Maantieteellisen alueen huomioiminen vahinkovakuutustuotteiden hinnoittelussa

SHV-harjoitustyö (suppea)

Teija Talvensaari

5.9.2014

Sisällysluettelo

Abstract 1
1 Johdanto 2
2 Yleistetyt lineaariset mallit 2
2.1 Teoria 2
2.2. Esimerkki vahinkotiheysmallista 4
3 Spatiaalinen analyysi 8
3.1 Taustaa ja käsitteitä 8
3.2 Spatiaaliset tasoitusmenetelmät 9
3.2.1. Painotettu etäisyyteen perustuva tasoitus 9
3.2.2. Vierekkäisyyteen perustuva tasoitus 10
4 Esimerkki kotivakuutuksen aluehinnoittelusta 11
5 Yhteenveto 18
Lähteet 19

Abstract

Geographical area is considered one of the primary drivers of claims experience and is widely used rating factor in non-life insurance. Typically territories are defined as a collection of small geographical units, for example postal codes. The problem with area as a rating factor is that there are too many territory categories to directly include in the statistical model. The data in each individual territory is often sparse so it is difficult to assess risk with any certainty.

Before analyzing claims experience by area, the effect of standard rating factors has to be removed. This can be done by fitting a generalized linear model (GLM) using all rating factors other than area and then considering the residual risk, the difference between the actual experience and that predicted by the model. Spatial smoothing techniques can be used to smooth the residuals and develop categorizations of regions. The smoothing techniques are based on the assumption that geographic risk tends to be similar for the neighbouring areas. Once the categories of postal codes have been derived, the new territorial boundaries can be included in the GLM along with the other rating factors to improve the predictive power of the model. In this work the above-mentioned methods are applied to a home insurance pricing data.

At the beginning of this report basic theory of GLMs is introduced. A multiplicative Poisson model with a logarithm link function is used in modelling accidental breakage claims frequency of home contents insurance. Age of the policyholder, size of the apartment and type of contents (household or holiday home) are used as rating factors in the model.

In this report adjacency-based smoothing technique is applied to smooth the residuals. The method weights the information from one postal code with the information from adjacent codes so that immediately adjacent codes get more weight. The level of smoothing is controlled by altering the smoothing parameter. The purpose of smoothing is to remove the noise from the signal, making any remaining geographical variation clearer. Once the residuals have been smoothed, the postal codes are grouped into territories and included in the GLM along with the other factors.

Territorial ratemaking based on spatial smoothing is subjective and therefore validation is important. Results of the analyzes must be considered from several perspectives. For example out-of-sample testing can be used to examine the suitability of the model and confidence intervals to examine the reliability of the estimates. Risk areas can be visually looked at the maps.

1 Johdanto

Vakuutussopimuksella vakuutuksenottaja siirtää taloudellista riskiä vakuutuksenantajalle vakuutusmaksua vastaan. Vakuutusmaksun tulee vastata odotettavissa olevaa, keskimääräistä vahingosta aiheutuvaa tappiota lisättynä liikekulu- ym. kuormituksilla. Vakuutuksen riskimaksu saadaan jakamalla vahinkomeno vakuutusvuosilla tai vaihtoehtoisesti vahinkotiheyden ja keskivahingon tulona. Vakuutusvuodella tarkoitetaan sitä aikaa, jonka vakuutus on voimassa kalenterivuonna. Vahinkotiheys on vahinkojen lukumäärä jaettuna vakuutusvuosilla. Keskivahinko saadaan puolestaan jakamalla vahinkomeno vahinkojen lukumäärällä.

Vakuutusten hinnoittelussa tarvitaan tilastollisia menetelmiä, sillä odotettavissa olevat tappiot vaihtelevat vakuutuksittain. Vahingon sattumisen todennäköisyys ei ole kaikilla vakuutuksenottajilla sama, ja kun vahinko on sattunut, odotettavat tappiot vaihtelevat vakuutuksenottajien välillä. Hinnoittelussa käytetään yhtiön omaa historiadataa vakuutuksista ja vahingoista, ja pyritään löytämään malli, joka parhaiten kuvaa, miten vahinkomeno riippuu selittävistä muuttujista. Yleistettyjen lineaaristen mallien (generalized linear models, GLM) käyttö yleistyi 1990-luvulla, ja nykyään mallit ovat laajasti käytössä vahinkovakuutuksen hinnoittelussa useissa maissa. Aiemmin hinnoittelussa käytettiin muun muassa yksisuuntaisia analyysejä, mutta ne eivät ota huomioon tariffitekijöiden välisiä riippuvuuksia. GLM-malleilla tutkitaan, miten vahinkotiheys ja keskivahinko riippuvat eri tariffitekijöistä.

Tyypillisesti omaisuusvakuutuksissa tariffitekijöinä käytetään vakuutettavaan kohteeseen liittyviä ominaisuuksia, maantieteellistä sijaintia ja vakuutuksenottajan ikää. Esimerkiksi kotivakuutuksen hinnoittelussa usein käytettyjä tariffitekijöitä rakennuksilla ovat rakennuksen tyyppi, ikä, pinta-ala, rungon rakennusaine ja rakennuksen maantieteellinen sijainti. Muuttujista täytyy olla saatavilla luotettavaa dataa, jotta niitä voidaan käyttää tariffitekijöinä, ja muuttujien arvoja täytyy mahdollisesti luokitella.

Postinumerotasolla olevan aluetekijän käyttäminen GLM-mallissa on ongelmallista, koska yksittäisiä postinumeroita on liikaa sisällytettäväksi malliin sellaisenaan, joten aluetekijän arvot täytyy saada ryhmiteltyä. Aineistoon voidaan sovittaa ensin GLM-malli ilman aluetekijää ja sen jälkeen hyödyntää mallin antamia residuaaleja eli jäännöstermejä. Residuaaleihin sovelletaan spatiaalisia tasoitusmenetelmiä (spatial smoothing techniques), joiden avulla saadaan muodostettua alueellisia luokkia. Kun alueet on saatu luokiteltua, voidaan aineistoon sovittaa uudelleen GLM-malli, jossa on alkuperäisten selittävien muuttujien lisäksi mukana myös aluetekijä. Tässä työssä sovelletaan edellä kuvattuja menetelmiä kotivakuutuksen hinnoitteluaineistoon.

2 Yleistetyt lineaariset mallit

2.1 Teoria

Lineaarisilla malleilla estimoidaan tarkasteltavan vastemuuttujan y lineaarista riippuvuutta selittävistä muuttujista $x_j, j = 1, \dots, p$. Vastemuuttujan arvoja y_i havaintoyksiköissä $i = 1, \dots, n$ käsitellään satunnaismuuttujan Y_i realisaatioina. Klassisessa lineaarisessa mallissa Y_i on odotusarvonsa μ_i ja virhetermin ε_i summa. Mallissa oletetaan virhetermien ε_i olevan normaalisti

jakautuneita odotusarvolla 0 ja varianssilla σ^2. Vastemuuttujan Y_i odotusarvon μ_i oletetaan olevan lineaarinen funktio selittävistä muuttujista, ja lisäksi muuttujien Y_i oletetaan olevan keskenään riippumattomia ja samoin jakautuneita. Lineaarinen malli voidaan siis esittää muodossa

$$Y_i = \mu_i + \varepsilon_i, \qquad \mu_i = \mathrm{E}(Y_i) = \sum_{j=1}^{p} \beta_j x_{ij},$$

missä $\beta_1, \dots, \beta_p$ovat mallin tuntemattomia parametreja. Mallin systemaattinen osa kertoo, miten odotusarvot μ_i riippuvat selittävien muuttujien arvoista x_{ij}, ja satunnainen osa sisältää muun tiedon vastemuuttujan Y_i jakaumasta. Tavoitteena on löytää parametreille $\beta_1, \dots, \beta_p$ sellaiset arvot, jotka minimoivat virhetermien ε_i neliösumman. [5], 5

Klassinen lineaarinen malli ei sovellu vahinkovakuutuksen hinnoitteluun, sillä vahinkojakaumat ovat usein oikealle vinoja ja vastemuuttuja saa vain positiivisia arvoja. Yleistetyt lineaariset mallit (generalized linear models, GLM) ovat klassisten lineaaristen mallien laajennus. GLM-malleissa ei ole normaalijakaumaoletusta vaan vastemuuttujan Y_i jakaumat oletetaan kuuluvan samaan eksponenttiperheeseen, joka sisältää muun muassa Poisson- ja Gamma-jakaumat. GLM-malleissa systemaattinen osa on edelleen lineaarinen, mutta se ei välttämättä ole suoraan vastemuuttujan Y_i odotusarvo μ_i, vaan jokin tunnettu muunnos

$$\eta_i = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Funktiota g kutsutaan linkkifunktioksi ja sen oletetaan olevan monotoninen ja derivoituva. Lineaarinen malli on yleistetyn lineaarisen mallin erikoistapaus kun $g(\mu_i) = \mu_i$. Eräs paljon käytetty linkkifunktio on log-linkkifunktio $g(\mu_i) = \log(\mu_i)$. Tällöin

$$\mu_i = g^{-1}\big(\beta_1 x_{i1} + \cdots + \beta_p x_{ip}\big) = \exp(\beta_1 x_{i1}) \cdot \exp(\beta_2 x_{i2}) \cdots \exp(\beta_p x_{ip})$$

eli estimoidaan additiivisten vaikutusten sijaan multiplikatiivisia vaikutuksia. [5], 9–10, [1], 20

Yleistetyssä lineaarisessa mallissa satunnaismuuttujien Y_i tiheysfunktioiden oletetaan olevan eksponenttiperheen muotoa

