



Stockholms
universitet

Uppskattning av en automatisk pollen- fällas felklassificeringssannolikhet

Stefano Aziz

Kandidatuppsats 2025:5
Matematisk statistik
Juni 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Uppskattning av en automatisk pollenfällas felklassificeringssannolikhet

Stefano Aziz*

Juni 2025

Sammanfattning

I denna kandidatuppsats utvärderas noggrannheten hos en automatisk pollenfälla, *Poleno Jupiter*, genom att skatta sannolikheten för felklassificering av pollen Korn från fyra olika arter: Al (*Alnus*), Björk (*Betula*), Hassel (*Corylus*) och Ek (*Quercus*). Studien använder en statistisk modell baserad på multinomialfördelning i kombination med en föräxlingsmatris, vilken skattas med maximum likelihood-metoden. Genom att jämföra automatiskt klassificerade pollenproportioner med manuella referensmätningar, analyseras modellens passform och identifierbarhet. Resultaten visar att vissa arter, som *Quercus* och *Alnus*, kan klassificeras med hög precision, medan andra, särskilt *Corylus*, uppvisar betydande problem med felklassificering. Särskild vikt läggs vid hur temporal aggregering, månadsvis kontra veckovis, påverkar modellens prestanda. Slutsatsen är att finmaskig tidsupplösning är avgörande för att säkerställa tillförlitliga skattningar av felklassificeringssannolikheter, särskilt för arter med smala eller överlappande pollensäsonger.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: staz3712@student.su.se. Handledare: Martin Sköld.

Förord

Denna rapport har skrivits som en del av mitt kandidatarbete vid Stockholms Universitet. Arbetet har gett mig möjlighet att fördjupa mina kunskaper inom statistisk modellering och tillämpad dataanalys, med fokus på klassificering och validering av automatiska system.

Jag vill rikta ett stort och uppriktigt tack till min handledare Martin Sköld för hans värdefulla vägledning, insiktsfulla återkopplingar och kontinuerliga stöd under projektets gång. Jag vill även tacka mina klasskamrater Sebastian och Alex för givande diskussioner, praktisk hjälp och god stämning under hela arbetets gång.

Jag är också tacksam för det stöd jag fått från min familj, vars tålamod, uppmuntran och förståelse varit ovärderliga under arbetets gång.

Jag har även använt AI-verktyget ChatGPT som stöd för att förbättra språkstruktur, grammatik och tydlighet i rapportens text. Verktyget har även varit till hjälp vid felsökning av R-kod under projektets gång.

Innehåll

1	Introduktion	4
2	Data	5
2.1	Datainsamling och ursprung	5
2.2	Filtrering av data	5
2.3	Visualisering av data	6
3	Teori	7
3.1	Inledning till teoridelen	7
3.2	Multinomialfördelning i Klassificeringssammanhang	7
3.3	Förväxlingsmatris π	8
3.4	Lagen om Total sannolikhet	8
3.5	Log-likelihood och Parameterestimering	9
3.6	Numerisk optimering	9
3.7	Hessianen och osäkerhetsuppskattning	10
4	Resultat	11
4.1	Månadsvisa analyser	11
4.1.1	Uppskattning av förväxlingsmatrisen för månadsvisa räkningar	11
4.1.2	Resultat av Förväxlingsmatrisen	11
4.2	Veckovisa analyser	12
4.2.1	Uppskattning av felklassificeringsmatrisen för veckovisa räkningar	12
4.2.2	Resultat av Förväxlingsmatrisen	12
4.3	Osäkerhetsuppskattning och Modellvalidering	13
4.3.1	Jämförelse av predikterade och observerade proportioner	13
4.3.2	Osäkerhetsuppskattning via Hessianen	14
5	Diskussion och slutsatser	15
5.1	Tolkning av felklassificeringsmatris – månadsvis	15
5.1.1	Artberoende prestanda och svårigheter med månadsvis analys	15
5.1.2	Begränsningar med månadsvis aggregering	15
5.1.3	Slutsats för månadsvis analys	15
5.2	Tolkning av felklassificeringsmatris – veckovis	15
5.2.1	Högre diagonalelement i den veckovisa modellen	15
5.2.2	Tolkning och säsongsvariation	16
5.2.3	Praktiska begränsningar	16
5.3	Sammanfattning av artvis prestanda med fokus på <i>Corylus</i>	16
5.4	Reflektion kring antagandet om multinomialfördelning	17
5.5	Modellvalidering och osäkerhetsuppskattning	17
5.5.1	Jämförelse mellan q_{pred} och q_{auto}	17
5.5.2	Osäkerhetsuppskattning via Hessianen	18
5.5.3	Tolkning av konfidensintervall – månadsvis modell	18
5.5.4	Tolkning av konfidensintervall - veckovis modell	19
5.6	Praktisk tillämpning av maximum likelihood i R	19
5.7	Slutsats	19
5.8	Förslag till framtida förbättringar	19
6	Appendix	21
7	Referens lista	24

Sammanfattning

I denna kandidatuppsats utvärderas noggrannheten hos en automatisk pollenfälla, *Poleno Jupiter*, genom att skatta sannolikheten för felklassificering av pollen Korn från fyra olika arter: Al (Alnus), Björk (Betula), Hassel (Corylus) och Ek (Quercus). Studien använder en statistisk modell baserad på multinomialfördelning i kombination med en föräxlingsmatris, vilken skattas med maximum likelihood-metoden. Genom att jämföra automatiskt klassificerade pollenproportioner med manuella referensmätningar, analyseras modellens passform och identifierbarhet. Resultaten visar att vissa arter, som *Quercus* och *Alnus*, kan klassificeras med hög precision, medan andra, särskilt *Corylus*, uppvisar betydande problem med felklassificering. Särskild vikt läggs vid hur temporal aggregering, månadsvis kontra veckovis, påverkar modellens prestanda. Slutsatsen är att finmaskig tidsupplösning är avgörande för att säkerställa tillförlitliga skattningar av felklassificeringssannolikheter, särskilt för arter med smala eller överlappande pollensäsonger.

1 Introduktion

Under senare år har automatiserade metoder för miljöövervakning blivit allt viktigare, inte minst inom områden som luftkvalitet och pollenmätning. En ny teknik är att använda automatiska fällor baserade på bildanalys, en av dem är *Swisens Poleno Jupiter* Swisens AG (2025a), som använder bildanalys för att identifiera och klassificera pollenkorn i realtid. Genom att fotografera luftburna partiklar kan fällan snabbt avgöra vilka pollenarter som förekommer i luften. Tekniken är särskilt lovande eftersom den möjliggör kontinuerlig och snabb datainsamling utan behov av manuellt arbete.

Automatiserade pollenfällor har under de senaste åren utvecklats för att möta bristerna i traditionella manuella metoder, som ofta är arbetsintensiva och har flera dagars fördröjning. En översiktsartikel av Buters et al. (2024) visar att automatiserad övervakning inte bara kan förbättra den tidsmässiga upplösningen, utan även minska mänskliga felkällor och öppna nya möjligheter inom såväl hälsa som klimatforskning.

Poleno Jupiter har testats i Stockholm under pollensäsongen 2024, och systemet har genererat omfattande data över pollenförekomsten under den perioden. Samtidigt har manuella mätningar genomförts parallellt, där experter klassificerat pollenarter under mikroskop, en metod som är långsam men betraktas som tillförlitlig. Dessa manuella mätningar utgör därmed en referenspunkt för att bedöma noggrannheten i den automatiska klassificeringen.

Ett viktigt förhållande som möjliggör denna typ av studie är att olika pollenarter inte förekommer samtidigt, utan tenderar att spridas under olika perioder av året. Denna säsongsvariation är avgörande för att statistiskt kunna särskilja arter och korrekt skatta klassificeringssannolikheter. Vi kommer att återkomma till denna aspekt senare, dels i teorikapitlet 3.1 och mer i diskussionen kring modellens identifierbarhet 5.1.2, 5.2.3.

Projektets syfte är att undersöka klassificeringsprestandan hos den automatiska pollenfällan *Poleno Jupiter*, det vill säga hur väl maskinen kan skilja mellan olika pollenarter jämfört med manuella räkningar. Ett centralt begrepp i denna analys är *felklassificering*, risken att ett pollenkorn från en viss art tolkas som tillhörande en annan. För att kvantifiera detta introduceras en *förväxlingsmatris*, där varje element beskriver sannolikheten att en viss art klassificeras som en annan.

