



Stockholms  
universitet

# Analys av PaiCo-metoden

Alexander Ketonen

Kandidatuppsats 2014:12  
Matematisk statistik  
Oktober 2014

[www.math.su.se](http://www.math.su.se)

Matematisk statistik  
Matematiska institutionen  
Stockholms universitet  
106 91 Stockholm

# Analys av PaiCo-metoden

Alexander Ketonen\*

Oktober 2014

## Sammanfattning

I den här rapporten analyserar vi två olika tillvägagångsätt på den nya metoden som används för att rekonstruera klimatet (temperaturen) från förr, med hjälp av proxies. En proxy kan vara data från trädringar eller från borrhärdar från isen. Datasetet består av simulerade temperaturer, över en period om tusen år, från en klimatmodell. Metoden är Pairwise Comparison (PaiCo), och de två tillvägagångsätten är Hanhijärvis metod och Brattströms metod. För att kunna använda PaiCo måste vi sortera varje proxy-serie i storleksordning. När vi försöker återskapa kurvan från vår givna tidsserie, använder vi en Quasi-Newtonmetod (BFGS). Dess likelihood funktion för Hanhijärvis metod bygger på produkten av alla observerade olikheter i vår sorterade data. Brattströms metod använder endast olikheter i den sorterade data som ligger intill varandra. Det visar sig att vi har en signifikant skillnad för de två tillvägagångsätten när vi jämför dem med hjälp av Spearmans rangkorrelation. Men eftersom Hanhijärvis metod tar längre tid för att rekonstruera en kurva, än tiden som Brattströms metod tar, är det önskvärt att man gör en större analys än vad som har gjorts i det här arbetet, för att kunna säga att en metod är bättre än den andra.

---

\*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.  
E-post: alexander.ketonen@hotmail.com. Handledare: Gudrun Brattström.

## Abstract

In this paper we examine two different approaches of the new method to reconstruct the climate (temperature) from the past, by the help of proxies. A proxy can be data from a tree ring or from an icecore. The dataset contain simulated temperatures, over an period of thousand years, from an climate model. The method is called Pairwise Comparision (PaiCo), and the two approaches are Hanhijärvi's method and Brattström's method. In order to use PaiCo we have to sort each proxy-serie in decreasing order. When we reconstruct the curve of our given timeserie, we use a Quasi-Newton-method (BFGS). The likelihood function for Hanhijärvi's method is built on the product of all observed differences in the sorted data. Brattström's method use only the observed differences in the sorted data that are next to each other. We find that there is a significant difference between the two approaches, when we compare them by the help of Spearman correlation. But since Hanhijäriv's method takes longer time to reconstruct the curve, than the time required by Brattström's method, it's desired to extend the analysis of PaiCo in order to say that one approach is better than the other one.

## Förord

Först och främst så vill jag tacka min handledare Gudrun Brattström för all hjälp som jag har fått. Utan hennes hjälp hade det här arbetet inte varit möjligt. Jag vill även tack Ida Hed Myrberg, som har läst igenom mitt arbete och kommit med värdefulla tips.

# Innehåll

<b>1</b>	<b>Introduktion</b>	<b>5</b>
1.1	Syfte . . . . .	5
<b>2</b>	<b>Beskrivning av data</b>	<b>5</b>
<b>3</b>	<b>Teori</b>	<b>6</b>
3.1	Maximum likelihood . . . . .	6
3.2	Quasi-Newton-metod . . . . .	7
3.3	Spearman's rangkorrelation . . . . .	9
3.4	Teckenrangtest . . . . .	9
<b>4</b>	<b>Metod</b>	<b>10</b>
4.1	Metod 1 . . . . .	11
4.2	Metod 2 . . . . .	12
<b>5</b>	<b>Statistisk modellering och analys av data</b>	<b>13</b>
5.1	Brusets varians = $0.1 \cdot \sigma_f^2$ . . . . .	15
5.1.1	Startvärde 1 . . . . .	16
5.1.2	Startvärde 2 . . . . .	18
5.1.3	Startvärde 3 . . . . .	19
5.2	Brusets varians = $0.9 \cdot \sigma_f^2$ . . . . .	20
5.2.1	Startvärde 1 . . . . .	21
5.2.2	Startvärde 2 . . . . .	23
5.2.3	Startvärde 3 . . . . .	24
5.3	Brusets varians = $4 \cdot \sigma_f^2$ . . . . .	25
5.3.1	Startvärde 1 . . . . .	26
5.3.2	Startvärde 2 . . . . .	28
5.3.3	Startvärde 3 . . . . .	29
5.4	Brusets varians = $10 \cdot \sigma_f^2$ . . . . .	30
5.4.1	Startvärde 1 . . . . .	31
5.4.2	Startvärde 2 . . . . .	32
5.4.3	Startvärde 3 . . . . .	34
<b>6</b>	<b>Diskussion</b>	<b>36</b>
	<b>Referenser</b>	<b>38</b>

# 1 Introduktion

På senare år har klimatet fått ett allt större utrymme i samhällsdebatten, och för att få en bättre uppfattning om vilket klimat vi kan vänta oss under det här århundradet kan det vara av intresse att studera klimatet från förr. Paleoklimatologi är läran om forna tiders klimat på jorden. Oftast är det temperaturen som man vill rekonstruera, men ibland kan det även vara av intresse att rekonstruera nederbörd och lufttryck från förr.

Ett sätt som används idag för att rekonstruera klimatet från förr, är med hjälp av tidsserieanalys av diverse proxies. En proxy kan vara data från till exempel trädringar eller borrhärdor från isen. Ett vanligt antagande är att temperaturen beror linjärt av de uppmätta proxyvärdena, det är även vad dagens metoder för rekonstruktion av temperaturen kräver. En nyligen framtagen metod, som är mindre känslig för avvikelser från linjäritet, är Pairwise Comparison (PaiCo). PaiCo antar att temperaturen är monotont ökande mot de uppmätta proxyvärdena, och utnyttjar endast ordinala relationer inom varje tidsserie.

Då klimatet har en stor inverkan på allt liv här på jorden, är det av stort intresse att finna nya och mer tidseffektiva metoder till bland annat forskningen inom paleoklimatologin. Eftersom andra forskningsområden kan få mer resurser, då det går snabbare att rekonstruera klimatet. Förhoppningen är att det visar sig att PaiCo-metoden är förhållandevis bra på att rekonstruera klimatet från förr.

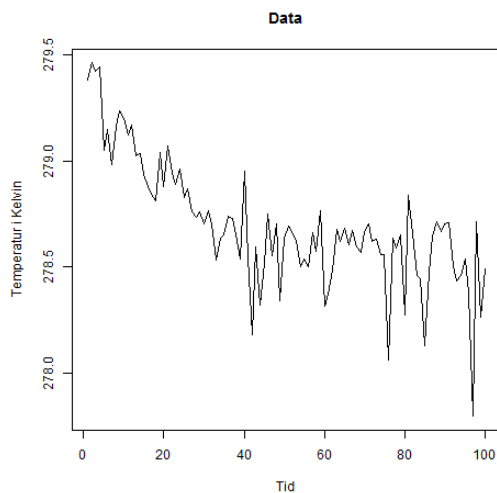
## 1.1 Syfte

Frågor som är av intresse är hur mycket information metoden tar till vara på, hur känslig den är för små förändringar i indata samt vad den ger för resultat då indata endast består av slumpmässigt brus (ingen information alls). Vi kommer försöka besvara dessa frågor i det här arbetet. Förhoppningen är att resultatet av det här arbetet kan vara till hjälp för personer som arbetar med klimatfrågor, och att läsaren kan få idéer om hur man kan bygga vidare på analysen.

## 2 Beskrivning av data

Data som vi har att arbeta med är simulerade temperaturer, i Kelvin, från perioden år 1000 till och med år 2001. Simuleringen utfördes på Goddard Institute for Space Studies, GISS, se [5]. Temperaturerna är skattade med någon av de äldre metoderna för att rekonstruera temperaturerna från förr. Vilken metod som har använts har ingen betydelse i det här arbetet. Eftersom temperaturvektorn är lång, väljer vi att bilda 10års-medelvärden och får då vektorn  $\mathbf{f} = (f_1, f_2, \dots, f_{100})$ .  $f_1$  är alltså medelvärdet av temperaturerna från januari 1000 till och med december 1009 och så vidare. Dock

förlorar vi data från år 2001, men det gör ingen skillnad i det här arbetet. Elementen i vektorn  $\mathbf{f}$  kommer vi se som de sanna värdena. I figuren nedan ser vi en tidsserie av  $\mathbf{f}$ .



Figur 1: Figuren visar tidsserien av den sanna temperaturvektorn.

Under anysen kommer vi arbeta med störda versioner av  $\mathbf{f}$ . Detta beskriver vi i sektion 4.

