Stockholms universitet

Analys av röd kängurupopulation med thin-plate splines

Daniel Sadeghei Fazel

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:25
Matematisk statistik
September 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:25**
http://www.math.su.se

Analys av röd kängurupopulation med thin-plate splines

Daniel Sadeghei Fazel*

September 2015

Sammanfattning

I denna uppsats bekantar vi oss med thin-plate splines teorin för att sen tillämpa den på ett dataset. Datasetet har samlats genom observation från ett litet flygplan både på vänster och höger sida uppdelat i par och består av antalet röd kängurun per km^2 som en funktion av *longitud*, *latitud* och *år* från Sydaustralien, men vi fokuserar på vänsterdata. Vi tittar på tre olika modeller med och utan samspel för att undersöka hur den totala populationen över området varierar med tiden, men också för att undersöka ifall populationstätheten varierar som en funktion av koordinaterna över tiden. Höger data används slutligen för att undersöka modellernas prediktionsförmåga.
Resultatet är att de tre modellerna genererar årsvisa populationsstorlekar som ligger inom 90% konfidensintervall av varandra, att samspel mellan koordinat och år verkar finnas och att det finns en tendens för en god prediktionsförmåga.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: danne.fazel@gmail.com. Handledare: Martin Sköld.

Abstract

In this thesis we get familiar with thin-plate theory and then we apply it on a dataset. The dataset has been collected through observation from a small airplane, both on the left and the right side divided in pair and consists of the number of red kangaroos per km^2 as a function of *longitud*, *latitud* and *year* from South Australia, but we focus on left data. We look at three different models with and without interaction to examine how the total population over the area varies with time, but also to examine whether the population density varies as a function of coordinates over time. Finally right data is used to examine the models prediction ability.

The result is that the three models generates yearly populations that lies within 90% confidence interval of each other, furthermore there seems to be interaction between *coordinates* and *year* and that there is a tendency for good prediction.

Förord

Denna uppsats utgör ett självständigt arbete om 15 hp vilket leder till en kandidatexamen i matematisk statistik vid Stockholms Universitet.

Jag skulle vilja tacka min handledare Martin Sköld för uthållighet och all vägledning, och jag vill också tacka Tom Britton för värdefull rådgivning.

Innehåll

1 Inledning

Det har egentligen alltid varit av intresse för människan att bilda sig en uppfattning om djurlivets populationsstorlek och variation, inte minst för utrotningsrisken, men först på senare tid har det varit möjligt att göra tillförlitliga skattningar. I denna uppsats intresserar vi oss för populationen av röd kängurun i Sydaustralien.

Syftet med uppsatsen är till stor del att bekanta sig med thin-plate splines teorin, men också att tillämpa den på ett dataset för att undersöka den årsvisa variationen i populationen, vidare undersöker vi ifall populationstätheterna varierar i förhållande till varandra som en funktion av koordinater över åren och slutligen tittar vi på prediktionsförmågan.

2 Teori

Modeller som vi använder i denna uppsats tillhör en mera allmän klass av modeller, nämligen generaliserade additiva modeller (GAM). Då denna klass av modeller är en generalisering av generaliserade linjära modeller (GLM), definieras denna först innan vi definierar GAM.

2.1 Generaliserade linjära modeller (GLM)

Generaliserade linjära modeller är en klass av modeller för respons variabeln Y och består av tre olika komponenter:

1. **Slumpkomponenten**

   Låt $\{Y_1, ..., Y_n\}$ beteckna observationerna av Y. Varje komponent tilldelas en täthets- eller sannolikhetsfunktion som tillhör exponentiala spridningsfamiljen, vars fördelning har formen,

   $$f(y_i; \theta_i, \phi) = exp([y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)), \qquad i = 1, ..., n$$

   där θ_i kallas den naturliga parametern, ϕ spridnings parametern. Vanligtvis har $a(\phi)$ formen $a(\phi) = \phi/w_i$, där w_i är en känd vikt. [1]

2. **Systematisk komponent**

   Systematiska komponenten kopplar ihop vektorn $\eta_1, ..., \eta_n$ med förklarande variabler genom en linjär model. Låt x_{ij} beteckna värdet av prediktor j (j=1,2,..,p). Då gäller

   $$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, ..., n.$$

   Denna linjära kombination av förklarande variabler kallas för en linjär prediktor.

[1] Se [1] sid 133.

3. **Länkfunktionen**

   Länkfunktionen kopplar ihop slumpkomponenten med systematiska komponenter. Låt $\mu_i = E(Y_i), i = 1, ..., n$. Kopplingen sker genom $\eta_i = g(\mu_i)$, där länkfunktionen g är monoton och deriverbar funktion. Länkfunktionen $g(\mu) = \mu$ kallas identitetslänken, och är länkfunktionen för ordinär regression med normalfördelad Y. [2]

2.2 Generaliserade additiva modeller (GAM)

Generaliserade additiva modeller är generaliserade linjära modeller med en linjär prediktor innehållande minst en slät funktion, vilket också sägs vara icke-parametrisk. I allmänhet kan prediktorn exempelvis ha följande struktur

$$g(\mu_i) = \mathbf{X}_i^*\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + ... \tag{2.2.1}$$

där $\mu_i = E(Y_i)$ och Y_i tillhör någon fördelning från exponentiala spridningsfamiljen. Y_i är respons variabel, $\boldsymbol{X}_i^*$ en radvektor av modellmatrisen med strikt parametriska komponenter, $\boldsymbol{\theta}$ korresponderande parameter vektor, och f_j är släta funktioner av kovariater. [3]

2.3 Negativ Binomial fördelning

Negativ binomial fördelning har följande sannolikhetsfunktion på exponential form:

$$\begin{aligned} f(y; k, \mu) &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y \\ &= exp(y \log\frac{\mu}{\mu+k} + k \log\frac{k}{\mu+k} + \log\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}) \\ &= exp(y\theta + k \log(1 - exp(\theta)) + c(y, k)) \qquad y = 0, 1, 2, ... \end{aligned}$$

där $\theta = log\frac{\mu}{\mu+k}$, $a(\phi) = 1$, $b(\theta) = -k\, log(1 - exp(\theta))$ och $c(y, k) = log\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}$. Väntevärde och varians är $E(Y) = \mu$, $Var(Y) = \mu + \mu^2/k$, där k är form parametern, som måste antas vara känt för att ovanstående ska vara på exponential form. Men oftast är den inte känd och behöver skattas. [4]

2.4 Deviance och Akaikes informationskriterium

Deviance kan ha olika betydelser men här tänker vi oss att vi har en mättad modell med lika många parametrar, n, som antalet unika datakombinationer och en alternativ inkapslad modell med färre parametrar p. Deviancen beräknas då på följande sätt:

[2] Se [1] sid 116-117.
[3] Se [2] sid 163.
[4] Se [1] sid 131.

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

där $L(\boldsymbol{y}; \boldsymbol{y})$ är log-likelihood med väntevärde $\boldsymbol{y}$ och $L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ är log-likelihood med anpassad väntevärde $\hat{\mu}$. En skalad deviance definieras som $\boldsymbol{D^*(y, \mu)} = \boldsymbol{D}(\boldsymbol{y}, \boldsymbol{\mu})/\phi$, och har en asymptotisk chi-kvadrat fördelning hos fördelningar i GLM med $n - p$ frihetsgrader, d.v.s. $\boldsymbol{D^*} \sim \chi^2(n - p)$. Skalad deviance kan användas för att avgöra om den enklare modellen är tillräckligt bra givet en signifikansnivå α. Testet kallas för likelihood-ratio test, och den enklare modellen är tillräckligt bra om $\boldsymbol{D}^* \leq x_{1-\alpha}$ d.v.s. icke signifikant under nollhypotesen, där $x_{1-\alpha}$ är $(1 - \alpha)$-kvantilen.

