell egyik lényeges momentuma az a feltételezés, hogy a megfigyeléseink exponenciális eloszláscsaládból származnak. Ezért röviden bemutatjuk a két-paraméteres exponenciális eloszláscsaládot és ismertetjük néhány fontos tulajdonságát.

Definíció: Azt mondjuk, hogy az Y valószínűségi változó exponenciális eloszláscsaládból származik, ha sűrűségfüggvénye

$$f_{\theta,\delta}(y) = exp[\frac{y\theta - b(\theta)}{a(\delta)} + c(y,\delta)] \tag{2.4}$$

alakú, ahol θ a kanonikus paraméter, δ a skála paraméter, továbbá

$a(\delta)$: pozitív és folytonos függvény

$b(\theta)$: konvex függvény

$c(y,\delta)$: független θ-tól.

Az $a(\delta), b(\theta), c(x,\delta)$ függvények alkalmas megválasztásával láthatjuk, hogy sok nevezetes eloszlás az exponenciális eloszláscsaládba tartozik. Ilyen például a normális, a poisson és a gamma eloszlás is. A három eloszláshoz tartozó függvényeket a 2. táblázat tartalmazza.

-	**Normális**	**Poisson**	**Gamma**
$a(\delta)$	$\frac{\delta}{\omega}$	$\frac{\delta}{\omega}$	$\frac{\delta}{\omega}$
$b(\theta)$	$\frac{\theta^2}{2}$	e^{θ}	$-ln(-\theta)$
$c(y,\delta)$	$-\frac{1}{2}(\frac{\omega y^2}{\delta} + ln(\frac{2\pi\delta}{\omega}))$	$-ln(y!)$	$\frac{\omega}{\delta}ln(\frac{\omega y}{\delta}) - ln(y) - ln(\Gamma(\frac{\omega}{\delta}))$

2. táblázat. Nevezetes eloszlások exponenciális családban (ω konstans)

Belátható, hogy exponenciális eloszláscsaládban a várható értékre (θ függvénye) és a szórásnégyzetre teljesülnek az alábbi összefüggések:

$$\mu(\theta) = b'(\theta) \tag{2.5}$$

$$\sigma^2 = b''(\theta)a(\delta) \tag{2.6}$$

Definiáljuk a $V(\mu)$ függvényt a $V(\mu(\theta)) = b''(\theta)$ összefüggéssel. Ekkor például $V(\mu) = 1$ normális eloszlás estén és $V(\mu) = \mu$ poisson eloszlás esetén. Ezzel a definícióval a (2.6) össefüggés felírható

$$\sigma^2 = V(\mu)\frac{\delta}{\omega} \tag{2.7}$$

alakban.

Példa: Az $a(\delta) = \frac{\delta}{\omega}$, $b(\theta) = e^{\theta}$, $c(y, \delta) = -ln(y!)$ függvényválasztások esetén, a "sűrűségfüggvény" megkapható (2.4)-ből:

$f_{\theta,\delta}(y) = exp[(y\theta - e^{\theta}) - ln(y!)] = \frac{(e^{\theta})^{y}}{y!}e^{-e^{\theta}}$,

ugyanis $a(\delta) = \frac{\delta}{\omega} = 1$, azaz valóban Poisson eloszlást kapunk.

Az általánosított lineáris modell

Az általánosított lineáris modell két ponton tér el lényegesen a lineáris modelltől. Itt nem azt tesszük fel, hogy a $\mu = E(Y)$ várható érték felírható a magyarázó változók lineáris kombinációjaként, hanem azt, hogy $g(\mu)$ áll elő a magyarázó változók lineáris kombinációjaként. A g függvényt a modell link függvényének nevezzük. Továbbá a lineáris modellben szereplő hibatag (ε) normalitását is enyhítjük, itt a hibatag eloszlásának megválasztásában szabadságunk van.

Az általánosított lieáris modell:

$$\underline{Y} = g^{-1}(X\underline{\beta}) + \underline{\varepsilon} \tag{2.8}$$

ahol $\underline{Y} = (Y_1, \ldots, Y_n)$ és az Y_i megfigyelések exponenciális családból származnak. Feltettük, hogy $g(\mu)$ előáll a magyarázó változók lineáris kombinációjaként, azaz

$$\mu_i = g^{-1}[(X\underline{\beta})_i] \tag{2.9}$$

alakú. Továbbá minden megfigyelés szórásnégyzetére igaz (2.7), azaz

$$\sigma_i^2 = V(\mu)\frac{\delta}{\omega_i}. \tag{2.10}$$

Ahol

- $V(x)$: a variancia függvény
- δ : az exponenciális eloszláscsalád skálaparamétere
- θ : az exponenciális eloszláscsalád kanonikus paramétere
- ω_i : az i-edik megfigyelés súlya.

A kárgyakoriság és átlagkár beágyazása a modellbe:

- kárgyakoriság:

  Jelölje n_{ik} az i-edik cellában a k-adik szerződés kárszámát. Tegyük fel, hogy n_{ik} Poisson eloszlású λ_i paraméterrel. Továbbá jelölje ω_i az i-edik cella szerződéseinek számát. Ekkor

  $$Y_i = \frac{1}{\omega_i} \sum_{k=1}^{\omega_i} n_{ik} \tag{2.11}$$

  jelöli az i-edik cella kárgyakoriságát. Ezekkel a jelölésekkel látható, hogy $\mu_i = \lambda_i$ és $\sigma_i^2 = \frac{\mu_i}{\omega_i}$, amiből (2.7) alapján azonnal következik, hogy $V(\mu_i) = \mu_i$ és $\delta = 1$.

- átlagkár:

  Jelölje n_{ik} az i-edik cellában a k-adik kár mértékét. Tegyük fel, hogy n_{ik} Gamma eloszlású, u_i várható értékkel és $u_i v$ szórással. Ha ω_i jelöli az i-edik cella kárszámát, akkor az i-edik cella átlagkára legyen

  $$Y_i = \frac{1}{\omega_i} \sum_{k=1}^{\omega_i} n_{ik}. \tag{2.12}$$

  Ekkor $\mu_i = u_i$ és $\sigma_i^2 = \frac{u_i^2 v^2}{\omega_i}$, amiből az előzőhöz hasonlóan adódik, hogy $V(\mu_i) = \mu_i^2$ és $\delta = v^2$.

A kárgyakoriság és átlagkár vonatkozásában a tipikus modellparaméterezést a 3.táblázat foglalja össze.

-	**Kárgyakoriság**	**Átlagkár**
Link függvény $(g(x))$	$ln(x)$	$ln(x)$
Hibatag (ε)	$Poisson$	$Gamma$
Skálaparaméter (δ)	1	$\hat{\delta}$
Variancia függvény $(V(x))$	x	x^2
Súly (ω)	szerződésszám	kárszám

3. táblázat. Tipikus modellparaméterezés (GLM)

Az általánosított lineáris modell paraméterbecslését megkaphatjuk a *Maximum Likelihood* (ML) becsléssel. A gyakorlatban ez nem egy gyors eljárás, hiszen $dim(\underline{\beta})$ elég nagy lehet. Ennek következtében a paraméterbecsléseket iteratív úton számítják. Általában a *Newton-Raphson* iterációs technikát használják, amelyet a [8] könyv részletekbemenően tárgyal. Az iteráció $(k+1)$-ik lépése:

$$\underline{\beta}_{k+1} = \underline{\beta}_k - H^{-1} s, \tag{2.13}$$

ahol

- H : a log-likelihood függvény második deriváltjának mátrixa
- s : a log-likelihood függvény első deriváltjának vektora

Az iteráció konvergenciájához szükséges, hogy a kezdőértéket a keresett megoldás közeléből vegyük. Ezért a paraméterek becslését először faktoronként elvégezzük és az így kapott vektor lesz az iteráció kezdőértéke, azaz $\underline{\beta}_0$.

A *GLM* a legtöbb programban megtalálható beépített függvényként. Ilyenek például az *R, Splus, SAS*.

2.3. Bailey és Simon módszere

Az egyszerűség kedvéért tegyük fel, hogy a kockázat két kategorikus változóval jellemezhető. Az első változónak k, a másodiknak, pedig l kategóriája van. Ekkor a két faktor, közösen $k \times l$ díjosztályt határoz meg. Az (i, j) díjosztály esetén a jelölések:

- Kárgyakoriság vizsgálata esetén:
  - r_{ij} : a kárgyakoriság
  - ω_{ij} : a szerződések száma
- Átlagkár vizsgálata esetén:
  - r_{ij} : a átlagkár
  - ω_{ij} : a kárszám

Két modellt különböztetünk meg, az alapján, hogy az egyes cellák várható értéke, milyen alakban áll elő:

- multiplikataív modell:
$$E(r_{ij}) = x_i y_j \tag{2.14}$$
- additív modell:
$$E(r_{ij}) = x_i + y_j \tag{2.15}$$

ahol x_i, y_j a modell paraméterei, $i = 1, \ldots, k$ és $j = 1, \ldots, l$.

A modell a diszkrét illeszkedésvizsgálat próbastatisztikáját veszi alapul. Itt az osztályoknak a díjosztályok felelnek meg ($k \times l$ db.), az osztályok elemszámának, pedig az r_{ij} értékek. Ezekkel a jelölésekkel a próbastatisztika

$$\sum_{i,j} \frac{\omega_{ij}(r_{ij} - E(r_{ij}))^2}{E(r_{ij})} \tag{2.16}$$

alakú és χ^2_{kl-1} eloszlású.

a. Multiplikatív modell:
A (2.16)-ba behelyettesítve (2.14)-et, a veszteségfüggvény a következő alakra módosul:

$$Q = \sum_{i,j} \frac{\omega_{ij}}{x_i y_j}(r_{ij} - x_i y_j)^2 \tag{2.17}$$

Az illeszkedésvizsgálati feladatból adódóan, a paraméterek becslését a Q veszteségfüggvény minimalizálásával kapjuk, ugyanis azon paraméterek mellett lesz a legjobb a modell illeszkedése, amelyekre Q minimális. A Q függvénynek számoljuk ki a paraméterek szerinti parciális deriváltjait. Az egyenleteket megoldva 0-ra a következő egyenletrendszert kapjuk:

$$\begin{cases} x_i = \left(\dfrac{\displaystyle\sum_j \dfrac{\omega_{ij} r_{ij}^2}{y_j}}{\displaystyle\sum_j \omega_{ij} y_j} \right)^{\frac{1}{2}} & (i = 1, ..., k) \\ y_i = \left(\dfrac{\displaystyle\sum_i \dfrac{\omega_{ij} r_{ij}^2}{x_i}}{\displaystyle\sum_i \omega_{ij} x_i} \right)^{\frac{1}{2}} & (j = 1, ..., l) \end{cases} \tag{2.18}$$

Az egyenletrendszert megoldhatjuk iterációval. Az y_j-nek megadva egy kezdőértéket és ezt behelyettesítve az x_i képletébe egy új x_i értéket nyerünk, ezzel elindítva az iterációt.

b. Additív modell:
Az előző ponthoz hasonlóan (2.16)-ba behelyettesítve (2.15)-öt:

$$Q = \sum_{i,j} \frac{\omega_{ij}}{x_i + y_j}(r_{ij} - (x_i + y_j))^2 \tag{2.19}$$

Ugyanúgy, mint a multiplikatív modellben, itt is Q minimalizálásából kapjuk a paraméterek becslését. A minimalizálást elvégezve a következő egyenletrendszer adódik:

$$\begin{cases} \sum\limits_j \dfrac{\omega_{ij} r_{ij}^2}{(x_i + y_j)^2} = \sum\limits_j \omega_{ij} \quad (i = 1, ..., k) \\ \sum\limits_i \dfrac{\omega_{ij} r_{ij}^2}{(x_i + y_j)^2} = \sum\limits_j \omega_{ij} \quad (j = 1, ..., l) \end{cases} \tag{2.20}$$

Ezt az egyenletrendszert a *Newton-Raphson* iterációs módszer segítségével oldhatjuk meg. Az iteráció $(k+1)$-ik lépése a következő:

$$\begin{cases} x_i^{(k+1)} = x_i^{(k)} + \dfrac{\sum\limits_j \omega_{ij} (g_{ij}^{(k)})^2 - \sum\limits_j \omega_{ij}}{2 \sum\limits_j (\omega_{ij} (g_{ij}^{(k)})^3 / r_{ij})} \quad (i = 1, ..., k) \\ y_j^{(k+1)} = y_j^{(k)} + \dfrac{\sum\limits_i \omega_{ij} (g_{ij}^{(k)})^2 - \sum\limits_i \omega_{ij}}{2 \sum\limits_i (\omega_{ij} (g_{ij}^{(k)})^3 / r_{ij})} \quad (j = 1, ..., l) \end{cases} \tag{2.21}$$

ahol $g_{ij}^{(k)} = \dfrac{r_{ij}}{x_i^{(k)} + y_j^{(k)}}$

2.4. Peremátlag módszer

Megtartva az előző pont jelöléseit, a peremátlag módszer alapgondolata, hogy a peremátlagok jó közelítést adnak a megfelelő marginális eloszlás várható értékére, azaz

$$\begin{cases} \sum\limits_j \omega_{ij} E(r_{ij}) = \sum\limits_j \omega_{ij} r_{ij} \quad (i = 1, ..., k) \\ \sum\limits_i \omega_{ij} E(r_{ij}) = \sum\limits_i \omega_{ij} r_{ij} \quad (j = 1, ..., l) \end{cases} \tag{2.22}$$

a. Multiplikatív modell

Tegyük fel, hogy $E(r_{ij}) = x_i y_j$ alakú. Ekkor (2.22) alapján a

$$\begin{cases} \sum\limits_j \omega_{ij} x_i y_j = \sum\limits_j \omega_{ij} r_{ij} \quad (i = 1, ..., k) \\ \sum\limits_i \omega_{ij} x_i y_j = \sum\limits_i \omega_{ij} r_{ij} \quad (j = 1, ..., l) \end{cases} \tag{2.23}$$

egyenletrendszer adódik. Ez egy $k+l$ egyenletből álló $k+l$ ismeretlenes egyenletrendszer. Ezt megoldhatjuk iterációval. Vegyük észre, hogy (2.23)-ban az első egyenletből

x_i, a másodikból pedig y_j kifejezhető. Így átrendezéssel a következő egyenletrendszert kapjuk:

$$\begin{cases} x_i = \dfrac{\sum\limits_j \omega_{ij} r_{ij}}{\sum\limits_j \omega_{ij} y_j} & (i = 1, ..., I) \\ y_j = \dfrac{\sum\limits_i \omega_{ij} r_{ij}}{\sum\limits_i \omega_{ij} x_i} & (i = 1, ..., I) \end{cases} \tag{2.24}$$