### 3 Teori

Innan vi presenterar teorin som vi kommer använda för att analysera PaiCo-metoden, vill vi illustrera vad vi gör med de störda vektorerna med ett exempel som vi senare kommer referera till.

#### *Exempel*

Antag att vi har en störd vektor  $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = (1.2, 1.5, 1.0, 0.9, 2.1, 3.0, 0.5, 1.6)$ . Vi ordnar dessa i storleksordning; störst, näst störst och så vidare. Vi får då

$$x_6 > x_5 > x_8 > x_2 > x_1 > x_3 > x_4 > x_7,$$

Den störda vektorn innehåller kontinuerlig data. Men efter att vi ordnar dess element i storleksordning, har vi ordinal information (diskret data).

#### 3.1 Maximum likelihood

Vanligtvis när man arbetar med maximum likelihood (ML) så använder man täthetsfunktion, (1), för det kontinuerliga fallet. Men eftersom vi rangordnar



elementen i de störda vektorerna, kan vi definiera likelihooden som (2). Detta genom att genomföra parvisa jämförelser mellan de observerade värdena, se [2] (sida 2044-2045). Vi antar att  $X_i \sim N(f_i, \sigma^2)$ .

$$L(\theta, x) = \prod_{i=1}^n f(x|\theta), \quad (1)$$

$$\begin{aligned} Pr(\{\mathbf{x}_i\}|\mathbf{f}) &= \prod_{i,j:\mathbf{x}_i < \mathbf{x}_j} Pr(X_i < X_j) = \\ &= \prod_{i,j:\mathbf{x}_i < \mathbf{x}_j} Pr(\mathbf{Y}_i \cdot (\mathbf{f} + \mathbf{Z}) < \mathbf{Y}_j \cdot (\mathbf{f} + \mathbf{Z})) = \\ &= \prod_{i,j:\mathbf{x}_i < \mathbf{x}_j} \Phi((\mathbf{Y}_j - \mathbf{Y}_i) \cdot \mathbf{f}, 0, \sigma_{ij}^2), \end{aligned} \quad (2)$$

där  $\Phi(y, 0, \sigma^2)$  är den kumulativa normalfördelningsfunktionen med väntevärde noll och varians  $\sigma^2$ , och  $\sigma_{ij}^2 = (\mathbf{Y}_j - \mathbf{Y}_i) \Sigma (\mathbf{Y}_j - \mathbf{Y}_i)$ . Kovariansmatrisen,  $\Sigma$ , är en diagonalmatris för bruset med elementen  $\sigma_{\mathbf{f}}^2$ . Varje vektor  $\mathbf{Y}_i$  innehåller en etta och resten nollor, där ettans position är sådan att  $\mathbf{Y}_i \cdot (\mathbf{f} + \mathbf{Z}) = X_i$ .

När vi sedan försöker skatta kurvan, noterar vi att vi har ett skattningsproblem. Hessian matrisen är, eller nästan är, singulär. För att komma runt problemet lägger vi till en liten regulariseringsterm för  $\mathbf{f}$ , se [2] (sida 2045). Vi regulariserar med tätheten för den multivariata normalfördelningen med väntevärdesvektor noll och kovariansmatris  $\Sigma_0$ . Där kovariansmatrisen är en diagonalmatris med samtliga element satta till ett litet tal som vi väljer till fem. Vi vet inte vilket tal som ska vara i diagonalen, men fem är litet i förhållande till elementen i  $\mathbf{f}$ .

### 3.2 Quasi-Newton-metod

Bakgrunden till de olika Quasi-Newton-metoderna är den klassiska Newton-metoden. I det här arbetet använder vi oss av Broyden-Fletcher-Goldfarb-Shanno-metoden (BFGS). Men vi börjar med att beskriva den klassiska metoden som utvecklas med hjälp av taylorutveckling. Metoden steepest-descent är en första ordningens metod, eftersom den är en linjär approximation av taylorserien. Newton-metoden är en andra ordningens metod. Vi antar därför att följande approximation genom taylorutveckling av andra ordningen håller

$$f(\mathbf{x} + \boldsymbol{\delta}) \approx f(\mathbf{x}) + \sum_{i=1}^n \frac{\partial f}{\partial x_i} \delta_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j} \delta_i \delta_j, \quad (3)$$

där  $\boldsymbol{\delta}$  är en förändring i  $\mathbf{x}$ . Om inget annat nämns är  $n$  dimensionen av  $\mathbf{x}$ . Derivering av  $f(\mathbf{x} + \boldsymbol{\delta})$  med avseende på  $\delta_k$ , för  $k=1,2,\dots,n$ , ger oss de

värden på  $\delta_k$  som minimerar  $f(\mathbf{x} + \delta)$  då vi sätter derivatan till noll. Det ger följande

$$\frac{\partial f}{\partial x_k} + \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i \partial x_k} \delta_i = 0 \quad , \quad k = 1, 2, \dots, n,$$

eller på matris form

$$\mathbf{g} = -\mathbf{H}\delta,$$

där  $\mathbf{H}$  är hessian matrisen.  $\mathbf{H}$  är en matris som består av andraderivatorna från uttrycket ovan, och  $\mathbf{g}$  är en vektor med förstaderivator från uttrycket ovan. Den optimala förändringen i  $\mathbf{x}$  är

$$\delta = -\mathbf{H}^{-1}\mathbf{g}.$$

Lösningen håller endast om och endast om hessian matrisen inte är singular, samt att (3) håller. Det här tillvägagångssättet kan användas genom att sätta  $\mathbf{x}_{k+1}$  till:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \delta_k = \mathbf{x}_k + \alpha_k \mathbf{d}_k,$$

där

$$\mathbf{d}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k,$$

och  $\alpha_k$  är det värde av  $\alpha$  som minimerar  $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ . Den intresserade läsaren hänvisas till [4] (sida 128-140).

### Broyden-Fletcher-Goldfarb-Shanno (BFGS) metod

En fördel med Quasi-Newton-metoder är att de inte kräver explicita uttrycker av andraderivatorna i  $\mathbf{H}$ . Uttrycket för  $\mathbf{x}_{k+1}$  kan skrivas på följande form:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{S}_k \mathbf{g}_k,$$

där

$$\mathbf{S}_k = \begin{cases} \mathbf{I}_n & \text{för steepest-descent metoden} \\ \mathbf{H}_k^{-1} & \text{för Newton-metoden} \end{cases}$$

$\mathbf{S}$  har samma funktion som inversen av hessian matrisen i Newtons metod.  $\mathbf{S}$  är genererad av tillgänglig data, och är en approximation av  $\mathbf{H}^{-1}$ . Allteftersom antalet upprepningar ökar, blir approximationen allt bättre. Efter  $n + 1$  upprepningar är den identisk med  $\mathbf{H}^{-1}$ .

Vi undersöker möjligheten att använda en positiv definit matris  $\mathbf{S}_k$  som lösning till

$$f(\mathbf{x}) = a + \mathbf{b}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x},$$

där  $a = f(\mathbf{0})$  och  $\mathbf{b}$  är gradienten av  $f(\mathbf{x})$  i  $\mathbf{x} = \mathbf{0}$ . Genom att derivera  $f(\mathbf{x}_k - \alpha \mathbf{S}_k \mathbf{g}_k)$  med avseende på  $\alpha$ , och sätta resultat lika med noll, ger

det oss värdet på  $\alpha$  som minimerar  $f(\mathbf{x}_k - \alpha \mathbf{S}_k \mathbf{g}_k)$ . Uttrycket för  $\alpha$  kan då skrivas som

$$\alpha_k = \frac{\mathbf{g}_k^T \mathbf{S}_k \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{S}_k \mathbf{H} \mathbf{S}_k \mathbf{g}_k},$$

där  $\mathbf{g}_k = \mathbf{b} + \mathbf{H} \mathbf{x}_k$ . För hur man genererar  $\mathbf{S}_k$  hänvisas läsaren till [4] (sida 177-181), och för en mer detaljerad beskrivning av teorin bakom BFGS-metoden hänvisas läsaren till [4] (sida 175-177).

### 3.3 Spearmans rangkorrelation

Eftersom observationerna, som vi kommer arbeta med, endast kommer in i analysen via sin inbördes rangordning, väljer vi att använda den icke-parametriska versionen av Pearsons korrelation, Spearmans rangkorrelation.

#### Definition

Rangordna  $x$ -värdena och  $y$ -värdena var för sig från 1 till  $n$ . Låt  $u_i := \text{rang}(x_i)$  och  $v_i := \text{rang}(y_i)$ . Spearmans rangkorrelation definieras som  $r_s := r_{uv}$ .

