



Stockholms
universitet

Optimering av Svergies pollenmättningsnätverk: En analys av rumslig täckning och förbättrings- potential

Alexander Sebek

Kandidatuppsats 2025:4
Matematisk statistik
Juni 2025

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Optimering av Svergies pollenmättningsnätverk: En analys av rumslig täckning och förbättringspotential

Alexander Sebek*

Juni 2025

Sammanfattning

Denna studie undersöker hur lika pollensäsongerna är mellan olika mätstationer i Sverige i syfte att förbättra pollenprognoser för allergiker. Målet är att identifiera stationer som förmedlar överlappande information, så att automatiska fällor kan placeras mer effektivt. Vi analyserar pollentyperna al, alm, björk, ek, gräs, gråbo, hassel, sälg och vide, vilka rapporteras på *Pollenrapporten.se*. Stationerna grupperas med hierarkisk klustring där avstånden baseras på Pearson- respektive cosinusavstånd och där Wards länkmetsod används. Optimal gruppindelning fastställs med silhuettanalys. Klassisk multidimensionell skalning (MDS) ger tvådimensionella representationer av avstånden, vilka kompletteras med tabeller som visar de närmaste stationsparen. Resultaten visar att stationer som ligger nära varandra geografiskt också uppvisar de minsta avstånden i både Pearson- och cosinusavstånden, vilket tyder på att flera stationer i samma område kan vara överflödiga.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: alexsebek7@gmail.com. Handledare: Martin Sköld.

Abstract

This study examines how similar pollen seasons are across measurement stations in Sweden so that pollen forecasts for people with allergies can improve. The goal is to find stations that give overlapping information, making it possible to place the automatic traps more efficiently. We analyze the pollen types alder, elm, birch, oak, grass, mugwort, hazel, willow and goat willow reported on pollenrapporten.se. The stations are grouped with hierarchical clustering based on Pearson- and cosine-distances and with Ward's Method as linkage method. The best number of groups is found with silhouette analysis. Classical MDS provides two dimensional views of the distances, and we showcase tables that list the closest station pairs. The results show that stations that lie close to each other in geography also have the smallest distances in both Pearson and cosine, which suggests that several stations in the same area can be redundant.

Förord

Jag vill börja med att varmt tacka min handledare Martin, för idéer, inspiration och stöd under hela arbetets gång. Tack också till Naturhistoriska riksmuseet för att ha tillhandahållit den pollendata som utgör grunden för denna studie. Ett stort tack riktar jag till mina klasskamrater Stefano och Sebastian, för vänskap och emotionellt stöd under skrivprocessen.

Slutligen vill jag redogöra för hur jag använt moderna språkmodeller och verktyg i arbetet: jag har använt ChatGPT för att generera och finslipa LaTeX kod, rätta och optimera R kod samt för stavnings och klarhetshjälp i texten.

Innehåll

Sammanfattning	1
Abstract	2
Förord	3
1 Inledning	6
2 Datamaterial	8
2.1 Insamling av datamaterial	8
2.1.1 Pollen	8
3 Teori och metod	10
3.1 Agglomerativ hierarkisk klustering	10
3.1.1 Avstånds- och likhetsmått	10
3.1.2 Formell definition av hierarkisk klustering	12
3.1.3 Länkmeter	13
3.2 Silhuettmetoden	13
3.2.1 Tolkning av silhuettbredd	14
3.3 Multidimensionell skalning (MDS)	14
3.4 Databearbetning	16
3.4.1 Åtgärder	16
3.4.2 Databearbetning	16
3.4.3 Kod för metoder	19
4 Resultat	20
4.1 Cosinus-avstånd	20
4.2 Pearson-avstånd	23
4.3 Karta Över stationer	26

5	Diskussion	28
5.1	Jämförande analys av avståndsmått	28
5.1.1	Cosinusavstånd	28
5.1.2	Pearson-avstånd	29
5.2	Geografisk tolkning av kluster och placering av nya mätstationer . .	29
5.3	Problematiska stationer	30
5.4	Problemet med en enhetlig start- och slutpunkt	30
5.4.1	Alternativ: Zonindelad klustring	30
5.5	Viktningsstrategi baserat på population	30
5.6	Länkmetod i hierarkisk klustring	31
5.7	Utvärdering av Klusterkvalitet	31
5.8	Klassisk MDS	31
5.9	Rekommendationer	31
A	Bilaga	33

1 Inledning

I Sverige drabbas ungefär var tredje person av pollenallergi [Astma- och Allergiförbundet, 2025] – cirka 3,1 miljoner människor [Statistiska centralbyrån (SCB), 2025]. Allergiker får nedsatt livskvalitet under stora delar av pollensäsongen; vanliga besvär är rinnande eller kliande ögon, nästäppa, nysningar och svullna slemhinnor. Dessa symtom kan leda till huvudvärk, sömnstörningar och allmän trötthet [Karolinska Institutet, 2014]. Därför är snabba och pålitliga varningar för höga halter av allergent pollen viktiga, exempelvis via webbplatser som *Pollenrapporten.se*.

Idag räknas och identifieras varje pollenkorn manuellt under mikroskop. Nackdelen med denna manuella metod är att datainsamlingen är mycket tidskrävande, vilket gör det omöjligt att tillhandahålla information i realtid. Ofta uppstår en fördröjning på flera dagar mellan det faktiska pollensläppet och den rapporterade pollenhalten [Oteros et al., 2019]. Nyligen har bildanalysstyrda, automatiska pollenfällor utvecklats som kan räkna pollenkorn i realtid. Dessa fällor är dock dyra, så det är angeläget att välja ut optimala geografiska placeringar. Man kan antingen optimera för allergivarningar eller ren vetenskaplig övervakning. Det är inte säkert att de stationer som bäst varnar allergiker ligger där man får mest ny kunskap om pollensäsongernas rumsliga och tidsmässiga dynamik.

Denna studie fokuserar på de pollenarter som är mest allergena. De pollenarter som undersöks är al, alm, björk, ek, gräs, gråbo, hassel, sälg och vide [Naturhistoriska riksmuseet, 2025].

En liknande studie har genomförts och kommer att utgöra underlag för denna analys. År 2015 etablerades ett tätt nätverk av 27 manuella Hirst-typ fällor i Bayern, Tyskland, och hierarkisk klusteranalys genomfördes för att välja platserna för det slutliga pollennätverket [Oteros et al., 2019].

För att identifiera vilka stationer som ger överflödiga information och vilka som bäst kompletterar varandra tillämpar vi även agglomerativ hierarkisk klustring. Metoden grupperar stationer efter likhet i pollendata utan att kräva några förhands-

antaganden om antalet kluster. Vid varje steg slås de två mest lika stationerna samman tills samtliga stationer är inkluderade i ett dendrogram. Syftet är dels att urskilja stationer som rapporterar i stort sett samma mönster, dels att dela in stationerna i meningsfulla kluster för att kunna välja ett lämpligt antal optimala mätplatser med hjälp av multidimensionell skalning (MDS) och tabeller som visar de stationspar som har mest likartade säsonger.

2 Datamaterial

2.1 Insamling av datamaterial

2.1.1 Pollen

Naturhistoriska riksmuseet har övervakat pollen runt om i Sverige sedan 1973. Den avdelning som ansvarar för detta är Palynologiska laboratoriet i Stockholm. Mätningarna görs med hjälp av en Burkard-pollenfälla, *7-Day Recording Volumetric Spore Trap* [Burkard Scientific, 2025]. Maskinen är konstruerad för att samla luftprover under sju dygn. Inuti instrumentet sitter en roterande tejprulle monterad på en trumma. Trumman är kalibrerad så att tejpens rullning sker med konstant hastighet, motsvarande sju dygns provtagning. Partiklar sugas in och fastnar på tejpens rulle, som är förtryckt med tidsmarkeringar. Efter provperioden kan man avläsa var på tejpens rulle partiklarna har samlats och därmed fastställa vilken dag och tid de inkom. Därefter skärs tejpens rulle i sektioner som monteras på mikroskopglas, och partiklarna räknas och identifieras i varje sektion under mikroskop. Eftersom man vet vilket klockslag varje segment representerar kan man beräkna koncentrationen över tid. Resultaten förs in i Naturhistoriska riksmuseets databassystem, varifrån tidsserier som visar variationer i pollenhalterna genereras.