Az y_j kezdőértékeket alkalmasan megválasztva, iterációval megoldhatjuk a feladatot.

b. Additív modell
Tegyük fel, hogy $E(r_{ij}) = x_i + y_j$ alakú. Ekkor a

$$\begin{cases} \sum\limits_j \omega_{ij}(x_i + y_j) = \sum\limits_j \omega_{ij} r_{ij} & (i = 1, ..., k) \\ \sum\limits_i \omega_{ij}(x_i + y_j) = \sum\limits_i \omega_{ij} r_{ij} & (j = 1, ..., l) \end{cases} \tag{2.25}$$

egyenletrendszert adódik. Az elsőből kifejezve x_i-t, a másodikból pedig y_j-t az

$$\begin{cases} x_i = \dfrac{\sum\limits_j \omega_{ij}(r_{ij} - y_j)}{\sum\limits_j \omega_{ij}} & (i = 1, ..., k) \\ y_j = \dfrac{\sum\limits_i \omega_{ij}(r_{ij} - x_i)}{\sum\limits_i \omega_{ij}} & (j = 1, ..., l) \end{cases} \tag{2.26}$$

egyenletrendszert kapjuk, amelyet iteratív úton megoldhatunk.

2.5. Legkisebb négyzetek módszere

A modellezés alapgondolata, hogy a várható érték $\mu = E(r_{ij})$ *legkisebb négyzetes becslése* az a μ, ami az

$$S = \sum_{i,j} \omega_{ij}(r_{ij} - \mu)^2 \tag{2.27}$$

veszteségfüggvényt minimalizálja.

a. Multiplikatív modell

Tegyük fel, hogy $E(r_{ij}) = x_i y_j$ alakú. Ekkor a fenti veszteségfüggvény

$$S = \sum_{i,j} \omega_{ij}(r_{ij} - x_i y_j)^2 \tag{2.28}$$

alakban írható fel. A modell paramétereinek becslését az S minimalizálásával kapjuk. S parciális deriváltjait kiszámítva a következő egyenletrendszer adódik:

$$\begin{cases} x_i = \dfrac{\sum\limits_j \omega_{ij} r_{ij} y_j}{\sum\limits_j \omega_{ij} y_j^2} & (i = 1, \ldots, k) \\ y_j = \dfrac{\sum\limits_i \omega_{ij} r_{ij} x_i}{\sum\limits_i \omega_{ij} x_i^2} & (j = 1, \ldots, l) \end{cases} \tag{2.29}$$

Ezt iteratívan megoldva nyerjük a paraméterek becslését.

b. Additív modell

Tegyük fel, hogy $E(r_{ij}) = x_i + y_j$ alakú. Ezt behelyettesítve az eredeti veszteségfüggvénybe

$$S = \sum_{i,j} \omega_{ij}(r_{ij} - (x_i + y_j))^2 \tag{2.30}$$

alakú. A paraméterek becslését S minimalizálásával kapjuk. Az előzőekhez hasonlóan kiszámítva a parciális deriváltakat a (2.26) iterációs formula adódik.

2.6. A területi faktor és modellezésének problémái

A területi faktor modellezése nem végezhető el a fenti módszerek egyikével sem. Ennek oka, hogy a két dimenzióban mért adatokat nem tudjuk egyszerű módszerekkel csoportosítani. Ez egy dimenzióban nem nehéz feladat, ilyen például az életkor faktor kategorizálása. Az adatok nagyon kusza, átláthatatlan képet mutatnak. Ebben az értelemben a feladat felfogható úgy is, mint egy szemcsés kép kitisztítása.

Természetesen, ha a területi faktort, valamilyen módszerrel kategorizálni tudjuk, akkor minden további nélkül beépíthető a GLM-be.

A kezdeti feltételezéseink:

1. A modellezés során feltesszük, hogy a területi kockázat várható értékét egy "sima" felület írja le. A megfigyeléseink hibával terheltek, ezért mi nem látjuk tisztán ezt a felületet.

2. A felület simaságát arra a feltételezésünkre alapozzuk, hogy a közeli régiók kockázata hasonlít egymásra és ez a hasonlóság a távolság valamilyen monoton növő függvényében csökken. Ez éppen egy lokális simasági tulajdonságot eredményez a felületen.

3. A területi faktor modellezése során a feladatunk mindig a fent említett felület becslése. A becsült felület egy régióhoz tartozó értékét, a régió simított értékének nevezzük.

A következő pontokban három modell kerül ismertetésre, amelyek alkalmasak lehetnek az eredeti felület becslésére.

3. Whittaker modell

A területi simítás egyik lehetséges megoldása a Whittaker féle simítás. A *Whittaker kritérium* két részből áll, az egyik a *simító kritérium*, a másik az *illeszkedési kritérium*. A simító kritérium minimalizálásával törekszünk a felület simaságára, az illeszkedési kritérium minimalizálásával, pedig arra, hogy a becsült felület minél pontosabban illeszkedjen a megfigyelt adatokra. Érzékelhető, hogy bármely kritérium minimalizálása során a másik kritérium értéke nő. Például a simító kritérium önmagában akkor minimális, ha a becsült felület konstans, de így az illeszkedési kritérium értéke nagyon megnő. A cél, hogy a két kritérium értékét együtt minimalizáljuk, ezzel kompromisszumot teremtve a felület simasága és illeszkedése között.

Legyen adott n darab jól beazonosítható terület, mondjuk az irányítószámok alapján. Tekintsünk egy olyan μ várható értékű X valószínűségi változót, amelynek várható értéke r darab változó függvényeként áll elő, azaz

$$\mu = f(u_1, ..., u_r) \tag{3.1}$$

Például biztosítási területen, X lehet a kockázat (pl.: kárszám, kárgyakoriság, stb. . .), a változók pedig a különböző faktorokat reprezentáló valószínűségi változók. Természetesen az r értéke előre rögzített, hiszen ez adja meg a faktoraink számát. Az egyszerűség és átláthatóság kedvéért r=3 esetén kerül bemutatásra a modell. A későbbiekben láthatjuk, hogy ez semmilyen hátránnyal nem jár, mert a modellben minden lépés könnyedén általánosítható tetszőleges számú faktor esetére.

Tehát az X valószínűségi változó várható értéke, r=3 változó függvénye, amely változók a faktorokat reprezentálják. Ekkor ezen változók lehetséges értékei díjosztályokat alakítanak ki. Egy ilyen osztály

$$(\omega_{ijk}, X_{ijk}) \tag{3.2}$$

ahol

- i: a területi faktor egy kategóriája
- j, k: két másik, nem területi faktor egy-egy kategóriája
- X_{ijk}: az (i, j, k) osztály kockázata
- ω_{ijk}: az (i, j, k) osztály súlya (pl.: szerződésszám)

Ezek után legyen $\mu_{ijk} = E(X_{ijk})$ és tegyük fel, hogy

$$\mu_{ijk} = \alpha_i \beta_{jk} \tag{3.3}$$

ahol α_i az i-edik régióban a területi faktor által indukált kockázat várható értéke, amit meg szeretnénk becsülni és β_{jk} a nem területi faktorokhoz tartozó kockázat várható értéke, amiről feltesszük, hogy már rendelkezésünkre áll a becslése egy másik modellből.

A következő lépésben definiáljuk Y_i-t a következőképpen:

$$Y_i = \frac{\sum_{j,k} \omega_{ijk} (\frac{X_{ijk}}{\beta_{jk}})}{\sum_{j,k} \omega_{ijk}}. \tag{3.4}$$

Ekkor $E(Y_i) = \alpha_i$, $D^2(Y_i) = \sigma^2/\omega_i$, ahol $\omega_i = \sum_{j,k} \omega_{ijk}$ és $\sigma^2 > 0$ konstans. Ezzel valójában kiszűrtük az adatokból a nem területi hatást. A feladatunk így arra módosult, hogy becslést adjunk Y_i várható értékére, azaz α_i-re.

A becslést a Whittaker kritérium minimalizálásával nyerjük, amely

$$F = D + lS \tag{3.5}$$

alakú. D jelöli az illeszkedési kritériumot, S pedig a simító kritériumot. Az l konstanst tapasztalati úton szokták meghatározni. A két kritérium vizsgálatával a következő pontokban részletesen foglalkozunk. Ehhez szükségünk lesz egy kis jelölésbeli változtatáshoz a továbblépés érdekében. Jelölje x_i az i-edik régió egy pontját, amit a koordinátáival adunk meg. A továbbiakban ezt a pontot fogjuk használni a régió beazonosítására, ezért célszerű lenne ezt azon település koordinátáinak választani, amely szerint a régiót meghatároztuk. Ennek legfőbb oka az, hogy az adott régióban mért adatok nagy része onnan származik. A továbbiakban az Y_i, α_i jelölések helyett az $Y(x_i), \alpha(x_i)$ jelöléseket fogjuk használni.

3.1. Az illeszkedési kritérium és formalizálása

Az illeszkedési kritérium azt hivatott mérni, hogy a becsült felületünk mennyire illeszkedik jól a megfigyelt adatokra. Legyenek adottak $x_1, ..., x_n$ ($x_i \in \mathbb{R}^2$) pontok a síkon. Minden i-re az x_i ponthoz tartozik egy valószínűségi változó Y_i, ami az i-edik régió kockázatának azon részét reprezentálja, amelyet a területi faktor magyaráz. A

célunk, hogy megbecsüljük az $\alpha(x_i)$ várható értéket. Ehhez tekintsük a következő veszteségfüggvényt:

$$D = \sum_{i=1}^{n} \omega_i [Y(x_i) - W(x_i)]^2. \tag{3.6}$$

Ez azt jelenti, hogy ha a régiókban simított értékeknek a $W(x_i)$-ket $(i = 1, ..., n)$ választjuk, akkor a vizsgált tartományon az illesztés során vétett hibánk összértéke éppen D. Tehát elsődlegesen az a feladatunk, hogy a D értékét minimalizáljuk, azaz azon $W(x_i)$ $(i = 1, ..., n)$ értékeket válasszuk, amikre a D minimális.

A minimalizálás érdekében írjuk fel a D illeszkedési kritériumot mátrixos alakban. Ehhez vezessük be a következő jelöléseket:

- $z_i = W(x_i)$: az i-edik régió simított értéke (ezt szeretnénk megbecsülni, minden régió esetén)
- $z = [z_1, \ldots, z_n]$
- $Y = [Y_1, \ldots, Y_n]$: a régiók megfigyeléseinek vektora
- $\Gamma = diag(\omega_1, \ldots, \omega_n)$: a súlyok diagonális mátrixa

Ezekkel a jelölésekkel az illeszkedési kritérium (3.6) mátrixos alakja:

$$D = (Y - z)^T \Gamma (Y - z). \tag{3.7}$$

3.2. A simító kritérium

A simító kritérium a felület simaságáért felelős. A megfigyelt adatok a valóságban nagyon zavaros képet mutatnak, ezért szükség van egy olyan felület becslésére, amelyről már megállapítható a területi kockázat eloszlása és így elvégezhető a régiók díjosztályokba sorolása.

A simító kritérium a *TPS* spline interpolációs elméletből lett kölcsönözve, ahol ezt a *felület energiájának* szokták nevezni és a következő formula írja le:

$$S = \sum_{i=1}^{n} \underbrace{[\ (\Delta_{11}^2 W(x_i))^2 + 2(\Delta_{12}^2 W(x_i))^2 + (\Delta_{22}^2 W(x_i))^2\ ]}_{S(x_i)}, \tag{3.8}$$

ahol $\Delta_{p,q}^2$ a megfelelő parciális deriváltakat jelöli.

A legfőbb problémát számunkra $\Delta_{pq}^2 W(x_i)$ kiszámítása jelenti. Ennek az oka az, hogy $W(x_i)$-re nem ismerünk olyan formulát, amely differenciálható x_i két koordinátája szerint. Valójában $W(x_i)$-re semmilyen formula nem áll a rendelkezésünkre. Ez a probléma feloldható.

Legyen h előre rögzített természetes szám. Vegyük az x pont h darab szomszédját, azaz a hozzá legközelebb eső h darab pontot, jelölje ezeket $v_1, v_2, ..., v_h$, $x = v_1$. Ezen pontokra illesszünk egy qvadratikus formát, jelölje ezt $Q_x(.)$. Ekkor ez a felület közelítése a $W(x)$ értékeknek a $v_1, v_2, ..., v_h$ pontokban.

