Stockholms universitet

Hur påverkar pollenhalten söktrender på internet?

Elin Magnusson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2024:7
Matematisk statistik
Maj 2024

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2024:7**
http://www.math.su.se/matstat

Hur påverkar pollenhalten söktrender på internet?

Elin Magnusson*

Maj 2024

Sammanfattning

Denna studie ämnar att undersöka sambanden mellan uppmätta pollenhalter och söktrender genom att använda data tillhandahållen av Naturhistoriska riksmuseet och Google. Det ena syftet med denna studie är att ta reda på vilka pollenarter som påverkar söktrenderna mest medan det andra syftet är att jämföra de olika metoderna för att på så sätt avgöra vilken som ger bäst resultat. De pollentyper som undersöks är de som övervakas på pollenrapporten.se, det vill säga al, alm, björk, ek, gräs, gråbo, hassel samt sälg och viden. Gällande söktrender har vi specifikt valt ut ordet ”pollen” som sökord. Sambanden undersöks med hjälp av enkel- och multipel linjär regression, Poisson-regression, Quasi-Poisson-regression, stegvis regression, minsta kvadratmetoden med icke-negativa begränsningar och beräkning av korrelationskoefficienter. För att avgöra vilken modell som beskriver datan bäst används AIC och den justerade förklaringsgraden R^2_{adj}. Våra resultat påvisar att samtliga pollenarter utöver gråbo har en signifikant och positiv påverkan på pollensökningarna, varav björk och gräs är de två pollenarter som ger högst ökning till pollensökningarna.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: elin.karolina@outlook.com. Handledare: Martin Sköld.

Abstract

This study aims to investigate the relationships between measured pollen levels and search trends, using data provided by the Swedish Museum of Natural History and Google. One purpose of this study is to find out which pollen species affect the search trends the most, of which the second purpose is to compare the different methods in order to determine which one gives the best results. The pollen species that are examined are the those that are mentioned at `pollenrapporten.se`, namely alder, elm, birch, oak, grass, mugwort, hazel and willow. Regarding search trends, we have specifically selected the word "pollen" as a search term. The relationships are examined using simple and multiple linear regression, Poisson regression, Quasi-Poisson regression, stepwise regression, the least squares method with non-negative constraints and calculation of correlation coefficients. To determine which model describes the data best, AIC and the adjusted coefficient of determination R^2_{adj} are used. Our results suggests that all pollen species apart from mugwort have a significant and positive influence on pollen searches, of which birch and grass are the two pollen species that give the highest increase to pollen searches.

Förord

Denna studie är skriven vid matematiska institutionen på Stockholms universitet i samarbete med Naturhistoriska riksmuseet och utgör ett examensarbete om 15 högskolepoäng. Jag vill rikta ett stort tack till Martin Sköld för att han under arbetets gång stöttat och bidragit med idéer, hjälp och återkoppling.

Innehåll

1 Inledning

Cirka 30 procent av Sveriges befolkning lider av pollenallergi. Nästäppa, trötthet och rinnande ögon är tre vanliga symptom vid allergi orsakad av luftburen pollen. De flesta med pollenallergi lider av dessa symptom under en kortare tid motsvarande en månad, men det finns de som lider av svåra allergiska reaktioner och kan således inte klara av exempelvis jobb eller skola under flera månaders tid (Luther, 2024).

Användningen av internet har ständigt ökat och kommer mest troligen även fortsätta göra det då det har blivit en viktig del i vår vardag. Genom att undersöka trender för diverse sökord så som ”pollen” eller ”snuva” och finna deras toppar skulle man kunna utveckla varningssystem för framför allt de som är allergiska och därmed kunna informera om olika åtgärder för att de som lider av pollenallergi ska exponeras för pollen så lite som möjligt.

Denna studie avser att undersöka huruvida det råder ett samband mellan uppmätta pollenhalter och söktrender genom att använda data tillhandahållen av Naturhistoriska riksmuseet och Google med hjälp av korrelationsanalys och olika regressionsmodeller. Vidare har denna studie två syften. Den ena är att ta reda på vilka pollenarter som påverkar söktrenderna mest medan det andra syftet är att jämföra de olika metoderna för att avgöra vilken metod som erhåller bäst resultat. De pollenarter som kommer undersökas är al, alm, björk, ek, gräs, gråbo, hassel och sälg och viden. Vidare är studien geografiskt begränsad till Stockholm.

Det har tidigare genomförts liknande studier. Holmer (u.å.) genomförde en studie om pollenhalters relation till olika webbsökningar genomförda på 1177, tidigare kallad Vårdguiden, där han bland annat beräknade korrelationskoefficienter. Holmer använde sig av flera olika webbsökningar så som ”pollen” och ”hösnuva” varav hans slutsats var att björk var den pollenart som hade högst samband med webbsökningen ”pollen” samt att det fanns ett samband för både al och ek med samma webbsökning. I en annan studie genomförd av Sitaru et al. (2019) undersökte de sambanden mellan söktrender, pollen och försäljning av antihistamin i Tyskland och Sverige med hjälp av korrelationsanalys. Precis som i denna studie har de tagit del av data från Google Trends, men istället för daglig data som vi ska använda oss av har de använt sig av månadsvis data. Vidare har de även tagit med klimatfaktorer som nederbörd och temperatur vilket vi inte gör. Även här fann de att björkpollen hade ett tydligt samband med pollensökningar. Till skillnad från Holmer (u.å.) och Sitaru et al. (2019) kommer denna studie genomföra korrelationsanalys för säsongsjusterad data.

2 Datamaterial

2.1 Insamling av datamaterial

2.1.1 Pollen

Sedan 1973 har Naturhistoriska riksmuseet kontinuerligt övervakat pollen vid Palynologiska laboratoriet i Stockholm. Under 2023 uppmättes det pollenhalter på 22 olika platser runt om i Sverige, varav mätstationen i Stockholm är den mätstation som har en av de längsta mätserierna i världen av luftburen pollen. Denna mätning sker med hjälp av en Burkard pollenfälla av Hirst-typ (Hirst, 1952). Processen börjar med att fällan öppnas och förses med en ny, klistrig tejprulle. Denna rulle sitter på en urverksdriven rotor som får gå ett varv på 48 timmar. Samtidigt suger en fläkt in cirka 10 liter luft per minut genom en smal spalt som ska motsvara hur en människa andas. Detta gör att pollenkornen fastnar på tejpen med ett varierande mönster med olika densitet beroende på hur mycket pollen det befann sig i luften vid de olika tidpunkterna. Varje morgon under pollensäsongen töms fällan och därefter byts trumman ut mot en ny trumma med ny tejp som inte har blivit exponerad för pollen. Därefter klipps den exponerade tejpen i 48 millimeter långa bitar där varje bit motsvarar ett dygn för att sedan placeras på ett objektglas där det tillsätts ett färgämne och till sist ett täckglas. Detta preparat studeras sedan i ett mikroskop med 400 gångers förstoring där pollenkornen räknas till art och antal varav summan motsvarar ett dygnsmedelvärde som anger mängden av en specifik pollen per kubikmeter i luften (Pollenrapporten, 2017). För denna studie kommer vi använda dagliga pollenräkningar i Stockholm mellan åren 2014-2023. De pollenarter som kommer undersökas är al, alm, björk, ek, gråbo, gräs, hassel och sälg och viden vilka är våra vanligaste allergiframkallande pollenarter (Pollenrapporten, 2015).

2.1.2 Google Trends

Google Trends data finns fritt tillgängligt att ladda ned från internet och rapporteras som ett slumpmässigt urval av historiska sökningar gjorda på Google där sökresultaten är proportionell till tid och plats för den gjorda sökningen. För att kunna bestämma den relativa populariteten divideras varje datapunkt med det totala antalet sökningar inom det valda tidsintervallet respektive plats. Därefter justeras sökresultaten att falla mellan ett intervall från 0-100 där 100 motsvarar det maximala sökintresset för den valda tidsperioden och platsen. Datan utesluter dubblettsökningar vilket innebär att upprepade sökningar gjorda av samma person under en kort tidsperiod inte finns representerad. Vidare är det möjligt att olika platser visar samma sökintresse men att de inte har samma totala sökvolym (Google, u.å.).

Sökintresse för termen ”pollen” från 2014-2023 i Stockholms län med daglig data hämtades med hjälp av Pytrends som är en inofficiell Google Trends API som tillhandahåller olika metoder för att ladda ned resultat från Google Trends. Vidare gäller det att sökresultaten inkluderar sökningar som innehåller fler ord

än just det valda sökordet. Exempelvis skulle sökningen ”pollen symptom” ingå i sökresultaten och inte endast sökningen ”pollen”.

