Matematisk modellering

Statistisk dataanalys

Emmie Fahlström

December 2022

Innehållsförteckning

1. Beskrivning av data

Data som denna rapport grundar sig på är hämtad från SMHI [1] och innehåller lufttemperaturer för tre stycken olika väderstationer i Sverige. Lufttemperaturerna mäts i grader Celsius vid mätstationer en gång i timman och datan hämtas dynamiskt från SMHI:s "open data API" från dagens datum och ett antal månader bakåt i tiden. Datan sparas sedan till CSV-filer för att göra det enkelt att kunna exekvera koden och rita ut grafer och tabeller. Urvalet till denna rapport är begränsat till väderstationen i Gladhammar [2], väderstationen på Gotska Sandön [3] och väderstationen Målilla A [4]. Urvalet är baserat på att Gladhammar och Målilla ofta benämns i nyheter att ha slagit värmerekord [5], [6] och därav bidrar med en intressant vinkel att studera. Gotska Sandön är en ö som ligger på ostkusten, inte så långt ifrån de andra väderstationerna, och därav kan det vara givande att se om det finns några skillnader mellan den och mätstationerna på fastlandet. I datan som returneras från SMHI kan det i vissa fall vara saknad data på grund av att stationen eller givaren har varit ur funktion. De meteorologiska observationer som görs betygsätts i kvalitét med antingen ett "G" som innebär att värdena är kontrollerade och godkända, eller ett "Y" som innebär misstänkta eller aggregerade värden. "Value" i csv-filen förespråkar den uppmätta temperaturen och "date" datumet och tiden för den uppmätta temperaturen. Se tabell 1 för ett urval av data.

1.1 Metod

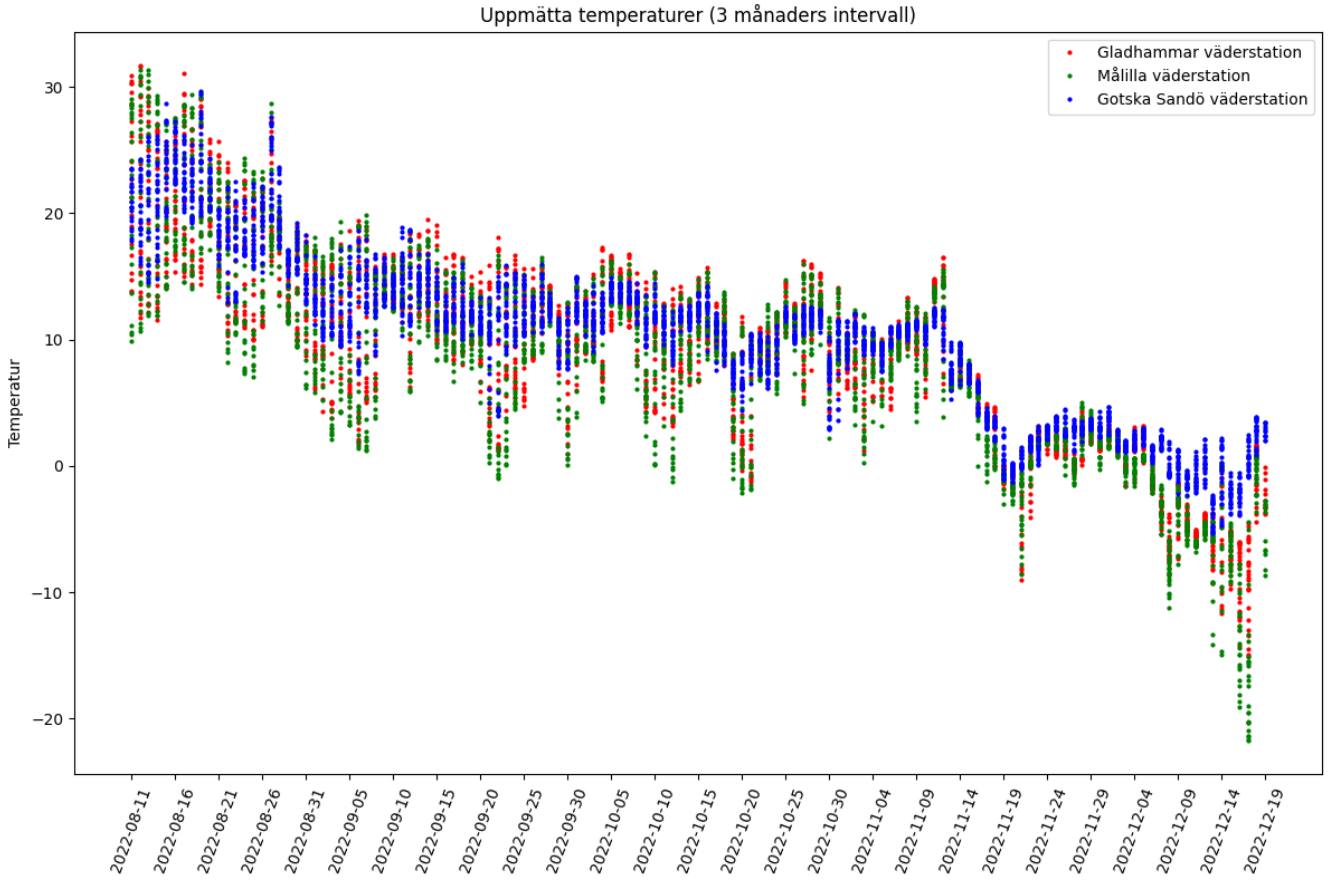
För att kunna analysera data via tabeller och grafer används Python och olika moduler såsom Pandas, Numpy, Matplotlib, Seaborn, Statsmodels. Koden finns som ett github-repo här där även varje diagram finns sparade. Dessa diagram kommer presenteras i denna rapport för att kunna analysera datan och dra slutsatser. Då datan innehåller datumobjekt som ska tas med i beräkningarna har dessa datum gjorts om till nummer via matplotlib.dates. Nummer 19215.125000 representerar exempelvis datumet 2022-08-11.