3.3. Qvadratikus forma lokális illesztése

Legyenek adottak a $v_1, v_2, ..., v_h \in \mathbb{R}^2$ pontok a síkon. Továbbá jelölje $z_i = W(v_i)$, $(i = 1, ..., h)$ a simított értékeket. Ekkor a következő formula írható fel:

$$z_i = Q(v_i) + \varepsilon_i \qquad (i = 1, ..., h), \tag{3.9}$$

ahol $Q : \mathbb{R}^2 \to \mathbb{R}$ qvadratikus forma, ε pedig a hiba. Ez a formula azt hivatott reprezentálni, hogy $Q(.)$ qvadratikus forma, olyan felületet definiál, ami a $v_1, v_2, ..., v_h$ pontokban a $z_1, z_2, ..., z_h$ értékekhez közeli értékeket vesz fel. Azaz a $Q(.)$ a z_i érték becslése a v_i pontban.

Egy qvadratikus formát egyértelműen meghatároznak az együtthatói. Azaz, ha $q = (q_1, q_2, ..., q_6)$ jelöli az együtthatók vektorát, akkor

$$Q(v) = q^T \bar{v} \tag{3.10}$$

alakú, ahol $\bar{v} = (v_1^2, v_1 v_2, v_2^2, v_1, v_2, 1)^T$. Ekkor (3.9) és (3.10) alapján a

$$z_i = q^T \bar{v}_i + \varepsilon_i \qquad (i = 1, ..., h) \tag{3.11}$$

többváltozós lineáris regressziós egyenlet adódik. Ebben az esetben létezik ismert becslése az együtthatóknak, mégpedig

$$\hat{q} = Az, \tag{3.12}$$

ahol $A = (X^T X)^{-1} X^T$, $z = (z_1, z_2, ..., z_h)^T$ és $X = (\bar{v}_1, \ldots, \bar{v}_h)^T$ $(h \times 6)$-os mátrix. Ezzel (3.10) alapján a lokálisan illesztett qvadratikus forma

$$Q(v) = \hat{q}^T \bar{v}, \tag{3.13}$$

ahol $\hat{q}$, a regressziós együtthatók becsléseinek vektorát jelöli.

3.4. A simító kritérium formalizálása

Az eddigiekhez hasonlóan, legyen h előre rögzített természetes szám. Vegyük az x pont h darab szomszédját, azaz a hozzá legközelebb eső h darab pontot, jelölje ezeket

$v_{(1)}, v_{(2)}, ..., v_{(h)}$, $x = v_{(1)}$. Ezen pontokra illesszünk egy qvadratikus formát, jelölje ezt $Q_x(.)$. Továbbá jelölje $z_{(i)} = W(v_{(i)})$, $(i = 1, ..., h)$ a simított értékeket. A $Q_x(.)$ qvadratikus formára kapott eredmény felhasználásával felírjuk a simító kritériumot $(S(x))$ mátrixos alakban. Ehhez tekintsük a (3.13)-ban felírt

$$Q_x(v) = q_x^T \bar{v}, \tag{3.14}$$

qvadratikus formát, ahol $\bar{v} = (v_1^2, v_1 v_2, v_2^2, v_1, v_2, 1)^T$, és $v = (v_1, v_2)$. Ekkor az együtthatók becslése

$$q_x = A_x z_x, \tag{3.15}$$

ahol $z_x = (z_{(1)}, z_{(2)}, ..., z_{(h)})^T$, $A_x = (X^T X)^{-1} X^T$ és $X = (\bar{v}_1, \ldots, \bar{v}_h)^T$ $(h \times 6)$-os mátrix. Ekkor az $A_x = (X^T X)^{-1} X^T$ egy $(6 \times h)$-as mátrix.

A továbbiakban tekintsük a simított értékek teljes

$$z = [z_1, z_2, ..., z_n]^T \tag{3.16}$$

vektorát. Az eddigi jelölések szerint, $z_x = [z_{(1)}, z_{(2)}, ..., z_{(h)}]^T$ a $z = [z_1, z_2, ..., z_n]^T$ részvektora, amelyet a $v_{(1)}, v_{(2)}, ..., v_{(h)}$ pontoknak megfelelő régiók jelölnek ki z-ből. Azaz, ha a $v_{(1)}, v_{(2)}, ..., v_{(h)}$ pontoknak megfelelő régiókat rendre $x_{k_1}, x_{k_2}, ..., x_{k_h}$ jelöli, ahol $\{k_1, k_2, ..., k_h\}$ az $\{1, 2, ..., n\}$ valamely h elemű részhalmaza, akkor

$$z_x = [z_{k_1}, z_{k_2}, ..., z_{k_h}]^T. \tag{3.17}$$

Továbbá tekintsük, azt a B_x $(6 \times n)$-es mátrixot, amelynek a $\{k_1, k_2, ..., k_h\} \subset \{1, 2, ..., n\}$ oszlopain kívül minden oszlopa 0, és a k_j-edik oszlopa megegyezik az A_x mátrix j-edik oszlopával $(j = 1, ..., h)$. Azaz, ha

$$A_x = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1h} \\ a_{21} & a_{22} & \ldots & a_{2h} \\ a_{31} & a_{32} & \ldots & a_{3h} \\ a_{41} & a_{42} & \ldots & a_{4h} \\ a_{51} & a_{52} & \ldots & a_{5h} \\ a_{61} & a_{62} & \ldots & a_{6h} \end{bmatrix}$$

akkor ebből

$$B_x = \begin{bmatrix} 0 & \dots & 0 & a_{11} & 0 & \dots & 0 & a_{12} & 0 & \dots & 0 & \dots & \dots & a_{1h} & \dots & 0 \\ 0 & \dots & 0 & a_{21} & 0 & \dots & 0 & a_{22} & 0 & \dots & 0 & \dots & \dots & a_{2h} & \dots & 0 \\ 0 & \dots & 0 & a_{31} & 0 & \dots & 0 & a_{32} & 0 & \dots & 0 & \dots & \dots & a_{3h} & \dots & 0 \\ 0 & \dots & 0 & a_{41} & 0 & \dots & 0 & a_{42} & 0 & \dots & 0 & \dots & \dots & a_{4h} & \dots & 0 \\ 0 & \dots & 0 & a_{51} & 0 & \dots & 0 & a_{52} & 0 & \dots & 0 & \dots & \dots & a_{5h} & \dots & 0 \\ 0 & \dots & 0 & \underbrace{a_{61}}_{k_1} & 0 & \dots & 0 & \underbrace{a_{62}}_{k_2} & 0 & \dots & 0 & \dots & \dots & \underbrace{a_{6h}}_{k_h} & \dots & 0 \end{bmatrix}$$

$(6 \times n)$-es mátrix. Ekkor (3.15) felírható

$$q_x = B_x z \tag{3.18}$$

alakban.

$\Delta^2_{pq} W(x)$-et közelítsük $Q_x(.)$ qvadratikus forma megfelelő parciális deriváltjaival. A

$$Q_x(v) = q_x^T \bar{v} = q_1 v_1^2 + q_2 v_1 v_2 + q_3 v_2^2 + q_4 v_1 + q_5 v_2 + q_6 \tag{3.19}$$

felírásból egyből kiszámolhatóak a megfelelő parciális deriváltak:

- $\Delta^2_{11} Q_x(v) = 2q_1$
- $\Delta^2_{12} Q_x(v) = q_2$
- $\Delta^2_{22} Q_x(v) = 2q_3$

A továbbiakban jelölje $d = [2q_1, q_2, 2q_3]^T$ a parciális deriváltak vektorát és D_x $(3 \times n)$-es mátrix, azt a mátrixot, amelyet úgy kapunk meg a B_x mátrixból, hogy elhagyjuk az utolsó 3 sorát. Továbbá legyen $G = diag(1/2, 1, 1/2)$ mátrix. Ekkor (3.18) alapján érvényes a

$$G d_x = D_x z \tag{3.20}$$

formula, melynek segítségével lokálisan felírható a simító kritérium

$$S(x) = d_x^T G^T C G d_x = z^T D_x^T C D_x z \tag{3.21}$$

alakban, ahol $C = diag(1, 2, 1)$ diagonális mátrix.

Ezt minden x_i $(i = 1, ..., n)$ pontra elvégezve és behelyettesítve a (3.8) formulába, amely megadja a simító kritérium globális értékét a vizsgált tartományon,

$$S = z^T \underbrace{[\sum_{i=1}^{n} D_{x_i}^T C D_{x_i}]}_{M} z \tag{3.22}$$

alakot kapjuk.

3.5. A Whittaker kritérium és minimalizálása

A Whittaker kritériumot (3.5)

$$F = D + lS$$

két nagy összetevő határozza meg, D az illeszkedési kritérium és S a simasági kritérium. Mindkét kritérium mátrixos alakját megadtuk az előző pontokban. Így (3.22) és (3.7) alapján a Whittaker kritérium mátrixos alakja

$$F = (Y - z)^T \Gamma (Y - z) + l z^T M z. \tag{3.23}$$

F minimalizálását elvégezve z-ben, megkapjuk a $z_i = W(x_i)$ simított értékeket az egyes régiókra. A minimalizálásához azonban szükségünk van rá, hogy tudjunk mátrix szerint deriválni. Ezen a ponton kitérünk a mátrixok szerinti deriválásra, de csupán amennyire ezt a jelenlegi feladatunk megköveteli.

Definíció: Legyen A egy (n x p)-es mátrix, és $f : \mathbb{R}^{n \times p} \to \mathbb{R}$ számértékű függvény. Ekkor legyen definíció szerint

$$\frac{\partial f(A)}{\partial A} = (\frac{\partial f(A)}{\partial A_{ij}}) \tag{3.24}$$

$(n \times p)$-es mátrix.

A minimalizálási feladatunkhoz a mátrix szerinti deriválás két alapvető tulajdonságát fogjuk használni:

1. $\frac{\partial c^T a}{\partial a} = a$ ahol $a, c \in \mathbb{R}^n$
2. $\frac{\partial a^T C a}{\partial a} = 2Ca$ ahol $a \in \mathbb{R}^n$, $C \in \mathbb{R}^{n \times n}$

Állítás: A (3.23)-ben megadott F Whittaker kritérium minimumhelye

$$z = (I_n + l\Gamma^{-1} M)^{-1} Y. \tag{3.25}$$

Bizonyítás:
Tekintsük tehát a Whittaker kritérium

$F = (Y - z)^T \Gamma (Y - z) + l z^T M z$

mátrixos alakját. Ha ezt szorzatra bontjuk, akkor az

$F = Y^T \Gamma Y + z^T \Gamma z - Y^T \Gamma z - z^T \Gamma Y + l z^T M z$

formula adódik, ezt kell deriválnunk z szerint. A mátrix szerinti deriválás, említett két tulajdonságát használjuk.
Ekkor

$$\frac{\partial F}{\partial z} = 2\Gamma z - \Gamma Y - \Gamma Y + 2lMz$$
$$\Gamma z - \Gamma Y + lMz = 0$$
$$(\Gamma + lM)z = \Gamma Y$$
$$(I_n + l\Gamma^{-1}M)z = Y$$
$$z = (I_n + l\Gamma^{-1}M)^{-1}Y$$

adódik a minimalizálás eredményeként.

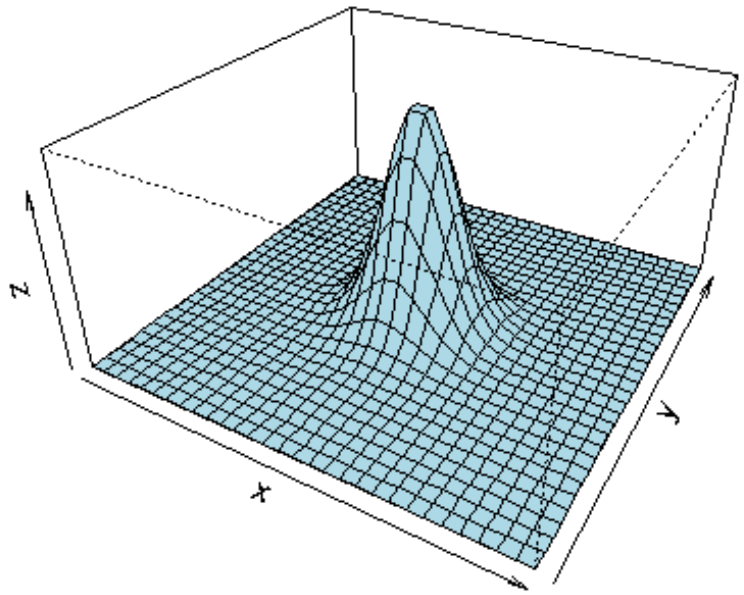
A modell explicit eredményt ad a régiók simított értékeire. Előfordulhat, hogy a kapott eredmény még nem megfelelő a területi faktor kategóriáinak meghatározására. Ekkor a modell újraillesztésével finomíthatjuk az eredményt, ha azokat a szomszédos régiókat összevonjuk, amikre a z értéke nagyon közeli.

