NÄR HAR ALL SVERIGES SKOG BRUNNIT NED?

En studie över svenska skogsbränder under åren 1998 till 2018.

Jakob Fjellström och Johan Byström

LUNDS UNIVERSITET
Ekonomihögskolan

Kandidatuppsats i statistik, 15 HP, HT-19
Statistiska institutionen, Ekonomihögskolan vid Lunds Universitet
Handledare: Jonas Wallin

Abstract

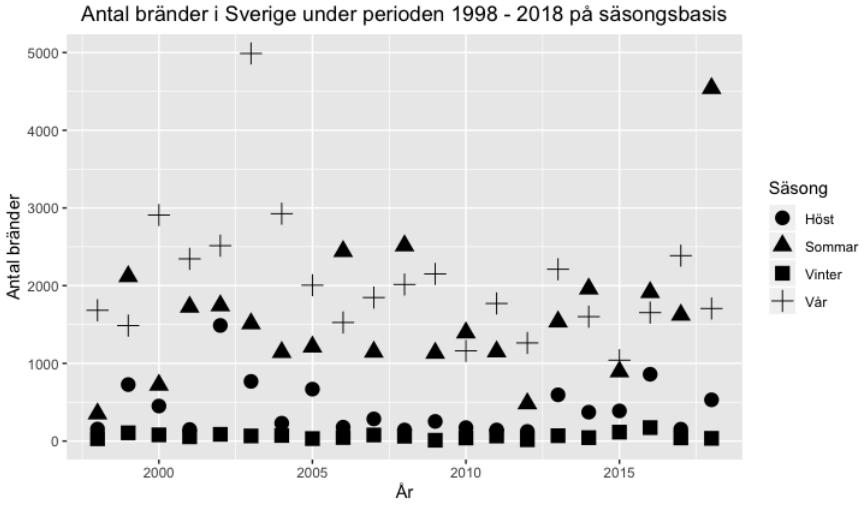
This thesis uses a data material from the Swedish Civil Contingencies Agency (MSB) to analyze forest fires in Sweden between 1998 and 2018. The thesis also uses weather data (precipitation and average temperature) as explanatory variables in two different regression models. The purpose of this thesis is to make predictions for the number of forest fires and the total burned down hectare of forest for the period 2019 to 2021. This is a research field where, for Swedish conditions, there is a lack of research and this thesis aims to fill that gap. Since the number of forest fires is a discrete variable a GLM model, that incorporates auto-regressive terms, is used and the observations are assumed to follow a Negative Binomial Distribution. To make predictions for the total burned down hectare of forest a multiple linear regression model is used. Our results suggest minor deviations in future forest fires from the overall pattern over the last two decades. However, our predictions carry some uncertainty since our prediction intervals are quite wide.

Keywords: GLM, Linear Regression, Time-Series, Negative Binomial Distribution.

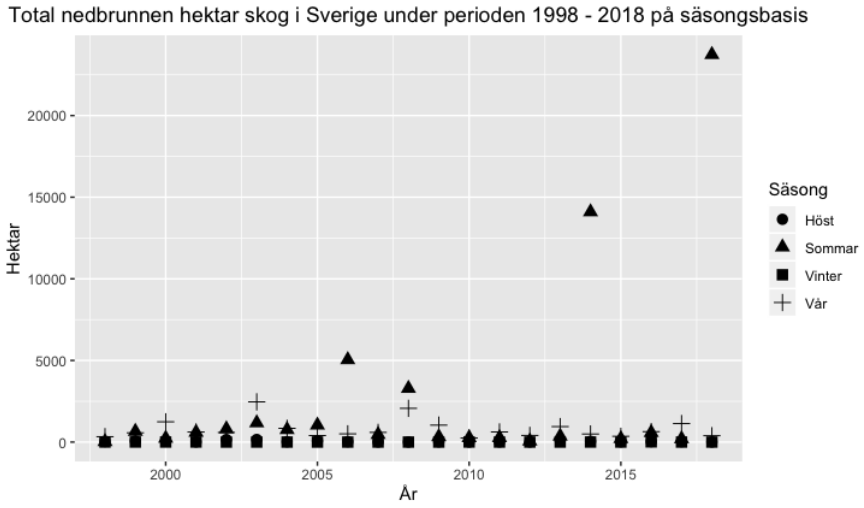
Innehållsförteckning

1 Inledning

Bränder inträffar i skog och mark med jämna mellanrum, men mer omfattande skogsbränder har historiskt inträffat i Sverige en till två gånger per decennium. Stora bränder kan medföra stora ekonomiska skador då det kan vara svårt att få ersättning för brandskadat virke och röjningsarbetet efter en brand innebär höga kostnader. Bränder kan även leda till att människor måste evakueras från sina hem och i värsta fall får sätta livet till, vilket inträffade under branden i Västmanland 2014 då en timmerbilsförare fastnade i lågorna och omkom. (SkogsSverige, 2019 och Skogsstyrelsen, 2019). I figur 1 och figur 2 nedan visas antalet skogsbränder samt den totala hektaren nedbrunnen skog i Sverige indelat per säsong under perioden 1998-2018.



Figur 1: Figuren visar antalet skogsbränder i Sverige under åren 1998-2018.



Figur 2: Figuren visar total nedbrunnen hektar skog i Sverige under åren 1998-2018.

Den stora branden i Västmanland sommaren 2014 var fram till 2018 den största branden i modern tid då ett område som omfattade drygt 13 000 hektar brann. Denna sommar sticker tydligt ut i figur 2, men i figur 1 avviker inte sommaren 2014 från det generella mönstret. Sommaren 2014 var varm och torr och SMHI hade varnat för att brandrisken var extremt stor. En brand som vid larmtillfället var ca 30 gånger 30 meter växte snabbt och okontrollerat främst på grund av höga temperaturer och starka vindbyar, men även på grund av att räddningstjänsten hade utdaterade kartor vilket gjorde att brandbilarna körde fel och släckningsarbetet blev ca en halvtimme försenat. Under de kommande dagarna steg temperaturen samtidigt som luftfuktigheten sjönk vilket ledde till att branden snabbt växte och omfattade omkring 13 000 hektar. När vädret till slut vände till räddningstjänstens fördel hade omkring 1000 människor och 1700 boskapsdjur evakuerats. Räddningsinsatsen kunde officiellt avslutas den 11 september. (Skogsstyrelsen, 2019).

Under sommaren 2018 drabbades dock Sverige ännu hårdare av skogsbränder än 2014, vilket också tydligt kan ses i figur 2. Förutsättningarna var mycket speciella då det kom väldigt lite regn vilket gjorde marken mycket uttorkad. Brandrisken var extremt hög i nästan hela landet. Sommaren 2018 var också speciellt i avseendet att bränderna var utspridda över stora delar av Sverige samt att de började redan i maj och den höga brandrisken, med några få nedgångar, bestod ända till augusti. Detta kan man tydligt se i figur 1 där antalet skogsbränder sommaren 2018 tydligt sticker ut från det generella mönstret. Västmanland drabbades återigen av bränder, men även Gävleborgs, Jämtlands och Dalarnas län drabbades av omfattande bränder. Totalt brann ungefär 25 000 hektar skog under sommaren 2018. (SOU 2019:7).

Således är skogsbränder ett högst aktuellt ämne och något som kan få mycket stora och allvarliga konsekvenser för samhället. Därför vill vi försöka prognosticera de kommande årens utveckling avseende skogsbränder i Sverige.

1.1 Syfte

Syftet med uppsatsen är att prognosticera antalet skogsbränder och den totala nedbrunna hektaren skog i Sverige för åren 2019, 2020 och 2021.

Fortsättningen av uppsatsen är indelad i 4 delar. Den första delen behandlar datainsamling och databearbetning och den andra delen behandlar uppsatsens metoder. I den tredje delen presenteras uppsatsens resultat och uppsatsen avslutas sedan med en diskussion.

2 Data

Det datamaterial som den här uppsatsen använder hämtar vi från två källor, dels Myndigheten för samhällskydd och beredskaps (MSB) databas, samt Sveriges meteorologiska och hydrologiska instituts (SMHI) databas. Båda databaserna är tillgängliga via respektive myndighets hemsida. Från MSBs databas samlar vi in data över hur många kvadratmeter skog och mark som har brunnit ned sedan januari 1998 fram till och med december 2018 samt antalet insatser från räddningstjänsten för att bekämpa dessa bränder (MSB, 2019). Den totala arealen nedbrunnen skog innefattar det MSB benämner som "brand i skog och mark". Detta innefattar bränder i "Produktiv skogsmark", "Annan trädbevuxen mark" samt "Mark utan träd". Vi väljer hädanefter att använda "skogsbränder" som samlingsnamn för samtliga tre kategorier då vi refererar till total areal nedbrunnen mark. Data över total areal nedbrunnen skog anges i m^2, men för att underlätta framtida databehandling räknar vi om arealen till hektar genom att dividera antal m^2 med 10 000.

Vi är som tidigare beskrivet intresserade av att undersöka antalet skogsbränder, men varken MSB eller räddningstjänsten för statistik över detta. Istället använder vi data för totala antalet insatser i skog och mark som en proxy för att approximera antalet skogsbränder. Vi kommer att använda benämningen "antal skogsbränder" i fortsättningen. Enligt MSB bör antalet räddningsinsatser överrensstämma väldigt väl med antalet bränder, men det finns viss osäkerhet i jämförelsen av ett par anledningar. Bland annat kan en brand starta och slockna av sig själv utan att räddningstjänsten får vetskap om den och det förekommer även att räddningstjänsten inte skrivit en händelserapport och därför syns inte den insatsen i statistiken. Vidare kan anlagda bränder ofta anläggas på flera närliggande platser samtidigt och det blir närmast en filosofisk fråga hur många bränder man ska se det som. Det är också rätt vanligt att en brand startar och släcks för att sedan starta igen några dygn senare varpå räddningstjänsten kan välja att fortsätta på samma händelserapport (räknas då som en insats) eller börja på en ny händelserapport (räknas då som två insatser). MSB har naturligt svårt att kvantifiera dessa avvikelser, men deras egna forskare brukar använda antalet insatser som en approximation av antalet bränder varför vi också väljer att göra detta. Det är troligtvis den bästa approximationen man kan göra.

När det kommer till hur stor areal som brunnit ned utgår vi ifrån att vår nedladdade data väl beskriver hur verkligheten ser ut. I kontakt med MSB framkom inga osäkerheter kring påligheten i deras data över nedbrunnen areal.

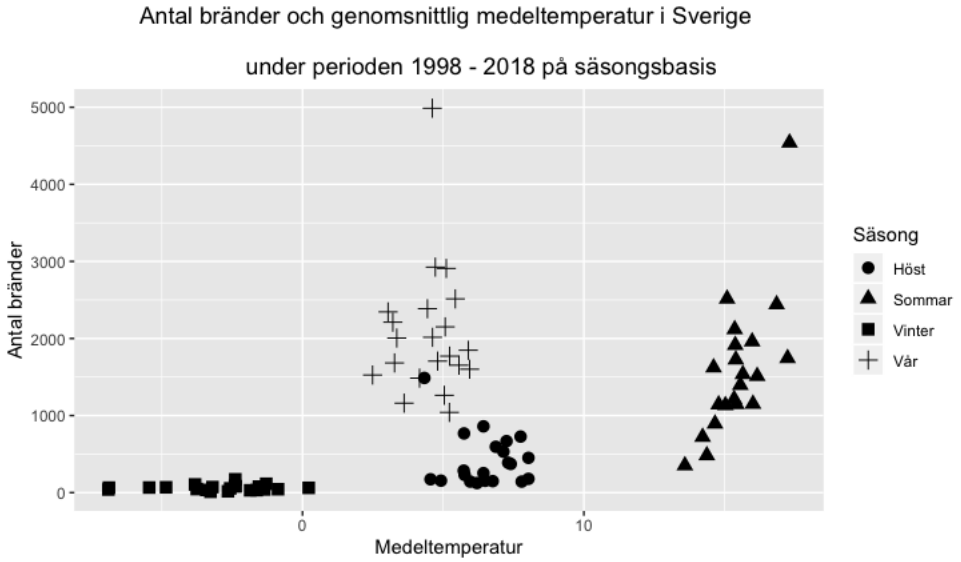
Från SMHI samlar vi in data för två förklarande variabler, nämligen temperatur och nederbörd, för åren 1998-2018. Data för temperatur samlar vi in i form av genomsnittlig temperatur (celsius) på säsongsbasis där mars-maj utgör vår, juni-augusti utgör sommar, september-november utgör höst och december-februari utgör vinter (SMHI, 2019a). Den genomsnittliga temperaturen per säsong är ett medelvärde av månadsmedeltemperaturen för de tre månaderna som ingår i respektive säsong.