Tabell 2.1: Översikt av variabler i pollendata

Variabel	Typ	Beskrivning
datum	kontinuerlig	Datum för pollenobservation
station	kategorisk	Plats i Sverige
koordinater	kontinuerlig (par)	Geografiska koordinater (latitud, longitud)
art	kategorisk	Pollenart
antal	diskret	Antal räknade pollenkorn
faktor	kategorisk	Multiplicerande faktor
koncentration	kontinuerlig	Uppskattad pollenkoncentration (korn/m ³ luft)

I Tabell 2.1 visas de variabler som ingår i det datamaterial som tillhandahålls av Naturhistoriska riksmuseet. Den fortsatta analysen bygger framför allt på följande variabler:

- datum
- station
- art
- koncentration

Dessa variabler behandlas mer ingående i avsnittet *Databearbetning* längre fram.

3 Teori och metod

3.1 Agglomerativ hierarkisk klustering

Hierarkisk klustering är en lämplig metod eftersom den gör det möjligt att gruppera stationerna utifrån likheter i deras pollensäsonger, vilket passar väl med den tidsseriedata vi använder. Multidimensionell skalning (MDS) är samtidigt lämpligt då metoden låter oss visualisera hur stationerna förhåller sig till varandra i ett lågdimensionellt rum, vilket underlättar tolkningen av klustringsresultat.

I detta avsnitt redogörs för den agglomerativa hierarkiska klustringsmetod som används i studien, samt för de avstånds- och likhetsmått som ligger till grund för grupperingarna. Huvudreferensen är [Miyamoto, 2022], som formellt behandlar teorin bakom agglomerativ hierarkisk klustering.

Enligt [Miyamoto, 2022] är agglomerativ hierarkisk klustering en bottom-up-metod där varje observation inledningsvis utgör ett eget kluster. Därefter slås vid varje steg de två kluster samman som minskar den totala inomklustervariansen (Wards metod) eller som har minst avstånd enligt något annat länkkriterium. Processen upprepas tills alla objekt ingår i ett och samma kluster.

3.1.1 Avstånds- och likhetsmått

Innan vi kan utföra agglomerativ hierarkisk klustering måste vi välja ett avståndsmått. Hur avståndet definieras styr hur ”lika” eller ”olika” två observationsvektorer uppfattas vara och påverkar därför klusterstrukturen. I denna studie använder vi två mått: Pearson- och cosinusavstånd. Båda är känsliga för skillnader i skala och amplitud, vilket är viktigt då pollenhalter kan skilja sig kraftigt i nivå mellan stationer. När vi beräknar Pearsons korrelationsmatris sätter vi alla korrelationskoefficienter som är mindre än 0,5 till noll, i enlighet med den tyska studien [Oteros et al., 2019].

Allmän definition

Låt $x, y \in \mathbb{R}^M$ beteckna två datapunkter. Ett avståndsmått $D(x, y)$ ska uppfylla

$$D(x, y) \geq 0, \quad D(x, y) = 0 \Leftrightarrow x = y, \quad D(x, y) = D(y, x),$$

det vill säga det är icke-negativt, symmetriskt och blir noll endast om punkterna sammanfaller [Miyamoto, 2022, s. 4].

Ett likhetsmått $S(x, y)$ bör anta värden i enhetsintervallet $[0, 1]$, där 1 motsvarar maximal likhet och 0 ingen likhet alls. Givet ett avstånd $D(x, y)$ med lägsta respektive högsta observerade värde $D_{\min}$ och $D_{\max}$ fås

$$S(x, y) = 1 - \frac{D(x, y) - D_{\min}}{D_{\max} - D_{\min}},$$

vilket ger $S(x, y) = 1$ när $D(x, y) = D_{\min}$ och $S(x, y) = 0$ när $D(x, y) = D_{\max}$.

Givet ett normaliserat avstånd där $D_{\min} = 0$ och $D_{\max} = 1$ (t.ex. Pearson- eller cosinusavstånd) förenklas formeln till

$$S(x, y) = 1 - D(x, y) \quad \Longleftrightarrow \quad D(x, y) = 1 - S(x, y).$$

Cosinusavstånd

Cosinusmåttet fångar vinkelskillnaden oberoende av längd:

$$S_{\cos}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}, \quad D_{\cos}(x, y) = 1 - S_{\cos}(x, y).$$

[Miyamoto, 2022, s. 8]

Två stationer med pollenprofiler som uppvisar samma tidsförlopp men ligger på olika nivåer placeras därför nära varandra. Detta är relevant när den relativa säsongskurvan är mer informativ än de absoluta halterna.

Pearson-avstånd

Pearsons korrelationskoefficient definieras som

$$\rho(x, y) = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2} \sqrt{\sum_j (y_j - \bar{y})^2}},$$

där $\bar{x}$ respektive $\bar{y}$ är medelvärdet för varje tidsserie. Koefficienten mäter det linjära sambandet mellan två serier efter centrerat kring sina medelvärden. Ett korrelationsbaserat avstånd kan därefter definieras som

$$D_{\rho}(x, y) = 1 - \rho(x, y).$$

[Everitt et al., 2011, s. 8]

Koefficienten fångar alltså hur två kurvor rör sig i samma takt över tid, oavsett deras medelnivå och skala.

Sammanfattningsvis fångar *cosinusavståndet* hur lika säsongskurvornas form är, medan *Pearson-avståndet* jämför formen efter att båda kurvorna har flyttats till sitt medelvärde och skalats till samma spridning. Tillsammans ger de två måtten en bredare bild av hur lika pollenprofilerna är.

3.1.2 Formell definition av hierarkisk klustring

En partition av en mängd X är en familj $\mathcal{G} = \{G_1, \dots, G_K\}$ av icke-tomma, parvis disjunkta delmängder som uppfyller

$$\bigcup_{i=1}^K G_i = X, \quad G_i \cap G_j = \emptyset \quad (i \neq j).$$

Denna definition räcker dock inte för att beskriva hierarkiska klusterindelningar. För att ange en familj av kluster $\mathcal{G}(\alpha)$ som är hierarkisk måste vi först definiera begreppet förfining mellan två klusterindelningar [Miyamoto, 2022, s. 2].

Låt X vara en mängd och låt

$$\{\mathcal{G}(\alpha)\}_{\alpha \geq 0}$$

vara en familj av partitioner av X . Familjen kallas *hierarkisk* om det för alla $\alpha < \alpha'$ gäller att

$$\mathcal{G}(\alpha) \text{ är en förfining av } \mathcal{G}(\alpha'),$$

det vill säga

$$\forall G \in \mathcal{G}(\alpha) \exists G' \in \mathcal{G}(\alpha') \text{ så att } G \subseteq G'.$$

Antag nu att vi har ett avståndsmått $D(x, y)$. Vid den initiala tröskeln $\alpha = D_{\max} = \max_{x, y \in X} D(x, y)$ bildar varje punkt sitt eget kluster:

$$\mathcal{G}(D_{\max}) = \{\{x\} \mid x \in X\}.$$

Därefter minskas α successivt från $D_{\max}$ mot 0. Vid varje tröskelvärde förenas de klusterpar vilkas minsta mellanklusterdissimilaritet är $\leq \alpha$. Den iterativa processen genererar en agglomerativ hierarkisk klustring $\{\mathcal{G}(\alpha)\}_{\alpha \geq 0}$ [Miyamoto, 2022, s. 2].

