Stockholms universitet

Variation i populationsskattningar från sälinventeringar: Effekten av störningar, rörelse och val av estimator

Shiba Afshar

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2026:6
Matematisk statistik
Maj 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2026:6**
http://www.math.su.se/matstat

Variation i populationsskattningar från sälinventeringar: Effekten av störningar, rörelse och val av estimator

Shiba Afshar*

Maj 2026

Sammanfattning

Syftet med denna studie är att undersöka hur variation i populationsskattningar från sälinventeringar påverkas av observationsprocessen och hur olika estimatorer fungerar under varierande förhållanden. För att analysera detta använder vi en simuleringsbaserad modell där population, rörelse mellan områden och observationer beskrivs separat.

Resultaten visar att störningar i observationsprocessen är den viktigaste orsaken till variation i skattningarna och att även små störningar kan leda till tydligt ökade osäkerheter. Rörelse mellan områden har däremot mindre betydelse för estimatorer som bygger på totalsummor. När vi jämför olika estimatorer framstår den nuvarande metoden som ett relativt stabilt alternativ, särskilt när observationerna påverkas av störningar.

Jämförelser med verkliga data visar också att variationen i praktiken är högre än i idealiserade scenarier utan störningar. Studien bidrar därmed till en bättre förståelse för hur osäkerhet uppstår i sälinventeringar och hur valet av estimator kan påverka populationsskattningar i praktisk övervakning.

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

Förord

Jag vill rikta ett stort tack till min handledare, Martin Sköld, för värdefull vägledning och stöd under arbetets gång. Hans synpunkter och förslag har varit betydelsefulla för att utveckla både modelleringen och den statistiska analysen i projektet.

Jag vill även nämna användningen av ChatGPT som ett hjälpmedel under arbetets gång. Verktyget har använts som stöd vid programmering i R samt för att strukturera och språkligt förbättra texten.

Innehåll

1 Introduktion

Att uppskatta populationsstorlek är en viktig uppgift inom ekologi och naturvårdsbiologi. För marina däggdjur, som knubbsäl (*Phoca vitulina*), är detta dock särskilt svårt eftersom djuren rör sig mellan land och vatten och därför inte alltid kan observeras.

I praktiken baseras populationsuppskattningar ofta på räkningar av sälar som ligger uppe på land vid så kallade haul-out platser. Dessa räkningar underskattar dock den verkliga populationen, eftersom en del av djuren befinner sig i vattnet vid observationstillfället och därmed inte räknas [1]. För att få mer tillförlitliga uppskattningar behöver vi därför ta hänsyn till denna osäkerhet i observationsprocessen.

Antalet observerade sälar påverkas dessutom av flera faktorer, till exempel tid på dygnet, årstid, väder och djurens beteende. Tidigare studier visar att variationen i observationerna ofta är stor både inom och mellan år [2]. Det innebär att förändringar i data inte nödvändigtvis speglar verkliga förändringar i populationen, utan i stället kan bero på hur och när observationerna görs.

För att hantera detta använder vi statistiska modeller som skiljer mellan den sanna populationen och de observerade data. I sådana modeller kan vi inkludera faktorer som rörelse mellan områden, rumslig fördelning och sannolikheten att en individ observeras. På så sätt kan vi både förbättra skattningarna och få en bättre förståelse för osäkerheten i dem.

I denna studie utvecklar vi en simuleringsbaserad modell för att undersöka hur olika faktorer påverkar variationen i populationsuppskattningar av sälar. Vi fokuserar särskilt på observationsprocessen och hur störningar, till exempel en ökad sannolikhet att missa individer, påverkar olika estimatorer.

Fokus ligger på variation snarare än bias. Eftersom den sanna populationen inte kan observeras direkt i verkligheten är det svårt att utvärdera systematiska fel. Däremot kan variation analyseras både i simuleringar och i verkliga data, vilket gör den till ett praktiskt mått på estimatorernas egenskaper.

Syftet med studien är därför att analysera hur variation uppstår i sälinventeringar och att jämföra hur robusta olika estimatorer är under varierande observationsförhållanden.

2 Teoretisk bakgrund

För att beskriva både den sanna populationen och hur observationerna uppstår använder vi sannolikhetsmodeller. I detta avsnitt introducerar vi de

statistiska begrepp som ligger till grund för vår modell och förklarar hur vi tänker kring data och osäkerhet.

2.1 Poisson- och binomialmodellen

Vi börjar med en förenklad modell där vi tydligt skiljer mellan den sanna populationen och det vi faktiskt observerar, och antar att den sanna populationen kan beskrivas som $N \sim \text{Poisson}(\lambda)$, där λ är det genomsnittliga antalet individer.

Vid varje observationstillfälle antar vi att varje individ har en sannolikhet p att bli upptäckt. Detta leder till modellen $n \mid N \sim \text{Binomial}(N, p)$, där n är det antal individer vi observerar och p är detektionssannolikheten.

Detta innebär att varje individ observeras oberoende av de andra med sannolikhet p. Givet N får vi därför $E(n \mid N) = Np$ och $\text{Var}(n \mid N) = Np(1-p)$. Om vi dessutom antar att N följer en Poissonfördelning kan vi visa att även de observerade värdena blir $n \sim \text{Poisson}(\lambda p)$, vilket ger $E(n) = \lambda p$ och $\text{Var}(n) = \lambda p$.

Den här modellen beskriver en situation där vi bara observerar en del av den verkliga populationen. I praktiken innebär det att våra observationer systematiskt underskattar det sanna antalet individer. Detta är vanligt i ekologiska studier. Vid sälinventeringar räknar vi exempelvis bara de sälar som befinner sig uppe på land, medan andra kan befinna sig i vattnet vid samma tidpunkt. Observationerna påverkas dessutom av faktorer som tid på dygnet, väder och störningar, vilket gör att antalet observerade individer kan variera mellan olika tillfällen [4].

Vi kan därför se modellen som två steg: först beskriver vi den sanna men okända populationen (N), och därefter observationerna (n) givet denna. Den sanna populationen är alltså inte direkt observerbar utan fungerar som en latent variabel. Denna typ av uppdelning kallas en hierarkisk modell och används ofta inom ekologisk statistik [5]. Detta innebär att variation i observerade data inte nödvändigtvis speglar verkliga förändringar i populationen, utan även kan uppstå genom variation i observationsprocessen. Tidigare studier visar också att detektionssannolikheten kan variera mellan olika observationstillfällen och påverkas av både miljöfaktorer och hur data samlas in [6].

