Stockholms universitet

Prissättning för skadeförsäkring med postnummer som kredibilitetsfaktor

Fredrik Bjärnek

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:13
Matematisk statistik
Oktober 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2014:13**
http://www.math.su.se/matstat

Prissättning för skadeförsäkring med postnummer som kredibilitetsfaktor

Fredrik Bjärnek*

Oktober 2014

Sammanfattning

I många försäkringar är geografi en riskförklarande variabel. Risken att råka ut för en skada och skadekostnadens storlek varierar beroende på var vi befinner oss geografiskt. Vi kommer att använda dataunderlag från en personbilsförsäkring och visa hur vi i prissättning kan använda postnummer som kredibilitetsfaktor i en multiplikativ modell för skadefrekvensen.
Nyckelord: Skadeförsäkring, prissättning, geografi, kredibilitetsfaktorer

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: fredrik.bjarnek@folksam.se. Handledare: Andreas Nordvall Lagerås.

Abstract

In many different insurance products the risk varies depending on the geography. Both the risk of being involved in a claim and the cost of the claim. We will use data from a carinsurance and show how we with credibility can use postcodes in pricing when modeling the claim frequency.

Tack

Jag vill speciellt tacka Jesper Andersson och Robert Nygren på Folksam som har hjälpt till med problemställningen, kommit med synpunkter och inte minst envist insisterat på att min uppsats ska bli klar. Jag vill rikta ett stort tack till Andreas Lagerås som har varit min handledare. Han har varit ett utmärkt bollplank och har tagit sig tid att läsa igenom och komma med mycket värdefull feedback.

Innehållsförteckning

Inledning

Grundtanken med försäkringar är att en försäkringstagare mot en avgift (premie) som betalas till försäkringsbolaget erhåller rätt till ersättning för vissa i förväg överenskommande händelser. För skadeförsäkring kan det exempelvis vara ett skydd mot att ens motorcykel blir stulen och att man ersätts med en ny, eller att fritidshuset får vattenskador och att man behöver byta golvet.

Vilket pris försäkringsbolag väljer att ta ut för olika försäkringar är upp till bolagen själva. Det ligger i försäkringsbolagens intresse att ta ut ett så rättvisande pris som möjligt, varken ligga för lågt eller för högt jämfört med konkurrenterna. Om priset är för högt är det troligt att kunden tecknar försäkringen hos ett annat bolag istället och är det är för lågt finns risken att många avtal som är för lågt prissatta tecknas och det kan generera en ekonomisk förlust.

Syftet med uppsatsen är att visa hur vi genom tillgång till information om försäkringstagarens geografiska adress kan använda det i prissättningen, och att visa vilka styrkor och svagheter som finns. Dels för att vi vill fånga den riskmässiga skillnaden som finns geografiskt, dels för att ta hänsyn till andra aspekter, som exempelvis att det ur ett kundperspektiv borde finnas en logik och i prissättningen.

Avgränsningar och definitioner av centrala begrepp

Av den premie som en försäkringstagare betalar är en stor del den förväntade skadekostnaden för avtalet och det är hur den kan modelleras som detta arbete handlar om. Övriga delar av premien som exempelvis driftskostnader och vinst behandlas inte.

Prissättning inom skadeförsäkring kan ske på olika sätt, vi tar upp frågeställningar som kan uppstå när generaliserade linjära modeller (GLM) används och hur vi kan hantera de frågorna. Hur detaljerad prissättningen av produkten är varierar kraftigt beroende på dess komplexitet. Det kan vara allt från samma premie för alla (enhetspremie) till att premien bestäms av utseendet på över tio olika premieargument. I de fall premieargumentet är en kontinuerliga variabel så kan det delas in i olika klasser med hjälp av intervall.

För en villaförsäkring skulle ett exempel på hur det kan se ut vara:

Premieargument	Klass	Intervall
Försäkringstagarens ålder	1	0-25 år
	2	26-40 år
	3	41-60 år
	4	61 år och äldre
Villans storlek, boyta	1	0-80 kvm
	2	81-200 kvm
	3	201 kvm och större
Geografi	1	Skåne, Västra Götaland och Stockholms Län
	2	Övriga Län

Tabell 1

En kombination av premieargumenten anger premien. De olika kombinationerna kan i exemplet för villa visas genom tabellen:

Försäkringstagarens ålder	Villans storlek, byggyta	Geografi
1	1	1
1	1	2
1	2	1
1	2	2
1	3	1
1	3	2
2	1	1
2	1	2
.	.	.
.	.	.
.	.	.
4	2	2
4	3	1
4	3	2

Tabell 2

Alla kunder som hamnar i samma tariffcell antas vara lika risker. Det är på den nivån som skadorna och beståndet analyseras och modelleras. Det för att på bästa sätt få fram en modell som förklarar de skadekostnader som uppstår med de åtaganden som försäkringsbolaget har.

Det som påverkar skadekostnaderna är:

- hur ofta sker en skada? (skadefrekvens)
- när en skada sker, hur mycket förväntas den kosta? (medelskada)

För att kunna mäta skadefrekvensen behöver vi ett mått på hur länge en försäkring har varit gällande. Två försäkringar i samma tariffcell är bara lika stora risker förutsatt att de gäller lika länge. Om den ena gäller tolv månader och den andra i fyra månader är det mer troligt att den som gäller i tolv månader råkar ut för en skada än den som gäller fyra månader. Vi inför duration som ett mått för hur lång tid en försäkring gäller (exponering), duration mäter vi i antal försäkringsår. Vi låter *x* vara antalet dagar som försäkringen är gällande och vi får durationen genom $^{x}/_{365}$ (om skottår 366).

De här två nyckeltalen kan uttryckas

$$\text{Skadefrekvens} = \frac{\text{Antal skador}}{\text{Duration}}$$

$$\text{Medelskada} = \frac{\text{Skadekostnad}}{\text{Antal skador}}$$

Det vi vill få fram är förväntad skadekostnad per försäkringsår. Vi kallar det nyckeltalet för riskpremie

$$\text{Riskpremie} = \frac{\text{Skadekostnad}}{\text{Duration}}$$

Vi ser att det är samma sak som att multiplicera skadefrekvensen med medelskadan så vi har sambandet

$$\text{Riskpremie} = \text{Skadefrekvens} \times \text{Medelskada}$$

Man kan antingen modellera riskpremien direkt eller göra en separat för skadefrekvensen och en för medelskadan och sedan sammanföra dem till en slutlig tariff. Vi kommer i det här arbetet att fokusera på skadefrekvensen.

En introduktion till personbilsförsäkring

Personbilsförsäkring består i huvudsak av tre olika moment (exklusive tilläggsförsäkring):

Trafikförsäkring

Trafikförsäkring måste enligt svensk lag alla personbilar som är i trafik ha. Trafikförsäkringen ger ersättning för personskador på förare, passagerare samt utomstående. Den ersätter även skador som orsakas av fordonet på annans egendom. Trafikförsäkringen ger ingen ersättning för skador som du själv orsakat på ditt eget fordon.

Delkaskoförsäkring

Delkaskoförsäkringen säljs ofta som ett paket och innehåller skydd mot följande skador (avvikelser kan förekomma):

- Brand
- Glas
- Stöld
- Räddning
- Rättsskydd
- Maskin- & elektronik

Vagnskadeförsäkring

Vagnskadeförsäkringen ger ersättning för skador på karossen som försäkringstagaren själv åstadkommit på sitt eget fordon.

De här momenten kan försäkringstagaren själv kombinera efter det behov som finns och välja trafikförsäkring, halvförsäkring, helförsäkring eller avställningsförsäkring.

	Trafikförsäkring	Delkaskoförsäkring	Vagnskadeförsäkring
Trafikförsäkring	Ja	Nej	Nej
Halvförsäkring	Ja	Ja	Nej
Helförsäkring	Ja	Ja	Ja
Avställningsförsäkring	Nej	Ja	Nej

Tabell 3

Teori

Innan vi ansätter en statistisk modell och går närmare in på teorin får vi från [1] några grundläggande antaganden som ska vara uppfyllda.

Antagande 1: Avtalsoberoende

Utfallen för olika försäkringsavtal är oberoende av varandra.

Vi antar att skadeutfallet för försäkring f_1 inte påverkas av skadeutfallet för försäkring f_2. För personbilsförsäkring är det inte svårt att hitta fall när detta inte är uppfyllt. Exempelvis är det inte otänkbart att två fordon försäkrade i samma försäkringsbolag är involverade i en olycka. Vi anser att effekterna av detta är försumbara.

Antagande 2: Tidsoberoende

Utfallen i disjunkta tidsintervall är oberoende av varandra

Även mot detta antagande går det att hitta fall där detta kan ifrågasättas. Exempelvis om ett fordon parkeras på gatan över natten och utsätts för ett stöldförsök är det inte otänkbart att försäkringstagaren i framtiden väljer att parkera i ett garage istället. Eller att en förare som kört vårdslöst och orsakat en trafikolycka framöver ändrar sitt körbeteende och kör lugnare. För de flesta fall känns dock detta antagandet rimligt.

En konsekvens av de här två antaganden om oberoende blir att kostnaderna för enskilda skador är oberoende av varandra då de antingen berör olika försäkringsavtal eller sker vid olika tidspunkter.

Antagande 3: Homogenitet

Inom en tariffcell har två utfall med samma exponering samma fördelning

Antagandet kan upplevas som att vi tvingas till att skapa väldigt små indelningar av våra tariffceller för att detta ska gälla, det begränsas dock av att vi inte i alla produkter har tillräckligt mycket dataunderlag för att kunna skatta parametrarna.

Väntevärde och varians

Anta att vi vill titta på ett nyckeltal för en viss tariffcell, ett nyckeltal Y med exponering w som ger upphov till ett utfall X, $Y = \frac{X}{w}$. Vi låter Z vara ett utfall med exponeringen w=1 och definierar

$$\mu \coloneqq E[Z] \qquad\qquad \sigma^2 \coloneqq Var(Z)$$

Vi noterar att en följd av antagande 3 ovan är att alla dessa Z har samma väntevärde och varians.

Om vi antar att Z är medelskadan innebär det att w är antal skador och x är skadekostnad. Då blir totala skadebeloppet summan av $Z_1, Z_2, \dots, Z_w$, där Z_j är skadekostnaden för den j:te skadan, alla Z har exponeringen $w = 1$. Av antagande 1-3 följer att dessa Z_k är oberoende och likafördelade, då de antingen sker vid olika tidpunkter eller kommer från olika försäkringsavtal.