#### Sats

Låt  $d_i := u_i - v_i = \text{rang}(x_i) - \text{rang}(y_i)$  och  $D := \sum_{i=1}^n d_i^2$ . Om det inte förekommer några dubletter gäller  $r_s := 1 - \frac{6D}{n(n^2-1)}$ , se [1] (sida 399).

### 3.4 Teckenrangtest

Vi har följande observationer  $(x_1, y_1), \dots, (x_n, y_n)$ , där  $(x_i, y_i)$  är en observation från den kontinuerliga tvådimensionella stokastiska variabeln  $(X_i, Y_i)$ . Beräkna differenserna  $z_i = x_i - y_i$ , och rangordna dem efter  $|z_1|, \dots, |z_n|$ . Sedan beräknar vi rangsumman för de positiva ( $T_+$ ) eller negativa ( $T_-$ ) differenserna på följande sätt

$$T_+ = \sum_{k=1}^n k \cdot U_k,$$

där  $U_k = 1$  om rang  $k$  svara mot en positiv differens, annars är  $U_k = 0$ . Väntevärdet av  $T_+$  är

$$E[T_+] = E\left[\sum_{k=1}^n k \cdot U_k\right] = \sum_{k=1}^n k \cdot E[U_k] = \sum_{k=1}^n \frac{k}{2} = \frac{n(n+1)}{4},$$

och variansen

$$V(T_+) = V\left(\sum_{k=1}^n k \cdot U_k\right) = \sum_{k=1}^n k^2 \cdot V(U_k) = \frac{n(n+1)(2n+1)}{24}.$$

Vi vill testa nollhypotesen  $H_0 : X$  och  $Y$  har samma fördelning. Differenserna  $Z = X - Y$  är symmetriskt fördelade kring noll under  $H_0$ . Vi testar nollhypotesen mot den alternativa hypotesen,  $H_a : Y$ :s fördelning är förskjuten i förhållande till  $X$ :s fördelning. För stickprov där  $n > 20$  kan man utnyttja, att under  $H_0$  gäller, följande normalapproximation

$$T \approx N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right),$$

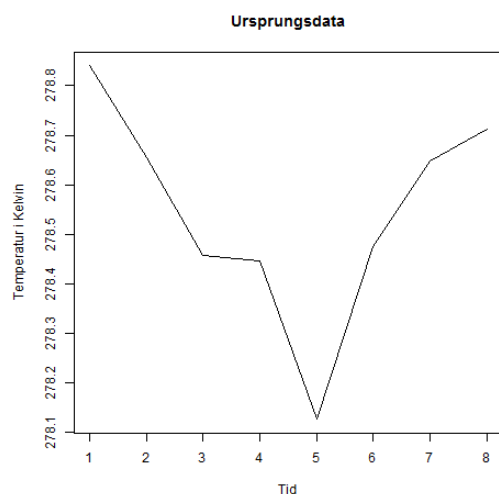
se [1] (sida 385, 390-391).

## 4 Metod

Data som vi kommer arbeta med under analysen är, som vi nämde tidigare, störda versioner av  $\mathbf{f}$ . Det gör vi eftersom det inte är rimligt att anta att vi känner de sanna värdena, samt för att om vi känner till de sanna värdena så är det ingen poäng med att försöka skatta dem (eller snarare formen på dess kurva). Stickprovsvariansen av  $\mathbf{f}$  är  $s^2 = \sigma_{\mathbf{f}}^2 = 0.0841$ , och den kommer vi utgå ifrån när vi nu bildar våra 50 störda vektorer. De störda vektorerna får vi genom att lägga till normalfördelat brus,  $\mathbf{Z}$ , till den sanna temperaturvektorn.  $\mathbf{Z}$  är en vektor som innehåller 50 stokastiska variabler  $Z$ . Där  $Z \sim N(0, \sigma^2)$ . Det är rimligt att anta att  $\sigma^2$  ligger någonstans mellan  $0.1s^2$  och  $10s^2$ . Vi får alltså den störda temperaturvektorn  $\mathbf{X} = \mathbf{f} + \mathbf{Z} = (X_{[i,1]}, \dots, X_{[i,100]})$ ,  $i = 1, \dots, 50$ . Målet är, som nämnt, att våra maximum likelihood skattningar ska efterlikna kurvan för  $\mathbf{f}$ . I analysdelen kommer vi variera  $\sigma^2$  för några olika värden.

Vi kommer nu beskriva de två tillvägagångssätten som vi kommer analysera PaiCo med. De är Hanhijärvis metod och Brattströms metod. Vilka vi nu kommer att referera till metod 1 respektive metod 2. Metoderna är inte exakta, utan de bortser från ett beroende då vi tar differensen av två stokastiska variabler. Fördelen med detta är att vi då kan använda univariata normalfördelningar.

Målet är inte att återskapa de sanna temperaturerna, utan istället att försöka återskapa formen på kurvan som tidsserien har. Det vi syftar på är kurvan upp till en linjär transformation,  $ax + b$ , ( $a > 0$ ), av tidsserien. Två tidsserier som alltså endast skiljer sig åt med en dylik transformation betraktas som ekvivalenta. På grund av att beräkningarna som vi kommer utföra i  $\mathbf{R}$  är tidskrävande, kommer vi endast titta på en tidsserie som är åtta tidssteg lång. Detta kommer vi diskutera senare i rapporten. I figuren nedan ser vi hur kurvan som vi vill återskapa ser ut.



Figur 2: Figuren visar den del av kurvan,  $f_{81}, \dots, f_{88}$ , i figur 1 som vi ska försöka återskapa.

Det ligger ingen bakomliggande tanke med att vi väljer att försöka återskapa kurvans form från följderna  $f_{81}, \dots, f_{88}$ , utan vi skulle kunna välja vilka åtta värden som helst (som ligger i följd).

Vi kommer nu endast titta på elementen  $X_{[i,81]}, \dots, X_{[i,88]}$  i  $\mathbf{X}$ ,  $i = 1, \dots, 50$  i storleksordning. Säg att vi observerar följande utfall för  $X_{[1,j]}$ ,  $j = 81, \dots, 88$ :

$$x_{[1,82]} > x_{[1,81]} > x_{[1,85]} > x_{[1,88]} > x_{[1,83]} > x_{[1,84]} > x_{[1,86]} > x_{[1,87]}, \quad (4)$$

Vi kommer använda (4) som exempel för att förklara hur de olika varianterna av PaiCo-metoden fungerar. Antalet möjliga utfall för (4) är  $8! = 40320$ , men alla utfall är inte lika troliga eftersom

$$X_{[i,j]} \sim N(f_j, \sigma^2), \quad i = 1, \dots, 50, \quad j = 81, \dots, 88.$$

Därför kommer små värden på  $\sigma^2$  ge utfall som påminner om storleksordningen av  $f_{81}, \dots, f_{88}$ , medan stora värden på  $\sigma^2$  kommer ge utfall som alla är lika sannolika. Vi söker alltså likelihooden

$$\begin{aligned} L(\mathbf{f}) &= \\ &= P(X_{[1,82]} > X_{[1,81]} > X_{[1,85]} > X_{[1,88]} > X_{[1,83]} > X_{[1,84]} > X_{[1,86]} > X_{[1,87]} | \mathbf{f}). \end{aligned} \quad (5)$$

#### 4.1 Metod 1

Med Hanhijärvis tillvägagångssätt, se [2] & [3], ersätter vi högerledet i (5) med produkten av alla observerade olikheter på följande sätt

$$\begin{aligned}
L_1(\mathbf{f}) = & P(X_{[1,82]} > X_{[1,81]}|\mathbf{f}) \cdot P(X_{[1,82]} > X_{[1,85]}|\mathbf{f}) \cdot P(X_{[1,82]} > X_{[1,88]}|\mathbf{f}) \\
& \cdot P(X_{[1,82]} > X_{[1,83]}|\mathbf{f}) \cdot P(X_{[1,82]} > X_{[1,84]}|\mathbf{f}) \cdot P(X_{[1,82]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,82]} > X_{[1,87]}|\mathbf{f}) \\
& \cdot P(X_{[1,81]} > X_{[1,85]}|\mathbf{f}) \cdot P(X_{[1,81]} > X_{[1,88]}|\mathbf{f}) \cdot P(X_{[1,81]} > X_{[1,83]}|\mathbf{f}) \cdot P(X_{[1,81]} > X_{[1,84]}|\mathbf{f}) \\
& \cdot P(X_{[1,81]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,81]} > X_{[1,87]}|\mathbf{f}) \cdot P(X_{[1,85]} > X_{[1,88]}|\mathbf{f}) \cdot P(X_{[1,85]} > X_{[1,83]}|\mathbf{f}) \\
& \cdot P(X_{[1,85]} > X_{[1,84]}|\mathbf{f}) \cdot P(X_{[1,85]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,85]} > X_{[1,87]}|\mathbf{f}) \cdot P(X_{[1,88]} > X_{[1,83]}|\mathbf{f}) \\
& \cdot P(X_{[1,88]} > X_{[1,84]}|\mathbf{f}) \cdot P(X_{[1,88]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,88]} > X_{[1,87]}|\mathbf{f}) \cdot P(X_{[1,83]} > X_{[1,84]}|\mathbf{f}) \\
& \cdot P(X_{[1,83]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,83]} > X_{[1,87]}|\mathbf{f}) \cdot P(X_{[1,84]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,84]} > X_{[1,87]}|\mathbf{f}) \\
& \cdot P(X_{[1,86]} > X_{[1,87]}|\mathbf{f}). \quad (6)
\end{aligned}$$

där vi skriver om den första termen i (6) i högerledet till  $P(X_{[1,82]} - X_{[1,81]} > 0|\mathbf{f})$  och de andra termerna skrivs om på samma sätt. Differenserna har följande fördelning:  $X_{[1,p]} - X_{[1,q]} \sim N(f_p - f_q, 2\sigma^2)$ ,  $p, q = 81, 82, \dots, 88$ ,  $p \neq q$ .