I många verkliga tillämpningar räcker dock inte Poissonmodellen för att beskriva variationen i data. Ofta är variansen större än vad modellen tillåter, vilket kallas överdispersion. För att hantera detta används mer flexibla modeller, till exempel negativ binomialfördelning [8]. I vissa situationer förekommer dessutom fler nollobservationer än vad standardmodeller kan förklara. Detta kan bero på att observationer helt uteblir, exempelvis på grund av

störningar i observationsprocessen och för att modellera detta används så kallade zero-inflated modeller.

2.2 Zero-inflated modeller

I många tillämpningar, särskilt inom ekologiska data, observeras fler nollor än vad standardmodeller kan förklara. Detta kan bero på att nollor uppstår från olika processer, till exempel att inga individer observeras trots att de finns, eller att observationer helt uteblir på grund av störningar.

För att hantera detta används zero-inflated modeller, där observationer antas genereras från en blandning av två processer. Den första processen beskriver situationer där observationen misslyckas helt och ger upphov till en nolla. Den andra processen beskriver den vanliga observationsmekanismen.

I fallet med binomial data innebär detta att en observation kan vara noll antingen på grund av låg detektionssannolikhet, eller på grund av att observationen störs och därför inte registreras alls.

I denna studie använder vi en zero-inflated binomial modell där en extra parameter q introduceras för att modellera störningar i observationsprocessen. Modellen gör det möjligt att skilja mellan nollor som beror på att individer inte observeras och nollor som uppstår på grund av störningar, vilket ger en mer realistisk beskrivning av data [10].

2.3 Lagen om total varians

För att analysera variationen i observationerna använder vi lagen om total varians,$\mathrm{Var}(n) = E\big(\mathrm{Var}(n \mid N)\big) + \mathrm{Var}\big(E(n \mid N)\big)$ se till exempel [3]. Genom att sätta in uttrycken för väntevärde och varians får vi, $\mathrm{Var}(n) = E(Np(1-p)) + \mathrm{Var}(Np)$.

Detta uttryck visar att den totala variationen i observationerna består av två delar. Den första delen kommer från själva observationsprocessen, det vill säga osäkerheten i om en individ observeras eller inte. Den andra delen beror på variation i den underliggande populationen.

Detta ger en förklaring till varför variationen i data ofta är större än vad enklare modeller, såsom Poissonmodellen, kan fånga.

I den fortsatta modellen antar vi att detektionssannolikheten p är konstant, medan ytterligare variation introduceras genom störningsparametern q i den zero-inflated observationsmodellen.

2.4 Multinomialfördelning

För att modellera hur individer fördelas mellan olika områden använder vi multinomialfördelningen.

Vi antar att ett antal N individer fördelas mellan k kategorier (eller polygoner) med sannolikheter $p_1, \ldots, p_k$. Detta skrivs som $(X_1, \ldots, X_k) \sim \text{Multinomial}(N, p_1, \ldots, p_k)$, där $\sum_{j=1}^{k} X_j = N$. Detta innebär att varje individ tilldelas en kategori med sannolikhet p_j, och att det totala antalet individer bevaras.

Multinomialfördelningen används också i ekologiska modeller för att beskriva hur individer förflyttar sig mellan olika geografiska områden över tid [7]. I denna studie använder vi multinomialfördelningen för att modellera hur populationen omfördelas mellan olika områden mellan inventeringstillfällen.

2.5 Modeller för förflyttning

Vid inventeringar av sälar kan antalet individer i ett område förändras mellan olika dagar eftersom sälar rör sig mellan närliggande områden. Även om den totala populationen är oförändrad kan sådana förflyttningar påverka observationerna och bidra till variation i populationsskattningarna.

För att beskriva detta använder vi en förenklad rörelsemodell där individer antingen stannar kvar i samma polygon eller förflyttar sig till en angränsande polygon mellan två inventeringsdagar. Modellen används för att undersöka hur förflyttning mellan områden kan påverka de observerade populationerna och de estimatorer som används i studien.

2.6 Estimatorer

En estimator är en funktion av observerade data som används för att uppskatta en okänd parameter. I denna studie använder vi flera olika estimatorer för att uppskatta populationsstorleken baserat på observerade data.

Eftersom olika estimatorer utnyttjar informationen i data på olika sätt kan de också ha olika egenskaper. I synnerhet kan de skilja sig åt när det gäller variation och robusthet, vilket gör det viktigt att jämföra deras prestanda under olika förhållanden.

2.7 Mått på variation

För att jämföra variation mellan olika estimatorer använder vi variationskoefficienten (CV), definierad som $\text{SD}(\hat{N})$ genom $E[\hat{N}]$. Den mäter relativ

variation och är särskilt användbar när estimatorer har olika medelvärden. På så sätt kan vi göra rättvisa jämförelser även när estimatorerna inte skattar exakt samma storhet.

Utöver detta använder vi även den log-transformerade standardavvikelsen, $\text{logSD} = \text{SD}(\log \hat{N})$, vilket är ett mått som är särskilt relevant vid analys av relativa förändringar, till exempel vid studier av populationstrender.

I den empiriska analysen använder vi båda dessa mått för att studera variationen i skattningarna. I praktiken visar de mycket likartade mönster, och därför redovisar vi i vissa fall endast ett av måtten för att undvika onödig upprepning.

Dessa mått är centrala i studien eftersom de gör det möjligt att analysera hur variation uppstår i modellen, och särskilt hur förändringar i observationsprocessen, till exempel genom störningsparametern q, påverkar variationen i skattningarna.

2.8 Trendanalys och statistisk styrka

För att analysera förändringar i populationsstorlek över tid använder vi linjära regressionsmodeller på log-transformerade skattningar. Vi utgår från modellen $\log(\hat{N}_t) = a + bt + \varepsilon_t$, där $\hat{N}_t$ är den skattade populationsstorleken vid tidpunkt t, b beskriver förändringstakten (trenden), och ε_t är en slumpterm med varians σ^2.

Log-transformeringen gör det möjligt att tolka förändringar i relativa termer. En linjär trend på log-skalan motsvarar därmed en exponentiell förändring i den ursprungliga skalan.

Ett centralt mål i trendanalys är att avgöra om en verklig förändring i populationen kan detekteras. Den statistiska styrkan (power) definieras som sannolikheten att förkasta nollhypotesen $b = 0$ när en verklig trend finns.

Den statistiska styrkan påverkas av flera faktorer, framför allt trendens storlek, antalet observationer och variationen i skattningarna. En hög variation i data leder till större osäkerhet i skattningen av lutningen b, vilket gör det svårare att skilja en verklig trend från slumpmässiga variationer.

Detta innebär att även om en population förändras över tid kan hög variabilitet i observationerna göra att förändringen inte upptäcks. Detta samband är välkänt inom ekologisk statistik, där möjligheten att detektera trender beror på både variansen i data och längden på tidsserien [9].