3.6. A Whittaker modell tesztelése

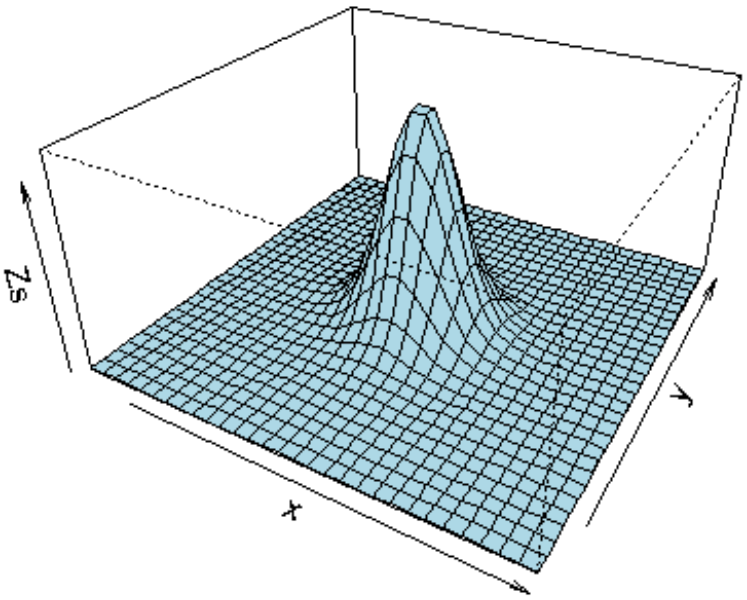
A Whittaker modellt két tesztfeladaton vizsgáljuk. Az egyik egy domb (sima ill. szórt) a másik pedig egy lépcső, ami véletlen számokkal van megszórva. Az első tesztfelület a 3. ábrán látható. Ennek simítását a 4. ábra mutatja. Megállapíthatjuk, hogy a simítás nem változtatta meg jelentősen az eredeti felületünket.

Az eredeti tesztfelület szórt állapota látható a 5. ábrán. A simított felület (6. ábra) mutatja, hogy a modell bizonyos mértékig érzékeny a koordinátákra, ugyanis a kicsi koordinátájú tartományban jelentősebb simítást tapasztalunk. Ennek egyik oka lehet a qvadratikus forma lokális illesztésének használatából fakadó hiba. Ennek ellenére az eredeti felület alakja jól kivehetően kirajzolódik és az adatok szórása is csökken, még a nagyobb koordinátájú tartományban is.

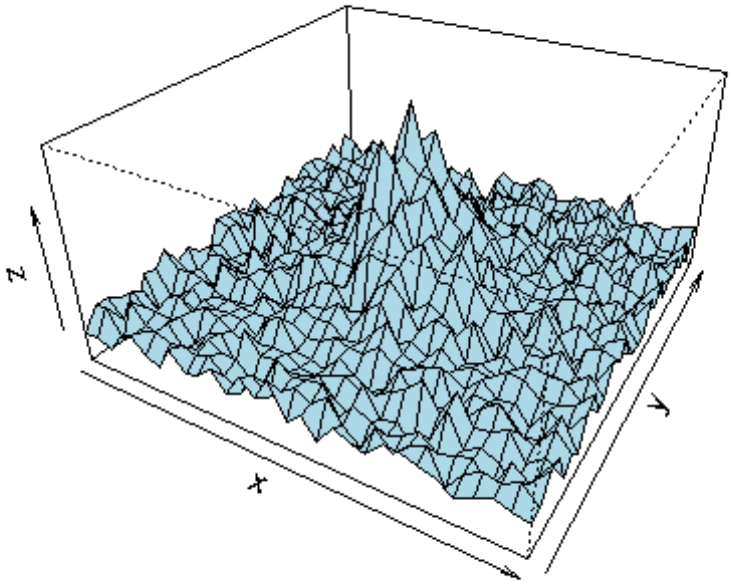
A második tesztfelület egy szórt lépcső (7. ábra), amelyet a törésvonal detektálása miatt választottunk. A simított felület (8. ábra) az előzőhöz hasonlóan azt mutatja, hogy a kisebb koordinátájú tartományban erősebb a simítás mértéke. A simítás azonban a lokális hibáktól eltekintve megtalálja a törésvonalat és tisztítja a képet is .



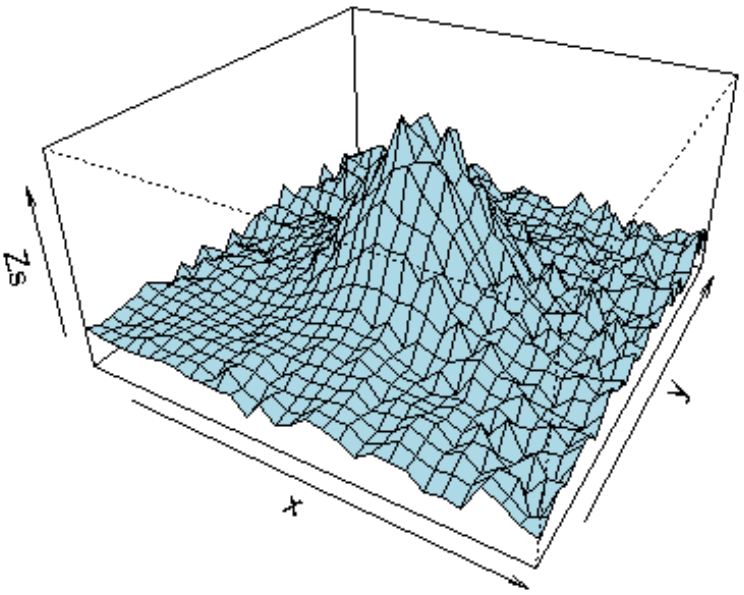
3. ábra. Whittaker: eredeti felület



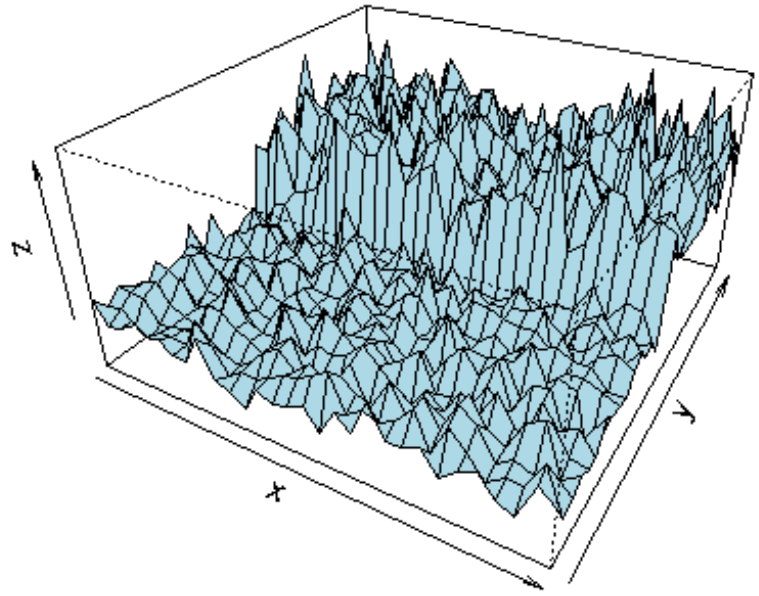
4. ábra. Whittaker: eredeti felület simítása



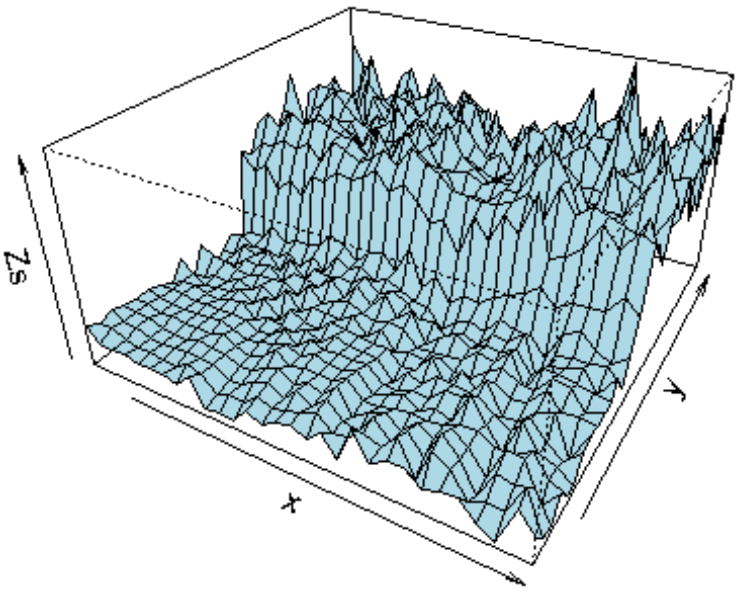
5. ábra. Whittaker: első tesztfelület



6. ábra. Whittaker: első tesztfelület simítása



7. ábra. Whittaker: második tesztfelület



8. ábra. Whittaker: második tesztfelület simítása

4. Credibility modell

4.1. Credibility becslések

A biztosítási matematika egyik fontos kérdése, hogy a biztosító mennyire támaszkodjon saját kártapasztalatára és mennyire az általános információra. Ez a problémakör nagyon sok területet felölel. Ennek egyik iskolapéldája, hogy az újonnan piacra lépő biztosító, mennyire vegye figyelembe saját kártapasztalatát és mennyire azokat, amikhez más biztosítóktól fér hozzá. Akkor is ezzel a problémával állunk szemben, ha díjosztályok kockázatát szeretnénk becsülni és valamelyik díjosztályban kevés szerződés áll rendelkezésünkre. Ekkor az a kérdés, hogy a többi díjosztály kártapasztalatát mekkora súllyal számítsuk be. Hasonló problémák feloldására nyújthat lehetőséget a *credibility elmélet*, aminek kezdeti formája

$$T = zM + (1 - z)M_0 \tag{4.1}$$

alakú, ahol

- T : a credibility becslés
- z : a Bühlmann faktor, $0 \leq z \leq 1$
- M : a saját tapasztalatokon alapuló becslés
- M_0 : külső információkon alapuló becslés.

A nehézséget a Bühlmann faktor meghatározása jelenti, amely arra ad választ, hogy mennyire kell figyelembe venni a saját tapasztalatot és mennyire a külső információt.

A credibility elmélet a Bayesi hozzáállást veszi alapul. Tegyük fel, hogy X valószínűségi változó az $\mathcal{F} = \{\mathrm{F}_\theta : \theta \in \Theta\}$ (Θ a paramétertér) eloszláscsaládból származik. Jelölje $X_1, X_2, \ldots, X_n$ a független, azonos eloszlású mintát. A klasszikus becsléselmélethez hasonlóan itt is az a feladatunk, hogy a minta alapján becslést adjunk az eloszlás paraméterére, azaz θ-ra, vagy annak egy függvényére $g(\theta)$-ra. A Bayesi hozzáállásban a θ paramétert is valószínűségi változónak tekintjük. A Bayesi statisztika alapvető fogalmai és jelölései:

- **a priori eloszlás** (Q): θ eloszlása, a sűrűségfüggvényt q-val jelöljük
- (P_t): az $X|\theta = t$ eloszlása, a sűrűségfüggvényt f_t-vel jelöljük
- **prediktív eloszlás** (P): X feltétel nélküli eloszlása, a sűrűségfüggvényt f-el jelöljük

- **a poszteriori eloszlás** (Q_x): $\theta|X = x$ eloszlása, a sűrűségfüggvényt q_x-el jelöljük

A fenti srűségfüggvények nem feltétlenül léteznek. Ha léteznek, akkor felírható a Bayes formula, amely

$$q_x(t) = \frac{f_t(x)q(t)}{f(x)} \tag{4.2}$$

alakú.

Legyen $w : \Theta \times \mathbb{R}^k \to \mathbb{R}$ tetszőleges veszteségfüggvény, ekkor $g(\theta)$ Bayes becslése az a T statisztika, ami minimalizálja az $R_T(Q) = E[w(T, g(\theta))]$ *a priori rizikót*, azaz

$$R_{\hat{g}(\theta)}(Q) = \min_T R_T(Q) \tag{4.3}$$

Ez azt jelenti, hogy $g(\theta)$ Bayes becslése minimalizálja az átlagos veszteséget. Ismert, hogy négyzetes veszteségfüggvény esetén a $g(\theta)$ Bayes becslése, megegyezik az *a poszteriori eloszlás* várható értékével, azaz

$$\hat{g}(\theta) = E(g(\theta)|X_1, X_2, \ldots, X_n). \tag{4.4}$$

A továbbiakban ezt nevezzük credibility becslésnek. A becsléshez meg kell határoznunk az a poszteriori eloszlást, azonban ez sokszor nem egyszerű feladat. Ezért a megfigyelések lineáris kombinációi között keressük $g(\theta)$ credibility becslését, amire teljesül, hogy

$$R_{\hat{g}(\theta)}(Q) = \min_T R_T(Q)$$

ahol a minimumot olyan T statisztikákra keressük, amelyek

$$T = c_0 + \sum_{i=1}^{n} c_i x_i \tag{4.5}$$

alakúak. Ezt nevezzük $g(\theta)$ lineáris credibility becslésének. A legtöbb esetben minket $\mu(\theta) = E(X|\theta)$ becslése érdekel, az ilyen jellegű becslésekkel a Bühlmann modellek foglalkoznak. A könnyebb érthetőség kedvéért először a klasszikus Bühlmann modellt tárgyaljuk, majd rátérünk a Bühlmann-Straub modellre és alkalmazására.