Akaikes informationskriterium (AIC) är ett relativt mått för att jämföra modeller med varandra i syfte att komma fram till en enkel modell med få parametrar och förklarande variabler. AIC bygger på log-likelihood, men straffar för fler frihetsgrader och är definierad som följer:

$$AIC = 2[-L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) + p].$$

Ju mindre AIC är desto bättre, men AIC säger ingenting om hur bra modellen i sig är, utan är endast ett relativt mått mot andra modeller.

2.5 Full thin-plate splines

Full thin-plate splines är en spline baserad metod för att skatta en slät funktion av en eller flera variabler. Vi gör det nu enkelt för oss när vi väljer en linjär modell för att introducera full thin-plate splines.

Låt $f(\mathbf{x})$ vara en slät funktion som vi vill skatta från n observationer $(y_i, \mathbf{x}_i)$ Låt vidare $y_i = f(\mathbf{x}_i) + \epsilon_i$, där ϵ_i är oberoende slumptermer och $\mathbf{x}$ är en d-dimensionell vektor $(n \geq d)$. Thin plate splines används här för att skatta funktionen f genom att minimera med avseende på g:

$$||\mathbf{y} - \mathbf{g}||^2 + \lambda J_{md}(g) \tag{2.5.1}$$

där $\mathbf{y}$ är vektorn av y_i, $\mathbf{g} = (g(\mathbf{x}_1), (g(\mathbf{x}_2), ..., (g(\mathbf{x}_n))$, $J_{md}(g)$ är en straffterm och λ kontrollerar trade-off mellan hur väl anpassad data är till funktionen och hur pass linjär funktionen av g är. Ju större värde på λ för en slät funktion av en variabel, desto mer linjär blir funktionen, medans ett litet λ oftast medför en funktion som 'slingrar' sig så att den är välanpassad till data. Straff termen är definerad som:

$$J_{md}(g) = \int ... \int_{\mathbb{R}^d} \sum_{\nu_1 + ... + \nu_d = m} \frac{m!}{\nu_1! ... \nu_d!} \left(\frac{\partial^m g}{\partial x_1^{\nu_1} ... \partial x_d^{\nu_d}} \right)^2 dx_1 ... dx_d. \tag{2.5.2}$$

Givet att $2m > d$, så kan man visa att det g som minimerar (2.5.1) har formen

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(||(\boldsymbol{x} - \boldsymbol{x}_i)||) + \sum_{j=1}^{M} \alpha_j \phi_j(\boldsymbol{x}) \tag{2.5.3}$$

där $\boldsymbol{\delta}$ och $\boldsymbol{\alpha}$ är okända parametrar med bivillkoren $\boldsymbol{T'\delta = 0}$ och $T_{ij} = \phi_j(\boldsymbol{x}_i)$, där $\boldsymbol{T}$ är en $n \times M$-matris. Vidare gäller att $M = \binom{m+d-1}{d}$ och $\{\phi_i \ 1 \leq i \leq M\}$ är linjärt oberoende polynomer som spänner rummet av polynom av grad lägre än m, i $\mathbb{R}^d$. Det gäller att

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} log(r) & \text{om d jämn} \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & \text{om d udda} \end{cases}$$

Definiera en $n \times n$-matris $\boldsymbol{E}$ med $E_{ij} = \eta_{md}(||\boldsymbol{x}_i - \boldsymbol{x}_j||)$ då löser vi problemet med följande:

$$\text{minimera } ||\boldsymbol{y - E\delta - T\alpha}||^2 + \lambda \boldsymbol{\delta' E \delta} \text{ med bivillkoret } \boldsymbol{T'\delta = 0} \tag{2.5.4}$$

med avseende på $\boldsymbol{\delta}$ och $\boldsymbol{\alpha}$

2.6 Låg rang thin-plate spline smooths

Då full thin-plate splines kräver lika många parametrar som datapunkter, är det en långsam metod för en stor dataupsättning. Man använder istället en trunkerad bas, som gärna stör lösningen så litet som möjligt. En ideal bas är en som leder till en minimal förändring av både goodness-of-fit och strafftermen, för varje $\boldsymbol{\delta}$. Vi ska härleda en optimal trunkerad bas för full thin-plate splines, och dessa splines kallas då thin-plate regression splines (t.p.r.s).

Låt $\boldsymbol{\Gamma}_k$ vara av en rang k matris, så att kolonnerna bildar en k-dimensionell ortonormal bas för ett delrum till $\boldsymbol{\delta}$, så att $\boldsymbol{\delta} = \boldsymbol{\Gamma}_k \boldsymbol{\delta}_k$, där $\boldsymbol{\delta}_k$ är en k-dimensionell vektor. Inom rummet spänt av $\boldsymbol{\Gamma}_k$ fås att (2.5.4) blir

$$\text{minimera } ||\boldsymbol{y - E\Gamma}_k\boldsymbol{\delta}_k - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta}_k' \boldsymbol{\Gamma}_k' \boldsymbol{E \Gamma}_k \boldsymbol{\delta}_k \text{ med bivillkoret } \boldsymbol{T'\Gamma}_k\boldsymbol{\delta}_k = 0 \tag{2.6.1}$$

Definera $\tilde{\boldsymbol{E}}_k = \boldsymbol{E\Gamma}_k\boldsymbol{\Gamma}_k'$ och $\hat{\boldsymbol{E}}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\boldsymbol{E\Gamma}_k\boldsymbol{\Gamma}_k'$, (2.5.4) kan då skrivas som:

$$\text{minimera } ||\boldsymbol{y - \tilde{E}}_k\boldsymbol{\delta - T\alpha}||^2 + \lambda \boldsymbol{\delta' \hat{E}}_k \boldsymbol{\delta} \text{ med bivillkoret } \boldsymbol{T'\delta = 0} \tag{2.6.2}$$

Istället för att hitta en bas som minimerar skillnaden i anpassade värden, väljer man det genomförbara, att minimera 'värsta' möjliga förändring:

$$\epsilon_k = \max_{\boldsymbol{\delta \neq 0}} \frac{||\boldsymbol{(E - \tilde{E})\delta}||}{||\boldsymbol{\delta}||}$$

På liknande sätt väljer man ett mått som minimerar 'värsta' möjliga förändring för strafftermen:

$$e_k = \max_{\boldsymbol{\delta \neq 0}} \frac{\boldsymbol{\delta'(E - \hat{E}}_k)\delta}{||\boldsymbol{\delta}||^2}$$

Givet målet att samtidigt minimera ϵ_k och e_k så visar det sig att den lämpliga basen är en egenbas av $\boldsymbol{E}$. Låt $\boldsymbol{E = UDU'}$ där $\boldsymbol{D}$ är en diagonalmatris av egenvärdena av $\boldsymbol{E}$

arrangerade så att $|D_{i,i}| \geq |D_{i+1,i+1}|$, $i = 1, ..., n-1$, och $\boldsymbol{U}$ är motsvarande egenvektor matris till $\boldsymbol{D}$. Det går att visa att den 'bästa' basen av rang k ges av $\boldsymbol{U_k}$, de första k kolumnerna av $\boldsymbol{U}$ (så $\boldsymbol{\Gamma}_k = \boldsymbol{U}_k$). Detta implicerar $\hat{\boldsymbol{E}} = \tilde{\boldsymbol{E}} = \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{U}_k'$, där $\boldsymbol{D}_k$ är en $k \times k$-matris nu. Problem 2.5.4 blir:

$$\text{minimera } ||\boldsymbol{y} - \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{\delta}_k - \boldsymbol{T}\boldsymbol{\alpha}||^2 + \lambda \boldsymbol{\delta}_k' \boldsymbol{D}_k \boldsymbol{\delta}_k \text{ med bivillkoret } \boldsymbol{T}'\boldsymbol{U}_k \boldsymbol{\delta}_k = 0 \quad (2.6.3)$$