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{\omega_i[y_i\theta_i - b(\theta_i)]}{\phi} + c(y_i, \phi_i)\right\},$$

missä $\theta_i, i = 1, \dots, n$, ovat tuntemattomia parametreja, hajontaparametri ϕ voi olla tunnettu tai tuntematon, sama kaikilla i, $\omega_i, i = 1, \dots n$, ovat havaintoyksiköihin liittyviä tunnettuja prioripainoja, $\phi_i = \phi/\omega_i$ ja b ja c ovat tunnettuja funktioita, samat kaikilla i. Prioripainoilla ω_i saadaan malliin sisällytettyä tietoa kunkin havainnon uskottavuudesta. Vastemuuttujan Y_i varianssin ei yleistetyssä lineaarisessa mallissa tarvitse olla vakiofunktio, vaan se on muotoa

$$\mathrm{var}(Y_i) = \frac{\phi b''(\theta_i)}{\omega_i}$$

ja odotusarvo on $\mu_i = b'(\theta_i)$. [5], 14–16, [1], 14, 17

Jos Y_i noudattaa Poisson-jakaumaa, niin sen tiheysfunktio on

$$f_{Y_i}(y_i; \mu_i) = \frac{e^{-\mu_i}{\mu_i}^{y_i}}{y_i!}, \quad y_i = 0,1,2, \dots; \ \mu_i > 0.$$

Jos Y_i noudattaa Gamma-jakaumaa, niin sen tiheysfunktio on

$$f_{Y_i}(y_i; \lambda_i, \nu) = \frac{\lambda_i^\nu}{\Gamma(\nu)} y_i^{\nu-1} e^{-\lambda_i y_i}, \quad y_i > 0; \lambda_i > 0, \nu > 0.$$

Tyypillisesti vahinkojen lukumäärien tai vahinkotiheyksien mallintamisessa käytetään multiplikatiivista Poisson-mallia ja keskivahinkoa mallinnetaan puolestaan multiplikatiivisella Gamma-mallilla. Molemmissa malleissa linkkifunktiona käytetään log-linkkifunktiota. Prioripainona vahinkotiheysmallissa on vakuutusvuodet ja keskivahinkomallissa puolestaan vahinkojen lukumäärä. Kun sovellettava malli on määritelty, tuntemattomien parametrien $\beta_1, \dots, \beta_p$ estimaatit johdetaan suurimman uskottavuuden menetelmällä maksimoimalla log-uskottavuusfunktio

$$l(\beta; y) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \beta, \phi).$$

[5], 16, [1], 22–24, [3], 67–68

GLM-malliin liittyy paljon erilaista diagnostiikkaa mallin sopivuuden tarkastelemiseksi. Voidaan tutkia muun muassa mallin jäännösten vaihtelua, selittävien muuttujien kykyä selittää vastemuuttujan vaihtelua ja vertailla kilpailevia malleja. Jokainen malliin sisällytetty selittävä tekijä yleensä parantaa mallin sopivuutta, mutta tarpeettomien muuttujien lisääminen malliin huonontaa estimaattien tarkkuutta. On olemassa useita kriteereitä, joiden avulla voidaan vertailla kilpailevia malleja ottamalla samalla huomioon parametrien lukumäärä. Tällaisia ovat Akaiken informaatiokriteeri (Akaike's Information Criterion) AIC ja Bayesin informaatiokriteeri (Bayesian Information Criterion) BIC, jotka ovat muotoa

$$\text{AIC} = -2l + 2p, \quad \text{BIC} = -2l + p \ln n,$$

missä l on mallin log-uskottavuusfunktio. Mallia, jolla on pienin AIC tai BIC, pidetään parhaana [3], 62–63.

GLM-mallin devianssi määritellään kaavalla

$$D(y; \hat{\mu}) = 2\phi[l(y; y) - l(\hat{\mu}; y)],$$

missä $l(y; y)$ on saturoidun eli täydellisesti aineistoon sopivan mallin uskottavuusfunktio. Devianssia voidaan hyödyntää mallin selittävien muuttujien merkitsevyyden tarkastelussa uskottavuusosamäärätestin muodossa, jolloin testataan nollahypoteesia $H_0: \beta_{k_1} = \beta_{k_2} = \cdots = \beta_{k_{p-q}} = 0$ annetuilla $k_1, k_2, \dots, k_{p-q} \in \{1, \dots, p\}$. Testisuure on muotoa

$$\frac{D(y; \hat{\mu}_0) - D(y; \hat{\mu})}{\hat{\phi}(p-q)}.$$

Jos hajontaparametri ϕ on tunnettu, niin testisuure noudattaa asymptoottisesti χ^2-jakaumaa parametrein $p - q$. Jos hajontaparametri on estimoitava, noudattaa testisuure asymptoottisesti F-jakaumaa parametrein $p - q, n - p$. [5], 29–32, [3], 71–75

2.2. Esimerkki vahinkotiheysmallista

Tariffitekijöiden vaikutus on usein erilaista erityyppisissä vahingoissa, joten eri vahinkotyypit kannattaa analysoida erillisissä malleissa. Kotivakuutuksessa eri vahinkotyyppejä ovat esimerkiksi palo-, varkaus- ja vuotovahingot.

Yleistettyä lineaarista mallia varten tarvitaan yhdistetty vakuutus- ja vahinkoaineisto. Tässä työssä käytetään kotivakuutuksen irtaimiston rikkoutumisvahinkoaineistoa. Aineistoon sovitetaan vahinkotiheydelle multiplikatiivinen Poisson-malli linkkifunktiona log-linkkifunktio käyttäen Towers Watsonin Emblem – hinnoitteluohjelmistoa [11]. Tariffitekijöinä mallissa ovat kohteen pinta-ala, vakuutuksenottajan ikä ja kohteen tyyppi eli onko kyseessä koti-irtaimisto vai vapaa-ajan asunnon irtaimisto. Pinta-ala ja vakuutuksenottajan ikä on luokiteltu sopiviin luokkiin.

Taulukossa 1 on mallissa käytetyt tariffitekijät ja niitä vastaavat parametrit. Mallissa on mukana vakiotermi eli malli on muotoa

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Jokaisen tariffitekijän yhdeltä luokalta puuttuu parametri, jotta malli on yksikäsitteisesti määritelty. Näitä tariffitekijöiden luokkia kutsutaan niin sanotuiksi perustasoiksi.

Vakuutuksenottajan ikä (vuosina)

Arvo	Parametri
Alle 30	$\beta_{1,1}$
30-32	$\beta_{1,2}$
33-35	$\beta_{1,3}$
36-38	$\beta_{1,4}$
39-41	$\beta_{1,5}$
42-44	$\beta_{1,6}$
45-47	$\beta_{1,7}$
48-50	$\beta_{1,8}$
51-53	$\beta_{1,9}$
54-56	$\beta_{1,10}$
57-59	$\beta_{1,11}$
60-62	
63-65	$\beta_{1,12}$
66-68	$\beta_{1,13}$
69-71	$\beta_{1,14}$
72-74	$\beta_{1,15}$
75-77	$\beta_{1,16}$
78-80	$\beta_{1,17}$
Yli 80	$\beta_{1,18}$

Pinta-ala (m^2)

Arvo	Parametri
Alle 30	$\beta_{2,1}$
30-39	$\beta_{2,2}$
40-49	$\beta_{2,3}$
50-59	
60-69	$\beta_{2,4}$
70-79	$\beta_{2,5}$
80-89	$\beta_{2,6}$
90-99	$\beta_{2,7}$
100-109	$\beta_{2,8}$
110-119	$\beta_{2,9}$
120-129	$\beta_{2,10}$
130-139	$\beta_{2,11}$
140-149	$\beta_{2,12}$
150-159	$\beta_{2,13}$
160-169	$\beta_{2,14}$
170-179	$\beta_{2,15}$
180-189	$\beta_{2,16}$
190-199	$\beta_{2,17}$
Yli 199	$\beta_{2,18}$

Kohdetyyppi

Arvo	Parametri
Koti-irtaimisto	
Vapaa-ajan asunnon irtaimisto	$\beta_{3,1}$

Vakiotermi	β_0

Taulukko 1. Vahinkotiheysmallin tariffitekijät ja parametrit.

Mallin antamat tulokset on esitetty taulukossa 2. Parametrien estimaateista on otettu eksponentti, jolloin tulokset voidaan esittää kertoimina. Tariffitekijöiden luokilla, joille ei estimoitu parametria, kerroin on 1. Vakiotermi antaa mallin ennustaman vahinkotiheyden sellaiselle vakuutukselle, jonka tariffitekijöiden arvot ovat kaikki perustasolla. Näin ollen malli ennustaa vahinkotiheyden olevan 1,96 prosenttia, kun vakuutuksenottaja on 60–62-vuotias, asunnon pinta-ala on 50–59 m^2 ja kyseessä on koti-irtaimisto. Jos vakuutuksenottaja on puolestaan 30–32-vuotias ja kyseessä on 60–69 m^2 koti-irtaimisto, niin vahinkotiheys saadaan kertomalla vakiotermiä ao. tariffitekijöiden luokkien kertoimilla, eli $1{,}96\ \% \cdot 2{,}1521 \cdot 1{,}0721 \cdot 1 = 4{,}52\ \%$. Vakiotermi ei siis ole keskimääräinen

vahinkotiheys, sillä sen arvo riippuu täysin siitä, mitkä tariffitekijöiden luokat valitaan mallissa perustasoiksi.