Från [1] sida 14 får vi följande hjälpsats

Lemma 1
Under antagande 1-3 gäller att om X är ett utfall av skadekostnad eller antal skador, samt att w>0 och $Y = {X}/{w}$ ges dessa variablers väntevärde och varians av:

$$E[X] = w\mu\,, \qquad Var(X) = w\sigma^2$$

$$E[Y] = \mu, \qquad Var(Y) = {\sigma^2}/{w}$$

där μ och σ^2är väntevärde och varians för ett utfall med exponeringen $w = 1$.

Multiplikativa modeller

Den modell som är mest använd för prissättning av sakförsäkringar är multiplikativa modeller, vilket även är det som vi kommer att använda oss av. Anta att vi har K premieargument där k_i anger hur många klasser det i:te premieargumentet har. Låt oss börja med fallet där vi har två premieargument, det ger oss en tariffindelning där vi låter (i, j) ge oss information om vilken klass av första och andra premieargumentet som cellen tillhör. Vi vill nu ställa upp en modell för hur ett godtyckligt nyckeltal förklaras av våra två premieargument och deras indelningar. Vi inför $w_{i,j}$ för exponeringen i cell (i, j) och $X_{i,j}$ för utfallet i cell (i, j), det ger oss nyckeltalet $Y_{i,j}$ för cell (i, j) ty

$$Y_{i,j} = \frac{X_{i,j}}{w_{i,j}}$$

Från Lemma 1 får vi att $E[Y_{i,j}] = \mu_{i,j}$, där $\mu_{i,j}$ är väntevärdet vid exponeringen $w = 1$. Vi ansätter en multiplikativ modell för väntevärdet

$$\mu_{i,j} = \gamma_0 \gamma_{1,i} \gamma_{2,j}$$

Där $\gamma_{1,i}$ $(i = 1,2, \dots, k_1)$ är de parametrar som svarar mot klasserna i premieargument 1 och $\gamma_{2,j}$ $(j = 1,2, \dots, k_2)$ mot premieargument 2. Den här modellen ger inte någon möjlighet att tillåta samspel mellan premieargumenten, det innebär att relationen mellan klass 1 och klass 2 för premieargument 1 är densamma oavsett om vi befinner oss i samma eller olika klass i premieargument 2.

Det är viktigt att poängtera att det är just det som den här modellen förklarar, den relativa riskskillnaden för de olika klassindelningarna inom olika de två premieargumenten. För att vi ska ha något att relatera till väljer vi ut en klass i respektive premieargument som vi använder som bascell och ger värdet 1,00, det bör vara den klass i respektive premieargument som innehåller mest exponering så att vi jämför mot den statistiskt säkraste skattningen. Då får vi en cell med värdet 1,00, nämligen den cell som är basklass i både premieargument 1 och 2, och då kan vi tolka γ_0 som den nivån som gäller i bascellen, oavsett om vårt nyckeltal är skadefrekvens, medelskada eller riskpremie.

Generaliserade linjära modeller (GLM)

Vanligt förekommande för prissättning inom sakförsäkringsbranschen är generaliserade linjära modeller (GLM), det kommer även vi i detta arbete att använda oss av. Vi kommer i detta avsnitt att översiktligt gå igenom den teorin.

Exponentiella dispersionsmodeller

Vi har vårt data på vektorform $\mathbf{y}' = (y_1 y_2, \dots, y_n)$ där n i detta fall anger antal observationer.

Vi inför

μ_i	väntevärdet för observation nr i
w_i	exponering i observation nr i
Y_i	motsvarande nyckeltal

där vi benämner w för vikter och Y för respons. Ett viktigt antagande för GLM är att frekvensfunktionen för varje Y_i kan skrivas på formen

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right\} \quad (1)$$

Funktionen gäller bara för de y_i som kan antas av den stokastiska variabeln Y_i, vidare gäller restriktionerna att $w_i \geq 0$ och $\phi > 0$. Vi antar också att funktionen $b(\theta_i)$ är två gånger kontinuerligt deriverbar. Sannolikhetsfördelningar som kan skrivas på det här sättet kallas exponentiella dispersionsmodeller (EDM).

Sannolikhetsfördelning för skadefrekvensen

I det här arbetet kommer vi att gå in på skadefrekvensen eftersom att det är den som vi i slutändan kommer att analysera. För information om modeller för medelskadan och riskpremien hänvisar vi till [1] kapitel 2.

Om X_i är antalet skador i cell i som har durationen w_i får vi nyckeltalet skadefrekvensen Y_i genom

$$Y_i = \frac{X_i}{w_i}$$

Under de antaganden som vi tidigare gjort om oberoende och homogenitet kommer X_i, se [1] sida 22, att vara Poissonfördelade om skadorna kommer en åt gången. Låt μ_i vara väntevärdet av antalet skador i cell i under förutsättning att durationen $w_i = 1$, det vill säga ett riskår.

Enligt tidigare Lemma 1 får vi att

$$E[X_i] = w_i \mu_i$$

och sannolikhetsfördelningen för X_i kan skrivas som

$$f_{X_i}(x_i; \mu_i) = e^{-w_i \mu_i} \frac{(w_i \mu_i)^{x_i}}{x_i!}$$

Sannolikhetsfördelningen för skadefrekvensen Y_i kallas i [1] för relativ Poisson och vi får även därifrån att sannolikhetsfunktionens utseende

$$\begin{aligned} f_{Y_i}(y_i; x_i) &= P(Y_i = y_i) = P(X_i = w_i y_i) = e^{-w_i \mu_i} \frac{(w_i \mu_i)^{w_i y_i}}{(w_i y_i)!} \\ &= \exp\{-w_i \mu_i + w_i y_i \log(w_i \mu_i) - \log(w_i y_i!)\} \\ &= \exp\{w_i (y_i \log(\mu_i) - \mu_i) + w_i y_i \log(w_i) - \log(w_i y_i!)\} \end{aligned}$$

Om vi inför $c(y_i, w_i) = w_i y_i \log(w_i) - \log(w_i y_i!)$ och får då

$$f_{Y_i}(y_i; x_i) = \exp\{w_i (y_i [\log(\mu_i) - \mu_i) + c(y_i, w_i)\}$$

För att visa att det är en EDM parametriserar vi om med parametern $\theta_i = \log(\mu_i)$ och får då

$$f_{Y_i}(y_i; \theta_i) = \exp\{w_i (y_i \theta_i - e^{\theta_i}) + c(y_i, w_i)\}$$

Vi ser att ekvationen stämmer överens med ekvationen (1) med $\phi = 1$ och $b(\theta_i) = e^{\theta_i}$ vilken är deriverbar hur många gånger som helst.

Ett rimligt krav att ställa är att om vi, efter ihopslaging av två celler, fortfarande är kvar i samma familj av fördelningar. Anta att vi har skadefrekvenserna Y_1 och Y_2 relativt Poissonfördelade med exponeringarna w_1 och w_2 samt att båda har parameter μ. Slår vi ihop dessa celler får vi den sammanvägda skadefrekvensen

$$Y = \frac{Y_1 w_1 + Y_2 w_2}{w_1 + w_2}$$

Då är Y relativt Poissonfördelad med exponering $w_1 + w_2$ och parameter μ, det får vi också genom att $w_1 Y_1 + w_2 Y_2$ är summan av två oberoende Poissonfördelade variabler och den summan är också Poissonfördelad. Denna egenskap hos sannolikhetsfördelningar kallas reproducerbarhet.

Kumulantgenererande funktion – väntevärde och varians

För att erhålla väntevärde och varians för önskat nyckeltal ska vi använda oss av den kumulantgenererande funktionen som är logaritmen av den momentgenererande funktionen.
I det fall då vårt nyckeltal tillhör en diskret fördelning får vi följande kumulantgenererande funktion

$$\psi(t) = \frac{b\left(\theta + {}^{t\phi}/_{w}\right) - b(\theta)}{{}^{\phi}/_{w}}$$

De två första kumulanterna är väntevärde och varians, för att få fram dem så deriverar vi den kumulantgenererande funktionen två gånger

$$\psi'(t) = b'(\theta + {}^{t\phi}/_{w})$$

$$\psi''(t) = b''(\theta + {}^{t\phi}/_{w})\,{}^{\phi}/_{w}$$

Väntevärdet ges av $\psi'(0)$ och variansen av $\psi''(0)$ vilket i vårt fall där skadefrekvensen är det nyckeltal som vi analyserar ger oss

$$E[Y] = \psi'(0) = b'(\theta) = e^{\theta} = \mu$$
$$Var(Y) = \psi''(0) = \frac{\mu}{w}$$

Vilket stämmer bra i fallet med att skadefrekvensen är relativt Poissonfördelad.
För det generella fallet kan vi sammanfatta det enligt

Lemma 2
Antag att Y_i följer EDM, då gäller att dess kumulantgenererande funktion ges av

$$\psi(t) = \frac{b(\theta_i + \frac{t\phi}{w_i} - b(\theta_i))}{{}^{\phi}/_{w_i}}$$

Samt att

$$\mu_i \coloneqq E[Y_i] = b'(\theta_i)$$
$$\sigma_i^2 \coloneqq Var(Y_i) = \phi v(\mu_i)/w_i$$

Där $v(\mu_i)$ är variansfunktionen $b''(.)$ uttryckt som en funktion av μ_i, nämligen
$v(\mu_i) = b''(b'^{-1}(\mu_i))$

Sats 1
Alla EDM är reproducerbara, det vill säga att om vi har två oberoende variabler Y_1 och Y_2 från samma EDM-familj (samt samma väntevärde) och vi bildar det w-viktade medelvärdet

$$Y = \frac{Y_1 w_1 + Y_2 w_2}{w_1 + w_2}$$

Då får Y en ny fördelning inom samma familj med samma väntevärde men med en ny vikt som i detta fallet är $w_1 + w_2$.