Nu letar vi efter en $k \times k$-dimensionell ortogonal kolumn bas $\boldsymbol{Z}_k$ så att den $k \times k$-dimensionella matrisen $\boldsymbol{T}'\boldsymbol{U}_k\boldsymbol{Z}_k = 0$. Genom att begränsa $\boldsymbol{\delta}_k$ till rummet, genom $\boldsymbol{\delta}_k = \boldsymbol{Z}_k\tilde{\boldsymbol{\delta}}_k$, fås slutligen ett icke-begränsat problem:

$$\text{minimera } ||\boldsymbol{y} - \boldsymbol{U}_k \boldsymbol{D}_k \boldsymbol{Z}_k \tilde{\boldsymbol{\delta}} - \boldsymbol{T}\boldsymbol{\alpha}||^2 + \lambda \tilde{\boldsymbol{\delta}}' \boldsymbol{Z}_k^{'} \boldsymbol{D}_k \boldsymbol{Z}_k \tilde{\boldsymbol{\delta}}$$

med avseende på $\tilde{\boldsymbol{\delta}}$ och $\boldsymbol{\alpha}$. Genom insättning av $\boldsymbol{\delta} = \boldsymbol{U}_k\boldsymbol{Z}_k\tilde{\boldsymbol{\delta}}$ i (2.5.3) kan man nu få ett uttryck för den släta funktionen $f(\boldsymbol{x})$. Parametervektorn och modellmatrisen är $\boldsymbol{\beta}' = [\tilde{\boldsymbol{\delta}}', \boldsymbol{\alpha}']$ respektive $\boldsymbol{X} = [\boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{Z}_k, \boldsymbol{T}]$, och straffmatrisen $\boldsymbol{S}$ består av $\boldsymbol{Z}_k'\boldsymbol{D}_k\boldsymbol{Z}_k$ i vänstra hörnet och nollor i övrigt. Strafftermen ges nu av $\boldsymbol{\beta}'\boldsymbol{S}\boldsymbol{\beta}$. [5]

2.7 Tensorproduktbas

En nackdel eller en fördel beroende på situation med ovanstående straffterm konstruktion är att strafftermen medför att modellen blir isotrop, det innebär att strafftermen straffar varje enhets förflyttning av varje variabel lika mycket. Detta skulle inte vara ett problem om motsvarande variabler är på samma skala. Men skulle kunna utgöra ett problem om de var på olika skalor, exempelvis om variablerna är *longitud* och *år*. Å andra sidan en fördel med en isotrop egenskap är att byte av plats av kovariater inte ger ett annorlunda resultat.

För att kringgå problemet med ett resultat beroende på skalor, kan man istället bilda tensorprodukter av marginal baserna. Då kan man komma runt problemet genom att bilda olika strafftermer för variabler som inte är på samma skala. Följande konstruktion kan beskrivas mera allmänt, men vi konstruerar en tensorprodukt bas av två marginal baser, vilket räcker för denna uppsats. Låt x och z vara två kovariater, som också kan vara vektorer. Vi definierar de motsvarande släta funktionerna med:

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x) \qquad\qquad f_z(z) = \sum_{l=1}^{L} \delta_l d_l(z)$$

där α_i och δ_l är parametrarna, och $a_i(x)$ och $d_l(z)$ basfunktionerna. Genom att låta parametern α_i variera som en funktion av z, $\alpha_i(z) = \sum_{l=1}^{L} \beta_{il} d_l(z)$, fås en slät funktion av x och z:

$$f_{xz}(x, z) = \sum_{i=1}^{I} \sum_{l=1}^{L} \beta_{il} d_l(z) a_i(x)$$

[5] Se [3].

där β_{il} är okända parametrar. Givet en lämplig ordning av β_{il}, kan man visa att i:te raden av modellmatrisen är Kronecker produkten av i:te raden av respektive marginal modellmatris:

$$\boldsymbol{X}_i = \boldsymbol{X}_{xi} \otimes \boldsymbol{X}_{zi}$$

Vi antar nu att strafftermerna för respektive marginal term, kan uttryckas med:

$$J_x(f_x) = \boldsymbol{\alpha'} \boldsymbol{S}_x \boldsymbol{\alpha} \qquad\qquad J_z(f_z) = \boldsymbol{\delta'} \boldsymbol{S}_z \boldsymbol{\delta}$$

där $\boldsymbol{\alpha}$ och $\boldsymbol{\delta}$ är respektive koefficient vektor. I den här uppsatsen använder vi $m = 2$ i (2.5.2), vilket också antas för naturliga kubiska splines. Om $d = 1$ i (2.5.2), dvs kovariaterna är 1-dimensionella, så gäller att $J_x(f_x) = \int (\partial^2 f_x / \partial x^2)^2 dx$. Under dessa förutsättningar definierar vi strafftermen som:

$$J(f_{xz}) = \int_{x,z} \lambda_x \left(\frac{\partial^2 f_{xz}}{\partial x^2} \right)^2 + \lambda_z \left(\frac{\partial^2 f_{xz}}{\partial z^2} \right)^2 dxdz$$

och här kan man se att vi har en utjämningsparameter för varje riktning vilket möjliggör att 'straffet' är invariant av skalning av kovariaterna. Tensorprodukt ger dock utrymme för samspelseffekter. [6]

2.8 Penaliserad maximum likelihood

Penaliserad maximum likelihood är en variant av maximum likelihood (ML),som används vid skattning av basparametrar. Vi konstruerar nu formeln. Låt $\{f_j \;\; 1 \leq j \leq n\}$ vara släta funktioner så att de har formen $\boldsymbol{f_j} = \boldsymbol{\tilde{X}_j} \boldsymbol{\tilde{\beta}_j}$, där $\boldsymbol{\tilde{X}_j}$ är modellmatrisen och $\boldsymbol{\tilde{\beta}_j}$ är parametervektorn. Vi antar här att modellmatrisen redan är given och att endast skattning av parametervektorn återstår.

Vanligtvis är (2.2.1) inte identifierbar, så man inför ett villkor så att $\boldsymbol{1'} \boldsymbol{\tilde{X}_j} \boldsymbol{\tilde{\beta}_j} = 0$. Man kan med en omparametrisering uppnå detta direkt, vilket vi nu antas göra. Låt de nya modellmatriserna och parametervektorerna vara $\boldsymbol{X_j}$ respektive $\boldsymbol{\beta_j}$. En lämplig variant av (2.2.1) kan nu skrivas:

$$g(\mu_i) = \boldsymbol{X}_i \beta$$

där $\boldsymbol{X} = [\boldsymbol{X^*} : \boldsymbol{X_1} : \boldsymbol{X_2} : ... : \boldsymbol{X_n}]$ och $\boldsymbol{\beta'} = [\boldsymbol{\theta'}, \boldsymbol{\beta_1'}, \boldsymbol{\beta_2'}, ..., \boldsymbol{\beta_n'}]$. Låt $\boldsymbol{\bar{S}_j}$ vara de omparametriserade straffmatriserna. Låt vidare $\boldsymbol{S_j}$ vara $\boldsymbol{\bar{S}_j}$ med tillagda nollor upp till dimensionen av $\boldsymbol{\beta\beta'}$, så att om $dim(\boldsymbol{X^*}) = 0 \times 0$, $dim(\boldsymbol{\bar{S}_1}) = r \times s$ och $dim(\boldsymbol{\bar{S}_2}) = u \times v$, gäller att $\boldsymbol{S_1}[2 : (1 + r), 2 : (1 + s)] = \boldsymbol{\bar{S}_1}$ och
$\boldsymbol{S_2}[(2 + r) : (1 + r + u), (2 + s) : (1 + s + v)] = \boldsymbol{\bar{S}_2}$ o.s.v., om vi antar att vi har en intercept parameter i modellen. Då fås att $\boldsymbol{\beta' S_j \beta} = \boldsymbol{\beta_j' \bar{S}_j \beta_j}$.

Slutligen definieras penaliserad maximum likelihood, som vi vill maximera, med:

[6] Se [2] sid 158-163.

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j \boldsymbol{\beta}' \boldsymbol{S}_j \boldsymbol{\beta}$$

där $l(\boldsymbol{\beta})$ är likelihood och λ_j är utjämningsparametrarna, som kontrollerar trade-off mellan en välanpassad modell och modellens släthet [7]. Dessa skattas med en annan metod.