3.1.3 Länkmetod

Efter att ha valt ett passande avståndsmått för data måste vi välja en länkmetod som styr hur klustren successivt slås ihop. Valet av länkmetod motiveras av den tyska pollenstudien [Oteros et al., 2019], som använder sig av Wards metod. Kort sagt fungerar Wards metod så att den i varje steg slår ihop de två kluster som medför den minsta ökningen av den totala inomklustervariansen. Metoden börjar med att varje station utgör ett eget kluster. Därefter beräknas hur mycket den samlade spridningen inom kluster skulle öka om man slog ihop varje tänkbart klusterpar. Det par som ger den minsta ökningen väljs, och processen upprepas tills alla stationer ingår i ett och samma kluster.

Ward-metoden

Sätt $E(G) = \sum_{x \in G} \|x - m(G)\|^2$. Dissimilariteten definieras som

$$\Delta E(G, G') = E(G \cup G') - E(G) - E(G').$$

Vid sammanslagning gäller

$$\Delta E(G_r, G_j) = \frac{(|G_p| + |G_j|) \Delta E(G_p, G_j) + (|G_q| + |G_j|) \Delta E(G_q, G_j) - |G_j| \Delta E(G_p, G_q)}{|G_r| + |G_j|}.$$

[Miyamoto, 2022, s. 24]

3.2 Silhuettmetoden

Silhuettmetoden ger för varje objekt i (i vårt fall en pollenstation) ett mått $s(i) \in [-1, 1]$ på hur väl det passar i sitt tilldelade kluster jämfört med intilliggande kluster. Antag att i tillhör kluster A och att $|A| \neq 1$. Definiera

$$a(i) = \frac{1}{|A| - 1} \sum_{\substack{j \in A \\ j \neq i}} d(i, j)$$

som medeldissimilariteten från i till övriga objekt i A . För varje annat kluster $C \neq A$ beräknas

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j),$$

och den minsta av dessa kallas

$$b(i) = \min_{C \neq A} d(i, C).$$

Silhuettbredden $s(i)$ definieras sedan som

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i), \\ 0, & a(i) = b(i), \\ \frac{b(i)}{a(i)}, & a(i) > b(i). \end{cases}$$

För enkelhets skull kan uttrycket skrivas

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

I de fall ett objekt ensamt bildar ett kluster ($|A| = 1$) sätts $s(i) = 0$. [Kaufman and Rousseeuw, 1990, s. 85]

Medelsilhuettbredden för ett kluster G_r med n_r punkter definieras som

$$\hat{s}_r = \frac{1}{n_r} \sum_{i \in G_r} s(i).$$

3.2.1 Tolkning av silhuettbredd

Nedan ges intervallen för silhuettbredden s och motsvarande tolkning.

Tabell 3.1: Tolkning av silhuettbreddens värdeintervall

Intervall	Tolkning
0,71 – 1,00	Stark struktur har identifierats.
0,51 – 0,70	Rimlig struktur har identifierats.
0,26 – 0,50	Strukturen är svag.
$s \leq 0,25$	Ingen väsentlig struktur har identifierats.

[Kaufman and Rousseeuw, 1990, s. 88]

3.3 Multidimensionell skalning (MDS)

Multidimensionell skalning (MDS) är en metod som utgår från en matris av likheter eller avstånd mellan objekt och placerar objekten som punkter i ett lågdimensionellt rum. Koordinaterna väljs så att de inbördes avstånden mellan punkterna

överensstämmer med de ursprungliga värdena så väl som möjligt. Denna grafiska representation gör det lättare att upptäcka mönster och grupperingar som annars är svåra att se [Borg and Groenen, 2005, s. 3].

Antag att de valda avståndsmåtten (Pearson- och cosinusavstånd) ger den symmetriska avståndsmatrisen

$$\mathbf{D} = (d_{ij})_{i,j=1}^N, \quad d_{ij} = d_{ji} \geq 0, \quad d_{ii} = 0.$$

Klassisk (metrisk) MDS syftar då till att konstruera en konfigurationsmatris

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times k},$$

så att det euklidiska avståndet mellan raderna i $\mathbf{X}$ reproducerar elementen i $\mathbf{D}$:

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2 \approx d_{ij}, \quad i, j = 1, \dots, N.$$

Varje rad $\mathbf{x}_i$ utgör därmed en k -dimensionell representation av station i som bevarar de ursprungliga cosinus- respektive Pearson-avstånden så bra som möjligt.

Algoritmen går i huvudsak till på följande sätt [Borg and Groenen, 2005, ss. 262–263]:

1. Beräkna den kvadrerade avståndsmatrisen

$$\mathbf{D}^{(2)} = [d_{ij}^2].$$

2. Utför *dubbelcentrering* med hjälp av centreringsmatrisen

$$\mathbf{C} = \mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \quad [\text{Borg and Groenen, 2005, s. 570}],$$

och bilda därefter matrisen

$$\mathbf{B} = -\frac{1}{2} \mathbf{C} \mathbf{D}^{(2)} \mathbf{C}.$$

3. Gör egenvärdesuppdelningen $\mathbf{B} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$, där $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N)$ och $\lambda_1 \geq \dots \geq \lambda_N$.
4. Välj de k största egenvärdena och motsvarande egenvektorer: $\mathbf{\Lambda}_k = \text{diag}(\lambda_1, \dots, \lambda_k)$, $\mathbf{Q}_k = [\mathbf{q}_1, \dots, \mathbf{q}_k]$. Koordinaterna i den lågdimensionella kartan fås som

$$\mathbf{X} = \mathbf{Q}_k \mathbf{\Lambda}_k^{1/2} \in \mathbb{R}^{N \times k}.$$

3.4 Databearbetning

3.4.1 Åtgärder

Det finns några problem med data som måste åtgärdas innan analysen kan fortsätta. Eftersom det rör sig om tidsseriesdata är det viktigt att varje station har en tidsserie av samma längd, så att pollenhalter kan jämföras dag för dag.

Ett första problem är att olika stationer har samlat in data under olika många år, vilket leder till tidsreor av varierande längd.

Under stora delar av året (vinter och höst) rapporteras inga pollenhalter alls. Därför behöver vi fastställa gemensamma start- och slutdatum för pollensäsongen i hela Sverige. Utan detta innehåller serierna många nollvärden, vilket kan få stationer att framstå som mer lika än de egentligen är. Eftersom Sverige är ett avlångt land med varierande klimat kan säsongen börja och sluta vid olika tidpunkter beroende på geografi.

Ett ytterligare problem är att vissa dagar under säsongen helt saknar mätvärden, trots att det borde finnas pollen. Detta kan bero på bemanningsrelaterade orsaker, som helgdagar, då inga prover har samlats in. Följden blir att en dag tilldelas värdet noll, trots att både dagen före och efter uppvisar höga halter.

Stationen Abisko uteslöts på grund av inkonsekventa mätningar, långa avbrott i årsserierna och många saknade datum.

Metoden som presenteras nedan syftar till att hantera dessa problem och skapa mer jämförbara och kompletta data för vidare analys. Arbetet är inspirerat av den tyska studien [Oteros et al., 2019]. Vilka moment som är hämtade från den studien och vilka som har utvecklats inom ramen för detta arbete klargörs i nästa avsnitt.

3.4.2 Databearbetning

Låt x_{ijkl} beteckna det observerade pollenvärdet vid

- station $i = 1, \dots, N$,
- pollenart $j = 1, \dots, P$,
- år $k = 1, \dots, J_i$ (antal år med mätdata för station i),
- dag på året $l = 1, \dots, L$.

Här är L det maximala antalet dagar (365), och J_i kan variera mellan stationerna.