Data för nederbörd samlar vi in i form av den summerade nederbörden i antal mm på säsongsbasis, där säsongerna är desammma som för temperatur, och de tre månadernas nederbörd har sum-

merats (SMHI, 2019b).

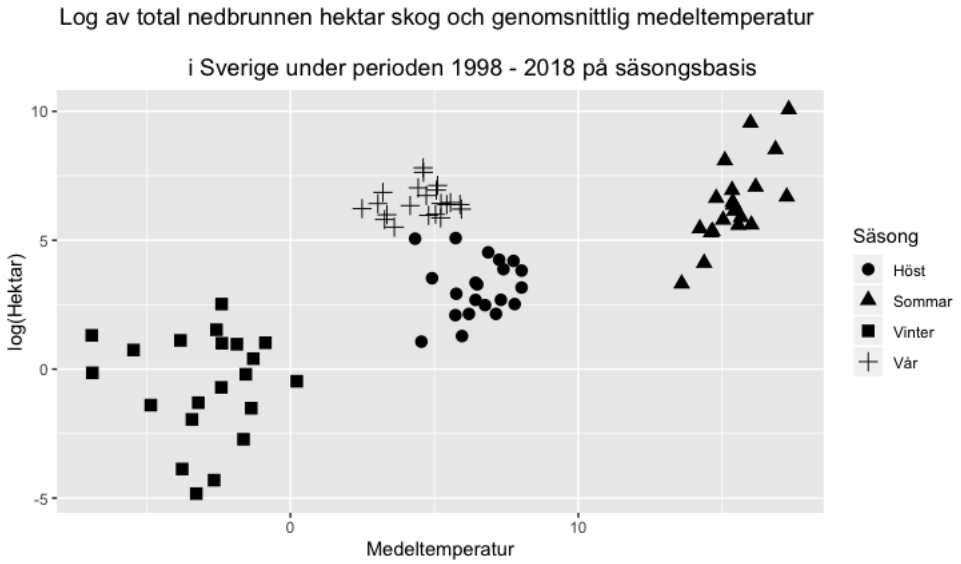
Eftersom att vår nedladdade data för medeltemperatur och nederbörd var uppdelad per säsong, och inte per månad, sammanställer vi hela datamaterialet från MSB över antalet insatser och total nedbrunnen areal på säsongsbasis. Detta gör vi genom att summera antal insatser och antal hektar för de månader som ingår i respektive säsong.

Vi väljer att använda nederbörd och temperatur som förklarande variabler för att dessa har stor påverkan på en skogsbrands möjligheter att starta och sprida sig. För att en skogsbrand ska kunna uppstå behövs lämpligt och torrt bränsle på marken. Ju torrare bränsle, desto större möjligheter har branden även att sprida sig. (Granström & Axelsson, 2018). Torrheten påverkas både av temperaturen och mängden nederbörd. Varmare somrar med lite nederbörd ökar risken för bränder eftersom att högre temperaturer leder till större avdunstning och till att marken torkar ut (SVT, 2019). I figur 3 kan vi se antal bränder mot medeltemperaturen. Det finns ett tydligt samband mellan högre medeltemperaturer och ett högre antal skogsbränder.



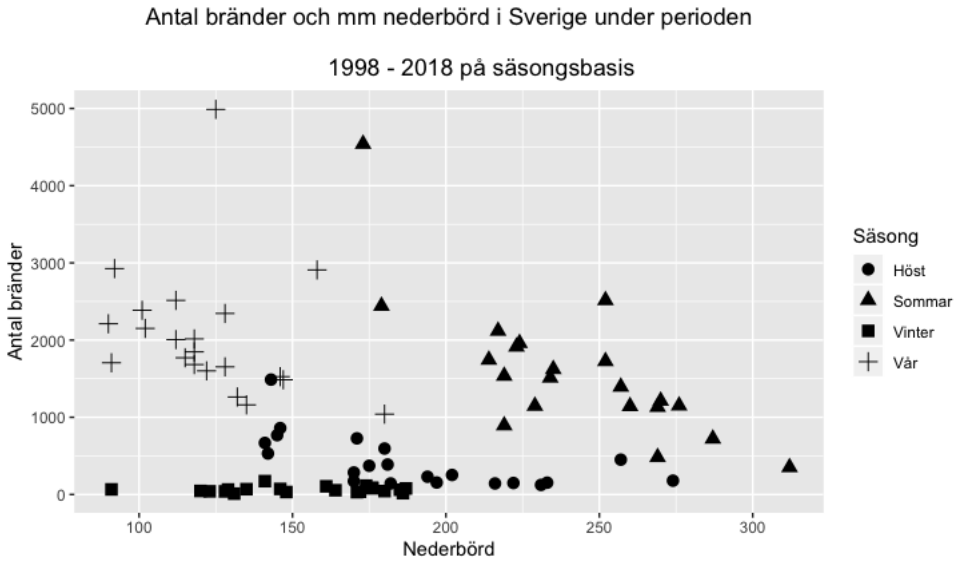
Figur 3: Figuren visar antalet bränder och medeltemperatur (celsius) under åren 1998-2018.

I figur 4 kan vi se den logaritmerade nedbrunna hektaren skog mot medeltemperatur. Även här kan vi se ett tydligt samband mellan högre medeltemperaturer och en större nedbrunnen areal skog. Vi har logaritmerat hektaren för att tydligare åskådliggöra sambandet. Vi använder logaritmerade värden i analysen vilket beskrivs mer utförligt i metodavsnittet.

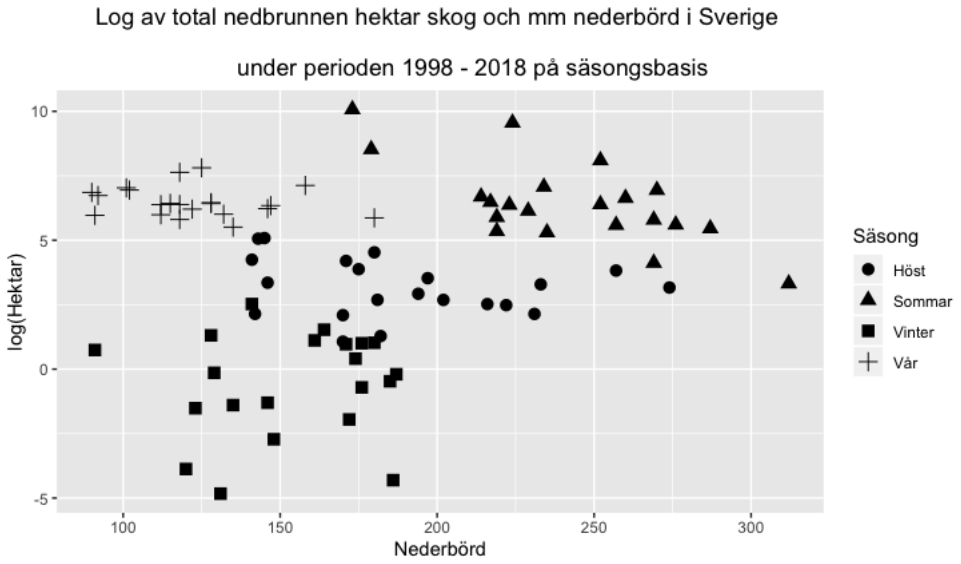


Figur 4: Figuren visar log av total nedbrunnen hektar skog och medeltemperatur (celsius) i Sverige under åren 1998-2018.

I figur 5 och 6 kan vi se antal bränder och den logaritmerade nedbrunna hektaren skog mot mängden nederbörd i mm. Även här kan vi se ett samband mellan mängden nederbörd och brändernas omfattning. I figur 5 kan vi se att ökad nederbörd tenderar att hålla nere antalet bränder, men att sommaren ändå har relativt många bränder. Våren tenderar även att ha lite nederbörd och relativt många bränder, det finns alltså någon form av säsongseffekt.



Figur 5: Figuren visar antalet skogsbränder och mängden nederbörd i mm i Sverige under åren 1998-2018.



Figur 6: Figuren visar log av total nedbrunnen hektar skog och mängden nederbörd i mm i Sverige under åren 1998-2018.

När det kommer till en skogsbrands förmåga att sprida sig har även vindens hastighet stor betydelse (Granström & Axelsson, 2018). Vi har sökt efter data över vind, men SMHI har ingen sammanställd data på månads- eller säsongsbasis för vindhastighet då detta är svårt att mäta och vindens hastighet kan variera väldigt mycket från dag till dag varför ett medelvärde av något slag lätt blir missvisande (SMHI, 2019c).

För att ha möjlighet att utvärdera hur precisa våra prognoser är delar vi upp datamaterialet i två delar. 90%, d.v.s 19 år, av datamaterialet använder vi för att bygga våra modeller och resterande 10%, d.v.s 2 år, använder vi för att utvärdera våra prognoser. För att kunna göra prognoser för år 2019, 2020 och 2021 behöver vi värden på nederbörd och medeltemperatur för dessa år. Prognoser för kommande års nederbörd och medeltemperatur hämtar vi också från SMHI. De värden vi kommer att använda för nederbörd och medeltemperatur är genomsnittliga värden som beräknats från 9 klimatmodellers upskattade värden för de kommande årens nederbörd och medeltemperatur.

Sammanställningen av all data utförs i Excel och samtliga statistiska analyser utförs med hjälp av R. I huvudsak använder vi standardpaketen i R samt paketen *MASS* och *tscount*.

3 Metod

3.1 Kort om tidsserier

Mot bakgrund av att det datamaterial som vi har är insamlat över tid ska vi kort redogöra för viktiga begrepp och metoder som vi använder oss utav vid analys av tidseriedata. Data som samlas in över tid är väldigt vanligt inom diverse områden. Man brukar definiera två syften med tidsserieanalys; det första är att fånga den stokastiska komponenten som skapar en given tidsserie och det andra syftet är att göra prognoser för framtiden. (Cryer & Chan, 2008, s.1). Först ska en distinktion göras mellan två olika typer av tidsserier. En deterministisk tidsserie anses vara en tidsserie vars framtida värden kan prognistiseras helt baserat på tidigare värden. Komplementet till en deterministisk tidsserie är en stokastisk sådan vars framtida värden antingen delvis eller inte alls påverkas av tidigare värden. Detta kallas då för en stokastisk process, Y_t, där y_t är det observerade värdet vid tidpunkt t, $t = 1, 2, 3, .., T$. (Chatfield, 2011, s.33).