Vakuutuksenottajan ikä (vuosina)

Arvo	Kerroin
Alle 30	2,2126
30-32	2,1521
33-35	2,1366
36-38	1,9582
39-41	2,0119
42-44	1,9663
45-47	1,8981
48-50	1,7784
51-53	1,5052
54-56	1,2803
57-59	1,0857
60-62	1,0000
63-65	0,9663
66-68	0,8435
69-71	0,8599
72-74	0,7550
75-77	0,6220
78-80	0,4586
Yli 80	0,4278

Pinta-ala (m^2)

Arvo	Kerroin
Alle 30	0,6203
30-39	0,5435
40-49	0,8418
50-59	1,0000
60-69	1,0721
70-79	1,3995
80-89	1,4241
90-99	1,6841
100-109	1,5594
110-119	1,6190
120-129	1,7631
130-139	1,7141
140-149	1,7734
150-159	1,9121
160-169	1,9579
170-179	2,0671
180-189	2,1068
190-199	2,3445
Yli 199	2,2129

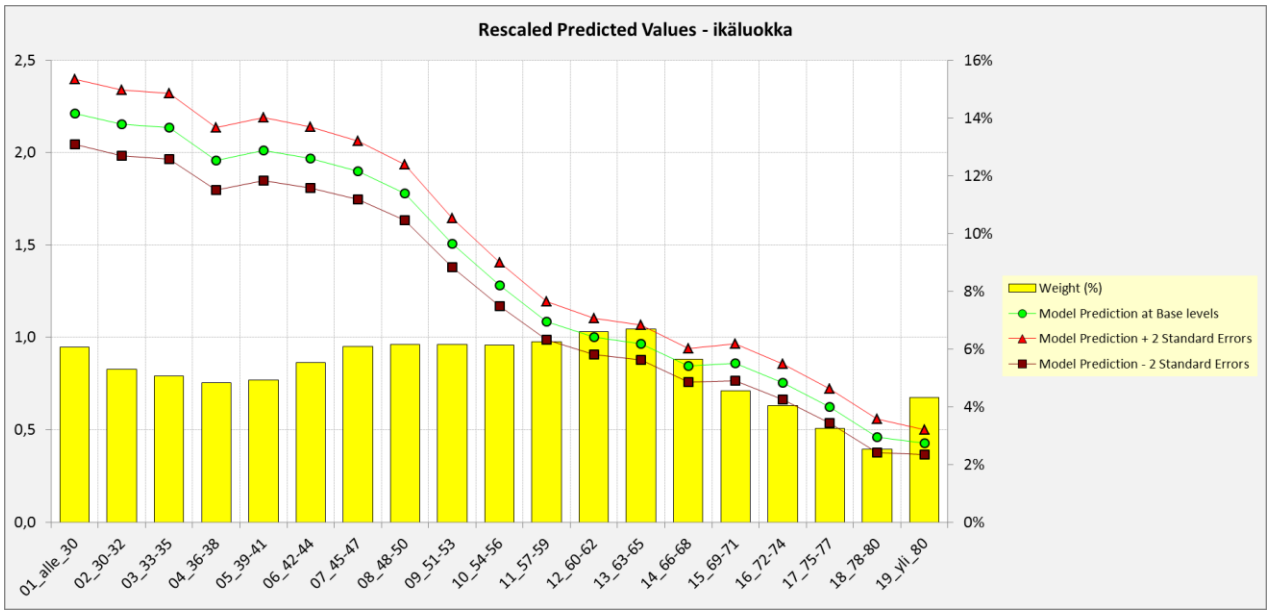
Kohdetyyppi

Arvo	Kerroin
Koti-irtaimisto	1,0000
Vapaa-ajan asunnon irtaimisto	0,1531

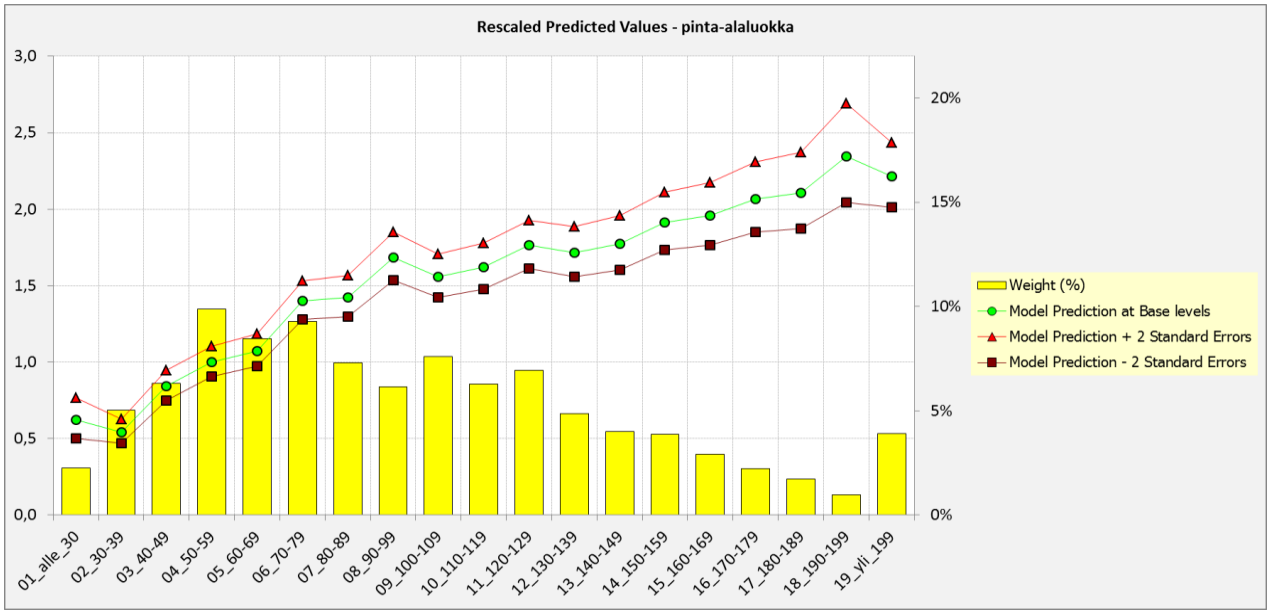
Vakiotermi	0,0196

Taulukko 2. Vahinkotiheysmallin antamat kertoimet.

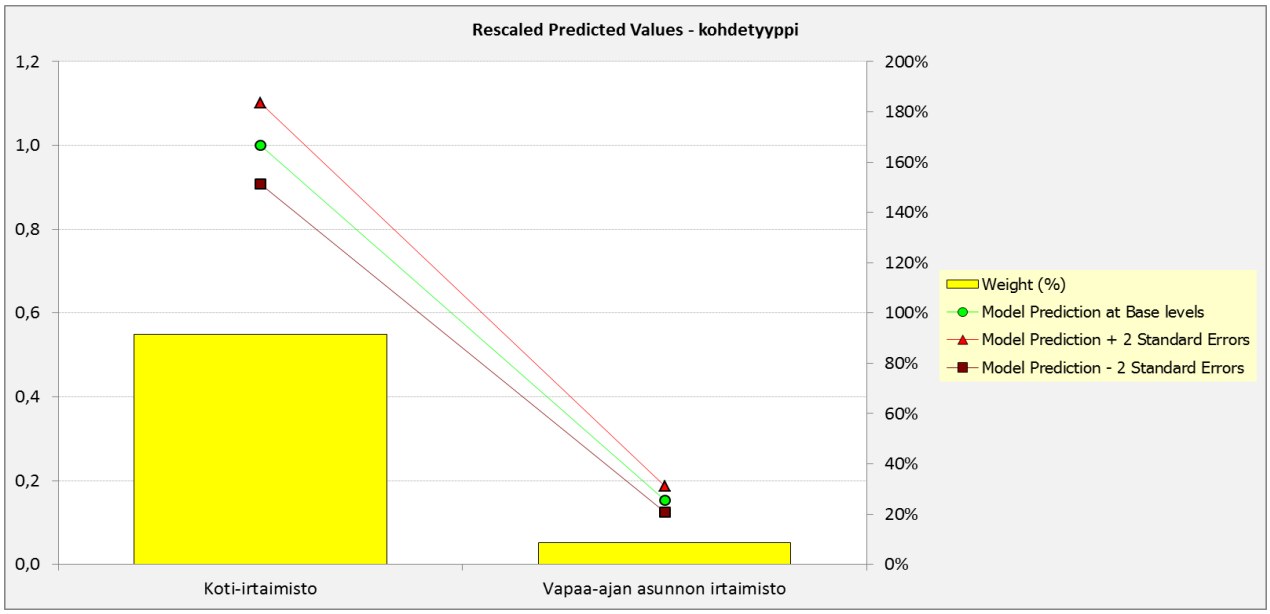
Mallin antamat kertoimet kullekin tariffitekijälle on esitetty myös kuvissa 1, 2 ja 3. Kuvissa on kertoimien lisäksi näiden keskihajonnat kahdella kerrottuna sekä pylväinä vakuutusvuosien jakautuminen aineistossa tariffitekijöiden eri luokille. Keskihajonta kuvaa estimaatin epävarmuutta ja keskihajontakäyrien rajaamaa aluetta kutsutaan luottamusväliksi. Karkeasti voidaan sanoa, että aineiston perusteella kunkin tariffitekijäluokan kerroin on luottamusvälin sisällä noin 95 prosentin varmuudella. Kuvien avulla saa käsitystä siitä, kuinka merkitsevä kyseinen tariffitekijä on mallissa. Jos keskihajontakäyrät ovat kaukana kerroinkäyrästä, toisin sanoen luottamusväli on leveä, niin estimaattiin liittyy paljon epävarmuutta. Kuvista 1, 2 ja 3 nähdään, että kaikki mallissa käytetyt tariffitekijät ovat tilastollisesti merkitseviä.



Kuva 1. GLM-malli vahinkotiheydelle, vakuutuksenottajan iän vaikutus.



Kuva 2. GLM-malli vahinkotiheydelle, pinta-alan vaikutus.



Kuva 3. GLM-malli vahinkotiheydelle, kohdetyypin vaikutus.