2.2 Förståelse av data

2.2.1 Pollen

Datamaterialet tillhandahållet består av observationer av pollenräkningar mätt från 1973-03-13 fram till 2024-02-03 motsvarandes 10 815 unika dagar från 8 stycken olika pollenarter. De dagar där det inte har uppmätts några pollenkorn finns inte med i datamaterialet. Vidare finns det 5 variabler, nämligen `date`, `station`, `swe_name`, `count` och `factor`. Av dessa används inte `factor` (Tabell 1).

Tabell 1: Beskrivning av de olika variablerna i datamaterialet om pollen.

Variabel	Typ	Beskrivning
date	Diskret	Det datum pollenräkningarna registrerades.
station	Kategorisk	Den geografiska plats som pollenkornen fångades in.
swe_name	Kategorisk	Det svenska namnet för de registererade pollenräkningarna.
count	Diskret	Antalet pollenkorn räknat.
factor	Kontinuerlig	Referensvariabel för storleken på det använda mikroskopet.

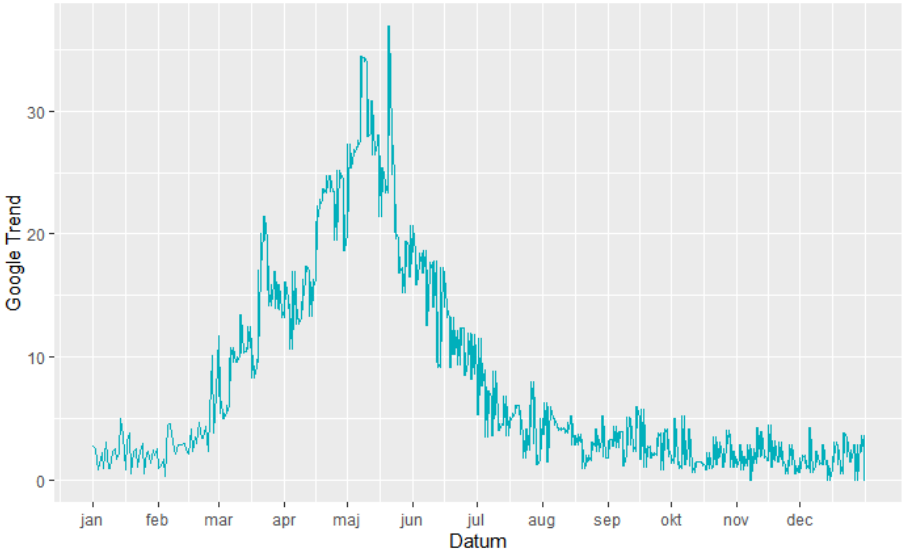
Figur 1: Genomsnittlig dagligt antal av räknade pollenkorn under perioden 1973-2024.

2.2.2 Google Trends

Datamaterialet innehållandes sökvärden från Google består av 3652 observationer mätt från 2014-01-01 fram till 2023-12-31 där varje observation motsvarar en unik dag med sökvärden. Materialet består av de två variablerna `date` och `trend_value` (Tabell 2).

Tabell 2: Beskrivning av de olika variablerna i datamaterialet om söktrender på Google.

Variabel	**Typ**	**Beskrivning**
date	Diskret	Det datum pollenräkningarna registrerades.
trend_value	Kontinuerlig	Sökvärde för termen pollen.



Figur 2: Genomsnittligt dagligt sökvärde för pollen under perioden 2014-2023.

3 Teori och metod

3.1 Säsongsjustering

Teorin som beskrivs nedan fram till det utökade Dickey-Fuller testet beskrivs enligt Mazzi (2018). Tidsseriedata är data där samtliga observationer har gjorts över tid vid en specifik tidpunkt. Observationerna är dessutom insamlade med jämna mellanrum. Vidare gäller det att ordningen på observationen är viktig. Denna data kan därefter användas för att analysera mönster eller förändringar. En viktig skillnad mellan tidsseriedata och andra datatyper är tidskomponenten som tillåter oss att upptäcka trender och eventuellt kunna göra förutsägelser om framtiden. Syftet med säsongsjustering är att eliminera regelbundet återkommande säsongsvariationer för att ge en tydligare bild av det underliggande beteendet som annars skulle överskuggas av säsongsmässiga skillnader och därmed ge felaktiga resultat.

Trenden representerar antingen en ökning eller en minskning av värdet på tidsserien över tid och således säger vi att den har en uppåtgående eller nedåtgående trend. Datamaterialet med söktrenderna från Google har exempelvis en uppåtgående trend eftersom antalet sökningar har ökat med takt att allt fler använder sig av Google för att genomföra sökningar. Trender kan resultera i att medelvärdet varierar över tid. För att säkerställa att medelvärdet inte varierar kan man ta bort trenden vilket man kan göra på flera olika sätt. I denna studie kommer vi ta bort den genom att först uppskatta det årliga medelvärdet av observationerna för att sedan dividera varje observation med det.

Säsongsvariation är en egenskap hos en tidsserie där observationerna upprepas med regelbundna tidsintervall vilket innebär att variationen är periodisk. Dessa perioder kan exempelvis röra sig om ett dygn, en månad eller ett år. Som i fallen med pollenräkningarna är det tydligt att varje pollenart har en cykel på ett år vilka kan ses i Figur 1. Säsongsvariationen kan resultera i att variansen förändras över tid. För att ta bort denna variation kommer vi subtrahera det dagliga medelvärdet av observationerna från varje observation.

Stationäritet hos en tidsserie uppnås om den har konstant medelvärde och varians. Om tidsserien innehåller trender eller säsongsvariationer är dessa två krav ej uppfyllda vilket medför att tidsserien är icke-stationär. Om tidsserien ej uppnår stationäritet kommer de dragna slutsatserna endast vara giltiga för den aktuella perioden som undersökts. Således är det inte möjligt att tillämpa sina resultat och på så sätt generalisera dem på andra tidsperioder. Dessutom finns risken att slutsatserna är felaktiga om man använder sig av data som ej är stationär. Framtida och säkra prediktioner kommer därmed inte vara möjligt.

För att säkerställa att tidsserien är stationär kommer vi tillämpa det utökade Dickey-Fuller testet, vilket är uppnåt när tidsserien har konstant medelvärde och varians. Detta test kallas även för ADF-test och är ett av de vanligaste statistiska testerna för att kontrollera huruvida en tidsserie är stationär eller inte.

ADF-testet undersöker närvaron av en sådan kallad enhetsrot i tidsseriedatan som är en stokastisk trend som ger ett mönster som ej går att förutspå. Nollhypotesen är att det existerar en enhetsrot som gör datan icke-stationär medan den alternativa hypotesen är att det inte finns en enhetsrot vilket gör datan stationär. Om p-värdet för ADF-testet understiger signifikansnivån 0.05 förkastas nollhypotesen om icke-stationäritet. Mer specifikt kan modellen för tidsserien skrivas som

$$Y_t = pY_{t-1} + \epsilon_t,\ t = 1, 2, ..., n$$

där n är antalet observationer, $Y_0 = 0$, p är enhetsroten och $\{\epsilon_t\}$ är sekvensen av oberoende, normalfördelade residualer med medelvärdet 0 och varians σ^2. När $t \to \infty$ konvergerar tidsserien mot en stationär tidsserie om $|p| < 1$. Om $|p| \geq 1$ är tidsserien inte stationär. Vidare gäller det att

$$\hat{p} = \left(\sum_{t=1}^{n} Y_{t-1}^2 \right)^{-1} \sum_{t=1}^{n} Y_t Y_{t-1}$$

(Dickey & Fuller, 1979).

3.2 Enkel linjär regression

Med enkel linjär regression undersöks sambandet mellan två variabler, nämligen en responsvariabel och en förklarande variabel. Modellen för enkel linjär regression definieras som

$$y_i = \beta_0 + \beta_i x_i + \epsilon_i$$

där y_i motsvarar responsvariabeln och x_i den förklarande variabeln. Vidare är β_0 interceptet och β_1 är koefficienten för den förklarande variabeln medan ϵ_i är residualen, det vill säga differensen mellan det observerade och predikterade värdet i regressionsmodellen. Vidare antas det ofta att residualerna är normalfördelade och oberoende med konstant varians samt väntevärde 0 (Sundberg, 2023). Dessa antaganden tillåter oss att dra pålitliga statistiska slutsatser om koefficienterna eftersom det säkerställer att p-värdena är korrekta.