Återigen så poängterar vi att detta inte är en exakt metod. De ursprungliga variablerna är oberoende, medan differenserna mellan variablerna inte är det. Vi bortser dock från det här beroendet.

## 4.2 Metod 2

Brattstöms tillvägagångssätt påminner om Hanhijärvis, men nu ersätter vi högerledet i (5) med produkten av de observerade olikheterna som (endast) ligger intill varandra, se [3]. Det vill säga

$$\begin{aligned}
L_2(\mathbf{f}) = & P(X_{[1,82]} > X_{[1,81]}|\mathbf{f}) \cdot P(X_{[1,81]} > X_{[1,85]}|\mathbf{f}) \cdot P(X_{[1,85]} > X_{[1,88]}|\mathbf{f}) \\
& \cdot P(X_{[1,88]} > X_{[1,83]}|\mathbf{f}) \cdot P(X_{[1,83]} > X_{[1,84]}|\mathbf{f}) \cdot P(X_{[1,84]} > X_{[1,86]}|\mathbf{f}) \cdot P(X_{[1,86]} > X_{[1,87]}|\mathbf{f}). \quad (7)
\end{aligned}$$

Precis som för metod 1 skriver vi om termerna i högerledet av (7) som differenser som är större än noll. Differenserna har följande fördelning:  $X_{[1,p]} - X_{[1,q]} \sim N(f_p - f_q, 2\sigma^2)$ ,  $p, q = 81, 82, \dots, 88$ ,  $p \neq q$ .

Vi poängterar även här att den här metoden inte är exakt av samma anledning som för metod 1.

## 5 Statistisk modellering och analys av data

Som vi nämnde i sektion 2, är ett rimligt antagande att bruset (vilket vi adderar till  $\mathbf{f}$ ) har en varians som ligger någonstans mellan  $0.1\sigma_{\mathbf{f}}^2$  och  $10\sigma_{\mathbf{f}}^2$ . Då vi inte känner till det sanna värdet, kommer vi utföra analysen för några olika värden på  $\sigma^2$ . Närmare bestämt  $0.1 \cdot \sigma_{\mathbf{f}}^2$ ,  $0.9 \cdot \sigma_{\mathbf{f}}^2$ ,  $4 \cdot \sigma_{\mathbf{f}}^2$  och  $10 \cdot \sigma_{\mathbf{f}}^2$ .

När vi använder funktionen *mle* för att beräkna maximum likelihoodskattningarna behöver vi ange startvärden för parametrarna som ska skattas. Det beror på att *mle*, i vårt fall, använder sig av en Quasi-Newton-metod vid parameterskattningen. Metoden beskrevs i sektion 3.2. För de två varianterna av PaiCo-metoden kommer vi utföra analysen för tre olika uppsättningar startvärden. Dessa kommer givetvis analyseras med samma störda data. Startvärdena är följande

1. Den första uppsättningen startvärden är de sanna temperaturerna. Detta är förstås fusk, då vi i verkligheten inte känner till de sanna värdena. Men då vi vill återskapa formen av kurvan i figur 2, är det intressant att se hur väl den skattade kurvan stämmer överens med den sanna.
2. Den andra uppsättningen startvärden vi väljer, är medelvärdet av de 50 störda vektorerna, det vill säga för element  $f_{82}$  har vi startvärdet som medelvärdet av de 50 störda versionerna av  $f_{82}$  och på motsvarande sätt för övriga element. Detta är även det mest sannolika att man använder, eftersom det är det här man har tillgång till.
3. Den tredje uppsättningen startvärden sätter vi till fem för elementen. Anledningen är att det är intressant att se hur utfallet blir då startvärdena inte påminner om den sanna kurvans form. Utan endast är ett resultat utifrån informationen från ordningsrelationerna mellan de observerade värdena.

Startvärdena för  $f_{81}$  och  $f_{88}$  låser vi till ett respektive noll. Vi gör det för att lättare kunna jämföra metoderna (vi har då en standardiserad version av skattningarna), då vi utför analysen för olika värden på brusets varians,  $\sigma^2$ .

I tabellen nedan presenterar vi den första och tredje uppsättningen startvärden. Vilka vi kommer att referera till startvärde 1 respektive startvärde 3.

	Startvärde 1	Startvärde 3
$f_{81}$	1	1
$f_{82}$	278.6558	5
$f_{83}$	278.4576	5
$f_{84}$	278.4464	5
$f_{85}$	278.1271	5
$f_{86}$	278.4760	5
$f_{87}$	278.6485	5
$f_{88}$	0	0

Tabell 1: Tabellen visar den första och den tredje uppsättningen startvärden.

I tabellen nedan presenterar vi den andra uppsättningen startvärden, för olika värden på brusets varians. Den här uppsättningen refererar vi till startvärde 2.

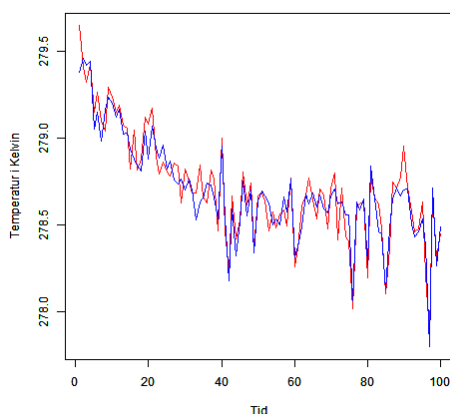
Startvärde 2				
Brusets varians	$0.1 \cdot \sigma_f^2$	$0.9 \cdot \sigma_f^2$	$4 \cdot \sigma_f^2$	$10 \cdot \sigma_f^2$
$f_{81}$	1	1	1	1
$f_{82}$	278.6348	278.5815	278.5929	278.4993
$f_{83}$	278.4735	278.3937	278.4948	278.6068
$f_{84}$	278.4569	278.4330	278.3755	278.4664
$f_{85}$	278.1176	278.1785	278.1227	278.0587
$f_{86}$	278.4745	278.5018	278.4529	278.4912
$f_{87}$	278.6583	278.6207	278.7160	278.5734
$f_{88}$	0	0	0	0

Tabell 2: Tabellen visar den andra (startvärde 2) uppsättningen startvärden för de olika värdena på brusets varians.



### 5.1 Brusets varians = $0.1 \cdot \sigma_{\mathbf{f}}^2$

Den första delen av analysen börjar med att vi lägger på det normalfördelade bruset med variansen  $\sigma^2 = 0.1 \cdot \sigma_{\mathbf{f}}^2$ ,  $\mathbf{Z} \sim N(0, \sigma^2 = 0.0084)$ . I figuren nedan ser vi skillnaden (i ett möjligt utfall) mellan den sanna och den störda temperaturvektorn.



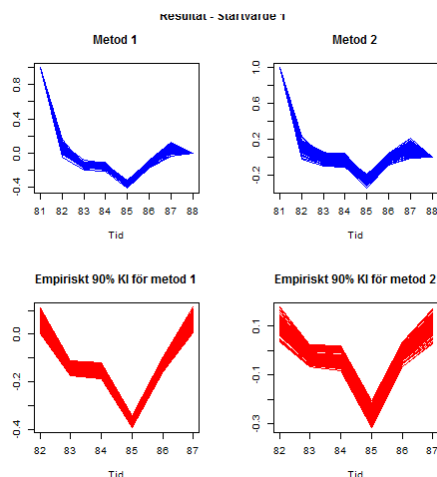
Figur 3: Den blåa kurvan motsvarar den sanna temperaturvektorn, och den röda den störda temperaturvektorn.