1.2 Urval

date	value	quality
2022-08-11 03:00:00	18.6	G
2022-08-11 04:00:00	17.8	G
2022-08-11 05:00:00	18.0	G
2022-08-11 06:00:00	18.8	G
2022-08-11 07:00:00	19.4	G

Tabell 1. Ett urval av data hämtad från mätstationen i Gotska Sandön.

1.3 Visuell representation av data



Figur 1. Diagrammet visar en visuell representation av hur datamängden ser ut. Färgerna representerar de olika mätstationer som går att se i högra hörnet högst upp och visar hur temperaturen ser ut mellan perioden 2022-08-11 till 2022-12-19. Som det går att se sjunker temperaturen relativt jämt i samband med datumen. Det finns några outliers för framför allt Målilla väderstation där "extremvärden" i temperaturen är uppmätta.

2. Beskrivande statistik

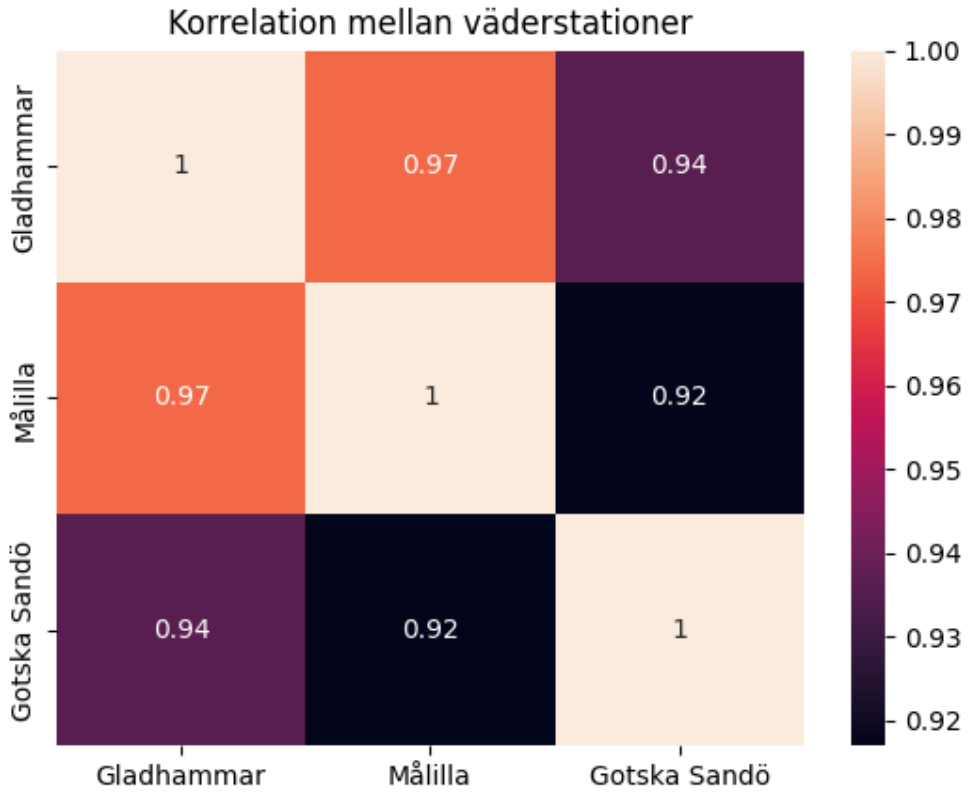
2.1 Tabell över data

	Gladhammar	Målilla	Gotska Sandö	Korr(Gldhmr-Mlla)	Korr(Gldhmr-Gtsk Snd)	Korr(Mlla-Gtsk Snd)
mean	9	8,45	10,21	0,97	0,94	0,92
std	7,71	8,15	6,52	0,97	0,94	0,92
min	−15	−21,7	−5,3	0,97	0,94	0,92
max	31,7	31,3	29,7	0,97	0,94	0,92

Tabell 2. Tabellen visar medelvärde, standardavvikelse, minsta- respektive högsta temperatur uppmätt på mätstationerna samt korrelationen mellan olika mätstationer.

I tabell 2 går det att se att det inte är stora skillnader i värden mellan de olika mätstationerna. En orsak kan bero på att de ligger relativt nära varandra geografiskt sett. Gladhammar och Målilla ligger på fastlandet med ungefär 75 kilometers avstånd från varandra och därav är korrelationen hög (0,97). Deras medelvärde, standardavvikelse och minsta och högsta värde är också ungefär likadana. Mätstationen på Gotska Sandö ligger som tidigare nämnt på en ö och lite längre bort från Målilla och Gladhammar, men fortfarande på samma kust. Det som är den största skillnaden är att Gotska Sandöns lägsta uppmätta temperatur är mycket lägre än både Gladhammar och Målilla. Korrelationen mellan Gotska Sandö och de andra är fortfarande relativt hög då den ligger på 0,94 och 0,92, men inte lika högt som 0,97.