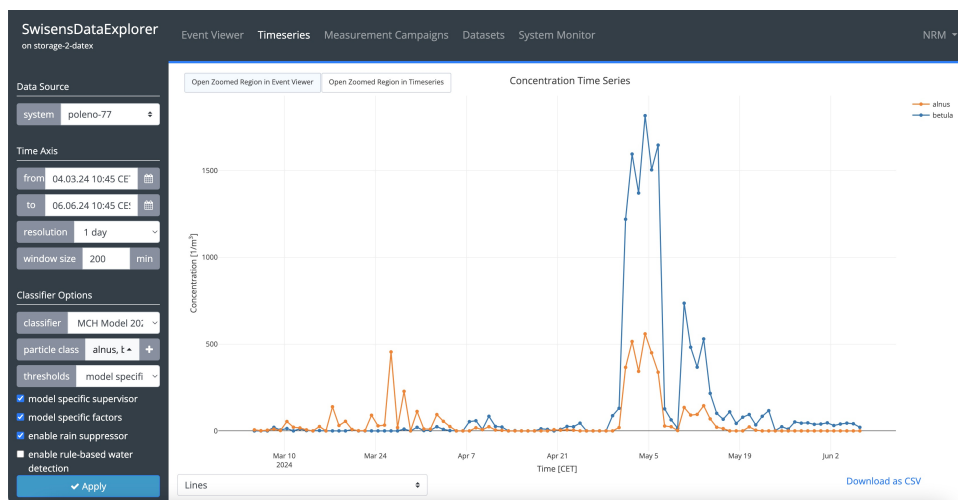
Mer konkret kommer arbetet att:

- Analysera skillnader mellan de proportioner som rapporteras av den automatiska respektive den manuella metoden.
- Uppskatta en förväxlingsmatris med hjälp av statistisk modellering.
- Genomföra numerisk optimering för att maximera sannolikheten att modellen stämmer med observerade data.
- Bedöma osäkerheten i skattningarna med hjälp av den observerade Fisherinformationen, som approximeras med Hessianen från den numeriska optimeringen.

Slutligen kommer resultaten att diskuteras i relation till praktiska användningsområden, till exempel hur stor tillförlitlighet man kan förvänta sig från automatiska pollenfällor i forsknings- eller hälsoövervakningssyfte.

2 Data

2.1 Datainsamling och ursprung



Figur 1: Fällans användargränssnitt (SwisensDataExplorer)

I denna studie klassificeras pollen från fyra arter: Al (*Alnus*), Björk (*Betula*), Hassel (*Corylus*) och Ek (*Quercus*). Data som ligger till grund för analysen kommer från två olika källor, varav den ena representerar de automatiska pollenmätningarna och den andra de manuella räkningarna. De automatiska mätningarna har hämtats via den automatiska fällans användargränssnitt Swisens AG (2025b). Detta system registrerar varje enskilt händelseförlopp där pollen Korn detekteras av instrumentet, och de olika händelserna dokumenteras med information om tidpunkt, plats samt klassificerad pollentyp. Genom att utvinna data ur detta gränssnitt får vi en omfattande tidsseriedata över de automatiskt uppmätta pollenräkningarna, vilket möjliggör en detaljerad analys av systemets prestanda över tid.

Den manuella datan, som utgör en referens för de ”sanna” pollenfördelningarna har samlats in genom traditionell räknemetodik, där forskare manuellt undersökt och räknat pollen Korn under mikroskop, se Naturhistoriska riksmuseet (2024). Detta ger en grundlig och noggrann uppskattning av de proportioner som finns i verkligheten, vilket används för att validera och justera den automatiska mätningens resultat.

Genom att jämföra den automatiska datan med de manuellt insamlade referensvärdena kan modellen utvärderas, och eventuella systematiska fel i den automatiska klassificeringen identifieras och kvantifieras. Genom att kombinera regelbundna automatiska mätningar med den exakta information som traditionella metoder ger, kan vi på ett stabilt sätt följa upp systemets prestanda över tid.

2.2 Filtrering av data

För att säkerställa att endast relevanta observationer ingår i de vidare analyserna genomfördes en systematisk förbehandling av rådata. För de manuella mätningarna extraherades enbart de rader som tillhörde arterna *Alnus*, *Betula*, *Corylus* och *Quercus* och som dessutom var tagna på stationen i Stockholm. Observationerna begränsades till tidsperioden 2024-03-01 till 2024-06-31.

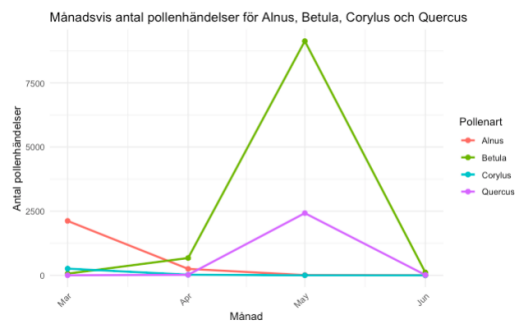
På liknande sätt importerades och kombinerades de automatiska mätningarna från flera filer där varje fil motsvarade en specifik art. Beroende på om en månadsvis eller veckovis period analyserades, skapades en tidsvariabel med respektive tidsenhet, exempelvis genom att aggregera data per månad eller vecka. Genom att gruppera data per tidsenhet och art summerades de

räknade händelserna, vilket möjliggjorde att både de manuella och automatiska mätningarna lätt kunde jämföras över tid.

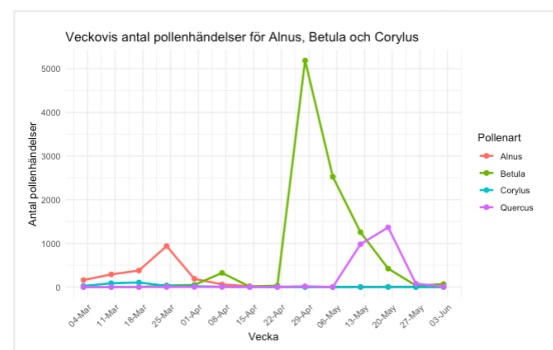
För att minska risken för felklassificering filtrerades de automatiska observationerna med en sannolikhetsgräns på 0.99, vilket innebar att endast observationer där modellen var minst 99% säker på arttillhörighet inkluderades i analysen. Denna tröskel är inte hårdkodad i systemet utan kan justeras manuellt via fällans användargränssnitt Swisens AG (2025b), där användaren själv väljer önskad säkerhetsnivå innan export (se Figur 1). I denna studie har tröskeln 99% valts för att uppnå en balans mellan datakvalitet och datamängd.

Efter att data aggregerats och pivoterats till brett format erhöles fyra tabeller med proportioner av respektive pollenart, både för manuella och automatiska mätningar. Tabellerna täcker både månadsvisa och veckovisa sammanställningar: Tabell 5 och 6 visar de månadsvisa proportionerna, medan Tabell 7 och 8 visar motsvarande veckovisa data.

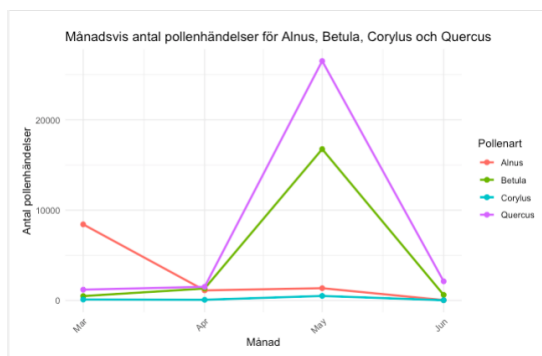
2.3 Visualisering av data



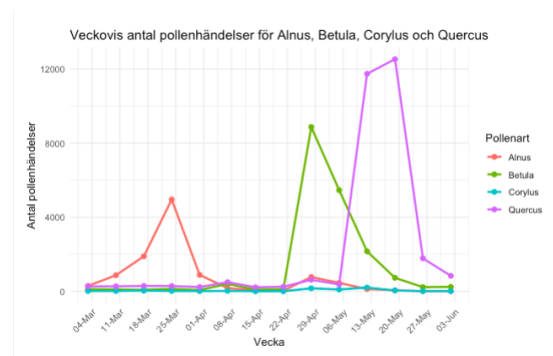
Figur 2: Månadsvisa manuella räkningar (Mars-Juni 2024)



Figur 3: Veckovisa manuella räkningar (Mars-Juni 2024)



Figur 4: Månadsvisa automatiska räkningar (Mars-Juni 2024)



Figur 5: Veckovisa automatiska räkningar (Mars-Juni 2024)

3 Teori

3.1 Inledning till teoridelen

För att kunna analysera felklassificeringar i en automatisk pollenfälla krävs en teoretisk ram som beskriver sambandet mellan observerade och sanna pollenproportioner. Denna ram bygger på sannolikhetsfördelningar, särskilt multinomialfördelningen, samt en statistisk modell för att uppskatta sannolikheten att en pollenart felklassificeras som en annan (Hastie et al., 2009, s.120).