Årssnittad profil

I den tyska studien användes endast säsongen 2015 [Oteros et al., 2019]. Eftersom vi har data från flera år definierar vi den genomsnittliga pollenprofilen som

$$\bar{x}_{ijl} = \frac{1}{J_i} \sum_{k=1}^{J_i} x_{ijkl}, \quad i = 1, \dots, N, \quad j = 1, \dots, P, \quad l = 1, \dots, L.$$

Genom att indexera med dag-på-året l får alla stationer serier av samma längd och saknade dagar fylls via aggregering över övriga år.

Urval av pollenarter

Urvalet av pollenarter baseras på vilka som orsakar flest allergiska reaktioner i befolkningen [Naturhistoriska riksmuseet, 2025]. Denna metod skiljer sig från den tyska studien, där man för varje art j beräknade den totala årskoncentrationen vid varje station och därefter valde de arter vars medianvärde över alla stationer översteg 1000.

- *Alnus* (al)
- *Betula* (björk)
- *Poaceae* (gräs)
- *Quercus* (ek)
- *Artemisia* (gråbo)
- *Corylus* (hassel)
- *Salix* (sälgt, vide)
- *Ulmus* (alm)

Dessa åtta arter står för majoriteten av pollenallergin i Sverige och är därför mest relevanta för allergivarningar och folkhälsa. I analysen definierar vi

$$K' = \{Alnus, Betula, Poaceae, Quercus, Artemisia, Corylus, Salix, Ulmus\}$$

som mängden arter som ingår i den hierarkiska klusteranalysen.

Pollensäsongens start och slut

Pollensäsongens gränser definieras som de dagar då 10% respektive 90% av den totala årskoncentrationen har uppmätts [Oteros et al., 2019]. Den tyska studien

använde medianen av start- och slutdag för varje art j ; vi breddar istället intervallet för att bättre täcka Sveriges geografiska variation.

För varje $j \in K'$ och station i definiera den kumulativa funktionen

$$C_{ijl} = \sum_{u=1}^l \bar{x}_{iju}, \quad l = 1, \dots, L,$$

så att $C_{ijL} = T_{ij}$. Sätt

$$s_{ij} = \min\{l : C_{ijl} \geq 0.1 T_{ij}\}, \quad e_{ij} = \min\{l : C_{ijl} \geq 0.9 T_{ij}\}.$$

Dessa dagar markerar när 10% respektive 90% av årsbelastningen uppnåts.

För att få ett gemensamt intervall för Sverige samlar vi $\{s_{1j}, \dots, s_{Nj}\}$ respektive $\{e_{1j}, \dots, e_{Nj}\}$, beräknar första kvartilen $Q_{1,s}$ för startdagarna och tredje kvartilen $Q_{3,e}$ för slutdagarna och sätter

$$S_j = Q_{1,s}, \quad E_j = Q_{3,e}.$$

Tabell 3.2: Kvartiler för start- och slutdag på året, per art

Art	S_j	Median		E_j
	Startdag	Startdag	Slutdag	Slutdag
<i>Alnus</i>	55.25	58.50	106.00	111.00
<i>Artemisia</i>	190.75	198.50	256.50	263.75
<i>Betula</i>	111.00	112.50	136.00	141.75
<i>Corylus</i>	44.00	57.00	118.00	130.50
<i>Poaceae</i>	156.00	159.50	214.00	218.25
<i>Quercus</i>	127.00	129.50	153.50	161.50
<i>Salix</i>	89.00	96.00	148.00	153.50
<i>Ulmus</i>	79.25	89.00	142.50	187.75

Slutlig beskärning

Beskär varje profil $\bar{x}_{ijl}$ till intervallet $l = S_j, \dots, E_j$. För varje station i och art $j \in K'$ erhålls därmed tidsserien

$$\{\bar{x}_{ijl}\}_{l=S_j}^{E_j},$$

vilken ligger till grund för avstånds- och likhetsberäkningar före klustring, silhuettanalys och MDS.

3.4.3 Kod för metoder

Analysen genomfördes i R [R Core Team, 2023]. För klustringen användes paketet `cluster` [Maechler et al., 2023]. Nedan visas ett exempel på hur hierarkisk klustring kan utföras på en avståndsmatris:

```
hc_corr <- hclust(dist_corr, method = "ward.D2")
```

Multidimensionell skalning (MDS) genomförs också i R, med funktionen `cmdscale`. Exempel på kod för en avståndsmatris visas här:

```
mds_corr <- cmdscale(dist_corr, k = 3)
```

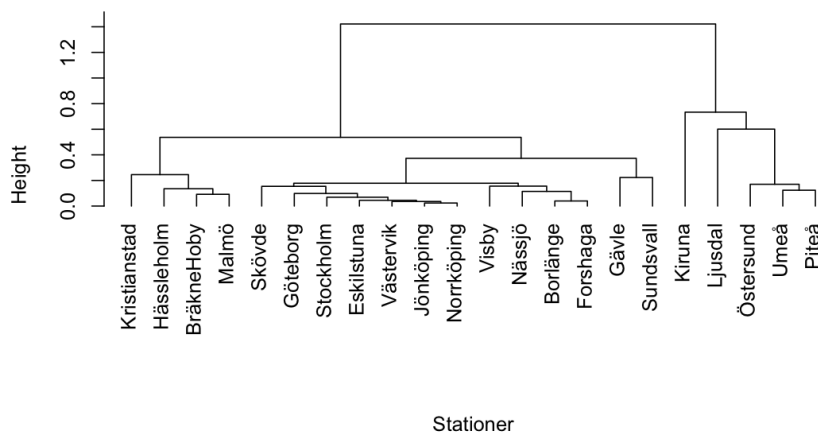
För silhuettanalys användes åter paketet `cluster`. Så här kan siluettbredden beräknas för k grupper:

```
cluster_labels <- cutree(hc_corr, k = k)
sil             <- silhouette(cluster_labels, dist_corr)
```

4 Resultat

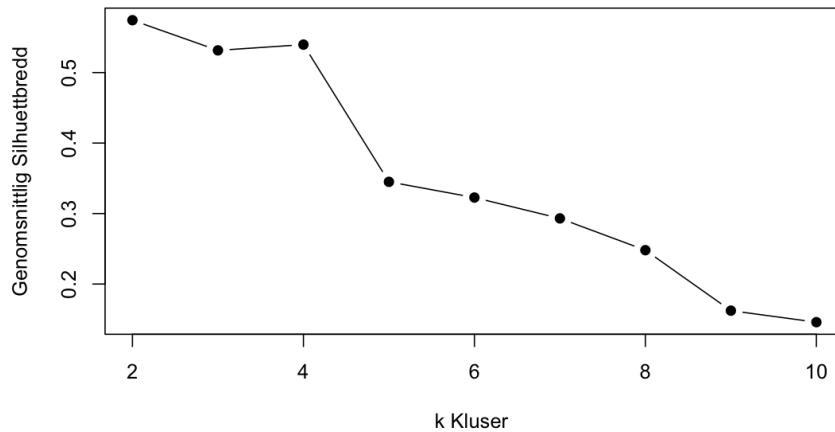
Nedan presenteras resultaten från hierarkisk klustring med två olika avståndsmått (Pearson- och cosinusavstånd). Först visas ett dendrogram samt en silhuettanalys som bestämmer det optimala antalet kluster för respektive mått. Därefter följer en MDS-plot i två dimensioner, indelad i k kluster (där k har valts utifrån silhuettanalysen), för att tydligare illustrera klustrens inbördes relationer. Slutligen presenteras en tabell med de fem närmaste stationsparen för vart och ett av avståndsmåtten, samt en karta över Sverige där stationerna är färgkodade efter klustertillhörighet, från två upp till fem kluster.