Skattningen av interceptet β_0 och koefficienten β_1 ges av

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

respektive

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} y_i x_i - \left(\sum_{i=1}^{n} y_i \right) \left(\sum_{i=1}^{n} x_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

där

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

och

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Den uppskattade regressionslinjen blir därför

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

vilket medför att paret av observation i uppfyller

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

där

$$\epsilon_i = y_i - \hat{y}_i$$

är residualen som beskriver felet mellan det anpassade värdet och det faktiska värdet för y_i (Montgomery & Runger, 2011).

I den här studien är pollensökningarna responsvariabeln och pollenräkningarna den förklarande variabeln där vi vill undersöka om pollenräkningarna har en signifikant inverkan på pollensökningarna.

3.3 Multipel linjär regression

Montgomery och Runger (2011) beskriver multipel linjär regression enligt nedan. Multipel linjär regression används för att undersöka om det finns ett samband mellan responsvariabeln och flera olika förklarande variabler till skillnad från enkel linjär regression där det endast finns en förklarande variabel. Låt n vara antalet observationer och k vara antalet förklarande variabler. Modellen som definierar multipel linjär regression ges av

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i, \ i = 1, 2, ..., n$$

där β_0 är interceptet och β_j betecknar effekten av de förklarande variablerna x_{ij} på responsvariabeln medan ϵ_i är residualen för den i:te observationen. Denna regressionsmodell går även att skriva i matrisform som

$$Y = X\beta + \epsilon$$

där

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_1} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

För att hitta koefficienterna β_j vill vi minimera

$$\begin{aligned} L &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \epsilon^{\mathrm{T}} \epsilon \\ &= (Y - X\beta)^{\mathrm{T}} (Y - X\beta) \end{aligned}$$

där skattningarna av $\hat{\beta}$ ges av lösningarna till

$$\frac{\partial L}{\partial \beta} = 0.$$

Således ges skattningarna av $\hat{\beta}$ av

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y.$$

Därför blir den uppskattade regressionslinjen

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$$

där

$$\epsilon_i = y_i - \hat{y}_i$$

är residualen för observation i.

Multipel linjär regression kommer tillämpas vid stegvis regression och minsta kvadratmetoden med icke-negativa begränsningar.

3.4 Poisson-regression

Teorin för Poisson-regression fram till log-likelihoodskattningarna kommer från Agresti (2013). Poisson-regression tillämpas vid analys av räkningsdata som inträffar slumpmässigt över tid och antar att den förklarande variabeln Y följer en Poissonfördelning. Således kan den beroende variabeln endast anta värden som är positiva heltal och noll. Dess sannolikhetsfunktion ges av

$$P(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}$$

med egenskapen

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu > 0.$$

Vidare är Poissonfördelningen unimodal och positivt skev för små värden på μ där dess skevhet beskrivs av

$$\frac{\mathrm{E}(Y - \mu)^3}{\sigma^3} = \frac{1}{\sqrt{\mu}}.$$

När μ ökar minskar skevheten och följaktligen närmar sig Poissonfördelningen en normalfördelning. Detta börjar inträffa redan när $\mu \approx 10$. För låga värden på μ är således Poissonfördelning fördelaktig att använda då skattningarna blir pålitligare eftersom Poissonfördelningen tar hänsyn till att den förklarande variabeln inte kan anta negativa värden. Om antagandet om normalfördelning tas finns möjligheten att skatta sannolikheten för icke-positiva tal vilket i sin tur bidrar till falska utfall eftersom de i verkligheten inte kan inträffa.

Poisson-regression bygger på antagandet att responsvariabeln följer en Poissonfördelning och använder sig av den logaritmiska länkfunktionen vars syfte är att den alltid garanterar positiva värden för μ. Således är sambandet mellan responsvariabeln och den förklarande variabeln inte linjär utan loglinjär. Vidare predikteras logaritmen på μ. Låt k vara antalet förklarande variabler och n antalet observationer. Låt vidare $j = 1, 2, ..., k$ och $i = 1, 2, ..., n$. Då ges modellen för Poisson-regression av

$$\log\ \mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

vilket innebär att väntevärdet uppfyller det exponentiella sambandet

$$\mu_i = e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}$$

där β_0 är interceptet och β_j är koefficienten för den förklarande variabeln.

Låt $\boldsymbol{\beta}$ vara vektorn som innehåller regressionskoefficienterna β_j. För att beräkna de skattade koefficienterna $\hat{\boldsymbol{\beta}}$ börjar vi med att ta fram log-likelihood-funktionen för Poisson-regressionen som ges av

$$\begin{aligned} \ln\ L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i x_i^{\mathrm{T}} \boldsymbol{\beta} - \mu_i - \ln\ y_i! \right) \\ &= \sum_{i=1}^{n} \left(y_i x_i^{\mathrm{T}} \boldsymbol{\beta} - e^{x_i^{\mathrm{T}} \boldsymbol{\beta}} - \ln\ y_i! \right). \end{aligned}$$

Därefter deriverar vi ovanstående uttryck med avseende på $\boldsymbol{\beta}$ och sätter derivatan lika med 0 vilket ger oss

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln\ L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - e^{x_i^{\mathrm{T}} \boldsymbol{\beta}} \right) x_i = 0.$$

Att lösa ekvationen ovan för $\boldsymbol{\beta}$ kommer ge de maximerade likelihood-skattningarna för $\boldsymbol{\beta}$. Denna ekvation kräver dock iterativa metoder eftersom den inte har en analytisk lösning (Winkelmann, 2008).

3.5 Quasi-Poisson-regression

Ett problem med verklig räkningsdata är antagandet om att väntevärdet ska vara lika med variansen sällan är uppfyllt. Istället är det vanligt förekommande att variansen är större än väntevärdet. Detta kallas för överspridning vilket resulterar i felaktiga standardfel så väl som felaktiga p-värden för koefficienterna vilket gör att Poisson-regression inte är lämplig. Istället kan Quasi-Poisson-regression tillämpas för att hantera problemet med överspridning genom att man inkluderar den i modellen. Där tillåter man spridningsparametern variera istället för att vara fix till 1 vilket medför att sambandet mellan variansen och väntevärdet är

$$\mathrm{Var}(Y) = \phi \mathrm{E}(Y) = \phi\mu$$

där ϕ motsvarar en konstant som uppskattas med hjälp av den givna datan (Agresti, 2013). Vi kan uppskatta överspridningskonstanten med

$$\hat{\phi} = \frac{1}{n-p} \sum \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

där n motsvar antalet observationer och p antalet parametrar. Om $\phi > 1$ råder överspridning och om $\phi = 1$ är det inte överspridning. Vid Quasi-Poisson-regression erhålls samma parameterskattningar som vid Poisson-regression. Däremot ändras standardavvikelserna till

$$\sqrt{\hat{\phi}\frac{\mu}{n}}$$

där n motsvarar antalet observationer (Roback & Legler, 2021).

3.6 Stegvis regression

I en regressionsmodell kan det vara av intresse att välja ut specifika förklarande variabler istället för att inkludera alla för att på så sätt hitta en bra modell. Tillvägagångssättet är då att man successivt en i taget antingen minskar eller ökar antalet förklarande variabler fram till ett stoppkriterium är uppfyllt (Sundberg, 2023). I denna studie kommer vi börja med att inkludera alla förklarande variabler i modellen för att därefter, en efter en, fokusera på att exkludera de förklarande variabler som uppvisar negativa koefficienter. Anledningen till att vi gör detta är för det är orimligt att högre halter av pollenkorn bidrar till färre pollensökningar. Den stegvisa regressionen kommer att genomföras för de tre regressionsmetoderna multipel linjär regression, Poisson-regression så väl som Quasi-Poisson-regression.

3.7 Minsta kvadratmetoden med icke-negativa begränsningar

Ibland händer det att man vill använda sig av linjär regression men att man vill framtvinga teckenbegränsningar till sin modell. Detta är användbart om

man har starka föraningar om att vissa koefficienter ska ha ett visst tecken vilket gör minsta kvadratmetoden med icke-negativa begränsningar användbar. Denna metod kallas även för NNLS vilket står för non-negative least-square. Låt X vara matrisen som består av de förklarande variablerna, β kolumnvektorn innehållandes regressionskoefficienterna och Y vara kolumnvektorn med responsvariabeln. Då kan NNLS uttryckas som

$$\hat{\beta} = \min_{\beta} ||Y - X\beta||^2, \ \beta \geq 0$$

där $\beta \geq 0$ anger att varje komponent i vektorn β ska vara icke-negativ. Syftet med denna metod är att anpassa en modell som minimerar kvadratsumman av residualerna för att således ge den bästa modellen (Lawson & Hanson, 1995).