Länkfunktionen

Vi har tidigare nämnt att vi har tänkt att ansätta en multiplikativ modell på formen

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j}$$

För att vi skall kunna använda GLM och de färdiga programvaror som hanterar de modellerna för att skatta parametrarna behöver modellen ha en linjär struktur enligt

$$\mu_{ij} = \gamma_0 + \gamma_{1i} + \gamma_{2j}$$

Genom att logaritmerera vårt multiplikativa uttryck får vi

$$\log(\mu_{ij}) = \log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j})$$

GLM tillåter att vi genom en så kallad länkfunktion övergår från vår ursprungliga multiplikativa modell till en linjär modell, i vårt fall kallas det att vi använder en log-länk.

Kredibilitetsteori

En viss typ av premieargument har väldigt många indelningar och det finns inte något bra sätt att gruppera. Det vi ska titta på är ett sådant premieargument, nämligen postnummer. Vi vill dela in landet i ett antal olika riskområden och behandla det som övriga premiepåverkande faktorer, men vi måste ta reda på vilka postnummer som ska höra till vilket riskområde.

I Sverige har vi ungefär 16 500 olika postnummer, om vi bortser från boxadresser vilket vi gör så är det omkring 11 000. För att kunna behandla geografin på samma sätt som övriga premieargument skulle det krävas mycket data i respektive postnummer för att få modellen att konvergera. Så mycket data har vi inte tillgång till. Premieargument av den här typen kallas för multiklassargument eller Multi-Level Factor (MLF). Vi ska titta på ett sätt som vi kan hantera multiklassargument, det kallas för kredibilitetsteori.

Ansatsen är att vi har ett multiklassargument. För vår del gäller det att postnummer är ett multiklassargument och vi kommer i fortsättningen att prata om just det fallet. Anta att vi vill bestämma skadefrekvensen Y och vi har observationer Y_{kt} , där t är antal upprepningar i multiklass k och w_{kt} är exponeringen. Det viktade medelvärdet $\overline{Y}_k$ ger en individuell skattning av skadefrekvensen, men för många postnummer baseras den på alldeles för lite data för att ge oss en pålitlig skattning. Kredibilitetsteorins grundidé är att kompromissa mellan det observerade $\overline{Y}_k$ och hela kollektivets väntevärde μ

$$\widehat{Y}_k = z_k \overline{Y}_{k\cdot} + (1 - z_k)\mu$$

där $0 \leq z_k \leq 1$ är en vikt och kallas kredibilitetsfaktorn och $\widehat{Y}_k$ kallas kredibilitetsskattning.

Dessa kredibilitetsfaktorer vill vi ska ha ett antal egenskaper:

- Ju mer exponering, desto pålitligare observationer och större vikt
- Ju mindre variation inom i klassen, desto större vikt
- Ju större variation mellan klasserna, desto större vikt

Vi har som tidigare vårt nyckeltal Y_{ikt} med exponeringsvikt w_{ikt}, index t står för upprepade observationer i kombinationer av i och k. De försäkringar från multiklass k antas ha en multiplikativ avvikelse från μ_i. Denna avvikelse väljer vi att betrakta som en stokastisk variabel U_k, med $E[U_k] = 1$. Det gör att vi kan skriva

$$E[Y_{ikt}|U_k] = \mu_i U_k$$

Tidigare har vi sett hur vi kan få skatta parametrarna i GLM för våra relationstal och därmed även ett värde på μ_i, det vi nu vill är att även skatta den stokastiska variabeln U_k.

En algoritm för skadefrekvensen med multiklassargument

Vi kan med en multiplikativ GLM-modell i standardprogramvaror, exempel SAS[1], skatta μ_i.
Nu har vi även en stokastisk effekt U_k som motsvarar en multiklass k, vi vet hur vi skattar den givet μ_i. Den enda skillnaden mot innan är att vi givet $U_k = u_k$ ska skatta $\mu_i u_k$ som ingår i väntevärdet. I GLM kallas en variabel som ska adderas till det linjära uttrycket för offset, och i vårt fall med en log-länk är det $\log(u_k)$ som är offset. Resultatet av analysen blir relationstal och skattningar $\hat{\mu}_i$. För att skatta $\hat{u}_k$ behöver vi känna till μ_i, naturligt är att iterera mellan $\hat{\mu}_i$ och $\hat{u}_k$ tills vi uppnår konvergens, man kan beskriva den iterativa processen med följande algoritmen från [1].

(0) Sätt initialt $\hat{u}_k = 1$ för alla k.

(1) Gör en GLM-analys med alla standardargument samt med $\log(\hat{u}_k)$ som offset. (Poissonfördelad och log-länk i GLM)

(2) Skatta $\alpha = \frac{\sigma^2}{\sigma_U^2}$ använd $\hat{\mu}_i$ från (1).

(3) Beräkna $\hat{u}_k$ för alla $k = 1,2, \ldots K$, använd $\hat{\mu}_i$ och $\hat{\alpha}$ från (1) och (2).

(4) Återvänd till (1) med log $(\hat{u}_k)$ som offset.

Upprepa sedan steg (1)-(4) tills konvergering erhålls.

Beräkningen av $\hat{u}_k$ görs enligt

$$\hat{u}_k = z_k \bar{u}_k + (1 - z_k)$$

där vi har kredibilitetsfaktorn

$$z_k = \frac{\tilde{w}_{\cdot k}}{\tilde{w}_{\cdot k} + \sigma^2 / \sigma_U^2}$$

och erfarenhetsvärdet

$$\bar{u}_k = \frac{\sum_i \tilde{w}_{ik} y_{ik} / \mu_i}{\tilde{w}_{\cdot k}}$$

samt

$$\tilde{w}_{ik} = w_{ik} \frac{\mu_i^2}{v_i} \qquad \tilde{w}_{\cdot k} = \sum_i \tilde{w}_{ik}$$

[1] *SAS, http://www.sas.com/*

Analys

Vi har valt att studera delkaskoförsäkringen för personbil och se hur försäkringstagarens geografi tillsammans med några andra argument påverkar skadefrekvensen. Vi kommer alltså att titta på hur sannolikt det är att en skada inträffar. Att använda den adress som försäkringstagaren är skriven på är rimligt då bilen oftast borde stå parkerad nära bostaden och då man ofta färdas på de vägar som leder till och från det området.

Vi kommer att använda en multiplikativ modell med Poissonfördelning och log-länk, samt 5-ställigt postnummer som multiklassargument. Övriga premiepåverkande argument som vi har med är

- Försäkringstagarens ålder Klass 1-3
- Fordonets ålder Klass 1-4
- Bilklassning Klass 1-3

För både försäkringstagarens, samt fordonets ålder hamnar de yngsta i klass 1 och de äldsta i klass 3 respektive 4. Bilklassning är en gruppering av fordon med liknande egenskaper. Alla försäkringsbolag har möjlighet att ha olika indelningar. Indelningar som är olika beroende på om det är trafik, delkasko eller vagnskada som analyseras. Vi kommer i det här arbetet att fokusera på delkaskoklassningen. Exempel på egenskaper som ligger som grund för grupperingarna kan vara:

- Priser på reservdelar
- Utrustningen i bilen
- Marknadsvärdet på bilen

För bilklassningen finns de fordon med lägst risk i klass 1 medan de högsta riskerna är i klass 3.

Det angreppssätt som vi använder oss är att vi först gör analyser enligt tidigare förslagen algoritm och sedan så tar vi hjälp av de närliggande postnumren och jämnar ut skattningarna. Det ger oss bättre skattningar för de områden där vi har ingen/lite exponering. Även i de fall som vi har relativt mycket information är det rimligt att ta hjälp från närliggande postnummer. Stora premieskillnader mellan två närliggande postnummer är i många fall ologiskt.

Regelbundna översyner av skattningarna för postnumren bör ske. Om man använder sig av en utjämning av skattningarna så minskar risken för att få stora förändringar från år till år vilket också är att föredra. Annars så säger vi att ena året så är ett visst postnummer ett lågriskområde och året efter ett högriskområde. Stabilitet när det gäller prissättning är viktigt för att få ett förtroende, både hos personalen internt men främst hos kunderna. Att ett bolag chockhöjer en premie från ett år till ett annat gör att många ser sig om efter att hitta en ny försäkring.

GLM-analys utan geografi

Med hjälp av teorin för GLM så börjar vi med att göra en GLM-analys utan geografi i modellen. Det gör vi för att se hur de andra faktorerna kommer att påverkas och ändras när vi senare tar med geografin. Vi har som nämnts tidigare en modell med försäkringstagarens ålder, fordonets ålder samt bilklassning. Vi har delat in dessa argument i ganska stora grupper för att förenkla analysen. Ofta finns det ytterligare argument i modellen men det har vi har utelämnat här. Om vi ansätter en modell för att förklara skadefrekvensen får vi följande parameterskattningar

Premieargument	Grupp	Observerad skadefrekvens %	Parameterskattning
Försäkringstagarens ålder	1	13,6	1,00
Försäkringstagarens ålder	2	11,0	0,79
Försäkringstagarens ålder	3	8,2	0,60
Fordonets ålder	1	11,7	1,00
Fordonets ålder	2	10,6	0,90
Fordonets ålder	3	8,6	0,70
Fordonets ålder	4	5,9	0,49
Bilklassning	1	9,1	1,00
Bilklassning	2	12,9	1,48
Bilklassning	3	14,2	1,80

Tabell 4

Den observerade skadefrekvensen kan vara vilseledande om man inte tänker på att varje försäkring och skada finns med på tre olika rader i tabellen ovan, en observation som finns i cell (försäkringstagarens ålder=1,fordonets ålder=1,bilklassning=1) bidrar till den observerade skadefrekvensen för rad 1,5 samt 10.