Det är inga stora skillnader mellan den störda tidsserien och den sanna i figuren ovan.

Vi ordnar nu de observerade utfallen i storleksordning och beräknar loglikelihooden (på de olika sätt som nämns i sektion 4.1 och 4.2) för samtliga störda vektorer. Dessa kan vi summera eftersom vi approximerar loglikelihooden med den loglikelihood som vi skulle ha fått om utfallen hade varit oberoende. Vi kör programmet i  $\mathbf{R}$  100 gånger för att kunna konstruera ett empiriskt konfidensintervall för de bägge metoderna.

### 5.1.1 Startvärde 1

Figuren nedan visar oss resultatet från körningen av programmet för de två tillvägagångssätten av PaiCo-metoden, då vi använder oss av startvärde 1.



Figur 4: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Som vi ser i figur 4 påminner kurvornas utseende om kurvan i figur 2, vilken vi vill återskapa. Vi noterar att konfidensintervallet för metod 1 är mindre än för metod 2.

För att kunna jämföra metoderna beräknar vi nu Spearmans rangkorrelation (den icke-parametriska motsvarigheten till Pearsons korrelation). Som vi nämnde tidigare kommer endast observationerna in i analysen via sin inbördes rangordning, då kan det vara rimligt att använda ett rangbaserat mått.

I tabellen nedan presenterar vi bland annat minimum, medel och maximum av korrelationen från de 100 körningarna.

Spearmans rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.8095	0.9024	0.9762	[0.8444 , 0.9604]
Metod 2	0.5476	0.8252	0.9762	[0.5830 , 1]

Tabell 3: Tabellen visar resultat av korrelationen från de 100 körningarna.

Eftersom vi har stickprov i par kan vi inte direkt titta på konfidensintervallen och se om de överlappar varandra. Utan vi utför ett teckenrangtest och bildar ett konfidensintervall för att se om intervallet innehåller värdet

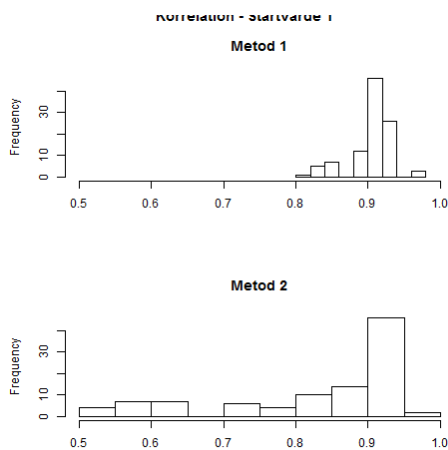
0. Vi utför testet på samma sätt som är beskrivet i sektion 3.4.  $x$ -värdena respektive  $y$ -värdena kommer från metod 1 respektive metod 2. Tabellen nedan presenterar resultatet från teckenrangtestet.

Teckenrangsumma	P-värde	95% konfidensintervall
1815.5	$3.875 * 10^{-9}$	[0.0834 , 0.1548]

Tabell 4: Tabellen visar resultatet från teckenrangtestet.

Ur tabellen ovan ser vi att det är en statistisk signifikant skillnad mellan de bägge metoderna i det här fallet.

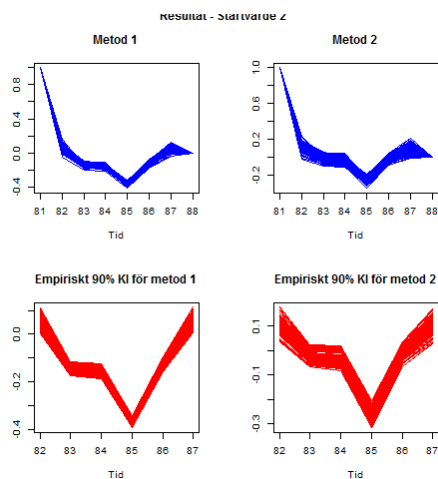
I histogrammet nedan kan vi se en skillnad i spridningen för de beräknade korrelationerna, för de två metoderna.



Figur 5: Histogram över de 100 framräknade korrelationerna.

### 5.1.2 Startvärde 2

Resultatet då vi använder den andra uppsättningen startvärden presenterar vi i figuren nedan.

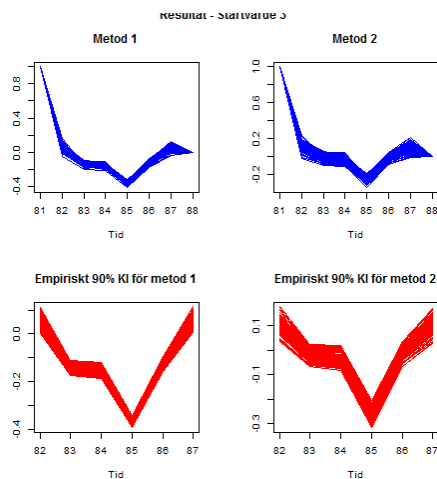


Figur 6: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Formen på dessa kurvor liknar de i figur 4. Tittar man närmre på figurerna är det endast små skillnader mellan dem. Resultatet från korrelationsberäkningarna ger oss samma resultat som i tabell 3 och figur 5. Det är alltså en signifikant skillnad mellan metoderna.

### 5.1.3 Startvärde 3

Till sist tittar vi på resultatet från analysen då vi använder startvärde 3.



Figur 7: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Även kurvorna i den här figuren liknar kurvorna i figur 4 och 6. Man ser alltså inga större skillnader.

Resultatet från korrelationsberäkningarna är närmast identiska med de tidigare beräknade korrelationerna. Histogrammet ser likadant ut som den i figur 5, men tabellen skiljer sig åt något. Resultatet presenterar vi i tabellen nedan.

Spearman's rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.8095	0.9024	0.9762	[0.8444 , 0.9604]
Metod 2	0.5476	0.8250	0.9762	[0.5831 , 1]

Tabell 5: Tabellen visar resultat av korrelationen från de 100 körningarna.

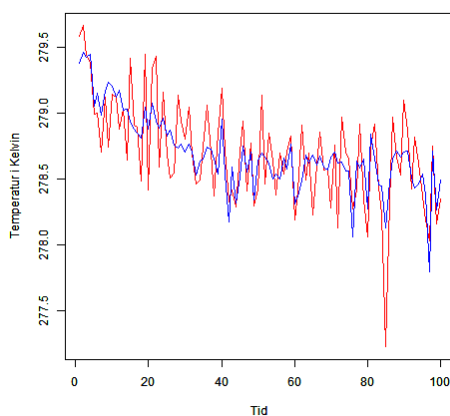
Resultatet från teckenrangtestet ser vi i tabell 6, och vi kan avläsa att det är en signifikant skillnad mellan metoderna, då intervallet inte innehåller värdet noll.

Teckenrangsumma	P-värde	95% konfidensintervall
1767	$3.489 \cdot 10^{-9}$	[0.0952 , 0.1548]

Tabell 6: Tabellen visar resultatet från teckenrangtestet.

## 5.2 Brusets varians = $0.9 \cdot \sigma_f^2$

Vi fortsätter nu analysen med att undersöka vad som händer då vi stör  $\mathbf{f}$  med ett normalfördelat brus med väntevärde noll och varians  $\sigma^2 = 0.9 \cdot \sigma_f^2 = 0.0757$ . I figuren nedan illustreras hur skillnaden mellan den sanna temperaturvektorn och den störda (ett möjligt utfall) ser ut.

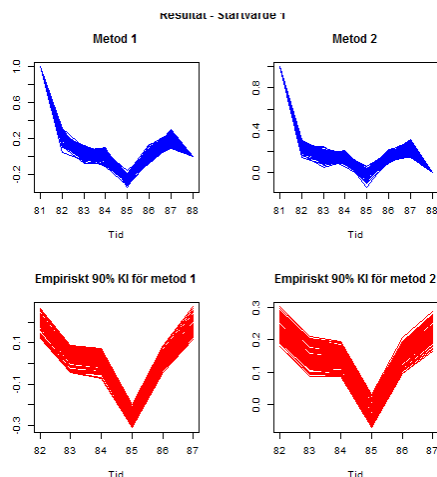


Figur 8: Den blåa kurvan motsvarar den sanna temperaturvektorn, och den röda den störda temperaturvektorn.

Den störda tidsserien skiljer sig nu från den sanna, men den följer den i stora drag.

### 5.2.1 Startvärde 1

I figuren nedan presenterar vi resultatet från körningen med startvärde 1.



Figur 9: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Med blotta ögat kan vi se att den större variansen på bruset gör en skillnad om man jämför figur 4 och 9. I figuren ovan ser vi att de skattade kurvorna har en (något) ökad spridning. Konfidensintervallen har även de blivit bredare. Både den första och andra metoden har skattat kurvor som påminner om kurvan i figur 2.