Målet är att skapa en modell som, givet de observerade automatiska klassificeringarna och de manuella mätningarna, kan uppskatta klassificeringsbeteendet hos pollenfällan. En viktig komponent är att uppskatta sannolikheten att ett pollenkorn klassificeras rätt eller som en annan art. Detta kräver en modell som tar hänsyn till osäkerhet i klassificeringen, vilket gör att vi använder en multinomialfördelning i kombination med en felklassificeringsmatris.

Avsnittet inleds med en översikt av de centrala begreppen:

- **Multinomialfördelning**, för att modellera hur klassificerade händelser fördelar sig över flera kategorier (pollenarter).
- **Förväxlingsmatris** π , som beskriver sannolikheten att en pollenart j klassificeras som art k .
- **Lagen om Total sannolikhet**, används för att uttrycka de observerade automatiska proportionerna som en viktad summa av manuella proportioner och felklassificeringssannolikheter.
- **Log-likelihood-funktionen**, som används för att skatta parametervärdena i π .
- **Hessianen**, används för att uppskatta varians-kovariansmatrisen och därigenom beräkna standardfel och konfidensintervall för parametervärdena.

Dessa teoretiska element ligger till grund för formuleringen av den statistiska modellen och utgör basen för den numeriska optimering som genomförs i programvaran R Core Team (2024).

En viktig egenskap i det tillgängliga datamaterialet är att de olika pollenarterna sprider sig under olika delar av våren och sommaren. Enligt Pollenrapporten (2024) har *Hassel* (Corylus) sin huvudsakliga säsong mycket tidigt på året, ofta redan i februari, följt av *Al* (Alnus), *Björk* (Betula) och slutligen *Ek* (Quercus), som blommar senare under våren.

Denna tidsmässiga separation mellan arternas säsonger är avgörande för modellens identifierbarhet. Om alla arter förekom samtidigt i varje period, skulle det vara mycket svårt, eller till och med omöjligt att avgöra vilka klassificeringsmönster som hör till vilken art. Då skulle parametrarna i felklassificeringsmatrisen π inte kunna identifieras entydigt.

I denna studie utnyttjas det faktum att arterna har olika säsonger: modellen drar nytta av att arterna dominerar olika veckor eller månader för att skilja mellan dem. På så vis hjälper säsongsvariationerna till att särskilja arterna statistiskt, vilket är centralt för att kunna skatta felklassificeringssannolikheter på ett meningsfullt sätt.

3.2 Multinomialfördelning i Klassificeringssammanhang

I denna studie antar vi att varje pollenobservation är oberoende, vilket innebär att observationerna från samma period kan modelleras med en multinomialfördelning. Antag nu att vi observerar K arter av pollen under N perioder, i vårt fall veckor eller månader. De manuella mätningarna betecknas med y_{ij} , där $i = 1, \dots, N$ och $j = 1, \dots, K$, vilket representerar antalet pollen av art j som observerats manuellt under period i . Vi definierar vektorn:

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iK}) \sim \text{Multinomial}(n_i, \mathbf{p}_i),$$

där $n_i = \sum_{j=1}^K y_{ij}$ är det totala antalet manuellt räknade pollen under period i , och $\mathbf{p}_i = (p_{i1}, \dots, p_{iK})$ är den sanna proportionen av varje art under period i . Dessa proportioner uppskattas via de relativa frekvenserna:

$$p_{ij} = \frac{y_{ij}}{n_i}.$$

På motsvarande sätt definieras de automatiska mätningarna som:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iK}) \sim \text{Multinomial}(m_i, \mathbf{q}_i),$$

där x_{ij} är antalet pollen som automatiskt klassificeras som art j under period i , $m_i = \sum_{j=1}^K x_{ij}$ är totala antalet automatiskt räknade pollen, och $\mathbf{q}_i = (q_{i1}, \dots, q_{iK})$ är sannolikheterna att ett pollen klassificeras som varje art.

Sannolikhetsfunktionen för den multinomiala fördelningen ges av:

$$P(\mathbf{x}_i) = \frac{m_i!}{\prod_{k=1}^K x_{i,k}!} \prod_{k=1}^K q_{i,k}^{x_{i,k}}, \quad (1)$$

där $\sum_{k=1}^K q_{i,k} = 1$ för varje i (Held and Bové, 2020, s.365).

Denna fördelning utgör grunden för att formulera den log-likelihoodfunktion som används vid parameterestimering av felklassificeringsmatrisen $\boldsymbol{\pi}$.

3.3 Förväxlingsmatris $\boldsymbol{\pi}$

En central idé i analysen är att den automatiska pollenfällan ibland felklassificerar pollenkorn, det vill säga att den angivna arten inte alltid stämmer med den verkliga. För att modellera detta används en felklassificeringsmatris $\boldsymbol{\pi}$, där varje element $\pi_{j,k}$ anger sannolikheten att ett pollenkorn från art j klassificeras som art k :

$$\pi_{j,k} = P(\text{klassificeras som } k \mid \text{är av art } j).$$

Med fyra arter kan $\boldsymbol{\pi}$ -matrisen beskrivas som en 4×4 -matris:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_{1,1} & \pi_{1,2} & \pi_{1,3} & \pi_{1,4} \\ \pi_{2,1} & \pi_{2,2} & \pi_{2,3} & \pi_{2,4} \\ \pi_{3,1} & \pi_{3,2} & \pi_{3,3} & \pi_{3,4} \\ \pi_{4,1} & \pi_{4,2} & \pi_{4,3} & \pi_{4,4} \end{pmatrix},$$

där varje rad summerar till 1:

$$\sum_{k=1}^4 \pi_{j,k} = 1, \quad j = 1, 2, 3, 4.$$

3.4 Lagen om Total sannolikhet

Utöver felklassificeringsmatrisen $\boldsymbol{\pi}$ behöver vi ett uttryck för de "sanna" proportionerna av varje art vid ett visst tillfälle. Vi antar att de manuella mätningarna ger uppskattningar av dessa proportioner $p_{i,j}$, det vill säga sannolikheten att ett slumpmässigt valt pollen under period i tillhör art j .

Eftersom den automatiska fällan kan felklassificera, är sannolikheten att ett pollen klassificeras som art k lika med den totala sannolikheten att det kommer från någon art j och sedan klassificeras som k . Vi använder alltså lagen om total sannolikhet (Held and Bové, 2014, s.319):

$$q_{i,k} = \sum_{j=1}^K P(\text{klassificeras som } k \mid \text{är av art } j) \cdot P(\text{är av art } j) = \sum_{j=1}^K \pi_{j,k} \cdot p_{i,j}. \quad (2)$$

Här tolkas $\pi_{j,k}$ som den betingade sannolikheten att ett pollen av art j klassificeras som art k , och $p_{i,j}$ som sannolikheten att ett slumpmässigt valt pollen under period i faktiskt är av art j .

Denna relation visar att de observerade proportionerna $\mathbf{q}_i$ är en viktad blandning av de sanna proportionerna $\mathbf{p}_i$, där vikterna ges av felklassificeringssannolikheterna i $\boldsymbol{\pi}$.

3.5 Log-likelihood och Parameterestimering

För att skatta π -matrisen utgår vi från log-likelihoodfunktionen för vår multinomialmodell. Om vi betraktar en samling av N perioder (veckor eller månader), kan vi skriva log-likelihood för en enskild period i som:

$$\log P(x_i) = \log(m_i!) - \sum_{k=1}^K \log(x_{i,k}!) + \sum_{k=1}^K x_{i,k} \log(q_{i,k}). \quad (3)$$

Summerar vi detta över samtliga perioder $i = 1, \dots, N$ får vi den totala log-likelihooden:

$$\ell(\boldsymbol{\pi}) = \sum_{i=1}^N \log P(x_i) = \sum_{i=1}^N \left[\log(m_i!) - \sum_{k=1}^K \log(x_{i,k}!) + \sum_{k=1}^K x_{i,k} \log\left(\sum_{j=1}^K p_{i,j} \pi_{j,k}\right) \right].$$

Observera att $\log(m_i!)$ och $\log(x_{i,k}!)$ är konstanter med avseende på $\boldsymbol{\pi}$. Därför är det centrala ledet i optimeringen:

$$\ell(\boldsymbol{\pi}) \propto \sum_{i=1}^N \sum_{k=1}^K x_{i,k} \log\left(\sum_{j=1}^K p_{i,j} \pi_{j,k}\right). \quad (4)$$

En optimeringsstrategi baserad på att maximera denna funktion är central för att skatta parametrarna i modellen, (Held and Bové, 2020, s.15).