I denna studie är detta särskilt relevant eftersom variationen i skattningarna påverkas av observationsprocessen, till exempel genom störningsparametern q. Genom att analysera hur variationen förändras kan vi därför också förstå

hur möjligheten att detektera trender påverkas.

3 Data

3.1 Beskrivning av data

De data som används i studien består av sälräkningar som har tillhandahållits av handledaren och som ursprungligen kommer från Naturhistoriska riksmuseet.

Studieområdet är uppdelat i flera polygoner längs den svenska västkusten (Figur 1). Indelningen bygger på en shapefile som har använts för att organisera observationerna

Data omfattar åren 2014–2023. För varje år finns tre inventeringstillfällen, vilket innebär att populationen observeras vid tre separata tidpunkter per år. Inventeringarna genomförs under sommaren, oftast i augusti, då en stor andel av sälarna befinner sig uppe på land och därmed är möjliga att observera.

Observationerna är organiserade per polygon och dag, där varje datapunkt anger hur många sälar som observerats i en given polygon vid ett visst tillfälle. Dessa data används både för att beskriva variationen i observationerna och som utgångspunkt för de simuleringar som genomförs i studien.

3.2 Implementation i R

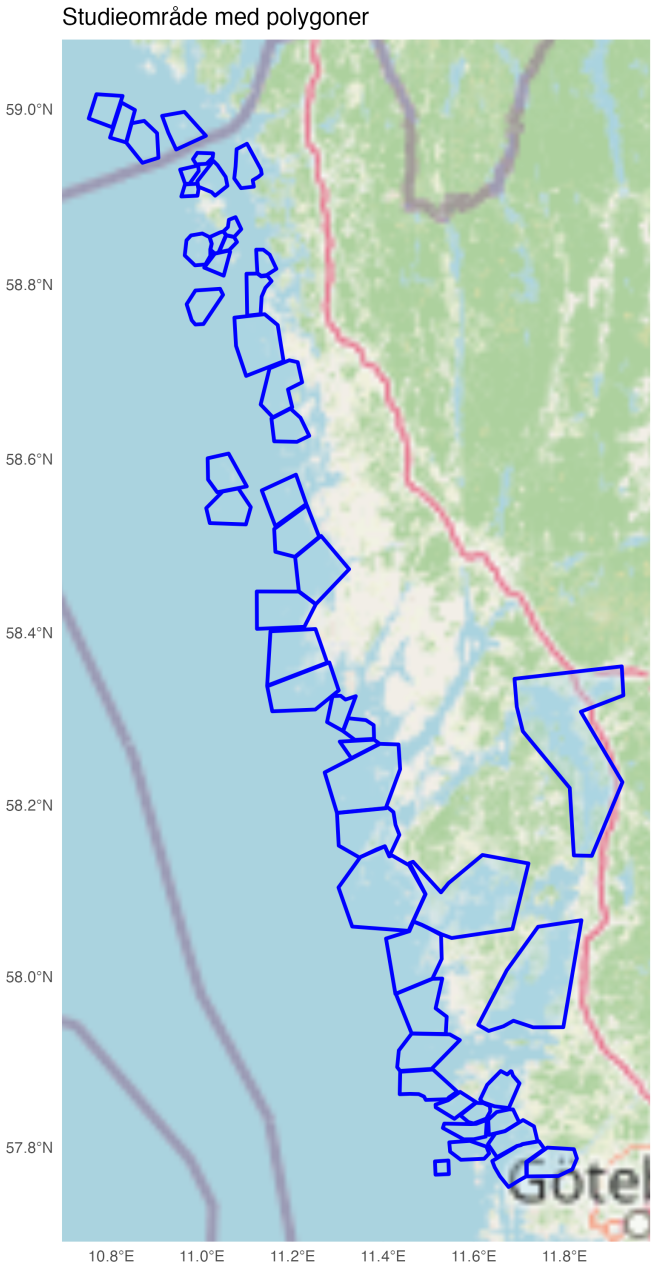
All datahantering, simulering och analys har genomförts i programmeringsspråket R [13].

Vi läser in data från en fil med polygonvisa observationer samt en grannmatris som anger vilka polygoner som är angränsande. Denna information används för att modellera rörelser mellan polygoner i simuleringarna.

För databehandling och summering använder vi funktioner från paketet `tidyverse`, vilket gör det möjligt att effektivt gruppera data och beräkna totalsummor samt olika estimatorer.

Simuleringen av observationsprocessen, inklusive den zero-inflated binomiala modellen, implementeras med hjälp av funktioner från paketet `VGAM` [11].

Själva simuleringsmodellen är implementerad genom egna funktioner i R. I varje simulering genererar vi först en population, därefter modellerar vi förflyttning mellan polygoner, genomför observationer och beräknar estimatorer. För varje parameterkombination upprepas simuleringen ett stort antal gånger, vilket gör det möjligt att stabilt uppskatta variationen i estimatorerna.



Figur 1: Studieområdet uppdelat i polygoner. Kartan har skapats i R baserat på shapefile från handledaren.

4 Modell

Vi modellerar sälinventeringar som ett stokastiskt system bestående av tre komponenter: den sanna populationen, förflyttning mellan polygoner samt observationsprocessen.

4.1 Motivering av modellval

Syftet med modellen är att förstå hur variation uppstår i populationsskattningar. Eftersom de observerade data påverkas av flera olika processer delar vi upp modellen i tre delar: population, förflyttning och observation.

Vi modellerar populationen stokastiskt för att fånga naturlig variation i antalet individer mellan polygoner. En Poissonfördelning används som en enkel och väletablerad modell för räknevariabler, vilket ger en rimlig utgångspunkt.

Vi inkluderar även förflyttning mellan polygoner, eftersom sälar kan röra sig mellan områden mellan inventeringsdagar. Detta modelleras med en övergångsmatris som beskriver sannolikheten att individer stannar kvar eller förflyttar sig till angränsande polygoner. På så sätt kan modellen fånga rumslig variation som annars skulle påverka observationerna.

Observationsprocessen modelleras separat eftersom det observerade antalet inte motsvarar den sanna populationen. Vi använder en binomialmodell för att beskriva ofullständig detektion, där varje individ observeras med sannolikhet p.

För att även ta hänsyn till störningar, som kan leda till att inga observationer görs trots att sälar finns närvarande, använder vi en zero-inflated komponent. Denna representeras av parametern q, som beskriver sannolikheten för att en observation helt uteblir. Detta gör det möjligt att skilja mellan låg detektion och faktiska störningar i datainsamlingen.

Slutligen inkluderar vi flera estimatorer för att analysera hur olika sätt att sammanfatta observationerna påverkar variationen i populationsskattningarna.