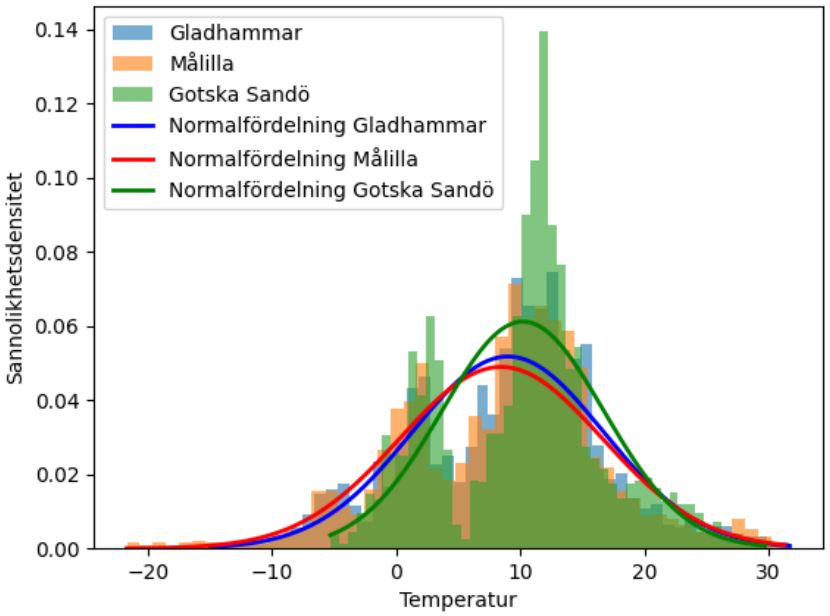
2.2 Heatmap



Figur 2. Diagrammet föreställer en heatmap över korrelationen över väderstationerna. Korrelationen är som starkast mellan Gladhammar och Målilla på 0.97 och lite svagare mellan Gladhammar och Gotska Sandö samt Målilla och Gotska Sandön.

3. Beskrivande histogram

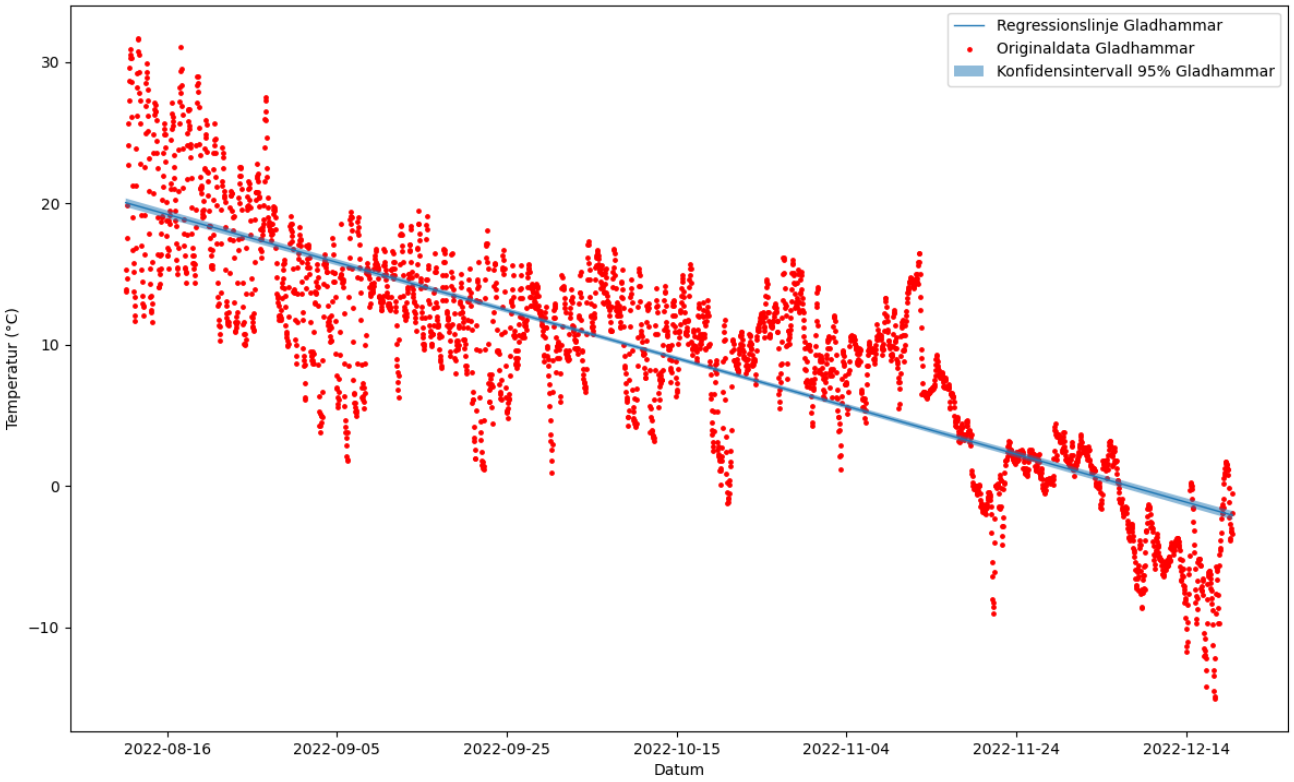
I figur 3 går det att se ett histogram över den uppmätta datan av temperaturer i respektive mätstation. Med hjälp av ett histogram går det att få en överblick hur vanligt förekommande datan är i varje intervall. Genom att lägga en normalfördelningskurva ovanpå histogrammet går det att se hur noga data i histogrammet matchar formen på normalfördelningskurvan som kan hjälpa till att avgöra om datan ungefär följer en normalfördelning eller om den avviker från normalfördelningen på något sätt.