GLM-analys med kredibilitetsskattningar

Vi fortsätter med att lägga till geografi och postnummer som en offset-variabel och arbetar efter den algoritm som vi tidigare gått igenom. Vi undersöker först hur snabbt vi får konvergens i den för att skatta variansparametern α och kan konstatera att vi uppnår det efter relativt få iterationer

Iteration	$\hat{\alpha}$
1	12,74
2	12,68
3	12,67
4	12,67
5	12,66

Tabell 5

För att få ett grepp om hur data ser ut så summerar vi över våra multiklassargument, och sorterar fallande med avseende på exponering (som mäts i försäkringsår). Då ser data ut enligt tabell 6 nedan

Postnummer **k**	Exponering $w_{\cdot k}$	Kredibilitetsfaktorer z_k	Kredibilitetsskattningar $\hat{u}_k$
1	2 372,87	0,95	0,67
2	1 592,58	0,92	0,95
3	1 571,39	0,92	0,58
.	.	.	.
.	.	.	.
269	539,94	0,80	0,69
270	539,53	0,80	1,24
271	538,42	0,81	0,93
.	.	.	.
.	.	.	.
10 492	0,01	0,00	1,00
10 493	0,01	0,00	1,00
10 494	0,01	0,00	1,00

Tabell 6

Vi ser att vi har några postnummer med nästan 0 i kredibilitetsfaktor, vilket beror på att vi har mycket låg exponering där, alltså väldigt få försäkrade bilar där de senaste åren. Det innebär att vi litar väldigt lite på den erfarenhet som vi har från det postnumret och mycket på genomsnittet vilket leder till att de får kredibilitetsskattningar som hamnar väldigt nära 1. Vi noterar även att postnummer $k = 2$, med mycket exponering, också får en kredibilitetsskattning nära 1, det är alltså inte bara de områden med låg exponering som skattas nära 1,00 utan även områden med hög exponering men som riskmässigt är som genomsnittet.

Höga kredibilitetsskattningar innebär höga risker och tvärtom. När vi har fått skattningarna delar vi in dem i 6 olika klasser, se tabell 7 nedan, med låga risker i geografiklass 1 och höga risker i geografiklass 6. Sedan betraktar geografiklass som ett premieargument tillsammans med de övriga tre i en GLM-analys och erhåller relationstal för geografiklasserna.

Att vi väljer sex olika klasser är en förenkling som vi gör det för att det blir lättare grafiskt att se. Vi har valt gränserna så att vi får ungefär lika många postnummer i de olika klasserna. I det här fallet så skulle vi egentligen föredra fler klasser än sex. Nu grupperar vi de med kredibilitetsskattningar 1,19 och 1,97 tillsammans, trots att det skiljer 66% mellan de olika skattningarna (1,97/1,19).

Kredibilitetsskattning	Geografiklass
0.01-0.81	1
0.82-0.91	2
0.92-0.98	3
0.99-1.06	4
1.07-1.18	5
1.19-1.97	6

Tabell 7

Diagram 1 och 2 visar hur kredibilitetsfaktorer z_k och kredibilitetsskattningar $\hat{u}_k$ fördelar sig.

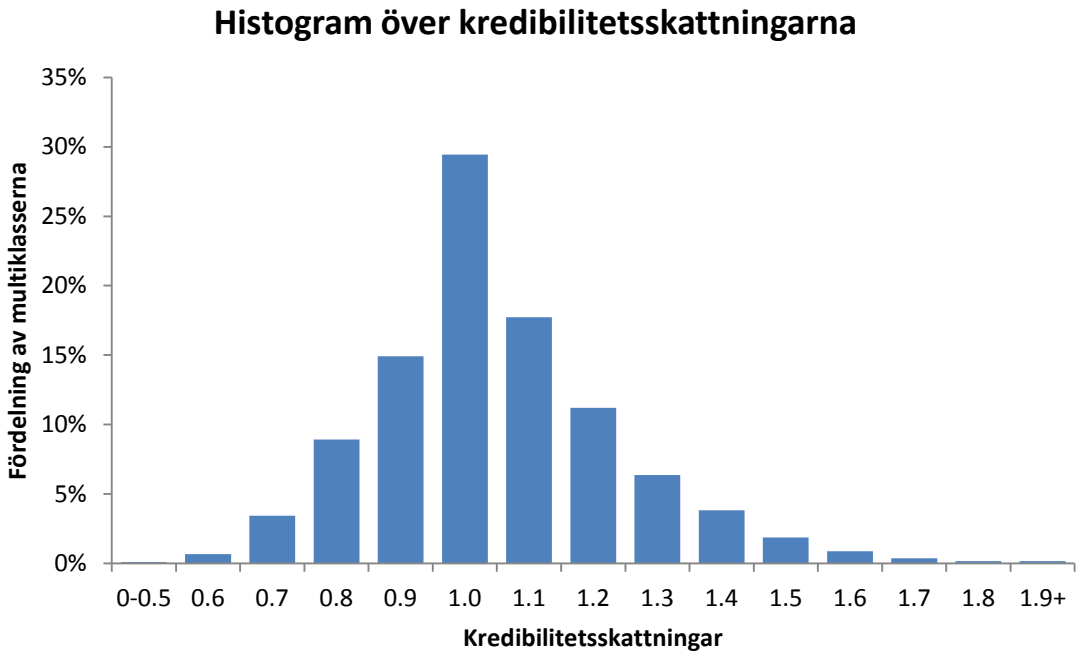


Diagram 1

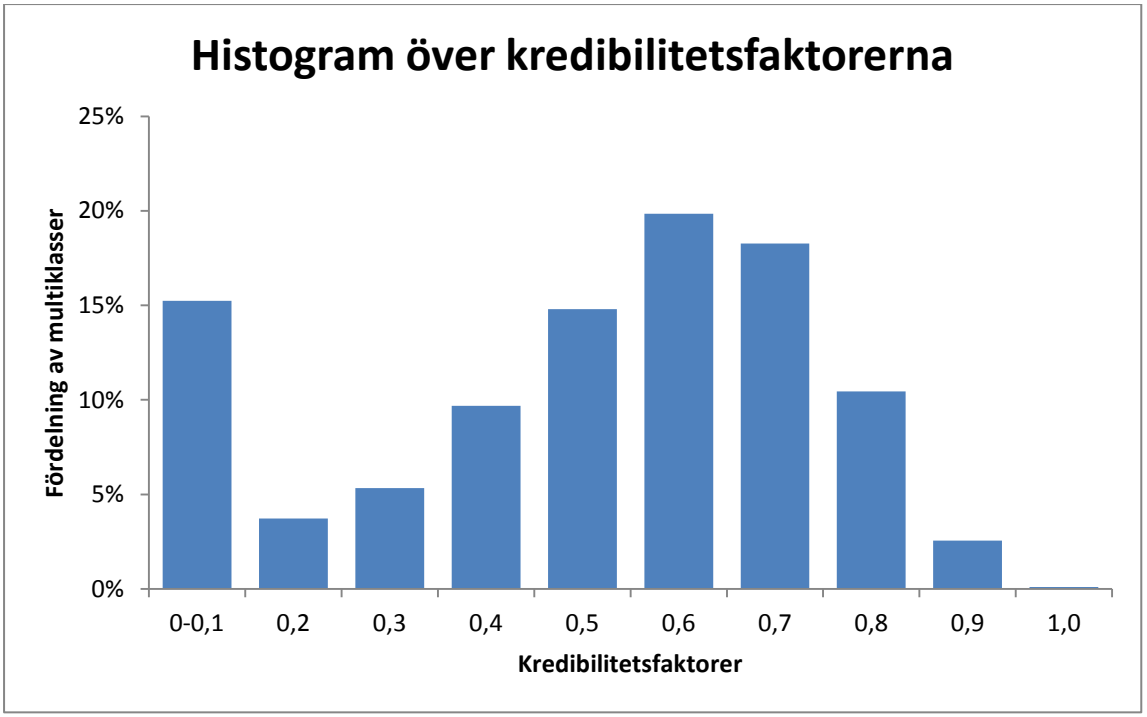


Diagram 2

I histogrammen är de postnummer där vi saknar data exkluderade – det motsvarar ca 5% av totala antalet postnummer. För dem får vi inte heller någon skattning i den föreslagna algoritmen. Ett alternativ är att sätta dessa till 1,00, men vi låter de vara utan skattningar tills vidare och håller reda på de områdena.

Som avslutande steg kör vi en GLM-analys med geografiklassen som ett nytt argument i modellen och jämför med de parameterskattningar som vi fick när vi inte hade med något som tog hänsyn till geografiska skillnader i modellen.

Premieargument	Grupp	Parameterskattning, utan geografi	Parameterskattning, med geografi
Försäkringstagarens ålder	1	1,00	1,00
Försäkringstagarens ålder	2	0,79	0,79
Försäkringstagarens ålder	3	0,60	0,60
Fordonets ålder	1	1,00	1,00
Fordonets ålder	2	0,90	0,92
Fordonets ålder	3	0,70	0,73
Fordonets ålder	4	0,49	0,52
Bilklassning	1	1,00	1,00
Bilklassning	2	1,48	1,48
Bilklassning	3	1,80	1,79
Geografiklass	1		1,00
Geografiklass	2		1,34
Geografiklass	3		1,56
Geografiklass	4		1,80
Geografiklass	5		2,11
Geografiklass	6		2,74

Tabell 8

Vi ser att det är relativt små förändringar i skattningarna för de parametrar som vi har haft med från början nu när vi lägger till geografin i modellen. Det är främst skattningarna för fordonets ålder som ändras.

Trots de små förändringarna så noterar vi att mellan högsta och lägsta geografiklassen skiljer det en faktor 2,74 riskmässigt för skadefrekvensen. Geografin fångar alltså till stor del upp något som inte tidigare förklarats av de övriga variablerna. Nästan 3 ggr så många skador förväntas i geografiklass 6 jämfört med geografiklass 1, förutsatt att de andra parametrar som är med i modellen är lika.

Den modell som vi ansätter är på formen $\mu_{ijkm} = \gamma_0\gamma_{1i}\gamma_{2j}\gamma_{3k}\gamma_{4m}$. Där γ_0 är en skalparameter som justerar den totala nivån och γ_{1i}, γ_{2j}, γ_{3k} och γ_{4m} är parameterskattningarna. Vi kommer inte att redovisa hur γ_0 förändras utan fokuserar på hur skattningarna för premieargumenten ändras.

För att visa vad vi menar så förtydligar vi med ett exempel.