4.2. Klasszikus Bühlmann modell

Tegyük fel, hogy n különböző biztosítási módozatból származó megfigyeléseink vannak h éven keresztül. Jelölje a j-edik módozat megfigyeléseit

$$\underline{X}_j = (X_{j1}, \ldots, X_{jh}) \qquad (j = 1, \ldots, n). \tag{4.6}$$

Tegyük fel továbbá, hogy az i-edik módozat kockázatát szeretnénk felmérni, de nem áll rendelkezésünkre elegendő adat, hogy ezt megtehessük. Ezért valamilyen módon fel szeretnénk használni a többi módozat kártapasztalatát is. A feladat elvégzéséhez először vezessünk be néhány jelölést:

- θ_j: a j-edik módozat rizikóparamétere, amely nem más, mint a megfigyelések eloszlásának paramétere (feltesszük, hogy ez módozatonként különböző)
- $\mu(\theta_j) = E(X_{jt}|\theta_j)$: valószínűségi változó, a paraméterünk egy függvénye, amire becslést szeretnénk adni
- $\sigma^2(\theta_j) = D^2(X_{jt}|\theta_j)$: a j-edik módozat kockázatának feltételes szórásnégyzete
- $m = E(X_{jt}) = E(\mu(\theta_j))$: a közös várható érték az összes módozat káradata alapján, ha ez nem ismert, akkor helyette az ő becslését

  $$M_0 = \frac{1}{nh} \sum_{j,t} x_{jt} \tag{4.7}$$

  átlagot használjuk
- $a = D^2(\mu(\theta_j))$: a módozatokhoz tartozó kockázat várható értékének szórásnégyzete, a j-edik módozatban a kockázat várható értéke szórásnégyzetének általános értéke, amelynek becslése

  $$a = \frac{1}{n} \sum_{k} (m - m_k)^2, \tag{4.8}$$

  ahol m_k jelöli a k-adik módozat kockázata várható értékének becslését a saját kártapasztalata alapján
- $s^2 = E(\sigma^2(\theta_j))$: a módozatokhoz tartozó kockázat szórásnégyzetének várható értéke, a j-edik módozatban a kockázat szórásnégyzete várható értékének általános értéke, amelynek becslése

  $$s^2 = \frac{1}{n} \sum_{j} \sigma_j^2, \tag{4.9}$$

  ahol σ_j^2 jelöli a j-edik módozat kockázata szórásnégyzetének becslését a saját kártapasztalata alapján

$\mu(\theta_i)$ lineáris credibility becslése során két esetet különböztetünk meg az alapján, hogy a közös várható érték ismert, vagy pedig meg lett becsülve. Azonban mindkét esetben két feltételezéssel kell élni:

1. (B1):
   $E(X_{jt}|\theta_j = y) = \mu(y) \qquad \forall i,t$-re
   $cov(\underline{X}_j|\theta_j = y) = \sigma^2(y)I_h \qquad \forall j$-re

2. (B2):
   $(\underline{X}_1, \ldots, \underline{X}_n)$ függetlenek és azonos eloszlásúak.

A (B1), (B2) feltételek teljesülése esetén, ha a közös várható érték (m) ismert, akkor $\mu(\theta_i)$ lineáris credibility becslése, amennyiben a becslést a $T = c_0 + \sum_{k=1}^{n}\sum_{l=1}^{h} c_{kl}x_{kl}$ alakú függvények között keressük

$$\hat{\mu}(\theta_i) = z\frac{1}{h}\sum_{t=1}^{h} x_{it} + (1-z)m, \tag{4.10}$$

ahol a Bühlmann faktort a $z = \frac{na}{s^2 + na}$ formula adja meg.
A második esetben, azaz ha m nem ismert, akkor $\mu(\theta_i)$ lineáris credibility becslése, az $E(\sum_{k=1}^{n}\sum_{l=1}^{h} c_{kl}X_{kl}|\theta) = \mu(\theta)$ feltétel mellett

$$\hat{\mu}(\theta_i) = z\frac{1}{h}\sum_{t=1}^{h} x_{jt} + (1-z)M_0, \tag{4.11}$$

ahol a becslést a $T = \sum_{k=1}^{n}\sum_{l=1}^{h} c_{kl}x_{kl}$ alakú függvények között kerestük.
Az állítások bizonyításai a [4] könyv 92-95 oldalain olvashatók.

Példa:
A 4. táblázat 3 biztosítási módozat kárszámait tartalmazza 5 évre visszamenőleg és az első módozat kárszámára szeretnénk lineáris credibility becslést adni.

Év/Módozat	**1.módozat**	**2.módozat**	**3.módozat**
2005	110	75	68
2004	85	81	86
2003	117	98	86
2002	98	97	73
2001	85	83	83

4. táblázat. Példa kárszám adatokra

Jelölje $\theta_1, \theta_2, \theta_3$ módozatok rizikóparamétereit. A feladatunk $\mu(\theta_1)$ lineáris credibility becslésének megadása.

Tegyük fel, hogy a módozatok megfigyelései Poisson(λ_i) eloszlásúak $i = 1, 2, 3$, így $\mu(\theta_1) = \lambda_1$ a becsülendő paraméter. Ekkor a közös várható érték becslése, az összes kárszám átlaga $M_0 = 88,33$. A kárszámok átlaga módozatonként: $m_1 = 99$, $m_2 = 86,8$, $m_3 = 79,2$. Ez a Poisson eloszlás miatt megegyezik a módozatok tapasztalati szórásával, azaz $\sigma_1^2 = 99$, $\sigma_2^2 = 86,8$, $\sigma_3^2 = 79,2$. Így

$$a = \frac{1}{3}\sum_j (M_0 - m_j)^2 = 66,51$$

és

$$s^2 = \frac{1}{3}\sum_j \sigma_j^2 = 88,3.$$

Ezekkel már megadható a Bühlmann faktor értéke:

$$z = \frac{3a}{s^2 + 3a} = 0,69$$

Ekkor a paraméter credibility becslése:

$$\hat{\lambda}_1 = 0,69 * 99 + 0,31 * 88,33 = 95,69,$$

amely jelentős eltérést mutat a módozat saját kártapasztalatán alapuló becsléséhez képest.

4.3. Bühlmann-Straub modell területi faktorokra

A Bühlmann-Straub modell alkalmazási területe megegyezik a klasszikus modell alkalmazásával. A biztosítási matematika területén sokszor kerülhetünk szembe azzal a problémával, hogy az egyes megfigyelésekhez különböző súlyok tartoznak. Ilyen például a kárgyakoriság vizsgálata során az, hogy a megfigyelésünk hány szerződésből származik, illetve átlagkár esetén a károk száma is egy ilyen súlyt rendel a megfigyelésekhez. Ezeket a súlyokat a klasszikus Bühlmann modell nem tudta kezelni, de a Bühlmann-Straub modell ezeket is figyelembe veszi. Ezzel párhuzamosan a (B1),(B2) feltételek értelemszerű módosítására van szükség:

- (BS1):
  $E(X_{jt}|\theta_j = y) = \mu(y) \qquad \forall j, t\text{-re}$
  $cov(X_{jk}, X_{jl}|\theta_j = y) = \frac{\sigma^2(y)}{\omega_{jk}}\delta_{kl} \qquad \forall j, k, l\text{-re}$
- (BS2):
  $(\underline{X}_1, \ldots, \underline{X}_n)$ függetlenek és $\theta_1, \ldots, \theta_n$ azonos eloszlásúak.

A (BS1) feltételben ω_{jk} jelöli az j-edik módozat k-adik évi megfigyelésének súlyát.

Ez a modell nem eredeti formájában kerül ismertetésre, hanem az eredeti feladatunk, azaz a területi simítás témakörén keresztül, de semmi olyan változtatást nem hajtunk végre, amelytől a modell alkalmazhatósága romlana.

A modellezés során módozatok helyett régiók megfigyelései állnak a rendelkezésünkre. A térstatisztika alapvető feltevése, hogy a közeli régiók kockázata, a távolság valamilyen függvényében hasonlít egymásra. Ezzel a feltevéssel a (BS1) és (BS2) feltélek teljesülését garantáljuk, abban az értelemben, hogy az i-edik település kockázatának felmérése során az ő "szomszédainak" kártapasztalatát tekintjük az általános információ forrásának.

Nevezzünk szomszédosnak két települést, ha közelebb vannak egymáshoz, mint r km. A távolság ebben a szituációban nagyon érzékeny pontja lehet a modellnek. Nem célszerű légvonalban mért távolságot használni, hiszen előfordulhatnak olyan természeti akadályok (pl.: folyó, hegy), amelyek a tényleges utat jelentősen megnövelik. Ez indokolja, hogy úthálózaton alapuló távolság lenne optimális. Az r meghatározása illetve becslése statisztikai módszerek felhasználásával végezhető el, ezt egy adatspecifikus paraméternek tekintjük. Intuitív meggondolások alapján a valóságban értéke 20-30 km közé tehető. A modellt egyetlen település kockázatának felmérésére fogjuk bemutatni, ugyanis minden település esetén ugyanaz a módszer, attól eltekintve, hogy a szomszédok száma településenként változik.

A Bühlmann-Straub modell alkalmazása

Tegyük fel, hogy adott n régió, amelynek h évi megfigyelése áll a rendelkezésünkre. Az i-edik régió kockázatát szeretnénk felmérni, feltéve, hogy a többi n-1 régió az i-edik régió szomszédja. Az előző pont jelöléseinek megtartása mellett vezessük be a következő jelöléseket:

- ω_{jt}: az j-edik régió t-edik évi megfigyelésének súlya
- $\omega_j = \sum_{t=1}^{h} \omega_{jt}$: az j-edik régió összsúlya
- $z_j = \frac{a\omega_j}{s^2 + a\omega_j}$: az j-edik régió Bühlmann faktora
- $z = \sum_{j=1}^{n} z_j$
- $M_j = \sum_{t=1}^{h} \frac{\omega_{jt}}{\omega_j} x_{jt}$: az j-edik régió kockázatának becslése a saját kártapasztalata alapján

- $M_0 = \sum_{j=1}^{n} \frac{z_j}{z} M_j$: a kockázat várható értékének általános értéke az összes régió kártapasztalata alapján

Ekkor $\mu(\theta_i)$ lineáris credibility becslése:

$$\hat{\mu}(\theta_i) = z_i M_i + (1 - z_i) M_0 \tag{4.12}$$

feltéve, hogy a (BS1), (BS2) feltételek érvényesek.

A modell javítási lehetőségei

A modell két intuitív feltevésünknek nem tesz eleget. Az egyik, hogy korlátozott évre visszamenőleg használhatjuk a megfigyeléseinket, hiszen az évek során sok külső tényező megváltoztathatja a kockázati viszonyokat. Például egy autópálya megépítése jelentős mértékben csökkentő hatással van a kockázatra. Ez indokolja, hogy az idő függvényében a korábbi évek megfigyeléseit egyre kisebb súllyal szeretnénk figyelembe venni.

A másik feltételezett jelenség, amit nem tartalmaz a modell, hogy egy régió szomszédainak kockázatai nem egyformán hasonlítanak a célterület kockázatára. A feltételezések szerint a hasonlóság a távolság függvényében csökken.

Az első probléma egy lehetséges megoldása lehet, ha az ω_{jt} súlyokat önkényesen megváltoztatjuk. A változtatáshoz az idő függvényében egy gyorsan lecsengő függvényt célszerű választani. Egy lehetséges választás lehet a

$$f(\omega_{jt}) = e^{-ak}\omega_{jt} \tag{4.13}$$

új súlyok választása, ahol k jelöli, hogy az adott súly hány évvel korábbi megfigyeléshez tartozik és $a > 0$ paraméter. Jelölje $\omega_j = \sum_{t=1}^{h} f(\omega_{jt})$ az új súlyok összegét. Az ω_{jt} súlyok helyett mindenhol az $f(\omega_{jt})$ súlyokat használva elérhető, hogy a korábbi évek megfigyeléseit bizonytalanabbnak tekintsük. Így a régebbi megfigyelések kisebb súllyal kerülnek beszámításra, hiszen a Bühlmann faktor meghatározása során a megfigyelések bizonytalansága jelentős szerepet játszik

A második probléma is hasonlóan oldható meg, de ebben az esetben a régiókhoz tartozó Bühlmann faktorokat módosítjuk. A módosítást a már meghatározott Bühlmann faktorokon hajtjuk végre, így a modell alkalmazhatósága nem sérülhet. Egy lehetséges választás lehet a módosításra a

$$g(z_j) = e^{-bd_{im}} z_j, \tag{4.14}$$

ahol d_{im} jelöli az i-edik és m-edik régió távolságát. Ennek megfelelően jelölje $z = \sum_{j=1}^{n} g(z_j)$ a Bühlmann faktorok összegét. Értelemszerűen ekkor az i-edik régió kockázatának lineáris credibility becslésében lévő súlyt változatlanul hagyjuk. A súlyok megváltoztatásával csupán az általános információ (M_0) összetétele és értéke változik.