3 Spatiaalinen analyysi

3.1 Taustaa ja käsitteitä

Maantieteellistä sijaintia hyödynnetään nykyään vahinkovakuutuksen hinnoittelussa, vaikkakin aluehinnoittelu voi erota merkittävästi yhtiöiden välillä. Ongelmana aluehinnoittelussa on yleensä se, että yhtiöllä on yhdeltä alueelta usein liian vähän dataa, jotta alueen riskiä voitaisiin mallintaa luotettavasti. Spatiaalisen tilastotieteen menetelmiä käytetään paikkatiedon analysoinnissa silloin, kun havaintojen sijainnilla on aineistossa erityismerkitys. [13], 188

Jos havaintoihin sisältyy mittausvirhettä tai satunnaisvaihtelua, sen suodattamisessa voidaan käyttää hyväksi myös muiden kuin kohdealueen havaintoja. Tätä kutsutaan spatiaaliseksi tasoittamiseksi. Spatiaaliset tasoitusmenetelmät perustuvat oletukseen, että jos luotettavaa tietoa ei ole käytössä, voidaan vierekkäisillä alueilla olettaa olevan samanlaiset riskitasot. Tämä oletus ei tietenkään välttämättä toteudu aina käytännössä, sillä vierekkäiset alueet voivat joskus olla luonteeltaan hyvinkin erilaisia. Tasoitusmenetelmiä hyödyntämällä saadaan muodostettua alueista luokkia, jotka ennustavat paremmin riskiä kuin ilman tasoitusta tehdyt analyysit. Analyyseissa voidaan käyttää yhtiön oman aineiston lisäksi ulkoista dataa, kuten väkilukua tai asukastiheyttä, jotka usein ennustavat myös melko hyvin alueiden riskiä. [10], [13], 188–189

Oletetaan, että tarkastellaan h maantieteellistä havaintoyksikköä eli aluetta, joihin viitataan indekseillä $k = 1,2, \dots, h$. Alueilla voi olla jollakin tavalla määritellyt keskipisteet $s_k = (s_{k1}, s_{k2})$, missä s_{k1} ja s_{k2} ovat reaalilukuja. Alueittain havaittavien muuttujien Y_k ja Y_l oletetaan olevan keskimäärin sitä samankaltaisempia, mitä lähempänä toisiaan alueet k ja l sijaitsevat. Tätä kutsutaan spatiaaliseksi autokorrelaatioksi. Naapuruus määritellään yleensä niin, että alueet k ja l ovat naapureita, jos niillä on yhteinen raja. Toinen tapa määritellä naapuruus on, että alueet k ja l ovat naapureita, jos keskipisteiden välinen etäisyys $d_{kl} = \|s_k - s_l\|$ on alle jonkin annetun rajan d_0. [6], 4–5, 11–12

Aluehinnoittelussa maantieteellisenä yksikkönä voidaan käyttää esimerkiksi postinumeroa, kuntaa, lääniä tai väestökeskittymää. Kun valitaan käytettävää yksikköä, täytyy huomioida monia asioita. Yksikön täytyy olla riittävän homogeeninen alueellisten eroavaisuuksien suhteen, mutta ei kuitenkaan liian tarkalla tasolla, jotta havaintoja on riittävästi luotettavien tulosten saamiseksi. Vahinkoaineisto pitää helposti pystyä jakamaan yksikkötasolle ja myös mahdollinen ulkoinen data pitää pystyä yhdistämään tällä tasolla. Lisäksi käytettävän yksikön tulisi olla helposti ymmärrettävä eikä se saisi muuttua ajan myötä. [13], 189

Alueellisten yksiköiden välillä on vaihtelua havaitussa vahinkokehityksessä, ja vaihtelusta voidaan tunnistaa niin sanottu varsinainen signaali (signal) ja satunnaiskohina (random noise). Signaali voidaan edelleen jakaa maantieteellisiin piirteisiin ja muista tekijöistä johtuviin piirteisiin. Havaitut vahinkotiheydet ja keskivahingot perustuvat juuri näihin vaihteluihin; niissä on aluetekijästä riippumattomia ja riippuvia vaikutuksia sekä jäännösvaikutuksia. [13], 189–190

Havaitusta datasta täytyy poistaa aluetekijästä riippumattomat vaikutukset ennen kuin voidaan estimoida kuhunkin maantieteelliseen yksikköön liittyvää riskiä. Tällöin aineistoon sovitetaan ensin GLM-malli, jossa on mukana kaikki muut tariffitekijät paitsi aluetekijä. Tämän jälkeen tarkastellaan alueittain niin sanottua jäännösriskiä eli havaitun ja mallin antaman ennusteen erotusta, josta pyritään poistamaan satunnaiskohina niin, että jäljelle jää vain maantieteellisestä sijainnista johtuvaa systemaattista vaihtelua. Jäännöstermien eli residuaalien tasoittamiseen voidaan käyttää useita eri tasoitusmenetelmiä. [13], 190

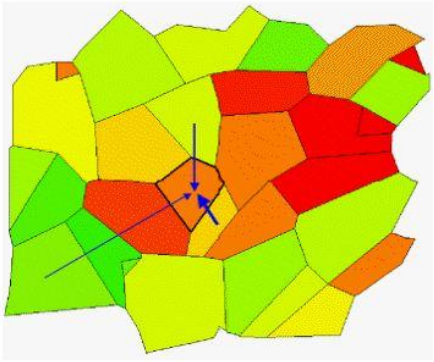
3.2 Spatiaaliset tasoitusmenetelmät

Maantieteellisen riskin oletetaan siis olevan samanlaista lähekkäin sijaitseville alueille. Tämän ansiosta voidaan yksittäisen havainnon estimaattia parantaa hyödyntämällä lähistöllä olevia havaintoja. Spatiaalisen tasoittamisen menetelmiä on olemassa useita. Tasoitukseen voidaan käyttää muun muassa splini-funktioita [8] tai Whittaker-menetelmää [9].

Seuraavissa luvuissa esitellään kaksi spatiaalista tasoitusmenetelmää, niin sanottu painotettu etäisyyteen perustuva menetelmä (weighted-distance smoothing) ja vierekkäisyyteen perustuva menetelmä (adjacency-based smoothing). Vierekkäisyyteen perustuva menetelmä pohjautuu bayesilaiseen todennäköisyysteoriaan. Etäisyyteen perustuvassa menetelmässä puolestaan hyödynnetään niin sanottua kredibiliteettiteoriaa, joka on alun perin aktuaarien kehittämä menetelmä riskimaksun laskentaan. [10], [13], 190

3.2.1. Painotettu etäisyyteen perustuva tasoitus

Etäisyyteen perustuvassa tasoitusmenetelmässä maantieteellisen havaintoyksikön informaatiota painotetaan kaikkien lähistöllä olevien havaintoyksiköiden informaatiolla perustuen yksiköiden väliseen etäisyyteen. Tasoituksen tulos on painotettu keskiarvo yksiköstä ja sen ympäröivistä yksiköistä siten, että painotus vähenee sitä mukaa kun etäisyys yksikköön kasvaa (kuva 4). Etäisyyteen perustuvan menetelmän heikkoutena on se, että menetelmässä oletetaan tietyllä välimatkalla, esimerkiksi kilometrillä, olevan samanlainen vaikutus riskin samankaltaisuuteen riippumatta siitä, onko kyseessä kaupunkialue vai maaseutu. Lisäksi etäisyyttä määriteltäessä ei oteta huomioon yksiköiden välisiä rajoja kuten jokia ja valtateitä. Etäisyyteen perustuva menetelmä soveltuu parhaiten tilanteisiin, joissa ollaan kiinnostuneita absoluuttisesta etäisyydestä, esimerkiksi tutkittaessa säähän liittyviä vahinkoja. [10], [13], 190



Kuva 4. Painotettu etäisyyteen perustuva tasoitusmenetelmä. [10]

Painotetun etäisyyteen perustuvan tasoitusmenetelmän antama tulos R_k havaintoyksikölle k on muotoa

$$R_k = Z_k r_k + (1 - Z_k)\bar{r}_k.$$

Tässä Z_k on havaintoyksikön k niin sanottu kredibiliteetti-muuttuja ja on muotoa

$$Z_k = \left(\frac{w_k}{w_k + w_0}\right)^{p^*},$$

missä w_0 on offset-paino, w_k on havaintoyksikön k kredibiliteetti-paino ja p^* on kredibiliteetti-potenssi. Muuttujan R_k kaavassa esiintyvä $\bar{r}_k$ on etäisyydellä painotettu keskiarvo muiden havaintoyksiköiden tasoittamattomista arvoista ja on muotoa

$$\bar{r}_k = \frac{\sum_{l \neq k} r_l \cdot e^{-d_{kl}P} w_l}{\sum_{l \neq k} e^{-d_{kl}P} w_l},$$

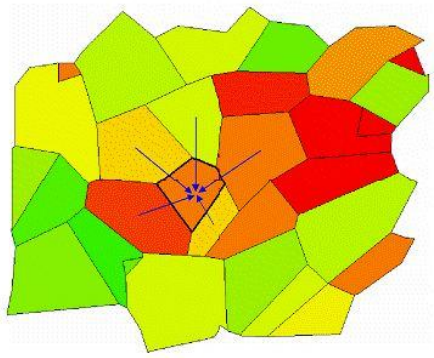
missä P on etäisyyden potenssi ja havaintoyksiköiden k ja l välinen etäisyys d_{kl} määritellään kaavalla

$$d_{kl} = \sqrt{(s_{k1} - s_{l1})^2 + (s_{k2} - s_{l2})^2},$$

missä (s_{k1}, s_{k2}) ja (s_{l1}, s_{l2}) ovat havaintoyksiköiden k ja l keskipisteitä. Kun parametreja w_0 ja p^* kasvatetaan, annetaan tasoituksessa vähemmän painoa havaintoyksikön omalle informaatiolle. Kun parametria P kasvatetaan, annetaan suhteessa enemmän painoa havaintoyksikön lähellä kuin kauempana oleville yksiköille. [10]