För en stokastisk process gäller att den antingen är stationär eller inte. Stationaritetsbegreppet är fundamentalt vid analys av stokastiska processer och ska här i korthet förklaras. En stationär stokastisk process kan antingen vara strikt stationär eller svagt stationär. Vid en strikt stationär tidsserie är dess statistiska egenskaper konstanta för alla förskjutna tidslagg. Det innebär att sannolikhetsfördelningen för $Y_{t_1}, Y_{t_2}, ..., Y_{t_p}$ är samma som för $Y_{t_1-\tau}$, $Y_{t_2-\tau}$, $Y_{t_3-\tau}$,...,$Y_{t_p-\tau}$, där τ är tidsförskjutningen. För en svagt stationär process gäller att dess väntevärde och varians är konstant för alla värden t, däremot beror dess autokovarians på storleken i tidsförskjutningen. (Chatfield, 2011, s.34f). Oftast finns det trender i tidsserien som gör att dessa stationaritetskrav inte uppfylls och då finns det metoder för att modifiera tidsserien så att den blir stationär. Nedan följer en kort genomgång av AR(p)-processen, vilken vi använder oss utav för analys av antalet skogsbränder.

3.1.1 Autoregressiva processer

En autoregressiv process är en process vars värde vid tidpunkten t beror på värden innan tidpunkten t. Dessa processer brukar definieras enligt:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t \tag{1}$$

Det nuvarande värdet av processen vid tidpunkten t, Y_t, är alltså en linjär kombination av de p senaste värdena plus en slumpmässig komponent, e_t, som innehåller all ny information i tidsserien som de tidigare värdena inte förklarar. Vi antar därför att e_t och Y_{t-1}, Y_{t-2} etc är oberoende för alla tidpunkter t. (Cryer & Chan, 2008, s.66).

3.2 Modellval för antalet skogsbränder

Datamaterialet över antalet skogsbränder är räknedata, d.v.s data som räknar antalet händelser under en tidsperiod och som endast kan anta positiva heltal. Ett vanligt tillvägagångssätt för att analysera den typen av datamaterial är genom att använda en Poisson-regressionsmodell, vilket är en form av generaliserade linjära modeller (Generalized linear model - GLM) (Agresti, 2007, s.74-75).

3.2.1 Generaliserade Linjära Modeller (GLM)

Alla GLM-modeller har en sak gemensamt, nämligen att de är alla uppbyggda av tre komponenter; den slumpmässiga- och den systematiska komponenten samt länk-funktionen. Den slumpmässiga komponenten består av det observerade x_i-värdet och bestämmer en sannolikhetsfördelning för dessa. I många fall kan den beroende variabeln, X_i, vara den andel lyckade försök av ett bestämt antal, då antar man att X_i är binomialfördelat. I andra fall, såsom i vårt fall, är X_i positiv heltalsdata och det är då vanligt att anta att X_i är Poissonfördelat. (Agresti, 2007, s.66).

Den systematiska komponenten definierar vilka förklarande variabler som ska ingå i modellen. Dessa variabler ingår linjärt i modelldefinitionen och utgörs av x_i. Denna linjära kombination, på formen $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$, kallas för *the linear predictor*. (Agresti, 2007, s.66).

Definiera väntevärdet av X, väntevärdet av dess sannolikhetsfördelning, med $\mu = \mathrm{E(X)}$, då specificerar den tredje komponenten i en GLM-modell, länk-funktionen, en funktion $g()$ som visar hur μ är relaterad till *the linear predictor*, enligt $g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$. Länk-funktionen, g(), kopplar alltså ihop den slumpmässiga komponenten med den systematiska komponenten. Den mest enkla länk-funktionen är $g(\mu) = \mu$ och modellerar medelvärdet direkt, den kallas för *the identity link*. Andra länk-funktioner låter μ vara icke-linjärt relaterad till de förklarande variablerna. Ett exempel är länk-funktionen $g(\mu) = log(\mu)$ som modellerar log av medelvärdet. Eftersom log-funktionen bara är tillämplig på positiva värden är log-länk-funktionen lämplig när μ inte kan vara negativ, vilket är fallet med räknedata. En GLM-modell som använder log-länk-funktionen kallas för en log-linjär modell och är definierad i ekvation 2 nedan. (Agresti, 2007, s.66-67). Eftersom vi modellerar räknedata använder vi oss genomgående av log-länk-funktionen:

$$log(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k. \tag{2}$$

3.2.2 Poissonfördelning och Poissonregression

En Poisson-regressionsmodell grundar sig i Poissonfördelningen, som är en diskret sannolikhetsfördelning och är därför lämplig att använda när man arbetar med diskreta variabler. Låt X vara en slumpvariabel som följer en Poissonfördelning ifall den har följande sannolikhetsfunktion (Hogg et al., 2014, s.88):

$$P\left(X = x\right) = \frac{e^{-\mu}\mu^{x}}{x!} \tag{3}$$

för $x = 0, 1, 2,...$ och 0 annars.

Poissonfördelningen ger sannolikheten för en händelse under ett givet tidsintervall. Exempelvis antalet skogsbränder som sker under juli månad. Vidare har Poissonfördelningen följande antaganden:

a) Variansen är detsamma som väntevärdet, $V[X] = E[X] = \mu$.
b) Observationerna X_i kan enbart anta positiva heltal, $x_i \in [0, 1, 2...]$
c) $\mu > 0$.
(Hogg et al., 2014, s.87-89).

Poisson-regressionsmodellen är en GLM som grundar sig i Poissonfördelningen och följer därav samma grundantaganden, exempelvis om att observationerna ska vara positiva heltal och att medelvärdet ska vara detsamma som variansen. Därtill följer dock ytterligare ett antagande ifall log-länkfunktionen används: det logaritmerade medelvärdet, $\log(\mu)$ måste vara en linjär funktion av x. (Agresti, 2007, s. 74-75).

Den log-linjära Poisson-regressionsmodellen, med en förklarande variabel x och en beroende variabel y, ser ut enligt följande:

$$log(\mu) = \beta_0 + \beta_1 x \tag{4}$$

I modellen ovan är variablerna logaritmerade. Den icke-logaritmerade motsvarigheten ges nedan.

$$\mu = e^{\beta_0}(e^{\beta_1})^{x} \tag{5}$$

Av (4) ser vi att en enhetsökning i x har en multiplikativ effekt av e^{β_1} på μ. Medelvärdet av Y vid $x + 1$ motsvarar medelvärdet av Y vid x multiplicerat med e^{β_1}. Låt säga att $\beta_1 = 0$, vilket ger $e^{\beta_1} = e^{0} = 1$ och därav är multiplikatoreffekten 1, vilket gör att väntevärdet av Y inte ändras när x förändras. Således, om $\beta_1 > 0$ är $e^{\beta_1} > 0$ och detta gör att Y ökar när x ökar. Slutligen, om $\beta_1 < 0$, minskar medelvärdet av Y när x ökar. (Agresti, 2007, s. 74ff).

3.2.3 Den Negativa Binomialfördelningen

Ifall slumpvariabeln X antas vara Poissonfördelad antas även variansen vara lika med medelvärdet, men detta är sällan fallet i verkligheten. När variansen är större än förväntat (alltså större än medelvärdet), talar man om överdispersion (Agresti, 2007, s.80). Detta är en vanligt förekommande företeelse, och man kan då istället anta att observationerna följer en Negativ binomialfördelning.

Den Negativa binomialfördelningens sannolikhetsfunktion definieras enligt:

$$P(X = x) = \binom{x-1}{r-1} p^{r-1}(1-p)^{x-r} \tag{6}$$

för $x = r, r+1, r+2,...$ och 0 annars.

Den Negativa binomialfördelningen används vid en situation då vi har en sekvens av oberoende Bernoulli-försök och där vi vill veta när exakt det r lyckade utfallet sker. Värdet för slumpvariabeln X är ett positivt heltal och anger vilken upprepning som det r lyckade utfallet sker i. (Hogg et al. 2014, s. 82).

3.2.4 Negativ binomial-regression med autoregressiva termer

Paketet *tscount* möjliggör inkluderingen av en eller flera autoregressiva termer i en GLM-regressionsmodell vilket medför att man kan fånga beroendet bakåt i tiden ett valfritt antal perioder. En skillnad mellan paketen *MASS* och *tscount* är hur skattningarna av modellernas parametrar utförs. I paketet MASS utförs en full maximum-likelihood-skattning av parametrarna vilket skiljer sig från *tscount* som utför quasi-likelihood-skattningar av parametrarna. (Liboschik et al., 2017). Vidare är den Negativa binomialfördelningens sannolikhetsfunktion definierad i *tscount* enligt nedan.

$$P(X_t = x|\mathcal{F}_{t-1}) = \frac{\Gamma(\phi + x)}{\Gamma(\phi)\Gamma(x+1)} \cdot (\frac{\phi}{\phi + \mu_t})^{\phi} \cdot (\frac{\mu_t}{\phi + \mu_t})^{x} \tag{7}$$

för $x = 0, 1, 2,...$ och 0 annars.

Här är ϕ fördelningens dispersionsparameter och μ_t är väntevärdet vid tidpunkt $t, t = [1, 2, 3, ..., T]$. Fördelningen har ett väntevärde $E[X_t|\mathcal{F}_{t-1}] = \mu_t$ och $V[X_t|\mathcal{F}_{t-1}] = \mu_t + \frac{\mu_t^2}{\phi}$, vilket medför att variansen ökar kvadratiskt med μ_t. Genom att addera en extra term till variansen hanterar man problemet med överdispersion. (Liboschik et al., 2017).

3.3 Modellvalidering

I residualanalysen av våra GLM-modeller använder vi standardiserade Pearsonresidualer (då detta är standardvalet i paketet *tscount*), PIT-histogram, kumulativt periodogram över residualerna, autokorrelationsfunktionen samt en *Marginal Calibration Plot*. Standardiserade Pearsonresidualer, för Negativ binomial-regression, beräknas enligt nedan.

$$\hat{e_t} = \frac{x_t - \hat{\mu_t}}{\sqrt{(\hat{\mu_t} + \hat{\mu_t}^2/\hat{\phi})(1 - h_{tt})}} \tag{8}$$

För Poisson-regressionsmodeller beräknas standardiserade Pearsonresidualer enligt följande:

$$\hat{e_t} = \frac{x_t - \hat{\mu_t}}{\sqrt{\hat{\mu_t}(1 - h_{tt})}}. \tag{9}$$

Pearsons standardiserade residualer som det är definierat subtraherar det skattade värdet från det observerade värdet och dividerar med medelfelet i skattningen, d.v.s. $(var[x_t](1-h_{tt}))^{1/2}$, där h_{tt} är leveragevärdet för observation t. Dessa värden ska fluktuera kring värdet noll. (Agresti, 2007, s. 87).

Det kumulativa periodogrammet över de standardiserade Pearsonresidualerna används för att avgöra huruvida slumpkomponenten i modellen, X_t, är vitt brus eller ej. De två linjerna inom vilka de kumulativt observerade Pearsonresidualerna ligger inom är ett 95%-konfidensintervall som, ifall det överträds, anger om det finns en trend i datamaterialet som modellen inte fångar upp. Periodogrammet är konstruerat efter frekvensintervallet mellan $0 < f < \frac{1}{2T}$, där T är utfallsrummet för tidsvariabeln. Här är T = 1, och periodogrammet går från 0 till 0.5. Det optimala är en rät linje genom origo, men huvudsaken är att frekvensen ligger inom det 95%-konfidensintervallet. (O'Leary & Rust, 2008).