4.1 Cosinus-avstånd



Figur 4.1: Dendrogram för hierarkisk klustring med Wards metod och cosinusavstånd

I figur 4.1 ser vi hur stationerna grupperas. Dock är det svårt att avgöra hur många kluster som bör skapas, därför analyserar vi den genomsnittliga silhuettbredden för varje antal kluster k .

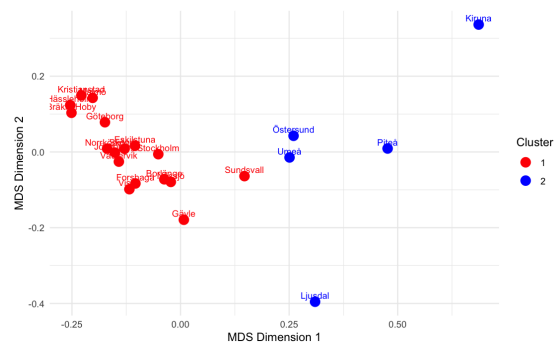


Figur 4.2: Silhuettbredd för klusteranalys med cosinusavstånd

Tabell 4.1: Genomsnittlig silhuettbredd för olika k

k	Genomsnittlig Silhuettbredd
2	0.57
3	0.53
4	0.54
5	0.34
6	0.32
7	0.29
8	0.25
9	0.16
10	0.15

I Figur 4.2 ser vi att två kluster ger den högsta genomsnittliga silhuettbredden, vilket enligt Tabell 4.1 motsvarar 0,57. Därför delar vi in stationerna i två grupper när vi presenterar MDS-plottarna.



Figur 4.3: MDS-plot baserad på cosinusavstånd, $k=2$ kluster

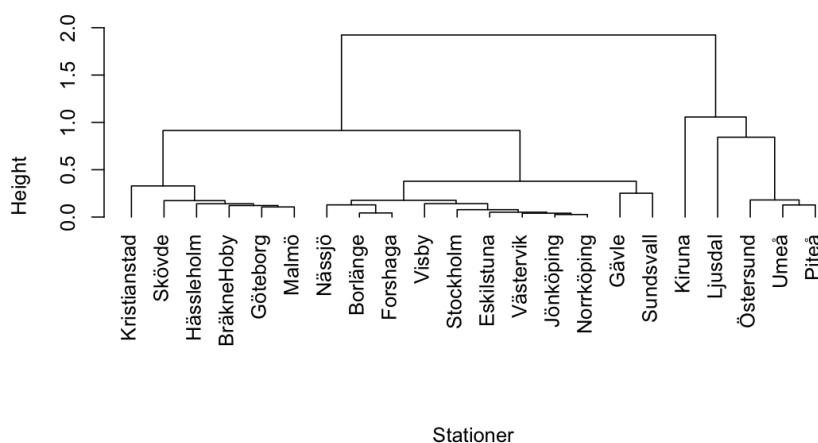
I Figur 4.3 visas hur stationerna förhåller sig till varandra när de delas in i två kluster ($k = 2$).

Tabell 4.2: De 5 närmaste stationparen enligt Cosinusavstånd

Station1	Station2	Avstånd
Jönköping	Norrköping	0.024
Eskilstuna	Norrköping	0.031
Norrköping	Västervik	0.032
Jönköping	Västervik	0.034
Borlänge	Forshaga	0.040

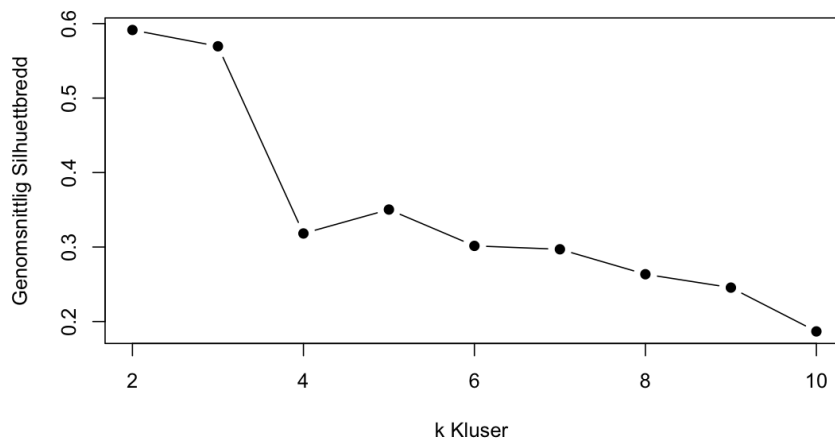
I Tabell 4.2 visas de fem närmaste stationsparen mätt i cosinusavstånd.

4.2 Pearson-avstånd



Figur 4.4: Dendrogram för hierarkisk klustring med Wards metod och Pearson-avstånd

I Figur 4.4 visas hur stationerna grupperas. Dock är det återigen svårt att avgöra hur många kluster som bör skapas, därför analyserar vi den genomsnittliga silhuettbredden för varje värde på k .

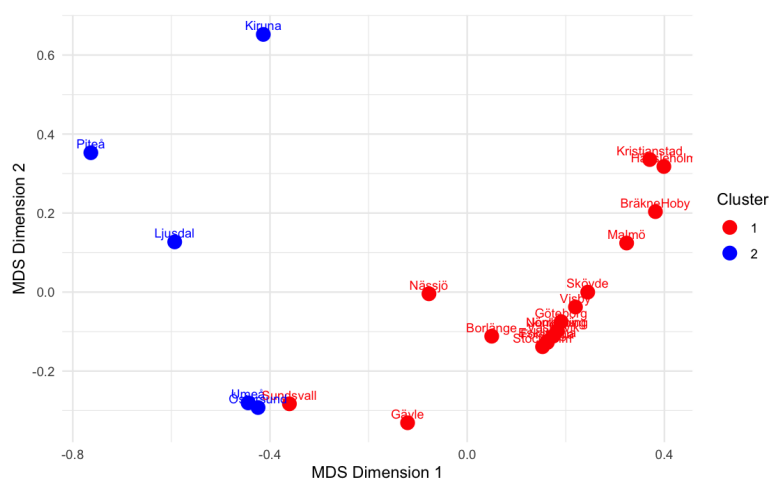


Figur 4.5: Silhuettbredd för klusteranalys med Pearson-avstånd

Tabell 4.3: Genomsnittlig silhuettbredd för olika k

k	Genomsnittlig Silhuettbredd
2	0.59
3	0.57
4	0.32
5	0.35
6	0.30
7	0.30
8	0.26
9	0.25
10	0.19

I Figur 4.5 visas att två kluster ger den högsta genomsnittliga silhuettbredden, vilket enligt Tabell 4.3 motsvarar 0,59. Därför delar vi in stationerna i två grupper när vi presenterar MDS-plotten.



Figur 4.6: MDSplot baserad på pearsson korrelation, $k=2$ kluster

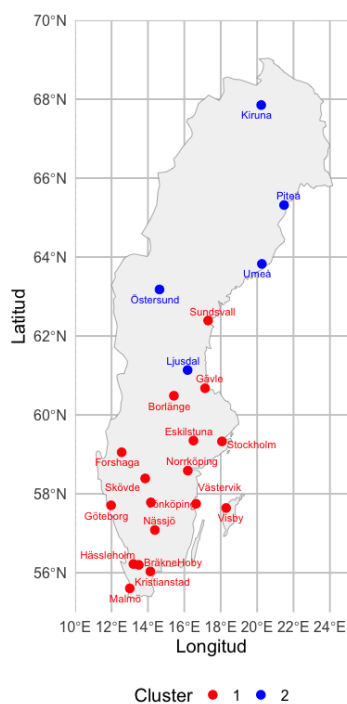
I Figur 4.6 visas hur stationerna förhåller sig till varandra när de delas in i två kluster ($k = 2$).