3.2.2. Vierekkäisyyteen perustuva tasoitus

Vierekkäisyyteen perustuvassa menetelmässä maantieteellisen havaintoyksikön oletetaan käyttäytyvän samoin kuin sen naapuriyksiköt. Yksikön informaatiota painotetaan naapurihavaintoyksiköiden informaatiolla, joten yksikköön vaikuttaa siten välittömässä läheisyydessä olevat yksiköt, mutta myös muut ympärillä olevat yksiköt naapuriyksiköidensä kautta (kuva 5). Vierekkäisyyteen perustuva menetelmä ottaa paremmin huomioon luonnonmukaiset ja keinotekoiset rajat sekä kaupunki- ja maaseutualueet kuin etäisyyteen perustuva menetelmä, joten se soveltuu paremmin esimerkiksi varkausvahinkojen mallintamiseen. [10], [13], 190



Kuva 5. Vierekkäisyyteen perustuva tasoitusmenetelmä. [10]

Vierekkäisyyteen perustuva menetelmä yhdistää havaintoyksikön informaation naapuriyksiköiden informaatioon käyttäen bayesilaista todennäköisyysteoriaa. Menetelmä olettaa, että tasoitettava tunnusluku on Poisson-jakaumasta vahinkotiheysmallissa ja vastaavasti Gamma-jakaumasta keskivahinkomallissa. Vahinkotiheysmallissa vahinkojen lukumäärän uskottavuusfunktio N_k havaintoyksikölle k on muotoa

$$N_k \sim \text{Poisson}\left(\text{vakuutusvuodet} \cdot e^{(m+b_k)}\right),$$

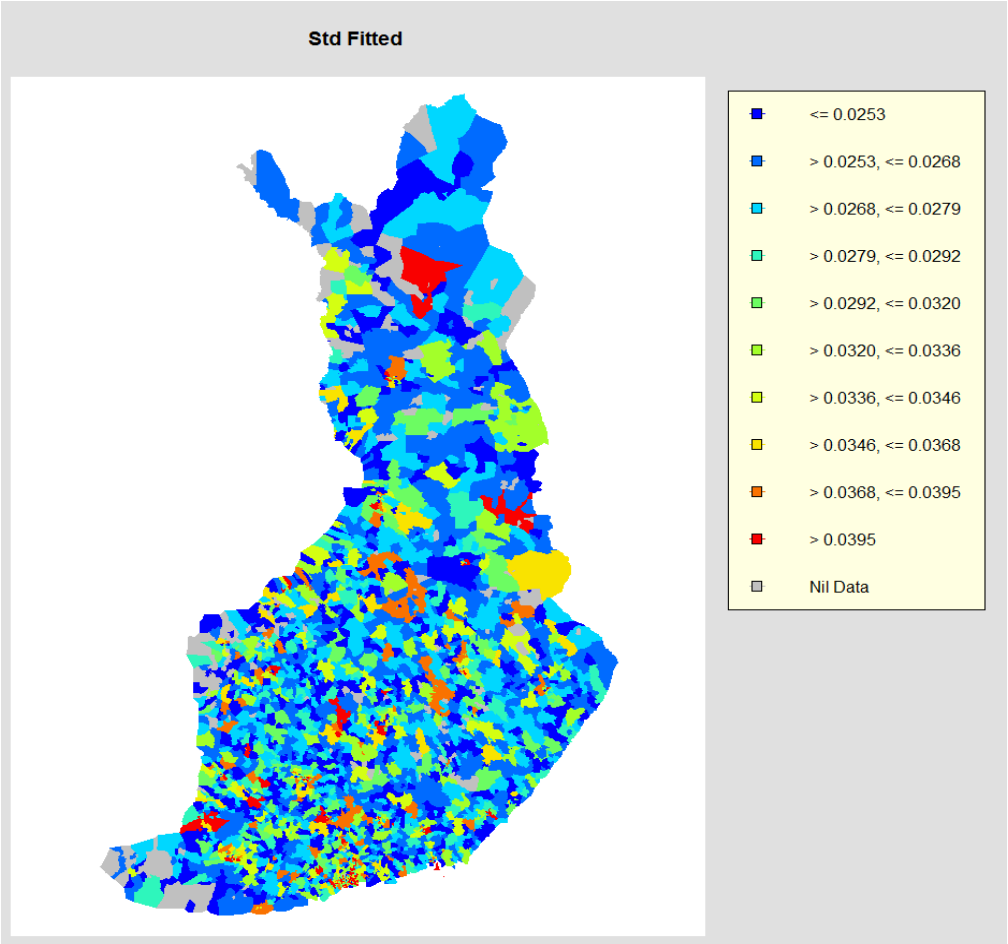
missä m on GLM-mallista saatava vakiotermin estimaatti ja b_k on havaintoyksikköön k liittyvä aluevaikutus. Kullekin vaikutukselle b_k saadaan estimaatti posteriori-jakaumasta, kun priori-jakauma on normaalijakauma

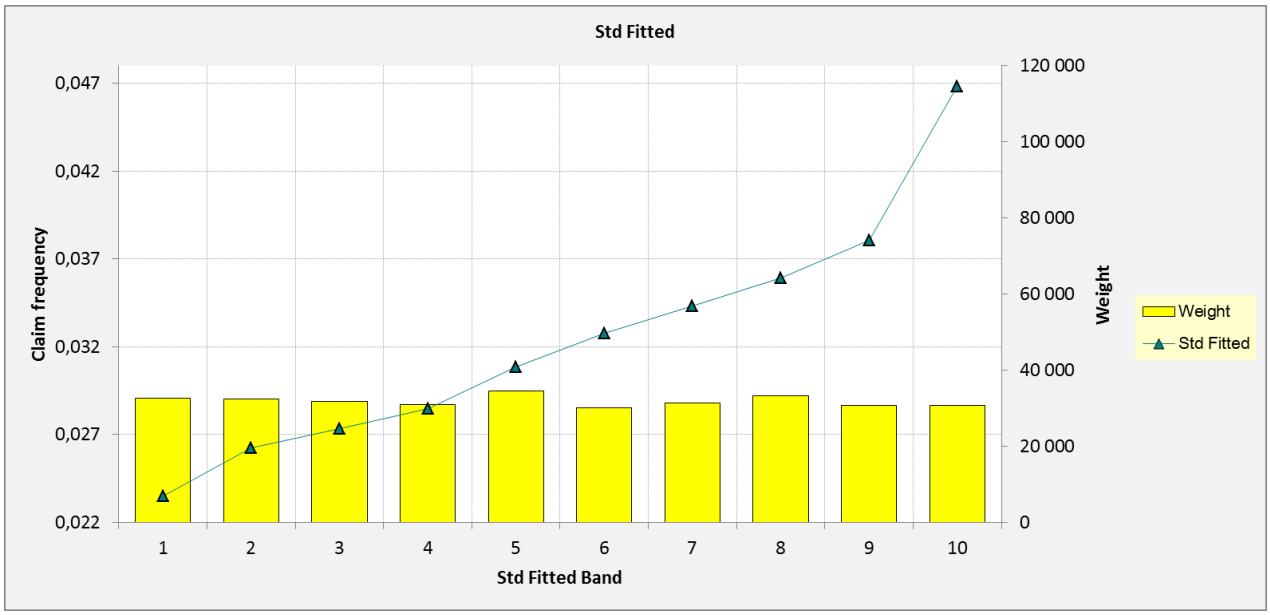
$$b_k \sim \text{N}\left(\overline{b_l}, \sigma^2\right), \tag{1}$$

missä $\overline{b_l}$ on vaikutusten b_l keskiarvo ja l on havainnon k välitön naapuri. Tasoituksen määrää säädellään parametrin σ^2 avulla. [10]

4 Esimerkki kotivakuutuksen aluehinnoittelusta

Luvussa 2.2 kotivakuutuksen rikkoutumisvahinkoaineistoon sovitettiin vahinkotiheysmalli, jossa tariffitekijöinä olivat kohteen pinta-ala, tyyppi ja vakuutuksenottajan ikä. Tarkastellaan nyt riskiä maantieteellisestä näkökulmasta ja käytetään maantieteellisenä yksikkönä postinumeroa. Yksittäisiä postinumeroita on liikaa sisällytettäväksi vahinkotiheysmalliin sellaisenaan, joten postinumerot täytyy ryhmitellä ensin ennen kuin aluetekijää voidaan käyttää tariffitekijänä mallissa. Tässä hyödynnetään spatiaalista tasoittamista ja analyysien tekemiseen käytetään Towers Watsonin Classifier – ohjelmistoa [10]. GLM-mallissa olleiden tariffitekijöiden lisäksi käytetään ulkoisena datana postinumerotasolla olevaa väkilukumuuttujaa. Alkuperäisestä aineistosta otetaan mukaan 70 prosentin satunnaisotos analysoitavaksi ja loput 30 prosenttia aineistosta jätetään analyysien ulkopuolelle testiaineistoksi mallin sopivuustarkasteluja varten.



