Säkrare skördeprognoser

Kvantifiering av osäkerhet, spatiala samband och regularisering av Jordbruksverkets modeller

Erik Göransson Gaspar

Examensarbete för Kandidatexamen
2024:K15

Lunds universitet

Naturvetenskaplig fakultet
Matematikcentrum
Matematisk statistik

Centrum Scientiarum Mathematicarum

Abstrakt

Jordbruksverket presenterar varje sommar en prognos av riksskördarna av 13 viktiga spannmåls- och oljegrödor. Hektarskörden av varje gröda i varje län modelleras var för sig utifrån väderobservationer m.h.a. linjär regression och empiriskt kända samband. I denna uppsats utreds och förbättras dessa modeller. För det första replikeras prognoserna tillsammans med simuleringsbaserade prediktionsintervall för att kvantifiera deras osäkerhet. Dessutom föreslås modifierade modeller med nästan hälften så stort prediktionsfel som Jordbruksverkets. Detta görs på två sätt; dels genom att modifiera regressionsmatriserna för att reducera dess kolinearitet, dels genom tillämpningen av regulariserad regression. Slutligen visar jag att en spatial modell som använder data från flera län och modellerar interaktionseffekter mellan dessa kan prestera bättre än att modellera varje län och gröda separat. Detta är en lovande inriktning för fortsatt utredning.

Abstract

Every summer the Swedish Department of Agriculture publishes a forecast of the national harvest of 13 key cereal and oilseed crops. The yield of each crop in each region is modeled individually based on weather observations using linear regression and empirically known relationships. This thesis investigates and improves these models. First, the forecasts are replicated along with simulation-based prediction intervals to quantify their uncertainty. Additionally, modified models are proposed, achieving nearly half the prediction error of the Department of Agriculture's. This is accomplished in two ways: by modifying the regression matrices to reduce their collinearity and by applying regularized regression. Finally, I demonstrate that a spatial model, which uses data from multiple counties and models the interaction effects between them, can perform better than modeling each county and crop separately. This is a promising direction for further research.

Populärvetenskaplig sammanfattning

Varje sommar publicerar Jordbruksverket en prognos för hur stora skördarna av tretton viktiga grödor kommer att bli det året. Den låter både bönder och livsmedelsföretag planera sin verksamhet. Därför är det viktigt att prognoserna är så bra som möjligt. Hur stora skördarna blir per odlad hektar (den s.k. hektarskörden) varierar mellan grödor och beror bl.a. på jordmånen och vädret. Jordbruksverket använder sig av linjär regression för att modellera sambandet mellan hektarskörden och vädret under året. Jag visar hur denna modell kan förändras för att ge bättre prognoser. Som bäst har dessa modifierade modeller hälften så mycket fel som Jordbruksverkets.

Alla prognoser osäkra. När man ska tolka prognoser är det därför viktigt att veta hur osäkra de är. Jordbruksverkets modell ger oss ett (mer eller mindre riktigt) samband mellan årets väder och hektarskörden av en viss gröda i ett visst län. Detta låter oss simulera hur stora skördarna kommer vara om detta samband håller. Säg att vi gör 1000 sådana simulationer. Då kan vi avgöra inom vilket intervall som 95 % av de simulerade skördarna faller inom. Detta intervall är ett mått på osäkerheten inneboende i modellen.

Årets väder fångas av ett antal mått på temperatur och nederbörd. Utöver medeltemperatur och total nederbörd för växtsäsongens månader anges också sådant som antalet dagar med nederbörd under en månad och medeltemperaturen kl. 12:00 för månadens fem varmaste dagar. Ju fler mått desto mer information har modellen att utgå ifrån. Vissa mått är däremot starkt kopplade: regnar det många dagar under en månad kommer ofta den totala nederbörden vara hög, t.ex. När sådana kopplingar, s.k. kolinearitet, finns förvirrar det regressionsmodellen och den ger sämre prognoser. Därför visar det sig att skördeprognosen förbättras av att ta bort vissa vädermått och slå ihop andra till ex. medeltemperatur för hela våren. Detta är att ändra data som modellen utgår ifrån.

Kolinearitet mellan olika vädermått kan också åtgärdas genom att ändra hur själva regressionsmodellen fungerar. Vanlig linjär regression hittar det samband bland data som ger minst prognosfel. Är flera mått kolineära så är risken att detta samband inte gäller för kommande år, d.v.s. för data som modellen inte har sett tidigare. Istället väljer vi ett samband som tar så lite hänsyn till så få vädermått som möjligt, så länge som prognosfelet fortfarande blir lågt. Detta sätt att reducera modellen, som statistiker säger, kallas för regulariserad regression. I uppsatsen visar jag att modellens prognosfel kan halveras antingen genom att ändra data så att korrelationen mellan mått minskar, eller genom att använda regulariserad regression.

I uppsatsens sista del utforskar jag en väsentligt annorlunda sorts modell. Jordbruksverkets modell (och de andra modeller som jag har diskuterat) behandlar varje gröda i varje län var för sig. Tyvärr har vi för flera grödor i flera län inte statistik för så många år. Detta gör att många samband bygger på väldigt få data; dåligt underbyggda samband leder till dåliga prognoser. Jag lägger fram ett utkast till en modell som tittar på hur en gröda växer i flera län samtidigt. På så sätt bygger dess prognoser på många gånger fler data än Jordbruksverkets modell. Preliminära resultat visar att denna modell presterar bättre än någon annan som undersöks i uppsattsen. Detta är en lovande riktning för vidareutveckling av Jordbruksverkets skördeprognos.

Innehåll

1. Introduktion

Statens Jordbruksverk är Sveriges förvaltningsmyndighet för jordbruk, fiske och landsbygd. De har i uppgift att främja en hållbar matproduktion och gott djurskydd i Sverige. Som en led i detta arbete ansvarar myndigheten för den officiella statistiken om jordbruk, trädgårdsodling och vattenbruk.

1.1. En prognos av svenska skördar

Varje sommar publicerar Jordbruksverket en skördeprogonos. I den förutsägs storleken av riksskördarna för tretton viktiga grödor. Prognosen berhandlar nio spannmålsgrödor:

- höstvete
- vårvete
- höstråg
- höstkorn
- vårkorn
- havre
- blandsäd
- höstrågvete
- vårrågvete

och fyra oljeväxtgrödor:

- höstraps
- vårraps
- höstrybs
- vårrybs

Dessa grödor odlas i vitt skilda omfattningar; höstvete stod ensamt för en majoritet av spannmålsskörden år 2022 och höstraps för 90 % av oljeväxtskörden. [5] Tabell 1 anger hur stora riksskördarna av de tretton grödorna var år 2022. Märk väl att Jordbruksverket inte publicerar prognoser för andra viktiga grödor, ex. potatis och grönsaker. Utifrån observationer av temperatur och nederbörd under året i alla Sveriges län modelleras storleken av skördarna. En modell passas för varje gröda i varje län. Detta låter prognosen ta hänsyn till att olika grödor har olika växtmönster och reagerar olika på samma väderförhållande. På samma sätt varierar jordbrukets förutsättningar inom landet. Jordmånen i vissa län är särskilt gynnsamma för vissa sorters grödor och i de stora lantbruksområdena (ex. Skåne och Västra Götaland) är jordbruket mer storskaligt, vilket också påverkar avkastningen. I och med att prognosen publiceras innan årets skördesäsong har avslutats låter den både bönder och livsmedelsföretag planera driften av sin verksamhet. Därför är det viktigt att prognosen har både låg osäkerhet och lågt prediktionsfel.

1.2. Frågeställningar

Denna uppsats behandlar tre frågeställningar. För det första undersöker jag osäkerheten i Jordbruksverkets modell. Myndigheten presenterar inte själv någon sådan analys. Om osäkerheten är stor riskerar avsaknaden av ett osäkerhetsmått ge en missvisande bild av modellens förutsägelser. För det andra utreder jag huruvida förändringar i modellen skulle kunna ge lägre prediktionsfel. Jag presenterar modeller med nästan hälften så stort

Tabell 1: Riksskördarna år 2022 av de grödor som behandlas i Jordbruksverkets skördeprognos. Andelen beräknas som del av den totala skörden av dessa tretton grödor. Statistiken är hämtad från [5].

Gröda	Riksskörd år 2022 (tusen ton)	Andel av den totala riksskörden
Höstvete	3018	48 %
Vårvete	211	3 %
Höstråg	129	2 %
Höstkorn	130	2 %
Vårkorn	1379	22 %
Havre	735	12 %
Höstrågvete	155	2 %
Vårrågvete	8	<1 %
Blandsäd	46	<1 %
Höstraps	388	6 %
Vårraps	37	<1 %
Höstrybs	1	<1 %
Vårrybs	3	<1 %

prediktionsfel som den Jordbruksverket använder. Slutligen utforskar jag huruvida en spatial modell som tar hänsyn till interaktionseffekter mellan län och därmed kan nyttja mångfaldigt fler data än när hektarskördarna av varje gröda i varje län modelleras var för sig.

1.3. Uppsatsens upplägg

I nästa avsnitt redogör jag i detalj för Jordbruksverkets modell. Sedan presenterar jag ett osäkerhetsmått för densamma. Resten av uppsattsen ägnas åt alternativa modeller. I avsnitt 4 visar jag att prognosen kan förbättras avsevärt genom att heuristiskt reducera kolineariteten av de oberoende variablerna. Avsnitt 5 innehåller en undersökning av regulariserade regressionsmodeller som både är mindre känsliga för kolinearitet och kräver mindre mängd data. Slutligen presenterar jag en spatial modell som utnyttjar att storleken av skördarna i angränsande län är korrelerade.

2. Jordbruksverkets modell

2.1. Regressionsmodellen

Jordbruksverket modellerar separat *hektarskörden* (kg/ha) av varje gröda i alla län som den odlas i. Vi vet hur stora arealer som varje gröda odlas på i varje län, eftersom lantbrukare rapporterar in detta till Jordbruksverket som en del av EU:s jordbruksstöd. Dessa arealer tillsammans med de skattade hektarskördarna ger oss en prognos för den totala länsskörden av alla grödor. Vissa arealer anges vara ”ej regionbestämda”; för dem

använder Jordbruksverket rikshektarskörden, ett länsvis medelvärde viktat efter odlad areal. Summan av prognoserna för länsskördarna ger oss en uppskattad totalskörd för hela riket. Följande redogörelse av Jordbruksverkets modell är baserad på [6].