3.6 Numerisk optimering

Eftersom uttrycket

$$\sum_{k=1}^K x_{i,k} \log\left(\sum_{j=1}^K p_{i,j} \pi_{j,k}\right)$$

i ekvation (4) inte har en sluten form för $\boldsymbol{\pi}$ (dvs. man kan inte enkelt lösa analytiskt), används numerisk optimering. I R-koden använder vi funktionen `optim` med metod "L-BFGS-B" för att minimera negativ log-likelihood (Held and Bové, 2020, s.32).

De diagonala elementen i $\boldsymbol{\pi}$ förväntas vara "rättklassificerings"-sannolikheter, medan övriga element står för "felklassificering".

Eftersom varje rad av π -matrisen summerar till 1, så vi gör en parameterisering som innehåller $(K - 1)$ parametrar per rad och deducerar den sista kolumnen som $1 -$ (summan av övriga kolumner).

3.7 Hessianen och osäkerhetsuppskattning

För att kunna uppskatta osäkerheten i de skattade felklassificeringssannolikheterna använde vi funktionen `optim` i R, med argumentet `hessian = TRUE`. Detta gör att R automatiskt beräknar en numerisk approximation av den så kallade Hessianen, det vill säga en matris av andraderivator av log-likelihoodfunktionen (Held and Bové, 2020, s.31).

Vid maximum likelihood-skattning motsvarar den negativa Hessianen den observerade Fisherinformationen. Under vissa antaganden kan man använda inversen av denna för att approximera kovariansmatrisen för parameterestimatet, vilket i sin tur gör det möjligt att beräkna standardfel och konfidensintervall (Held and Bové, 2020, s.28).

4 Resultat

4.1 Månadsvisa analyser

I detta avsnitt redovisas resultaten från den månadsvisa analysen för arterna *Alnus*, *Betula*, *Corylus* samt *Quercus*.

4.1.1 Uppskattning av förväxlingsmatrisen för månadsvisa räkningar

Genom att minimera den negativa log-likelihood-funktionen med R:s `optim`-funktion (metoden L-BFGS-B) erhöles en skattad förväxlingsmatris π enligt:

$$\pi = \begin{pmatrix} 0.860 & 0.039 & 0.012 & 0.089 \\ 0.010 & 0.375 & 0.014 & 0.600 \\ 0.772 & 0.039 & 0.000 & 0.189 \\ 0.099 & 0.339 & 0.000 & 0.562 \end{pmatrix}.$$

4.1.2 Resultat av Förväxlingsmatrisen

Rad 1: *Alnus*

- $\pi_{1,1} = 0.860$ visar att cirka 86% av pollen från *Alnus* klassificeras korrekt.
- $\pi_{1,2} = 0.039$ innebär att ungefär 3.9% felklassificeras som *Betula*.
- $\pi_{1,3} = 0.012$ visar att endast 1.2% hamnar i kategorin *Corylus*.
- $\pi_{1,4} = 0.089$ indikerar att 8.9% felklassificeras som *Quercus*.

Rad 2: *Betula*

- $\pi_{2,1} = 0.010$: Endast 1% felklassificeras som *Alnus*.
- $\pi_{2,2} = 0.375$: 37.5% klassificeras korrekt som *Betula*.
- $\pi_{2,3} = 0.014$: 1.4% hamnar som *Corylus*.
- $\pi_{2,4} = 0.600$: En stor andel, 60%, felklassificeras som *Quercus*.

Rad 3: *Corylus*

- $\pi_{3,1} = 0.772$: Cirka 77.2% av *Corylus*-pollen klassificeras felaktigt som *Alnus*.
- $\pi_{3,2} = 0.039$: 3.9% felklassificeras som *Betula*.
- $\pi_{3,3} = 0.000$: Inga korrekt klassificerade som *Corylus*.
- $\pi_{3,4} = 0.189$: Resterande 18.9% hamnar som *Quercus*.

Rad 4: *Quercus*

- $\pi_{4,1} = 0.099$: 9.9% av *Quercus*-pollen klassificeras som *Alnus*.
- $\pi_{4,2} = 0.339$: 33.9% klassificeras som *Betula*.
- $\pi_{4,3} = 0.000$: Ingen felklassificering mot *Corylus*.
- $\pi_{4,4} = 0.562$: 56.2% klassificeras korrekt som *Quercus*.

4.2 Veckovisa analyser

Denna del presenterar resultaten från den veckovisa modelleringen, där en mer finmaskig tidsindelning används för att analysera modellens klassificeringsförmåga.

4.2.1 Uppskattning av felklassificeringsmatrisen för veckovisa räkningar

Genom att minimera den negativa log-likelihood-funktionen med R:s `optim`-funktion (metoden L-BFGS-B) erhöles en skattad felklassificeringsmatris π utifrån de veckovisa data enligt:

$$\pi = \begin{pmatrix} 0.930 & 0.000 & 0.002 & 0.068 \\ 0.056 & 0.694 & 0.022 & 0.229 \\ 0.000 & 0.297 & 0.072 & 0.632 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}.$$

4.2.2 Resultat av Förväxlingsmatrisen

Rad 1: *Alnus*

- $\pi_{1,1} = 0.930$ visar att cirka 93.0% av pollen från *Alnus* klassificeras korrekt.
- $\pi_{1,2} = 0.000$ innebär att en försumbar andel felklassificeras som *Betula*.
- $\pi_{1,3} = 0.002$ visar att endast 0.2% hamnar i kategorin *Corylus*.
- $\pi_{1,4} = 0.068$ indikerar att 6.8% felklassificeras som *Quercus*.

Rad 2: *Betula*

- $\pi_{2,1} = 0.056$: 5.6% felklassificeras som *Alnus*.
- $\pi_{2,2} = 0.694$: 69.4% klassificeras korrekt som *Betula*.
- $\pi_{2,3} = 0.022$: 2.2% hamnar som *Corylus*.
- $\pi_{2,4} = 0.229$: 22.9% felklassificeras som *Quercus*.

Rad 3: *Corylus*

- $\pi_{3,1} = 0.000$: Ingen *Corylus*-pollen klassificeras felaktigt som *Alnus*.
- $\pi_{3,2} = 0.297$: 29.7% felklassificeras som *Betula*.
- $\pi_{3,3} = 0.072$: 7.2% korrekt klassificerade som *Corylus*.
- $\pi_{3,4} = 0.632$: Resterande 63.2% hamnar som *Quercus*.

Rad 4: *Quercus*

- $\pi_{4,1} = 0.000$: Inga *Quercus*-pollen misstas för *Alnus*.
- $\pi_{4,2} = 0.000$: Ingen klassas som *Betula*.
- $\pi_{4,3} = 0.000$: Ingen hamnar i *Corylus*.
- $\pi_{4,4} = 1.000$: 100% klassificeras korrekt som *Quercus*.

4.3 Osäkerhetsuppskattning och Modellvalidering

I detta avsnitt presenteras hur vi uppskattar osäkerheten i de skattade parametrarna och validerar modellen. Vi uppskattar osäkerheten genom Hessianen (Fisherinformation) från optimeringen. Därefter jämförs de predikterade automatiska proportionerna med de observerade.

4.3.1 Jämförelse av predikterade och observerade proportioner

För att utvärdera modellens passform beräknades predikterade automatiska proportioner q_{pred} enligt:

$$q_{\text{pred}} = P \times \pi^\top,$$

där P är matrisen med observerade manuella proportioner och π är den skattade felklassificeringsmatrisen. Modellen bygger på antagandet att de observerade automatiska räkningarna $\mathbf{x}_i$ följer en multinomialfördelning:

$$\mathbf{x}_i \sim \text{Multinomial}(m_i, q_i).$$

Varje rad i P summerar till 1 och representerar den relativa förekomsten av olika pollenarter en viss månad/vecka. Genom att transponera π multipliceras varje proportion med sannolikheten att arten klassificeras som någon annan, enligt lagen om total sannolikhet.

De predikterade proportionerna q_{pred} jämfördes med de faktiskt observerade automatiska proportionerna q_{auto} . Skillnaderna (residualerna) visas i Tabell 4.3.1 och 4.3.1. Dessa representerar modellens avvikelse från verkligheten. Residualerna varierar mellan månader och arter, vilket tyder på att vissa arter klassificeras mer korrekt än andra.