PIT står för *Probability Integral Transform* och dess värden beskrivs enklast med ett histogram. Dessa värden används för att utvärdera en modells prediktiva förmåga. Det görs genom att ta det värde som den prediktiva fördelningsfunktionen uppnår vid en given observation och om observationen dras från en kontinuerlig och prediktiv fördelning får PIT-histogrammet utseendet av en uniform fördelning. I vårt fall är datamaterialet räknedata, vilket gör den prediktiva fördelningen diskret. Ett sätt att korrigera för detta är att randomisera PIT. Låt P vara den prediktiva fördelningen $x \sim P$, den observerade räknedatan, och v den uniforma fördelningen oberoende av x, får vi,

$$u = P_{x_{-1}} + v(P_x - P_{x_{-1}}), \tag{10}$$

där vi definierar $P_{x-1} = 0$ som uniformt fördelad. Vad paketet tscount gör är att tillämpa en icke-randomiserad PIT genom att byta det slumpmässiga PIT-värdet med dess betingade fördelningsfunktion givet dess observerade värde x enligt nedan.

$$F(u|x) = \begin{cases} 0 & , u \leq P_{x-1} \\ (u - P_{x-1})/(P_x - P_{x-1}) & , P_{x-1} \leq u \leq P_x \\ 1 & , u \geq P_x \end{cases} \tag{11}$$

Vidare kan kalibreringen härifrån göras genom att aggregera n prediktioner och jämföra med det genomsnittliga PIT, enligt nedan.

$$\bar{F}(u) = \frac{1}{n} \sum_{i=1}^{n} F^{(i)}(u|x^{(i)}),\ 0 \leq u \leq 1 \tag{12}$$

Här är $F^{(i)}$ baserad på den prediktiva fördelningen $P^{(i)}$ och det observerade värdet $x^{(i)}$ från den

uniforma fördelningsfunktionen. Detta visas via ett histogram, där idealet är ett rätblock likt den uniforma fördelningsfunktionen. (Czado et al., 2009).

Marginal Calibration Plot används som ett komplement till PIT-histogrammet. Vad den gör är att visualisera differensen mellan två fördelningsfunktioner, i.e. $\bar{F}_T(x) - \hat{G}_T(x), x \in R$. Här får $\hat{G}_T(x)$ illustrera den observerade fördelningsfunktionen, och $\bar{F}_T(x)$ den genomsnittliga prediktiva fördelningsfunktionen (d.v.s den Negativa binomialfördelningen), utvärderad vid $\hat{\mu}$ och $\hat{\phi}$. Om sambanden i datamaterialet är korrekt modellerat ska differensen ha mindre fluktueringar kring värdet 0. Diagrammet anger hur skarpa prediktioner man kan göra med den anpassade modellen. (Gneiting et al., 2007).

3.4 Modellval för nedbrunnen hektar

Till skillnad från antalet bränder, som är en diskret variabel, så är nedbrunnen hektar en kontinuerlig variabel och Poissionregression lämpar sig därför inte för att modellera den. Vi använder därför en multipel linjär regressionsmodell som lämpar sig när både den beroende och de oberoende variablerna är kontinuerliga.

3.4.1 Linjär regression

En enkel linjär regressionsmodell definieras enligt:

$$Y_t = \beta_0 + \beta_1 x_t + e_t. \tag{13}$$

Modellen bygger på antaganden om att
1) Residualerna, e_t, är oberoende varandra,
2) Residualerna, e_t, är normalfördelade,
3) Residualerna, e_t, har konstant varians, samt
4) Y förklaras av x enligt modellen ovan.

Den enkla linjära regressionsmodellen kan sedan utvecklas till en multipel linjär regressionsmodell genom att lägga till fler förklarande variabler, d.v.s fler x_t. Modellantagandena är då fortsatt desamma. (Sheather, 2010, s.17-21).

Vid linjär regression är syftet oftast att förklara variationen i Y, men för att lösa eventuella problem med icke-konstant varians kan man ibland behöva logtransformera Y. För att då kunna beräkna ett R^2-värde i originalskala behöver man transformera tillbaka Y. Vi använder följande tillvägagångssätt i uppsatsen för att göra detta:

1) Vi beräknar först *fitted values* för $U = log(Y)$ enligt en linjär regressionsmodell för U. Dessa värden transformerar vi sedan till skalan för Y enligt: $\hat{y}_t = e^{(\hat{u}_t + MSE/2)}$, där MSE är en skattning av residualvariansen.
2) Därefter beräknar vi residualer i originalskala enligt: $y_t - \hat{y}_t$. (Yang, 2012).
3) RSS och SST i originalskala beräknar vi enligt:

$$RSS = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \tag{14}$$

$$SST = \sum_{t=1}^{n} (y_t - \bar{y})^2. \tag{15}$$

Därefter beräknar vi ett justerat R^2 i originalskala enligt:

$$R^2_{Adj} = 1 - \frac{((RSS/(n-p-1))}{(SST/(n-1))} \tag{16}$$

där p = antal förklarande variabler och n = stickprovsstorleken. (Sheather, 2010, s.135-137).

4 Resultat

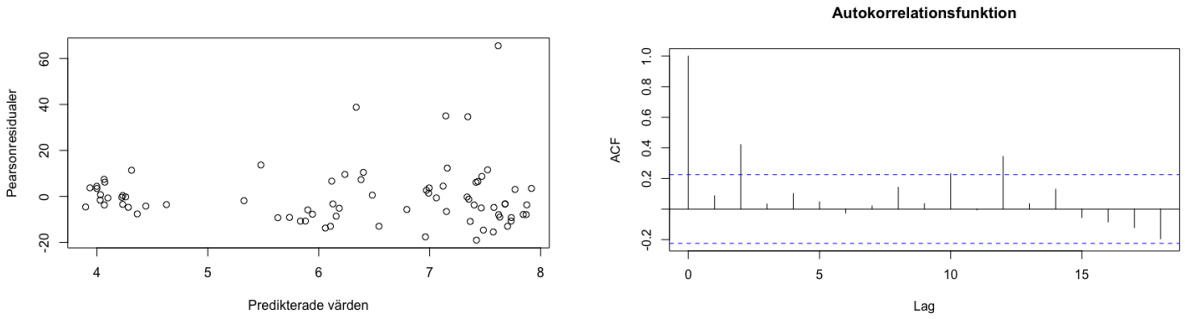
4.1 Antalet bränder

För att modellera vår data över antalet bränder anpassar vi en Poisson-regressionsmodell enligt nedan:

$$log(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 D_{1t} + \beta_4 D_{2t} + \beta_5 D_{3t}. \tag{17}$$

Variablerna x_{1t} och x_{2t} är medeltemperatur och nederbörd. Vidare är D_{1t}, D_{2t} samt D_{3t} dummyvariabler där $D_{1t} = 1$ om det är vår vid tidpunkten t, noll annars. $D_{2t} = 1$ om det är sommar vid tidpunkten t, noll annars. $D_{3t} = 1$ om det är höst vid tidpunkten t, noll annars.

Ovan modell innebär att det logaritmerade medelvärdet av antalet bränder för varje säsong från 1998 till 2016 förklaras utav medeltemperatur, nederbörd samt en säsongseffekt vi fångar upp genom att tilldela varje säsong en dummy-variabel. Nedan följer en residualanalys av modellen.



(a) Pearsonresidualer mot predikterade värden.

(b) Skattad korrelationsfunktion.

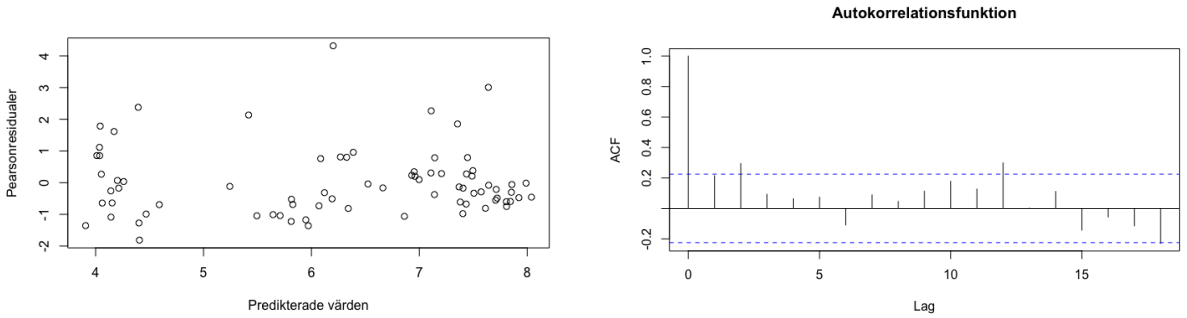
Figur 7: Figuren visar Pearsonresidualerna mot predikterade värden samt den skattade korrelationsfunktionen för Poisson-regressionsmodellen i ekvation 17.

I figur 7 ser vi residualerna mot predikterade värden samt den skattade korrelationsfunktionen. Vi kan se att det finns ett tydligt beroende mellan residualerna då det finns signifikant autokorrelation för flera laggade perioder. Vi kan även se att residualernas varians inte är konstant utan ökar över tid. Vi väljer därför att inte gå vidare med en Poisson-regressionsmodell. Istället tillpassar vi en regressionsmodell som antar att observationerna följer en Negativ binomialfördelning då en Negativ binomial-regressionsmodell inte förutsätter att variansen är lika med väntevärdet.

Vår nya anpassade modell är följande:

$$log(\mu_t) = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 D_{1t} + \beta_4 D_{2t} + \beta_5 D_{3t}. \tag{18}$$

Variablerna x_{1t} och x_{2t} är medeltemperatur och nederbörd. Vidare är D_{1t}, D_{2t} samt D_{3t} dummyvariabler där $D_{1t} = 1$ om det är vår vid tidpunkten t, noll annars. $D_{2t} = 1$ om det är sommar vid tidpunkten t, noll annars. $D_{3t} = 1$ om det är höst vid tidpunkten t, noll annars.



(a) Pearsonresidualer mot predikterade värden.

(b) Skattad korrelationsfunktion.

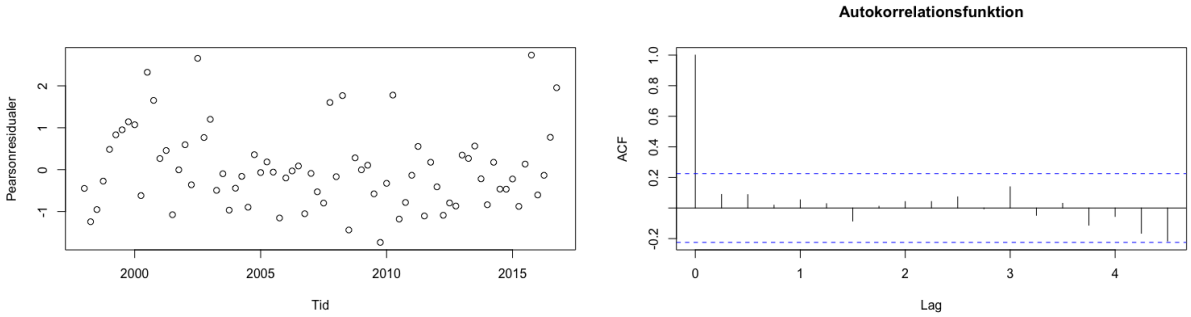
Figur 8: Figuren visar Pearsonresidualerna mot predikterade värden samt den skattade korrelationsfunktionen för den Negativa binomial-regressionsmodellen i ekvation 18.

I figur 8 ser vi att bytet till en Negativ binomial-regression ger residualer med jämn spridning som kan antas ha konstant varians. Det finns dock fortfarande signifikant autokorrelation hos residualerna vilket innebär att det finns en beroendestruktur mellan säsongerna över tid som modellen inte fångar upp. Vi väljer därför att utveckla modellen ytterligare ett steg genom att använda oss av metoder för tidsserieanalys.

För att lösa problemet med autokorrelationen adderar vi två autoregressiva termer till modellen i ekvation 18. Den nya anpassade modellen för antalet bränder blir således:

$$log(\mu_t) = \beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 D_{1t} + \beta_4 D_{2t} + \beta_5 D_{3t}. \tag{19}$$

Variablerna y_{t-1} och y_{t-2} är våra autoregressiva termer och x_{1t} och x_{2t} är medeltemperatur och nederbörd. Vidare är D_{1t}, D_{2t} samt D_{3t} dummyvariabler där $D_{1t} = 1$ om det är vår vid tidpunkten t, noll annars. $D_{2t} = 1$ om det är sommar vid tidpunkten t, noll annars. $D_{3t} = 1$ om det är höst vid tidpunkten t, noll annars. Således är skillnaden mellan modellerna i ekvation 18 och 19 att modellen i ekvation 19 inkluderar en AR(2)-process. Modellutvärdering och residualanalys följer nedan.