När Jordbruksverket efter varje årsskifte redovisar de faktiska skörderesultatet för förra årets säsong exkluderas de arealer som skördats som grönfoder, d.v.s. ännu omogna växter ämnade att bli djurfoder. Skördeprognosen tar ingen hänsyn till detta, vilket leder till en systematisk överskattning av riksskörden. Storleken av detta fel varierar år till år, beroende på andelen grönfoderuttag. Under torrår är bönder t.ex. tvungna att skörda grödor som foder för sin boskap, istället för att sälja den på marknaden. Skörden av blandsäd överskattas särskilt mycket, eftersom vissa sädesblandningar odlas just i syftet att skördas som grönfoder. Hektarskörden förutsägs med hjälp av en regressionsmodell för varje gröda i varje län. För län ℓ ställer vi upp ett system av regressionsmodeller med 35 beroende variabler och en konstant vardera:

$$\boldsymbol{Y}_\ell = \boldsymbol{X}_\ell \boldsymbol{\beta}_\ell + \boldsymbol{\epsilon}_\ell$$

Låt N vara antalet år det finns data för (som mest från 1961 till 2022) och G antalet grödor som odlas i län ℓ. Då är

- $\boldsymbol{Y}_\ell$ en $N \times G$ matris sådan att $[\boldsymbol{Y}_\ell]_{ij}$ är den observerade hektarskörden av gröda j år i i län ℓ.
- rad i av $\boldsymbol{X}_\ell$ är observationen av de 35 oberoende variablerna år i, tillsammans med en konstant sådan att $[\boldsymbol{X}_\ell]_{i1} \equiv 1$. $\boldsymbol{X}_\ell$ är alltså en $N \times 36$ matris.
- $[\boldsymbol{\beta}_\ell]_{kj}$ är regressionskoefficenten för gröda j och beroende variabel k.
- $[\boldsymbol{\epsilon}_\ell]_{ij} \sim \mathrm{NID}(0, \sigma^2_{\ell j})$ är slumpfelet av hektarskörden för gröda j år i i län ℓ som inte kan förklaras av de oberoende variablerna. Variansen $\sigma^2_{\ell j} > 0$ antags i ett och samma län vara gemensam för en gröda under alla år.

Eftersom grönfoderuttaget gör att modellen har systematiskt fel torde slumpfelen inte ha medelvärde noll. Detta borde däremot inte skada modellens prediktionsförmåga, eftersom det systematiska felet inkluderas i koefficenten av konstanten $[\boldsymbol{X}_\ell]_1 \equiv 1$. De oberoende variablerna i $\boldsymbol{X}_\ell$ består huvudsakligen av olika temperatur- och nederbördsmått, närmare bestämt:

- $[\boldsymbol{X}_\ell]_1$: en konstant $\equiv 1$.
- $[\boldsymbol{X}_\ell]_2$: en trendvariabel lika med året i fråga (2023, 2022, 2021, ...)
- $[\boldsymbol{X}_\ell]_3$ – $[\boldsymbol{X}_\ell]_{12}$: månadsmedelvärden för dygnsmedeltemperaturen (°C) för jan., feb., ..., sep., okt.
- $[\boldsymbol{X}_\ell]_{13}$–$[\boldsymbol{X}_\ell]_{22}$: den totala nederbörden (mm) under jan., feb., ..., okt.
- $[\boldsymbol{X}_\ell]_{23}$–$[\boldsymbol{X}_\ell]_{26}$: månadsmedelvärden av temperaturen (°C) kl. 12:00 för de fem dygn som då var varmast under mars, april, maj, och juni.

- $[\boldsymbol{X}_\ell]_{27}$–$[\boldsymbol{X}_\ell]_{30}$: månadsmedelvärden av temperaturen (°C) kl. 06:00 för de fem dygn som då var kallast under mars, april, maj, och juni.
- $[\boldsymbol{X}_\ell]_{31}$–$[\boldsymbol{X}_\ell]_{34}$: månadsvärden för antalet dagar med nederbörd > 0 mm under mars, april, maj, och juni.
- $[\boldsymbol{X}_\ell]_{34}$–$[\boldsymbol{X}_\ell]_{36}$: kombinationsvariabler lika med månadsvärdet för antal dagar med nederbörd multiplicerat med den totala månadsnederbörden under april och maj.

För att förbättra modellernas generaliseringsförmåga reduceras de med hjälp av bakåteliminering. Detta görs för alla grödor i de län där det finns tillräckligt med datapunkter att tillgå. Jordbruksverket beskriver inte deras implementering; jag har använt en implementering av algoritm 1.

Algoritm 1: Bakåteliminering

$\alpha \leftarrow 0.05$;
$S \leftarrow \{\text{alla oberoende variabler}\}$;
$m \leftarrow$ regressionsmodell passad till $\boldsymbol{Y}_\ell, \boldsymbol{X}_\ell$;
$p_{\max} \leftarrow \max\{\text{p-värdena av de oberoende variablerna i } m\}$;
Medan $p_{max} > \alpha$ **låt**
  $S \leftarrow S \setminus \{\text{den oberoende variabeln med p-värde } p_{\max}\}$;
  $m \leftarrow$ regressionsmodell passad till $Y_\ell, X_\ell[S]$;
  $p_{\max} \leftarrow \max\{\text{p-värdena av de oberoende variablerna i } m\}$;
Slut.

Det krävs alltså minst 37 datapunkter för att kunna ta fram entydiga modeller på detta sätt: 36 för de oberoende variablerna (inkl. en konstant) och en frihetsgrad till residualerna för att räkna ut p-värden. I de flesta län har vi för flera grödor inte så många datapunkter att tillgå; för blandsäd och höstrågvete har vi inte så många datapunkter i något län. Vi saknar fullständigt historiska hektarskördar för höstrybs och vårrågvete. Antalet tillgängliga datapunkter för varje gröda i varje län anges i tabell 2. I avsnitt 2.3 nedan redogörs för hur Jordburksverket löser detta problem. Jag förslår alternativa metoder i avsnitt 5. För de grödor i de län som vi har tillräckligt med data för kan vi göra en prognos. Låt $\boldsymbol{x}_\ell$ vara observationsvektorn av de oberoende variablerna för innevarande år i län ℓ. Då är den prognostiserade hektarskörden av gröda j i län ℓ

$$\hat{y}_{j\ell} = \boldsymbol{x}_\ell \cdot \hat{\boldsymbol{\beta}}_{j\ell}$$

där $\hat{\boldsymbol{\beta}}_{j\ell}$ är den relevanta koefficentvektorn. Eftersom Jordbruksverket publicerar skördeprognosen i mitten av augusti varje år så är observationer av vädret i augusti, september och oktober månader inte tillgängliga. Vid prediktion används därför genomsnittet av dessa oberoende variablers värden de senaste 30 åren. Denna imputering utgör givetvis en felkälla för prognosen; jag har inte undersökt hur stor denna effekt är.

Tabell 2: Antal tillgängliga datapunkter för hektarskörden av varje gröda i alla län. Grödor i län med minst 37 datapunkter är markerade. Historiska hektarskördar för höstrybs och vårrågvete saknas överhuvudtaget.

Län	Vårkorn	Höstvete	Vårvete	Blandsäd	Havre	Höstrågvete	Vårrybs	Höstråg	Höstraps	Vårraps	Höstkorn
Västernorrland	56	0	0	0	0	0	0	0	0	0	0
Jönköping	58	28	16	27	58	17	0	0	0	0	0
Dalarna	58	37	18	0	58	0	27	0	0	0	0
Jämtland	58	0	0	0	0	0	0	0	0	0	0
Uppsala	58	58	57	22	58	9	22	39	28	47	0
Halland	58	57	45	16	58	28	0	32	33	35	7
Örebro	58	58	58	0	58	15	23	39	18	40	0
Kronoberg	56	4	10	0	56	16	0	0	0	0	0
Gävleborg	58	7	16	0	58	0	0	0	0	0	0
Västerbotten	56	0	0	0	37	0	0	0	0	0	0
Västragötaland	58	58	55	38	58	28	29	58	48	43	14
Stockholm	58	58	41	0	58	7	22	28	36	42	0
Västmanland	58	57	57	15	58	4	31	0	14	46	0
Skåne	58	58	58	0	58	28	0	58	50	40	28
Blekinge	58	53	37	0	47	16	0	0	29	0	0
Kalmar	58	58	46	16	58	28	0	44	48	0	28
Värmland	58	55	20	0	58	8	29	0	0	24	0
Norrbotten	56	0	0	0	16	0	0	0	0	0	0
Gotland	58	58	26	0	58	25	0	52	48	32	25
Östergötland	58	58	56	32	58	28	24	57	50	41	14
Södermanland	58	58	58	22	58	26	26	44	33	42	0

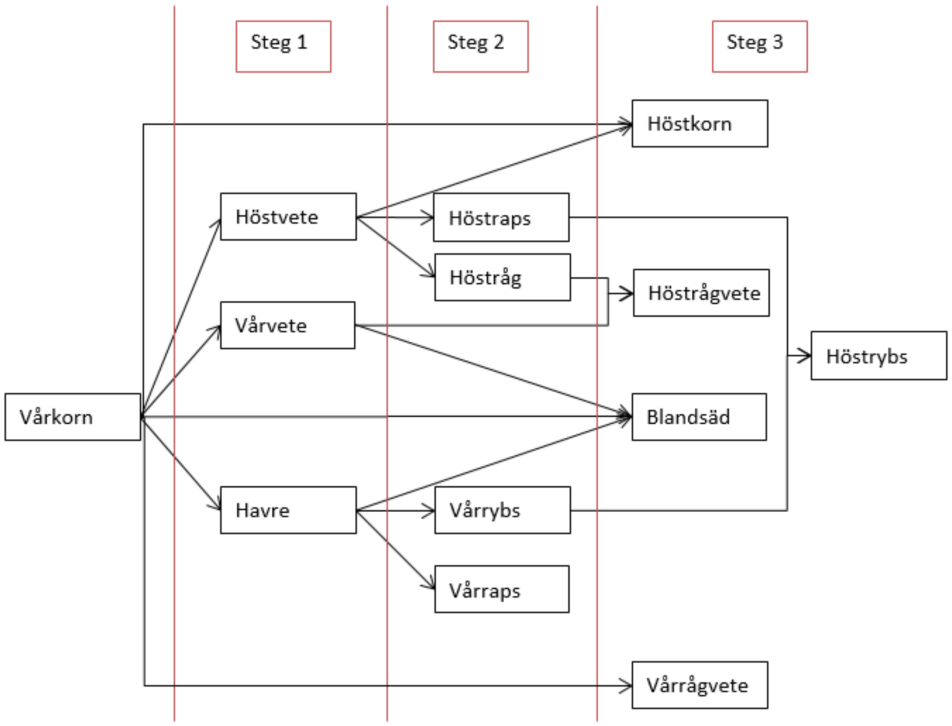
2.2. Datamängden

Jordbruksverkets prognoser bygger på historiska väderobservationer hämtade från SMHI. För varje län används temperatur- och nederbördsobservationer vid en väderstation. Vissa stationer har lagts ner och vissa har tillkommit sedan 1961 vilket gör att alla tidsserier inte utgörs av data från en och samma station; de stationer som 2023 års observationer är hämtade från anges i [6]. Jordbruksverket uppger inte alla väderstationer som använts för att konstruera tidsserien. Därför är det omöjligt för en utomstående att rekonstruera denna datamängd. Eftersom vädret kan förväntas variera även inom ett län så skulle kanske prognoserna förbättras om de byggde på observationer från flera väderstationer i varje län. Jordbruksverket uppger dock att de har undersökt denna möjlighet utan förbättrade resultat; jag har inte utforskat detta vidare. [4] Historiska hektarskördar för de tretton grödorna i alla Sveriges län finns att tillgå i Jordbruksverkets statistikdatabas. Jag har

fått tillgång till den datamängd som Jordbruksverket har använt när de gjort sin prognos. Den är nu publicerad på GitHub: `https://github.com/Furishon/skordeprognos`.