Tabell 4.4: De 5 närmaste stationparen enligt Pearsons- avstånd

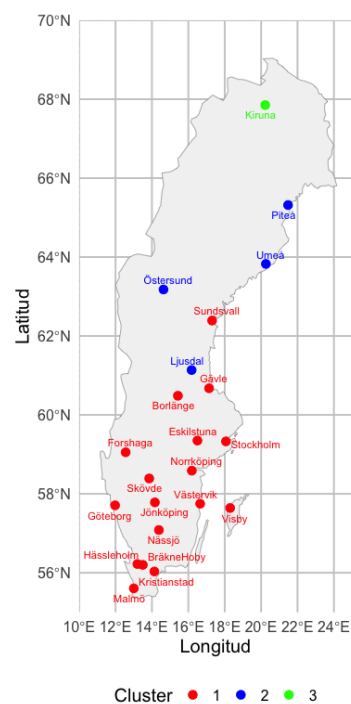
Station 1	Station 2	Avstånd
Jönköping	Norrköping	0.027
Norrköping	Västervik	0.035
Eskilstuna	Norrköping	0.036
Jönköping	Västervik	0.040
Borlänge	Forshaga	0.043

I Tabell 4.4 ser vi de 5 närmsta parvisa stationer i Pearson- avstånd.

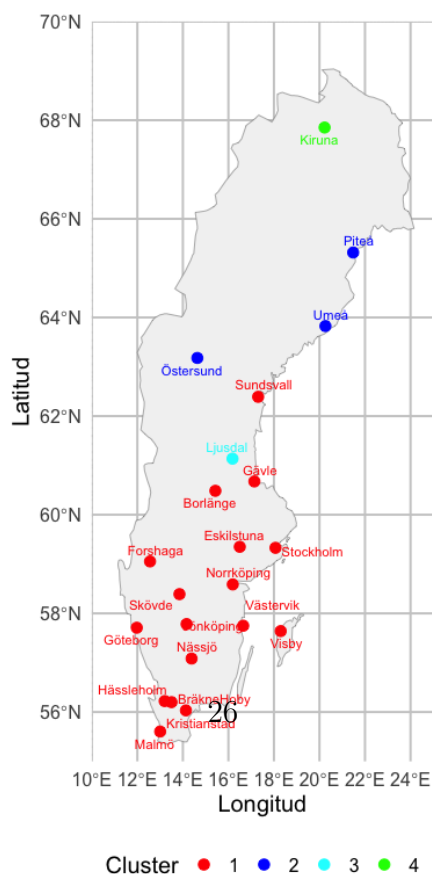
4.3 Karta Över stationer



(a) Karta över stationer enligt $k = 2$ kluster

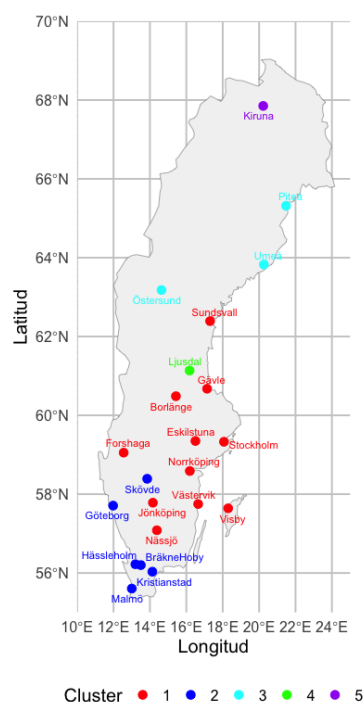
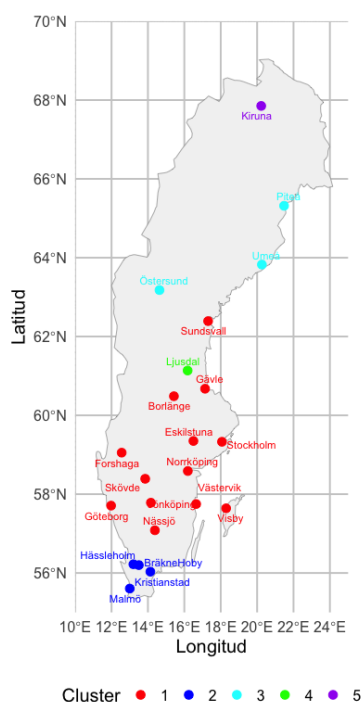


(b) Karta över stationer enligt $k = 3$ kluster



(c) Karta över stationer enligt $k = 4$ kluster

Figur 4.7: Kartor över stationer med kluster från $k = 2$ till $k = 4$.



(a) Karta över stationer enligt $k = 5$ kluster (Cosinusavstånd)

(b) Karta över stationer enligt $k = 5$ kluster (Pearson avstånd)

Figur 4.8: Kartor över stationer enligt $k = 5$ kluster med olika avståndsmått.

I Figur 4.7 visas stationernas geografiska positioner färgkodade efter klustertillhörighet, från $k = 2$ till $k = 4$ kluster. Kartorna är giltiga för både Pearson- och cosinusavstånd, eftersom båda metoderna resulterade i samma indelning av stationerna. I Figur 4.8 visas stationernas geografiska positioner färgkodade efter klustertillhörighet för $k = 5$ kluster. Till skillnad från tidigare nivåer gav cosinus- och Pearson-avstånd olika klusterindelningar, vilket illustreras i de två kartorna.

5 Diskussion

5.1 Jämförande analys av avståndsmått

5.1.1 Cosinusavstånd

Tolkning av dendrogram

I Figur 4.1 visas hur stationerna grupperas ungefär efter sin geografiska plats. Det är väntat, eftersom stationer som ligger nära varandra ofta har liknande pollen-säsonger. Det går dock inte att direkt avläsa hur många kluster dendrogrammet antyder.

Silhuettbredder

Enligt Tabell 4.2 ger $k = 2$ den högsta genomsnittliga silhuettbredden, 0,57. Enligt Tabell 3.1 innebär en bredd mellan 0,51 och 0,70 att klusterstrukturen är rimlig. Med cosinusavstånd verkar det därför bäst att dela Sveriges stationer i två grupper.

MDS-visualisering

Klustring med $k = 2$ skiljer en nordlig zon från en zon som täcker det mellersta och södra Sverige. I Figur 4.3 framgår att punkterna i kluster 2 ligger mer utspridda, medan punkterna i kluster 1 ligger tätare. Inom kluster 1 finns ändå undergrupper: stationerna i Skåne respektive de i Mellansverige ligger nära varandra.

Närmaste stationspar

Tabell 4.2 visar de fem stationspar som har lägst cosinusavstånd. Den minsta distansen finns mellan Norrköping och Jönköping (0,024), följt av Norrköping–Eskilstuna (0,031) och Norrköping–Västervik (0,032). Norrköping fungerar alltså som en central nod i förhållande till dessa stationer. Även Jönköping–Västervik har ett lågt avstånd (0,034). På femte plats kommer Borlänge–Forshaga (0,040),

vilket tyder på överlappning även i Mellansverige. Tabell A.1 listar de tjugo närmaste stationsparen.

5.1.2 Pearson-avstånd

Tolkning av dendrogram

Även Figur 4.4 visar grupperingar som speglar stationernas geografiska läge, vilket tyder på att cosinus- och Pearson-avstånd ger liknande resultat.

Silhuettbredder

Tabell 4.4 visar att $k = 2$ ger den högsta genomsnittliga silhuettbredden, 0,59. Enligt Tabell 3.1 representerar detta en rimlig klusterstruktur. Med Pearson-avstånd är alltså två kluster det bästa valet.