(a) Pearsonresidualer över tid.

(b) Skattad korrelationsfunktion.

Figur 9: Figuren visar Pearsonresidualerna över tid och den skattade korrelationsfunktionen för den Negativa binomial-modellen i ekvation 19.

Som vi kan se i autokorrelationsfunktionen i figur 9 finns det inte längre någon signifikant autokorrelation kvar för någon laggad period i modellen för antalet bränder efter adderandet av två autoregressiva termer. Vi väljer därför att fortsätta modellvalideringen för modellen i ekvation 19, hädanefter benämnd som den Negativa binomial-modellen.

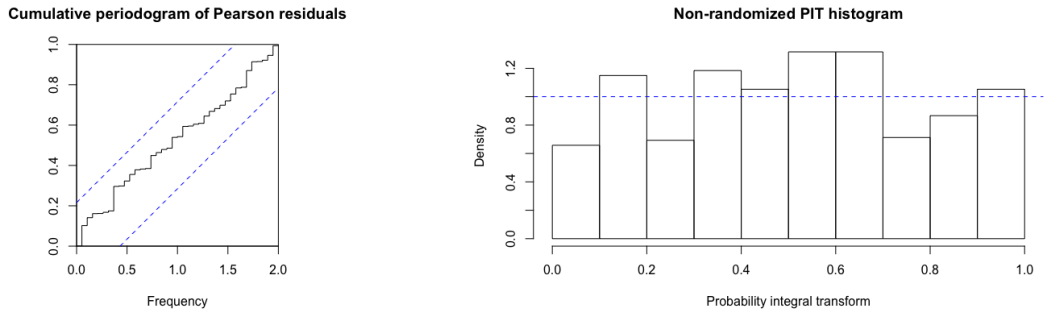
Den Negativa binomial-modellen har signifikanta parameterskattningar för nederbörd samt för samtliga dummyvariabler. Dessa parameterskattningar och tillhörande 95%-konfidensintervall återfinns i tabell 1 nedan. Som vi kan se är parameterskattningarna för de båda autoregressiva termerna inte signifikanta, även om adderandet av dessa har löst problemet med autokorrelation. Vi undersöker även en modell med endast en autoregressiv term, men detta räcker inte för att rensa bort all signifikant autokorrelation från modellen, oavsett om den autoregressiva termen är laggad en, två eller flera perioder tillbaka. Därför väljer vi att behålla två autoregressiva termer i modellen, även om skattningarna för de båda inte är signifikanta.

Parameter	Skattning	K.I 95% Nedre gräns	K.I 95% Övre gräns
Intercept	3,347	1,189	5,505
$\hat{\phi_1}$	0,111	-0,177	0,4
$\hat{\phi_2}$	0,264	-0,002	0,53
Nederbörd	-0,0107	-0,016	-0,0054
Medeltemperatur	0,0421	-0,101	0,185
Vår	3,433	2,115	4,75
Sommar	3,935	1,186	6,685
Höst	1,59	0,173	3,007

Tabell 1: Tabellen visar parameterskattningar samt tillhörande 95%-konfidensintervall för modellen i ekvation 19.

I figur 9 kan vi även se Pearsonresidualerna över tid. Vi undersöker som brukligt om det finns observationer som är ovanligt stora eller kan klassas som outliers och som bör studeras närmare. Om en Pearsonresidual har ett absolutvärde större än 2 kan den ses som en outlier. Som vi kan se i figur 9 finns det tre residualer som är aningen större än 2. En förklaring till dessa är att

de härrör från säsonger med ovanligt mycket bränder, t.ex våren 2003. Vi bedömmer inte att dessa beror på mätfel i vårt datamaterial utan att vissa år avviker från det generella mönstret vilket modellen misslyckas med att fånga upp. Därutöver finner vi inga uppenbara strukturer i residualerna och vi anser att det är rimligt att anta att residualerna har konstant varians.

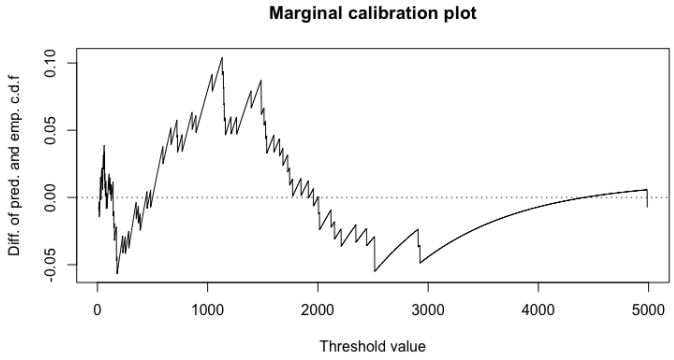


(a) Kumulativt periodogram.

(b) PIT-histogram.

Figur 10: Figuren visar ett kumulativt periodogram över Pearsonresidualerna samt ett PIT-histogram för den Negativa binomial-modellen i ekvation 19.

Som vi kan se i figur 10 över det kumulativa periodogrammet ligger samtliga residualer inom intervallet vilket tyder på att vi har fått fram residualer som är vitt brus. PIT-histogrammet i figur 10 för den Negativa binomial-modellen ser inte riktigt ut som en uniform fördelning, vilket antyder att dess prediktiva förmåga inte är helt tillförlitlig. Detta tyder på att vår anpassade modell inte modellerar datamaterialet på ett helt tillfredsställande sätt, men tillräckligt väl för att vi ska kunna acceptera modellen och gå vidare.



Figur 11: Figuren visar Marginal Calibration plot för Negativa binomial-modellen.

Eftersom att det ideala mönstret för en välanpassad modell, i en *Marginal Calibration Plot*, är mindre fluktuationer kring noll, kan vi efter att ha studerat figur 11 dra samma slutsats som från PIT-histogramet, nämligen att det finns viss osäkerhet i hur väl vår modell modellerar vårt datamaterial. Detta är något att ha i åtanke när vi använder modellen för att göra prognoser för framtiden.

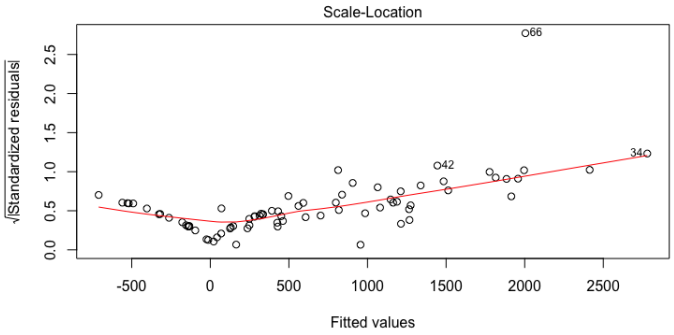
4.2 Total nedbrunnen hektar

För att modellera den totala nedbrunna hektaren skog börjar vi med att anpassa en multipel linjär regressionsmodell enligt följande:

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 D_{1t} + \beta_4 D_{2t} + \beta_5 D_{3t} + e_t. \tag{20}$$

Variablerna x_{1t} och x_{2t} är medeltemperatur och nederbörd. Vidare är D_{1t}, D_{2t} samt D_{3t} dummyvariabler där $D_{1t} = 1$ om det är vår vid tidpunkten t, noll annars. $D_{2t} = 1$ om det är sommar vid tidpunkten t, noll annars. $D_{3t} = 1$ om det är höst vid tidpunkten t, noll annars. e_t är residualerna.

Vi utvärderar modellen på sedvanligt sätt med att undersöka och bedömma modellutskriften och genomföra en residualanalys. Modellen som helhet är signifikant, men har inga signifikanta parameterskattningar och ett justerat R^2 på ca 14%. I figur 12 nedan återfinns en plot över de standardiserade residualerna. Residualerna uppvisar inte det slumpmässigt rektangelformade mönster vi vill se. Vi kan tydligt se en struktur i residualerna och vi drar därför slutsatsen att vi behöver justera modellen.



Figur 12: Figuren visar standardiserade residualerna mot fitted values för regressionsmodellen i ekvation 20.

För att komma tillrätta med strukturen hos residualerna anpassar vi en ny multipel linjär regressionsmodell, men med skillnaden att vi logaritmerar Y. Den nya modellen kommer i fortsättningen att benämnas log-modellen och är följande:

$$U_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 D_{1t} + \beta_4 D_{2t} + \beta_5 D_{3t} + e_t, \tag{21}$$

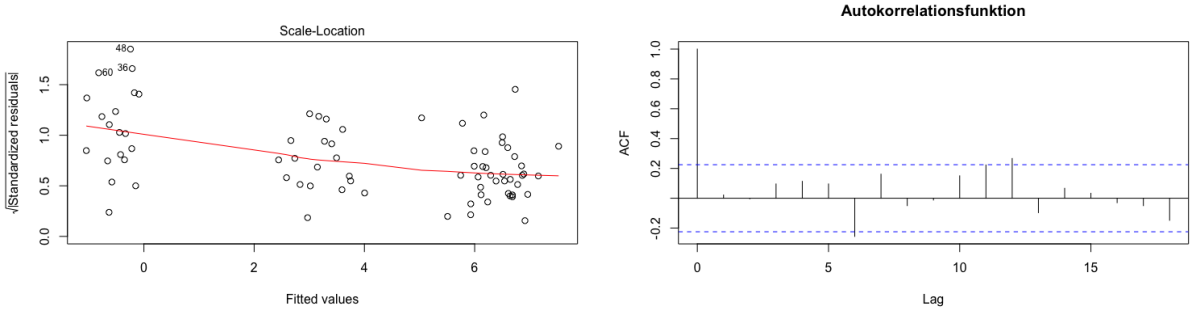
där $U_t = log(Y_t)$, d.v.s den naturliga logaritmen av Y_t. Variablerna x_{1t} och x_{2t} är medeltemperatur och nederbörd. Vidare är D_{1t}, D_{2t} samt D_{3t} dummyvariabler där $D_{1t} = 1$ om det är vår vid tidpunkten t, noll annars. $D_{2t} = 1$ om det är sommar vid tidpunkten t, noll annars. $D_{3t} = 1$ om det är höst vid tidpunkten t, noll annars. e_t är residualerna (det är dessa residualer vi använder för att beräkna MSE).

Vi utvärderar log-modellen på samma sätt som tidigare. Modellen som helhet är signifikant med signifikanta parameterskattningar för β_0, β_1 och β_5 samt ger oss ett justerat R^2-värde i originalskala på 8%. Parameterskattningarna är samlade i tabell 2 nedan. Sett utifrån antalet signifikanta parameterskattningar är log-modellen en klar förbättring jämfört med modellen i ekvation 21. Vi fortsätter med residualanalysen av logmodellen nedan.

Parameter	Skattning	P-värde	Signifikans
Intercept	4,185	0,0009	***
Nederbörd	-0,013	0,0221	*
Medeltemperatur	0,229	0,08	
Sommar	1,816	0,133	
Vinter	-1,967	0,127	
Vår	2,899	0,000	***

Tabell 2: Tabellen visar parameterskattningar samt p-värde för modellen i ekvation 21.
* - Signifikant på 5%-nivån. ** - Signifikant på 1%-nivån. *** - Signifikant på 0,1%-nivån.