2.9 Utjämningsparameter

Antalet frihetsgrader begränsas av basdimensionen, men i slutändan är det utjämningsparametern som avgör antalet frihetsgrader, som oftast är icke-heltal och heter på engelska 'effective degrees of freedom' (EDF). Utjämningsparametern är också som nämnts tidigare avgörande för hur formen på funktionen blir, där formen kan bli allt mellan linjär till en perfekt anpassad till datapunkterna. Men det sista är anledningen till varför man inför en straffterm, så att modellen straffar mot en perfekt anpassad modell, då detta skulle vara mycket osannolikt, men också för att få en ökad precision vid prediktion av ny data.

Utjämningsparameter kan skattas med flera olika metoder, men de vanligaste är ordinär korsvalidering (OCV) och generaliserad korsvalidering (GCV), beroende på om spridningsparametern ϕ är okänd eller känd. I denna uppsats används dock en metod som heter 'restricted maximum likelihood' (REML), då vi antar en negativ binomial fördelning med okänd form parameter. Men vi går inte närmre in på metoden i denna uppsats.

3 Beskrivning av datamaterial

Det ursprungliga datamaterialet är mer omfattande än vad vi väljer att fokusera på, för att förenkla analysen. Således fokuserar vi på det datamaterial som är samlat i Sydaustralien under åren 1978-1995 vid 18 olika tillfällen årsvis i juli/augusti på en rektangelyta av $4.9630 \cdot 10^4$ km^2 begränsad av koordinaterna 34.38° S 139.5° E och 31.13° S 140.97° E. Data består av antalet röd kängurun per km^2 som en funktion av *longitud*, *latitud* och *år*. Antalet *longitud-*, *latitud-* och *tidpunkter* är 137,19 respektive 18.

Insamlandet av data gick till på så sätt att man delade in ett gigantiskt område i Sydaustralien i 5 km långa horisontella transekter från norr till söder. Därefter flög man över området med observatörer samtidigt som man försökte hålla sig till en standard metod för fixed-wing surveys", så långt det gick. Metoden innebär bland annat: en höjd på 76 m ovan mark, en markhastighet på 185 km/h och en 200 m bred avsökningsarea från varje sida av flygplanet. Man delade in antalet kängurun på vänster respektive höger

[7] Se [2] sid 163-164.

sida av flygplanet i par. I figur 1 ser vi vart man samlat in data under åren.

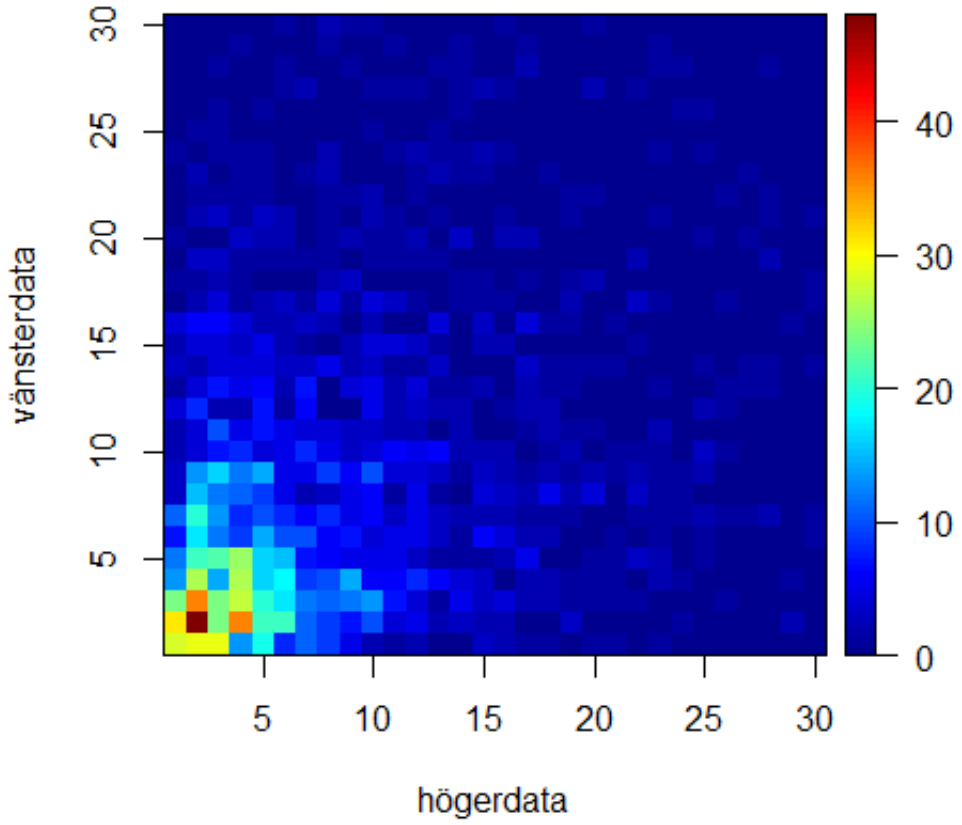
Ovanstående metod ger utrymme för räknefel, som kan bero på att man inte håller rätt höjd, dålig sikt, mänsklig faktor osv. Detta gör att man har manipulerat data med en korrektions term, som man räknar fram. Den mer intresserade av insamlad data, kan se [4]

Figur 1: Fördelning av kängurur markerat med svart.

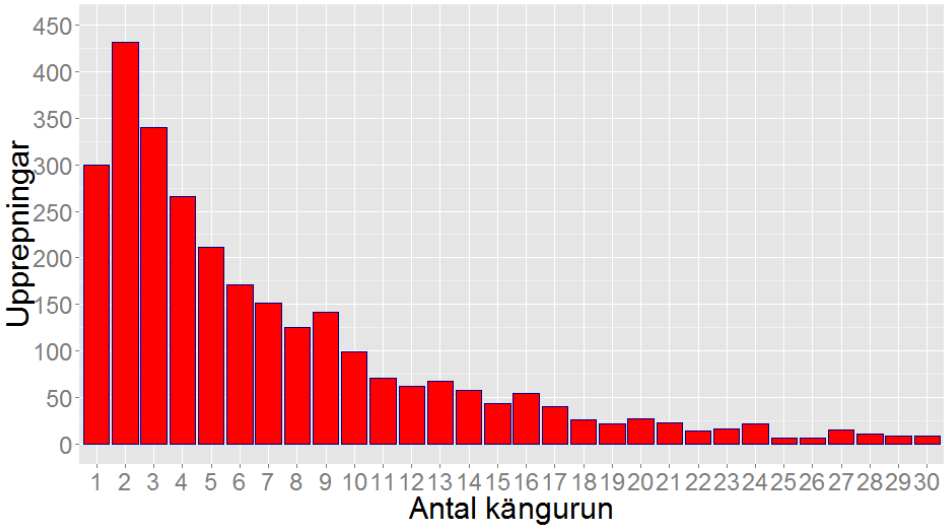
4 Analys av data och resultat

4.1 Inledande dataanalys

Vi har data på antalet kängurun på vänster och höger sida av flygplanet per km^2 i par som en funktion av *longitud*, *latitud* och *år*, men skillnaden verkar dock inte vara stor enligt Wilcoxon signed rank test, då man får ett p-värde på 0.5, även figur 2 verkar peka åt samma håll. Men huvudanalysen sker med vänsterdata, och därav ligger fokus på dessa data från nu. Något som karaktäriserar antalet kängurun är mängden nollor, som står för $\frac{3096}{6011} \approx 0.52$ andelar av data. I figur 3 ser vi antalet kängurun i intervallet 1-30.



Figur 2: Figur över antalet vänster och höger kängurun i par i intervallet 1-30.



Figur 3: Histogram över antalet kängurun i intervallet 1-30.