E**xempel**

Vi har två kunder som båda har följande egenskaper

Premieargument	Grupp
Försäkringstagarens ålder	1
Fordonets ålder	2
Bilklassning	2

Det som skiljer dem åt är att kund 1 bor i geografiklass 2 medan kund 2 bor i geografiklass 4. Vi får då följande relationstal för våra kunder:

$$\text{Kund } 1 = f_{\text{Förs.ålder}} * f_{\text{Ford.ålder}} * f_{\text{Bilklass}} * f_{\text{Geo.klass}} = 1{,}00 * 0{,}92 * 1{,}48 * 1{,}34 = 1{,}82$$

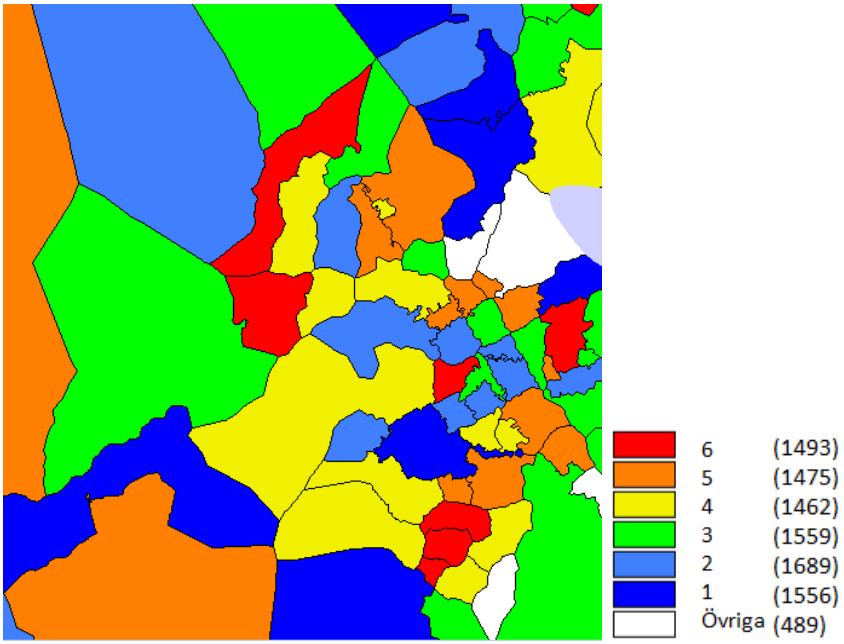
$$\text{Kund } 2 = 1{,}00 * 0{,}92 * 1{,}48 * 1{,}80 = 2{,}45$$

Vi ser att kund 2 är $2{,}45/1{,}82 = 1{,}34$ så hög risk jämfört med kund 1. Den här modellen säger alltså att det är 34% högre risk att kund 2 råkar ut för en skada jämfört med kund 1.

■

För att se hur det kan se ut geografiskt väljer vi ut ett område på kartan och zoomar in. Till höger om kartbilden visas indelningen av geografiklasserna samt inom parentes antalet postnummer i respektive grupp, områden som är vita är sådana som där vi saknar kredibilitetsskattningar.

Vi observerar att 489 postnummer saknar information, där vi i detta skede saknar skattningar.



Vi har också områden som är grannar och där det skiljer maximala 6 geografiklasser, det kan vara speciella omständigheter som råder och i somliga fall är riskskillnaderna så pass stora. Men det blir svårt att förklara för en kund som flyttar några hundra meter från det ena till det andra området och försäkringspremien för samma fordon förändras mycket uppåt eller nedåt. När förklaringen till det kan vara att vi har lite information i det ena området.

Vi har även många multiklasser med liten exponering vilket gör att de skattningarna inte är helt optimala heller. Kan vi göra någonting åt det?

Vi skall i efterföljande analyser se några alternativ som tittar på dessa bitar.

GLM-analys med kredibilitetsskattningar, utjämnade med medelvärde

Ett alternativ att minska stora slag mellan närliggande postnummer kan vara att vi istället för att enbart använda det observerade postnumrets kredibilitetsskattning, räknar ut ett medelvärde av både den observerade kredibilitetsskattningen samt de närmaste postnumrens skattningar. Vi viktar även med exponeringen så att en skattning med många årsrisker ger ett större bidrag än ett postnummer med få årsrisker.

Eftersom att vi har tillgång till koordinaterna för mittpunkten hos alla postnummer så kan vi räkna ut avståndet mellan dem. Vi använder oss av Pythagoras sats

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$d_{ij} = \text{avståndet mellan postnummer i och j}$$
$$(x_i, y_i) = \text{koordinaterför postnummer i}$$
$$(x_j, y_j) = \text{koordinaterför postnummer j}$$

Vi börjar med att testa fallet där vi använder information från de fem närmaste postnumren. Varför vi testar just de fem närmaste är för att vi vill bilda oss en uppfattning om vad konsekvensen blir. Det finns inte något statistiskt som styrker att det bör vara just fem. Vi använder det viktade medelvärdet från dem för att få en skattning och vi utgår från de kredibilitetsskattningar som vi fått fram ovan. Den nya skattningen för postnummer k, $\hat{u}_k^1$, får vi genom:

$$\hat{u}_k^1 = \frac{\hat{u}_k w_{.k} + \sum_i \hat{u}_i w_{.i}}{w_{.k} + \sum_i w_{.i}}$$

$$i = \text{postnummer k´s 5 närmsta grannar}$$
$$w_{.i} = \text{exponering postnummer i}$$
$$\hat{u}_i = \text{kredibilitetsskattning för postnummer i}$$

För de områden där vi saknar data från ett av de fem närmsta postnumren tar vi bara hjälp av de fyra närmsta istället, om det saknas hos två av dem så tar vi hjälp av tre osv.

Precis som innan så genomför vi en GLM-analys, och med samma indelningar på geografiklasserna som tidigare (tabell 7). Men istället för att bara använda postnumrets egna skattning använder vi alltså det viktade (med avseende på exponering) medelvärdet för att dela in dem i olika grupper.

Premieargument	Grupp	Parameterskattning, med geografi	Parameterskattning, med geografi, utjämnade
Försäkringstagarens ålder	1	1,00	1,00
Försäkringstagarens ålder	2	0,79	0,79
Försäkringstagarens ålder	3	0,60	0,60
Fordonets ålder	1	1,00	1,00
Fordonets ålder	2	0,92	0,92
Fordonets ålder	3	0,73	0,73
Fordonets ålder	4	0,52	0,53
Bilklassning	1	1,00	1,00
Bilklassning	2	1,48	1,47
Bilklassning	3	1,79	1,79
Geografiklass	1	1,00	1,00
Geografiklass	2	1,34	1,26
Geografiklass	3	1,56	1,47
Geografiklass	4	1,80	1,67
Geografiklass	5	2,11	1,94
Geografiklass	6	2,74	2,39

Tabell 9

Vi ser återigen att de övriga parametrarna inte förändras nämnvärt. Att vi får en mindre spridningen av skattningarna för geografiklasserna är väntat när vi jämnar ut med medelvärdet. Några som tidigare låg högt har sänkts och vice versa.

Det visas också i diagram 3 nedan att vi får en mindre spridning för våra skattningar, en större andel fördelas centrerat runt 1,0.

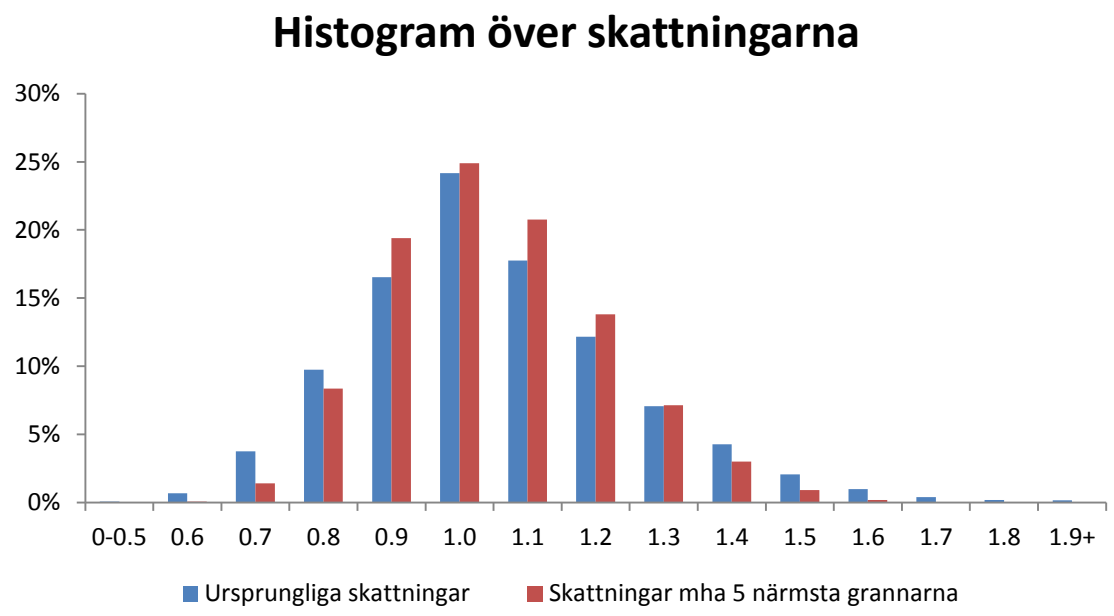
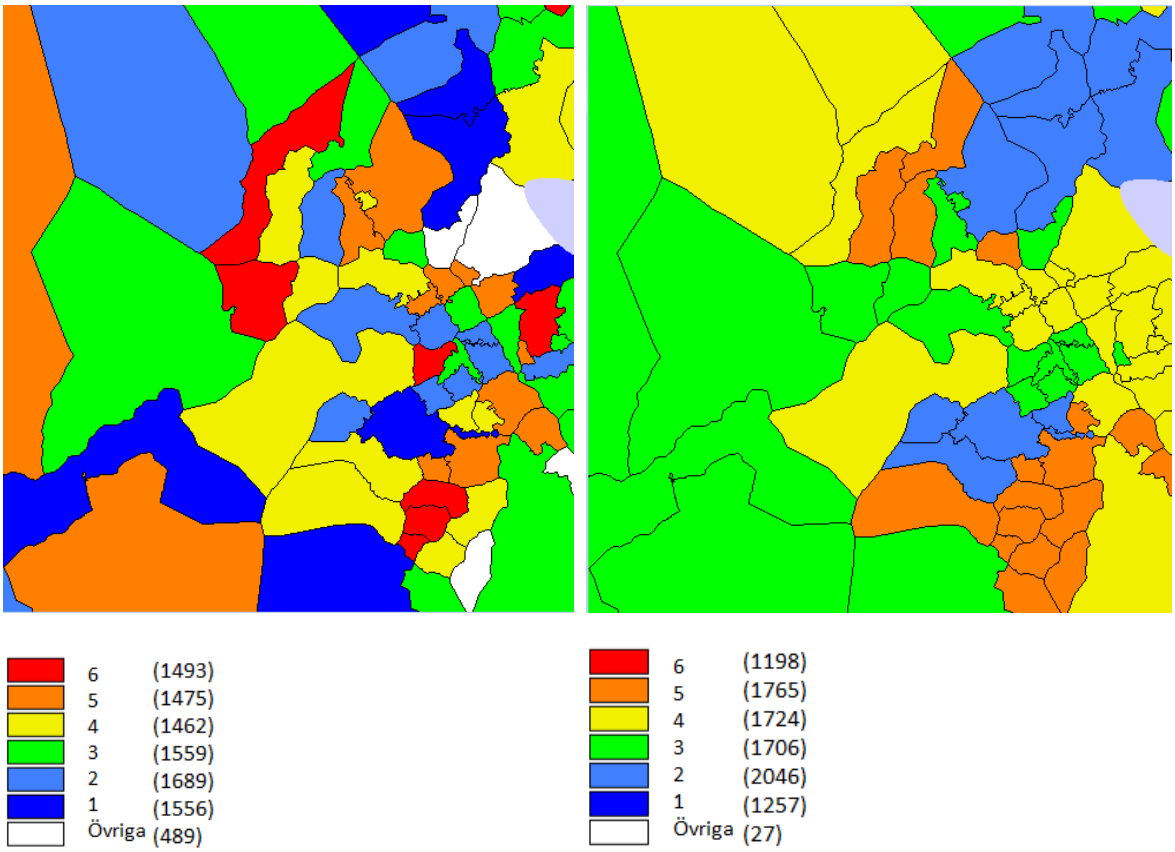