I tabellen nedan presenterar vi några numeriska värden för korrelationen från de 100 körningarna.

Spearman's rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.5238	0.6941	0.9048	[0.4949 , 0.8932]
Metod 2	0.3571	0.5460	0.6429	[0.3937 , 0.6983]

Tabell 7: Tabellen visar resultat av korrelationen från de 100 körningarna.

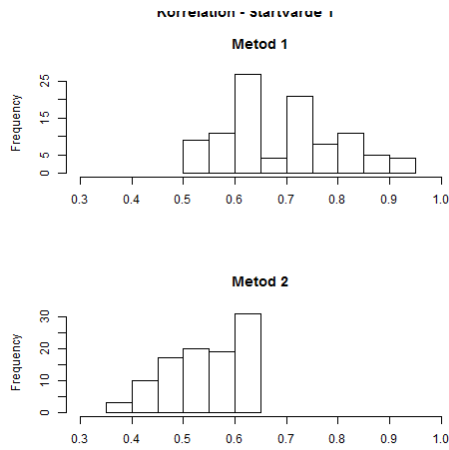
I tabellen nedan kan vi se resultatet från teckenrangtestet.

Teckenrangsumma	P-värde	95% konfidensintervall
4144	$4.726 * 10^{-16}$	[0.1429 , 0.1905]

Tabell 8: Tabellen visar resultatet från teckenrangtestet.

Ur tabellen ser vi att det är en signifikant skillnad mellan metoderna. Då brusets varians blev större, ser vi att konfidensintervallet har förflyttats längre bort från värdet noll än tidigare.

I histogrammet nedan kan vi se hur spridningen för de beräknade korrelationerna är för de bägge metoderna.



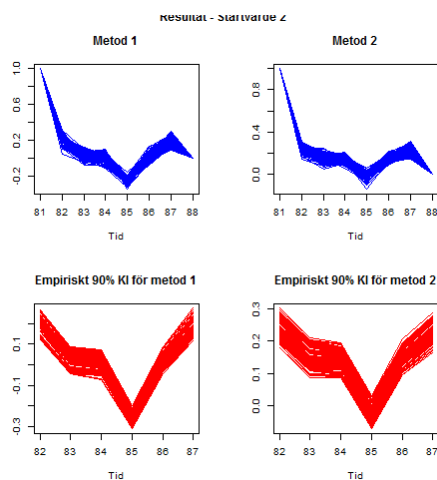
Figur 10: Histogram över de 100 framräknade korrelationerna.

Även i det här fallet ser vi en skillnad i spridningen mellan de bägge metoderna.



### 5.2.2 Startvärde 2

Resultatet från den andra uppsättningen startvärden presenteras i figuren nedan.



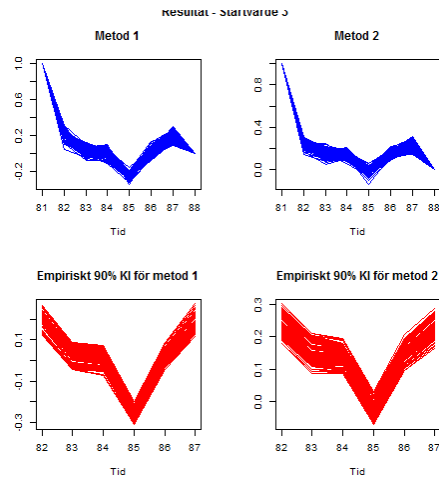
Figur 11: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Också i det här fallet liknar de skattade kurvorna i figur 11 den eftersökta kurvan i figur 2. De empiriska konfidensintervallen har blivit något bredare.

Resultatet från korrelationsberäkningarna ger oss samma resultat som i tabell 7 och figur 10. Vi har alltså en signifikant skillnad mellan metoderna.

### 5.2.3 Startvärde 3

Då vi använder oss av startvärde 3 får vi följande resultat.



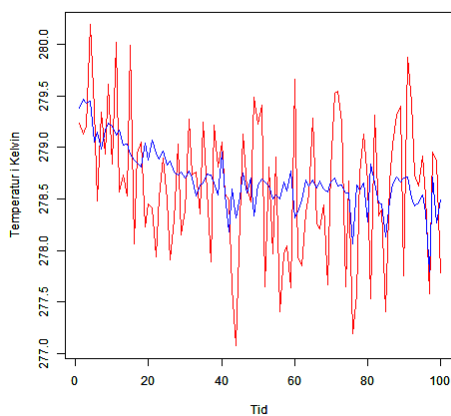
Figur 12: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Även i det här fallet liknar de skattade kurvorna i figur 12 den eftersökta kurvan i figur 2. De empiriska konfidensintervallen har blivit något bredare.

Resultatet från korrelationsberäkningarna ger oss samma resultat som i tabell 7 och figur 10. Även med startvärde 3 har vi en signifikant skillnad mellan de bägge metoderna.

### 5.3 Brusets varians = $4 \cdot \sigma_f^2$

Analysen fortsätter med att undersöka vad som händer då  $\mathbf{f}$  störs med ett normalfördelat brus med väntevärde noll och varians  $\sigma^2 = 4 \cdot \sigma_f^2 = 0.3364$ . Figuren nedan illustrerar hur skillnaden mellan den sanna temperaturvektorn och den störda (ett möjligt utfall) ser ut.

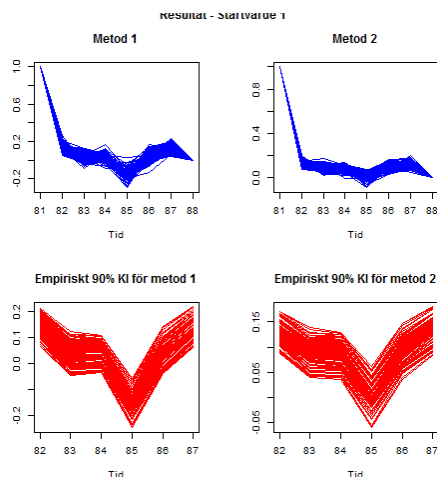


Figur 13: Den blåa kurvan motsvarar den sanna temperaturvektorn, och den röda den störda temperaturvektorn.

I figur 3 och figur 8 låg de störda vektorerna nästintill bredvid den sanna tidsserien. Först nu kan man se att tidsserien för den störda vektorn i figuren ovan avviker en hel del från de sanna värdena, men att den ändå (svagt) följer trenden hos den sanna tidsserien.

### 5.3.1 Startvärde 1

I figuren nedan presenterar vi resultatet från körningen då vi använder startvärde 1.



Figur 14: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Vi noterar att bägge metodernas skattade kurvor har tappat formen. Det är främst för metod 2, där kurvan har blivit plattare. Från figuren ser vi också att konfidensintervallen har blivit bredare.

I tabellen nedan presenterar vi några numeriska värden för korrelationen från de 100 körningarna.

Spearman's rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.1905	0.6483	0.8571	[0.4400 , 0.8567]
Metod 2	0.2143	0.4664	0.6429	[0.2737 , 0.6591]

Tabell 9: Tabellen visar resultat av korrelationen från de 100 körningarna.

I tabellen nedan presenteras resultatet från teckenrangtestet.

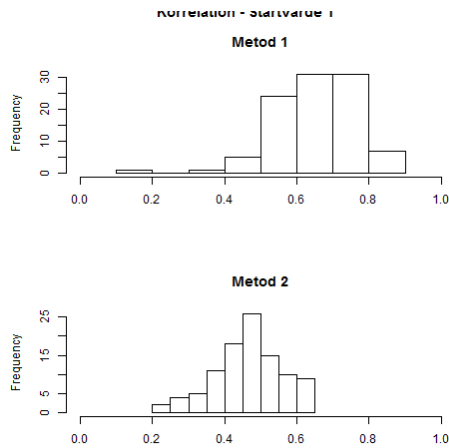
Teckenrangsumma	P-värde	95% konfidensintervall
4610	$2.200 \cdot 10^{-16}$	[0.1667 , 0.2143]

Tabell 10: Tabellen visar resultatet från teckenrangtestet.

Vi ser från tabellen att det är en signifikant skillnad mellan metoderna.

Med den större variansen på bruset har konfidensintervallet förskjutits åt höger från värdet noll.

I histogrammet nedan kan vi se hur spridningen för de beräknade korrelationerna är för de bägge metoderna.

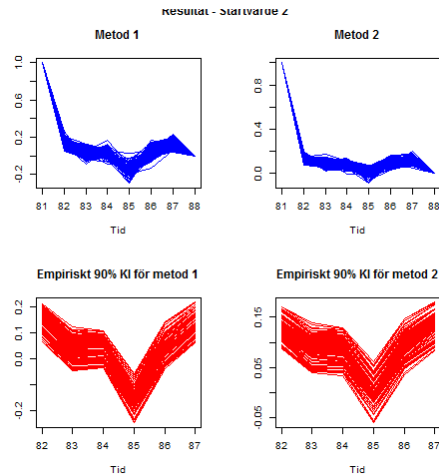


Figur 15: Histogram över de 100 framräknade korrelationerna.