2.3. Indirekt skattning

Vårkorn är den enda grödan som vi har tillräckligt många datapunkter att tillgå för att passa regressionsmodeller i varje län. När det inte går att skapa en regressionsmodell så skattar Jordbruksverket den prognostiserade hektarskörden indirekt genom empiriska samband med andra grödors avkastningsmönster. Skattningen görs i den ordning som anges i figur 1. Notera att indirekt skattning krävs eftersom regressionsmodellerna bakåtelimineras. Hade framåtinkludering istället använts hade regressionsproblemet varit lösbart för alla grödor i alla län som vi har data för. Jag har inte undersökt denna möjlighet vidare. I avsnitt 5 redogör jag däremot för hur regularisering kan användas för att undvika indirekt skattning överhuvudtaget.



Figur 1: För de grödor i de län som det inte finns tillräckligt mycket data att tillgå kan inte regressionsmodeller konstrueras. Istället skattas hektarskördarna indirekt m.h.a. empiriska samband enligt denna ordning. Diagramet är taget från [6].

Olika grödor skattas indirekt på olika sätt. De flesta indirekta skattningar antar att kvoten av två olika grödors hektarskördar är den samma i angränsande län. Låt a_ℓ vara hektarskörden i län ℓ som ska skattas indirekt och b_ℓ hektarskörden av en viss annan

gröda i samma län. Om a_k och b_k är hektarskördarna av dessa grödor i ett angränsade län k så ges den indirekta skattningen av följande samband:

$$\frac{a_\ell}{b_\ell} = \frac{a_k}{b_k} \quad \Longrightarrow \quad a_\ell = \frac{b_\ell a_k}{b_k}$$

I tabell 3 anges vilka par av huvudgrödor och alternativgrödor som Jordbruksverket skattar på detta sätt. Märk väl att denna metod kräver att det finns en direkt skattning av grödans hektarskörd i minst ett län. För blandsäd, höstrågvete, vårrågvete och höstrybs finns det så få datapunkter att tillgå att de måste skattas indirekt på andra sätt.

Tabell 3: Grödorna som används för indirekt skattning av Jordbruksverket. Kvoten mellan hektarskördarna av huvudgrödan och alternativgrödan antags vara densamma i angränsande län.

Huvudgröda	Alternativgröda
Höstvete	Vårkorn
Vårvete	Vårkorn
Höstkorn	Vårkorn
Havre	Vårkorn
Höstråg	Höstvete
Höstraps	Höstvete
Vårraps	Havre
Vårrybs	Havre

Blandsäd

Hektarskörden av blandsäd a_ℓ skattas i län ℓ som

$$a_\ell = 0{,}8 \times \frac{b_\ell + c_\ell + d_l}{3}$$

där b_ℓ, c_ℓ och d_ℓ är den prognostiserade hektarskörden av vårvete, vårkorn respektive havre i län ℓ.

Höstrågvete

Hektarskörden av höstrågvete a_ℓ skattas i län ℓ som

$$a_\ell = \max\{b_\ell, c_\ell\}$$

där b_ℓ och c_ℓ är prediktionen av hektarskörden av vårvete respektive vårkorn i län ℓ.

Vårrågvete

Hektarskörden av vårrågvete a_ℓ skattas i alla län ℓ som

$$a_\ell = 0{,}7 b_\ell$$

där b_ℓ är den prognostiserade hektarskörden av vårkorn i län ℓ.

Höstrybs

Hektarskörden av höstrybs a_ℓ skattas i alla län ℓ som

$$a_\ell = 0{,}75\, b_\ell + 0{,}25\, c_\ell$$

där b_ℓ och c_ℓ är den prognostiserade hektarskörden av vårrybs respektive höstraps i län ℓ.

2.4. Utvärdering och validering

För att säkerställa min implementering av Jordbruksverkets modell har jag replikerat deras prognos för riksskörden år 2023. Resultaten jämförs med den faktiska riksskörden i tabell 4. Min implementering ger något annorlunda resultat än Jordbruksverket. För alla grödor vars hektarskördar skattas indirekt är detta väntat, eftersom Jordbruksverket inte anger vilket av flera angränsande län som ska användas vid beräkningarna; jag har tagit ett ur mängden. Att min prognos för riksskörden av vårkorn är så mycket större än Jordbruksverkets måste däremot bero på någon annan skillnad mellan implementeringarna, eftersom dess hektarskörd bara skattas direkt m.h.a. regressionsmodeller. Skilnaderna i de andra grödorna lär också delvis vara propagerade från skillnaden i prognoserna för vårkorn. Eftersom jag inte har tillgång till Jordbruksverkets exakta implementering kan jag inte avgöra var skillnaden ligger.

Jordbruksverket ger prognoser för riksskörden av höst- och vårrybs, utan att förklara hur de kan göra detta. I inget län finns tillräckligt många (≥ 37) datapunkter att tillgå för vårrybs. Trots det uppger Jordbruksverket att de indirekt skattar hektarskörden av vårrybs på ett sätt som kräver en direkt skattning i minst ett län. Ingenstans beskriver de hur detta går ihop. Hektarskörden av höstrybs skattas sedan indirekt utifrån hektarskörden av vårrybs. Eftersom Jordbruksverket inte beskriver hur prognoserna för vårrybs och höstrybs har tagits fram kan jag inte replikera dem. I resten av uppsattsen kommer min implementering av Jordbruksverkets modell därför inte ge prognoser för höst- och vårrybs. Notera dock att dessa två grödor står för mindre än en procent vardera av den totala riksskörden av de tretton grödor som behandlas i Jordbruksverkets skördeprognos. Uteslutandet av höst- och vårrybs har därmed bara en liten påverkan på prognosens relevans för det svenska jordbruket.

För att avgöra hur bra prognoser Jordbruksverkets modell kan förväntas ge behöver vi kvantifiera hur väl modellen generaliserar, d.v.s. hur bra prognoserna blir för data som inte ingått i skattningarna. Detta kan göras på många olika sätt. Eftersom vi har

Tabell 4: Min och Jordbruksverkets prognos för riksskörden 2023 jämte storleken av de faktiska skördarna. Den faktiska riksskörden för år 2023 är hämtad från Jordbruksverkets Statistikdatabas. Jordbruksverkets prognos är hämtad från [5]. Notera att min implementering inte inkluderar höst- och vårrybs. Alla skördar anges i tusentals ton.

Gröda	Riksskörd 2023	Jordbruksverkets prognos	Min prognos
Höstvete	2636	2841	2889
Vårvete	132	175	183
Höstråg	139	148	151
Höstkorn	98	103	109
Vårkorn	757	996	1101
Havre	412	492	571
Blandsäd	18	19	21
Höstrågvete	115	168	152
Vårrågvete	4,8	7	7
Höstraps	277	383	397
Vårraps	26	24	24
Höstrybs	0,6	1	—
Vårrybs	1,2	2	—

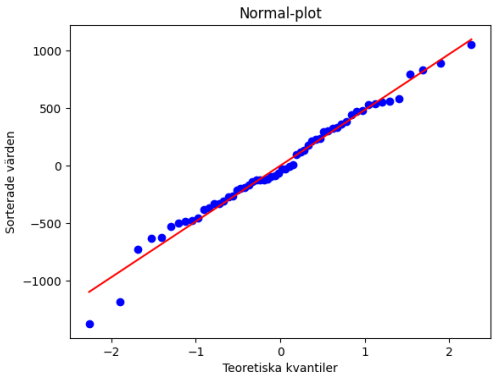
tillgång till relativt få datapunkter använder vi n-faldig korsvalidering som låter oss utnyttja all data. [3] Det går till på följande sätt: välj en datapunkt och träna modellen på all data förrutom denna datapunkt. Gör sedan en prediktion av denna datapunkten och utvärdera resultatet. Upprepa detta för alla datapunkter. Det genomsnittliga prediktionsfelet kallas för korsvalideringsfelet och är ett mått på modellens förmåga att generalisera till ny data. Denna metod ger oss korsvalideringsfel för varje gröda i varje län. Eftersom modellen syfte är att förutsäga riksskördar redovisas det genomsnittliga korsvalideringsfelet, viktat efter odlad areal, för varje gröda i tabell 5. Dessa viktas sedan samman utifrån varje grödas andel av den totala riksskörden, vilken anges i tabell 1. Detta ger oss ett mått på modellens prediktionsförmåga som tar hänsyn till att det är viktigare att ge bra prediktioner för de grödor som det odlas mycket av i de län som det odlas mycket i. Jordbruksverkets modell ger olika bra prediktioner för olika grödor. Särskilt allvarligt är att modellen ger sämst prognos för hektarskörden av höstvete, vilket är den avgjort vanligaste grödan efter vikt.

Det återstår att bekräfta modellantaganden genom en residualanalys. Låt oss först analysera ett representativt exempel grafiskt. Figur 2a visar residualerna av Jordbruksverkets modell $\hat{\boldsymbol{\epsilon}}_\ell = \boldsymbol{Y}_\ell - \hat{\boldsymbol{Y}}_\ell$ för höstvete i Skåne på normalfördelningspapper; den visar att residualerna är ungefärligt normalfördelade. Figur 2b och 2c visar residualerna plottade mot den uppskattade hektarskörden respektive observationsåret. De uppvisar inget systematiskt mönster och kan antags vara oberoende. Märk väl att höstvete i Skåne står för ca. 14 % av den totala riksskörden av alla tretton grödor. Vi kan vidare

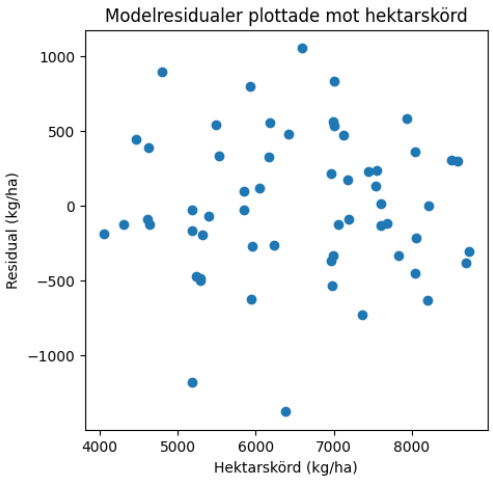
Tabell 5: Det n-faldiga korsvalideringsfelet för Jordbruksverkets modell. I avs. 2.4 beskrivs hur felen är sammanviktade.