MDS-visualisering

Klustring med $k = 2$ delar Sverige i en nordlig zon och en zon som omfattar Mellansverige och södra Sverige. I Figur 4.6 är punkterna i kluster 2 mer utspridda än i kluster 1. Inom kluster 1 återfinns stationerna i Skåne respektive Mellansverige i egna, tätare grupper. Resultatet liknar det som erhöles med cosinusavstånd.

Närmaste stationspar

Tabell 4.4 visar att Norrköping–Jönköping är närmast (0,027), följt av Norrköping–Västervik (0,035) och Norrköping–Eskilstuna (0,036). Detta bekräftar Norrköpings centrala roll. Jönköping–Västervik har också ett lågt avstånd (0,040), och Borlänge–Forshaga (0,043) visar viss överlappning i Mellansverige. De tjugo närmaste stationsparen finns i Tabell A.2.

5.2 Geografisk tolkning av kluster och placering av nya mätstationer

Om vi har k pollenfällor att placera ut, kan resultaten i Figur 4.7 och Figur 4.8 ge vägledning om var dessa bör placeras för att bäst täcka in Sveriges olika pollenregioner. Genom att betrakta kartorna ser vi att stationerna tenderar att gruppera sig naturligt i geografiskt sammanhängande kluster.

Vid $k = 5$ kluster, som visas i Figur 4.8, framträder flera tydliga grupperingar. Ett kluster bildas i södra Sverige, där flera stationer i Skåne ingår. Ett annat kluster

täcker delar av mellersta Sverige, medan stationen i Ljusdal avviker från detta mönster och verkar utgöra en outlier. I norra Sverige identifieras tre stationer som tillhör samma kluster, medan Kiruna bildar ett eget kluster.

Både Pearson- och cosinusavstånd ger liknande klusterindelningar. Skillnaden är att Göteborg och Skövde tillhör kluster 2 enligt Pearson-avstånd, medan de enligt cosinusavstånd placeras i kluster 1. För övriga stationer överensstämmer kluster-tillhörigheten mellan de två avståndsmåtten helt.

5.3 Problematiska stationer

Vissa stationer, exempelvis Ljusdal och Kiruna, har bara ett års mätningar. Det är rimligt att utesluta dessa, eftersom deras genomsnittliga profiler bygger på alltför korta tidsserier och därmed kan försämma klusterindelningen. Att vi bara får $k = 2$ kluster som bäst resultat från silhuettanalysen är dessutom inte särskilt hjälpsamt för beslut om placering av automatiska fällor.

5.4 Problemet med en enhetlig start- och slutpunkt

När vi använder samma start- och slutdatum för hela landet uppstår problem. Exempelvis infaller den nationella startdagen för *Alnus* (dag 55,25) mitt i säsongen i Visby (Figur A.1), medan *Betula* i Sundsvall avbryts mitt i säsongen (Figur A.2).

5.4.1 Alternativ: Zonindelad klustring

För att bättre bevara de lokala pollenmönstren skulle man kunna dela in mätstationerna i tre klimatzoner: norr, mellan och syd. Inom varje zon fastställer vi egna start- och slutdatum för pollensäsongen och beskär data därefter. Därefter körs hierarkisk klustring separat för varje zon, vilket ger tre individuella klusteruppställningar som speglar zonernas unika säsongskaraktär.

5.5 Viktningsstrategi baserat på population

Då vi optimerar i syfte att studera allergener kan ett förbättringsförslag vara att införa en viktningsstrategi baserad på befolkning. Man skulle kunna välja ut de stationer som täcker de största befolkningscentra, eftersom dessa är mest relevanta för allergipåverkan. Ett förslag på en alternativ strategi är att vikta stationerna baserat på befolkningsstorlek, så att mätningar från tätbefolkade områden får större genomslag i klusteranalysen.

5.6 Länkmetsod i hierarkisk klustring

Förutom Wards metoden som beskrivs i kapitel 3 finns många alternativ att välja mellan. Valet av länkmetsod påverkar direkt hur kluster bildas och därmed också resultaten. För att göra optimeringen mer robust kan man därför tillämpa flera olika länkmetsoder och sedan jämföra utfallet. Genom att analysera om metoderna ger likartade eller skiftande klusterindelningar och undersöka orsakerna till eventuella skillnader får man en djupare förståelse för hur länkmetsoden påverkar klusterstrukturen. Detta underlag gör det möjligt att motivera valet av länkmetsod på ett mer välgrundat sätt.

5.7 Utvärdering av Klusterkvalitet

För att bedöma klusterkvaliteten och bestämma hur många grupper stationerna ska delas in i använde vi silhuettmetoden. Den kan även kombineras med till exempel elbowmetoden och gapstatistiken [Borg and Groenen, 2005, s. 129] för att göra valet av optimalt antal klusterindelningar.

5.8 Klassisk MDS

Klassisk MDS optimerar ett globalt felmått (strain eller stress) för alla parvisa avstånd och kan därför offra precision i de kortaste avstånden. För att bättre identifiera närmaste grannar kan andra projektionstekniker behöva utvärderas.

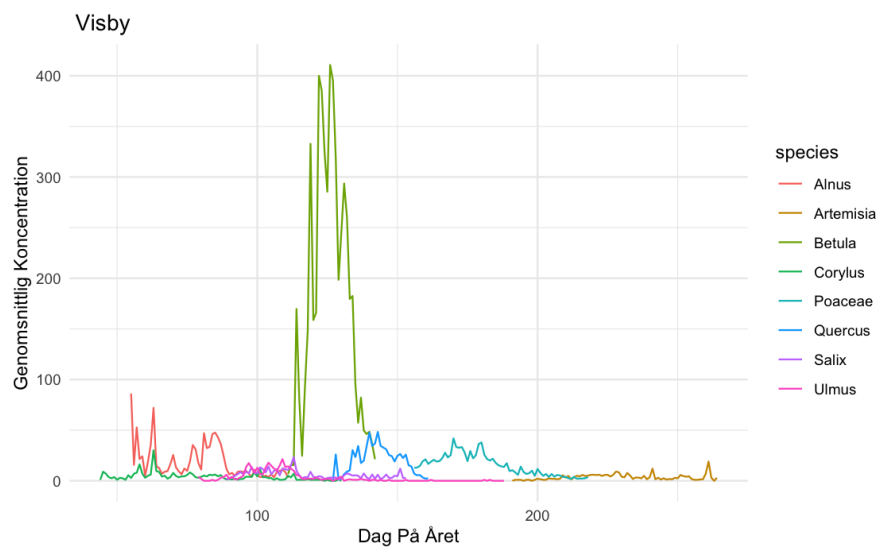
5.9 Rekommendationer

Resultaten visar att geografiskt närliggande stationer har kortast Pearson- och cosinusavstånd. Norrköping är en nyckelstation gentemot Jönköping, Västervik och Eskilstuna. Vi ser i Tabell A.1 och Tabell A.2 att Stockholm ligger också nära Norrköping ($d_{\cos} = 0,07$, $d_{\rho} = 0,078$), vilket gör en ny automatisk fälla där överflödig. Ett möjligt alternativ är att flytta den automatiska fällan från Stockholm till Norrköping och därmed förbättra täckningen utan extra kostnad.