4.2 Notation

Vi låter N_{yit} beteckna det sanna (okända) antalet sälar i polygon i, dag t och år y, och n_{yit} det observerade antalet. Den totala populationen vid dag t skrivs som $N_{yt} = \sum_{i=1}^{I} N_{yit}$, medan det totala observerade antalet skrivs som $n_{yt} = \sum_{i=1}^{I} n_{yit}$.

Varje år omfattar tre inventeringsdagar, vilket motsvarar $t = 1, 2, 3$. Vi antar

att den totala populationen är konstant inom varje år, vilket innebär att variationen mellan dagar främst beror på förflyttning mellan polygoner och variation i observationsprocessen.

4.3 Populationsmodell

Vi modellerar populationen på den första inventeringsdagen som $N_{yi1} \sim \text{Poisson}(\lambda_i)$, där $\lambda_i = c \cdot \bar{n}_i$. Här är $\bar{n}_i$ det genomsnittliga observerade antalet i polygon i, och $c > 1$ en skalningsfaktor som tar hänsyn till ofullständig detektion. Detta innebär att den sanna populationen antas vara proportionell mot de observerade medelvärdena.

Poissonfördelningen används som en standardmodell för räknevariabler. I praktiken uppvisar sådana data ofta överdispersion, vilket innebär att variansen överstiger väntevärdet. För att hantera detta används mer flexibla modeller, till exempel negativ binomialfördelning eller quasi-Poisson-modeller [8]. I denna studie används dock Poissonmodellen som en enkel utgångspunkt, eftersom fokus ligger på hur variation introduceras genom rörelse och observationsprocessen.

4.4 Förflyttning mellan polygoner

Vi modellerar förflyttning mellan polygoner eftersom sälar kan röra sig mellan olika områden mellan inventeringsdagar. För att beskriva vilka polygoner som är kopplade till varandra använder vi en grannmatris W, där $W_{ij} = 1$ om polygon i och j är grannar och $W_{ij} = 0$ annars. Antalet grannar till polygon i betecknas med k_i.

Förflyttningen styrs av parametern $\delta \in [0, 1]$, där $1 - \delta$ är sannolikheten att en individ stannar kvar i samma polygon och δ är sannolikheten att individen förflyttar sig till någon av sina grannar. Övergångssannolikheterna skrivs som $P_{ii} = 1 - \delta$ och $P_{ij} = \frac{\delta}{k_i}$ för grannpolygoner.

Givet att det finns N_{yit} individer i polygon i vid dag t modellerar vi fördelningen mellan polygonerna vid dag $t + 1$ med en multinomialfördelning, där $N_{yij,t+1}$ anger antalet individer som förflyttar sig från polygon i till polygon j.

$$(N_{yi1,t+1}, \ldots, N_{yiI,t+1}) \sim \text{Multinomial}(N_{yit}, P_{i1}, \ldots, P_{iI})$$

Populationen i polygon j vid dag $t+1$ beräknas som $N_{y,j,t+1} = \sum_{i=1}^{I} N_{yij,t+1}$. Den totala populationen bevaras därmed mellan dagarna, medan fördelningen mellan polygoner förändras över tid. Detta innebär att antalet individer i varje polygon kan variera mellan olika inventeringstillfällen, vilket i sin tur bidrar till variation i observationerna.

4.5 Observationsmodell

Vid varje inventering observeras endast en del av den sanna populationen. Det observerade antalet sälar påverkas därför både av detektionssannolikheten och av eventuella störningar i observationsprocessen.

Vi modellerar observationerna med en zero-inflated binomialfördelning:

$$n_{yit} = \begin{cases} 0, & \text{med sannolikhet } q, \\ \text{Binomial}(N_{yit}, p), & \text{med sannolikhet } 1-q. \end{cases}$$

Här är p sannolikheten att en enskild individ observeras, medan q beskriver sannolikheten för att en observation helt uteblir på grund av störningar.

Den binomiala komponenten modellerar ofullständig detektion, där endast en andel av individerna observeras. Zero-inflationen introducerar en extra sannolikhet för nollobservationer, vilket gör det möjligt att särskilja mellan två typer av nollor: de som uppstår på grund av låg detektion och de som beror på störningar i observationsprocessen.

Modellen kan därmed tolkas som en blandning av två processer: en som genererar nollor genom störningar och en som beskriver observationer givet att datainsamlingen fungerar.

Denna uppdelning är central i studien, eftersom den gör det möjligt att analysera hur variation i observationsprocessen, särskilt genom parametern q, påverkar variationen i populationsskattningarna. Liknande modeller används i både ekologiska och epidemiologiska tillämpningar för att hantera data med många nollobservationer [10].

4.6 Estimatorer

För varje dag summerar vi observationerna över alla polygoner, vilket ger den dagliga totalsumman $n_{yt} = \sum_{i=1}^{I} n_{yit}$ för $t = 1, 2, 3$. Dessa totalsummor används sedan för att konstruera olika estimatorer av populationen. Syftet är att undersöka hur olika sätt att sammanfatta observationerna påverkar variationen i skattningarna.

Top2-estimatorn baseras på de två största dagliga observationerna och syftar till att minska påverkan från dagar med låga värden, exempelvis till följd av störningar. Vi sorterar därför de dagliga totalsummorna så att $n_{(1)} \geq n_{(2)} \geq n_{(3)}$ och definierar estimatorn som $\hat{n}_y = \frac{n_{(1)}+n_{(2)}}{2}$. Den korrigerade estimatorn ges då av $\hat{N}_y^{\text{Top2}} = \frac{\hat{n}_y}{p}$.

Mean3-estimatorn använder istället samtliga observationer och definieras som $\hat{N}_y^{\text{Mean3}} = \frac{1}{3p}(n_{y1}+n_{y2}+n_{y3})$, medan **Max-estimatorn** baseras på den

största dagliga observationen och ges av $\hat{N}_y^{\text{Max}} = \frac{1}{p} \max(n_{y1}, n_{y2}, n_{y3})$.

Slutligen tar **SumMax-estimatorn** hänsyn till variation mellan polygoner genom att först välja det maximala värdet per polygon och därefter summera dessa värden. Estimatorn skrivs som $\hat{N}_y^{\text{SumMax}} = \frac{1}{p} \sum_{i=1}^{I} \max(n_{yi1}, n_{yi2}, n_{yi3})$.

De olika estimatorerna använder informationen i observationerna på olika sätt. Detta innebär att de kan reagera olika på variation i observationsprocessen, exempelvis vid störningar eller låg detektion, vilket gör det möjligt att jämföra deras robusthet.

5 Metod och simulering

Simuleringsstudien bygger direkt på den statistiska modell som beskrivs i föregående avsnitt. Syftet är att undersöka hur variation i populationsskattningarna uppstår och hur den påverkas av modellens olika komponenter.