Gröda	Absolut korsvalideringsfel (kg/ha)	Relativt korsvalideringsfel
Höstraps	459	21,6 %
Vårkorn	485	13,6 %
Blandsäd	391	13,1 %
Vårraps	457	35,2 %
Höstkorn	696	13,7 %
Höstrågvete	725	14,4 %
Höstråg	541	14,1 %
Vårvete	779	21,2 %
Höstvete	1255	31,2 %
Havre	564	20,0 %
Viktat genomsnitt	886	23,8 %

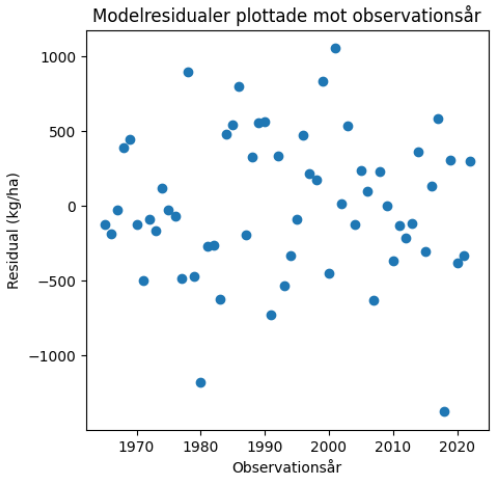
använda oss av hypotestest för att analysera alla 87 regressionsmodeller. Shapiro-Wilks-testet låter oss pröva huruvida residualerna är normalfördelade. Det beskrivs ingående i [9]. Noll-hypotesen av normalfördelning kunde endast förkastas för sju uppsättningar residualer med signifikans $p < 0{,}05$. Breusch-Pagan-testet, som beskrivs i [1], testar huruvida residualernas varians beror på en given oberoende variabel. Nollhypotesen att inget sådant beroende finns med de uppskattade hektarskördarna kunde förkastas 10 gånger ($p < 0{,}05$). Hypotesen av oberoende mot observationsåret kunde förkastas 9 gånger med samma signifikans. För alla tre tester förkastar förvisso nollhypotesen fler än $87 \times 0{,}05 = 4{,}35$ gånger, men inte tillräckligt ofta för att överge regressionsmodellens antagande.



(a)



(b)



(c)

Figur 2: Grafer över residualerna av Jordbruksverkets modell för hektarskörden av höstvete i Skåne. (a): Modellresidualerna på normalfördelningspapper. (b) och (c): Residualerna plottade mot den uppskattade hektarskörden respektive observationsåret.

3. Ett osäkerhetsmått

Jordbruksverket redovisar inget osäkerhetsmått för sin skördeprognos. Detta riskerar att ge en felaktig bild av modellens förutsägelser, särskilt om osäkerheten är stor. I detta avsnitt härleder jag därför prediktionsintervall för Jordbruksverkets modell. Eftersom regressionsmodeller passas för hektarskörden av vårkorn i alla län så kan vi räkna ut ett exakt prediktionsintervall för denna gröda. Sannolikhetsfördelningarna av de hektarskördar som istället skattas indirekt (redogjorda för i avs. 2.3 ovan) är däremot för komplicerade för att enkelt härleda exakta uttryck. Därför använder jag Monte Carlo-metodik för att beräkna simuleringsbaserade prediktionsintervall för dessa skördar.
Vi börjar med att härleda ett exakt prediktionsintervall för riksskörden S av vårkorn. Låt $S_\ell = a_\ell y_\ell$ vara den totala skörden av vårkorn i län ℓ, där a_ℓ är den areal som vårkorn odlas på i länet och y_ℓ hektarskörden av vårkorn. Regressionsmodellen antar att y_ℓ, och därmed också S_ℓ, är normalfördelad. Låt $\hat{S}_\ell = a_\ell \hat{y}_\ell$ vara en skattning av S_ℓ och $\hat{\sigma}^2_{S_\ell} = a_\ell^2 \hat{\sigma}^2_\ell$ dess skattade varians med n_ℓ frihetsgrader. Riksskörden S av vårkorn ges av $S = \sum_l S_l$ med skattat medelvärde $\hat{S} = \sum_\ell \hat{S}_\ell$ och skattad varians $\hat{\sigma}^2_S = \sum_\ell \hat{\sigma}^2_{S_\ell}$ med $n = \sum_\ell n_\ell$ frihetsgrader. 95 %-prediktionsintervallet för riksskörden av vårkorn är då

$$[\hat{S} + t_{0.025}(n)\,\hat{\sigma}_S, \hat{S} + t_{0.975}(n)\,\hat{\sigma}_S]$$

Med min implementering av Jordbruksverkets modell ger detta 95 %-prediktionsintervallet $[1065, 1135]$ (tusen ton) för riksskörden av vårkorn år 2023. Punktskattningen av riksskörden är 1120 tusen ton. För de grödor som vi inte kan konstruera regressionsmodeller för är den skattade hektarskörden en funktion av prediktioner (av de hektarskördar för vilka regressionsmodeller finns) som antags vara normalfördelade. Ta t.ex. hektarskörden a_ℓ av höstvete i län ℓ, som indirekt skattas med

$$a_l = \frac{b_l a_k}{b_k}$$

där a_k, b_ℓ, b_k är prognoser för andra grödor och/eller andra län. Det vore möjligt, om än omständligt, att härleda den exakta sannolikhetsfördelningen av a_ℓ. För grödor som skattas i senare steg (ex. de i steg tre i figur 1) vore detta däremot mycket svårt. Vi tar därför en genväg: vi har redan antagit att de hektarskördar som vi kan ställa upp regressionsmodeller för är normalfördelade; prognosen ger oss skattningar av deras medelvärde och varians. Vi kan nu simulera observationer av skördarna under det år vi vill prediktera, genom att tillämpa det indirekta skattningsförfarandet på dessa simulerade observationer. Detta ger oss för varje gröda i varje län en uppsättning observationer av hektarskörden. Tillsammans med de odlade arealerna det året låter detta oss beräkna riksskörden av varje gröda. De arealer som är 'ej regionbestämda' multipliceras med den genomsnittliga hektarskörden i riket för respektive gröda.

Låt $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$ vara en uppsättning simulerade observationer av riksskörden av en given gröda sorterad i stigande ordning. Om $n = 1000$ så är $[t_{(25)}, t_{(975)}]$ ett prediktionsintervall med approximativ konfidensgrad på 95 %. I tabell 6 redvisas sådana prediktionsintervall, simulerade med $n = 1000$, tillsammans med det exakta intervallet för vårkorn.

Vi ser att prediktionerna är rätt så säkra; den relativa bredden (bredden dividerad med punktskattningen) av prediktionsintervallen är mellan 5 och 10 %. Märk väl är att de simulerade observationerna tas från estimaterade sannolikhetsfördelingar. Om parametrarna är dåligt skattade, d.v.s. om modellerna ger dåliga prediktioner, eller de verkliga fördelningarna inte är normala så kommer dessa prediktionsintervall vara opålitliga. I avs. 2.4 såg vi däremot att vår data är förenlig med normalfördelningsantagandet.

Tabell 6: Prediktionsintervall med 95 % konfidensgrad och punktskattning för riksskörderna år 2023 enligt Jordbruksverkets modell. Prediktionsintervallet för vårkorn är exakt, medan resterande är simulationsbaserade med $n = 1000$. Alla skördar anges i tusentals ton. Den relativa bredden beräknas som bredden av konfidensintervallet dividerat med punktskattningen.

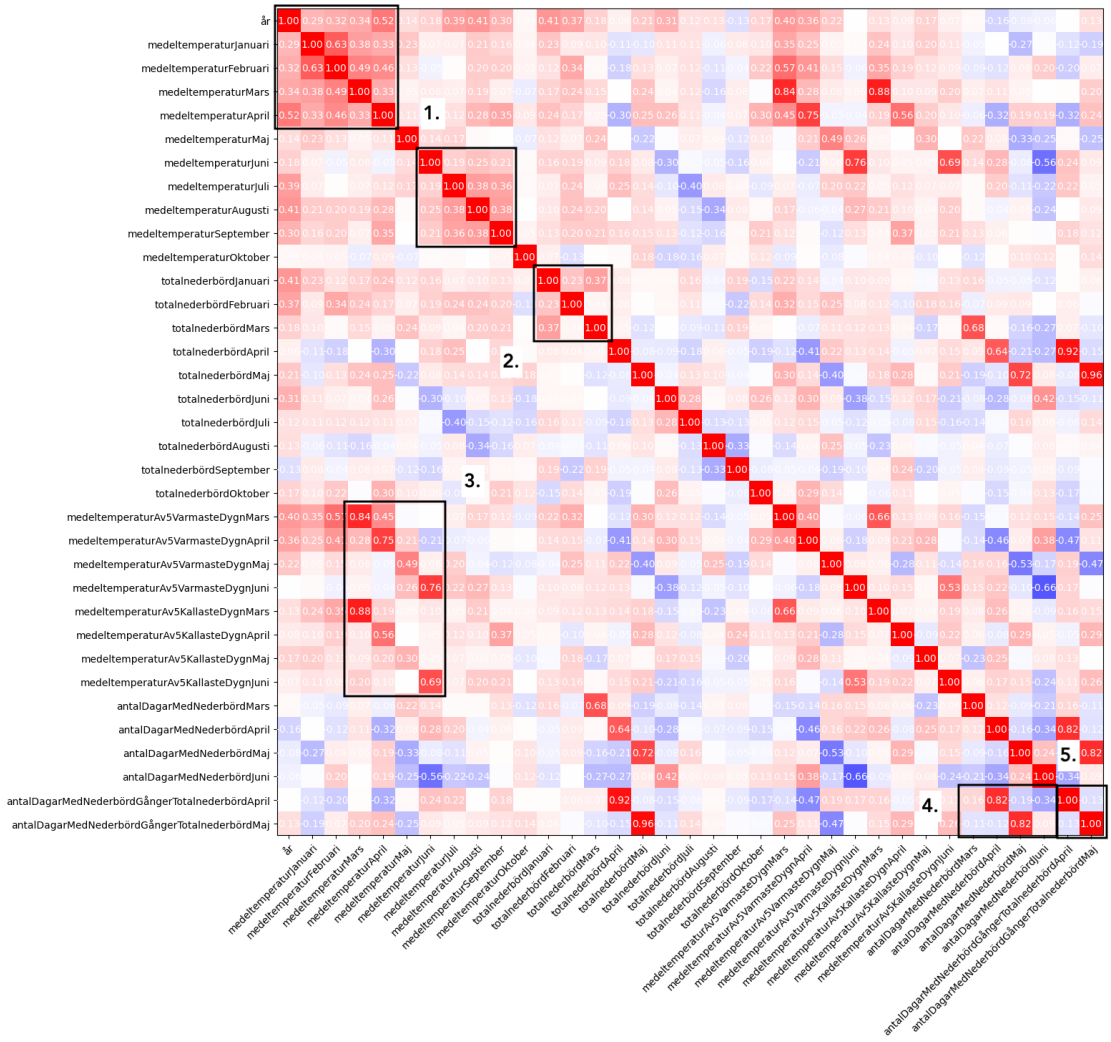
Gröda	Undre gräns	Punktskattning	Övre gräns	Relativ bredd
Vårkorn	1048	1101	1116	6,10 %
Höstvete	2814	2889	2963	5,13 %
Vårvete	173	183	194	11,2 %
Höstråg	145	151	158	8,25 %
Höstkorn	108	110	112	4,01 %
Havre	551	571	591	6,96 %
Blandsäd	20,8	21,3	21,8	4,39 %
Höstrågvete	148	152	157	6,29 %
Vårrågvete	7,1	7,3	7,5	5,57 %
Höstraps	380	397	414	8,40 %
Vårraps	22,3	23,8	25,5	13,3 %