Månad	Alnus	Betula	Corylus	Quercus
2024-03	-0.0777	-0.0264	0.6603	-0.0217
2024-04	-0.0257	-0.0524	0.2128	0.1027
2024-05	0.0200	0.0505	0.0600	-0.2021
2024-06	0.0298	0.1807	0.0472	-0.3923

Tabell 1: Skillnader mellan predikterade och observerade proportioner (månadsvisa räkningar)

Datum	Alnus	Betula	Corylus	Quercus
2024-03-03	0.3876	-0.1169	-0.0167	-0.3841
2024-03-10	0.0261	-0.3016	0.0015	-0.2160
2024-03-17	-0.0898	0.0105	-0.0006	-0.1263
2024-03-24	-0.0481	0.0533	0.0091	0.0530
2024-03-31	-0.0505	0.1125	0.0539	-0.1845
2024-04-07	-0.0142	0.2144	0.2301	-0.4473
2024-04-14	0.4461	0.0151	0.0827	-0.5862
2024-04-21	0.0932	0.3014	0.2210	-0.6263
2024-04-28	-0.0731	-0.1591	0.2216	-0.0562
2024-05-05	-0.0708	0.1615	0.2809	-0.0581
2024-05-12	0.0211	0.3383	0.4291	-0.3870
2024-05-19	0.1521	0.2838	0.5476	-0.1756
2024-05-26	0.0734	0.2484	0.5053	-0.2086
2024-06-02	-0.0071	0.4317	0.3174	-0.6631

Tabell 2: Skillnader mellan predikterade och observerade proportioner (veckovisa räkningar)

För att tydliggöra hur väl modellen överensstämmer med de observerade proportionerna visualiserar de predikterade och observerade proportionerna i Figur 6, 7. Figuren visar att modellen generellt har svårigheter att korrekt prediktera andelen Corylus, där proportionerna är särskilt höga i mars och maj. Quercus tenderar att underskattas i juni, vilket också framgår av modellen.

4.3.2 Osäkerhetsuppskattning via Hessianen

Vid optimering med `optim` i R används metoden L-BFGS-B, där vi även beräknade Hessianen genom att sätta `hessian = TRUE`, som nämnt i avsnittet 3.7. Inversen av Hessianen ger oss en approximativ varians-kovariansmatris, vilken kan användas för att beräkna standardfel och därmed 95% konfidensintervall för de fria parametrarna.

Nedan följer en sammanfattning av de skattade fria parametrarna (i vektorform, dvs. de 12 värdena som senare bygger upp π -matrisen), deras standardfel samt 95% konfidensintervall:

Parameter	Skattat värde	Standardfel	95% CI
$\pi_{1,1}$	0.9304	0.0062	[0.9183, 0.9426]
$\pi_{1,2}$	0.0000	0.0034	[-0.0066, 0.0066]
$\pi_{1,3}$	0.0017	0.0012	[-0.0006, 0.0040]
$\pi_{2,1}$	0.0556	0.0015	[0.0526, 0.0586]
$\pi_{2,2}$	0.6936	0.0038	[0.6862, 0.7010]
$\pi_{2,3}$	0.0218	0.0011	[0.0197, 0.0239]
$\pi_{3,1}$	0.0000	0.0446	[-0.0874, 0.0874]
$\pi_{3,2}$	0.2966	0.0245	[0.2486, 0.3445]
$\pi_{3,3}$	0.0716	0.0113	[0.0494, 0.0938]
$\pi_{4,1}$	0.0000	0.0025	[-0.0049, 0.0049]
$\pi_{4,2}$	0.0000	0.0055	[-0.0108, 0.0108]
$\pi_{4,3}$	0.0000	0.0010	[-0.0019, 0.0019]

Tabell 3: Skattade fria parametrar, standardfel och 95% konfidensintervall (baserat på Hessianen, månadsvisa data).

Parameter	Skattat värde	Standardfel	95% CI
$\pi_{1,1}$	0.8605	0.1389	[0.5883, 1.1326]
$\pi_{1,2}$	0.0387	0.2644	[-0.4796, 0.5571]
$\pi_{1,3}$	0.0117	0.0574	[-0.1007, 0.1241]
$\pi_{2,1}$	0.0104	0.0069	[-0.0032, 0.0239]
$\pi_{2,2}$	0.3749	0.0220	[0.3318, 0.4180]
$\pi_{2,3}$	0.0143	0.0049	[0.0047, 0.0238]
$\pi_{3,1}$	0.7723	1.1365	[-1.4552, 2.9999]
$\pi_{3,2}$	0.0391	2.1543	[-4.1834, 4.2616]
$\pi_{3,3}$	0.0000	0.4708	[-0.9228, 0.9228]
$\pi_{4,1}$	0.0991	0.0268	[0.0467, 0.1516]
$\pi_{4,2}$	0.3388	0.0845	[0.1731, 0.5045]
$\pi_{4,3}$	0.0000	0.0188	[-0.0368, 0.0368]

Tabell 4: Skattade fria parametrar, standardfel och 95% konfidensintervall (baserat på Hessianen, veckovisa data).

5 Diskussion och slutsatser

Syftet med denna studie var att utvärdera hur väl en automatisk pollenfälla, Poleno Jupiter, kan klassificera pollenkorn jämfört med manuella mätningar, samt att skatta sannolikheten för felklassificering. Genom att använda en multinomialmodell med en felklassificeringsmatris kunde vi kvantitativt uppskatta dessa sannolikheter för varje art.

5.1 Tolkning av felklassificeringsmatris – månadsvis

5.1.1 Artberoende prestanda och svårigheter med månadsvis analys

Den månadsvisa felklassificeringsmatrisen visar att vissa arter, som *Ek* (*Quercus*) och *Al* (*Alnus*), ofta klassificeras korrekt. Detta kan förklaras av att dessa arter har tydliga visuella kännetecken eller blommar under tydligt avgränsade perioder. För andra arter är det dock betydligt svårare. Till exempel tenderar *Björk* (*Betula*) att felklassificeras som *Quercus*, särskilt eftersom båda arterna har sin huvudsakliga blomningstid i maj (se Figur 2 och 4). Denna säsongsoverlappning försvårar modellens förmåga att skilja dem åt. Ett liknande, men ännu allvarigare fall gäller *Hassel* (*Corylus*), som ofta klassificeras felaktigt, ofta som *Alnus*. *Corylus* blommar tidigt, vanligtvis i mars, vilket gör att dess mönster försvinner när data slås samman på månadsnivå. I båda fallen leder aggregeringen till att viktiga temporala signaler går förlorade, vilket i sin tur försämrar parameteridentifieringen i modellen.

5.1.2 Begränsningar med månadsvis aggregering

Den månadsvisa aggregeringen döljer kortsiktiga variationer i pollenförekomst, vilket gör att förändringar i början eller slutet av en arts blomningsperiod inte fångas upp. Exempelvis kan en art som blommar i slutet av månaden få sitt signalvärde utspädd i periodens medelvärde. Detta syns tydligt vid jämförelse med de veckovisa mätningarna (se Figur 2 och 3), som ger en mer detaljerad bild av säsongsförloppet. Att ignorera dessa kortsiktiga variationer begränsar modellens förklaringskraft och riskerar att ge missvisande parameterestimater.

En ytterligare svårighet uppstår när arter med olika blomningstoppar samlas i samma tidsenhet. Den grova aggregeringen gör att separata säsongssignaler smälter samman, vilket försvårar tolkningen av proportionerna och ökar osäkerheten i parameteruppskattningarna. Det kan i sin tur leda till systematiska felklassificeringar och missvisande slutsatser.

Att ignorera dessa kortsiktiga variationer och artificiellt introducera överlapp begränsar modellens förklaringskraft och riskerar att ge missvisande parameterestimater.

5.1.3 Slutsats för månadsvis analys

Sammantaget visar resultaten att den månadsvisa analysen inte är tillräcklig för att uppnå tillförlitlig artidentifiering, särskilt för arter med smala eller tidiga säsonger som *Corylus*, eller arter som hamnar i samma period som andra, som *Betula* och *Quercus*. Resultaten bekräftar därmed teorin om att säsongsvariation är avgörande för parameteridentifiering (avsnitt 3.1). En mer finmaskig tidsupplösning, exempelvis veckovis, framstår som nödvändig för att möjliggöra noggranna och stabila skattningar av felklassificeringssannolikheter.

5.2 Tolkning av felklassificeringsmatris – veckovis

5.2.1 Högre diagonalelement i den veckovisa modellen

Den veckovisa felklassificeringsmatrisen visar tydliga förändringar i de diagonala elementen jämfört med den månadsvisa analysen. Detta blir särskilt tydligt vid jämförelse av de skattade π -matriserna, där högre diagonalelement innebär att modellen tillskriver högre sannolikhet för korrekt klassificering.