Även i det här fallet ser vi en skillnad mellan de bägge metoderna.

### 5.3.2 Startvärde 2

Resultatet då vi använder den andra uppsättningen startvärden presenteras i figuren nedan.



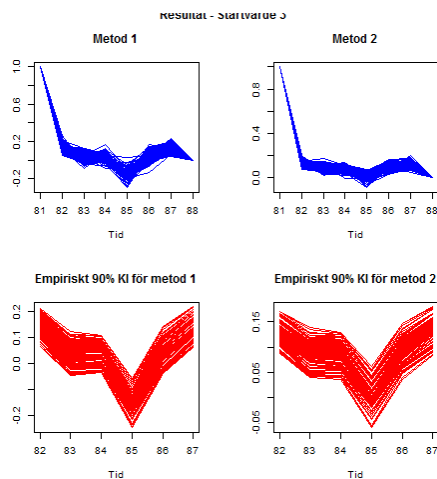
Figur 16: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Precis som i fallet ovan för startvärde 2, har de skattade kurvorna i figur 16 ändrat utseende något. Det är framförallt kurvan som metod 2 har skattat som ändrat form, då den har blivit plattare. Vilket syns tydligt då man jämför den med kurvan i figur 2. De empiriska konfidensintervallen har blivit något bredare.

Resultatet från korrelationsberäkningarna ger oss samma resultat som i tabell 9 och figur 15. Vi har alltså en signifikant skillnad mellan metoderna.

### 5.3.3 Startvärde 3

Då vi använder oss av startvärde 3 får vi följande resultat.



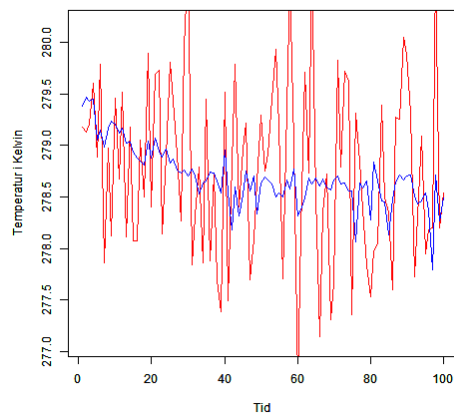
Figur 17: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Vi ser att kurvorna i figuren ovan har nästintill samma form som de i figur 16. Varför vi även kommenterar den på samma sätt som vi gjorde ovan.

Resultatet från korrelationsberäkningarna ger oss samma resultat som i tabell 9 och figur 15. Även i det här fallet har vi en signifikant skillnad mellan metoderna.

#### 5.4 Brusets varians = $10 \cdot \sigma_f^2$

Vi avslutar analysen med att undersöka vad som händer då  $\mathbf{f}$  störs med ett normalfördelat brus med väntevärde noll och varians  $\sigma^2 = 10 \cdot \sigma_f^2 = 0.8410$ . I figuren nedan illustreras hur skillnaden mellan den sanna temperaturvektorn och den störda (ett möjligt utfall) ser ut.



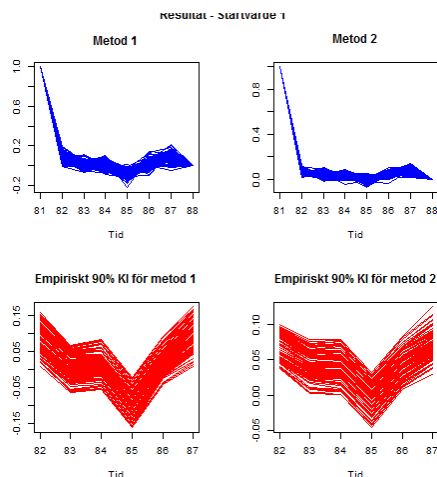
Figur 18: Den blåa kurvan motsvarar den sanna temperaturvektorn, och den röda den störda temperaturvektorn.

Nu ser tidsserien för den störda vektorn i figuren 18 ut att vara slumpmässigt brus.



### 5.4.1 Startvärde 1

I figuren nedan presenterar vi resultatet från körningen då vi använder startvärde 1.



Figur 19: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Vi noterar att de skattade kurvorna för de båda metoderna har blivit flackare i förhållande till den eftersökta kurvan. Från figuren ser vi också att konfidensintervallen har blivit bredare.

I tabellen nedan presenterar vi några numeriska värden för korrelationen från de 100 körningarna.

Spearman's rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.3333	0.6317	0.8333	[0.3890 , 0.8743]
Metod 2	0.0476	0.4452	0.7143	[0.1807 , 0.7098]

Tabell 11: Tabellen visar resultat av korrelationen från de 100 körningarna.

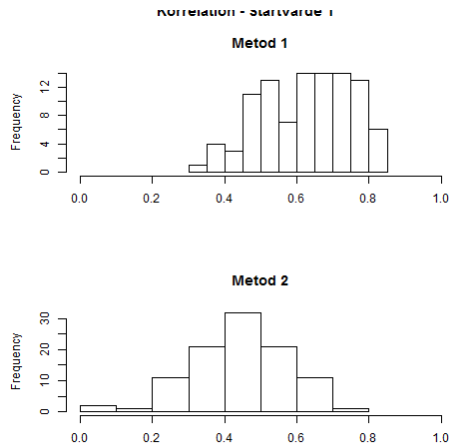
I tabellen nedan presenterar vi resultatet från teckenrangtestet.

Teckenrangsumma	P-värde	95% konfidensintervall
4526.5	$9.628 * 10^{-16}$	[0.1548 , 0.2262]

Tabell 12: Tabellen visar resultatet från teckenrangtestet.

Vi ser från tabellen att det är en signifikant skillnad mellan metoderna.

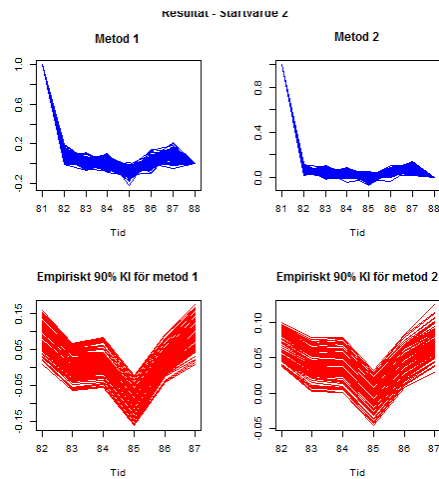
I histogrammet nedan kan vi se en skillnad i spridningen för de beräknade korrelationerna, för de två metoderna.



Figur 20: Histogram över de 100 framräknade korrelationerna.

### 5.4.2 Startvärde 2

Då vi använder den andra uppsättningen startvärden presenteras följande resultat i figuren nedan.



Figur 21: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

Precis som för de skattade kurvorna i figur 19 så har även kurvorna i

figuren ovan ändrat utseende efter att vi har ändrat storleken på brusets varians.

Då vi utför korrelationsberäkningarna noterar vi att det har blivit små förändringar jämfört med resultatet i tabell 11 och figur 20. I tabellen nedan presenteras resultatet för korrelationen.

Spearman's rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.3333	0.6321	0.8333	[0.3899 , 0.8744]
Metod 2	0.0476	0.4452	0.7143	[0.1807 , 0.7098]

Tabell 13: Tabellen visar resultat av korrelationen från de 100 körningarna.

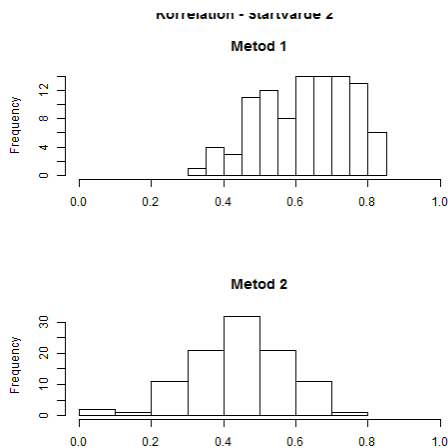
I tabellen nedan presenterar vi resultatet från teckenrangtestet. Vi ser endast små förändringar då vi jämför tabell 12 med tabellen nedan.

Teckenrangsumma	P-värde	95% konfidensintervall
4528.5	$8.976 * 10^{-16}$	[0.1548 , 0.2262]

Tabell 14: Tabellen visar resultatet från teckenrangtestet.

Vi ser från tabellen att det är en signifikant skillnad mellan metoderna.