4. Kolinearitetsproblem

Kolinearitet bland de oberoende variablerna riskerar att göra regressionsmodeller instabila. Små variationer i data kan ge stora förändringar i de skattade regressionskoefficenterna. Detta är framförallt ett problem i de fall där koefficenterna i sig är av huvudsakligt intresse. Modellens prediktionsförmåga skadas däremot också av kolineraitet, eftersom det reducerar rymden som de oberoende variablerna spänner upp. Därmed ökar sannolikheten att en prediktion faller utanför modellens observationsrymd, med osäkrare förutsägelser som följd. [7]

Det visar sig att bland de 35 oberoende variablerna (exkl. konstant) i ett och samma län är flertalet kraftigt kolineära i Jordbruksverkets modell. Figur 3 visar korrelationsmatrisen för de oberoende variablerna i Västmanland. Där går det att se tydliga mönster i hur variablerna kovarierar. Korrelationsmatriserna för andra län ser snarlika ut. Låt oss utifrån dessa korrelationsmatriser modifiera de oberoende variablerna för att minska kolinearitetet. Ingrep i de oberoende variablerna kan bestå av att medelvärdesbilda eller summera flera variabler. Eftersom medelvärdet är proportionerligt till summan är

dessa ekvivalenta; proportionalitetskonstanten kan förväntas tas upp av regressionskoefficenten. Vi kan också utesluta variabler från modellen helt och hållet. När korrelerade variabler anger väsentligt olika data (ex. medeltemperatur under olika månader) torde det vara passade att summera eller medelvärdesbilda, eftersom det tar vara på separata informationskällor. När två korrelerade variabler istället anger fysikaliskt lika storheter (ex. totalnederbörden och antal dagar med nederbörd > 0 mm.) representerar de snarlik information. I sådana fall anser jag det lämpligt att utesluta det ena av variablerna.



Figur 3: Korrelationsmatrisen för de oberoende variablerna i Västmanland. Avsaknaden av korrelation indikeras med ljusa nyanser. Det går att utläsa tydliga mönster i hur variablerna samvarierar. De delar av matrisen som motiverar ingrepp (1) - (5) är markerade och visas förstorat i figur 4.

Utifrån detta har jag identifierat fem ingrepp:

(1) Kombinera medeltemperaturerna för jan. t.o.m april och juli t.o.m aug. till medeltemperatur för vår respektive sommar.

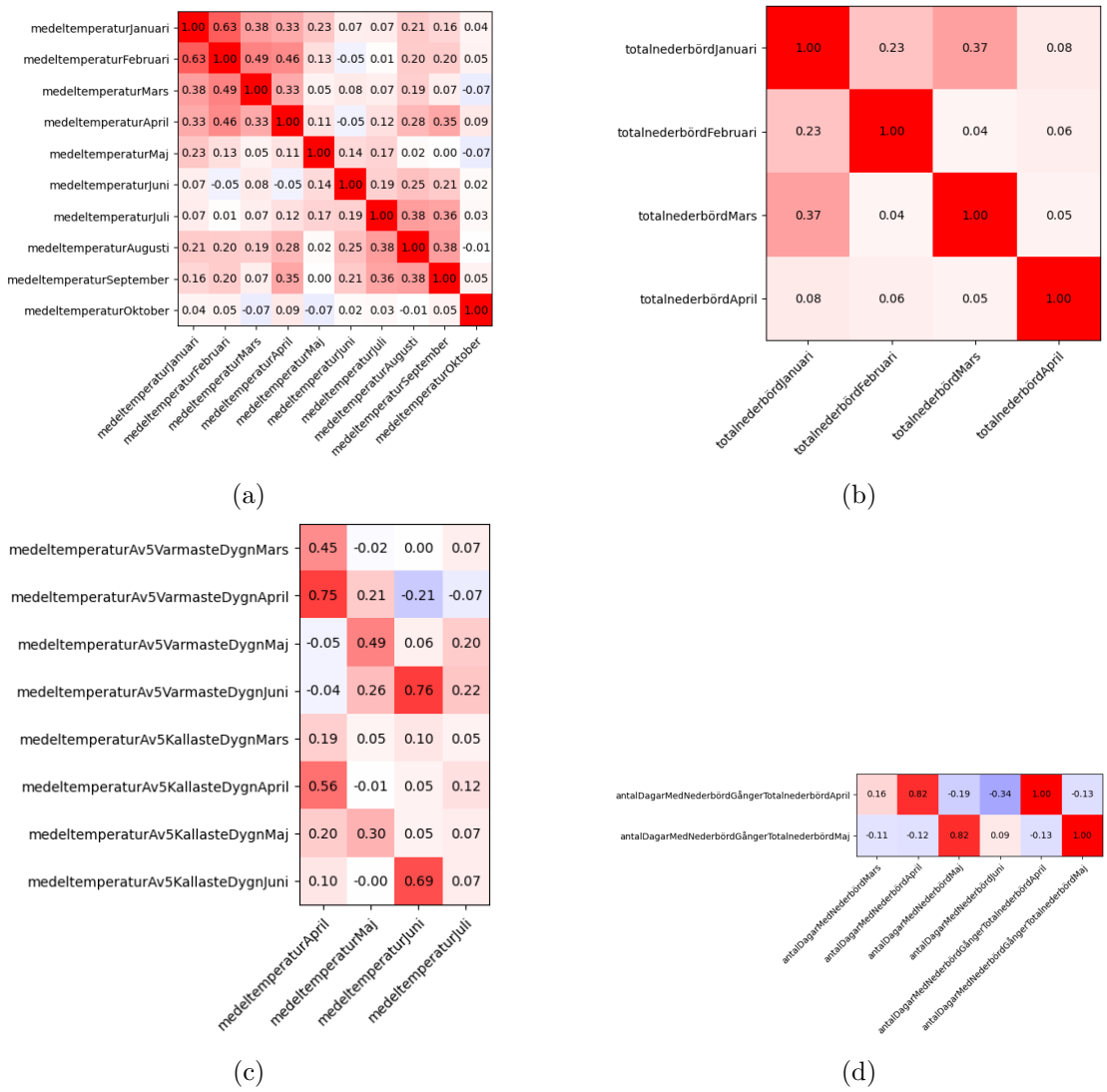
(2) Summera den totala nederbörden för jan., feb. och mars till total nederbörd för hela våren.

(3) Uteslut månadsmedelvärdena av de fem dygn med högst respektive lägst temperatur under mars t.o.m. juni.

(4) Uteslut månadsvärdena för antalet dagar med nederbörd > 0 mm under mars t.o.m. juni.

(5) Uteslut kombinationsvariablerna lika med total månadsnederbörd gånger antalet dagar med nederbörd under april och maj.

De delar av korrelationsmatrisen som motiverar dessa ingrepp visas förstorat i figur 4. Vi utvärderar modeller passade på en träningsmängd bestående av 85 % av all data behandlad med alla 32 möjliga kombinationer av dessa fem ingrepp. Resterande 15 % av data utgör en testmängd. Den indirekta skattningen görs på samma sätt som Jordbruksverkets modell; i de fall där ett färre antal oberoende variabler har möjliggjort direkt skattning m.h.a. bakåteliminering har detta gjorts. Den uppsättning oberoende variabler som ger lägst genomsnittligt korsvalideringsfel på träningsmängden, viktat så som beskrivs i avs. 2.4, ges av ingrepp (2), (3), (4) och (5) tillsammans. Dessa ingrepp ger i sin tur ett genomsnittligt testfel på 14,4 %. Jämför detta med det genomsnittliga korsvalideringsfelet på 23,8 % för Jordbruksverkets modell. Dessa ingrepp förbättrar alltså modellens generaliseringsförmåga markant, vilket pekar på att kolineariteten bland de oberoende variablerna störde regressionsmodellen nämnvärt. Testfelen för varje gröda redovisas i tabell 7. Ingreppen i de oberoende variablerna gör att modellen presterar bättre än Jordbruksverkets modell på alla grödor. Framförallt är prediktionsfelet för höstvete, som står för ca. hälften av den totala riksskörden, lägre.



Figur 4: Delar av korrelationsmatrisen för de oberoende variablerna i Västmanlands län. Figur (a) - (c) motiverar ingrepp (1) - (3) och figur (d) ingrepp (4) och (5).

Tabell 7: Testfel för modellerna passade på en datamängd utsatt för åtgärd (2), (3), (4) och (5), redogjorda för i avs. 4.