- *Alnus*: $\pi_{1,1}$ ökar från 0.860 till 0.930
- *Betula*: $\pi_{2,2}$ ökar från 0.375 till 0.694
- *Corylus*: $\pi_{3,3}$ ökar från 0.000 till 0.297
- *Quercus*: $\pi_{4,4}$ ökar från 0.562 till 1.000

Detta innebär att modellen, enligt skattningarna, i den veckovisa analysen i större utsträckning antar att pollen klassificeras korrekt. Det är särskilt tydligt för arter som tidigare var svåra att identifiera, såsom *Corylus* (Hassel), som i den månadsvisa modellen nästan aldrig klassificerades korrekt, men där den veckovisa modellen skattar en korrekt klassificering i nästan 30% av fallen.

Förbättringen i de skattade diagonalelementen kan bero på att den veckovisa modellen bättre fångar snäva och artunika säsongsmönster, såsom den tidiga toppen för *Corylus* i mars (se Figur 4 och Figur 5). Även *Betula* och *Alnus* uppvisar ökade diagonalskattningar, vilket tyder på att modellen är känslig för kortsiktiga variationer i artförekomst, något som döljs vid månadsvis aggregering.

5.2.2 Tolkning och säsongsvariation

Förbättringarna i den skattade klassificeringsnoggrannheten kan i stor utsträckning förklaras av att modellen i den veckovisa analysen får tillgång till mer detaljerad information om säsongsförloppet. I den månadsvisa modellen slås flera toppar och dalar samman till ett medelvärde, vilket gör att viktiga temporala signaler riskerar att gå förlorade.

I den veckovisa modellen blir de uppskattade säsongsvariationerna däremot tydligt urskiljbara. Exempelvis syns en skarp topp för *Betula* kring mitten av maj, medan *Quercus* når sin högsta nivå först under juni månad, särskilt runt dess andra hälft (se Figur 5). Denna tidsmässiga separation mellan arter gör att modellen får bättre förutsättningar att särskilja deras klassificeringsmönster, vilket i sin tur leder till högre skattade diagonalposter i π -matrisen. Detta bör dock tolkas med försiktighet, eftersom den sanna felklassificeringsförmågan inte är observerbar.

5.2.3 Praktiska begränsningar

Trots fördelarna med en finmaskig tidsupplösning innebär den veckovisa modellen också utmaningar. Vissa veckor innehåller få observationer, särskilt för arter med låg förekomst. Detta kan leda till större osäkerhet i skattningarna och ökade standardfel, vilket också bekräftas av de bredare konfidensintervallen i Tabell 3 och Tabell 4.

I praktiska tillämpningar krävs därför en avvägning mellan att fånga säsongsmönster med hög temporal upplösning och att ha tillräckligt många observationer för stabila skattningar. I denna studie visar dock resultaten att den veckovisa upplösningen ger en mer träffsäker och tillförlitlig analys, utan att osäkerheten blir orimligt stor.

5.3 Sammanfattning av artvis prestanda med fokus på *Corylus*

Klassificeringsprestandan varierar tydligt mellan olika arter. *Quercus* och *Alnus* klassificeras oftast korrekt, vilket kan bero på att de har tydliga säsongstoppar och särpräglade egenskaper. *Betula* uppvisar dock viss förväxlingsproblematik, särskilt med *Quercus*, vilket sannolikt hänger samman med att dessa arter blommar samtidigt i maj.

De största svårigheterna observeras för *Corylus*, som nästan aldrig klassificeras korrekt i den månadsvisa modellen. Detta kan förklaras av att arten har en mycket tidig och kort blomningsperiod i mars, som riskerar att förloras vid tidsmässig aggregering. I den veckovisa modellen syns däremot en tydlig förbättring, modellen lyckas bättre fånga den smala säsongstoppen för

Corylus, vilket antyder att en högre tidsupplösning är avgörande för korrekt klassificering av arter med korta och tidiga blomningsfönster.

5.4 Reflektion kring antagandet om multinomialfördelning

I denna studie antas att pollenräkningarna följer en multinomialfördelning, vilket innebär att vi modellerar antalet klassificeringar i varje tidsperiod som ett utfall från en kategoriindelad sannolikhetsmodell. Givet ett observerat antal pollen per vecka eller månad, antas att varje pollenhändelse tillhör exakt en art, och att den totala fördelningen följer en sannolikhetsvektor p_i .

Antagandet bygger på att varje observation är oberoende av de övriga, och att sannolikheten att tillhöra en viss art är konstant inom perioden. I praktiken är detta en förenkling, tekniska fel i klassificeringssystemet kan påverka flera observationer samtidigt, vilket skapar beroende. Dessutom kan miljöfaktorer som påverkar pollenfördelningen variera även inom en vecka eller månad, vilket bryter mot antagandet om konstant sannolikhet.

Trots dessa begränsningar ger modellen en rimlig approximation, särskilt när data aggregeras över flera dagar. Då tenderar kortsiktiga variationer att jämnas ut, och modellen kan ge meningsfulla skattningar av klassificeringsbeteendet. Den är också enkel att tolka och möjliggör en tydlig formulering av felklassificeringssannolikheter i matrisform.

Antagandet om en multinomialfördelning möjliggör dessutom likelihoodbaserad inferens, vilket gör det möjligt att skatta felklassificeringssannolikheter med maximum likelihood-metoden och att uppskatta osäkerheten genom Hessianen. Trots modellens förenklingar visar resultaten att den är tillräckligt bra för att dra meningsfulla slutsatser om klassificeringsförmågan för olika arter.

5.5 Modellvalidering och osäkerhetsuppskattning

Modellvalidering syftar i denna studie på att utvärdera hur väl vår statistiska modell lyckas återge verkligheten, alltså hur bra de predikterade automatiska proportionerna q_{pred} , som bygger på manuella mätningar och den skattade felklassificeringsmatrisen π , stämmer överens med de faktiskt observerade automatiska proportionerna q_{auto} .

Vikten av att validera automatiska mätinstrument med manuella referenser understryks även i tidigare forskning. I en översiktsartikel om automatiserad pollenigenkänning påpekar Buters et al. (2024) att det fortfarande saknas ett "gold standard"-system för validering, vilket gör jämförelser med manuella Hirst-data till det mest använda tillvägagångssättet. Det stärker relevansen i vår valideringsmetod, där manuella mått och modellerad korrigering jämförs mot automatiska observationer.

Att jämföra automatiskt klassificerade proportioner med manuella referensvärden är en etablerad metod i tidigare studier, till exempel i den tekniska rapporten från Swisens (2024), där en liknande validering av en automatisk pollenfälla genomfördes AG (2024).

I denna studie görs jämförelsen genom att beräkna residualer (skillnaden mellan modellens prediktion och faktisk observation) och analysera mönster i dessa. Om modellen fungerar väl bör residualerna vara små och inte systematiskt snedfördelade för vissa arter eller tidsperioder.

För att utvärdera modellens prestanda analyserades både hur väl modellen förklarar de observerade automatiska pollenproportionerna samt hur osäker parameteruppskattningen är. Denna validering baseras på två centrala metoder: (1) jämförelse mellan predikterade och observerade proportioner, och (2) konfidensintervall baserade på Fisherinformation via Hessianen.

5.5.1 Jämförelse mellan q_{pred} och q_{auto}

I avsnitt 4.3.1 definierades hur de predikterade proportionerna $q_{\text{pred}} = P \times \pi^{\top}$ beräknades, där P är de manuellt observerade proportionerna och π den skattade felklassificeringsmatrisen.

Jämförelsen med de observerade automatiska proportionerna q_{auto} ger residualer (modellens prediktion minus faktisk observation), vilka återges i Tabell 4.3.1 för månadsdata och Tabell 4.3.1 för veckodata.

Residualerna varierar mellan arter och tidpunkter, vilket tyder på att modellen inte är lika träffsäker för alla situationer. Exempelvis uppvisar *Corylus* stora positiva residualer i mars i månadsdatan (Tabell 4.3.1), vilket tyder på att modellen överskattar *Corylus* under denna period. I den veckovisa modellen syns dock mindre residualer i början av mars, medan förbättrad överensstämmelse uppnås i april och maj (Tabell 4.3.1). Det tyder på att den veckovisa modellen bättre fångar säsongsmönstret för *Corylus*.

Liknande mönster observeras för *Quercus*, där modellen konsekvent underskattar förekomsten i juni, både i månads och veckodata. Detta antyder att π -matrisen inte fullt ut fångar den sena säsongstoppen för *Quercus*.

Dessa mönster pekar på att modellen inte helt fångar de dynamiska säsongsförloppen, något som bör hanteras i framtida modeller, exempelvis genom att tillåta att π varierar över tid.