I histogrammet nedan kan vi se hur spridningen för de beräknade korrelationerna är för de två metoderna.

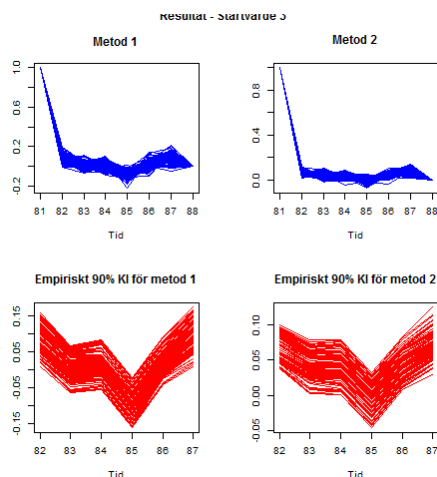


Figur 22: Histogram över de 100 framräknade korrelationerna.

Vi ser en skillnad mellan de två metoderna i figuren ovan.

### 5.4.3 Startvärde 3

Vi avslutar analysdelen med att undersöka resultatet då vi använder den tredje uppsättningen startvärden. Resultatet presenteras i figuren nedan.



Figur 23: De två övre figurerna visar de 100 skattade kurvorna för de olika metoderna. De två nedre figurerna visar det empiriska 90% konfidensintervallet för respektive metod.

De skattade kurvorna i figuren ovan, har precis som för kurvorna i figur 19 och figur 21 ändrat utseende efter att vi har ändrat storleken på brusets varians.

Korrelationsberäkningarna ger även i det här fallet en mindre förändring då man jämför resultatet med de, då vi använde startvärde 1 och startvärde 2. I tabellen nedan presenteras resultatet.

Spearman's rangkorrelation				
	Minimum	Medel	Maximum	95% konfidensintervall
Metod 1	0.3571	0.6326	0.8333	[0.3925 , 0.8727]
Metod 2	0.0476	0.4455	0.7143	[0.1812 , 0.7097]

Tabell 15: Tabellen visar resultat av korrelationen från de 100 körningarna.

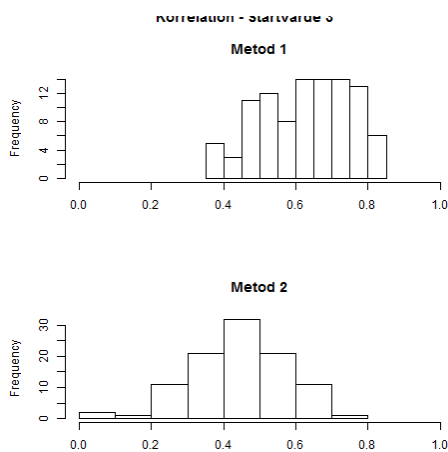
I tabellen nedan presenterar vi resultatet från teckenrangtestet. Det är endast små förändringar jämfört med tabell 12 och tabell 14.

Teckenrangsumma	P-värde	95% konfidensintervall
4528.5	$8.9765 * 10^{-16}$	[0.1548 , 0.2262]

Tabell 16: Tabellen visar resultatet från teckenrangtestet.

Vi ser från tabellen att det är en signifikant skillnad mellan metoderna.

I histogrammet nedan kan vi se hur spridningen för de beräknade korrelationerna är för de bägge metoderna.



Figur 24: Histogram över de 100 framräknade korrelationerna.

Även i det här fallet ser vi en skillnad mellan de bägge metoderna.

## 6 Diskussion

När man undersöker figurerna som visar de skattade kurvorna, ser man att både Hanhijärvis metod och Brattströms metod ger skattade kurvor som påminner om kurvan i figur 2, då brusets varians är  $0.1 \cdot \sigma_f^2$  respektive  $0.9 \cdot \sigma_f^2$ . Det gäller för samtliga uppsättningar startvärden. Först då brusets varians är  $4 \cdot \sigma_f^2$  ser man en skillnad från tidigare skattade kurvor. Det är främst de skattade kurvorna med Brattströms metod som har blivit märkbart flackare. Även detta gäller för samtliga uppsättningar startvärden. Då brusets varians är  $10 \cdot \sigma_f^2$  ser den störda temperaturvektorn ut som slumpmässigt brus, och då ser vi även en tydlig skillnad på de skattade kurvorna från Hanhijärvis metod.

Vi hade väntat oss att de skattade kurvorna skulle se sämre ut när brusets varians blev större. Då data som analyseras inte innehåller någon information om den eftersökta kurvan, är det inte konstigt att kurvorna ser dåliga ut.

Att vi valde att utföra analysen med 10års-medelvärden, istället för originaldata, påverkar inte resultatet av analysen då målet med det här arbetet inte var att rekonstruera temperaturen, utan den var att analysera PaiCo-metoden. Tanken var att genom att reducera temperaturvektorn, skulle vi kunna skatta (hela) kurvans form, för den sanna temperaturvektorn. Men på grund av tidsmässiga skäl blev det inte möjligt.

När vi jämför medelvärdena på de beräknade Spearmans korrelationerna noterar vi att då brusets varians är minimal ger metod 1 och metod 2 en korrelationskoefficient på nästan 0.90 respektive 0.83. Vid de större varianterna är korrelationen ungefär 0.64 respektive 0.45. Då vi har arbetat med stickprov i par utförde vi ett teckenrangtest och resultatet blev att för samtliga värden på brusets varians, och för samtliga uppsättningar startvärden, har vi en statistisk signifikant skillnad mellan metoderna.

Det är intressant att startvärde 3 ger i princip, genomgående, identiska konfidensintervall för korrelationsmedelvärdena, som de övriga två startvärdena. Alltså verkar det som att startvärdena inte har en (avgörande) betydelse för resultatet när vi jämför de skattade kurvornas form. Resultatet avgörs endast av de observerade olikheterna.

Tiden det tar för datorn (HP Pavillion dm1-4110eo) att skatta en kurva, av längd åtta med 50 störda vektorer, med Hanhijärvis metod är drygt 11 minuter för de två lägre nivåerna på brusets varians, och drygt 8 minuter för de högre nivåerna. För Brattströms tar det drygt 2.5 minuter respektive 2 minuter för de lägre respektive högre nivåerna. Det är som sagt inga fasta tider, utan ibland tar en körning längre tid och ibland mindre. Detta är av intresse då man väljer vilket tillvägagångsätt man vill använda.

Det visar sig, efter den här analysen, att Hanhijärvis metod är bättre än Brattströms metod på att rekonstruera temperaturen. Men det är önskvärt att fortsätta med analysen på en kraftfullare dator. Datorn som utförde

körningarna av koden i **R** hade sina begränsningar då vi endast kunde analysera åtta tidssteg åt gången. När man försöker rekonstruera temperaturer från förr, är det troligt att man vill göra det för längre tidssteg. I det här arbetet har vi sett att konfidensintervallen för korrelationsmedelvärdena är väldigt breda. Man bör därför också försöka öka antalet körningar, för att kunna minska konfidensintervallens bredd. Finns det tid bör den intresserade även fundera på att analysera den exakta metoden, och jämföra den med Hanhijärvis respektive Brattströms metod.

## Referenser

- [1] Stokastik, Liber AB, Sven Erick Alm och Tom Britton, Första upplagan, 2011.
- [2] Pairwise comparisons to reconstruct mean temperature in the Arctic Atlantic Region over the last 2,000 years, Springer-Verlag Berlin Heidelberg, Sami Hanhijärvi, Martin P. Tingley & Atte Korhola, 6 April 2013. Electronic supplementary material, doi:10.1007/s00382-013-1701-4
- [3] G. Brattström, personal communication, 2014.
- [4] Practical Optimization - Algorithms and Engineering Applications, Springer, Andreas Antoniou and Wu-Sheng Lu, 2007.
- [5] Schmidt, G.A., M. Kelley, L. Nazarenko, R. Ruedy, G.L. Russell, I. Aleinov, M. Bauer, S.E. Bauer, M.K. Bhat, R. Bleck, V. Canuto, Y.-H. Chen, Y. Cheng, T.L. Clune, A. Del Genio, R. de Fainchtein, G. Faluvegi, J.E. Hansen, R.J. Healy, N.Y. Kiang, D. Koch, A.A. Lacis, A.N. LeGrande, J. Lerner, K.K. Lo, E.E. Matthews, S. Menon, R.L. Miller, V. Oinas, A.O. Oloso, J.P. Perlwitz, M.J. Puma, W.M. Putman, D. Rind, A. Romanou, M. Sato, D.T. Shindell, S. Sun, R.A. Syed, N. Tausnev, K. Tsigaridis, N. Unger, A. Voulgarakis, M.-S. Yao, and J. Zhang, 2014: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, 6, 141-184, doi:10.1002/2013MS000265.