Gröda	Absolut testfel (kg/ha)	Relativt testfel
Höstraps	350	13,0 %
Vårkorn	502	14,9 %
Blandsäd	418	14,3 %
Vårraps	235	31,0 %
Höstkorn	620	12,8 %
Höstrågvete	741	15,6 %
Höstråg	359	9,79 %
Vårvete	580	16,7 %
Höstvete	593	13,9 %
Havre	488	15,4 %
Viktat genomsnitt	**536**	**14,4 %**

5. Regulariserad regression

Parametrarna $\boldsymbol{\beta}$ i Jordbruksverkets modell har hittills skattas genom att minimera kvadratfelet, d.v.s.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

Detta problem har som bekant en exakt lösning $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ om och endast om $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ är av full rang. Därför behövs initalt minst 37 datapunkter för att skatta de 36 parametrarna i Jordbruksverkets modell för en gröda i ett län. Med färre datapunkter än så går det inte att lösa det vanliga regressionsproblemet. Regulariserad regression låter oss passa modeller med färre datapunkter genom att lägga till en extra term jämte kvadratfelet och söka parametervektorn som minimerar summan av båda termerna. Detta nya regressionsproblemet har en lösning även om vi har få datapunkter att tillgå. Detta gör att vi i detta avsnitt kan prediktera också hektarskörden av vårrybs, vilket min implementering av Jordbruksverkets modell inte kunde. Däremot kan vi inte ge några prognoser för höstrybs och vårrågvete som vi helt saknar historiska hektarskördar för. Regulariserade regressionsmodeller är även mindre känsliga för kolinearitet än vanlig linjär regression. [3] Detta avsnitt redogör för två sorters regulariserade modeller: Ridge- och Lasso-regression, som straffar L^2- respektive L^1-normen av regressionskoefficenterna. Regulariserade regressionslösningar påverkas av att de oberoende variablerna skalas med en konstant. Därför normaliserar vi alla oberoende variabler till att ha medelvärde noll och varians ett.

5.1. Ridge-regression

Om vi istället för att minimera endast kvadratfelet låter

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

får vi Ridge-lösningen av regressionsproblemet. Termen $\lambda\|\boldsymbol{\beta}\|_2$ straffar stora koefficenter, med följd att lösningen krymper. Hur mycket den krymper beror på storleken av hyperparametern λ. Det går att visa att denna lösning är ekvivalent med att minimera kvadratfelet under begränsningen $\|\boldsymbol{\beta}\|^2 < c(\lambda)$, där $c(\lambda)$ är en funktion av λ. [3] Ridge-regressionsproblemet har den exakta lösningen $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$. Eftersom $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\mathbf{I})$ är av full rang så länge som $\boldsymbol{X}$ inehåller minst en observation ger ridge-regression alltid en entydig modell.

Låt oss använda ridge-regression för att modellera hektarskördarna av de tretton grödor som Jordbruksverkets skördeprognos behandlar. Detta låter oss undvika den indirekta skattningen som Jordbruksverket gör och borde vara mindre känslig för den kolinearitet som påvisats i föregående avsnitt. Varje modell har en hyperparameter λ; den väljs genom att optimera det viktade korsvalideringsfelet på en träningsmängd bestående av 85 % av all tillgänglig data. Testfelet på resterande data redogörs för i tabell 8. De optimala värdena på λ anges i appendix A. Det genomsnittliga felet på 16,1 % är markant bättre än det genomsnittliga korsvalideringsfelet för Jordbruksverkets modell på 23,0 %. Vi jämför här bara felen för hektarskördar av de grödor i de län som ridge-regressionen förutsäger. Detta innebär att vissa indirekta skattningar från Jordbruksverkets modell bortses från för att ge jämförbara mått.

Ridge-modellen skulle kunna passas på bara en datapunkt, men en sådan modell kan inte förväntas ge goda prognoser. Jag har endast passat modeller för hektarskördar av de grödor i de län för vilka vi har minst 10 datapunkter att tillgå. En lägre gräns hade tillåtit att modeller passades för fler hektarskördar, men dessa modeller kan förväntas ge sämre prediktioner. Historiska hektarskördar saknas för grödor som inte odlats i större utsträckning i respektive län. Det kan tänkas att en gröda skulle börja odlas i stor utsträckning i ett län där detta inte historiskt varit fallet och där det därför inte finns 10 datapunkter att tillgå. Att utesluta länet skulle då påverka prognosen av riksskörden nämnvärt. I sådant fall skulle en indirekt skattning kunna göras för dessa enstaka hektarskördar enligt den metod som redogörs för i avsnitt 2.3. I praktiken är det bara en enda gröda som denna gräns utesluter mer än 0,5 % av riksskörden av; 4 % av riksskörden av höstkorn faller bort.

5.2. Lasso

Ridge-regression straffar L^2-normen, med följden att stora koefficenter krymps kraftigt medan de nära noll knappt påverkas alls; styrkan av krympningseffekten är proportionell till kvadraten av koefficenten. Om vi istället straffar L^1-normen av $\boldsymbol{\beta}$ krymps alla koefficenter lika mycket, oavsett storlek. Effekten blir att många koefficenter blir lika med noll. Denna s.k. Lasso-regression (*Least absolute shrinkage and selection operator*) minimerar alltså modellen genom att kasta bort ett antal av de oberoende variablerna,

likt bakåteliminering gör. Formellt ges Lasso-modellen av lösningen på följande optimeringsproblem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Styrkan av krympningseffekten avgörs av hyperparametern λ. Detta optimeringsproblem saknar en exakt lösning om inte $\boldsymbol{X}$ är ortagonal och $\hat{\boldsymbol{\beta}}$ måste därför hittas med numeriska optimeringsmetoder. [3] Efter att lassoregressionen har gjort ett urval av oberoende variabler med nollskilda koefficenter passas en vanlig linjär regressionsmodell med endast dessa variabler; detta undviker att de nollskilda koefficenternas storlek, som också straffas, underskattas av modellen. [2]

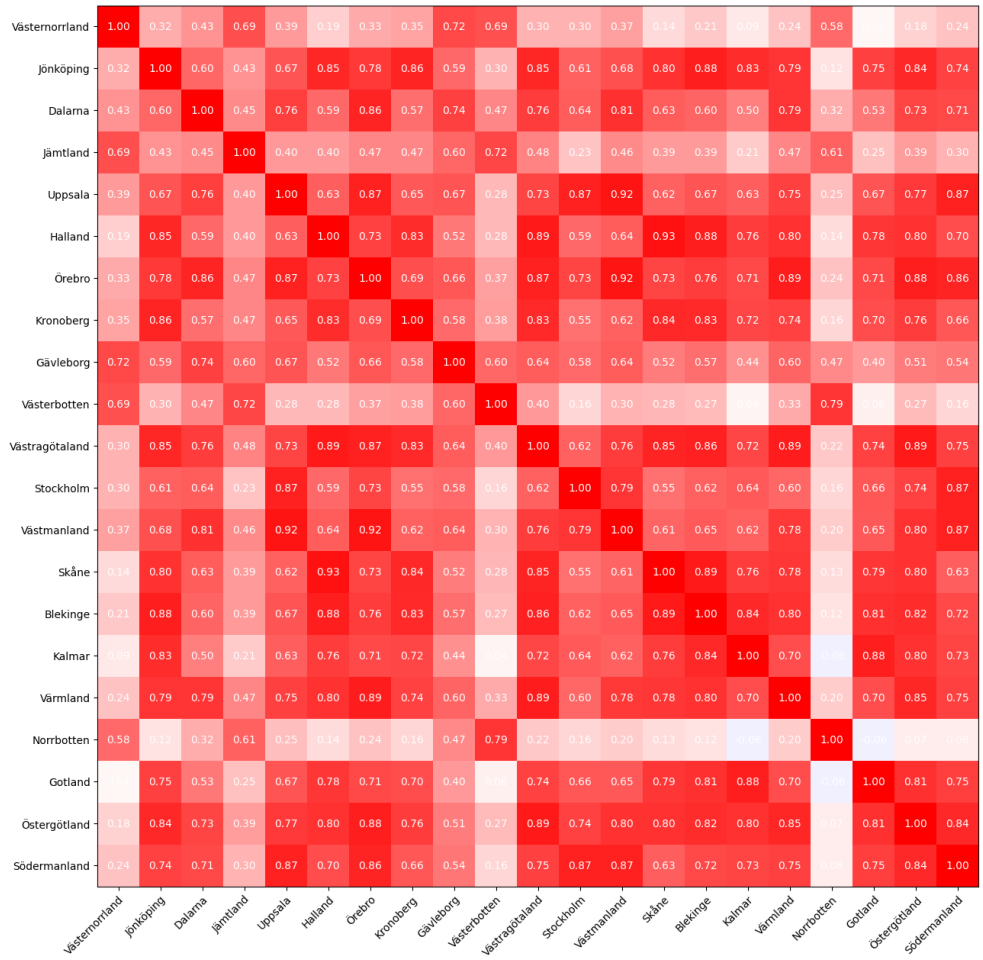
Lasso-regression är ytterligare ett sätt för oss att modellera hektarskörd. Liksom för ridge-regression begränsar vi oss till hektarskördar av de grödor i de län som vi har minst 10 datapunkter att tillgå för och hyperparametern λ för varje modell optimeras på en träningsmängd. Prediktionsfelet på testmängden anges i tabell 8. De optimala värde på λ och antalet nollskilda koefficenter i varje modell redogörs för i appendix B. Med ett genomsnittligt testfel på 14,7 % presterar Lasso-modellen bättre än både ridge-regression och Jordbruksverkets modell och ungefär lika bra som den bästa vanliga regressionsmodellen när kolineariteten mellan de oberoende variablerna har reducerats så som redogörs för i avs. 4.

Tabell 8: Testfel för ridge- och Lasso-modell med optimerade hyperparametrar, tillsammans med korsvalideringsfelet för Jordbruksverkets modell. Det sistnämnda är uträknat med bara de län och grödor som de regulariserade modellerna förutsäger, d.v.s. de med ≥ 10 datapunkter att tillgå. Motsvarande korsvalideringsfel för den bästa modellen med reducerad kolinearitet, redogjord för i avs. 4, anges inom parantes. Det viktade genomsnittsfelet är beräknat på det sätt som anges i avs. 2.4.