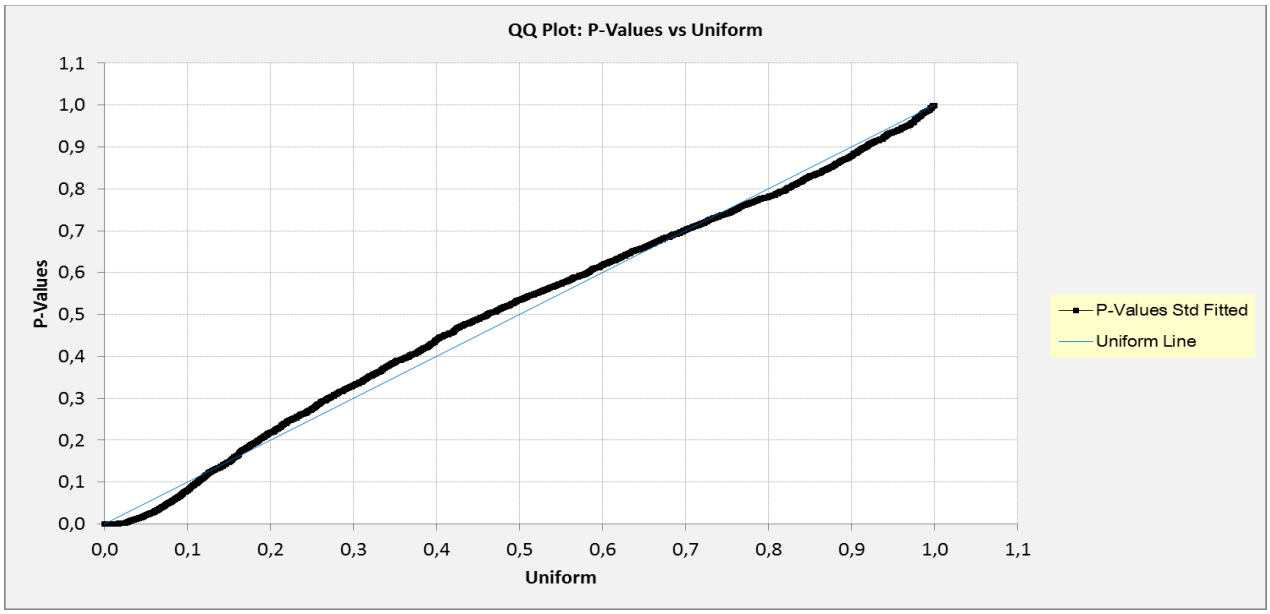


Kuva 6. GLM-mallin antamat standardoidut sovitteet vahinkotiheydelle ryhmiteltynä kymmeneen luokkaan.

Kuvassa 6 on esitetty GLM-mallin antamat sovitteet vahinkotiheydelle kartan ja kuvaajan avulla. Alueiden vahinkotiheydet on ryhmitelty kymmeneen eri luokkaan ja lajitteluperusteena on käytetty

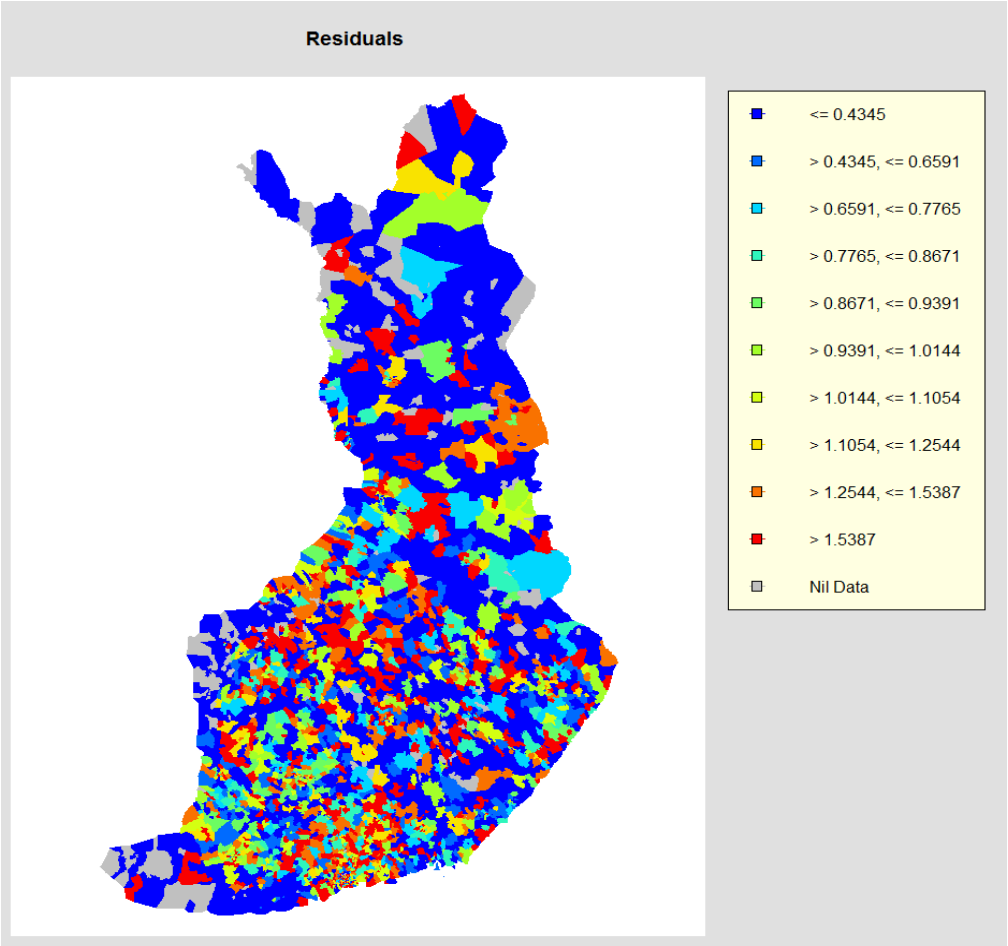
tasaista painotusta eli jokaisessa luokassa on vakuutusvuosia yhtä paljon. Sovitteet on standardoitu eli niistä on poistettu aluetekijästä riippumattomien tekijöiden vaikutus, jolloin jäljelle jää aluetekijästä johtuvaa vaihtelua ja satunnaisvaihtelua. Sovitteita tutkimalla saadaan alustava käsitys siitä, miten riski jakaantuu maantieteellisesti. Kuvan 6 kartalla alueiden vahinkotiheydet on esitetty eri värein niin, että pienimmät vahinkotiheydet on esitetty sinisellä ja suurimmat vahinkotiheydet punaisella. Kuvaajasta nähdään, että luokiteltu vahinkotiheys on 2,3 prosentin ja 4,7 prosentin välillä.

Kuvassa 7 on standardoitujen sovitteiden QQ-kuvio, joka kertoo, kuinka hyvin mallin antamat sovitteet sopivat yhteen havaitun datan kanssa. Jokaiselle postinumerolle on laskettu niin sanottu p-arvo, että saadaan havaittu arvo tai tätä suurempi arvo nollahypoteesin vallitessa. Näiden p-arvojen tulisi olla tasaisesti jakautuneita. QQ-kuviossa p-arvojen kvantiilit on piirretty tasa-(0,1)-jakauman kvantiileja vasten. Jos p-arvot sijaitsevat suoran alapuolella, malli alisovittaa dataa, ja jos taas p-arvot ovat suoran yläpuolella, niin malli yliarvioi dataa. Kuvasta 7 nähdään, että malli sopii aineistoon kaiken kaikkiaan melko hyvin. Tästä ei kuitenkaan voida vielä päätellä, sopiiko malli hyvin myös yksittäisiin alueisiin.



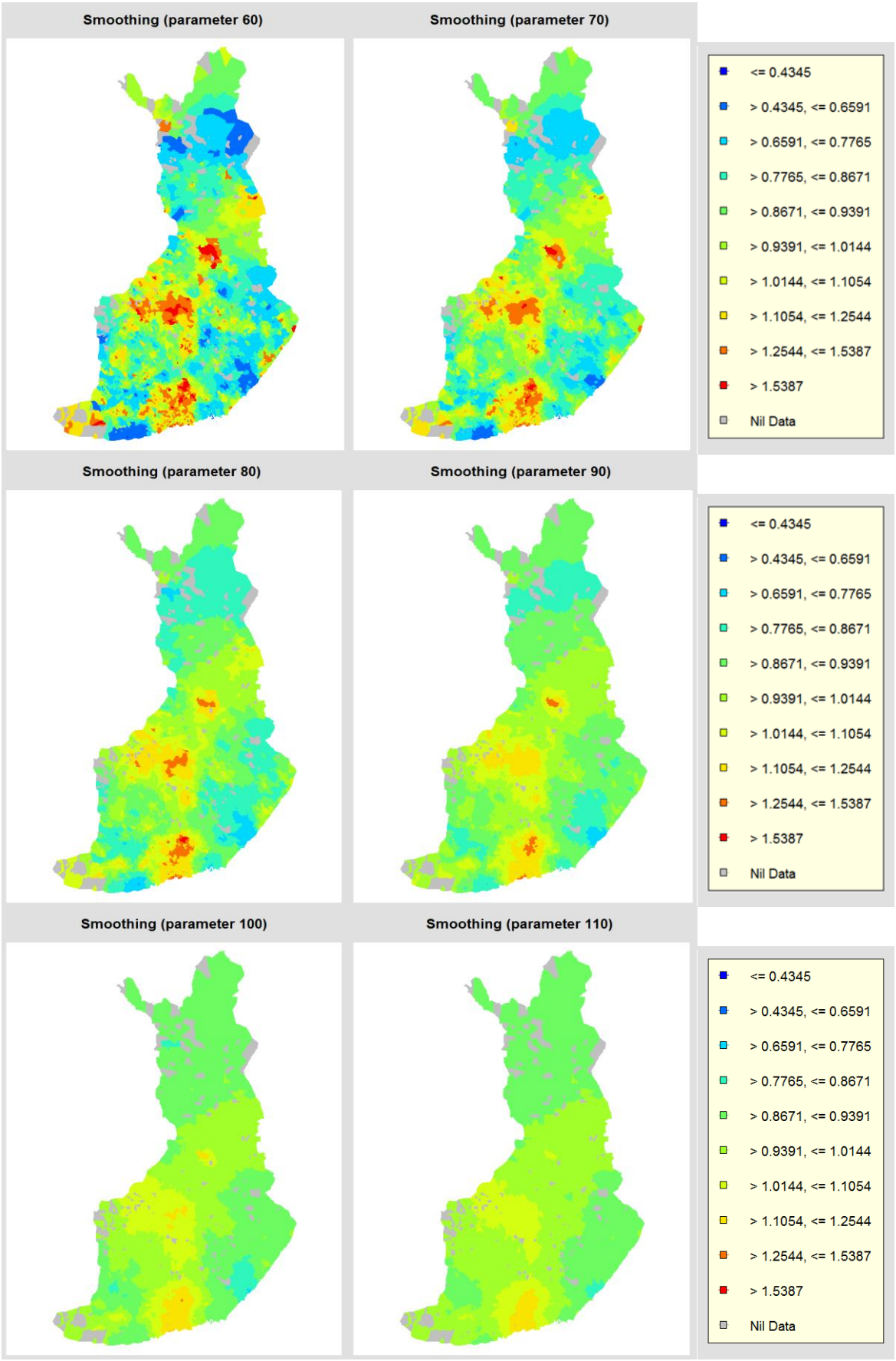
Kuva 7. Standardoitujen sovitteiden QQ-kuvio.