Matrisen X och kolumnvektorerna β respektive Y definieras som i multipel linjär regression vilka kan ses på sida 7.

3.8 Akaike informationskriterium

Akaikes informationskriterium, AIC, är ett statistiskt verktyg för att avgöra om en modell är lämplig för det givna datamaterialet. Den är användbar för att jämföra en modell med andra modeller för att bestämma vilken som är den mest lämpliga. Låt p vara antalet parametrar i modellen och $\hat{\theta}_{\mathrm{ML}}$ vara det maximerade värdet av likelihoodskattningen för modellen. För förtydligande är p antalet förklarande variabler inklusive intercept. Då definieras AIC som

$$\mathrm{AIC} = 2p - 2\ln\left(L(\hat{\theta}_{\mathrm{ML}})\right).$$

AIC belönar goodness of fit men bestraffar samtidigt överanpassning med antalet parametrar p. Eftersom en ökning av antalet parametrar i en modell generellt förbättrar goodness of the fit är denna bestraffning önskvärd. Således är den föredragna modellen den som erhåller lägst AIC-värde. AIC har ingen övre eller nedre begränsning och kan anta positiva så väl som negativa värden (Held & Bové, 2014).

3.9 Förklaringsgrad och justerad förklaringsgrad

Förklaringsgraden R^2 är ytterligare ett mått som förklarar hur väl en modell passar det givna datamaterialet där $0 \leq R^2 \leq 1$. Förklaringsgraden bestämmer andelen av variationen hos den beroende variabeln som förklaras av den oberoende variabeln. Ju högre värdet är för R^2, desto bättre passar modellen. Låt SSR beteckna regressionskvadratsumman och SST den totala kvadratsumman där

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

och

$$SST = \sum (y_i - \bar{y})^2$$

där $\bar{y}$ motsvarar medelvärdet av y. Då kan vi definiera förklaringsgraden som

$$R^2 = \frac{SSR}{SST}.$$

Förklaringsgraden brukar generellt sätt öka när fler förklarande variabler läggs till i modellen vilket kan medföra att förklaringsgraden blir missvisande. Således finns ett korrigerat R^2 som tar hänsyn till detta och kallas för justerad förklaringsgrad R^2_{adj} enligt

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1}$$

där n motsvarar antalet observationer och k antalet förklarande variabler (Sundberg, 2023). I denna studie kommer den justerade förklaringsgraden R^2_{adj} att tillämpas som mått.

3.10 Korrelation

Korrelationsanalys tillämpas för att undersöka om responsvariabeln och den förklarande variabeln samvarierar. I denna studie kommer Spearmans rangkorrelation tillämpas eftersom den är mer robust mot outliers än vad Pearsons rangkorrelation är. Dodge (2008) beskriver att Spearmans korrelationskoefficient ρ ges av

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

där n är antalet observationer och $d_i = x_i - y_i$ är differensen i rang för varje observation. Spearmans rangkorrelation kan anta värden mellan -1 och 1 där -1 anger maximalt negativt samband, 1 maximalt positivt samband och 0 inget samband. Tecknet på korrelationskoefficienten anger riktningen av det eventuella sambandet mellan de två variablerna. Korrelationskoefficienten ρ antar ett positivt värde om responsvariabeln tenderar att öka när den förklarande variabeln ökar medan ρ antar ett negativt värde om responsvariabeln tenderar att minska när den förklarande variabeln ökar. När $\rho \approx 0$ innebär det att responsvariabeln varken ökar eller minskar när den förklarande variabeln ökar. För att kunna dra en slutsats huruvida resultatet är statistiskt signifikant eller inte beräknas ett p-värde ut som anger hur sannolikt resultatet är. Anledningen till att korrelationsanalys tillämpas i denna studie är för att vi vill undersöka huruvida pollenräkningarna och pollensökningarna beror på varandra. Signifikansnivån som används i den här studien för att avgöra om det råder ett samband mellan de två variablerna är $p = 0.05$.

3.11 Rensning av data

Innan vi kan påbörja analysen behöver vi göra en rensning av datamaterialet för att resultatet inte ska påverkas av exempelvis saknade värden eller outliers. All datahantering sker i R (R Core Team, 2022). I och med att vi ska undersöka

åren 2014-2023 behöver vi först filtrera bort de rader som inte motsvarar datum från de åren. Eftersom vi har två olika datamaterial slår vi först ihop dem till ett datamaterial med hjälp av variabeln `date` för att värdena i kolumnerna ska matcha varandra och inte uppträda slumpmässigt. När vi gör detta får vi saknade värden i pollenräkningarna eftersom den inte innehåller nollräkningar utan endast registrerade värden som är större än noll. Därför ersätts saknade värden med 0.

3.11.1 Regressionsanalys

Till regressionsanalysen vill vi använda datamaterialet med noll-pollenräkningar. Det vill säga att vi kommer undersöka datamaterialet för alla dagar, oavsett om det har observerats pollen eller inte. Vidare kommer datamaterialet varken trend- eller säsongsjusteras, utan istället kommer den dagliga datan för pollenräkningarna att divideras med den totala summan för pollenräkningarna för respektive art. Syftet med detta tillvägagångssätt är för att få koefficienterna jämförbara mellan de olika pollenarterna eftersom vissa pollenarter är mer överrepresenterade än andra.

Anledningen till att vi inte säsongsjusterar det givna datamaterialet till regressionsanalyserna är för att vi kommer ta med årsvariabeln i regressionsmodellerna. Datum kan användas som en ersättare för en omätbar faktor. Som i exempelvis fallet med datamaterialet innehållandes sökvärdena från Google finns det en ökande trend eftersom antalet sökningar har ökat med takt att allt fler använder sig av Google. Eftersom det mest troligen inte finns ett mått för varumärkeskännedom kan tid istället användas som en förklarande faktor. Detta ger således de andra regressionskoefficienterna ett mer pålitligt resultat.

För att upptäcka och ta bort eventuella outliers inom sökvärdena för ”pollen” från Google används boxplot som använder sig av interquartile range criterion, IQR. Soetewey (2020) nämner att alla observationer som befinner sig ovan $q_{0.75} + 1.5 \cdot IQR$ eller nedan $q_{0.25} - 1.5 \cdot IQR$, där $q_{0.25}$ och $q_{0.75}$ motsvarar första respektive tredje kvartilen och IQR är differensen mellan den tredje och första kvartilen, ses som potentiella outliers. Sammanfattningsvis, alla observationer som inte befinner sig innanför intervallet

$$I = [q_{0.25} - 1.5 \cdot IQR, q_{0.75} + 1.5 \cdot IQR]$$

ses som potentiella outliers. Detta gör vi eftersom outliers har en tendens till att dra den skattade regressionslinjen bort från de övriga observationerna vilket i sin tur medför till försämrad passform och felaktiga parameteruppskattningar.

3.11.2 Korrelationsanalys

Korrelationsanalysen vill vi genomföra utan noll-pollenräkningar. Det vill säga att vi endast kommer undersöka datamaterialet för de dagar där det har observerats pollenräkningar som är större än noll. Ett skäl till varför nollorna

exkluderas från datan är för att vi vill fokusera på den tid på året där pollen är aktiv. Därför tar vi bort varje rad i datamaterialet där pollenräkningarna representeras av en nolla. Vidare vill vi ta bort trend- och säsongskomponenter eftersom datamaterialen är tidsseriedata då samtliga observationer har gjorts över tid vilket gör att dessa observationer kan uppvisa en trend. Exempelvis i vårt fall med söktrenderna från Google Trends kan antalet sökningar öka med takt att fler använder sig av Google för att söka på exempelvis "pollen" i och med att Google blir allt mer populär. Detta medför i sin tur att värdena för tidsserien kommer öka över tiden. För att kunna fånga upp andra mönster kommer vi behöva analysera datamaterialet utan denna trend. Därför tar vi bort den genom att dividera varje observation av pollenräkningarna med det årliga medelvärdet av pollenräkningarna och gör därefter samma sak med observationerna för sökvärdena från Google Trends. För att ta bort säsongskomponenterna subtraherar vi därefter det dagliga medelvärdet av pollenräkningarna från pollenräkningarna och upprepar sedan denna process för sökvärdena. För förtydligande säsongsjusterar vi datamaterialet mellan olika år och inte inom åren. Anledningen till att vi dividerar med det årliga medelvärdet är för att vissa år kan det vara extremt mycket pollen och att det i sådana fall kan generera ett år med mycket trend. Denna korrelation tar vi då bort för att begränsa oss till säsongsmönster istället.