Varje simulering motsvarar ett år och omfattar tre inventeringsdagar ($t = 1, 2, 3$). Först genererar vi populationen i varje polygon för den första inventeringsdagen enligt populationsmodellen. Därefter låter vi populationen förändras mellan dagarna genom förflyttning mellan polygoner enligt rörelsemodellen, vilket ger populationerna för dag 2 och dag 3.

När populationerna har genererats skapar vi observationer för varje polygon och dag enligt observationsmodellen. Observationerna summeras sedan över alla polygoner för att få dagliga totalsummor, vilka därefter används för att beräkna de olika estimatorerna av populationsstorleken.

För att undersöka hur variationen påverkas studerar vi detektionssannolikheten p, störningssannolikheten q samt rörelseparametern δ. Genom att systematiskt variera dessa parametrar kan vi analysera hur estimatorernas egenskaper förändras under olika förhållanden.

För varje parameterkombination upprepas simuleringen 1000 gånger. Detta gör det möjligt att approximera estimatorernas fördelningar och få stabila uppskattningar av variationen. Utöver simuleringarna analyserar vi även verkliga data. Estimatorerna beräknas då för varje år och jämförs, vilket gör det möjligt att relatera simuleringsresultaten till faktiska observationer.

6 Resultat och diskussion

I detta avsnitt presenterar vi resultaten från både simuleringar och analys av verkliga data. Fokus ligger på att undersöka hur variationen i populationsskattningar påverkas av modellens parametrar samt valet av estimator.

I simuleringarna analyserar vi särskilt effekten av detektionssannolikheten p, störningssannolikheten q och rörelseparametern δ. Genom att variera dessa parametrar kan vi studera hur olika delar av observationsprocessen bidrar till variation i skattningarna.

Resultaten från de verkliga data används för att sätta simuleringarna i ett praktiskt sammanhang och undersöka om de observerade mönstren stämmer överens med modellens förutsägelser.

Variation mäts med variationskoefficienten (CV) samt log-transformerad standardavvikelse (logSD). Eftersom dessa mått i stort sett visar samma mönster fokuserar vi främst på CV i resultatpresentationen, medan logSD används som ett kompletterande mått vid behov.

6.1 Effekt av detektionssannolikhet

För att undersöka effekten av detektionssannolikheten p studerar vi här ett scenario utan störningar ($q = 0$) och utan rörelse mellan polygoner ($\delta = 0$).

	CV				logSD			
p	Top2	Mean3	Max	SumMax	Top2	Mean3	Max	SumMax
0.3	0.013	0.013	0.015	0.014	0.013	0.013	0.015	0.014
0.4	0.013	0.012	0.013	0.013	0.013	0.012	0.013	0.013
0.5	0.012	0.011	0.012	0.012	0.012	0.011	0.012	0.012
0.6	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
0.8	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010

Tabell 1: Variationskoefficient (CV) och log-transformerad standardavvikelse (logSD) för olika estimatorer vid olika värden på detektionssannolikheten p.

Tabell 1 visar att variationen minskar något när p ökar. Skattningarna blir alltså lite mer stabila när en större del av populationen observeras. För Top2 minskar till exempel CV från 0.013 vid $p = 0.3$ till 0.010 vid $p = 0.8$.

Samtidigt är skillnaderna ganska små mellan de olika värdena på p. Det tyder på att detektionssannolikheten inte har så stor påverkan på variationen i detta scenario. Estimatorerna ger också liknande resultat. Max har genomgående något högre variation, medan Top2 och Mean3 är något stabilare. SumMax ligger mellan dessa och följer samma mönster.

Totalt sett verkar det alltså som att högre detektionssannolikhet ger något stabilare skattningar, men effekten är ganska svag när det inte finns några störningar eller rörelser i modellen.

6.2 Effekt av störningar

I detta scenario hålls detektionssannolikheten fast vid $p = 0.5$, samtidigt som ingen rörelse mellan polygoner tillåts ($\delta = 0$).

	CV				logSD			
q	Top2	Mean3	Max	SumMax	Top2	Mean3	Max	SumMax
0.00	0.012	0.011	0.012	0.012	0.012	0.011	0.012	0.012
0.05	0.028	0.034	0.022	0.012	0.028	0.035	0.023	0.012
0.10	0.045	0.051	0.042	0.015	0.046	0.052	0.043	0.015
0.15	0.058	0.063	0.057	0.017	0.059	0.064	0.059	0.018
0.20	0.067	0.071	0.069	0.025	0.068	0.072	0.070	0.026

Tabell 2: CV och logSD för olika estimatorer vid olika värden på störningsparametern q.

I Tabell 2 ser vi att variationen ökar när q blir större. Skillnaden syns redan vid låga värden på q. För Top2 ökar till exempel CV från 0.012 vid $q = 0$ till 0.028 vid $q = 0.05$. När störningarna blir större ökar också skillnaderna mellan simuleringarna. En möjlig förklaring är att högre värden på q ger fler nollobservationer, vilket gör skattningarna mindre stabila.

Mean3 har högst variation genom hela tabellen och verkar därför vara mest känslig för störningar. Top2 ger stabilare resultat, medan Max ligger mellan dessa. SumMax har lägst variation och påverkas minst när q ökar. Det här visar att störningar i observationsprocessen påverkar variationen mycket mer än förändringar i detektionssannolikheten.

6.3 Effekt av rörelse

Här studerar vi hur rörelse mellan polygoner påverkar variationen. I detta fall finns inga störningar ($q = 0$), och endast rörelseparametern δ ändras.

	CV				logSD			
δ	Top2	Mean3	Max	SumMax	Top2	Mean3	Max	SumMax
0.0	0.012	0.011	0.012	0.012	0.012	0.011	0.012	0.012
0.1	0.012	0.011	0.012	0.053	0.012	0.011	0.012	0.053
0.3	0.011	0.011	0.012	0.042	0.011	0.011	0.012	0.042

Tabell 3: CV och logSD för olika estimatorer vid olika värden på rörelseparametern δ.

Som framgår av Tabell 3 syns bara små förändringar för Top2, Mean3 och Max när δ ökar. Variationerna ligger nästan på samma nivå för alla värden på δ, vilket tyder på att dessa estimatorer påverkas ganska lite av rörelse mellan polygoner.

En möjlig förklaring är att rörelsen främst ändrar hur individerna fördelas mellan polygonerna, medan det totala antalet individer fortfarande är detsamma. Eftersom dessa estimatorer bygger på totalsummor över alla polygoner blir effekten därför liten.