Diagram 3

För att se hur det grafiskt blir tittar vi på samma område som tidigare och jämför. Till vänster ser vi den ursprungliga analysen och till höger den senare med utjämnade värden. Vi ser att den högra bilden visar jämnare övergångar, samt att vi inte får lika stora skillnader mellan närliggande postnummer. I bilden till vänster finns det områden som angränsar till varandra där den ena ligger i lägsta riskklassen och den andra i högsta, det kommer vi bort från på den högra bilden.



Ursprungliga skattningar *Skattningar med hjälp av de fem närmsta grannarna*

Är det alltid vi vill jämna ut dessa skillnader, eller kan det finnas skäl att låta den historik och den erfarenhet som vi har om områdena att vara kvar? Vi skall se några andra alternativ som vi kan använda oss av, varianter ger en mindre utjämning.

Noterbart är att antalet områden där vi saknar skattningar minskar från 489 till 27. De 27 områden som fortfarande saknar skattningar är områden där det inte finns någon data varken från de själva eller från de närmsta grannarna.

GLM-analys med kredibilitetsskattningar, utjämnade med hänsyn till avstånd

Att med hjälp av medelvärdet från de fem närmsta postnumren korrigera den observerade skattningen kan vara lämpligt i vissa områden i Sverige men mindre bra i andra. En egenskap som är önskvärd är att den information vi hämtar från närliggande postnummer avtar med avståndet.

Sveriges postnummerområden ser väldigt olika ut beroende på var i landet vi är. I storstäderna är det väldigt nära till det närmsta närliggande området och i norra Sverige kan det vara väldigt långt. För att få en blick över det skapar vi en tabell som vi sorterar så att vi överst får det postnummer som har kortast avstånd till närliggande postnummer samt även inom det postnumret sortera i fallande ordning.

Postnummer, k	Närliggande postnummer nr	Avstånd, d_{ki} (meter)
1	1	26,76
1	2	141,61
1	3	307,48
.	.	.
.	.	.
4 474	1	811,61
4 474	2	837,63
4 474	3	905,00
.	.	.
.	.	.
9 723	1	46 620,25
9 723	2	50 569,83
9 723	3	62 737,16
.	.	.
.	.	.

Tabell 10

Vi ser att vi har en stor spridning över hur nära det är till närmaste grannområde, från knappa 27 meter upp till dryga 4,6 mil och även ännu längre. Det medför att metoden att jämna ut mellan de fem närmsta grannarna får väldigt olika konsekvenser beroende på avståndet mellan områdena. Vi borde dock ta någon sorts hänsyn till avståndet.

Vi ställer oss också frågan hur mycket vikt vi skall fästa vid det observerade värdet för postnummer k och hur mycket information som vi skall hämta från de kringliggande postnumren. Det rimliga är att även det varierar med mer än enbart exponeringen. Om vi exempelvis har två postnummer som ligger nära varandra, där vi i båda områdena har ganska mycket exponering är det rimligt att lita mer på den statistik som vi känner till för det postnumret som vi fokuserar på och inte direkt vikta ihop det med exponeringen. Vi vill skriva skattningen $\hat{u}_k$ på formen

$$\hat{u}_k = v_k \hat{u}_k + (1 - v_k) \frac{\sum_i f(d_{ki}) w_{.i} \hat{u}_i}{\sum_i f(d_{ki}) w_{.i}}$$

$$0 \leq v_k \leq 1$$

Vi har samma notation som innan och har nu även infört v_k för den vikt som vi kommer att lägga vid kredibilitetsskattningen som vi har erhållt för postnummer k. Det är rimligt att denna vikt baseras på hur mycket information som vi har just där, exponeringen. Från uträkningen av de ursprungliga kredibilitetsskattnignarna har vi erhållit vikter baseras på delvis exponeringen, nämligen kredibilitetsfaktorn z_k

$$z_k = \frac{\tilde{w}_{\cdot k}}{\tilde{w}_{\cdot k} + \sigma^2 / \sigma_U^2}$$

För områden där vi saknar exponering får vi $z_k = 0$, vilket även det är en egenskap som vi vill att vår vikt skall ha, låter vi som utgångspunkt $v_k = z_k$ .Tabell och diagram över hur z_k och $w_{\cdot k}$ förhåller sig till varandra finns i Appendix.

Vi vill även ha en funktion $f(d_{ki})$ som avtar med avståndet. Vi känner till några avtagande funktioner, en är exponentialfunktionen. Vi skulle kunna använda någon annan avtagande funktion. Vi provar funktionen på formen $e^{-td_{ki}}$.

Funktionen har egenskapen att ta värdet 1,00 vid $d_{ki} = 0$ och avtar sedan olika snabbt beroende på värdet på konstanten t. Vi tänker testa några olika varianter och ser hur det påverkar analysen, vi väljer

$$t_1 = 0{,}05$$
$$t_2 = 0{,}70$$
$$t_3 = 1{,}50$$

Vi kan då skriva ekvationerna enligt

$$\hat{u}_k^2 = z_k \hat{u}_k + (1 - z_k) \frac{\sum_i f_1(d_{ki}) w_{\cdot i} \hat{u}_i}{\sum_i f_1(d_{ki}) w_{\cdot i}}$$

$$\hat{u}_k^3 = z_k \hat{u}_k + (1 - z_k) \frac{\sum_i f_2(d_{ki}) w_{\cdot i} \hat{u}_i}{\sum_i f_2(d_{ki}) w_{\cdot i}}$$

$$\hat{u}_k^4 = z_k \hat{u}_k + (1 - z_k) \frac{\sum_i f_3(d_{ki}) w_{\cdot i} \hat{u}_i}{\sum_i f_3(d_{ki}) w_{\cdot i}}$$

$$d_{ki} = \text{avståndet mellan postnummer k och i (i km)}$$
$$i = \text{postnummer k´s 5 närmsta grannar}$$
$$f_1(d_{ki}) = e^{-0{,}05 d_{ki}}$$
$$f_2(d_{ki}) = e^{-0{,}70 d_{ki}}$$
$$f_3(d_{ki}) = e^{-1{,}5 d_{ki}}$$
$$w_{\cdot i} = \text{exponering i postnummer i}$$
$$\hat{u}_i = \text{kredibilitetsskattning för postnummer i}$$

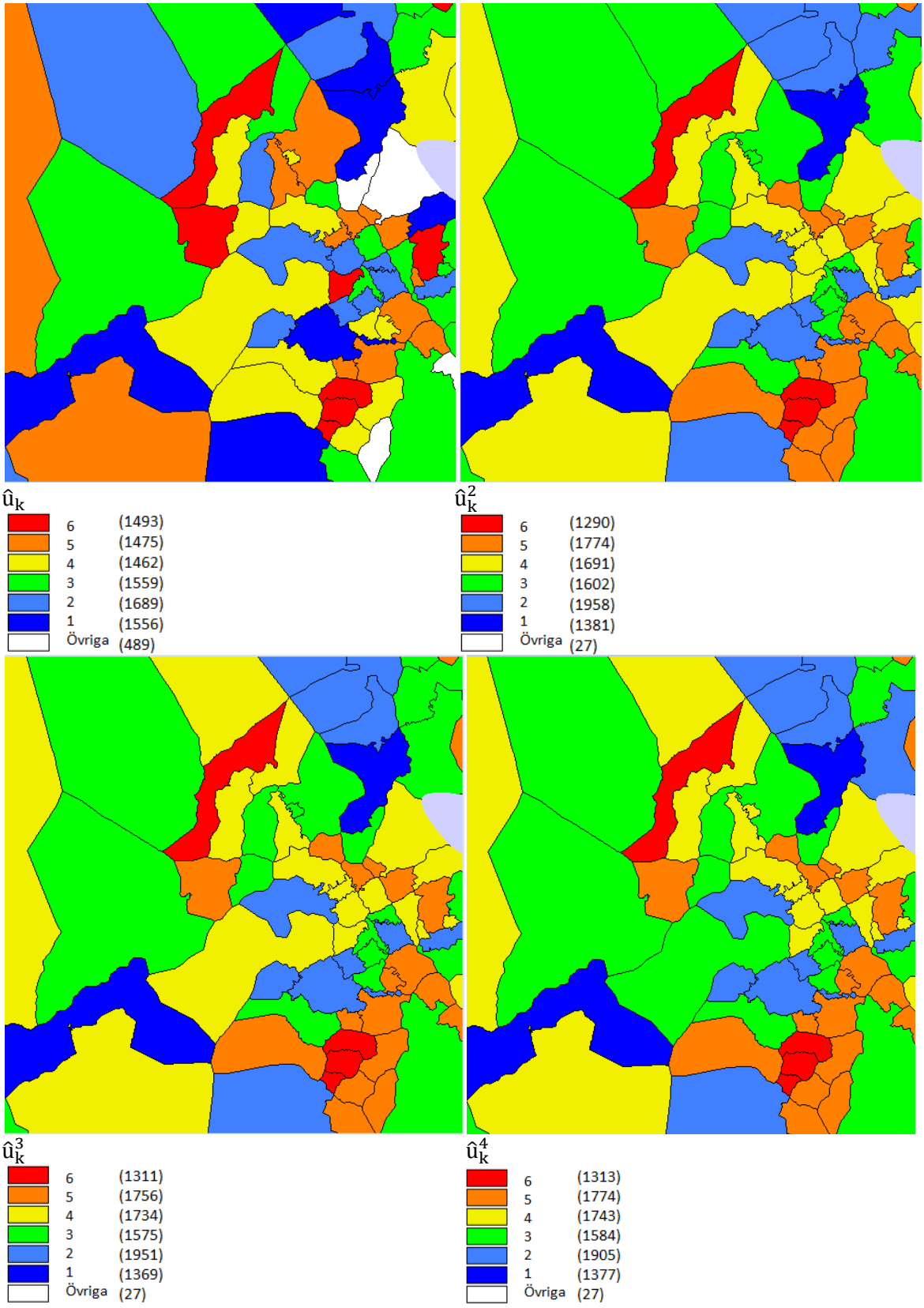
För att se hur funktionerna $f_1(d_{ki}) - f_3(d_{ki})$ ser ut se appendix diagram 3.