Kuvassa 8 on GLM-mallin antamat residuaalit ryhmiteltynä kymmeneen luokkaan. Lajitteluperusteena on käytetty jälleen tasaista painotusta. Alueet, joilla residuaalit ovat pieniä, on esitetty kuvassa sinisellä ja suurten residuaalien alueet on esitetty punaisella. Residuaalit on standardoitu eli niistä on poistettu muiden tekijöiden vaikutukset, jolloin jäljellä on aluetekijästä johtuvaa vaihtelua ja satunnaisvaihtelua. Kuvan 8 perusteella on kuitenkin hieman vaikea sanoa, onko jäljellä selvää aluevaikutusta vai ei, joten residuaaleja tasoitetaan. Tasoituksen ideana on poistaa satunnaiskohinaa, jotta nähdään, onko jäljellä aluetekijästä johtuvaa systemaattista vaihtelua.



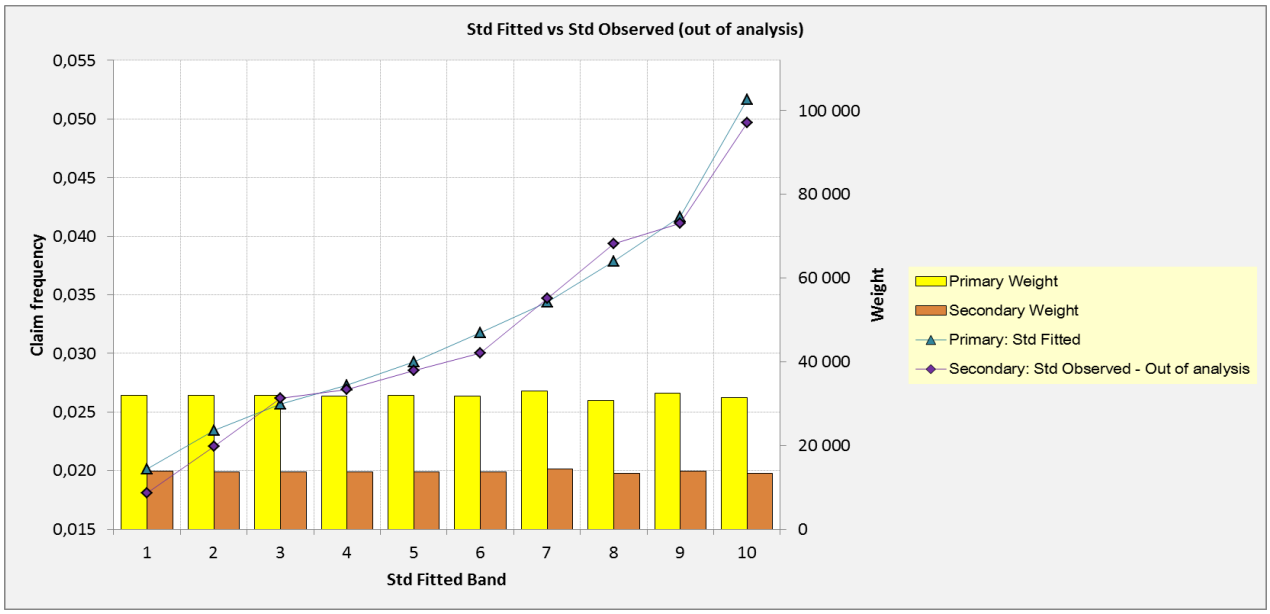
Kuva 8. GLM-mallin antamat residuaalit ryhmiteltynä kymmeneen luokkaan.

Kuvassa 9 on esitetty spatiaalisesti tasoitetut GLM-mallin antamat residuaalit. Tasoittamisessa on käytetty vierekkäisyyteen perustuvaa menetelmää (luku 3.2.2), joka olettaa, että yksittäisen alueen vahinkotiheys on naapurialueiden vahinkotiheyksien keskiarvo. Kuvan 9 vasemman yläkulman kartassa kaavaan (1) perustuva tasoitusparametri on 60 ja parametria on kasvatettu jokaisessa kuvassa kymmenellä niin, että oikean alakulman kartassa tasoitusparametri on 110. Tasoitusparametria kasvattamalla ympäröivien alueiden vaikutusta kasvatetaan. Kuvan 9 vasemman yläkulman kartassa on selviä punaisia ja sinisiä alueita, mikä on osoitus systemaattisesta aluejäännösvaihtelusta. Tasoitusparametrin kasvaessa alueiden väliset erot tasoittuvat ja lisäksi kohina vähenee. Sopivan tasoitusparametrin valinta on subjektiivista, mutta se pyritään valitsemaan niin, että oleelliset piirteet säilyvät, mutta kohinaa olisi mahdollisimman vähän. Kuvan 9 perusteella sopivin tasoituksen taso voisi olla parametrin arvolla 80 tai 90.



Kuva 9. Tasoitetut residuaalit kaavaan (1) perustuvan tasoitusparametrin arvoilla 60–110.

Edellä olevissa tarkasteluissa alkuperäisestä aineistosta oli käytössä 70 prosentin satunnaisotos. Analyysien antamien tulosten luotettavuutta voidaan tutkia vertaamalla ennusteita analyysin ulkopuolisen testiaineiston havaittuihin arvoihin. Kuvassa 10 on GLM-mallin antamat standardoidut sovitteet ja testiaineiston standardoidut havaitut arvot, kun vahinkotiheydet on ryhmitelty kymmeneen luokkaan. Nähdään, että havaittujen arvojen käyrä seuraa sovitteiden käyrää melko hyvin, joten mallin ennustuskykyä voidaan pitää hyvänä.



Kuva 10. Standardoidut sovitteet ja standardoidut havaitut arvot.

Residuaaleihin sovelletaan nyt vierekkäisyyteen perustuvan menetelmän mukaista tasoitusta tasoitusparametrin arvolla 80. Tasoituksen jälkeen residuaalit voidaan ryhmitellä luokkiin. Erilaisia ryhmittelytekniikoita on olemassa useita. Kvantiilimenetelmät muodostavat luokat niin, että kussakin luokassa on joko yhtä paljon postinumeroita tai vakuutusvuosia. Tässä työssä sovellettava, jo edeltä tuttu tasainen painotus kuuluu tähän ryhmään ja siinä kussakin luokassa on yhtä paljon vakuutusvuosia. Niin sanotut samankaltaisuus-menetelmät muodostavat luokat perustuen alueiden välisiin etäisyyksiin. [13], 191

Kun postinumeroista on saatu muodostettua alueluokat, voidaan vahinkotiheydelle sovittaa uudelleen GLM-malli niin, että tariffitekijänä on nyt aiempien tekijöiden lisäksi aluetekijä. Mallin antamat kertoimet on esitetty taulukossa 3. Perustasona aluetekijällä on alueluokka 4. Malli ennustaa vahinkotiheyden olevan 1,70 prosenttia, kun kaikkien tariffitekijöiden arvot ovat perustasolla, eli kun vakuutuksenottaja on 60–62-vuotias, asunnon pinta-ala on 50–59 m², kyseessä on koti-irtaimisto ja kohteen postinumero kuuluu alueluokkaan 4. Jos puolestaan kohteen postinumero kuuluisi alueluokkaan 10 ja muut tekijät pysyisivät muuttumattomina, niin malli ennustaa vahinkotiheydeksi $1{,}70\ \% \cdot 1{,}7505 = 2{,}98\ \%$.