Anledningen till att vi säsongsjusterar innan vi genomför korrelationsanalysen är för att vi vill uppnå stationäritet. Om datan inte är stationär kommer resultaten endast vara giltig för den aktuella perioden som undersöks, det vill säga åren 2014 till 2023. Således kommer det inte vara möjligt att generalisera och tillämpa dessa resultat på andra tidsperioder, åtminstone inte med säkra resultat. Vidare finns även risken att slutsatserna är felaktiga vilket medför att framtida och säkra prediktioner ej kommer vara möjligt. Syftet med att inte ta bort outliers till korrelationsanalysen är för att vissa observationer kan se ut som potentiella outliers trots att de egentligen inte är det eftersom de motsvarar riktig data och således bidrar till mer korrekt resultat. Dessutom kommer vi tillämpa Spearmans korrelationskoefficient vilken är robust mot eventuella outliers.

Korrelationsanalysen kommer genomföras för de enskilda pollenarterna, men vi kommer även slå samman olika pollenarter för att se om sambandet mellan dem och pollensökningarna möjligtvis ökar. De sammanslagningar som väljs ut baseras på de olika pollenarternas säsonger. Vi vill se till att täcka hela pollensäsongen istället för enskilda pollenarters säsonger. Exempelvis väljs björk, al, gräs och gråbo ut som en kandidat. Däremot kommer inte pollenarter med samma säsong slås ihop, så som björk och ek. För förtydligande kommer pollenräkningarna att adderas ihop för varje unik dag.

4 Resultat

Nedan kommer resultaten för de olika regressionsmodellerna respektive korrelationsanalysen att presenteras. Dessa regressionsmodeller kommer vi därefter jämföra, både med varandra och korrelationsanalysen, och diskutera i diskussionsdelen för att besvara denna studies två syften. Observera att de två kolumnerna year och count i Tabell 3, Tabell 4 och Tabell 5 avser de skattade koefficienterna för variablerna count respektive year där count motsvarar pollenräkningarna och year är det år som pollenräkningarna observerades. Även koefficienterna i Tabell 6, Tabell 7 och Tabell 8 är skattade. Vidare står SE för standardavvikelsen.

4.1 Enkel linjär regression

Tabell 3: Resultat av enkel linjär regression för att undersöka sambandet mellan olika pollenarter och söktrend för ordet "pollen".

Pollen	**year**	**count**	**SE(year)**	**SE(count)**	R^2_{ajd}	**AIC**
Al	0.3183	679.7283	0.0402	74.2416	0.0394	22 650
Alm	0.3268	405.1356	0.0405	56.1140	0.0306	22 681
Björk	0.3286	4256.3044	0.0393	277.2795	0.0795	22 505
Ek	0.3205	1032.2957	0.0400	95.2440	0.0486	22 618
Gråbo	0.3073	−181.9004	0.0406	87.8604	0.0170	22 729
Gräs	0.2990	1704.2730	0.0397	130.3383	0.0628	22 566
Hassel	0.3074	405.2267	0.0404	58.5024	0.0294	22 685
Sälg och viden	0.3183	1385.9869	0.0395	98.8519	0.0695	22 542

Tabell 3 representerar resultatet för enkel linjär regression där responsvariabeln är sökvärdena från Google och den förklarande variabeln är pollenräkningarna. Vidare har årsvariabeln tagits med i regressionsmodellen som ett mått för varumärkeskännedom för att få mer pålitliga koefficienter och ett pålitligare resultat. Varje rad motsvarar en modellanpassning. Exempelvis motsvarar första raden i tabellen en modellanpassning för alpollen. Det vill säga att endast pollenräkningarna för alpollen har tagits med i regressionsanalysen för just den modellen. Samtliga p-värden för koefficienter och modellen är 0.0000.

4.2 Poisson-regression

Tabell 4: Resultat av Poisson-regression för att undersöka sambandet mellan olika pollenarter och söktrend för ordet "pollen".

Pollen	**year**	**count**	**SE(year)**	**SE(count)**	R^2_{adj}	**AIC**
Al	0.0586	55.4200	0.0026	2.1610	0.0263	36 641
Alm	0.0601	43.9100	0.0026	2.0590	0.0234	36 724
Björk	0.0582	266.4000	0.0026	6.5930	0.0220	36 182
Ek	0.0573	87.5000	0.0026	2.9770	0.0314	36 482
Gråbo	0.0565	−42.5000	0.0026	7.0310	0.0162	36 991
Gräs	0.0541	174.6000	0.0026	4.8280	0.0462	36 121
Hassel	0.0571	26.2900	0.0026	1.3960	0.0170	36 844
Sälg och viden	0.0572	101.8000	0.0026	2.7100	0.0333	36 222

Precis som i Tabell 3 motsvarar varje rad i Tabell 4 en modellanpassning där responsvariabeln är sökvärdena och den förklarande variabeln är pollenräkningarna samt att årsvariabeln är med. Samtliga p-värden är 0.0000.

4.3 Quasi-Poisson-regression

Tabell 5: Resultat av Quasi-Poisson-regression för att undersöka sambandet mellan olika pollenarter och söktrend för ordet "pollen".

Pollen	**SE(year)**	**SE(count)**	**Överspridningskoefficient**
Al	0.0074	6.2380	8.3335
Alm	0.0075	5.9480	8.3424
Björk	0.0073	18.7200	8.0625
Ek	0.0073	8.5557	8.2620
Gråbo	0.0074	20.3900	8.4065
Gräs	0.0074	13.8900	8.2752
Hassel	0.0074	4.0420	8.3843
Sälg och viden	0.0073	7.7349	8.1465

Resultatet för Quasi-Poisson-regression skiljer sig lite åt i förhållande till Poisson-regression. Koefficienterna för year och count är identiska tillsammans med den justerade förklaringsgraden R^2_{adj}, men värdena för standardfelen är högre för Quasi-Poisson och dessutom har en överspridningskoefficient tillkommit. Vidare är samtliga p-värden 0.0000. Dessutom är inte AIC definierat för Quasi-Poisson-regression. Observera att varje rad motsvarar en modellanpassning. Se Tabell 9 i appendix.

4.4 Stegvis regression och minsta kvadratmetoden med icke-negativa begränsningar

Nedan kommer resultaten för de stegvisa regressionsmodellerna, där vi fokuserar på att exkludera de förklarande variabler som ger negativa koefficienter, och minsta kvadratmetoden med icke-negativa begränsningar att presenteras. Även här kommer sökvärdena för "pollen" att vara responsvariabeln medan pollenräkningarna för de olika pollenarterna är de förklarande variablerna.

4.4.1 Multipel linjär regression och minsta kvadratmetoden med icke-negativa begränsningar

Tabell 6: Resultat av stegvis regression med multipel linjär regression och minsta kvadratmetoden med icke-negativa begränsningar.

Variabel	**Koefficient**	**Koefficient(NNLS)**	**p-värde**	**Standardfel**
Year	0.3622	0.0020	0.0000	0.0362
Al	635.7155	610.1971	0.0000	70.3974
Alm	116.2151	72.2098	0.0461	58.2361
Björk	3705.6115	3609.9250	0.0000	258.3398
Ek	912.2217	881.9465	0.0000	86.3226
Gråbo	–	0.0000	–	–
Gräs	1933.5828	1940.5700	0.0000	118.5678
Hassel	259.2658	261.8672	0.0000	55.2054
Sälg och viden	1094.2711	1093.9660	0.0000	106.0548

Tabell 6 redogör resultatet för två modellanpassningar, nämligen stegvis regression där vi har tillämpat multipel linjär regression och minsta kvadratmetoden med icke-negativa begränsningar. Den andra kolumnen Koefficient redovisar resultatet för stegvis regression med multipel linjär regression medan den tredje kolumnen Koefficient(NNLS) redovisar resultatet för minsta kvadratmetoden med icke-negativa begränsningar. Vidare tillhör de två kolumnerna p-värde och Standardfel värdena erhållna från den stegvisa regressionen. För den förstnämnda modellanpassningen framkommer det att gråbo är den enda pollenart som ger negativ koefficient och är således den enda pollenart som tas bort från modellen. För den sistnämnda har vi begränsat modellen så att regressionskoefficienterna inte kan anta negativa värden vilket är varför koefficienten för gråbo är noll. Vidare är $R^2_{adj} = 0.2308$ och AIC = 21 908 för den stegvisa regressionsmodellen. För den minsta kvadratmetoden lyckades vi inte extrahera värden för varken R^2_{adj} eller AIC.