I figur 13 kan vi se de standardiserade residualerna mot fitted values för log-modellen. Nu ser vi inte längre någon tydlig struktur i residualerna likt i figur 12. Spridningen är nu jämnare om än inte helt perfekt då den är lite större i början för att sedan avta. Vi anser ändå att spridningen är tillräckligt jämn för att vi ska kunna acceptera det. I den skattade autokorrelationsfunktionen i figur 13 ser vi dock att modellen uppvisar signifikant autokorrelation för lagg 6 och lagg 12. Den signifikanta autokorrelationen kan ha påverkat hypotestesterna avseende om parameterskattningarna är signifikanta (Sheather, 2010, s.311). Dock letar sig autokorrelationen för dessa perioder precis ovanför konfidensbanden och är därmed precis signifikanta. Det är inte förvånande att det finns korrelation och beroende mellan närliggande säsonger, vilket vi kunde se tidigare när vi undersökte antalet bränder, men detta kan vi inte se här då autokorrelationen för lagg 1 till 5 inte är signifikant. Att det skulle finnas ett tydligt beroende 6 (d.v.s 1,5 år) och 12 (d.v.s 3 år) säsonger tillbaka finner vi däremot inga logiska förklaringar till. Eftersom att autokorrelationen för dessa perioder också är relativt svaga och precis signifikanta misstänker vi att det kan röra sig om spökkorrelation. Det ska tilläggas att vi saknar biologiska kunskaper rörande skogsbränder och det skulle därmed kunna finnas en rimlig förklaring till detta mönster, dock har vi inte kunnat finna någon sådan.

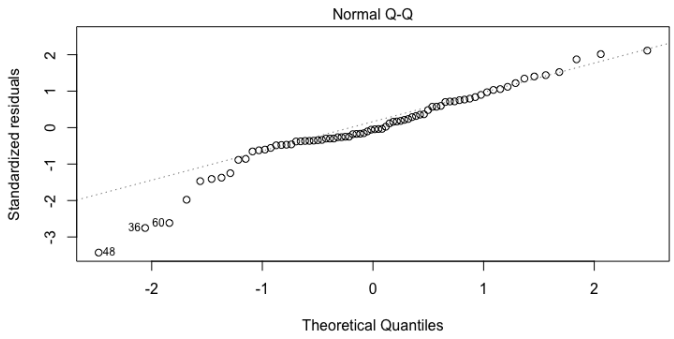


(a) Standardiserade residualer mot fitted values.

(b) Skattad korrelationsfunktion.

Figur 13: Figuren visar standardiserade residualer mot fitted values samt den skattade autokorrelationsfunktionen för modellen i ekvation 21.

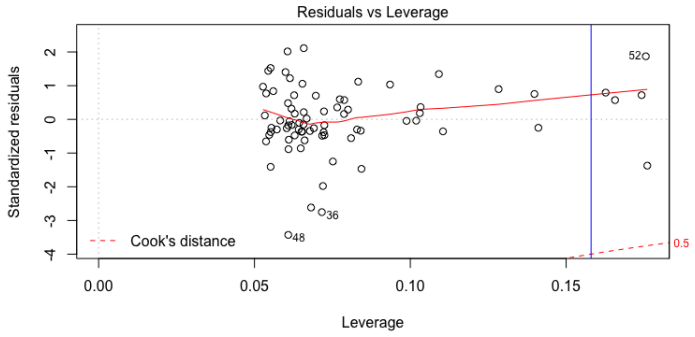
Vidare kontrollerar vi normalfördelningsantagandet hos residualerna med QQ-plotten i figur 14. Som vi kan se följer majoriteten av residualerna den räta linjen, men i ena svansen finns tre väldigt tydliga avvikelser från linjen. Vi har kontrollerat dessa värden och finner inga tecken på mätfel eller dylikt utan kan konstatera att dessa tre residualer samtliga härrör från vintersäsonger med ovanligt låg nedbrunnen areal skog. Vintern 2006 brann det 0,0208 hektar, vintern 2009 brann det 0,008 hektar och vintern 2012 brann det totalt 0,0135 hektar. Avvikelser likt de vi ser i figur 14 är vanligt vid datamaterial som har mer extrema värden än vad som väntas ifall de kommer från normalfördelningen. Eftersom att majoriteten av residualerna tydligt följer den räta linjen anser vi att antagandet om normalfördelade residualer är uppfyllt.



Figur 14: Figuren visar en QQ-plot för regressionsmodellen från ekvation 21.

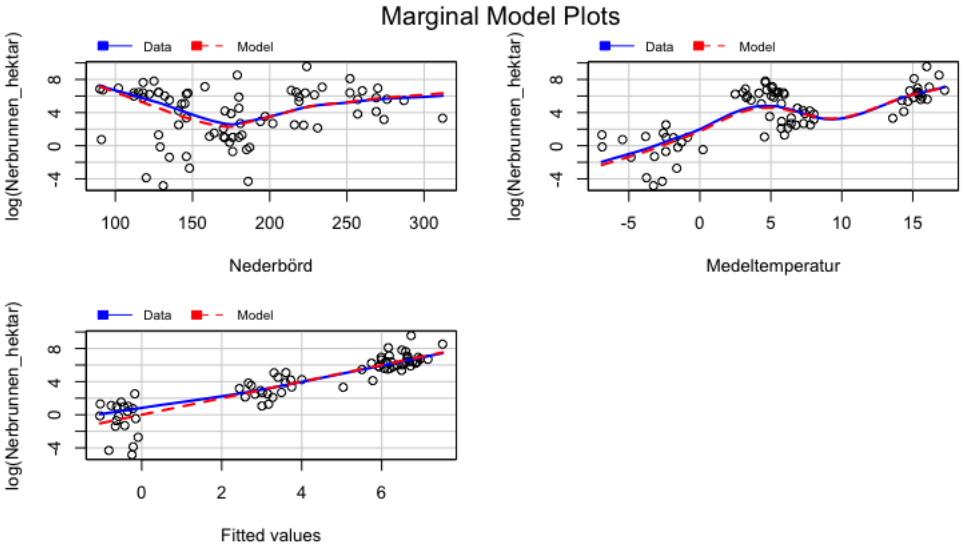
Vi fortsätter residualanalysen med att undersöka residualernas leveragevärden i kombination med Cook-avstånd, detta kan ses i figur 15. Som vi kan se är det 4 residualer som kan anses ha ett högt leveragevärde, beräknat som $2 \cdot \frac{(p+1)}{n}$, där p är antalet förklarande variabler (Sheather, 2010, s.154). Dessa residualer har dock alla ett absolutvärde som understiger 2 och ses därför inte som outliers (Sheather, 2010, s.155). Då dessa observationer inte är att betrakta som dåliga leveragepunkter och vi inte finner några mätfel eller dylikt väljer vi att gå vidare utan åtgärder, men håller i minnet att dessa kan ha ökat R^2-värdet och sänkt medelfelen för parameterskattningarna (Sheather, 2010, s.60). Även här kan vi se att observation 36 och 48 sticker ut av

samma anledningar som beskrevs i föregående stycke. Samtliga värden ligger inom gränserna för Cook-avstånden och sammantaget drar vi slutsatsen att vi inte ser något alltför oroväckande i figur 15.



Figur 15: Figuren visar de standardiserade residualernas leveragevärden och Cook-avstånd för regressionsmodellen från ekvation 21. Gränsen för vad som anses som ett högt leveragevärde är markerat i blått.

Figur 16 visar *Marginal Model Plots* för log-modellen. Vi kan se att linjerna följer varandra relativt väl i samtliga diagram, dock ser linjerna ut att glida ifrån varandra för låga värden på den förklarande variabeln nederbörd. För medeltemperatur ser linjerna ut att följa varandra väl för alla värden på den förklarande variabeln. Även i den tredje bilden följer linjerna varandra relativt väl, men för låga fitted values så glider linjerna isär. Sammantaget från figur 16 drar vi slutsatsen att vi har modellerat vår data på ett tillfredsställande sätt.



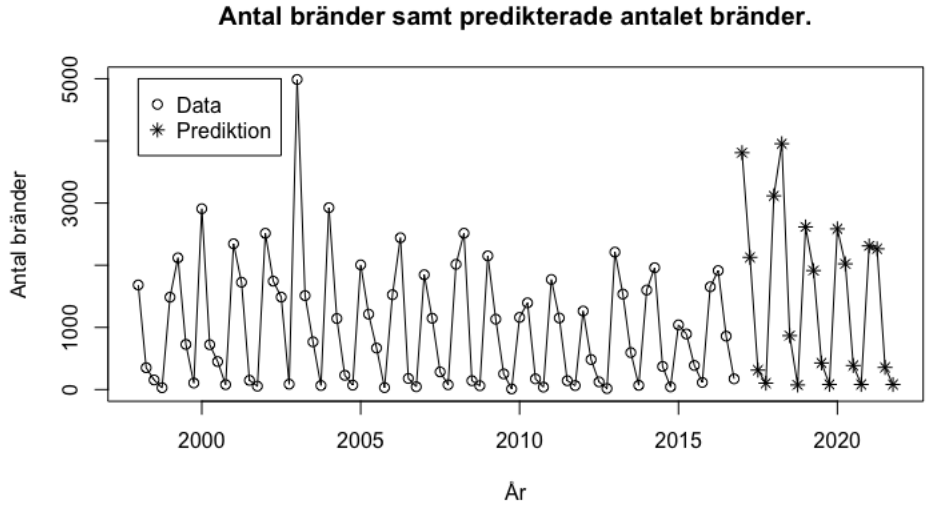
Figur 16: Figuren visar Marginal Model Plots för regressionsmodellen från ekvation 21.

Sammanfattningsvis ser log-modellen ut att uppfylla de flesta modellantaganden för en linjär regressionsmodell. Residualerna kan antas vara normalfördelade och med konstant varians. Dock finns svagt signifikant autokorrelation för sex och tolv perioder tillbaka, vilket kan komma att påverka tillförlitligheten i prognoserna. Vi finner inga oroväckande tecken på alltför inflytelserika observationer och Marginal Model Plots indikerar att vi har modellerat datamaterialet tillfredsställande. Eftersom att vi har logaritmerat Y och är intresserad av att förklara variationen i Y och inte log(Y), genomför vi en transformation av $\hat{y}$ och beräknar ett justerat R^2 i originalskala. Log-modellen förklarar 8% av variationen i Y vilket är relativt lågt. Våra parameterskattningar är heller inte alla signifikanta. Detta är tydliga tecken på att vår färdiga modell är långt i från den optimala.

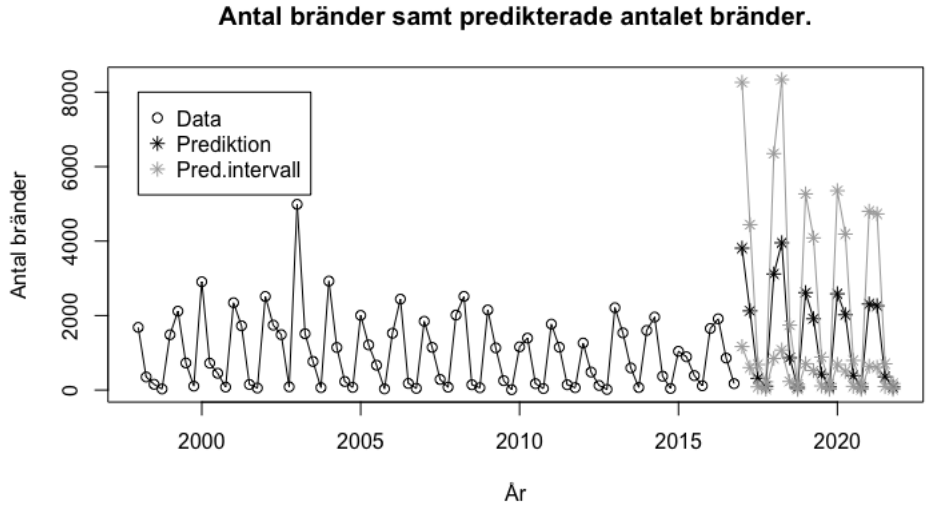
4.3 Prognoser

4.3.1 Antalet bränder

Figur 17 nedan visar antalet bränder samt våra predikterade värden för de kommande årens antal bränder. Figur 18 visar samma sak fast med tillhörande prediktionsintervall. Som vi kan se följer de prognosticerade värdena det generella mönstret, med undantag för våren och sommaren 2018 som sticker ut. 2018 var, som beskrevs i inledningen, det mest brandfyllda året i modern tid i Sverige. Modellen predikterar relativt höga antal för våren och sommaren 2018, men överskattar ändå det faktiska utfallet för våren och underskattar utfallet för sommaren, vilket kan ses i tabell 3 nedan.