Liknande mönster har rapporterats i en utvärdering av det automatiska klassificeringssystemet Swisens, där vissa arter, såsom *Quercus* och *Betula*, uppvisade hög överensstämmelse med manuella referensmätningar, medan andra arter var betydligt svårare att klassificera korrekt AG (2024).

5.5.2 Osäkerhetsuppskattning via Hessianen

Vid optimering med `optim` i R angavs argumentet `hessian = TRUE`, vilket gör att R approximerar Hessianen numeriskt. Inversen av Hessianen utgör en approximation till kovariansmatrisen för parameterestimaten. På så vis kan vi beräkna standardfel och 95% konfidensintervall för varje fri parameter i π -matrisen.

Tabell 3 och 4 sammanfattar resultaten: skattat värde, standardfel och konfidensintervall för de 12 fria parametrarna i π . Osäkerheten varierar kraftigt, parametrar som rör vanliga arter (som $\pi_{1,1}$ för *Alnus*) har mycket små standardfel, medan parametrar för sällsynta arter (t.ex. $\pi_{3,1}$ för *Corylus*) har mycket breda intervall och i vissa fall negativa gränser, vilket antyder svag identifikation.

Dessa resultat visar att det är svårt att skatta felklassificeringssannolikheter för arter med låg förekomst och smala säsongsfönster, som *Corylus* och till viss del *Betula*. Detta understryker vikten av att använda en högre tidsupplösning, som veckodata, för att förbättra träffsäkerheten och minska osäkerheten i skattningarna.

5.5.3 Tolkning av konfidensintervall – månadsvis modell

Konfidensintervallen som beräknats med hjälp av Hessianen (Tabell 3 och 4) ger viktig information om hur säkra våra skattningar av felklassificeringssannolikheter är. För den månadsvisa modellen ser vi att flera parametrar har mycket små standardfel och snäva intervall, exempelvis $\pi_{1,1}$ (för *Alnus*) ligger i intervallet $[0.9183, 0.9426]$, vilket tyder på hög säkerhet i denna uppskattning. Detta kan förklaras av att *Alnus* förekommer rikligt och under en tydlig period i datan, vilket ger god information till modellen.

Samtidigt är vissa konfidensintervall mycket breda eller till och med inkluderar negativa värden, vilket är tecken på stor osäkerhet. Ett exempel är $\pi_{3,1}$, som motsvarar sannolikheten att *Corylus* felklassificeras som *Alnus*. Detta intervall sträcker sig från -0.0874 till 0.0874, vilket inte är meningsfullt eftersom sannolikheter inte kan vara negativa. Det indikerar att modellen inte har tillräcklig information för att skatta denna parameter med precision, troligtvis på grund av att *Corylus* är ovanlig i datamängden och dessutom överlappar säsongsmässigt med andra arter.

Sådana resultat innebär att den skattade π -matrisen bör tolkas med försiktighet för vissa arter. Framför allt visar detta att modellen fungerar bättre för vanliga arter som *Alnus* och

Quercus, men är betydligt osäkrare för arter som *Corylus* och till viss del *Betula*.

5.5.4 Tolkning av konfidensintervall - veckovis modell

Tabell 4 visar de skattade fria parametrarna i felklassificeringsmatrisen π baserat på veckovisa data, tillsammans med standardfel och 95% konfidensintervall. Generellt sett ser vi att vissa parametrar skattas med hög precision, medan andra uppvisar mycket stor osäkerhet.

Exempelvis har $\pi_{2,2}$, som representerar sannolikheten att *Betula* korrekt klassificeras som sig själv, ett smalt konfidensintervall $[0.3318, 0.4180]$, vilket indikerar god precision. Detsamma gäller $\pi_{2,3}$ och $\pi_{4,1}$, som båda har små standardfel och tydliga intervall bort från noll. Detta tyder på att modellen med relativt hög säkerhet lyckas uppskatta vissa felklassificeringsmönster.

I kontrast visar parametrar som $\pi_{3,1}$ och $\pi_{3,2}$ mycket breda intervall, t.ex. $[-1.4552, 2.9999]$ respektive $[-4.1834, 4.2616]$, vilket antyder att modellen har svårt att identifiera felklassificeringssannolikheter för *Corylus*. Detta stämmer överens med tidigare observationer om att *Corylus* är en svårklassificerad art, särskilt eftersom den förekommer tidigt på säsongen med låg frekvens, vilket resulterar i få observationer och därmed hög osäkerhet.

Sammanfattningsvis visar denna tabell att modellen fungerar rimligt bra för de vanligare arterna med tydliga säsonger (*Betula*, *Quercus*), men att den är mycket osäker för ovanliga arter som *Corylus*. Detta bekräftar slutsatsen att sällsynta och tidiga arter kräver fler observationer eller alternativ modellering för att uppnå tillförlitliga skattningar.

5.6 Praktisk tillämpning av maximum likelihood i R

Metodologiskt visade det sig att numerisk optimering i R med `optim` fungerade effektivt, och att Hessianen kunde användas för att beräkna standardfel på ett tillförlitligt sätt. Denna metod stöds av kurslitteraturen Likelihood and Bayesian Inference (kap. 2.3.1), där det förklaras hur den negativa Hessianen kan tolkas som observerad Fisherinformation och därmed ge konfidensintervall för ML-skattningar.

5.7 Slutsats

Sammanfattningsvis visar resultaten att Poleno Jupiter har potential att automatisera pollenklassificering på ett tillfredsställande sätt för vissa arter, särskilt *Alnus* och *Quercus*, där den klassificeringsmässiga träffsäkerheten är relativt hög. För andra arter, som *Corylus*, kvarstår dock utmaningar med felklassificering. Den månadsvisa modellen har svårt att fånga artens tidiga säsong, medan den veckovisa modellen uppvisar en viss förbättring. Trots detta är osäkerheten i skattningarna fortsatt hög, vilket understryker behovet av finare tidsupplösning och eventuellt vidare metodutveckling.

Resultaten understryker vikten av att använda modeller som beaktar säsongsvariationer och finmaskigare tidsupplösning, eftersom sådana variationer är avgörande för att korrekt kunna identifiera artvisa klassificeringsmönster. Studien visar också att en statistisk modellering av felklassificering i kombination med numerisk optimering är en användbar metod för att utvärdera och förbättra automatiska klassificeringssystem, men att tillförlitligheten varierar beroende på artens frekvens och säsongsmönster.

5.8 Förslag till framtida förbättringar

Framtida förbättringar bör fokusera på att undersöka hur modellens identifierbarhet påverkas av andra grader av tidsmässig aggregering. En sådan analys skulle ge fördjupad förståelse för vilken tidsupplösning som krävs för att möjliggöra tillförlitliga parameteruppskattningar.

Utöver detta bör modellen göras mer flexibel över tid, exempelvis genom att låta felklassificeringsmatrisen π variera med säsong, månad eller andra yttre faktorer som temperatur. Detta

kan hjälpa till att fånga förändringar i klassificeringsprestanda som idag antas vara konstanta inom varje period.

Vidare hade modellen kunnat utökas till att omfatta fler pollenarterna än de som analyserats i denna studie, vilket skulle ge en mer heltäckande bild av den automatiserade klassificeringens prestanda. Det vore även värdefullt att inkludera fler kontextuella variabler, såsom geografisk plats och meteorologiska förhållanden, då dessa kan påverka både förekomsten av olika arter och risken för felklassificering.