Vi börjar med att se hur parameterskattningarna ändras och vi ser att det är marginellt, det beror delvis på att vi har ganska grova indelningar, försäkringstagarens ålder har ofta intervall på ett år och är inte indelade i tre grupper.

Premieargument	Grupp	Parameter-skattning, $\hat{u}_k$	Parameter-skattning, $\hat{u}_k^2$	Parameter-skattning, $\hat{u}_k^3$	Parameter-skattning, $\hat{u}_k^4$
Försäkringstagarens ålder	1	1,00	1,00	1,00	1,00
Försäkringstagarens ålder	2	0,79	0,79	0,79	0,79
Försäkringstagarens ålder	3	0,60	0,60	0,60	0,60
Fordonets ålder	1	1,00	1,00	1,00	1,00
Fordonets ålder	2	0,92	0,92	0,92	0,92
Fordonets ålder	3	0,73	0,74	0,74	0,74
Fordonets ålder	4	0,52	0,53	0,53	0,53
Bilklassning	1	1,00	1,00	1,00	1,00
Bilklassning	2	1,48	1,47	1,47	1,47
Bilklassning	3	1,79	1,79	1,79	1,79
Geografiklass	1	1,00	1,00	1,00	1,00
Geografiklass	2	1,34	1,32	1,32	1,31
Geografiklass	3	1,56	1,57	1,57	1,56
Geografiklass	4	1,80	1,82	1,81	1,81
Geografiklass	5	2,11	2,14	2,13	2,12
Geografiklass	6	2,74	2,75	2,73	2,72

Tabell 11

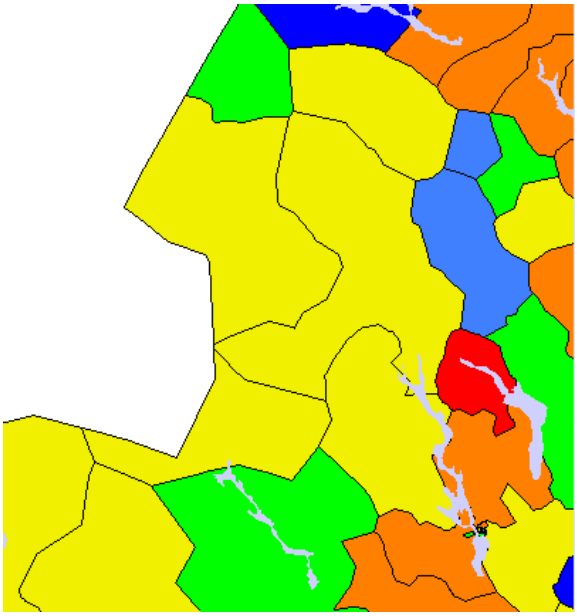
Följande grafer visar hur de olika viktningsmetoderna påverkar geografiklasserna inom samma geografiska område som vi betraktade tidigare.



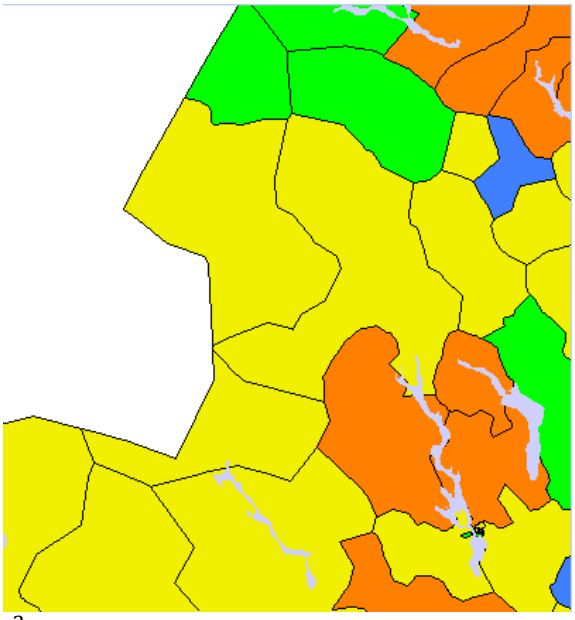
Jämfört med den icke utjämnade geografianalysen så ser parameterskattningarna väldigt lika ut. Däremot ser vi på kartbilderna att $\hat{u}_k^2$ och $\hat{u}_k^4$ skiljer sig mot $\hat{u}_k$, det är ganska många områden som har bytt klasser och det är mer utjämnat. Men mellan $\hat{u}_k^2$ och $\hat{u}_k^4$ skiljer det inte mycket, d.v.s. det är några få områden som har bytt geografiklass. En förklaring till det är att vi har ganska stora geografiklasser,

de täcker stora intervall för våra skattningar. En annan är att vi fokuserar på ett område där närliggande postnummer ligger ganska nära varandra. Det gör att funktionen $f(d_{ik})$ inte ”hinner avta” så mycket och därmed blir dess påverkan på hur mycket information som hämtas från övriga postnummer inte så stor.

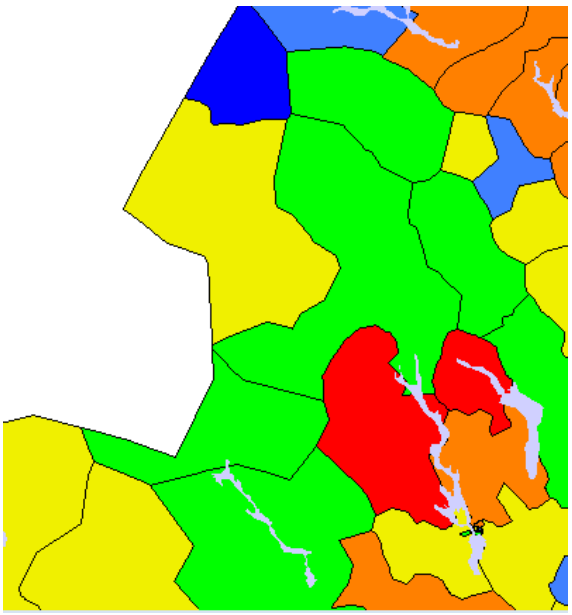
Tittar vi på områden där avståndet till närliggande postnummer är längre ser vi att värdet på t spelar större roll, ju högre värde på t desto snabbare avtar funktion och därmed även hur mkt information vi hämtar från grannarna.



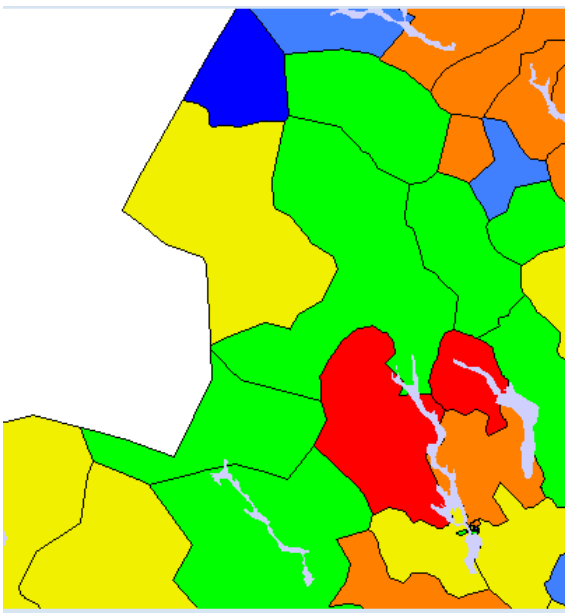
$\hat{u}_k$



$\hat{u}_k^2$



$\hat{u}_k^3$



$\hat{u}_k^4$

Att jämna ut den skattning som vi får från den föreslagna algoritmen med hjälp av information från närliggande områden är en metod som vi föreslår. Det är också rimligt att hur mycket information som vi hämtar in avtar när avståndet dit ökar. Vi har sett att Sveriges postnummerområden skiljer sig åt väldigt mycket när det gäller avstånd till andra, det medför att det är svårt att anpassa en och samma avtagande funktion som passar i hela landet.

Det finns några olika aspekter att tänka på när det gäller multiklassargument av den här typen, den primära frågan som vi vill lösa är- Hur förklarar vi skaderisken bäst?

För att få fram den modellen kan man plocka bort en delmängd av sitt dataunderlag innan man börjar att analysera och därmed inte låta det påverka modellen. När ett förslag till modell sedan är klar kan man applicera det på den datamängd som inte varit med i analysen och se hur väl den predicerar utfallet. Ett sätt att få fram rätt funktion är då att välja den modell som predicerar utfallet bäst.

Det krävs en kontinuerlig uppdatering av den här typen av argument. Dels för att när vi tittar på så pass små områden krävs det inte så mycket exponering i ett visst område för att de initiala skattningarna ska påverkas. Dels även för att det händer saker i vår omvärld som vi vill fånga upp. Det kan exempelvis vara att hyreshus som rivs och ersätts av villor och på sikt på kan det vara ett helt annat område än tidigare.