Ekkor az i-edik régió kockázatának lineáris credibility becslése

$$\hat{\mu}(\theta_i) = z_i M_i + (1 - z_i) M_0, \tag{4.15}$$

ahol

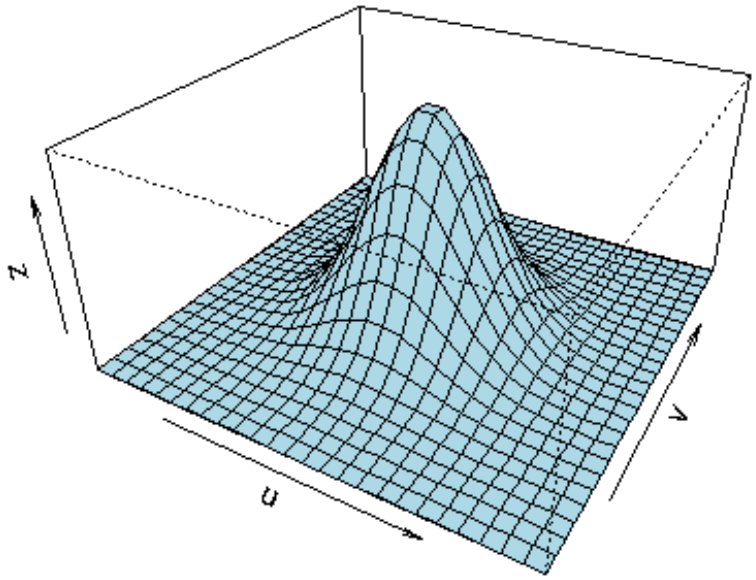
- $f(\omega_{jt})$: az j-edik régió t-edik évi megfigyeléséhez tartozó súly
- z_j : az új súlyokkal számolt Bühlmann faktor
- $M_j = \sum_{t=1}^{h} \frac{f(\omega_{jt})}{\omega_j} X_{jt}$
- $M_0 = \sum_{j=1}^{n} \frac{g(z_j)}{z} M_j$

4.4. A Bühlmann-Straub modell tesztelése

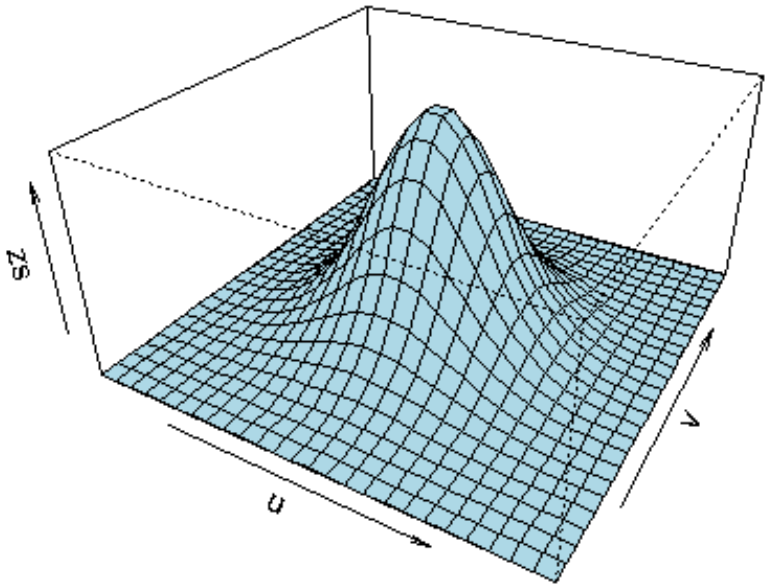
A modell képi tesztelését két generált felületen végezzük el. Az első tesztfeladat egy domb. Ezt eredeti és szórt formában is simítjuk. A második tesztfeladat egy szórt lépcső. Az eredeti felületet láthatjuk a 9. ábrán. A tesztelés ezen része arra irányul, hogy a modell mennyire rontja el a már sima felületünket. A 10. ábrán látható a simítás eredménye. A modell egy kicsit összenyomta a felületet, de nem jelentős mértékben.

Az első tesztfelületet véletlen generált számokkal megszórtuk, ez látható a 11. ábrán. Itt főként az eredeti felület alakjának visszanyerése a cél. A simított adatokat mutatja a 12. ábra. Láthatóan a szórást egészen jól kivette az adatokból és kirajzolódik az eredeti felület alakja is.

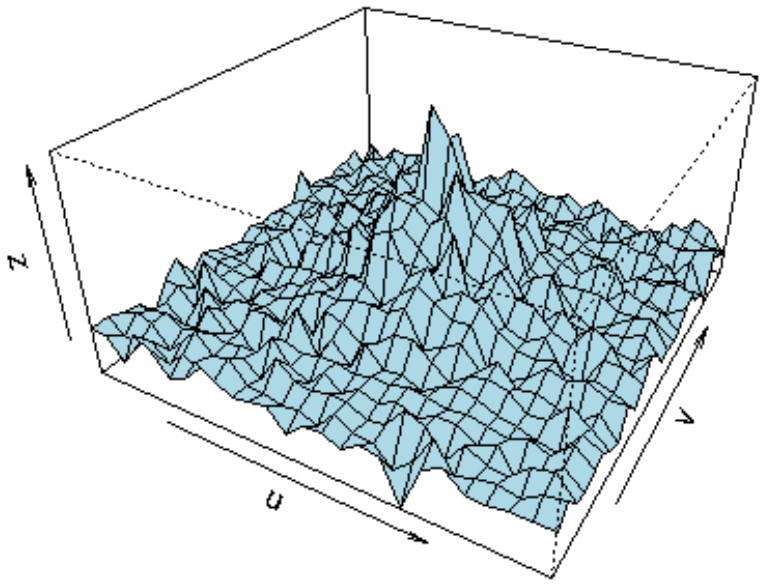
A második tesztfelület egy lépcső, amelyet már kezdetben megszórtunk véletlen számokkal (13. ábra). A kérdés, hogy a modell képes-e detektálni a törésvonalat. A simított felületen (14. ábra) érzékelhetően kirajzolódik a törésvonal és az adatok szórása is kisebb lett.



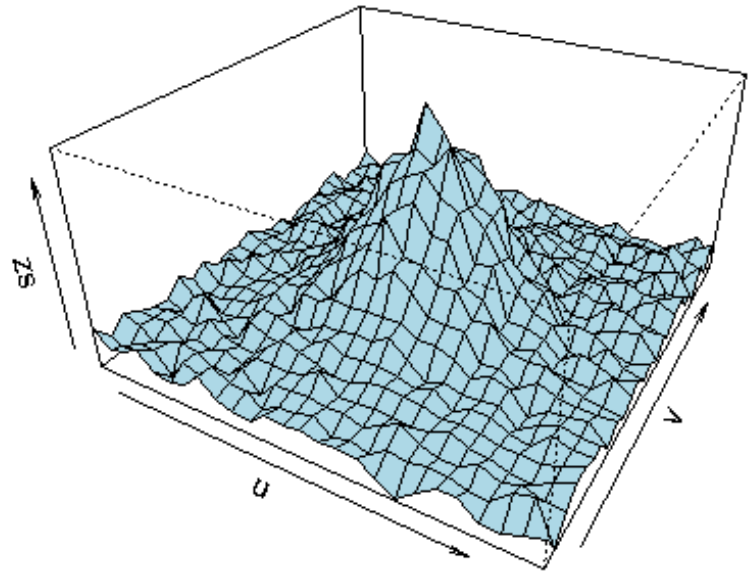
9. ábra. Credibility: eredeti felület



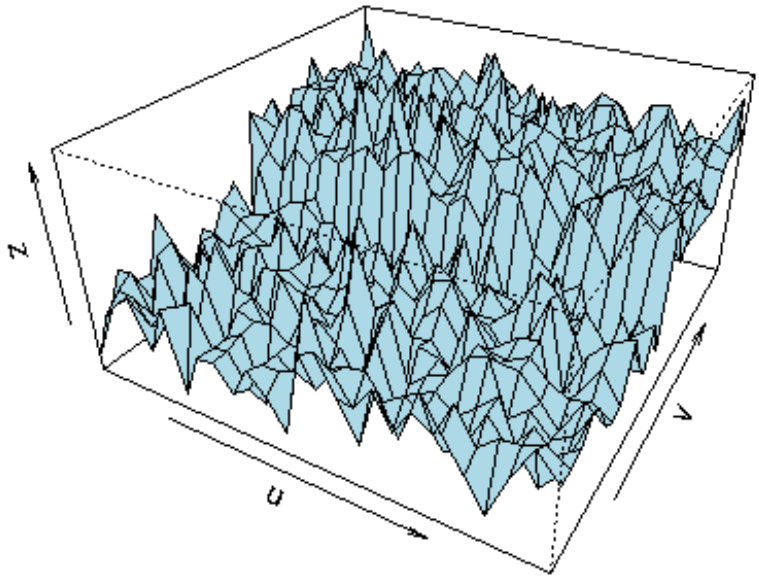
10. ábra. Credibility: eredeti felület simítása



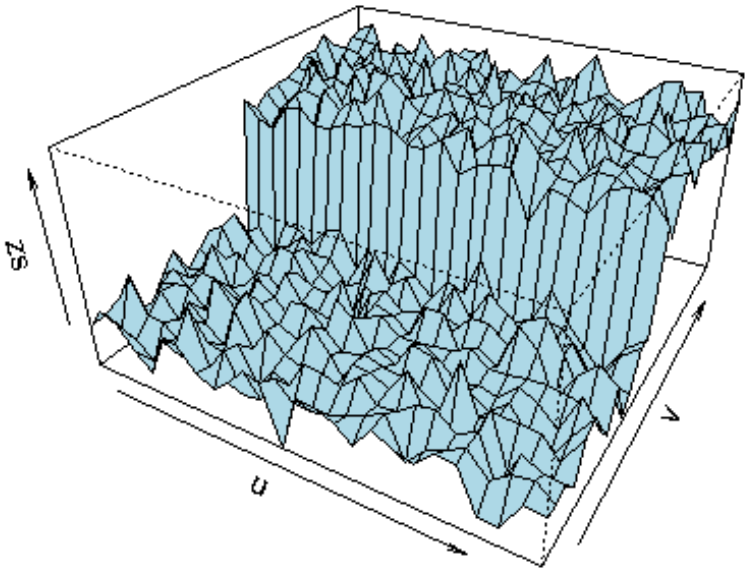
11. ábra. Credibility: első tesztfelület



12. ábra. Credibility: első tesztfelület simítása



13. ábra. Credibility: második tesztfelület



14. ábra. Credibility: második tesztfelület simítása

5. Bayes modell

A modellezés során a Bayesi statisztikát hívjuk segítségül. A credibility elmélet során már kaptunk egy kis ízelítőt a Bayesi hozzáállásból. Ez a modell tekinthető a *credibility modell* általánosításának is.

5.1. A modell bevezetése

Tegyük fel, hogy adott n darab régió (irányítószámok alapján), amelyeket díjosztályokba szeretnénk sorolni. Jelölje X_i azt a valószínűségi változót, amely az i-edik régió kockázatát reprezentálja és Y_i pedig az i-edik régióban mért káradatot (pl: károk számát). Továbbá jelölje δ_i az i-edik régió "szomszédainak" kockázatát, ez egy valószínűségi vektorváltozó, amely dimenzióját a szomszédok száma adja meg. A szomszédság itt jelentheti a tényleges szomszédokat, azaz akiknek van közös határuk, illetve egy régió szomszédainak tekinthetjük az összes olyan régiót, amely például nincs távolabb, mint 10 km. A Bayes becslések elméletébe beágyazva ez azt jelenti, hogy az $\underline{X} = (X_1, ..., X_n)$ paraméterekre szeretnénk becslést adni az $\underline{Y} = (Y_1, ..., Y_n)$ minta alapján. A modellben a kezdeti feltételezésünk az, hogy X_i eloszlása csak a szomszédos régiók kockázatától függ. Ezért az i-edik régió kockázatának feltételes sűrűségfüggvénye

$$g_i(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \tag{5.1}$$

a következő alakra egyszerűsödik:

$$g_i(x_i|\delta i) \tag{5.2}$$

A következő lépésben tekintsük a likelihood függvényt:

$$f(\underline{y}|\underline{x}) = \prod_{i=1}^{n} f(y_i|x_i) \tag{5.3}$$

Jelen helyzetben nem követünk el hibát, ha a fenti módon a likelihood függvényt szorzatra bontjuk, hiszen feltehetjük, hogy az i-edik mintaelem eloszlása csak az i-edik régió kockázatától függ. Ezután már fel tudjuk írni az a poszteriori sűrűségfüggvényt, amely Bayes tétele a alapján a következő alakú:

$$g(\underline{x}|\underline{y}) = f(\underline{y}|\underline{x})g(\underline{x}) \tag{5.4}$$

Az $\underline{X}$ Bayes becslése az a becslés, amely maximalizálja az a poszteriori sűrűséget. Itt a legnehezebb feladat természetesen az a poszteriori eloszlás meghatározása. Vizsgáljuk meg az egyenlet jobb oldalát. Tudjuk, hogy a Bayes becslések, abban különböznek a hagyományos becsléselmélettől, hogy itt a paramétereket is valószínűségi

változóknak tekintjük és ezeknek megadjuk az eloszlásukat is, amelyet a priori eloszlásnak nevezünk. Most tegyük fel, hogy rendelkezésünkre állnak a $g_i(x_i|\delta i)$ feltételes a priori sűrűségfüggvények. Ez sajnos nem azt jelenti, hogy ismerjük vagy meg tudjuk határozni a feltétel nélküli a priori eloszlást. Ugyanis ha ez $(g(\underline{x}))$ a rendelkezésünkre állna, akkor meg tudnánk határozni az a poszteriori sűrűséget, aminek a maximalizálásával adódna a kívánt becslés.

Ehelyett arra fogjuk használni ezeket a feltételes a priori sűrűségeket, hogy mintát nyerjünk az a poszteriori eloszlásból. A mintavételt sajnos nem mindig tudjuk analitikus úton elvégezni. Ez a probléma itt is jelentkezni fog, ezért szükségünk lesz egy jó mintavételezési eljárásra. Erre a célra a Gibbs módszert fogjuk használni. Ez az eljárás a Markov lánc Monte Carlo (MCMC) módszerek egyik alkalmazása. Úgy generál mintát egy tetszőleges f sűrűségfüggvényű eloszlásból, hogy algoritmikusan egy Markov láncot konstruál, amelynek létezik stacionárius eloszlása és ez éppen f lesz. Megfelelően nagy mintaelemszám után pedig egyszerűen a tapasztalati sűrűségfüggvénnyel helyettesítve az a poszteriori sűrűségfüggvényt, annak maximalizálásával kapjuk a kívánt becslést.