Gröda	Absolut korsvalideringsfel (kg/ha)	Relativt korsvalidersingsfel
	Ridge-modell	
Vårkorn	495	15,3 %
Höstkorn	751	15,3 %
Vårvete	570	17,9 %
Höstvete	590	15,5 %
Blandsäd	469	18,9 %
Höstråg	504	14,8 %
Havre	640	21,3 %
Höstrågvete	530	11,1 %
Vårraps	207	34,0 %
Höstraps	364	14,0 %
Vårrybs	215	17,9 %
Viktat genomsnitt	**557**	**16,1 %**
	Lasso-modell	
Vårkorn	446	13,5 %
Höstkorn	1085	21,68 %
Vårvete	629	18,7 %
Höstvete	573	13,8 %
Blandsäd	503	18,4 %
Höstråg	412	11,6 %
Havre	565	18,4 %
Höstrågvete	926	18,8 %
Vårraps	267	30,0 %
Höstraps	337	12,2 %
Vårrybs	276	23,0 %
Viktat genomsnitt	**544**	**14,7 %**
	Jordbruksverkets modell	
Vårkorn	485 (502)	13,6 % (14,9 %)
Höstkorn	640 (541)	13,0 % (12,4 %)
Vårvete	792 (536)	20,1 % (15,6 %)
Höstvete	1257 (583)	31,3 % (13,8 %)
Blandsäd	391 (418)	13,1 % (14,3 %)
Höstråg	530 (346)	13,7 % (9,41 %)
Havre	536 (488)	17,8 % (15,3 %)
Höstrågvete	710 (791)	14,0 % (16,5 %)
Vårraps	457 (235)	35,2 % (31,0 %)
Höstraps	454 (344)	20,2 % (12,6 %)
Vårrybs	—	—
Viktat genomsnitt	**868 (533)**	**23,0 % (14,3 %)**

6. En spatial modell

Försöken att modellera hektarskörd har alla varit begränsade av att vi har få datapunkter att tillgå. Detta tvingar Jordbruksverket att använda indirekt skattning och motiverade användningen av regulariserad regression. Detta problem uppstår eftersom hektarskörden av varje gröda och varje län modelleras var för sig. Det visar sig dock att hektarskördarna av samma gröda i olika län är kraftigt korrelerade. I figur 5 visas korrelationsmatrisen för hektarskördarna av vårkorn i Sveriges alla län. Notera att intilliggande län är särskilt starkt korrelerade. Det verkar därför lovande att modellera hektarskördar för angränsande län i *ensamble*, med delvis gemensamma parametrar. Då skulle prediktionen för varje län bygga på mångfaldigt fler data.



Figur 5: Korrelationsmatris för hektarskördarna av vårkorn i Sveriges alla län.

I detta avsnitt utforskar jag en spatial modell som modellerar interaktionseffekten mellan angränsande län. Jag begränsar modellen till endast hektarskörden av vårkorn i Götalands åtta län för att underlätta implementeringen. Jag modellerar hektarskördarna

$\boldsymbol{Y}_\ell$ av vårkorn i län ℓ som

$$\boldsymbol{Y}_\ell = \boldsymbol{X}_\ell\boldsymbol{\beta} + \boldsymbol{Z}_\ell\boldsymbol{\Delta}_\ell + \boldsymbol{\epsilon}_\ell$$

där

- $\boldsymbol{X}_\ell$ är observationer av samma 36 oberoende variabler (inkl. en konstant) som Jordbruksverkets modell använder.
- $\boldsymbol{\epsilon}_\ell \sim \text{NID}(\mathbf{0}, \sigma^2\mathbf{I})$ är slumpfel.
- $\boldsymbol{\beta}$ och $\boldsymbol{\Delta}_\ell$ är regressionskoefficenter.
- $\boldsymbol{Z}_\ell$ är ett elementvist medelvärde av observationerna av de oberoende variablerna i ℓ och angränsande län. D.v.s. att om S är mängden av de län som gränsar till län ℓ så är $\boldsymbol{Z}_\ell = (\boldsymbol{X}_\ell + \sum_{k\in S}\boldsymbol{X}_k)/|S|$.

På blockmatrisform är det tydligt att detta utgör en enda regressionsmodell:

$$\begin{bmatrix}\boldsymbol{Y}_1\\ \boldsymbol{Y}_2\\ \vdots\\ \boldsymbol{Y}_8\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{Z}_1 & \mathbf{0} & \dots & \mathbf{0}\\ \boldsymbol{X}_2 & \mathbf{0} & \boldsymbol{Z}_2 & \dots & \mathbf{0}\\ \vdots & & \ddots & & \vdots\\ \boldsymbol{X}_8 & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{Z}_8\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\Delta}_1\\ \vdots\\ \boldsymbol{\Delta}_8\end{bmatrix} + \begin{bmatrix}\boldsymbol{\epsilon}_1\\ \boldsymbol{\epsilon}_2\\ \vdots\\ \boldsymbol{\epsilon}_8\end{bmatrix}$$

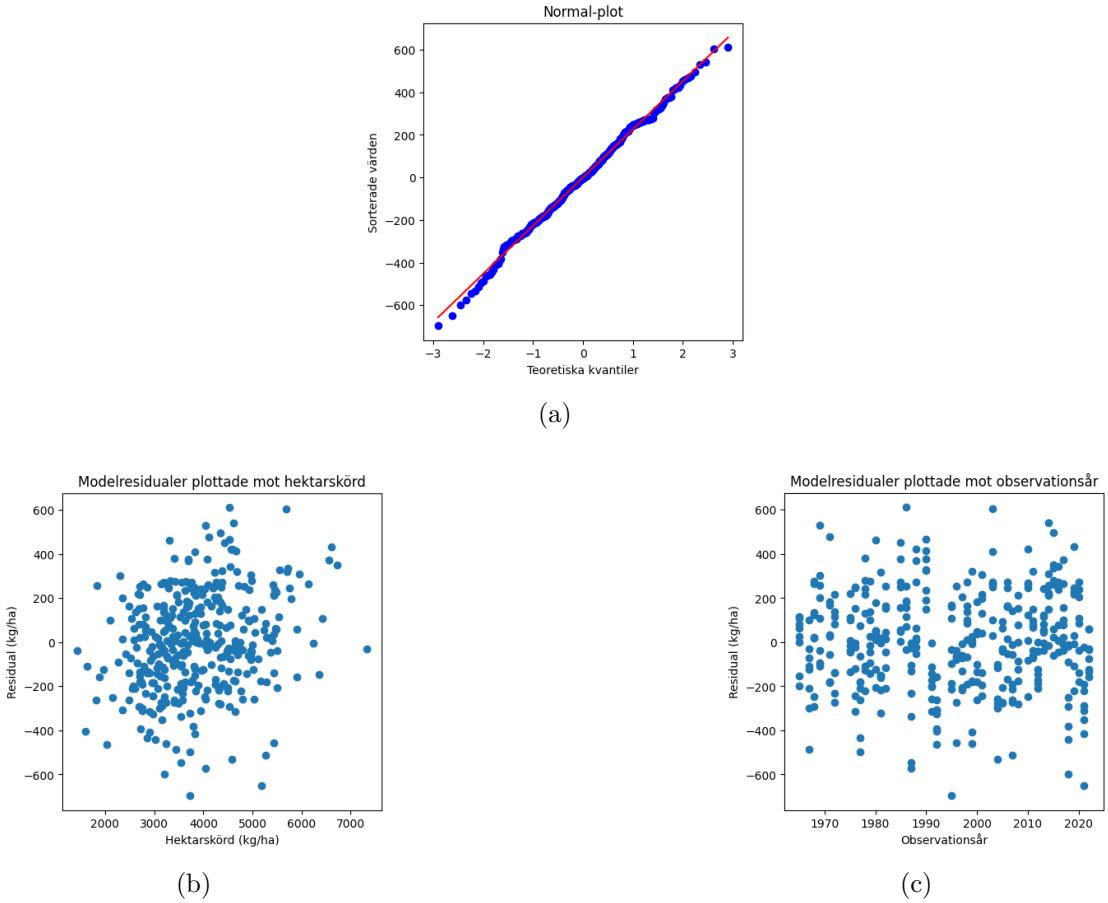
Märk väl att denna modell har $(1 + 8) \times 36 = 324$ parametrar och 464 datapunkter att tillgå. Modellen skulle alltså kunna passas med vanlig linjär regression, men för att förbättra modellens förmåga att generalisera lägger vi ett Lasso-straff på interaktionseffekterna $(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \dots, \boldsymbol{\Delta}_8)$. Resultatet är att endast ett par koefficenter är nollskilda i varje $\boldsymbol{\Delta}_\ell$ Hyperparametern λ för Lasso-regressionen optimeras på en träningsmängd bestående av 85 % av data. Utvärderingen visar att testfelet på resterande data är 9,05 %. Detta är lägre än de jämförbara testfelen för alla andra modeller som redogjorts för i denna uppsats, vilket framgår av tabell 9. Detta är lovande! Det återstår nu bara att validera modellantagandena. Normalplotten av modellresidualerna i figur 6a visar att de är ungefärligt normalfördelande. I figurerna 6b och 6c plottas residualerna mot

Tabell 9: Testfel för regulariserade modeller och spatial model. För Jorbruksverkets modell, med och utan reducerad kolinearitet, redovisas istället korsvalideringsfel. Alla fel är uträknade endast för vårkorn i Götalands åtta län.

Modell	Absolut fel (kg/ha)	Relativt fel
Jordbruksverkets	499	12,4 %
Jordbruksverkets med reducerad kolinearitet	512	13,9 %
Ridge-regression	510	13,9 %
Lasso	426	11,7 %
Spatial	332	9,05 %

den faktiska hektarskörden respektive observationsåret. De indikerar att residualernas varians inte uppvisar något beroende. Detta berättigar regressionsmodellens antaganden. Det börs märkas att det optimala parametervärdet $\lambda = 26{,}9$ innebär att alla Δ_ℓ innehåller ytterst få nollskilda koefficenter. Totalt bevarar Lasso-regressionen sex nollskilda parametervärden i $(\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2, \ldots, \boldsymbol{\Delta}_8)$. För restrerande fyra län bevaras inga. Det är anmärkningsvärt att modellen presterar bättre än alternativen även med så små interaktionseffekter.

Det vore i princip enkelt, om än tidskrävande, att låta modellen täcka hela landet. För att modelera andra grödor än vårkorn måste man lösa problemet att vi inte har data att tillgå för samma år i alla län. För hektarskörden av vårkorn har vi så pass många data att vi kan exkludera de år för vilka vi inte har historiska hektarskördar för alla län. Detta är en lovande inriktning för framtida arbete.



(a)

(b)

(c)

Figur 6: (a): Den spatiala modellens residualer på normalfördelningspapper. (b) och (c): Samma residualer plottade mot den faktiska hektarskörden respektive observationsåret.

7. Sammanfattning

I denna uppsats undersökte jag Jordbruksverkets skördeprognoser. Först undersökte prognosernas osäkerhet, vilket låter oss konstatera att modellen ger relativt säkra prediktioner. Bredden av 95 %-prediktionsintervall är 5–10 % av punktskattningens storlek.

Jag har också demonstrerat att den data som Jordbruksverket utgår ifrån uppvisar stor kolinearitet bland de oberoende variablerna. Detta skadar regressionsmodellernas prediktionskraft. Jag har visat att detta kan åtgärdas genom att heuristiskt modifiera regressionsmatrisen $\boldsymbol{X}$, vilket reducerar modellens prediktionsfel markant. Att bygga upp modellen genom framåtinkludering, istället för att reducera den med bakåteliminering, som Jordbruksverket gör, hade också mildrat effekten av kolinearitet. [8] Detta har däremot inte undersökts i denna uppsatts. Jag har istället använt regulariserad regression för att uppnå samma resultat.