4.4.2 Poisson-regression och Quasi-Poisson-regression

Tabell 7: Resultat av stegvis regression med Poisson-regression respektive Quasi-Poisson-regression.

Variabel	**Koefficient**	**p-värde**	**SE(Poisson)**	**SE(Quasi-Poisson)**
Year	0.0629	0.0000	0.0026	0.0069
Al	55.1195	0.0000	2.2900	6.1470
Alm	22.8151	0.0000	2.4962	6.7010
Björk	249.9254	0.0000	6.9762	18.7300
Ek	85.1566	0.0000	2.9900	8.0260
Gräs	193.6944	0.0000	4.6900	12.5900
Hassel	20.1554	0.0000	1.7140	4.6010
Sälg och viden	84.8702	0.0000	3.4373	9.2270

Precis som Tabell 6 representerar Tabell 7 en modellanpassning, men med resultatet för stegvis regression för både Poisson- och Quasi-Poisson-regression. Utöver standardfelen och värdet på överspridningskoefficienten som är 7.2059 är resultatet för Quasi-Poisson-regressionen identiskt med resultatet för Poisson-regression förutom AIC som inte är definierat för Quasi-Poisson-regression. Både Poisson- och Quasi-Poisson-regressiorna får $R^2_{adj} = 0.1029$ medan Poisson-regressionen får AIC $= 33\,388$. Även här finner vi att gråbo är den enda pollenart som erhåller negativt värde för koefficienten och är därmed den enda pollenart som tas bort från modellen.

4.5 Korrelation

Efter att ha säsongsjusterat datamaterialet och därefter genomfört det utökade Dickey-Fuller testet finner vi att datan är stationär eftersom samtliga variabler erhåller ett p-värde som understiger signifikansnivån på 0.05. Nedan redovisas resultatet för korrelationsanalysen där vi har tillämpat Spearmans rangkorrelation.

Tabell 8: Korrelation mellan olika pollenarter och söktrend för ordet "pollen".

Pollen	**Korrelation**	**p-värde**
Ej gråbo	0.6147	0.0000
Björk	0.5953	0.0000
Al, björk & sälg och viden	0.5750	0.0000
Björk, al, gräs & gråbo	0.5335	0.0000
Al, ek, gräs & gråbo	0.3426	0.0000
Hassel, ek, gräs & gråbo	0.2906	0.0000
Alla	0.2580	0.0000
Hassel	0.2388	0.0000
Ek	0.2203	0.0000
Hassel, sälg och viden & gräs	0.2067	0.0000
Gräs	0.1975	0.0000
Al	0.1930	0.0000
Gråbo	0.0793	0.0969
Sälg och viden	0.0790	0.0591
Alm	−0.0798	0.1675

5 Diskussion

Nedan kommer de olika resultaten från resultatsdelen att diskuteras för att besvara denna studies två syften. Nämligen vilka pollenarter som påverkar söktrenderna mest och vilken modell som ger bäst resultat. För att besvara dessa frågeställningar kommer vi att kolla på koefficienter, p-värden och anpassningsmått som beskriver hur väl modellerna överensstämmer med den ursprungliga datan.

5.1 Regression

5.1.1 Enkel linjär regression

För varje pollenart har det genomförts en enkel linjär regression där vi undersöker hur trenden i sökdata för pollen förändras över åren med hänsyn till pollenräkningar. För samtliga enkla linjära regressioner får vi att alla koefficienter är signifikanta där det högsta p-värdet motsvarar 0.0385. Vi finner att alla pollenarter har en positiv koefficient för år medan alla pollenarter förutom gråbo har positiv koefficient för pollenräkningar. Detta implicerar att söktrenderna ökar med tiden på ett sätt som inte kan förklaras av pollenräkningarna. Framför allt syns det tydligt från koefficienterna att björk och gräs påverkar pollensökningarna mest eftersom värdena på deras koefficienter är höga till skillnad från de andra. Detta resultat överensstämmer väl med D'Amato et Al. (2007) som betonar att särskilt björk och gräs är allergiframkallande.

Generellt sett verkar det finnas ett samband mellan pollensökningar och pollenräkningar, men däremot är förklaringsgraden R^2_{ajd} relativt låg för samtliga modeller då den endast antar värden mellan $R^2_{adj} = 0.0170 = 1.7\%$ och $R^2_{adj} = 0.1000 = 10\%$ vilket tyder på att modellerna endast förklarar en liten del av variansen i pollensökningarna. En anledning till detta resultat är att dag-till-dag variationen i pollenräkningarna är stor, och det är oklart vilken dag som korrelerar mest. Är det idag, igår eller imorgon?

En annan anledning skulle kunna vara att det finns andra faktorer som påverkar men som inte har tagits med i modellen. Sådana faktorer skulle kunna vara geografiska platser eller väderförhållanden. Det skulle även kunna bero på att medvetenheten har ökat kring pollens olika säsonger och även pollenallergi. Exempelvis har man upptäckt att säsongerna för al och hassel börjar tidigare än andra säsonger samt att pollensäsongen generellt startar tidigare vilket blir uppmärksammat. Detta bidrar till att folk förstår att deras symptom beror på pollenallergi och inte en förkylning vilket gör att de söker på internet och vet att de kan få hjälp i form av exempelvis antihistaminer.

5.1.2 Poisson-regression och Quasi-Poisson-regression

Precis som i fallet med de enkla linjära regressionerna får vi att alla koefficienter är signifikanta, men nu med signifikansnivån 0.0001 i alla Poisson-regressioner. Vidare fås det samma resultat att alla pollenarter har en positiv koefficient för år medan alla pollenarter förutom gråbo har positiv koefficient för pollenräkningar. Detta tyder än en gång på att årstrender och pollenhalter har en betydande effekt på pollensökningarna. Vidare har vi att björk och gräs än en gång erhåller högst värden på koefficienterna vilket indikerar att de bidrar mest till pollensökningarna.

5.1.3 Stegvis regression

Stegvis regression genomfördes för de tre regressionerna multipel linjär regression, Poisson-regression och Quasi-Poisson-regression där vi fokuserade på att exkludera de pollenarter som uppvisade negativa koefficienter istället för höga p-värden. Den enda pollenart som hade negativa koefficienter var gråbo och var således den enda pollenart som togs bort från datan i samtliga tre fall.

För samtliga tre regressionsmodeller fann vi att koefficienten för år var positiv och signifikant vilket antyder att det finns en ökande söktrend genom åren. Vidare är även alla koefficienterna för de olika pollenarterna positiva och signifikanta vilket indikerar att de erhåller en signifikant effekt på pollensökningarna.

Eftersom både Poisson-regression och Quasi-Poisson-regression är anpassad vid analys av räkningsdata är det möjligt att de ger ett säkrare resultat i jämförelse med multipel linjär regression, men eftersom alla tre nu visade samma resultat tyder resultaten på att pollensökningarna ökar genom åren och att gråbo är den enda pollenart som inte har en positiv effekt på pollensökningarna. Som i de

tidigare regressionerna har gräs och björk mest påverkan på pollensökningarna eftersom de erhåller högst koefficienter.

5.1.4 Minsta kvadratmetoden med icke-negativa begränsningar

Det framkommer att NNLS i princip ger samma resultat som stegvis regressionen för multipel linjär regression. Koefficienterna skiljer sig åt en aning, men marginellt. Anledningen till att koefficienterna skiljer sig åt skulle möjligtvis kunna bero på att metoden bestraffas om man sätter icke-negativa begränsningar vilket även Mehmud (2020) noterar i ett inlägg där han jämför NNLS med linjär regression. Vidare finner han även att förklaringsgraden blir lägre för NNLS än för linjär regression. Trots att det finns en liten skillnad mellan koefficienterna kan vi dra samma slutsats om att alla pollenarter utöver gråbo har en signifikant effekt på pollensökningarna och att gräs och björk är de två pollenarter som påverkar pollensökningarna mest.