Foglaljuk tehát össze a modell főbb pontjait:

- adott n régió (irányítószám alapján)
  X_i: az i-edik régió kockázatát reprezentálja (valószínűségi változó)
  Y_i: az i-edik régióból származó minta
  δ_i: az i-edik régió szomszédainak kockázatát reprezentáló vektor (valószínűségi változó)
  $\underline{X} = (X_1, ..., X_n)$
  $\underline{Y} = (Y_1, ..., Y_n)$

- tegyük, fel hogy ismerjük a $g_i(x_i|\delta i)$ feltételes a priori sűrűségeket

- felírjuk az a poszteriori sűrűségfüggvényt Bayes tételéből: $g(\underline{x}|\underline{y}) = f(\underline{y}|\underline{x})g(\underline{x})$, ahol $f(\underline{y}|\underline{x})$ a likelihood függvény
  és $g(\underline{x})$ az *a priori* eloszlás

- a Gibbs módszer segítségével mintát generálunk az *a poszteriori* eloszlásból

- elég nagy számú minta esetén a tapasztalati eloszlással dolgozunk tovább

- a Bayes becslés az a becslés, amely maximalizálja az a poszteriori sűrűséget, így a kívánt becslésünk az az $\underline{x}$ lesz, amire a tapasztalati eloszlás maximális

5.2. A Gibbs módszer

A Markov Lánc Monte Carlo (MCMC) módszer széleskörű alkalmazási területtel rendelkezik. Mintavételezésre is ezen módszerek egyikét fogjuk használni, amelyet az angol szakirodalom "Gibbs sampler" néven tárgyal. Az eljárás alapgondolat az, hogy konstruálunk egy olyan Markov-láncot, amelynek stacionárius eloszlása (h), azaz éppen a kívánt eloszlás, amiből mintát szeretnénk venni. A Markov-láncot elindítva egy idő után beáll egy egyensúlyi állapotba. Ezek után az egyes mintaelemek legyenek az aktuális állapotok, amibe a Markov-lánc lép.

Tegyük fel, hogy $h(x_1, \ldots, x_n)$ többdimenziós sűrűségfüggvényből szeretnénk mintát venni és ismerjük a $h(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ feltételes sűrűségeket $\forall i$-re. Ekkor az algoritmus a következő:

1. Adjuk meg a h függvény tartójának egy pontját, mint kiindulási pontot. Jelölje ezt $x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$.

2. Amennyiben $x^{(k)}$-t már meghatároztuk, akkor x^{k+1} koordinátái legyenek:

   - $x_1^{(k+1)} = h(x_1^{(k)}|x_2^{(k)}, \ldots, x_n^{(k)})$
   - $x_2^{(k+1)} = h(x_2^{(k)}|x_1^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)})$
   - $\ldots$
   - $x_{n-1}^{(k+1)} = h(x_{n-1}^{(k)}|x_1^{(k+1)}, \ldots, x_{n-2}^{(k+1)}, x_n^{(k)})$
   - $x_n^{(k+1)} = h(x_n^{(k)}|x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)})$

Belátható, hogy az így kapott mintaelemek egy Markov-láncon való bolyongásnak felelnek meg és az így konstruált Markov-láncnak létezik stacionárius eloszlása, ami éppen h. A kapott mintaelemek első néhány (kb.:1000) elemét eldobjuk, mert a tapasztalatok szerint bonyolultabb h esetén is ennyi idő alatt beáll az egyensúlyi állapot. Az erre vonatkozó szakirodalom ezt az időt "beégetési idő"-nek nevezi. Másrészről szükség van még a minta ritkítására is. Ennek az oka, hogy a mintaelemek egy Markov-lánc lépéseinek felelnek meg, ezért nem mondhatjuk, hogy az egymást követő elemek korrelálatlanok. Ezek a problémák a következő pontban kerülnek szemléltetésre.

A Gibbs módszer vizsgálata egy példán keresztül

Az algoritmust egy példán keresztül szemléltetjük. Ezen vizsgálatot és a hozzá tartozó ábrákat az **R** program használatával valósítjuk meg. A Gibbs módszer megtalálható a program beépített függvényei között, mégpedig "**gibbs.met**" néven. Ezen elemzés során is ezt fogjuk használni.

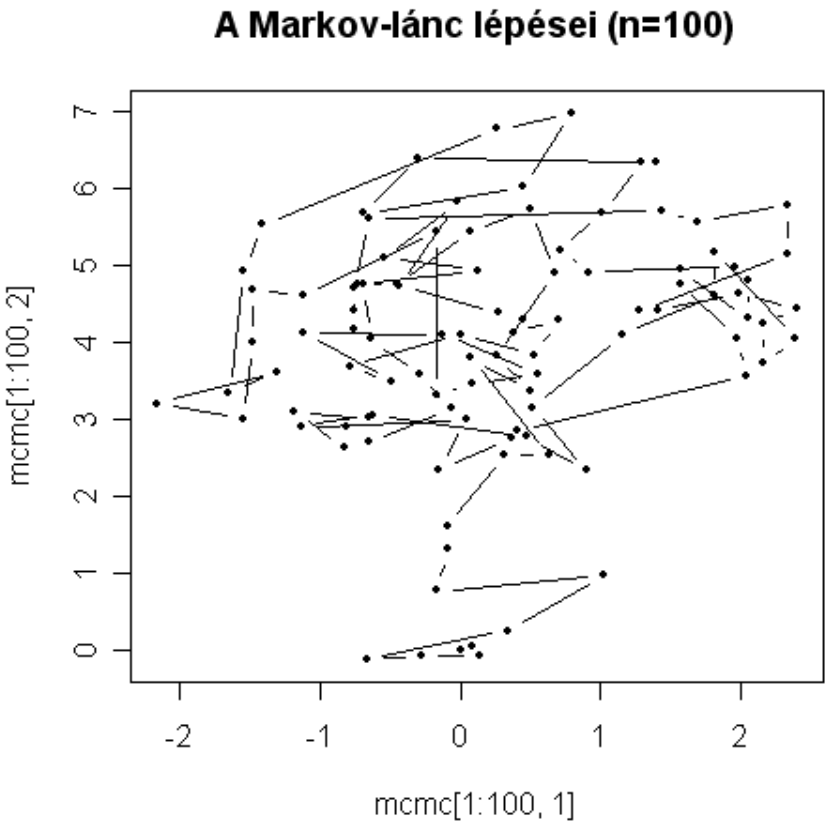
Legyen $X = (X_1, X_2)$ kétdimenziós normális eloszlású valószínűségi változó $(0, 5)$ várható értékkel és

$$M = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$$

Ekkor X sűrűségfüggvénye felírható

$$h(x) = \frac{1}{2\pi} det(M)^{-1/2} exp(\frac{1}{2}(x-m)^T M^{-1}(x-m))$$

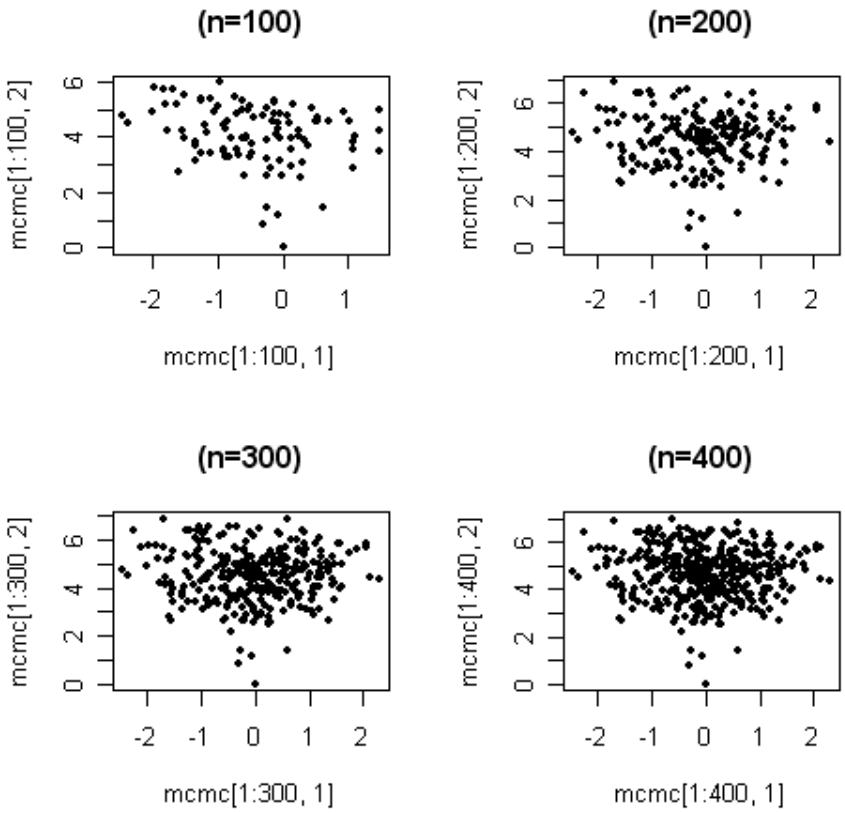
alakban. Ezen sűrűségfüggvényű eloszlásból szeretnénk mintát generálni. Nézzük meg az algoritmus által definiált bolyongás lépéseit a 15. ábrán.



15. ábra. Gibbs mintavételezés lépései

Az ábráról leolvasható, hogy a kezdeti néhány lépés után a Markov-lánc a $(0, 5)$ pont környezetében maradva bolyong, ami nem más mint az eloszlásunk várható értéke. Ez azt jelenti, hogy azt kaptuk, amit vártunk, hiszen ott található a legtöbb pont, ahol az eloszlás sűrűségfüggvénye a legnagyobb.

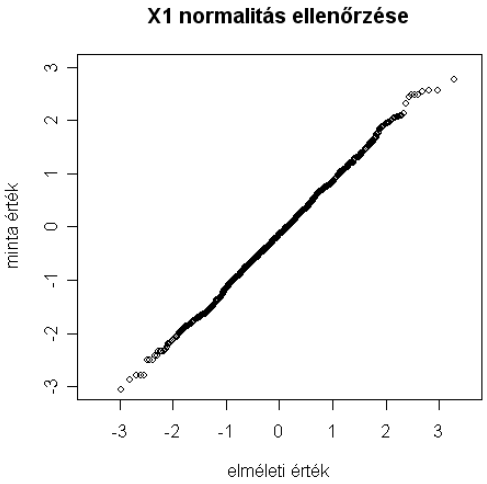
Most vizsgáljuk meg, hogy mi történik, ha növeljük a lépések számát. A 16. ábra sorozaton jól kivehetően kirajzolódik egy kétdimenziós normális eloszlás. Továbbra is a $(0, 5)$ pont köré kocentrálódnak az újonnan érkezett pontok.



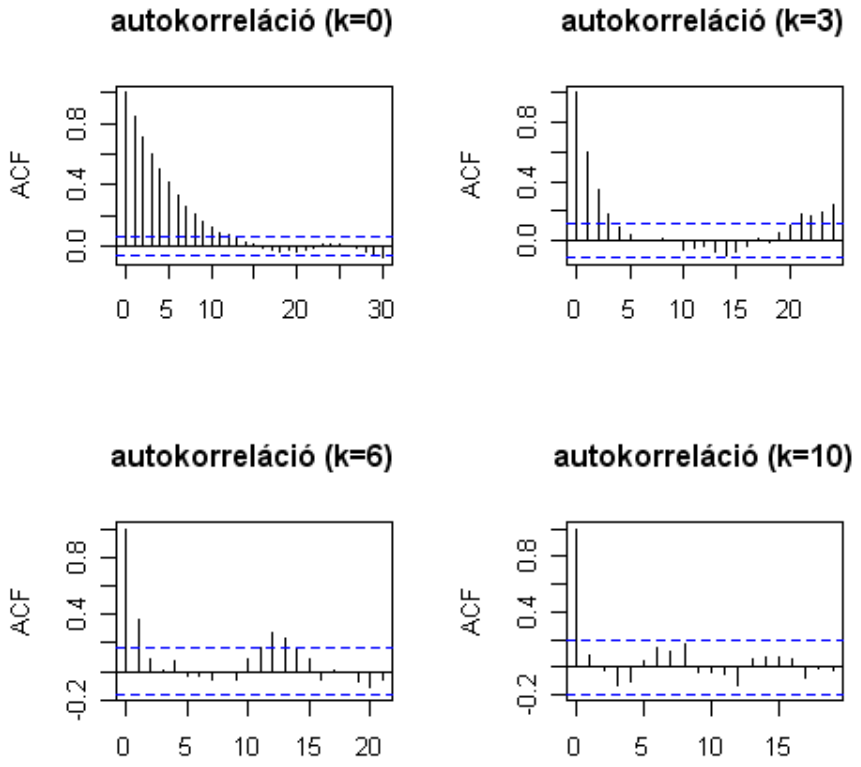
16. ábra. Gibbs mintavételezés

Most vizsgáljuk meg a minta normalitását. A normalitás vizsgálatát külön-külön végezzük el a marginálisokra. Az alábbi ábra a "**qqplot**" beépített függvénnyel készült, amely azt hivatott szemléltetni, hogy a generált mintánk mennyire tér el a normális eloszlás elméleti értékeitől. Tehát a normalitásnak az kedvez, ha az értékek minél inkább a főátló egyenesén helyezkednek el. A 17. ábráról leolvashatjuk, hogy a mintánk valóban normális az első marginális szerint. Ezen eljárás hasonló eredményre vezetett a második marginális vizsgálata során is.