Två sorters regularisering har utforskats: Ridge- och Lasso-regression, vilka reducerar storleken av regressionskoefficenterna, respektive gör ett urval av de oberoende variablerna. På denna datamängd presterar Lasso något bättre än ridge-regression och lika bra som Jordbruksverkets modell med reducerad kolinearitet.

I uppsatsens sista avsnitt presenterade jag en spatial modell som tillåter interaktionseffekter mellan angränsande län. Detta ger modellen mångfaldigt fler data. Denna modell ger bättre prediktioner av hektarskörden av vårkorn i Götaland än någon annan modell jag utvärderat. Märk väl att detta utgör 58 % av rikskörden av vårkorn och 13 % av den totala rikskörden av de tretton grödor som prognosen berör. Detta är lovande resultat! Ytterligare arbete borde läggas på att utveckla denna modell till att täcka hela landet och fler grödor. Det vore också möjligt att på liknande sätt modellera interaktionseffekter mellan grödor i samma län, men jag har inte undersökt detta. En modell som inkluderar interaktionseffekter mellan både olika grödor och olika län hade kunnat utnyttja all tillgänglig data. I sådant fall hade varje prediktion baserats på 10 till 100 gånger fler datapunkter. Detta förtjänar fortsatt uppmärksamhet.

Det är viktigt för både bönder och livsmedelsföretag att Jordbruksverkets skördeprognoser är så bra som möjligt. Jag har visat att prognosen kan förbättras, genom att reducera prediktionsfelet och att kvantifiera modellens osäkerhet. Jag hoppas att Jordbruksverket kommer överväga att implementera de metoder som jag föreslår i denna uppsatts.

Källförteckning

[1] Trevor S Breusch och Adrian R Pagan. "A simple test for heteroscedasticity and random coefficient variation". I: *Econometrica: Journal of the econometric society* (1979), s. 1287–1294.

[2] Evgenii Chzhen, Mohamed Hebiri och Joseph Salmon. "On Lasso refitting strategies". I: *Bernoulli* 25.4A (2019), s. 3175–3200.

[3] Trevor Hastie m. fl. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[4] Simon Lind. *Kvalitetsdeklaration. Skördeprognos för spannmål och oljeväxter*. Tekn. rapport JO0605. `https://jordbruksverket.se/download/18.a00253718a6314c97956882/1694078836712/JO0605_KD_2023-tga.pdf`, hämtad 13/12-2023. Statens jordbruksverk, aug. 2023.

[5] Simon Lind. *Skördeprognos för spannmål och oljeväxter 2023*. Tekn. rapport JO0605. `https://jordbruksverket.se/5.5baf3185189d2e2fc497d6c8.html`, hämtad 13/12-2023. Statens jordbruksverk, aug. 2023.

[6] Simon Lind. *Statistikens Framställning. Skördeprognos för spannmål och oljeväxter*. Tekn. rapport JO0605. `https://jordbruksverket.se/download/18.a00253718a6314c97956883/1694078836742/JO0605_STAF_2023-tga.pdf`, hämtad 13/12-2023. Statens jordbruksverk, aug. 2023.

[7] John O. Rawlings, Sastry G. Pantula och David A. Dickey. *Applied regression analysis: a research tool*. Springer, 1998.

[8] Ashish Sen och Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 2012.

[9] Samuel Sanford Shapiro och Martin B Wilk. "An analysis of variance test for normality (complete samples)". I: *Biometrika* 52.3-4 (1965), s. 591–611.

A. Ridge-modellernas hyperparametrar

Ridge-regression krymper storleken på regressionskoefficenterna. Styrkan av denna effekt bestäms av en hyperparameter λ, vald genom optimering på en träningsmängd. Dessa parametervärden för varje gröda i varje län anges i följande tabell. Ett streck (—) markerar att en ridge-model inte kunnat passa p.g.a avsaknad av data. Optimeringen av λ begränsades till intervallet $[0, 1000]$.

Län / Gröda	Vårrybs	Höstvete	Vårraps	Höstkorn	Höstråg	Vårkorn	Havre	Vårvete	Höstrågvete	Blandsäd	Höstraps
Blekinge	—	9,05	—	—	—	5,02	30,4	33,9	1000	—	9,36
Kalmar	—	4,60	—	65,1	8,36	16,8	18,4	10,3	174	0,00	15,9
Värmland	159	3,84	48,7	—	—	3,79	13,7	0,00	—	—	—
Östergötland	382	5,37	85,8	0,00	2,69	4,13	8,06	4,19	40,0	4,18	31,8
Jönköping	—	29,6	—	—	—	6,07	17,7	0,00	12,9	0,00	—
Norrbotten	—	—	—	—	—	70,1	0,0	—	—	—	—
Jämtland	—	—	—	—	—	12,3	—	—	—	—	—
Dalarna	35,0	33,7	—	—	—	63,8	154	69,0	—	—	—
Uppsala	173	52,2	103	—	71,7	59,4	30,0	217	—	155	5,45
Örebro	11,6	14,0	133	—	14,0	7,21	1,96	6,26	1000	—	0,00
Halland	—	12,5	1000	—	10,8	8,49	17,2	75,1	1000	0,00	39,8
Västernorrland	—	—	—	—	—	22,8	—	—	—	—	—
Västmanland	95,2	31,2	211	—	—	142	318	1000	—	1000	39,5
Kronoberg	—	—	—	—	—	6,08	13,0	1000	1000	—	—
Södermanland	370	0,462	55,3	—	54,4	73,8	34,9	88,1	0,00	35,9	0,551
Västerbotten	—	—	—	—	—	35,8	610	—	—	—	—
Skåne	—	5,82	95,2	1000	4,02	12,5	24,0	31,4	1000	—	49,6
Västragötaland	39,3	1,29	257	9,71	2,22	1,81	2,40	35,6	7,80	11,3	13,9
Gävleborg	—	—	—	—	—	20,6	33,3	376	—	—	—
Gotland	—	21,7	39,1	0,00	17,6	19,0	21,1	38,3	1000	—	63,6
Stockholm	537	22,9	37,0	—	710	59,5	18,2	278	—	—	0,198

B. Specifikationer för Lasso-modellerna

Vikten av Lasso-straffet bestäms av en hyperparameter λ. Det optimala parametervärdena, på en träningsmängd, anges i följande tabell. Där avsaknaden av data gjort att en Lasso-modell inte kunnat passas markeras detta med ett streck (—).

Län / Gröda	Vårrybs	Höstvete	Vårraps	Höstkorn	Höstråg	Vårkorn	Havre	Vårvete	Höstrågvete	Blandsäd	Höstraps
Blekinge	—	54,6	—	—	—	27,6	50,9	66,1	10,7	—	28,9
Kalmar	—	11,0	—	33,2	22,8	22,1	33,4	61,2	893	3,03	78,5
Värmland	32,9	24,1	28,3	—	—	13,0	9,04	16,9	—	—	—
Östergötland	—	13,8	102	8,92	36,1	42,0	25,5	6,80	30,3	28,2	55,0
Jönköping	—	35,6	—	—	—	12,3	28,1	11,3	53,0	28,9	—
Norrbotten	—	—	—	—	—	61,4	1,54	—	—	—	—
Jämtland	—	—	—	—	—	15,6	—	—	—	—	—
Dalarna	5,94	33,4	—	—	—	30,1	39,8	21,6	—	—	—
Uppsala	90,8	64,5	77,2	—	30,0	28,5	21,3	7,98	—	4,52	15,8
Örebro	34,2	40,3	73,9	—	2,26	78,9	20,0	68,4	164	—	4,26
Halland	—	26,7	11,7	—	67,6	31,3	24,4	33,1	29,7	10,7	56,5
Västernorrland	—	—	—	—	—	17,0	—	—	—	—	—
Västmanland	6,47	136	14,9	—	—	129	86,5	450	—	197	3110
Kronoberg	—	—	—	—	—	7,66	22,1	11,0	0,875	—	—
Södermanland	29,1	2,59	21,0	—	31,7	47,0	37,1	29,9	13,6	42,7	39,2
Västerbotten	—	—	—	—	—	32,0	32,0	—	—	—	—
Skåne	—	59,9	58,3	1130	78,2	57,3	49,0	53,2	48,5	—	52,7
Västragötaland	17,9	15,5	51,0	27,0	39,2	25,8	39,1	78,5	27,1	40,3	70,1
Gävleborg	—	—	—	—	—	17,7	46,4	35,5	—	—	—
Gotland	—	39,1	29,8	25,5	17,9	39,4	28,8	748	64,9	—	53,1
Stockholm	158	12,5	20,1	—	459	19,3	19,8	119	—	—	1,64

Lasso-regression gör ett urval av datamängdens 36 oberoende parametrarna (inkl. en konstant). Följande tabell anger hur många nollskilda parametrar som varje modell innehåller.

Län / Gröda	Vårrybs	Höstvete	Vårraps	Höstkorn	Höstråg	Vårkorn	Havre	Vårvete	Höstrågvete	Blandsäd	Höstraps
Dalarna	14	10	—	18	15	—	—	—	—	11	—
Norrbotten	10	—	—	—	—	—	—	—	—	—	—
Västerbotten	17	—	—	—	—	—	—	—	—	17	—
Västernorrland	20	—	—	—	—	—	—	—	—	—	—
Östergötland	11	26	17	23	—	9	12	13	13	20	3
Örebro	8	13	5	17	8	15	—	—	26	18	7
Jönköping	17	11	8	11	—	—	11	—	—	15	—
Halland	19	16	13	18	—	11	12	—	8	20	19
Jämtland	22	—	—	—	—	—	—	—	—	—	—
Skåne	17	17	11	13	—	6	—	—	10	18	6
Södermanland	14	14	20	34	12	13	9	—	12	15	13
Västragötaland	16	9	15	18	11	8	9	7	15	13	8
Stockholm	19	6	—	22	—	28	—	—	—	19	11
Västmanland	4	—	—	4	18	—	1	—	—	7	18
Värmland	23	16	—	21	6	—	—	—	—	23	13
Kalmar	18	13	—	27	—	5	11	17	18	17	—
Gotland	15	—	8	17	—	11	—	19	20	19	8
Gävleborg	17	9	—	—	—	—	—	—	—	12	—
Blekinge	17	9	14	13	—	13	—	—	—	13	—
Kronoberg	28	6	—	—	—	—	—	—	—	17	—
Uppsala	18	29	—	10	2	18	15	—	22	20	4

Bachelor's Theses in Mathematical Sciences 2024:K15
ISSN 1654-6229

LUNFMS-4074-2024

Matematisk statistik
Matematikcentrum
Lunds universitet
Box 118, 221 00 Lund

http://www.maths.lu.se/