5.1.5 Jämförelse av de olika regressionsmodellerna

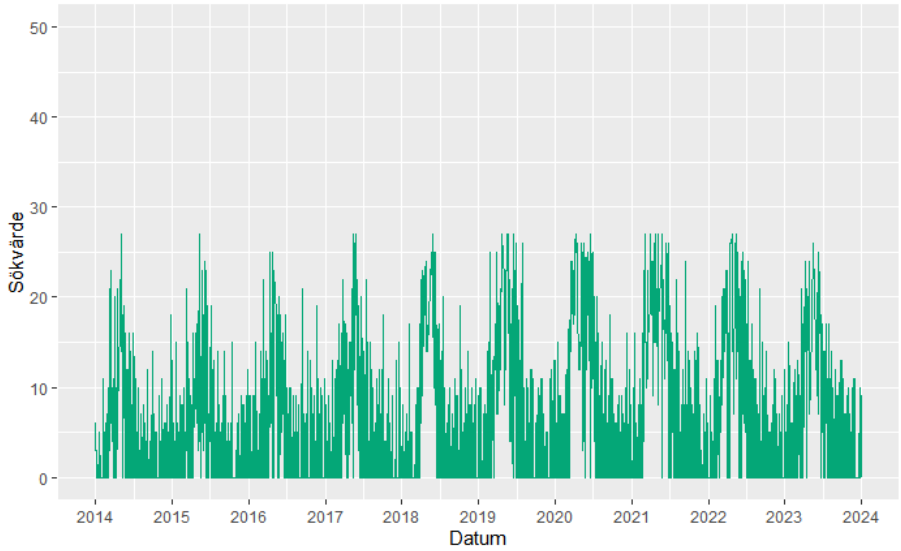
Samtliga regressionsmodeller ger oss samma slutresultat, nämligen att alla pollenarter utöver gråbo har en signifikant och positiv påverkan på pollensökningarna och att björk och gräs är de två pollenarter som ger högst ökning till pollensökningarna.

Gällande vilken eller vilka modeller som anpassar datan bäst kommer vi titta på förklaringsgraden R^2_{adj} och AIC. Ju högre värdet på R^2_{adj} är, där $0 \leq R^2_{adj} \leq 1$, desto bättre är modellen. För AIC gäller att ju mindre värdet är, desto bättre är modellen. AIC har ingen övre eller undre begränsning och kan anta positiva så väl som negativa värden. För samtliga modeller där linjär regression har tillämpats fås högre värden för R^2_{adj} till skillnad från Poisson- och Quasi-Poisson-regression som har lägre värden. Gällande AIC har linjär regression lägre värden än Poisson-regression. Således indikerar bägge måtten på att linjär regression gör bättre ifrån sig än Poisson- och Quasi-Poisson-regression. Det framkommer att den modell som erhåller den högsta förklaringsgraden på $R^2_{adj} = 0.2308$ är stegvis regression där vi har tillämpat multipel linjär regression. Dessvärre har jag inte lyckats ta fram varken R^2_{adj} eller AIC för NNLS, men eftersom NNLS approximerar linjär regression bör förklaringsgraden likna den för stegvis regression med multipel linjär regression.

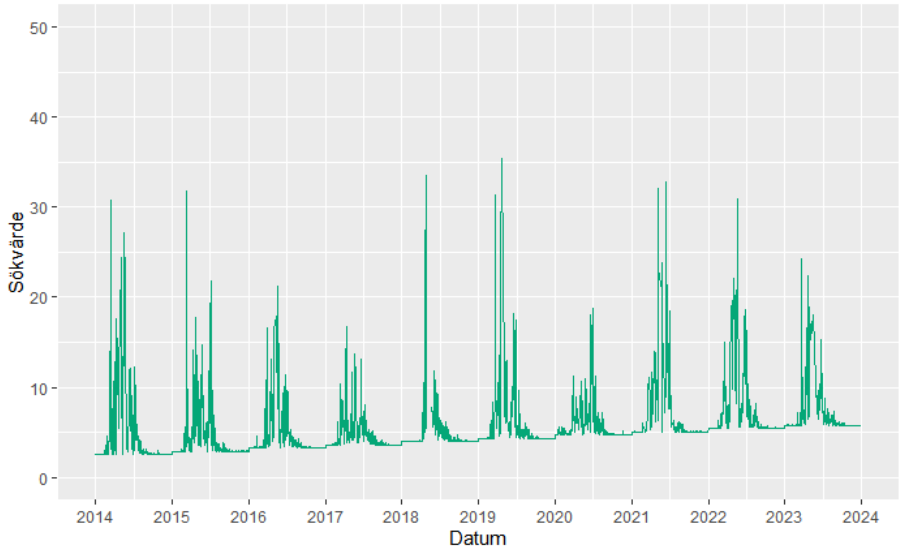
Detta resultat är intressant med tanke på att datamaterialen för pollenräkningar och pollensökningar har samlats in på olika sätt. Pollenräkningarna är räknedata vilka ofta följer en Poissonfördelning där Poisson-regression tenderar att passa datan bättre än linjär regression som antar att räknedatan följer en normalfördelning. Däremot antar inte Poisson-regression att datan måste vara Poisson-fördelad. Pollensökningarna är däremot redan normaliserade vilket indikerar på att linjär regression borde passa datan bättre. Om pollensökningarna hade varit räknedata istället för normaliserat skulle det vara tänkbart att Poisson-regression hade varit bättre lämpad. Något annat som

är intressant är att regressionerna för Poisson respektive Quasi-Poisson ger identiska resultat. Om jag skulle väntat mig någon skillnad mellan dem två så skulle p-värdena skilja sig åt vilket de inte gör. Detta skulle dock kunna bero på att datamaterialet består av många observationer vilket kan bidra till låga p-värden.

Nedan visas en jämförelse mellan de ursprungliga värdena för söktrenderna för ”pollen” med de predikterade värdena från stegvis regression där multipel linjär regression har tillämpats.



(a)



(b)

Figur 3: (a) Ursprungliga datan för de dagliga sökvärdena för ”pollen”. (b) De predikterade värdena för de dagliga sökvärdena för ”pollen” från stegvis regression med multipel linjär regression.

Trots den relativt låga justerade förklaringsgraden på R^2_{adj} kan vi från Figur 3 observera att de predikterade värdena överensstämmer godtyckligt med de ursprungliga värdena. Det finns lite skillnader som att vissa av topparna inte korresponderar, men i det stora hela anser jag att modellen predikterar förhållandevis väl med tanke på förklaringsgraden.

Sammanfattningsvis verkar stegvis regression med multipel linjär regression vara den bästa modellen för det givna datamaterialet. Det hade varit intressant att se om resultatet hade skiljt sig om både responsvariabeln och de förklarande variablerna rörde sig om räknedata.

5.2 Korrelation

Det framkommer att korrelation mellan samtliga pollenarter och sökvärdena för "pollen" på Google Trends varierar mellan $\rho = -0.0798$ och $\rho = 0.5953$. Den pollenart som har störst samband med pollensökningarna är björk där $\rho = 0.5953$ med den höga signifikansnivån på 0.0001. Detta är å andra sidan ett relativt starkt samband eftersom björk inte är den enda pollenart som bidrar till pollensökningarna i och med att björksäsongen delvis sammanfaller med säsongerna för både ek och sälg och viden vilket vi kan se i Figur 1. Den höga korrelationen indikerar att människor är mer benägna att söka på "pollen" när pollenhalterna av björken är höga än när de är låga. Det skulle kunna bero på att de som söker lider av allergiska reaktioner och därmed vill veta om pollensäsongen är aktiv. Vidare kan det även vara för att björkpollen är dominerande i Stockholm eller att den är särskilt allergiframkallande. En anledning till att korrelationen inte är högre skulle kunna bero på att pollensökningarna genomfördes någon eller några dagar efter björkpollentopparna vilket medför att pollenhalterna av björk är lägre i förhållande till sökvärdet. Om pollensökningarna faktiskt hade genomförts samma dag hade korrelationen mellan björkpollen och pollensökningarna troligtvis varit högre.

Vidare har vi att alm, gråbo och sälg och viden har svaga korrelationer på $\rho = -0.0798$, $\rho = 0.0793$ respektive $\rho = 0.0790$. Ingen av dessa samband är däremot signifikanta vilket antyder att det inte finns något tydligt samband mellan dessa tre pollenarter och sökbeteendet. Anledningen till detta resultat skulle kunna bero på att dessa pollenarter inte är lika allergiframkallande som exempelvis björk vilket i sin tur medför att färre människor kommer söka på pollen. En annan anledning skulle kunna vara att dessa pollenarter inte är lika dominanta som björkpollen är vilket implicerar att det är låg risk att utsättas och därmed liten risk för allergisymptom.