Figur 17: Figuren visar antalet bränder för perioden våren 1998 till och med vintern 2016 samt våra predikterade värden för perioden våren 2017 till och med vintern 2021 för modellen i ekvation 19.



Figur 18: Figuren visar antalet bränder för perioden våren 1998 till och med vintern 2016, våra predikterade värden samt tillhörande prediktionsintervall för perioden våren 2017 till och med vintern 2021 för modellen i ekvation 19.

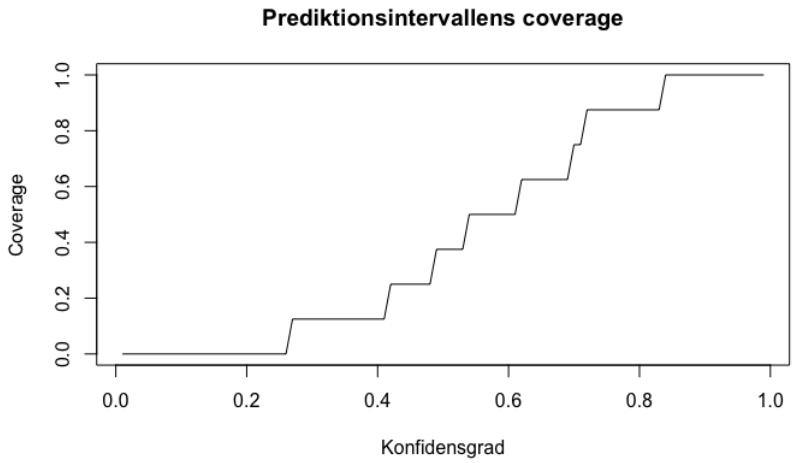
I tabell 3 nedan återfinns prognoser för antalet bränder de kommande fem åren från den Negativa binomial-modellen. Som tidigare nämnts delade vi upp datamaterialet i två delar där vi reserverade de sista två åren (i.e. 2017 och 2018) för att användas i validerande syfte av precisionen i våra prognoser. I tabell 3 redovisas antal prognostiserade bränder och ett 95% prediktionsintervall för dessa samt det faktiska antalet bränder för vardera säsong under åren 2017 och 2018. För att beräkna prediktionsintervallen använder vi 2000 simuleringar utförda med Bootstraping, likt Liboschik et al. (2017).

År		Punktskattning	Nedre gräns	Övre gräns	Faktiskt antal bränder
2017	Vår	3810	1213	8142	2386
	Sommar	2127	639	4566	1623
	Höst	313	82	690	153
	Vinter	103	28	224	43
2018	Vår	3116	883	6460	1706
	Sommar	3955	1022	8145	4540
	Höst	867	217	1825	531
	Vinter	74	18	153	35
*2019**	Vår	2616	675	5652	
	Sommar	1914	469	3892	
	Höst	428	113	902	
	Vinter	83	23	176	
*2020**	Vår	2585	652	5434	
	Sommar	2023	531	4267	
	Höst	387	95	805	
	Vinter	82	20	180	
*2021**	Vår	2311	620	4773	
	Sommar	2263	611	4747	
	Höst	356	99	769	
	Vinter	82	18	172	

Tabell 3. Tabellen visar punktskattningar för antalet bränder samt 95%-prediktionsintervall uppdelat per säsong för åren 2017 till 2021. Tabellen visar även de faktiska värdena för åren 2017 och 2018. De år markerade med asterix saknar faktiska värden.

Av tabell 3 ser vi att det egentliga antalet bränder under 2017 och 2018 täcks in av samtliga prediktionsintervall. Intervallen i sig är relativt breda och våra punktskattningar är, med undantag för sommaren 2018, genomgående högre än det faktiska antalet bränder. Sammantaget har vi skattat antalet bränder i genomsnitt 30% högre än vad det faktiska utfallet har varit.

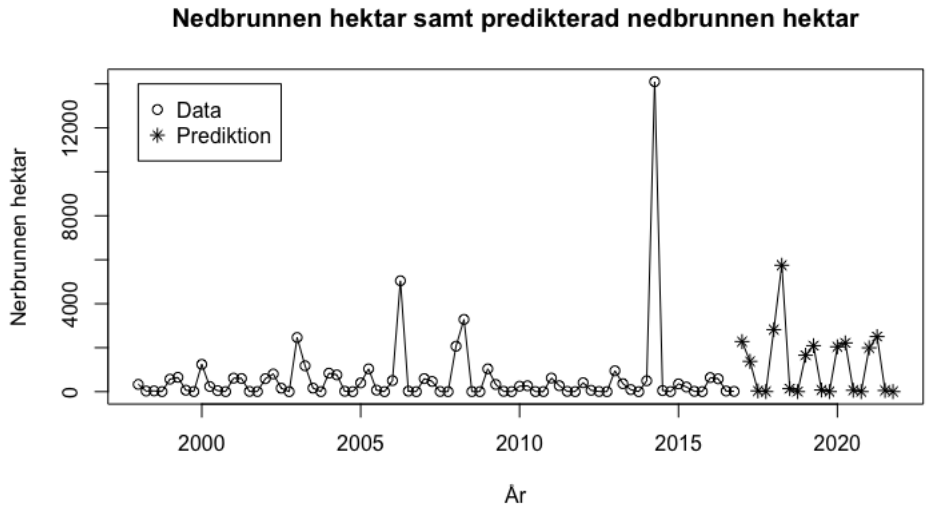
I figur 19 nedan har vi skapat en Coverageplot som visar hur mycket av de faktiska värdena under 2017 och 2018 som prediktionsintervallen täcker in för olika konfidensgrader med start på 1% och i steg om 1 procentenheter. Vi tittar på detta eftersom att våra beräknade prediktionsintervall är så breda att vi misstänkte att det kunnde blivit fel någonstans i beräkningarna. Från figur 19 finner vi dock inget som tyder på uppenbara fel.



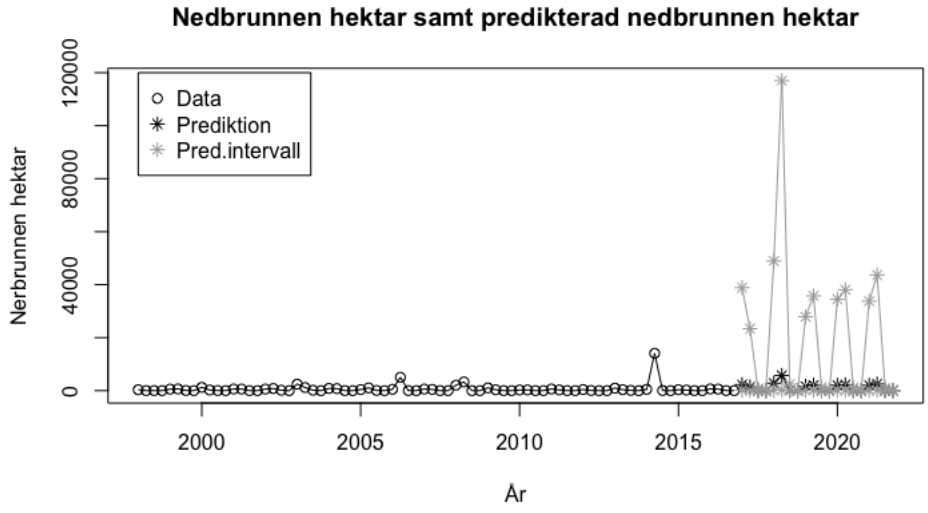
Figur 19: Figuren visar en plot över prediktionsintervallens *Coverage*, d.v.s hur många av de faktiska värdena som intervallen täcker in. Konfidensgraderna är i steg om 1 procentenheter.

4.3.2 Total nedbrunnen hektar

I figur 20 kan vi se den totala nedbrunna hektaren skog samt våra motsvarande prediktioner för kommande års säsonger, i figur 21 visas samma sak fast med tillhörande 95%-prediktionsintervall. Vi kan se att de flesta predikterade värden följer det generella mönstret. Dock kan man tydligt se att sommaren 2018 avviker från det generella mönstret vilket innebär att modellen predikterar ett relativt högt antal nedbrunnen hektar denna säsong. Som beskrivits tidigare var 2018 ett år där väldigt mycket skog brann. Både antalet bränder och den totala nedbrunna arealen var ovanligt stor. Log-modellen lyckas prediktera ett värde som avviker från det normala för sommaren 2018, men det faktiska utfallet var mer än fyra gånger så stort som det värde modellen predikterar.



Figur 20: Figuren visar antalet bränder för perioden våren 1998 till och med vintern 2016 samt våra predikterade värden för perioden våren 2017 till och med vintern 2021 för modellen i ekvation 21.



Figur 21: Figuren visar antalet bränder för perioden våren 1998 till och med vintern 2016, våra predikterade värden samt tillhörande prediktionsintervall för perioden våren 2017 till och med vintern 2021 för modellen i ekvation 21.

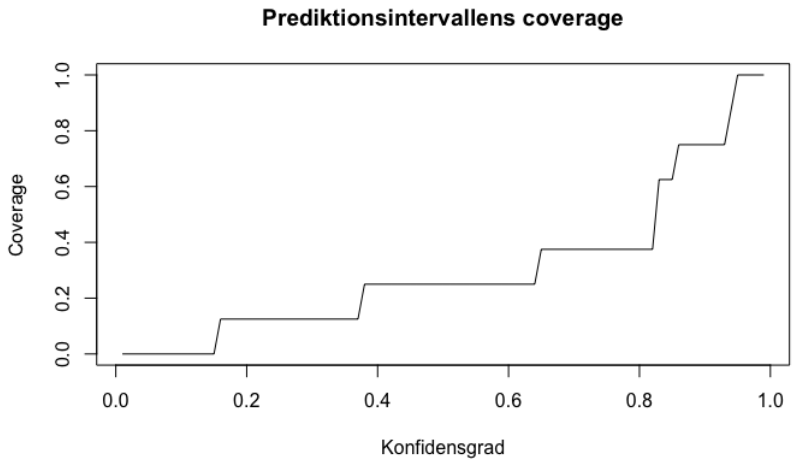
I tabell 4 nedan återfinns en sammanställning över log-modellens prognoser för den totala nedbrunna hektaren skog per säsong för perioden 2017 till 2021. Precis som för den Negativa binomial-modellen används de faktiska utfallen för 2017 och 2018 för att validera prognoserna. Tabellen innehåller även 95%-prediktionsintervall. I tabellen ser vi att de skattade prediktion-

sintervallen täcker samtliga faktiska värden för den nedbrunna hektaren under åren 2017 och 2018. Dock är prediktionsintervallen, precis som för den Negativa binomial-modellen, även här relativt breda. Detta utmärker sig väldigt tydligt i figur 21. Punktskattningarna ligger också konsekvent högre än vad det faktiskt har brunnit, med undantag för sommaren 2018 då det brann ovanligt mycket och då modellen underskattade det faktiska utfallet.