Figur 3. Y-axeln i detta histogram visar sannolikhetsdensiteten, det vill säga hur sannolikt det är att det går att hitta ett visst värde i datan. Sannolikhetsdensiteten är en mängd per enhet och är en mått på hur koncentrerad data är kring ett visst värde. Ju högre sannolikhetsdensitet, desto mer koncentrerad är data kring det värdet.

Som figur 3 visar skiljer sig histogrammet en del från normalfördelningen. Speciellt histogrammet för Gotska Sandön som har betydligt högre staplar än normalfördelningskurvan på många ställen. Det skulle eventuellt kunna gå att göra ett antagande om att datan inte följer normalfördelningen enligt histogrammet, men för att kunna dra en slutsats krävs ytterligare statistiska test.

4. Linjär regression



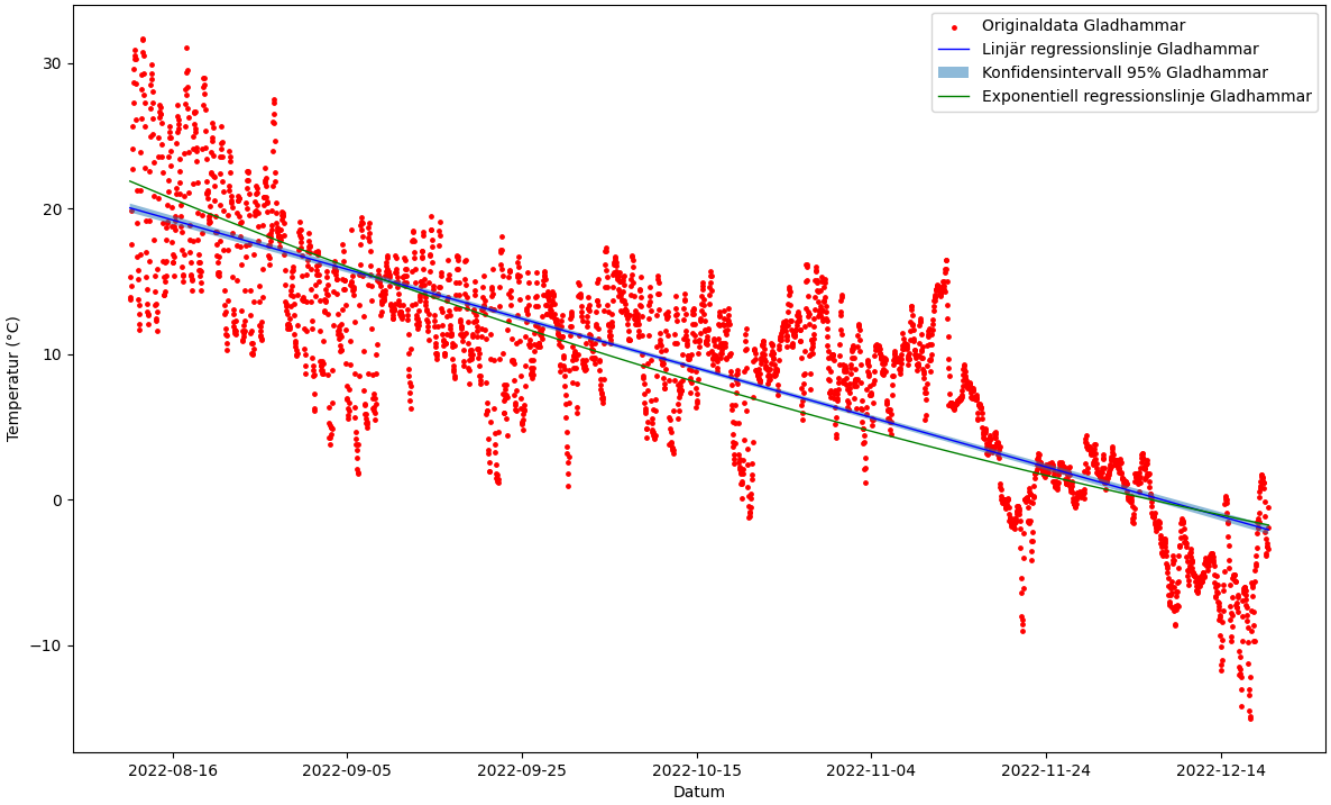
Figur 4. Linjär regression med tillhörande 95% konfidensintervall över data från Gladhammar väderstation.