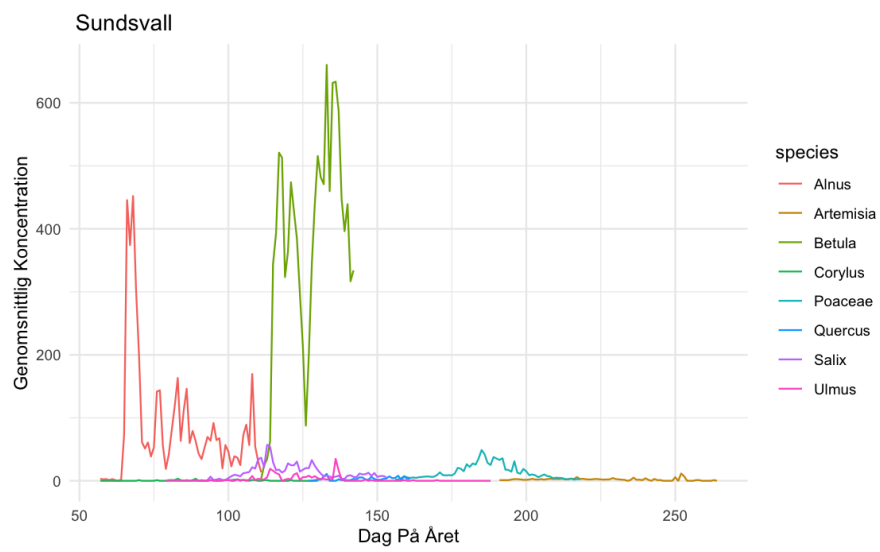
Stationerna i södra och mellersta Sverige liknar varandra. Det kan räcka med en gemensam station för södra Sverige och en för Mellansverige, till exempel i Stockholm där befolkningen är störst. Däremot behövs fler stationer i Norrland, eftersom avstånden där är större, vilket framgår av Figur 4.3 och Figur 4.6. Enligt Figur 4.7 och Figur 4.8 kan vi direkt se hur den hierarkiska klustringsmetoden delar in stationerna för olika val av k kluster. Om vi följer denna indelning vid

placering av nya pollenfällor, skulle det vara naturligt att placera en i Skåne, en i mellersta Sverige samt en eller två i norra Sverige. Det är dock osäkert hur tydligt stationerna i Ljusdal och Kiruna skiljer sig från övriga, då pollen endast har mätts där under ett år, vilket gör deras klustertillhörighet mer osäker.

A Bilaga



Figur A.1: Visbys fördelning av pollenkoncentration under definierad pollensäsong för Sverige



Figur A.2: Sundsvalls fördelning av pollenkoncentration under definierad pollen-säsong för Sverige

Tabell A.1: De 20 närmaste stationparen enligt Cosinusavstånd

Station1	Station2	Avstånd
Jönköping	Norrköping	0.02373150
Eskilstuna	Norrköping	0.03138825
Norrköping	Västervik	0.03183527
Jönköping	Västervik	0.03428305
Borlänge	Forshaga	0.03963074
Eskilstuna	Jönköping	0.04437364
Eskilstuna	Västervik	0.04474648
Göteborg	Norrköping	0.04520137
Stockholm	Västervik	0.04773083
Eskilstuna	Stockholm	0.04818216
Forshaga	Jönköping	0.05784008
Göteborg	Jönköping	0.05877913
Västervik	Visby	0.06454787
Jönköping	Stockholm	0.06509958
Forshaga	Norrköping	0.06755828
Eskilstuna	Göteborg	0.06997602
Norrköping	Stockholm	0.07076394
Forshaga	Västervik	0.07221567
Eskilstuna	Forshaga	0.08095424
Borlänge	Västervik	0.08971294
Borlänge	Jönköping	0.09127236
Jönköping	Skövde	0.09146753
BräkneHoby	Malmö	0.09181047
Forshaga	Nässjö	0.09388862
Norrköping	Visby	0.09448574
Jönköping	Visby	0.09674078
Jönköping	Nässjö	0.09681320
Forshaga	Stockholm	0.09915929
Eskilstuna	Nässjö	0.09935029
Göteborg	Malmö	0.10199333

Tabell A.2: De 20 närmaste stationparen enligt Pearsons- avstånd

Station1	Station2	Avstånd
Jönköping	Norrköping	0.02684929
Norrköping	Västervik	0.03549291
Eskilstuna	Norrköping	0.03573488
Jönköping	Västervik	0.03979043
Borlänge	Forshaga	0.04315131
Göteborg	Norrköping	0.05022143
Eskilstuna	Jönköping	0.05093742
Eskilstuna	Västervik	0.05110788
Eskilstuna	Stockholm	0.05275938
Stockholm	Västervik	0.05624781
Stockholm	Forshaga	0.06239303
Göteborg	Jönköping	0.06482761
Västervik	Visby	0.07209935
Forshaga	Norrköping	0.07434572
Jönköping	Stockholm	0.07484315
Forshaga	Västervik	0.07717520
Eskilstuna	Göteborg	0.07794626
Norrköping	Stockholm	0.07828447
Eskilstuna	Forshaga	0.08910264
Borlänge	Västervik	0.10244361
Forshaga	Nässjö	0.10294632
Forshaga	Stockholm	0.10371419
Borlänge	Jönköping	0.10427527
Jönköping	Skövde	0.10610778
Norrköping	Visby	0.10668349
Göteborg	Malmö	0.10672395
Jönköping	Visby	0.10952249
BräkneHoby	Malmö	0.11107439
Jönköping	Nässjö	0.11233497
Göteborg	Västervik	0.11383673

Litteraturförteckning

- Astma- och Allergiförbundet. Pollenallergi. <https://astmaoallergiforbundet.se/information-rad/allergi/pollenallergi/>, 2025. Hämtad 5 maj 2025.
- Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, NY, 2nd edition, 2005. ISBN 0-387-25150-2, 978-0387-25150-9. Printed on acid-free paper.
- Burkard Scientific. 7-day recording volumetric spore trap. <https://burkard.co.uk/product/7-day-recording-volumetric-spore-trap/>, 2025. Accessed: 2025-05-05.
- B. Everitt, S. Landau, M. Leese, D. Stahl, and an O'Reilly Media Company Safari. *Cluster Analysis, 5th Edition*. John Wiley & Sons, 2011. URL <https://books.google.se/books?id=3RFPzQEACAAJ>.
- Karolinska Institutet. Pollen & luftföroreningar – hälsorisker och råd. <https://www.imm.ki.se/Datavard/Rapporter/pollen%20luftfororeningar%20broschyr%202014-08.pdf>, 2014. Hämtad 5 maj 2025.
- Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. 01 1990. ISBN 0-471-87876-6. doi: 10.2307/2532178.
- M. Maechler, P.J. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2023. URL <https://CRAN.R-project.org/package=cluster>. R package version 2.1.4.
- Sadaaki Miyamoto. *Theory of Agglomerative Hierarchical Clustering*, volume 15 of *Behaviormetrics: Quantitative Approaches to Human Behavior*. Springer Singapore, Singapore, 2022. ISBN 978-981-19-0419-6. doi: 10.1007/978-981-19-0420-2.
- Naturhistoriska riksmuseet. Allergiframkallande pollen. <https://pollenrapporten.se/ompollen/allergiframkallandepollen.4.314e02dd13d69872ec083.html>, 2025. Hämtad 5 maj 2025.

Jose Oteros, Mikhail Sofiev, Matt Smith, Bernard Clot, Athanasios Damialis, Marje Prank, Matthias Werchan, Reinhard Wachter, Alisa Weber, Susanne Kutzora, Stefanie Heinze, Caroline E.W. Herr, Annette Menzel, Karl-Christian Bergmann, Claudia Traidl-Hoffmann, Carsten B. Schmidt-Weber, and Jeroen T.M. Buters. Building an automatic pollen monitoring network (epin): Selection of optimal sites by clustering pollen stations. *Science of The Total Environment*, 688:1263–1274, 2019. ISSN 0048-9697. doi: <https://doi.org/10.1016/j.scitotenv.2019.06.131>. URL <https://www.sciencedirect.com/science/article/pii/S0048969719327020>.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL <https://www.R-project.org/>.

Statistiska centralbyrån (SCB). Sveriges befolkning. <https://www.scb.se/hitta-statistik/sverige-i-siffror/manniskorna-i-sverige/sveriges-befolkning/>, 2025. Senast uppdaterad 21 februari 2025; hämtad 5 maj 2025.