År		Punktskattning	Nedre gräns	Övre gräns	Faktisk nedbrunnen hektar
2017	Vår	2278,81	133,6	38871,12	1132,4
	Sommar	1377,25	81,17	23368,75	200,7
	Höst	35,77	2,04	628,2	26,8
	Vinter	3,5	0,19	62,76	0,2
2018	Vår	2820,25	162,6	48916,13	389,6
	Sommar	5748,88	282,50	116988,2	23736
	Höst	136,37	7,16	2443,20	8,5
	Vinter	1,15	0,07	19,51	0,1
*2019**	Vår	154,53	98,2	27879,54	
	Sommar	2092,12	122,40	35759,54	
	Höst	74,97	4,37	1285,35	
	Vinter	1,85	0,11	31,38	
*2020**	Vår	2041,26	120,70	34521,83	
	Sommar	2217,06	129,1	38075,38	
	Höst	62,50	3,66	1066,90	
	Vinter	2,01	0,12	34,13	
*2021**	Vår	1994,19	117,81	33756,18	
	Sommar	2506,23	143,95	43635,64	
	Höst	53,5	3,14	911,09	
	Vinter	2,21	0,129	37,84	

Tabell 4. Tabellen visar punktskattningar för den totala nedbrunna hektaren skog samt 95%-prediktionsintervall uppdelat per säsong för åren 2017 till 2021. Tabellen visar även de faktiska värdena för åren 2017 och 2018. De år markerade med asterix saknar faktiska värden.

I figur 22 nedan har vi inkluderat en Coverageplot för log-modellen för olika konfidensgrader med start på 1% och i steg om 1 procentenheter. Även för log-modellen är våra beräknade prediktionsintervall så breda att vi misstänkte att det kunde blivit fel någonstans i beräkningarna. Från figur 22 drar vi dock samma slutsats som tidigare att vi inte ser några uppenbara tecken på fel.



Figur 22: Figuren visar en plot över prediktionsintervallens *Coverage*, d.v.s hur många av de faktiska värdena som intervallen täcker in. Konfidensgraderna är i steg om 1 procentenheter.

5 Diskussion

Som vi kunde se i resultatdelen så tenderar log-modellens punktskattningar att överskatta den totala nedbrunna hektaren skog, med undantag för sommaren 2018. Med tanke på att det justerade R^2-värdet i originalskala för log-modellen ligger på ca 8% är det kanske inte så förvånande att modellens prognoser inte är särskilt precisa. Eftersom att vi har ett antal värden med högt leverage i modellen kan det även vara så att det beräknade R^2-värdet är högre än vad det faktiskt är. Orsaken till att skogsbränder uppstår och sprider sig är komplext och det finns garanterat fler faktorer än temperatur och nederbörd som påverkar, t.ex vind. Inkluderandet av fler förklarande variabler skulle troligtvis förbättra precisionen i prognoserna och med utgångspunkt i det vi har läst om skogsbränders orsaker så tror vi att vind är högst relevant att inkludera då vindstyrka har stor påverkan på hur omfattande en skogsbrand blir. När vi i efterhand har studerat figur 3 och 4 i inledningen närmare tror vi att inkluderandet av en interaktionsterm mellan säsong och medeltemperatur även kan vara värt att beakta i en framtida undersökning. Även en ökning av datamaterialet skulle förbättra precisionen i prediktionsintervallen genom att sänka variansen, antingen genom att undersöka en längre tidsperiod eller genom att studera t.ex månadsdata istället för säsongsdata. Ett annat angreppssätt och en annan metod för att modellera hektaren nedbrunnen skog skulle också kunna vara ett sätt att förbättra precisionen i prognoserna, men på grund av uppsatsens tidsbegränsning har detta inte varit möjligt. Vi har prövat att bygga modellen på det fulla datamaterialet, d.v.s att inkludera 2017 och 2018, och detta ökar det justerade R^2-värdet i originalskala till ca 29%, vilket är en ganska drastisk ökning från 8%. Detta tyder på att 2017 och 2018 är viktiga år för att förklara det generella mönstret. Att inkludera dessa år i modellbyggandet hade dock inte gett oss några möjligheter att validera våra prognoser.

Fortsättningsvis så visade vi under resultatdelen figurer över PIT-histogrammet och *Marginal Calibration Plot* för vår Negativa binomial-modell som har till syfte att utvärdera en modells prognosförmåga. Till att börja med var PIT-histogrammet inte önskvärt uniformt fördelat och för det andra var det stora fluktuationer i *Marginal Calibration Plot* kring värdet noll. Tillsammans antyder detta att vår modell inte besitter den prediktiva precisionen som vi önskar. Detta återspeglas även i de prediktionsintervall som vi har beräknat som genomgående är väldigt breda. Vidare speglas även modellens nedsatta prediktionsförmåga i dess punktskattningar som för säsongerna under åren 2017 till 2018 genomgående överskattade de faktiska utfallen.

Det är väntat att punktskattningarna av våra prognosticerade värden ska missa de faktiska utfallen, även om vi anser att dessa avviker lite väl mycket för att vi ska vara nöjda med modellerna. Samtliga prediktionsintervall täcker de faktiska utfallen, men dessa är samtidigt väldigt breda, t.ex för sommaren 2018 då intervallet spänner mellan 282 och 116 988 hektar. I praktiken blir dessa intervall inte särskilt användbara då de täcker in de flesta möjliga utfall.

När det kommer till våra prognosticerade värden för åren 2019, 2020 och 2021 har vi inga möjligheter att validera dessa. Vi har inte lyckats finna några prognoser för kommande års skogsbränder från någon myndighet eller dylikt att jämföra med. Den slutsatsen vi däremot kan dra är att punktskattningarna för prognoserna för 2017 och 2018 genomgående har missat de faktiska utfallen och med största sannolikhet även kommer missa framtida utfall. En extra osäkerhet infinner sig även i dessa prognoser eftersom att de värden för våra förklarande variabler, medeltemperatur och nederbörd, som vi använder för att göra prognoser för dessa

år i sig är skattningar. Vi kan troligtvis vara rätt säkra på att de prognosticerade värden för kommande års säsongsnederbörd eller medeltemperatur vi har använt inte kommer att stämma överens med de faktiska utfallen. Därmed hade våra prognosticerade värden varit något annat om vi idag hade haft tillgång till de faktiska utfallen för temperatur och nederbörd i framtiden, vilket uppenbarligen är omöjligt. Dock kan det vara så att de faktiska utfallen för både antalet skogsbränder och nedbrunnen hektar skog för 2019, 2020 och 2021 täcks in av våra prediktionsintervall tack vare att de är så breda.

Eftersom vi inte har kunnat finna några forskningsrapporter rörande framtidens utveckling av skogsbränder lämnar vi följande förslag på vidare forskning inom det här ämnet.

Som vi i inledningen kunde se i figur 1 och 2 verkar både antalet skogsbränder och den totala nedbrunna ytan vara relativt konstant över tid. Däremot inträffar då och då extrema säsonger med antingen ett väldigt högt antal bränder eller en stor nedbrunnen areal eller både och, som exempelvis under sommaren 2018 som hade både ett högt antal bränder och var i omfattning den värsta branden i modern tid i Sverige. Därför skulle ett fortsatt forskningsarbete kunna handla om att utveckla våra modeller genom att använda sig av extremvärdesteori för att kunna fånga upp dessa extrema säsonger som inträffar relativt sällan. Ett andra förslag är att närmare undersöka huruvida det finns ett verkligt samband mellan säsonger som tidsmässigt är långt ifrån varandra. Eftersom vi visade i figur 13 att det finns en signifikant autokorrelation för säsonger långt tillbaka i tiden, men vi tror att denna kan avfärdas som spökkorrelation, skulle en sådan insikt kunna förbättra prognoserna för kommande år.

Referenslista

Agresti, A. (2007). *Introduction to Categorical Data Analysis*. 2. ed. Hoboken, N. J.: Wiley.

Chatfield, C. (2011). *Time Series Forecasting*. Boca Raton: Chapman Hall/CRC.

Cryer, J.D & Chan, K-S. (2008). *Time Series Analysis With Applications in R*. 2. ed. New York, NY: Springer.

Czado C., Gneiting T. & Held L. (2009). *Predictive Model Assessment for Count Data*. Biometrics. 65(4), s.1254-1261.

Gneiting, T., Balabdaoui, F. & Raftery, A. (2007). *Probabilistic forecasts, calibration and sharpness*. Journal Of The Royal Statistical Society. 69(2):243 - 268.

Granström, A. & Axelsson, A-L. (2018). *SLU-forskare svarar på frågor om skogsbränder*, SLU-nyhet. Tillgänglig via:
`https://www.slu.se/ew-nyheter/2018/8/skogsbrander/` [Hämtad 2019-10-02].

Hogg, Robert V., Tanis, Elliot A. & Zimmerman, Dale L. (2014). *Probability and statistical inference*. 9. ed., Global edition. Boston: Pearson.

O'Leary, D. & Rust, B. (2008). *Residual Periodograms for Choosing Regularization Parameters for Ill-Posed Problems*. Inverse Problems. 24(3).

Liboschik, T., Fokianos, K. & Fried, R. (2017). *tscount: An R package for analysis of count time series following generalized linear models*. Journal of Statistical Software 82(5), 1–51. Tillgänglig via:
`https://cran.r-project.org/web/packages/tscount/vignettes/tsglm.pdf` [Hämtad 2019-10-11].

Myndigheten för samhällsskydd och beredskap. (2019). *Bränder i skog eller mark*. Tillgänglig via:
`https://ida.msb.se/ida2#fbclid=IwAR3IdvDTUjdUNN-ImBKqxQRvBJfMmN9OeEWLNjGL8moj6j09_F6OiqgZk64&page=f5943eeb-e86c-4342-af56-f2de8c55883d` [Hämtad: 2019-09-15].

Sheather, S. (2010). *A Modern Approach To Regression With R*. New York: Springer-Verlag New York Inc.

Skogsstyrelsen. (2019). Mer om skog/Skogsbranden i Västmanland 2014/Fakta om branden. Tillgänglig via:
`www.skogsstyrelsen.se/mer-om-skog/skogsbranden-i-vastmanland-2014/fakta-om-branden/` [Hämtad 2019-10-02].

SkogsSverige. (2019). Hem/Skog/Fakta om skog/Skogsbrand. Tillgänglig via:

https://www.skogssverige.se/skog/fakta-om/skogsbrand [Hämtad 2019-10-02].

SMHI.(2019a). *Klimatindikator - Temperatur*. Tillgänglig via: https://www.smhi.se/klimat/klimatet-da-och-nu/klimatindikatorer/klimatindikator-temperatur-1.2430 [Hämtad: 2019-09-15].

SMHI. (2019b). *Klimatindikator - Nederbörd*. Tillgänglig via: https://www.smhi.se/klimat/klimatet-da-och-nu/klimatindikatorer/klimatindikator-nederbord-1.2887 [Hämtad: 2019-09-15].

SMHI. (2019c). *Ladda ner scenariodata*. Tillgänglig via: https://www.smhi.se/klimat/framtidens-klimat/ladda-ner-scenariodata/ [Hämtad: 2019-10-15].

SOU 2019:7. Skogsbränderna sommaren 2018. *Betänkande av 2018 års skogsbrandsutredning*. Tillgänglig via: https://www.regeringen.se/4906d2/contentassets/8a43cbc3286c4eb39be8b347ce78da16/skogsbranderna-sommaren-2018-sou-2019-7.pdf [Hämtad 2019-10-02].

SVT. (2019). *Prognos: Kraftig ökning av skogsbränder i Sverige*. Tillgänglig via: https://www.svt.se/nyheter/vetenskap/prognos-kraftig-okning-av-skogsbrander-i-sverige [Hämtad: 2019-09-23].

Yang, J. (2012). Interpreting Coefficients in Regression with Log-Transformed Variables. *Cornell Statistical Consulting Unit, Cornell University*. Tillgänglig via: https://www.yumpu.com/en/document/view/51760991/statnews-83-interpreting-coefficients-in-regression-with-log- [Hämtad 2019-10-18].

Zeileis, A., Kleiber, C. & Jackman, S. (2008). *Regression Models for Count Data in R*. Journal of Statistical Software 27(8).