	Gladhammar
a	3280,53
b	−0,17
y = a + b * x	y = 3280,53 − 0,17x
Konfidensintervall a	**Intercept**: 3202,83 **date**: -0,174
Konfidensintervall b	**Intercept**: 3358,234 **date**: -0,166

Tabell 3. Konfidensintervall för Gladhammar väderstation.

Då värdena på x-axeln innehåller datumobjekt har de omvandlats till nummer som för att kunna utföra beräkningar. Därav är siffrorna i intervallen höga för att kunna stämma överens med datumen som siffrorna representerar. Eftersom regressionskoefficienten b är minus innebär det att det finns en negativ linjär relation mellan de två variablerna y och x. Det vill säga, ju mer x ökar, desto mer minskar y vilket går att se i Figur 4.

5. Transformerade data

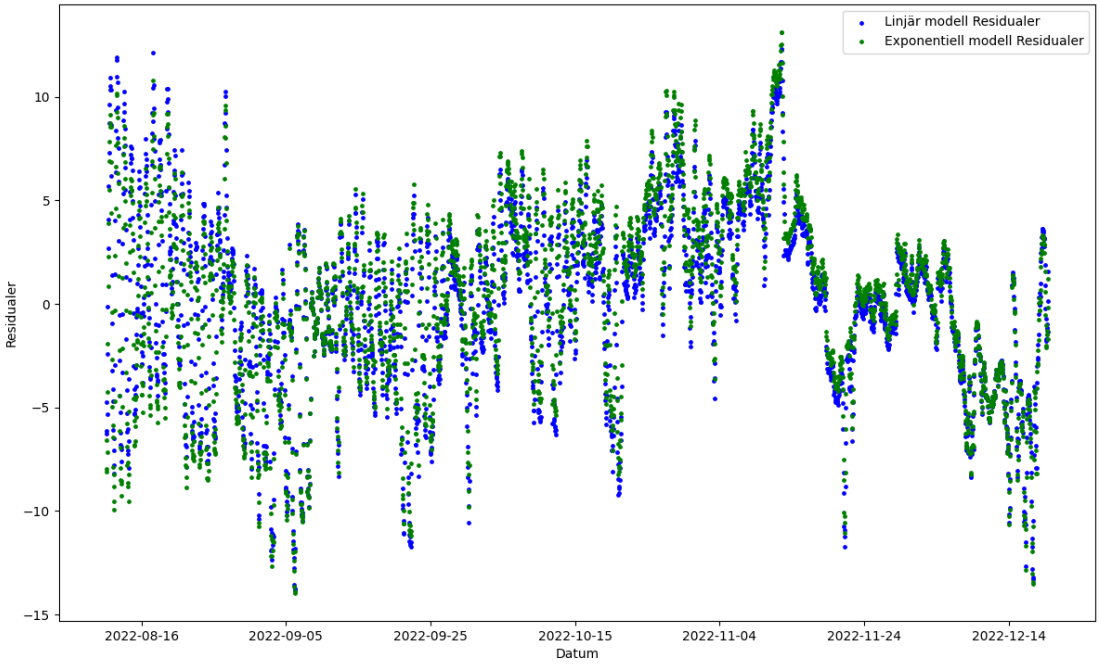


Figur 5. Linjär regression med tillhörande 95% konfidensintervall samt transformerad exponentiell regressionlinje från Gladhammar väderstation.