Utöver att få en modell som förklarar skaderisken så är det också viktigt att ha en konkurrenskraftig premie. Marknaden för försäkringar är konkurrensutsatt och premiens storlek spelar en stor roll när en försäkring ska tecknas. Sidor på nätet som jämför premier och villkor ökar i antal och det blir ännu viktigare att ha en uppdaterad prissättning, bland annat hur premiernas storlek varierar med avseende på geografi. Det kan finnas vissa geografiska områden som är högre prioriterade än andra och där är det ännu viktigare att ha en fördelaktig premie. Det är otroligt svårt att via en formel göra de justeringarna, utan där krävs det nog manuella kontroller och korrigeringar.

Diskussion

Att hantera multiklassargument för prissättning är inte helt enkelt och det kan ske på många olika sätt, vi har visat några. Att skatta multiklassargumenten enligt den metod som Ohlsson & Johansson [1] föreslår är grundarbetet och det är viktigt att det görs korrekt. När det sedan är gjort återstår en del andra frågor som vi har belyst:

- Hur skall vi hantera postnummer där vi har liten erfarenhet och får skattningar nära 1,00?
- Vad skall vi göra med postnummer där vi saknar tidigare data helt?
- Har kan vi skapa skattningar som är logisk, både för försäkringsbolag samt kunderna?

Vi har sett några alternativ som till viss del hanterar dessa frågorna, men inte fullt ut.

Vi har sett på kartbilder att kredibilitetsskattningarna mellan närliggande postnummer kan variera väldigt mycket. I vissa fall är det så vi vill att prissättningen ska vara och det är riskmässigt motiverat, men i andra fall kan det vara slumpen som spelar in. Den metod som vi har tittat på har fokuserat på att vi använder dels den informationen som vi har från det observerade postnumret och dessutom har vi tagit hjälp av det vi känner till hos närliggande postnummer. Vi har valt att begränsa oss och tittat på de 5 närmaste områdena men det är inte något som säger att just 5 är optimalt. Beroende på vad ändamålet är med den här typen av prissättning är så kan det vara olika metoder som passar olika bra.

Vi har på olika sätt analyserat hur en funktion som avtar med avståndet kan användas för att vikta hur mycket vi skall lita på de områden som ligger ”nära”. Vad avståndet till närmaste postnummer är varierar väldigt mycket beroende på vart i Sverige vi befinner oss, det gör att det blir svårt att använda en och samma funktion som fungerar i hela landet.

Vilken metod vi använder i olika delar av landet är även beroende av andra faktorer, det kan exempelvis vara input från de som arbetar med försäljning som indikerar att vi har svårt att konkurrera i olika geografier. Vi kanske ser att vi har en nedåtgående trend beståndsmässigt i vissa geografiska delar av landet och tror att för högt pris kan vara en orsak.

Syftet med arbetet är att visa olika sätt att på ett maskinellt sett få en grund att utgå ifrån, men även att det inte är klart efter det. Grundskattningar som kombinerar den tidigare erfarenhet riskmässigt som vi har från området samt även vissa andra aspekter. Det är svårt att helt maskinellt få till en geografiöversyn på ett bra sätt. Manuell arbete, input, och kontroller behövs också.

Det här arbetet har fokuserat på geografi som en multiklassvariabel, premieklassningen är även den en multiklass som kan hanteras på liknande sätt. Där kan vi istället för närliggande postnummer titta på liknande bilmodeller. Har man liten erfarenhet om en specifik modell av något märke kan en utgångspunkt vara att den modellen påminner om en liknande modell av samma märke som vi har mer erfarenhet om.

Appendix

Postnummer	Kredibilitetsfaktor z_k	Exponering $w_{\cdot k}$
1	0,95	2 372,87
2	0,92	1 571,39
3	0,92	1 592,58
4	0,91	1 259,23
.	.	.
.	.	.
4 091	0,56	180,08
4 092	0,56	179,41
4 093	0,56	166,58
4 094	0,56	162,10
.	.	.
.	.	.
9 001	0,03	3,32
9 002	0,03	2,38
9 003	0,03	3,63
9 004	0,03	3,76
.	.	.
.	.	.

Tabell 11

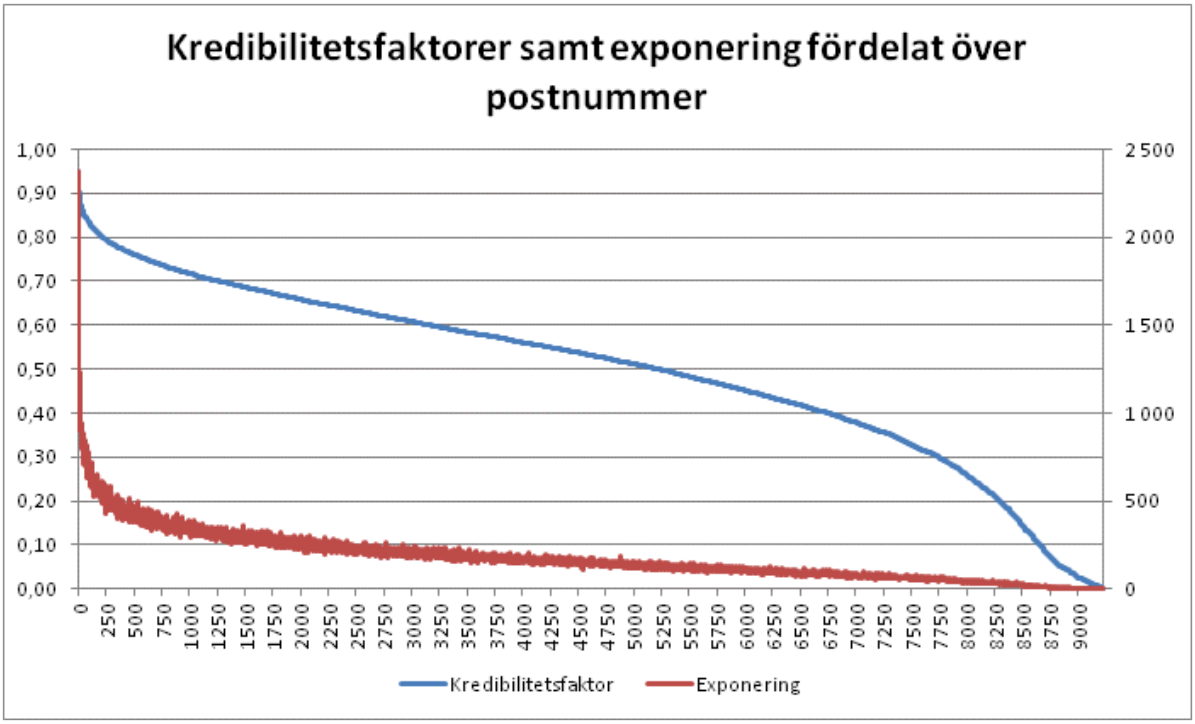


Diagram 4
(Kredibilitetsfaktorer på vänster axel och exponering på höger axel)

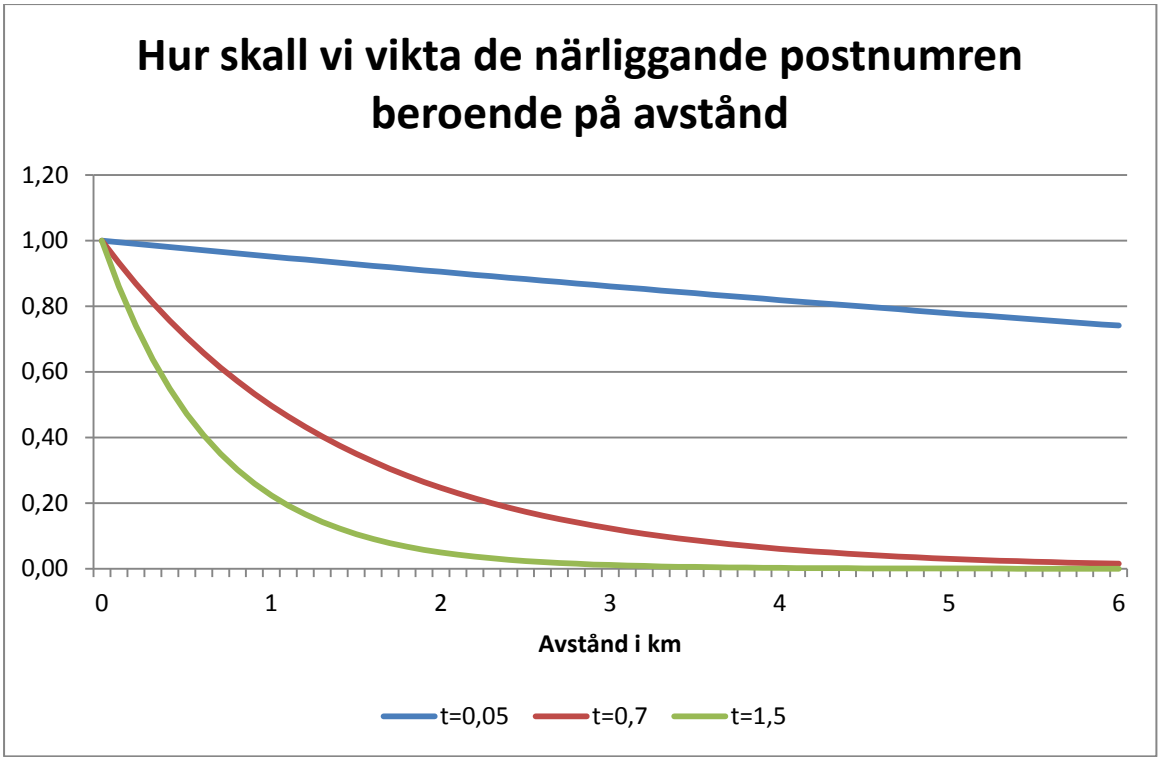

Hur skall vi vikta de närliggande postnumren beroende på avstånd
1,20
1,00
0,80
0,60
0,40
0,20
0,00
0
1
2
3
4
5
6
Avstånd i km
t=0,05
t=0,7
t=1,5

Diagram 5

Referenser

[1] E Ohlsson, B Johansson *Prissättning inom Sakförsäkring med Generaliserade linjära modeller*, version 4.0 Augusti 2006

Stockholms universitet/Stockholm University
SE-106 91 Stockholm
Telefon/Phone: 08 – 16 20 00
www.su.se

Stockholms universitet