När vi senare främst tittar på modell 1, ser vi från tabell 1, att varianserna är betydligt högre än medelvärdena. En poisson modell skulle inte lämpa sig. Vi använder oss istället av en negativ binomial fördelning när vi i fortsättningen modellerar, då denna tillåter överdispersion. Men även denna lämpar sig egentligen inte heller för data med en hög andel nollor, men stöds av paketet *mgcv* i **R**. [8]

År	Medelvärde	Stickprovsvarians
1978	2.56	29.5
1979	2.52	20.6
1980	3.02	29.9
1981	6.04	125.4
1982	3.35	43.0
1983	2.21	19.0
1984	1.70	9.8
1985	2.66	33.8
1986	2.46	29.3
1987	2.67	22.8
1988	4.66	43.9
1989	4.72	42.0
1990	6.12	84.4
1991	6.46	114.4
1992	5.15	112.9
1993	4.04	29.9
1994	4.52	44.7
1995	5.35	83.3

[8] Se [6].

Tabell 1: Årsvisa medelvärden och stickprovsvarianser.

4.2 Modeller

Vi kommer att titta på tre modeller på olika form, där vi konstruerar baser med thin-plate splines och tensorprodukter med hjälp av *mgcv* paketet i **R** [9]. Vi använder en log länk, antar en negativ binomial fördelning för $Y_i \sim NegBin(\mu_i, \theta)$ så att $log(\mu_i) = \eta_i$, $E(Y_i) = \mu_i$. Vi låter $m = 2$ i (2.5.2) Vi behöver hitta en lämplig basdimension för respektive modell, det gör vi genom att jämföra AIC och justerad R-kvadrat R^2_{adj} värden. Dessutom beräknar vi deviance för vänsterräkningar, men också för högerräkningar genom modellen anpassad för vänsterdata. Detta är av intresse då det är ett sätt att testa modellernas prediktionsförmåga givet att vänster och högerräkningarna kan beskrivas av samma modell.

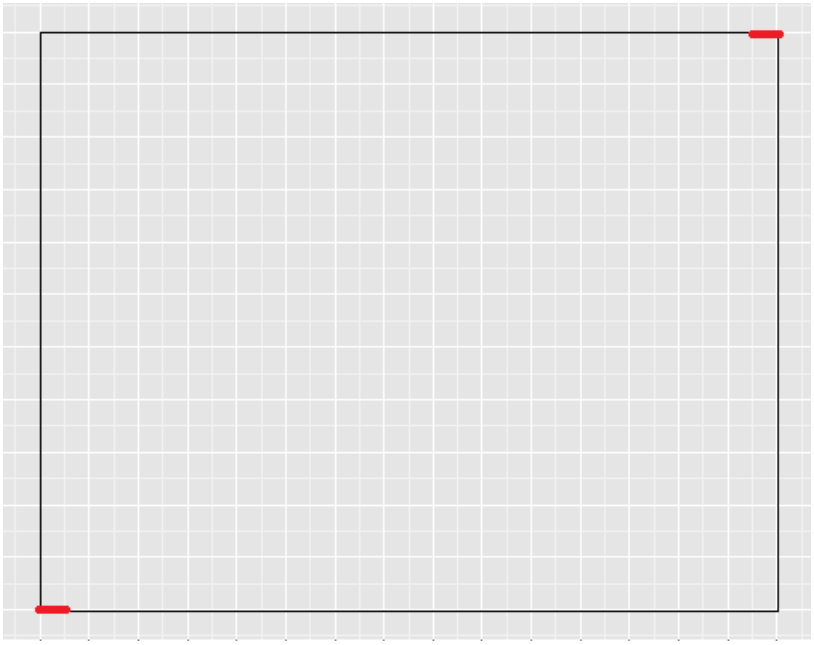
För modell 2 och 3 kommer vi också beräkna totala antalet kängurun för varje år (för modell 1 fås de nästan direkt). Detta gör vi genom att dela in området i ett tillräckligt fint rutnät, i detta fall räcker det med 100×100 rektanglar, och approximera antalet kängurun per km^2 inom varje rektangel med medelvärdet av prediktionerna av fyra hörnen multiplicerat med arean av rektangeln. Givet indelningen beräknar vi fram en genomsnittlig area, då en enhets förflyttning i *latitud* och *longitud* i *km* varierar med samma variabler. Vi approximerar rektangelarean

$$dlongitud \cdot dlatitud = \frac{140.97 - 139.5}{100} \cdot \frac{-31.13 - (-34.38)}{100} = 0.0147 \cdot 0.0325$$

$$\approx 4.78 \cdot 10^{-4} longitud \cdot latitud$$

med $1.377 \cdot 3.604 = 4.963$ km^2. Vi har beräknat fram omvandlingen från 0.0147 *longitud* till 1.377 *km*

[9] Se [6].



som medelvärdet av avståndet av de röda linjerna i ovanstående figur, som motsvarar sidan av en rektangel i hörnen på området. På ett analogt sätt beräknar vi motsvarande avstånd för *latitud*.

Formeln för uppskattad antal kängurun årsvis ges av följande:

$$[exp(\boldsymbol{X_1\beta}) \cdot w_1 + exp(\boldsymbol{X_2\beta}) \cdot w_2 + ... + exp(\boldsymbol{X_{(n+1)^2}\beta}) \cdot w_{(n+1)^2}]/4 \cdot 4.963 \qquad (4.2.1)$$

där n^2=100^2 är antalet rektanglar, $\boldsymbol{X_i}$ är en rad ur modellmatrisen och w_i tillhörande antalet gånger denna används i formeln beroende på var i rutnätet punkten finns, $1 \leq i \leq (n+1)^2$.

För att ta fram 90% konfidensintervall gör vi approximationen att parametervärdena $\boldsymbol{\beta}$ är multivariat normalfördelat med väntevärde och kovarians som tas från anpassad modell. Konfidensintervallen tas fram genom att simulera, $\boldsymbol{\beta}$, 5000 gånger, använda formel (4.2.1) för att slutligen beräkna motsvarande kvantiler. [10]

För modell 1 gör vi på liknande sätt med skillnaden att istället för ekvation (4.2.1) använder vi följande formel:

$$exp(\boldsymbol{X_i\beta}) \cdot 4.9630 \cdot 10^4.$$

4.2.1 Modell 1

I denna modell använder vi endast *år* som förklarande variabel, vilket kan skrivas på följande form:

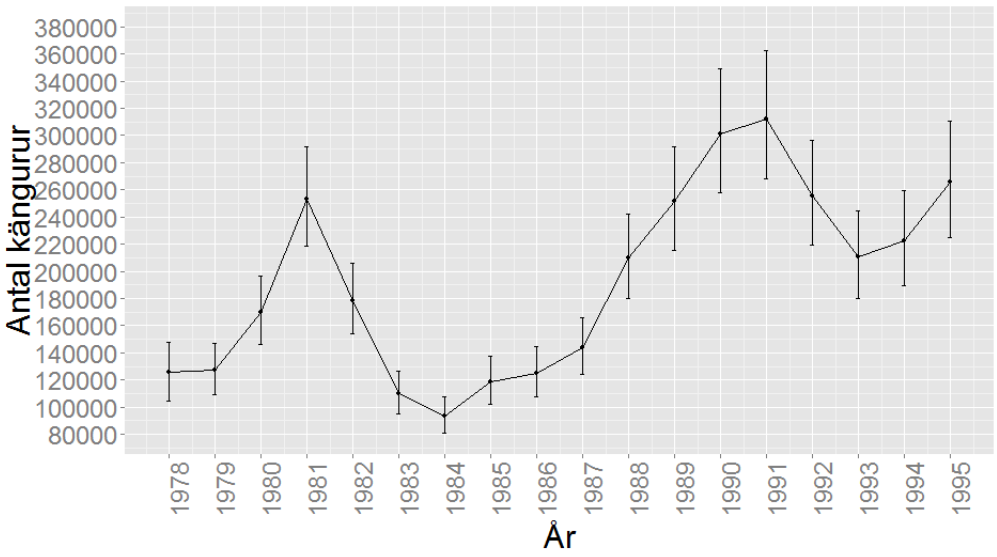
[10]Se [2] sid 190.