För SumMax ser mönstret annorlunda ut. Där ökar CV tydligt när δ går från 0 till 0.1. Det visar att estimatorn är mer känslig för rörelse. Det beror troligen på att SumMax använder de högsta observationerna från varje polygon. När individer flyttar sig mellan polygonerna förändras också vilka polygoner som har höga värden från dag till dag, vilket ger större variation i skattningarna. För de globala estimatorerna är effekten alltså liten, medan SumMax påverkas betydligt mer av rörelse mellan polygoner.

6.4 Kombination av rörelse och störningar

Här varierar både störningsparametern q och rörelseparametern δ, medan detektionssannolikheten hålls fast vid $p = 0.5$.

q	δ	Top2	Mean3	Max	SumMax
0.00	0.0	0.012	0.011	0.012	0.012
0.00	0.1	0.012	0.011	0.012	0.053
0.00	0.3	0.011	0.011	0.012	0.042
0.05	0.0	0.028	0.034	0.022	0.012
0.05	0.1	0.026	0.034	0.022	0.052
0.05	0.3	0.027	0.034	0.023	0.044
0.10	0.0	0.045	0.051	0.042	0.015
0.10	0.1	0.043	0.050	0.040	0.052
0.10	0.3	0.042	0.048	0.040	0.045
0.15	0.0	0.058	0.063	0.057	0.017
0.15	0.1	0.055	0.059	0.054	0.054
0.15	0.3	0.053	0.059	0.051	0.046
0.20	0.0	0.067	0.071	0.069	0.025
0.20	0.1	0.066	0.073	0.066	0.058
0.20	0.3	0.067	0.073	0.067	0.053

Tabell 4: CV för olika estimatorer vid olika kombinationer av q och δ ($p = 0.5$).

Det tydligaste mönstret i Tabell 4 är att variationen ökar när q blir större. För alla estimatorer syns högre CV-värden vid större störningar än när $q = 0$. Mean3 och Max påverkas särskilt mycket när störningarna ökar.

Rörelse mellan polygoner verkar däremot ha ganska liten betydelse för Top2, Mean3 och Max. För samma värde på q ligger resultaten nära varandra även när δ ändras. Det tyder på att dessa estimatorer är relativt stabila trots att individer flyttar sig mellan polygonerna.

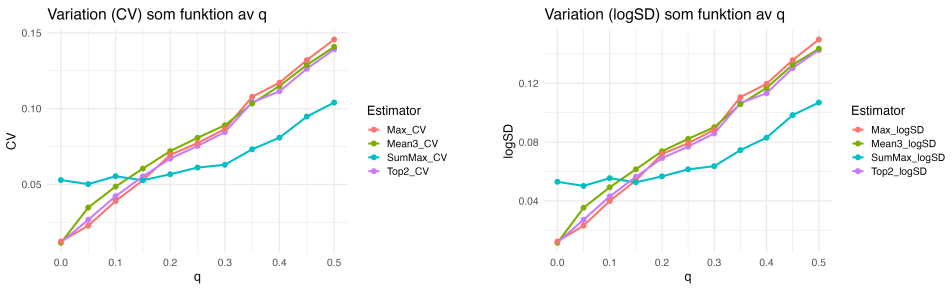
Samma mönster för SumMax i det här fallet också. Där ökar variationen tydligt så fort δ är större än noll. Skillnaden syns redan när $q = 0$, där CV ökar från 0.012 till 0.053 när δ ändras från 0 till 0.1. Tabellen visar därför att störningar påverkar alla estimatorer, medan rörelse främst påverkar estimatorer som bygger på lokala observationer.

6.5 Robusthet hos olika estimatorer

Till skillnad från avsnitt 6.2, där endast effekten av störningar studerades, analyseras här ett scenario där både störningar och rörelse mellan polygoner ingår. Rörelseparametern sätts till $\delta = 0.1$ för att undersöka hur estimatorerna påverkas när flera källor till variation verkar samtidigt. Och för att få en tydligare bild av utvecklingen studeras också ett större intervall av störningssannolikheten q.

q	Top2	Mean3	Max	SumMax
0.00	0.012	0.011	0.012	0.053
0.05	0.027	0.035	0.023	0.050
0.10	0.043	0.049	0.040	0.055
0.15	0.056	0.061	0.054	0.053
0.20	0.069	0.074	0.072	0.057
0.25	0.077	0.082	0.079	0.061
0.30	0.086	0.090	0.088	0.064
0.35	0.107	0.106	0.111	0.075
0.40	0.113	0.117	0.120	0.083
0.45	0.130	0.133	0.136	0.098
0.50	0.143	0.143	0.150	0.107

Tabell 5: CV för olika estimatorer som funktion av störningssannolikheten q.



(a) Variation (CV) som funktion av q.

(b) Variation (logSD) som funktion av q.

Figur 2: Variation för olika estimatorer som funktion av störningssannolikheten q.

Både tabellen och figuren visar att variationen ökar när q blir större. Skillnaden är särskilt tydlig vid höga värden på q, där alla estimatorer får betydligt högre CV än i fallet utan störningar. Vid låga nivåer av störningar har Mean3 lägst variation. När q ökar växer dock variationen snabbare för Mean3 än för flera av de andra estimatorerna. Max följer ett liknande mönster och får också hög variation vid stora värden på q. Eftersom estimatorn bygger på en enda observation påverkas den mer av slumpmässiga förändringar i data.

Top2 förändras mer gradvis över hela intervallet och ger stabilare resultat än Mean3 och Max vid högre nivåer av störningar. Det verkar därför som att estimatorn är mindre känslig för låga observationer. Och för SumMax ser utvecklingen annorlunda ut. Variationerna är redan relativt höga vid låga värden på q, vilket troligen hänger ihop med rörelse mellan polygonerna. Samtidigt ökar variationen långsammare när q blir större. Skillnaderna mellan estimatorerna blir alltså tydligare när både rörelse och störningar finns med i modellen.

6.6 Effekt i verkliga data

För de verkliga data används ett fast värde $p = 0.5$. Tidigare resultat visade att detektionssannolikheten har relativt liten påverkan på variationen jämfört med störningar och rörelse.

Estimator	CV	logSD
Top-2	0.169	0.169
Mean-3	0.173	0.173
Max	0.175	0.175
SumMax	0.134	0.134

Tabell 6: CV och logSD för olika estimatorer baserat på verkliga data.

Tabell 6 visar att Mean3 och Max har högst variation, medan SumMax har lägst CV. Top2 ligger mellan dessa estimatorer. Skillnaderna mellan Top2, Mean3 och Max är ganska små, men estimatorerna reagerar olika på variation i data. Mean3 och Max påverkas mer av extrema eller låga observationer, medan Top2 ger något stabilare resultat.

Variationerna i de verkliga data är också betydligt större än i simuleringarna utan störningar. CV-värdena ligger runt 0.17, vilket är nivåer som i simuleringarna främst syntes vid höga värden på q. Det tyder på att observationsprocessen i verkligheten innehåller mer störningar än i de enklaste scenarierna.