6 Appendix

Tabell 5: Manuellt observerade proportioner och antal pollenkorn per art och månad

Månad	p_{Alnus}	p_{Betula}	p_{Corylus}	p_{Quercus}	Alnus	Betula	Corylus	Quercus
2024-03-01	0.8664	0.0266	0.1062	0.0008	2121	65	260	2
2024-04-01	0.2588	0.6959	0.0247	0.0206	251	675	24	20
2024-05-01	0.0010	0.7893	0.0001	0.2096	11	9128	1	2424
2024-06-01	0.0000	0.8862	0.0000	0.1138	0	109	0	14

Tabell 6: Automatiskt klassificerade proportioner och antal pollenkorn per art och månad

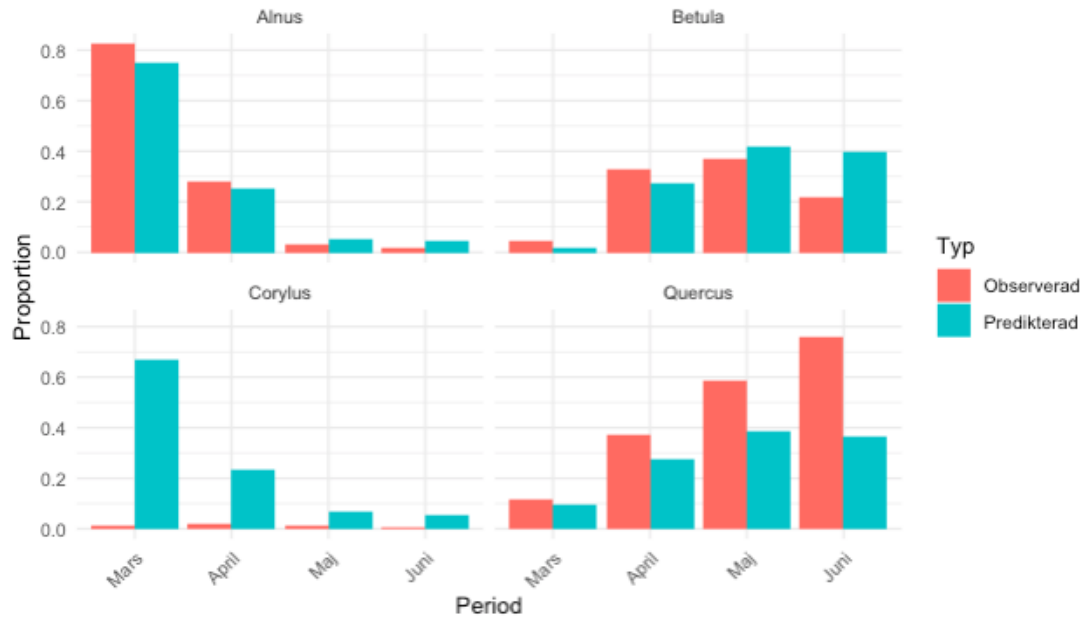
Månad	q_{Alnus}	q_{Betula}	q_{Corylus}	q_{Quercus}	Alnus	Betula	Corylus	Quercus
2024-03-01	0.8255	0.0473	0.0110	0.1170	8423	483	103	1194
2024-04-01	0.2775	0.3287	0.0182	0.3757	1116	1322	73	1511
2024-05-01	0.0301	0.3713	0.0111	0.5875	1358	16756	502	26509
2024-06-01	0.0147	0.2199	0.0089	0.7565	41	615	25	2116

Tabell 7: Manuellt observerade proportioner och antal pollenkorn per art och vecka

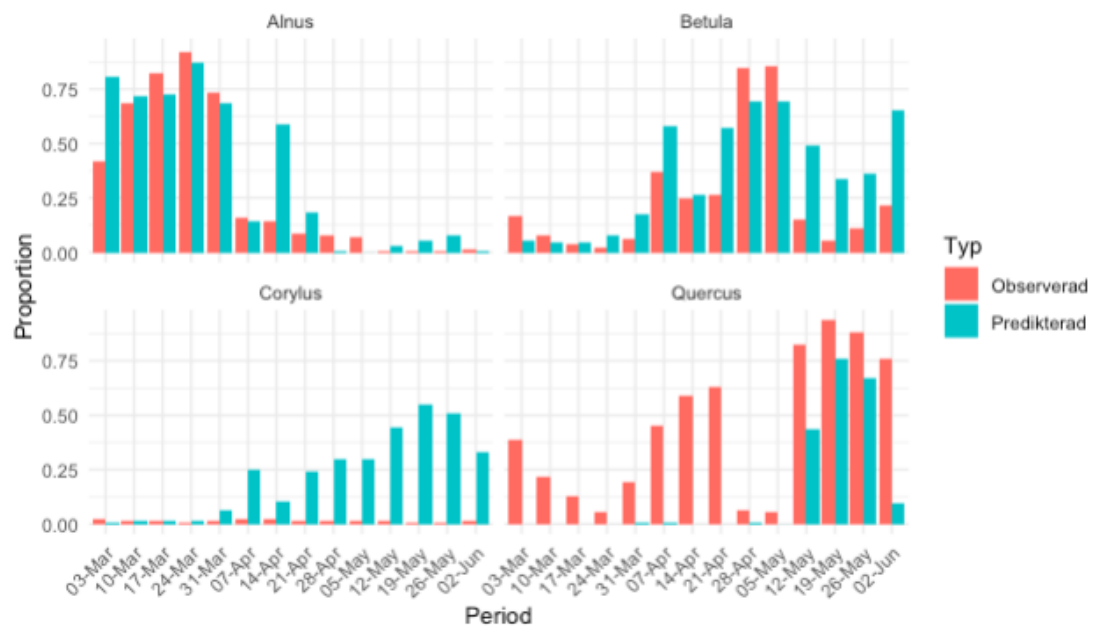
Vecka	p_{Alnus}	p_{Betula}	p_{Corylus}	p_{Quercus}	Alnus	Betula	Corylus	Quercus
2024-03-03	0.870	0.000	0.130	0.000	160	0	24	0
2024-03-10	0.769	0.000	0.231	0.000	289	0	87	0
2024-03-17	0.784	0.000	0.216	0.000	378	0	104	0
2024-03-24	0.936	0.036	0.028	0.000	938	36	28	0
2024-03-31	0.737	0.184	0.071	0.008	188	47	18	2
2024-04-07	0.154	0.824	0.016	0.008	60	323	6	3
2024-04-14	0.633	0.333	0.033	0.000	19	10	1	0
2024-04-21	0.195	0.805	0.000	0.000	8	33	0	0
2024-04-28	0.001	0.996	0.000	0.003	4	5190	0	17
2024-05-05	0.000	0.999	0.000	0.000	1	2527	0	1
2024-05-12	0.000	0.561	0.000	0.439	0	1256	0	980
2024-05-19	0.003	0.235	0.006	0.762	5	421	1	1367
2024-05-26	0.035	0.296	0.000	0.670	4	34	0	77
2024-06-02	0.000	0.905	0.000	0.095	0	67	0	7

Tabell 8: Automatiskt klassificerade proportioner och antal pollenkorn per art och vecka

Vecka	q_{Alnus}	q_{Betula}	q_{Corylus}	q_{Quercus}	Alnus	Betula	Corylus	Quercus
2024-03-03	0.4218	0.1681	0.0261	0.3841	291	116	18	265
2024-03-10	0.6894	0.0792	0.0151	0.2160	868	100	19	272
2024-03-17	0.8198	0.0378	0.0161	0.1263	1889	87	37	291
2024-03-24	0.9192	0.0243	0.0035	0.0530	4956	131	19	286
2024-03-31	0.7371	0.0597	0.0108	0.1924	889	72	13	232
2024-04-07	0.1572	0.3676	0.0202	0.4550	171	400	22	495
2024-04-14	0.1433	0.2519	0.0186	0.5863	54	95	7	221
2024-04-21	0.0884	0.2638	0.0177	0.6263	35	106	7	248
2024-04-28	0.0705	0.8506	0.0158	0.0595	772	8869	165	620
2024-05-05	0.0712	0.8546	0.0156	0.0585	455	5461	100	374
2024-05-12	0.0086	0.1517	0.0144	0.8252	123	2158	205	11740
2024-05-19	0.0044	0.0546	0.0034	0.9380	58	730	46	12531
2024-05-26	0.0044	0.1120	0.0054	0.8781	9	227	11	1780
2024-06-02	0.0136	0.2179	0.0108	0.7577	15	241	12	838



Figur 6: Månadsvisa proportioner



Figur 7: Veckovisa proportioner

7 Referens lista

Referenser

- AG, S. (2024). Performance improvement of new swisens classifier evaluated using autopollen first timeseries.
- Buters, J. T. M., Cecchi, L., Sofiev, M., Crouzy, B., Oteros, J., Galán, C., Thibaudon, M., and et al. (2024). Automatic detection of airborne pollen: an overview. *Aerobiologia*, 40(1):3–24. Hämtad 2025-05-13.
- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.
- Held, L. and Bové, D. S. (2014). Applied statistical inference. *Springer, Berlin Heidelberg*, doi, 10(978-3):16.
- Held, L. and Bové, D. S. (2020). Likelihood and bayesian inference. *Statistics for Biology and Health. Springer, Berlin, Heidelberg*.
- Naturhistoriska riksmuseet (2024). Palynologiska samlingar – pollenövervakning. <https://www.nrm.se/natur--och-miljoovervakning/pollenovervakning/palynologiska-samlingar>. Hämtad 2024-05-17.
- Pollenrapporten (2024). Pollenprognos och statistik. <https://www.pollenrapporten.se>.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Swisens AG (2025a). Real-time bioaerosol monitor. <https://www.swisens.ch/en/swisenspoleno-jupiter>.
- Swisens AG (2025b). Swisens data explore. <https://dataexplorer.swisens.ch/eventviewer>.