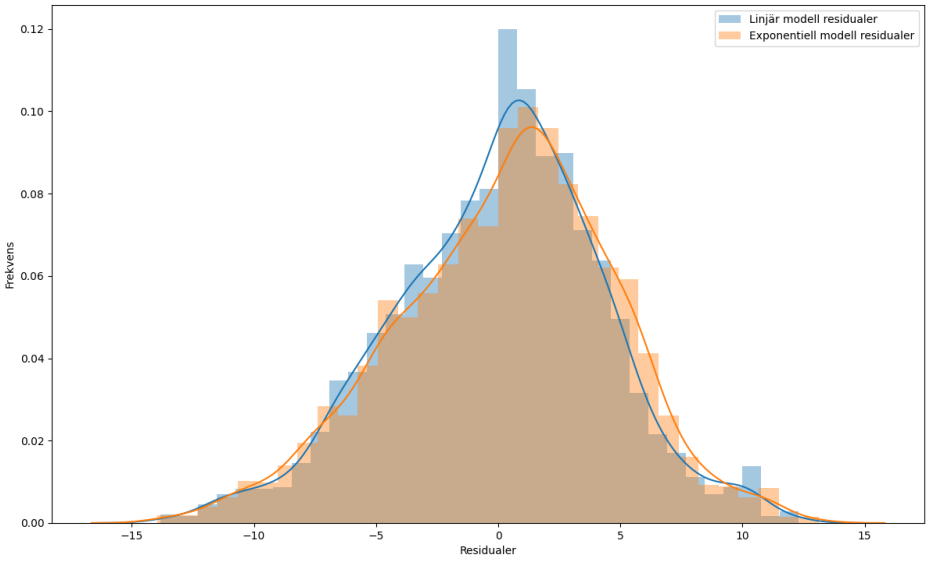
I figur 5 visas den linjära regressionslinjen tillsammans med en exponentiell regressionslinje. För att skapa den exponentiella regressionslinjen har först datan transformerats för att kunna få en annan form av data. En exponentiell modell är en modell som beskriver en relation mellan två variabler där en av variablerna är en exponentiell funktion av den andra variabeln. Med blotta ögat skulle det kunna gå att göra ett antagande om att temperaturen följer den linjära regressionsmodellen lite bättre än den exponentiella då originaldatan stämmer mer överens med den. En vidare jämförelse mellan dessa två olika modeller kommer vidareutvecklas i avsnitt 6.

6. Residualanalys

6.1 Visuell representation



Figur 6. Plottade residualer för den linjära och exponentiella modellen.



Figur 7. Normalfördelningskurvor över residualerna för den linjära och exponentiella modellen.