Vidare har hassel, al, gräs och ek alla ett måttligt men positivt och signifikant samband med pollensökningar med korrelationerna $\rho = 0.2388$, $\rho = 0.1930$, $\rho = 0.1975$ och $\rho = 0.2203$, i respektive ordning. Samtliga har dessutom den höga signifikansnivån 0.0001. Detta påvisar att människor har en tendens till att söka på "pollen" när respektive halt av de fyra olika pollentyperna är hög eftersom dessa pollenarter kan vara allergiframkallande för många människor

vilket i sin tur driver dem till att söka efter information om pollen och mest troligen även tillhörande symptom. I korrelationsanalysen framkommer det att det finns många relativt små korrelationer som ändå är väldigt signifikanta. En anledning till att många av p-värdena är så låga skulle kunna bero på att det finns många datapunkter. Med många datapunkter fås det låga p-värden även om korrelationen är väldigt svag.

Utöver de enskilda pollenarterna valde vi även att slå samman olika pollenarter för att se om sambandet mellan dem och pollensökningarna ökar. Det framkommer att de olika sammanslagningarna ger måttliga och höga samband mellan $\rho = 0.2067$ till $\rho = 0.6147$. Den sammanslagning som stegvis regression angav som bästa modell var den modell som inkluderade alla pollenarter utöver gråbo. Denna modell gav korrelationen $\rho = 0.6147$ vilken även är den modell med högst korrelation. Detta tyder på att när flera olika pollenarter har hög halt samtidigt ökar också sökaktiviteten för "pollen" vilket, som skrivet tidigare, kan bero på att de frammanar allergisymptom.

Det framkommer att Holmer (u.å.) erhöll liknande resultat för sin korrelationsanalys där han rapporterar att björk var den pollenart som korrelerade mest med pollensökningarna med $\rho = 0.79$ samt att det fanns ett samband för både al och ek med pollensökningarna. Även Sitaru et al. (2019) fann ett tydligt samband mellan björkpollen och pollensökningar med $\rho = 0.63$ vilket indikerar att björk är den pollenart som påverkar pollensökningar mest. Skillnaden mellan våra undersökningar är att i denna analys säsongsjusteras datan för att uppnå stationäritet medan både Holmer (u.å.) och Sitaru et al. (2019) använder regressionsanalysen på den ursprungliga datan. Det finns alltså risk att deras data inte är stationär vilket i sin tur kan medföra att deras resultat endast är giltig för de år de undersöker. Deras slutresultat är därmed inte giltiga för andra tidsperioder och det finns även risk för att deras slutsatser är felaktiga. Detta kan vara anledningen till att denna studie även finner samband mellan hassel och gräs och inte endast al, björk och ek. Det bör dock betonas att Holmer inte undersökte sambanden för pollenarterna hassel, alm eller sälg och viden medan Sitaru et al. genomförde korrelationsanalysen för de totala pollenräkningarna och björkpollen.

Sammanfattningsvis påvisar resultaten att sambanden mellan de olika pollenarterna och söktrenden för "pollen" varierar. Björkpollen är den art som verkar vara mest betydelsefull för söktrenden medan andra arter visar svagare eller inget samband alls. Än en gång överensstämmer detta resultat väl med D'Amato et al. (2007) som påpekar att särskilt björk och gräs är allergiframkallande följt av hassel och al. Vidare överensstämmer detta även väl med resultatet från minsta kvadratmetoden med icke-negativa begränsningar där multipel regression har tillämpats som också anger att björk verkar vara den mest betydelsefulla pollenarten för söktrenden.

Genom att fortsätta undersöka dessa korrelationer och eventuellt ta med andra relevanta variabler kan man erhålla en djupare förståelse för exempelvis människors reaktioner på olika pollensäsonger och deras sökbeteende. Detta

kan i sin tur vara användbart för att prediktera allergiska reaktioner genom att utveckla varningssystem för de som är allergiska och även informera om olika åtgärder för att den som är allergisk ska få så liten exponering som möjligt för pollen.

5.3 Framtida forskning och förbättringar

För framtida forskning skulle det vara intressant att inkludera ytterligare förklarande variabler i modellerna för att på så sätt kunna förstå sambandet mellan sökdata för ordet ”pollen” och pollenräkningarna på ett mer heltäckande sätt och eventuellt kunna förbättra förklaringsgraden i modellerna. Sådana förklarande variabler skulle kunna vara olika geografiska platser och väderförhållanden för att på sätt kunna jämföra om resultaten skiljer sig åt mellan olika platser och om olika väderförhållanden möjligen påverkar. Även andra söktermer så som ”snuva” och ”antihistamin” skulle vara av intresse att ta med i analysen. Vidare hade det även varit av intresse att genomföra analysen och jämföra mellan olika år. Skiljer det sig exempelvis åt mellan 2014 och 2023, och i sådana fall med hur mycket?

Appendix

Tabell 9: Resultat av Quasi-Poisson-regression för att undersöka sambandet mellan olika pollenarter och söktrend för ordet "pollen". Varje rad motsvarar en modellanpassning. Samtliga p-värden är 0.0000.

Pollen	**year**	**count**	R^2_{adj}	**AIC**
Al	0.0586	55.4200	0.0263	NA
Alm	0.0601	43.9100	0.0234	NA
Björk	0.0582	266.4000	0.0220	NA
Ek	0.0573	87.5000	0.0314	NA
Gråbo	0.0565	−42.5000	0.0162	NA
Gräs	0.0541	174.6000	0.0462	NA
Hassel	0.0571	26.2900	0.0170	NA
Sälg och viden	0.0572	101.8000	0.0333	NA

Referenser

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). John Wiley & Sons.

D'Amato, G., Cecchi, L., Bonini, S., Nunes, C., Annesi-Maesano, I., Behrendt, H., Liccardi, G., Popov, T. & van Cauwenberge, P. (2007). Allergenic pollen and pollen allergy in Europe. *Allergy*, 62(9), 976-990.

Dickey, D.A. & Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*. 74(366), 427-431.

Dodge, Y. (2008). *The Concise Encyclopledia of Statistics*. Springer.

Google (u.å.). *Vanliga frågor om data i Google Trends*. Google. `https://support.google.com/trends/answer/4365533?hl=sv` [2024-05-23].

Holmer, G.S. (u.å.). *Pollenhalters relation till webbsökningar*. Smittskyddsinstitutet.

Held, L. & Bové, D.S. (2014). *Likelihood and Bayesian Inference: With Applications in Biology and Medicine* (2nd ed.). Springer.

Hirst, J.M. (1952). An automatic volumetric spore trap. *Annals of Applied Biology*, 39(2), 257-265.

Lawson, C.L. & Hanson R.J. (1995). *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics.

Luther, M.-L. (2024). *Pollenallergi*. Astma och Allergiförbundet. `https://astmaoallergiforbundet.se/information-rad/allergi/pollenallergi/` [2024-05-01].

Mazzi, G.L. (2018). *Handbook on Seasonal Adjustment*. Publications Office of the EU.

Mehmud, S. (2020). *Non-Negative Least Squares Regression*. Predictive Modeler. `https://predictivemodeler.com/2020/05/09/non-negative-least-squares-regression/` [2024-05-01].

Montgomery, D. C., & Runger, G. C. (2010). *Applied statistics and probability for engineers* (5th ed.). John Wiley & Sons.

Pollenrapporten (2015). Allergiframkallande pollen. `https://pollenrapporten.se/ompollen/allergiframkallandepollen.4.314e02dd13d69872ec083.html` [2024-05-01].

Pollenrapporten (2017). Så gör vi prognoserna. `https://pollenrapporten.se/ompollen/sagorviprognoserna.4.314e02dd13d69872ec097.html` [2024-05-01].

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Roback, P. & Legler, J. (2021). *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. Chapman and Hall/CRC.

Sitaru, S., Tizek, L., Buters, J., Ekebom, A., Wallin, J.-E. & Zink, A. (2023). Assessing the national burden of allergic asthma by web-search data, pollen counts, and drug prescriptions in Germany and Sweden. *World Allergy Organization Journal*, 16(2), 100752.

Soetewey, A. (2020). *Outliers detection in R*. Stats and R. `https://statsandr.com/blog/outliers-detection-in-r/` [2024-05-23].

Sundberg, R. (2023). *Lineära statistika modeller*. Kompendium Stockholms universitet.

Winkelmann, R. (2008). *Econometric Analysis of Count Data* (5th ed.). Springer.