$$log(\mu_i) = intercept + f_1(år).$$

Vi får följande värden vid anpassning:

Basdimension (k)	Utjämningsparameter (λ)	sum(EDF)	p-värde (Termer)	R^2_{adj}	AIC	D (Vänster)	D (Höger)
5	8.55e-05	4.91	<2e-16	0.022	25981	5454	5311
10	8.75e-05	9.12	<2e-16	0.035	25922	5457	5338
15	1.70e-04	11.5	<2e-16	0.037	25913	5456	5335
18	1.01e-05	12.6	<2e-16	0.038	25911	5455	5337

Vi ser att både AIC och R^2_{adj} förbättras ju större dimension på basen är. Då beräkningstiden är kort, väljer vi modellen med högst dimension. Vi noterar även att deviancen för vänster och höger räkningarna inte skiljer sig stort procentuellt, vilket förstås är en fördel för modellens prediktionsförmåga. Dessutom är höger deviancen lägre, vilket implicerar högre likelihood vilket förstås är bättre. Dock är det bara $18 - 12.6 = 5.4$ frihetsgrader vilket gör deviancerna signifikant på signifikansnivån 0.01, men det är inte så konstigt då en förklarande variabel är ett naivt antagande, dessutom är det inte tiden i sig, utan andra faktorer som samvarierar med tiden som förklarar modellen.



Figur 4: Uppskattat totala antalet kängurun årsvis med 90% konfidensintervall.

Ovan i figur 4 har vi plottat totala antalet kängurun årsvis med 90% konfidensintervall. Vi ser att väntevärdena kan variera rätt mycket från år till år, men att konfidensintervallen är å andra sidan breda så att nästan alla värden överlappar sina grannar. Vi ser också en liten uppåtgående trend i antalet kängurun.

4.2.2 Modell 2

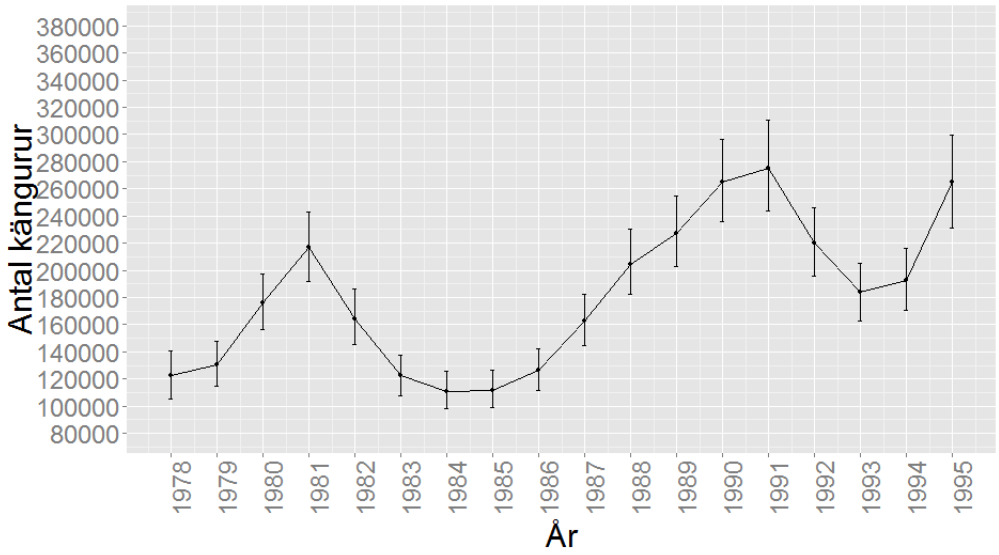
Här använder vi alla tre förklarande variabler, med följande modell:

$$log(\mu_i) = intercept + f_1(longitud, latitud) + f_2(år).$$

Vi får följande värden vid anpassning:

Basdimension (k)	Utjämningsparameter (λ)	sum(EDF)	p-värde (Termer)	R^2_{adj}	AIC	D (Vänster)	D (Höger)
30, 18	3.45e-04, 1.81e-05	39.7	<2e-16	0.22	23743	5443	5270
50, 18	3.93e-04, 1.96e-05	56.6	<2e-16	0.27	23584	5412	5386
70, 18	5.55e-04, 1.86e-05	71.5	<2e-16	0.28	23516	5379	5362
100, 18	6.36e-04, 2.09e-05	90.6	<2e-16	0.29	23449	5354	5405

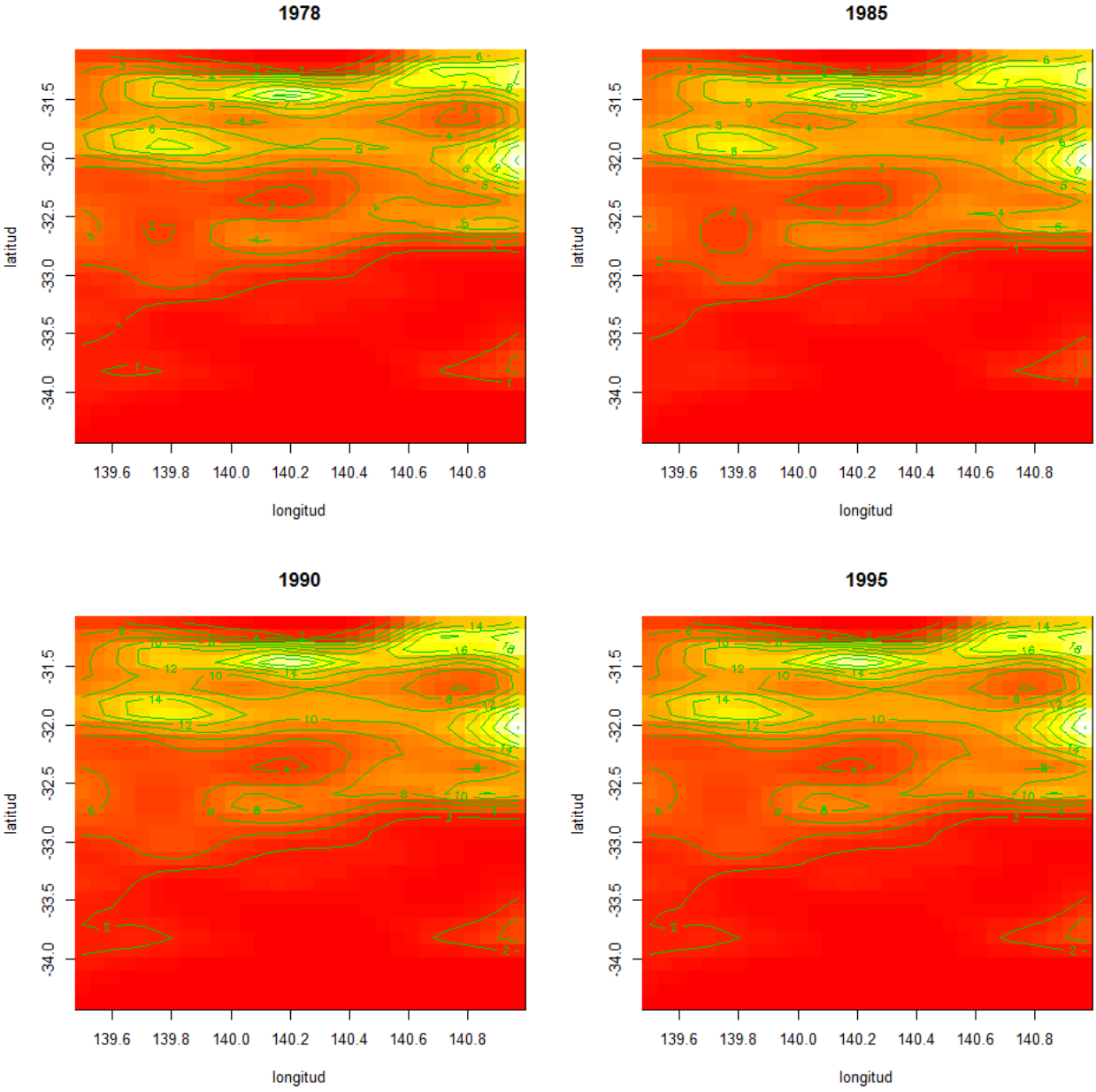
Även här fås bättre R^2_{adj} och AIC värden ju högre basdimensionen är, men vi väljer att stanna med basdimensionen då deviancen inte ändras mycket mer när vi ökar den, men också för att få samma högsta basdimension som i modell 3 längre ner. Således väljer vi att gå vidare med modellen med högst basdimension. Dock är denna gång vänster deviancen lägre än höger deviancen till skillnad mot modell 1, vilket är mer rimligt då modellen är anpassad till vänster data. Vi har nu 758+18 unika datakombinationer vilket ger $776 - 90.6 = 685.4$ frihetsgrader, men deviancerna är signifikant på signifikansnivån 0.01. Det kan bero på flera saker, och en anledning kan vara att modellen inte är tillräcklig komplex.