Residualanalys är en teknik som används för att undersöka om en statistisk modell passar väl till en given data. Som det går att se i figur 6 visas residualerna upp som plottar i diagrammet. Om residualerna ligger nära 0 kan det indikera att modellen i allmänhet passar väl till data. En modell med låga residualer kan också ha en låg varians, vilket kan indikera att modellen passar väl till data. Residualernas varians går att se i avsnitt 7.2. För både den linjära och exponentiella modellen finns det plottar som sticker iväg från värdet 0 men även en del som är koncentrerad runt det värdet. Det är svårt att dra en slutsats rakt ifrån bilden vilken som är den mest lämpliga modell att använda då det inte är en sådan stor markant skillnad mellan plottarna för modellerna. Även de båda residualerna följer normalfördelningen någorlunda bra. Det finns några värden som sticker iväg som kan bero på outliers i datan men annars följer normalfördelningen.

6.2 Residualernas varians

Varians	**Linjär data**	**Transformerad data**
Gladhammar	18,69	20,25

Tabell 4. Tabellen visar variansen för den linjära och den exponentiella modellen.

Som det går att se i tabell 4 är variansen lägre för den linjära datan än den transformerade. Om variansen är låg betyder det att residualerna är samlade kring en central täthet och har en liten spridning. Det kan innebära att det finns en stark relation mellan variablerna eller att modellen passar väl till data. Efter en sammanbedömning av figur 6 och 7 samt tabell 4 skulle det därav kunna gå att att göra ett antagande om att den linjära modellen är bättre att använda sig av än den exponentiella modellen i detta syfte.

7. Sammanfattning

Med hjälp av analysen skulle det kunna gå att dra följande slutsatser:

Det är inga märkvärdiga skillnader i temperaturer för de olika mätstationerna mer än en del extremvärden för Gladhammar och Målilla som hade betydligt högre lägsta temperatur än Gotska Sandön. Dock skiljer sig inte medelvärdet och standardavvikelserna speciellt mycket mellan stationerna som det går att se i tabell 2. Eftersom Målilla och Gladhammar hade högre extremvärden än Gotska Sandön bidrar det till att korrelationen mellan de två väderstationerna är högre. Dock går det att dra slutsatsen att korrelationen är hög mellan alla tre väderstationer eftersom de ligger nära 1. En orsak kan vara att de ligger nära varandra geografiskt sett och därav är sambandet i temperaturer större. Att Gotska Sandön skiljer sig lite grann mot Gladhammar och Målilla kan bero på att det är en ö. Om en mätstation hade legat i högst upp i norra Sverige och den andra längst ner i södra Sverige kanske resultaten hade sett annorlunda ut, vilket kan vara en intressant vinkel för framtida studier. Att den linjära modellen är att föredra än den exponentiella modellen i mätning av temperaturer kan vara givet då temperaturen stiger eller ökar i takt med årstider vilket inte sker exponentiellt. Ett bevis av det går att se visuellt i figur 4.

Referenslista

[1] SMHI, "Ladda ner meteorologiska observationer" 2022. [Online]. Tillgänglig: https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=airtemperatureInstant,stations=core (hämtad: 2022-12-09).

[2] SMHI, "Gladhammar A" 2022. [Online]. Tillgänglig: https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=airtemperatureInstant,stations=core,stationid=76420 (hämtad: 2022-12-09).

[3] SMHI, "Gotska Sandön A" 2022. [Online]. Tillgänglig: https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=airtemperatureInstant,stations=core,stationid=89230 (hämtad: 2022-12-09).

[4] SMHI, "Målilla A" 2022. [Online]. Tillgänglig: https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=airtemperatureInstant,stations=core,stationid=75250 (hämtad: 2022-12-09).

[5] Svt, "Nytt värmerekord i Småland: 16,7 grader" 2022. [Online]. Tillgänglig: https://www.svt.se/nyheter/lokalt/smaland/varmerekord-16-7-grader-i-smaland (hämtad: 2022-12-09).

[6] SMHI, "Svenska temperaturrekord" 2022. [Online]. Tillgänglig: https://www.smhi.se/kunskapsbanken/meteorologi/svenska-temperaturrekord (hämtad: 2022-12-09).