Vakuutuksenottajan ikä (vuosina)

Arvo	Kerroin
Alle 30	2,1438
30-32	2,0255
33-35	2,0010
36-38	1,8375
39-41	1,8981
42-44	1,8624
45-47	1,8047
48-50	1,7035
51-53	1,4574
54-56	1,2600
57-59	1,0766
60-62	1,0000
63-65	0,9669
66-68	0,8396
69-71	0,8556
72-74	0,7582
75-77	0,6297
78-80	0,4725
Yli 80	0,4514

Pinta-ala (m^2)

Arvo	Kerroin
Alle 30	0,5803
30-39	0,5423
40-49	0,8325
50-59	1,0000
60-69	1,0972
70-79	1,4182
80-89	1,4794
90-99	1,7562
100-109	1,6603
110-119	1,7503
120-129	1,9010
130-139	1,8559
140-149	1,9259
150-159	2,0579
160-169	2,1219
170-179	2,2312
180-189	2,2595
190-199	2,5143
Yli 199	2,3646

Kohdetyyppi

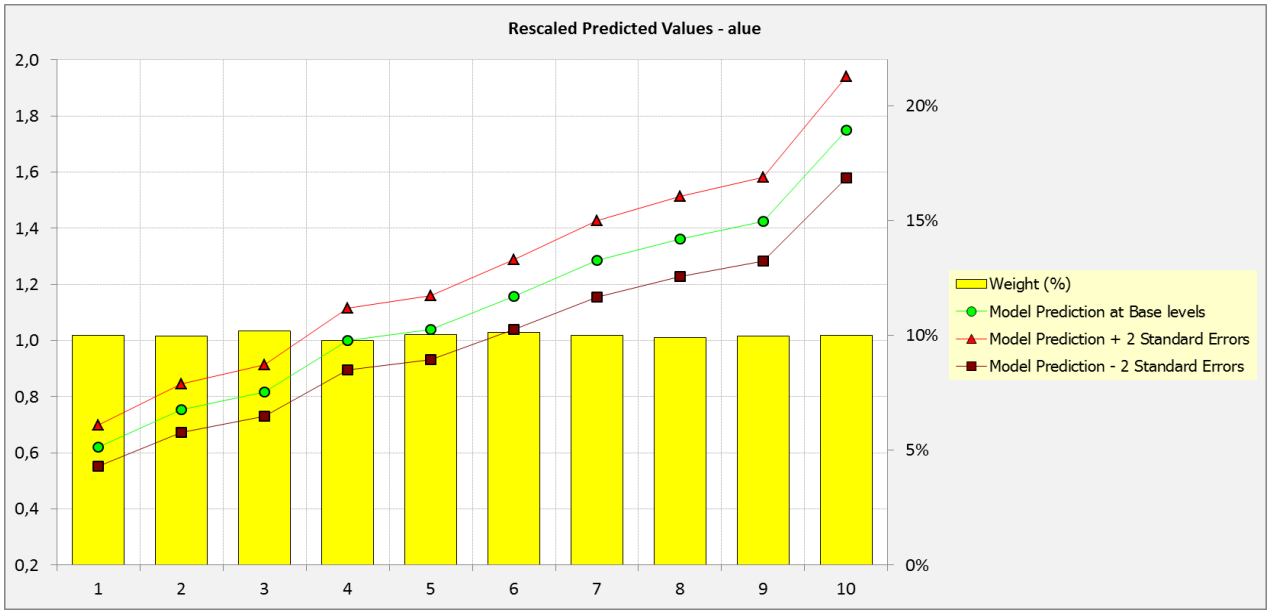
Arvo	Kerroin
Koti-irtaimisto	1,0000
Vapaa-ajan asunnon irtaimisto	0,1768

Alue

Arvo	Kerroin
1	0,6205
2	0,7545
3	0,8172
4	1,0000
5	1,0403
6	1,1564
7	1,2845
8	1,3625
9	1,4252
10	1,7505

Vakiotermi	0,0170

Taulukko 3. Vahinkotiheysmallin antamat kertoimet, kun malliin on lisätty tariffitekijäksi alue.



Kuva 11. GLM-malli vahinkotiheydelle, aluetekijän vaikutus.

Kuvassa 11 on esitetty GLM-mallin antamat kertoimet aluetekijälle, kertoimien luottamusvälit sekä vakuutusvuosien jakautuminen alueluokille. Malli ennustaa alueluokalle 10 vahinkotiheyden olevan noin 2,8-kertainen alueluokan 1 vahinkotiheyteen verrattuna muiden tekijöiden pysyessä

muuttumattomina. Luottamusväli on melko kapea, joten aluetekijää voidaan pitää mallissa tilastollisesti merkitsevänä tekijänä.

Kun aluetekijän sisältävää vahinkotiheysmallia verrataan malliin, jossa ei ole mukana aluetekijää, pienenee Akaiken informaatiokriteerin AIC arvo 1454 yksikköä. Lisäksi χ^2-testi antaa p-arvoksi nollan, kun nollahypoteesina on, että mallien välillä ei olisi eroa. Voidaan siis päätellä, että aluetekijän sisällyttäminen vahinkotiheysmalliin parantaa mallia.

5 Yhteenveto

Aluetekijä on nykyään laajasti käytössä vahinkovakuutuksen hinnoittelussa. Aiemmin yhtiöillä saattoi olla käytössä vain muutamia alueluokkia ja hinnoittelu oli samanlaista kaikissa yhtiöissä. Nyt hyödynnetään erilaisia menetelmiä maantieteellisen riskin analysoimiseksi ja alueellisten luokkien muodostamiseksi, joten hinnoittelu voi erota merkittävästi yhtiöiden välillä. Analyyseissa käytetään sekä yhtiön omaa dataa että ulkoisia lähteitä. Vahinkotiheydelle ja keskivahingolle muodostetaan omat mallinsa ja eri vahinkotyyppejä analysoidaan erillisissä malleissa.

Spatiaalisten tasoitusmenetelmien avulla voidaan hyödyntää ympäröivien alueiden informaatiota silloin, kun yksittäisestä alueesta ei ole tarpeeksi dataa. Menetelmät perustuvat oletukseen, että vierekkäisillä alueilla on samanlainen riskitaso, mikä ei välttämättä päde aina käytännössä. Tästä huolimatta menetelmien ansiosta saadaan alueista muodostettua luokkia, jotka ennustavat paremmin riskiä kuin ilman tasoitusta tehdyt analyysit. Tasoittamiseen voidaan käyttää useita erilaisia spatiaalisen analyysin menetelmiä.

Spatiaalisen tasoituksen ideana on poistaa jäännöstermeistä satunnaiskohinaa niin, että jäljelle jäisi vain aluetekijästä johtuvaa vaihtelua. Sopivan tasoitusparametrin valinnalla on tässä keskeinen merkitys. Jos tasoitusta tehdään liikaa, kadotetaan myös aluetekijästä johtuvaa vaihtelua. Jos taas tasoitusta tehdään liian vähän, jäljelle voi jäädä huomattava määrä satunnaiskohinaa.

Spatiaalisten tasoitusmenetelmien avulla tehty aluehinnoittelu on subjektiivista, joten validointi on erityisen tärkeää. Analyysien antamia tuloksia täytyy tarkastella useasta näkökulmasta ennen johtopäätösten tekemistä. Alkuperäisestä aineistosta voidaan jättää tietty otos analyysien ulkopuolelle testiaineistoksi ja verrata analyysien antamia tuloksia testiaineiston arvoihin. Tarkastelemalla GLM-mallin aluetekijälle antamien kertoimien luottamusvälejä saadaan käsitystä siitä, kuinka merkitsevä aluetekijä on mallissa ja näyttääkö estimaatteihin sisältyvän epävarmuutta. Riskin jakaantumista alueille voidaan visuaalisesti tarkastella karttojen avulla.

Lähteet

[1] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Thandi, N. 2007: A Practitioner's Guide to Generalized Linear Models. Casualty Actuarial Society. Viitattu 31.8.2014. http://www.casact.org/pubs/dpp/dpp04/04dpp1.pdf

[2] Boskow, M., Verrall, R. J. 1994: Premium Rating by Geographic Area Using Spatial Models. ASTIN BULLETIN, Vol 24, No 1, 131–143. Viitattu 31.8.2014. http://www.actuaries.org/LIBRARY/ASTIN/vol24no1/131.pdf

[3] de Jong, P., Heller, G. Z. 2008: Generalized Linear Models for Insurance Data. Cambridge University Press.

[4] Guven, S. 2004: Multivariate Spatial Analysis of the Territory Rating Variable. Casualty Actuarial Society. Viitattu 31.8.2014. http://www.casact.org/pubs/dpp/dpp04/04dpp245.pdf

[5] Heikkinen, J. 2005: Yleistetyt lineaariset mallit. Matematiikan ja tilastotieteen laitos, Helsingin yliopisto. Viitattu 31.8.2014. http://www.rni.helsinki.fi/~jmh/glm05/glm05.pdf

[6] Heikkinen, J. 2007: Alueittaisten aineistojen spatiaalinen analyysi. Matematiikan ja tilastotieteen laitos, Helsingin yliopisto. Viitattu 31.8.2014. http://www.rni.helsinki.fi/~jmh/mrf07/mrf07.pdf

[7] Ohlsson, E., Johansson, B. 2010: Non-Life Insurance Pricing with Generalized Linear Models. EAA Series, Springer.

[8] Taylor, G.C. 1989: Use of Spline Functions for Premium Rating by Geographic Area. ASTIN BULLETIN, Vol 19, No 1, 91–122. Viitattu 31.8.2014. http://casualtyactuarialsociety.net/library/astin/vol19no1/91.pdf

[9] Taylor, G.C. 2001: Geographic Premium Rating by Whittaker Spatial Smoothing. ASTIN BULLETIN, Vol 31, No 1, 147–160. Viitattu 31.8.2014. http://www.actuaries.org/LIBRARY/ASTIN/vol31no1/147.pdf

[10] Towers Watson Classifier (V1.6.5.89926) 2014.

[11] Towers Watson Emblem (V4.2.89.79700) 2014.

[12] Werner, G. 1999: The United States Postal Service's New Role: Territorial Ratemaking. Casualty Actuarial Society Forum. Viitattu 31.8.2014. http://www.casact.org/pubs/forum/99wforum/wf99287.pdf

[13] Werner, G., Modlin, C. 2010: Basic Ratemaking. Casualty Actuarial Society. Viitattu 31.8.2014. http://www.casact.org/library/studynotes/Werner_Modlin_Ratemaking.pdf