Figur 5: Uppskattat totala antalet kängurun årsvis med 90% konfidensintervall.

I figur 5 har vi plottat totala antalet kängurun, och vi noterar att formen liknar motsvarande figur för modell 1, men med skillnaden att figuren är mer hoptryckt och att

konfidensintervallen är mindre. I figur 6 har vi plottat täthetskarta för fyra olika år, och vi ser att de har ungefär samma form och täthetsförhållanden från år till år. Detta är inte förvånande då enligt modellen ska täthetsförhållanden inte variera, anledningen till att formen skiftar lite är helt enkelt att skärningen av ytan varierar.



Figur 6: Täthetskarta för model 2 över antal kängurun per km^2 över åren 1978, 1985,1990 och 1995.

4.2.3 Modell 3

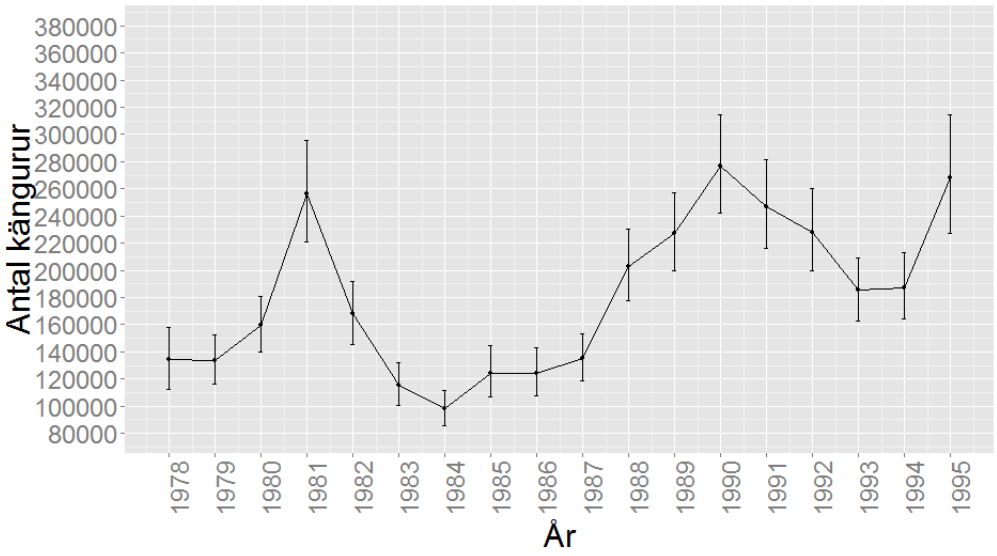
Här använder vi också alla tre förklarande variabler, fast med rumtids samspel:

$$log(\mu_i) = intercept + te(f_1(longitud, latitud), f_2(år))$$

där *te* står för tensorprodukt. Vi får följande värden vid anpassning:

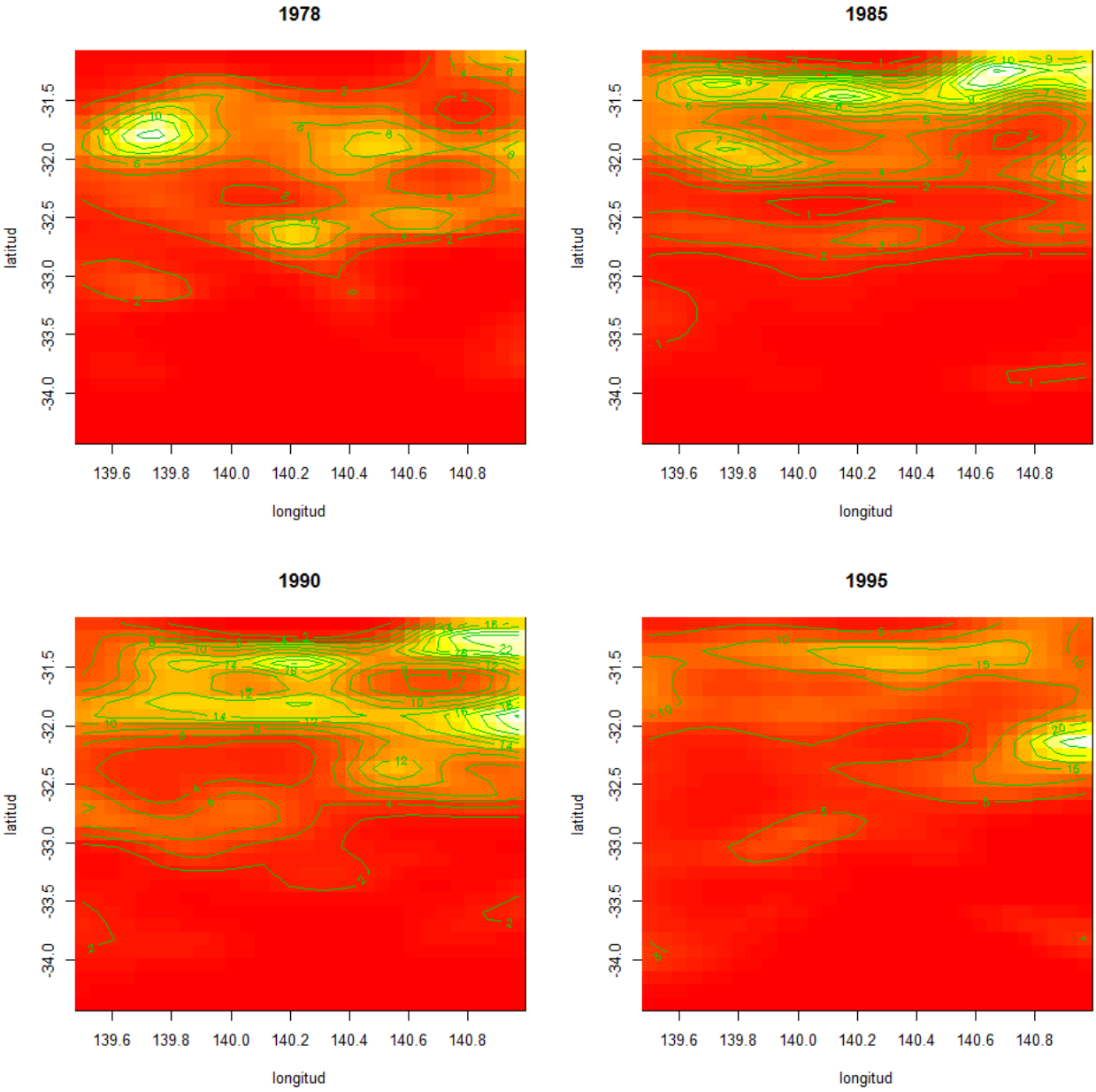
Basdimension (k)	Utjämningsparameter (λ)	sum(EDF)	p-värde (Termer)	R^2_{adj}	AIC	D (Vänster)	D (Höger)
30, 18	1.07e-03, 0.840	181	<2e-16	0.25	23578	5253	5295
50, 18	1.30e-03, 0.888	249	<2e-16	0.31	23429	5169	5456
70, 18	2.05e-03, 0.353	362	<2e-16	0.33	23295	5036	5620
100, 18	2.48e-03, 0.383	419	<2e-16	0.35	23269	4984	5692

Även här fås bättre R^2_{adj} och AIC värden ju högre basdimensionen är. Då vi använder tensorprodukt, har den mätta modellen hela $758 \cdot 18 = 13644$ parametrar, vilket är en enorm ökning från modell 2 och tar betydligt längre tid att modellerna och av denna anledning stannar vi vid den högsta basdimensionen enligt tabellen ovan, vars EDF är hela 419. Med $13644 - 419 = 13225$ frihetsgrader, är både vänster och höger deviancen icke signifikanta på signifikansnivån 0.01.