Samtidigt är modellen en förenkling av verkligheten, och flera faktorer kan bidra till variationen i data. Trots det liknar mönstren i de verkliga data resultaten från simuleringarna, särskilt scenarier med större störningar.

SumMax har lägst variation i detta datamaterial, men tidigare analyser visade att estimatorn påverkas mer av rörelse mellan polygoner. Top2 framstår därför som ett mer stabilt alternativ eftersom variationen är relativt låg samtidigt som estimatorn är mindre känslig för rörelse och extrema observationer.

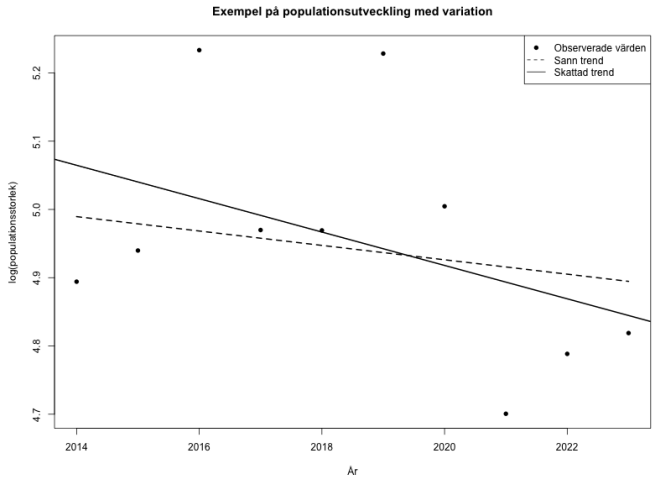
6.7 Detekterbarhet av populationstrender

För att koppla variationen i skattningarna till en praktisk tillämpning analyserar vi hur väl en förändring i populationsstorlek kan upptäckas över tid.

Vi antar att populationsutvecklingen följer en linjär trend på log-skalan, $\log(\hat{N}_t) = a + bt + \varepsilon_t$, där $\hat{N}_t$ är den skattade populationsstorleken vid tidpunkt t, b beskriver förändringstakten och $\varepsilon_t \sim N(0, \sigma^2)$ representerar slumpvariation. Här motsvarar σ de observerade logSD-värdena.

En minskning på 10% över tio år innebär att populationen multipliceras med 0.9. På log-skalan blir denna förändring additiv, som ger $b = \frac{\log(0.9)}{10} \approx -0.0105$, vilket motsvarar en årlig minskning på ungefär 1%.

Figur 3 visar ett simulerat exempel på en sådan utveckling för perioden 2014–2023. Punkterna motsvarar observationer med slumpvariation motsvarande den observerade logSD, medan den streckade linjen visar den sanna trenden och den heldragna linjen den skattade. Figuren illustrerar hur variation i observationerna kan dölja en svag trend.

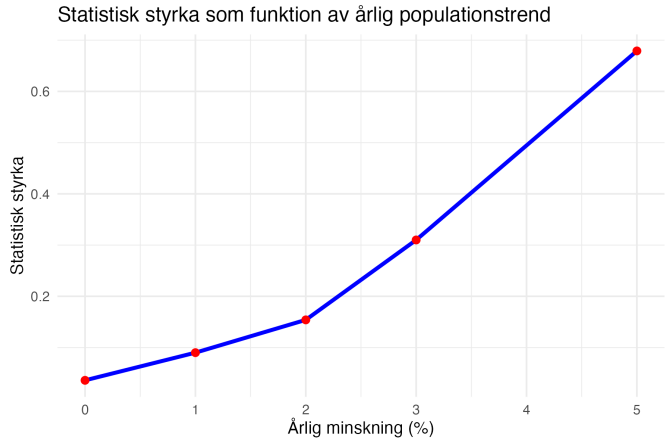


Figur 3: Simulerat exempel på populationsutveckling med 10% nedgång över tio år.

För att kvantifiera detta beräknar vi standardfelet för lutningen b, vilket kan

approximeras som $\mathrm{SE}(b) \approx \frac{\sigma}{\sqrt{\sum_{t=1}^{T}(t-\bar{t})^2}}$. Med $\sigma \approx 0.17$ och $T = 10$ så får vi $\mathrm{SE}(b) \approx 0.019$.

Eftersom den sanna lutningen är relativt liten, $b \approx -0.0105$, blir kvoten $\frac{|b|}{\mathrm{SE}(b)} \approx 0.55$, vilket ligger långt under det kritiska värdet 1.96. Detta innebär att vi i de flesta fall inte kan påvisa någon statistiskt signifikant trend. Plus visar simuleringarna också att den statistiska styrkan i detta fall är cirka 0.087, vilket innebär att en så liten trend endast upptäcks i en liten andel av simuleringarna.



Figur 4: Statistisk styrka som funktion av årlig populationstrend. Figuren visar att sannolikheten att upptäcka en trend ökar snabbt med trendens storlek.

I Figur 4 ser vi att styrkan beror starkt på trendens storlek. Vid en trend på omkring 1% per år är styrkan mycket låg, cirka 0.087, vilket innebär att sådana förändringar sällan upptäcks statistiskt. När trenden istället ökar till omkring 5% per år ökar styrkan till ungefär 0.67, vilket visar att större förändringar är betydligt lättare att identifiera.

Vi ser alltså att hög variation i skattningarna kraftigt begränsar möjligheten att upptäcka små förändringar i populationen. Även biologiskt relevanta trender kan därför förbli oupptäckta om variationen är stor.

Detta visar att låg variation i skattningarna inte bara är önskvärd, utan avgörande för att kunna upptäcka långsiktiga förändringar i populationen.

6.8 Koppling till teori

Den teoretiska modellen hjälper oss att förstå de mönster som syns i både simuleringarna och analysen av de verkliga data.

Som beskrevs i avsnitt 2.1 antas populationen följa en Poissonmodell, medan observationerna modelleras med en binomial observationsprocess. När detektionssannolikheten p får variera kan variansen delas upp enligt lagen om total varians i avsnitt 2.3.

Variationen består då av två delar. Den första kommer från själva observationsprocessen, medan den andra uppstår när detektionssannolikheten varierar mellan observationer. Detta förklarar varför förändringar i p gav relativt små skillnader i simuleringarna. När p ändras påverkas både väntevärde och varians ungefär proportionellt, vilket gör att den relativa variationen bara förändras lite.

För störningsparametern q ser situationen annorlunda ut. Högre värden på q ger fler låga eller noll observationer, vilket skapar extra variation i data. Därför ökar variationen snabbt när q blir större. Samma mönster syntes i simuleringarna, där även små ökningar i q gav tydligt högre CV och logSD. Observationerna blir då mer instabila, vilket gör populationsskattningarna mindre precisa.