A következő fontos tulajdonsága a mintának a függetlenség. A probléma abból adódik, hogy a mintánkat egy Markov-lánc lépései szolgáltatják. Ezért biztosan állíthatjuk, hogy a szomszédos mintaelemek korreláltak. Ennek kiküszöbölése érdekében a mintánkat ritkítani fogjuk és megnézzük ennek hatását a korrelációra. A ritkítás alatt azt értjük, hogy csak minden k-adik mintaelemet tarjuk meg. Itt az 18. ábrasorozat első ábrája a mintaelemek autokorrelációját mutatja a mintaelemek ritkítása előtt. Ezen vizsgálatot is csak az X_1 marginális szerint mutatjuk be, X_2 szerint hasonló eredményeket kapunk.



17. ábra. Normalitás vizsgálata



18. ábra. Ritkított minta autokorreláció függvénye

Az 18. ábráról leolvasható, hogy ritkítás előtt az egymáshoz közel eső mintaelemek erősen korrelálnak. Első két lépésben $k = 3$ illetve $k = 6$ választással ritkított-

tuk a mintánkat, ami ugyan csökkentette a korrelációt, de még mindig nem nyújtott kielégítő eredményt. Hiszen egyes értékek nagyobbak lettek mint a tűréshatár, amit a kék szaggatott vonal jelez az ábrákon. A $k = 10$ választás már megfelelőnek bizonyult. Itt minden érték a tűréshatáron belül helyezkedik el. Ekkor már nyugodtan feltehetjük, hogy az így megritkított mintaelemek nem korreláltak. Megjegyezzük, hogy az ezzel kapcsolatos szakirodalmak többsége a $k = 10$ választást már megfelelőnek tarja. Ezen kívül még szükség van az első kb. 1000 mintaelem elvetésére, ez az angol szakirodalomban "burn-in" néven található meg. Ez éppen a Markov-lánc azon része, amely még nincs megfelelően közel a stacionárius eloszláshoz, másképpen fogalmazva az az idő ameddig a Markov-láncban beáll az egyensúlyi állapot.

5.3. A feltételes a priori eloszlás formalizálása

A következő feladatunk az $X_i|\delta_i$ a priori eloszlás formalizálása. Ahhoz, hogy ezt megtehessük, térjünk vissza egy kicsit az X_i valószínűségi változóhoz. Először is tegyük fel, hogy az X_i három komponens összegére bomlik, azaz

$$X_i = t_i + u_i + v_i \tag{5.5}$$

alakú, ahol

- t_i : az i-edik régió kockázatának azon része, ami ismert vagy valahonnan már megbecsültük (pl.: GLM)
- u_i : az i-edik régió területi hatása
- v_i : az i-edik régió hibája

Tegyük fel, hogy az ismert faktorokat kiszűrték már az adatsorunkból, azaz (5.5) a következő alakra egyszerűsödik:

$$X_i = u_i + v_i. \tag{5.6}$$

Tegyük fel, hogy u_i, v_i függetlenek. Ekkor a feltételes *a priori* eloszlásuk formalizálását külön-külön is elvégezhetjük.

v_i a priori eloszlásának formalizálása

A hibatag eloszlásáról semmilyen információ nem áll a rendelkezésünkre. Ez indokolja, pontosabban nem gátolja, hogy az *a priori* eloszlását normálisnak válasszuk.

Tehát tegyük fel, hogy v_i normális eloszlású és sűrűségfüggvénye

$$g(v_i) = \lambda^{-1/2}\ exp(-\frac{v_i^2}{2\lambda}) \tag{5.7}$$

alakú, ahol λ egy hiperparaméter.

u_i feltételes a priori eloszlásának formalizálása

Kezdeti feltételezésünk szerint a területi kockázat hasonlóságot mutat a közeli régiók esetén. Ezt az információt szeretnénk beépíteni a modellbe a területi faktor *a priori* eloszlásán keresztül. A továbbiakban tekintsük az $u_{-i} = (u_1, ..., u_{i-1}, u_{i+1}, ..., u_n)$, $n-1$ dimenziós vektort. Feltesszük, hogy $u_i|u_{-i}$ eloszlásának sűrűségfüggvénye komponensekre bomlik, oly módon, hogy ezek a komponensek éppen a régiók egymásra kifejtett hatását mérik. Azaz u_i feltételes *a priori* sűrűsége:

$$g_i(u_i|u_{-i}) = exp(-\sum_{j\in\delta_i} \phi(u_i - u_j)) \tag{5.8}$$

alakú. Ahol a ϕ funkcionál a közeli régiók kockázatának hasonlóságát hivatott mérni. A modell során *M.Boskov és J.Verall - Premium Rating by Geographic Area Using Spatial Models* című cikkében ajánlott ϕ függvényt használjuk:

$$\phi(x) = \frac{x^2}{2\tau}, \tag{5.9}$$

ahol τ egy hiperparaméter. (5.8) és (5.9) alapján

$$g(u_i|u_{-i}) = \tau^{-n_i/2} exp(-\frac{1}{2\tau}\sum_{j\in\delta_i}(u_i - u_j)^2) \tag{5.10}$$

alakban írható fel, ahol n_i jelöli az i-edik régió szomszédainak számát.

(τ, λ) hiperparaméterek a priori eloszlása

A modell időközben két újabb paraméterrel bővült. Egyrészt v_i sűrűségfüggvényéből bejött paraméterként a λ, másrészt a ϕ funkcionál bevezetése során bevezettük a τ paramétert is. A feltételes a priori sűrűség meghatározásához szükségünk van még ezen paraméterek a priori eloszlásának megadására is.

Tegyük fel, hogy (λ, τ) *a priori* sűrűségfüggvénye:

$$prior(\tau, \lambda) = exp(-\frac{\varepsilon}{2\tau} - \frac{\varepsilon}{2\lambda}), \tag{5.11}$$

ahol $\varepsilon > 0$ egy kicsi konstans (pl.: $\varepsilon = 0,01$). Ezen választás helyességét [9] részletesen tárgyalja.

5.4. Az a poszteriori eloszlás formalizálása és maximalizálása

A (5.7), (5.10), (5.11) felhasználásával felírható az *a poszteriori* sűrűség a Bayes formula segítségével, amely

$$g(\underline{u}, \underline{v}, \lambda, \tau|\underline{y}) = \prod_{i=1}^{n} f(y_i|x_i) g_i(u_i|u_{-i}) g(v_i) prior(\tau, \lambda), \tag{5.12}$$

alakú. Ebbe behelyettesítve az általunk feltételezett feltételes *a priori* sűrűségeket a

$$g(\underline{u}, \underline{v}, \lambda, \tau|\underline{y}) = \tag{5.13}$$

$$= \prod_{i=1}^{n} f(y_i|x_i)\ \tau^{-n_i/2} exp(-\tfrac{1}{2\tau}\sum_{j\in\delta_i}(u_i - u_j)^2)\ \lambda^{-1/2}\ exp(-v_i^2/2\lambda)\ exp(-\frac{\epsilon}{2\tau} - \frac{\epsilon}{2\lambda})$$

formulát kapjuk.

Az *a poszteriori* sűrűség (5.13) által kijelölt eloszlásból szeretnénk mintát venni. A mintavételt a Gibbs módszer segítségével fogjuk elvégezni, amelyhez szükségünk van a marginális *a poszteriori* sűrűségek ismeretére. Ezek a marginális sűrűségek egyszerűen megkaphatóak az *a priori* marginális sűrűségek és a *likelihood* függvény segítségével. A számítás nem más, mint a marginális *a priori* és *a poszteriori* sűrűségekre felírt Bayes formula. Tekintsük át ezeket:

- $g(u_i|u_{-i}, \underline{v}, \lambda, \tau, \underline{y}) = f(y_i|x_i) g_i(u_i|u_{-i}, \underline{v}, \lambda, \tau) = f(y_i|x_i) g_i(u_i|u_{-i}, \tau)$ $(i = 1, ..., n)$
- $g(v_i|v_{-i}, \underline{u}, \lambda, \tau, \underline{y}) = f(y_i|x_i) g(v_i|v_{-i}, \underline{u}, \lambda, \tau) = f(y_i|x_i) g(v_i|v_{-i}, \lambda)$ $(i = 1, ..., n)$
- $g(\lambda|\underline{u}, \underline{v}, \underline{y}) = f(\underline{y}|\underline{x})\ prior(\lambda)$
- $g(\tau|\underline{u}, \underline{v}, \underline{y}) = f(\underline{y}|\underline{x})\ prior(\tau)$

Megjegyzzük, hogy a formulák jobb oldalán szereplő $f(y_i|x_i)$ tagokat ismertnek tekintjük. Például kárszám adatok esetén Poisson eloszlást feltételezve $c_i e^{x_i}$ várható értékkel

$$f(y_i|x_i) = \frac{exp(-c_i e^{x_i})(c_i e^{x_i})^{y_i}}{y_i!}, \tag{5.14}$$

formula adódik.

Ez azt jelenti, hogy alkalmazni tudjuk a Gibbs módszert az *a poszteriori* eloszlásból való mintavételezésre. Elegendően nagy minta esetén a paraméterek becslése a paramétertér azon pontja lesz, amelyre az *a poszteriori* tapasztalati eloszlás maximális.

6. Összefoglalás

A dolgozat a biztosítási kockázathoz tartozó területi faktorok modellezésével foglalkozott. A modellezés során minden esetben feltettük, hogy az adatokból már ki lettek szűrve a nem területi hatást reprezentáló faktorok. A faktorok szűrésére több módszert is ismertettünk, ezek közül a leggyakrabban alkalmazott az *általánosított lineáris modell*, amelyet a 2.2-es pontban részletesebben is bemutattunk. Továbbá ismertettük a területi hatás vizsgálatának nehézségeit a 2.6-os pontban.

Három modellt mutattunk be. Mindegyik esetén feltételeztük egy sima felület létezését, amely a régiók kockázatának várható értékét mutatja. Továbbá feltettük, hogy a régiókban mért adatok ezen felülethez tartozó kockázatok egy realizációja. A tapasztalatok azt mutatják, hogy a mért adatokból sokszor nem állapítható meg könnyen az eredeti felület, mert a szórásból fakadóan egy nagyon kusza felületet látunk. Az eredeti felület simasága implicit tartalmazza azt a feltételezést, hogy a közeli régiók kockázata hasonló.

Az első modell a *Whittaker kritérium* minimalizálásából adódik, amely az illesztendő felület illeszkedését és a simaságát egyszerre optimalizálva adja meg a régiók kockázata várható értékének becslését. A modell két véletlenszám generátorral előállított fiktív példán lett tesztelve. A tesztfeladatok során azt tapasztaltuk, hogy a modell nagyobb simítást eredményez a kisebb koordinátájú pontok esetén, de ennek ellenére az eredetileg feltételezett felület minden esetben kirajzolódott és a mutatott kép sokkal tisztább lett.

A második modell a *credibility elmélet*, pontosabban a Bühlmann-Straub modell alkalmazása a régiók kockázata várható értékének becslésére. A modellezés során az egyes régiókra adott becslést, a régió szomszédaival számolt lineáris credibility becslésből kaptuk. Itt két javítási lehetőséget is vázoltunk, amelyek a térbeli és időbeli távolság modellbe való beépítését tartalmazzák. Ebben az esetben a modellt szintén két próbafelületen teszteltük. A kapott eredmények az előző modellhez hasonlóan minden esetben visszaadták az eredeti felület alakját és csökkentették a közeli régiók szórását.

A harmadik modell a *Bayes statisztika* módszereivel dolgozik. A modellt csak teljes általánossággal mutattuk be. Ez a modell az *a poszteriori* eloszlás maximalizálásával szolgáltatja a régiókra vonatkozó becslést. A módszer nagyon kényelmes abból a szempontból, hogy az előismereteinket beépíthetjük a modellbe az *a priori* eloszláson keresztül. A Bayes formula alapján, az a poszteriori sűrűség megegyezik a likelihood függvény és az a priori sűrűség szorzatával. Sajnos az így kapott a poszteriori sűrűség sokszor nagyon bonyolult lesz, ezért kell egy általános módszer,

amellyel mintát tudunk venni ebből az eloszlásból. Mintavételezésre a Gibbs módszert választottuk, amely az MCMC módszerek egyik alkalmazása. Ezt követően a tapasztalati eloszlás maximalizálásával kapunk becslést a régiókban.

Ábrák jegyzéke

Irodalomjegyzék

[1] Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, Neeza Thandi: *A Practitioner's Guide to Generalized Linear Models* (2007)

[2] Greg Taylor: *Smoothness Criteria For Multi-Dimensional Whittaker Graduation* (1996) The University of Melbourne, Research Paper Number 37.

[3] Greg Taylor: *Geographic Premium Rating By Whittaker Spatial Smoothing* (2001) ASTIN BULLETIN, Vol. 31, No. 1, pp. 147-160

[4] Arató Miklós: *Nem-Élet Biztosítási Matematika* (2001) ELTE Eötvös Kiadó, Budapest

[5] M. Boskov, R.J. Verrall: *Premium Rating by Geograohic Area Using Spatial Models* (1994) ASTIN BULLETIN Vol 24 No I.

[6] Peter Beerli: *Markov chain Monte Carlo* (2005)

[7] Vitéz Ildikó Ibolya: *Térstatisztikai modellek alkalmazása a biztosításban* (2005) ELTE Alkalmazott Matematikus Diplomamunka

[8] Stoyan Gisbert, Takó Galina: *Numerikus Analízis I.*, (2005) Typotex

[9] Besag J., York J. and Mollie A.: *Bayesian Image Restoration, with Applications in Spatial Statistics*, (1991) AISM, Vol. 43, 1-59

[10] J. Van Eeghen, E.K. Greup, J.A. Nijssen: *Rate Making*, (1983) surveys of actuarial studies no. 2

[11] Dr. Katona Endre: *Automatikus térkép-interpretáció*, (2000) Szegedi Tudományegyetem, PhD értekezés