Figur 7: Uppskattat totala antalet kängurun årsvis med 90% konfidensintervall.

I figur 7 har vi plottat totala antalet kängurun, och formen påminner den in modell 2. Täthetskartan visas i figur 8. Vi noterar att de vita topparna varierar från år till år d.v.s. att populationstätheten varierar i förhållande till varandra, detta tyder på att det finns ett samspel mellan *koordinater* och *tid* vilket modellen ger utrymme för.



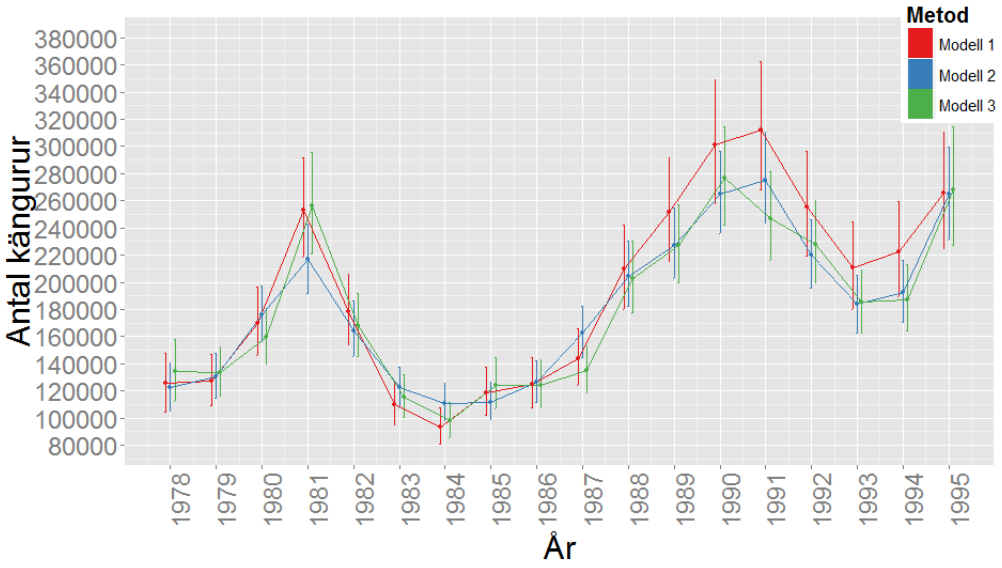
Figur 8: Täthetskarta för model 3 över antal kängurun per km^2 över åren 1978, 1985,1990 och 1995.

4.2.4 Allmänt om och jämförelse av modeller

I figur 9 kan vi åter igen se totala antalet kängurun för alla tre modeller. Vi noterar att kurvan för modell 2 och 3 ligger närmare varandra, dock påminner formerna för alla tre modeller varandra. Dessutom ligger alla kurvorna innanför varandras 90% konfidensintervall vilket är intressant därför att komplexiteten och antalet parametrar skiljer sig

enormt, ju enklare modellen är desto bättre är det.

Vi noterar att ibland kan höger deviancen vara lägre än vänster deviancen, förklaringen ligger i att vi maximerar penaliserad ML, därav maximeras inte ML, detta kan leda till att höger deviancen blir bättre.



Figur 9: Uppskattat totala antalet kängurun årsvis med 90% konfidensintervall för alla tre modeller.

5 Slutsats och diskussion

Vi har använt *mgcv* paketet i **R** för att skapa tre olika modeller för att försöka besvara vårt syfte med uppsatsen. Ju komplexare modell vi använt ju bättre AIC och R^2_{adj} värden har vi fått, vilket inte är en självklarhet då båda måtten straffar för varje extraparameter man har. Modell 1 har ett dåligt R^2_{adj} värde, men detta kunde man nästan ha räknat ut då modellen har *år* som förklarande variabel. Anledningen till att vi tog med modell 1 var snarare att få en uppfattning om hur stor skillnaden skulle bli mot modell 2 och 3 för totala antalet kängurun. Det blev ingen stor skillnad och formerna liknade varann.

Modell 3 påvisar i täthetskartan en variation i tätheten jämfört med modell 2, dessutom är parametern signifikant, vilket man självklart inte kan tolka bokstavligt, men det är ändå en tendens.

Det kan tyckas märkligt att antalet kängurun kan sjunka med 90000 på bara ett år mellan 1981-1982 för modell 3, men faktum är att när vi jämför med en Australiensk myndighets uppskattningar för hela Sydaustralien, så verkar det inte konstigt längre då populationen mellan åren 2002 och 2003 sjönk med hela 500000 från ca 1500000 till

1000000 [11]. Den stora variationen i populationsstorleken kan bero på flera orsaker, en orsak kan vara, vilket också kan vara en delorsak till samspelet i modell 3, är det faktum att yngre kängurun kan vara nomadiska d.v.s. att de springer bort stora sträckor och förändrar täthetsförhållanden. Även om större delen av de mogna uppfödande kängurna håller sig till sitt område kan vädret tvinga iväg de och på så sätt också påverka täthetsförhållanden.

Thin-plate splines är en svår metod, inte minst på grund av alla parametrar som behöver skattas och tolkas trots att vi bara har 3 förklarande variabler. Det finns mycket kvar att göra för att förbättra analysen. Man kan ta hänsyn till antalet nollor i data och anta en annan fördelning än en negativ binomial. Vidare har vi inte tittat på modellantagandena som till exempel deviance residualerna, utan vi har utgått från att de ska vara godtagbara och byggt analysen efter det. Med en kraftfullare dator kan man utöka basdimension, rutnät och simulering för att förbättra modellen. Fler förklarande variabler som mängden regn hade också varit givande.

Deviancerna för vänster- och högerdata för varje anpassning för respektive modell är signifikanta eller icke-signifikanta samtidigt. För modell 3 som verkar vara den bästa modellen är deviancerna icke-signifikanta samtidigt, vilket är en tendens till en god prediktionsförmåga. Man ska dock inte tolka dessa test bokstavligt då deviancen är asymptotiskt chi-kvadrat fördelat givet utjämningsparametern under icke-trunkerad bas [12]. Att kunna prediktera är en fördel då man vid nästa gång man samlar in data endast behöver titta åt ett håll vid observation.

[11] Se [5].
[12] Se [2] sid 200-201.

Referenser

[1] Agresti, Alan (2007) *Categorical Data Analysis*, 2. uppl., Wiley.

[2] N. Wood, Simon (2006) *Generalized Additive Models An Introduction with R*, Chapman & Hall/CRC.

[3] Wood, S.N. (2003) *Thin plate regression splines*, J. R. Statist. Soc. B 65, 95-114.

[4] G. C. Grigg, L. A. Beard, P. Alexander, A. R. Poplel & S. C. Cairns, 1999 *Aerial survey of kangaroos in South Australia 1978-1998: a brief report focusing on methodology*, Australian Zoologist, vol 31, no. 1,pp. 292-300.

[5] http://www.environment.gov.au/biodiversity/wildlife-trade/natives/wild-harvest/kangaroo-wallaby-statistics/kangaroo-population 2015-08-04.

[6] https://cran.r-project.org/web/packages/mgcv/index.html 2015-08-05

[7] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009, URL: http://had.co.nz/ggplot2/book 2015-08-05.