Det går också att koppla detta till trendanalysen. När variationen i observationerna är hög blir osäkerheten i den skattade trenden större. Det gör det svårare att upptäcka förändringar i populationen över tid, särskilt negativa trender. Teorin stämmer alltså väl överens med resultaten i både simuleringarna och de verkliga data.

7 Slutsats

I denna studie har vi undersökt hur variation i populationsskattningar från sälinventeringar påverkas av olika faktorer i observationsprocessen, samt jämfört hur olika estimatorer presterar under varierande förhållanden.

Resultaten visar tydligt att störningar i observationsprocessen, representerade av parametern q, är den dominerande källan till variation. Redan vid låga värden på q ökar variationen markant, och ökningen är dessutom icke-linjär. Detta innebär att även små störningar kan få stor påverkan på skattningarnas stabilitet. Detta är i linje med den teoretiska modellen, där variation i observationsprocessen bidrar med en extra varianskomponent enligt lagen om total varians. Vi kan därför tolka störningsparametern q som en mekanism som förstärker variationen i observationsprocessen och därmed ökar osäkerheten i populationsskattningarna.

Detektionssannolikheten p påverkar främst skattningarnas skala och har begränsad effekt på den relativa variationen, eftersom p fungerar som en skalningsfaktor som påverkar både väntevärde och varians proportionellt. Rörelse mellan polygoner har däremot liten påverkan på de globala estimatorerna,

eftersom rörelse främst innebär en omfördelning av individer mellan områden snarare än en förändring av den totala populationen. SumMax-estimatorn påverkas dock tydligare, vilket visar att estimatorer som bygger på lokal information är mer känsliga för rumslig dynamik.

Jämförelsen mellan estimatorerna visar tydliga skillnader i robusthet. Mean3 är effektiv i frånvaro av störningar men blir snabbt känslig när variationen ökar, eftersom alla observationer inkluderas i skattningen. Top2 uppvisar däremot en mer robust utveckling och ger en bra balans mellan stabilitet och effektivitet. Max har genomgående hög variation eftersom estimatorn baseras på en enskild observation, medan SumMax uppvisar ett annorlunda beteende med hög variation i vissa situationer men relativt stabil utveckling i relativa mått.

Användningen av Top2 har även stöd i tidigare studier. I [2] beskrivs hur populationsskattningar ofta baseras på flera räkningar där de högsta observationerna ges störst vikt, vilket överensstämmer med principen bakom Top2.

Analysen av verkliga data visar att variationen är betydligt högre än i simuleringar utan störningar. De observerade nivåerna motsvarar scenarier med relativt höga värden på q, vilket tyder på att störningar är en central faktor i praktiska tillämpningar. Resultaten från trendanalysen visar dessutom att denna variation kan göra det svårt att statistiskt upptäcka populationsförändringar över tid, även när förändringarna är biologiskt relevanta.

Dessa resultat ligger i linje med tidigare forskning. Det är välkänt att observerade populationsdata ofta påverkas mer av variation i observationsprocessen än av förändringar i den underliggande populationen [4]. Vidare visar modern metodutveckling att variation i detektionssannolikhet kan maskera verkliga populationstrender och därmed minska den statistiska styrkan i analyser [6]. Den modellstruktur som används, där populationsprocess och observationsprocess modelleras separat, är standard inom ekologisk modellering och möjliggör en mer realistisk tolkning av data [5]. Resultaten kan även relateras till överdispersion i ekologiska räknevariabler, där standardmodeller ofta underskattar variationen [8].

Det är samtidigt viktigt att betona modellens begränsningar. Faktorer som väderförhållanden, beteendevariation och tidsberoende effekter modelleras inte explicit, vilket kan påverka överförbarheten till verkliga data [12]. Vidare antas parametrarna vara konstanta inom varje år, vilket är en förenkling.

Sammantaget visar studien att variation i observationsprocessen är avgörande för osäkerheten i populationsskattningar från sälinventeringar. Vi ser också att valet av estimator har stor betydelse för hur denna variation påverkar resultaten, där Top2 framstår som ett robust alternativ i närvaro av störningar. Studien bidrar därmed till en ökad förståelse för hur osäkerhet

uppstår i sälinventeringar och kan användas som stöd vid praktisk populationsövervakning.

Referenser

[1] Thompson, P. M., & Harwood, J. (1990). Methods for estimating the population size of common seals, *Phoca vitulina*. *Journal of Applied Ecology*, 27(3), 924–938.

[2] Teilmann, J., Rigét, F., & Härkönen, T. (2010). Optimizing survey design for Scandinavian harbour seals. *ICES Journal of Marine Science*, 67(5), 952–958.

[3] Held, L., & Sabanés Bové, D. (2020). *Likelihood and Bayesian Inference* (2nd ed.). Springer.

[4] Hayward, J. L., Henson, S. M., Logan, C. J., Parris, C. R., Meyer, M. W., & Dennis, B. (2005). Predicting numbers of hauled-out harbour seals: a mathematical model. *Journal of Applied Ecology*, 42, 108–117.

[5] Royle, J. A., & Dorazio, R. M. (2008). *Hierarchical Modeling and Inference in Ecology*. Academic Press.

[6] Conn, P. B., Johnson, D. S., Boveng, P. L., & London, J. M. (2018). A guide to Bayesian hierarchical models for ecological data. *PLOS ONE*, 13(7), e0200253. https://doi.org/10.1371/journal.pone.0200253

[7] Carrizo Vergara, Ricardo and Kéry, Marc and Hefley, Trevor. (2025) The evolving categories multinomial distribution: introduction with applications to movement ecology and vote transfer. arXiv preprint arXiv:2505.20151.

[8] Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11), 2766–2772.

[9] Gerrodette, T. (1987). A power analysis for detecting trends. *Ecology*, 68(5), 1364–1372.

[10] Giardina, F., Gosoniu, L., Konate, L., Diouf, M. B., Perry, R., Gaye, O., Faye, O., & Vounatsou, P. (2012). Estimating the burden of malaria in Senegal: Bayesian zero-inflated binomial geostatistical modeling of the MIS 2008 data. *PLOS ONE*, 7(3), e32625. https://doi.org/10.1371/journal.pone.0032625

[11] Yee, T. W. (2024). VGAM: Vector Generalized Linear and Additive Models (R package documentation). https://search.r-project.org/CRAN/refmans/VGAM/html/zibinomUC.html

[12] Blanchet, M.-A., Vincent, C., Womble, J. N., Steingass, S. M., & Desportes, G. (2021). Harbour seals: population structure, status, and threats in a rapidly changing environment. *Oceans*, 2, 41–63.

[13] R Core Team. (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/