LUT-kauppakorkeakoulu

Kauppatieteiden kandidaatintutkielma

Liiketoiminta-analytiikka

ARIMA-mallin sovittaminen aikasarjadatalle – CASE: Lappeenrannan lentokentän ilmanlämpötila 1960–2020

ARIMA-model application for time series modeling – CASE: Lappeenranta's airport air temperature in 1960–2020

17.9.2021

Tekijä: Matias Heikkinen

Ohjaaja: Jyrki Savolainen

TIIVISTELMÄ

Tekijä:	Matias Heikkinen
Tutkielman nimi:	ARIMA-mallin sovittaminen aikasarjadatalle – CASE: Lappeenrannan lentokentän ilmanlämpötila 1960–2020
Akateeminen yksikkö:	LUT-kauppakorkeakoulu
Koulutusohjelma:	Kauppatieteet, Liiketoiminta-analytiikka
Ohjaaja:	Jyrki Savolainen
Hakusanat:	ARIMA-malli; data-analyysi; ilmasto

Tutkimuksen aiheena on tarkastella *ARIMA*-mallien toimintaa ja niiden matemaattista määrittelyä. Tutkimus keskittyy erityisesti *ARIMA*-mallien matemaattisen esitysmuodon tarvitsemien kerrointen määrittämiseen teoreettisesti, ja kuinka näitä voidaan tarkentaa kokeellisesti iteroiden. Työssä *SARIMA*-mallien yleinen yhtälö käydään läpi, selittäen mistä se koostuu, avaten myös viiveoperaattorin käyttöä. Lopulta dekomponointia hyödyntäen, ilmanlämpötila aikasarja datan stokastiseen sarjaan sovitetaan $SARIMA(0,0,1)(0,0,1)_{12}$-sovite havainnollistaen aikasarjamallin rakentamista käytännössä.

Tutkimuksen tuloksena esimerkki case-aineistoon Lappeenrannan ilmanlämpötilasta vuosina 1960-2020 onnistuttiin rakentamaan malli, jossa $SARIMA(0,0,1)(0,0,1)_{12}$-mallilla luotiin ennuste jäännöstermeille. *SARIMA*-mallille onnistuttiin laskemaan teoreettiset kertoimet, joita iteroimalla saatiin kertoimia tarkennettu pienintä neliösummaa minimoimalla parhaiksi mahdollisiksi. Työn tuloksena rakennettu malli ei välttämättä ollut paras vaihtoehto case-aineistona käytetyn aikasarjan mallintamiseen ja ennustamiseen, mutta mallin varsinainen rakennusprosessi tuki *ARIMA*-mallien toiminnan tarkastelua ja sen matemaattisen esitysmuodon kerrointen määrittämistä, joka oli itse tutkimuksen keskiössä.

ABSTRACT

Author:	Matias Heikkinen
Title:	ARIMA-model application for time series modeling – CASE: Lappeenranta's airport air temperature in 1960–2020
School:	School of Business and Management
Degree programme:	Business Administration, Business analytics
Supervisor:	Jyrki Savolainen
Keywords:	ARIMA-model; data-analytics; climate

The research subject is to study the function of *ARIMA*-models and the mathematical definition. The research focuses in particular on the theoretical determination of the coefficients required for the mathematical representation of *ARIMA*-models, specifying them experimentally by iteration. In this work, the general equation of the SARIMA models is reviewed by explaining what it consists of, also opening up the use of backshift operator. Finally, utilizing decomposition method, $SARIMA(0,0,1)(0,0,1)_{12}$-model can be fitted to the stochastic series of air-temperature time series data to illustrate the construction of the time series model in practice.

As the results of the study, the case-material of Lappeenranta's air temperature from 1960-2020 was used to build a model, in which the $SARIMA(0,0,1)(0,0,1)_{12}$-model was used to forecast residual terms. For the *SARIMA*-model, the theoretical coefficients were successfully calculated, and were iterated to optimum values by minimizing the least squares of the residual of model fit. As the result, the model itself may not have been the best for modeling and forecasting the time series used as case data, but the actual model building supported the review of *ARIMA* models and the determination of coefficients for the mathematical representation that was itself the key point of the study.

SISÄLLYSLUETTELO

1. Johdanto

Aikasarjadatan pohjalta luodut estimaatit tulevaisuuden ennustamisessa ovat entistä tärkeämmässä roolissa päätöksenteon taustalla nyky-yhteiskunnassa. (Box & Jenkins 1976, x) Yksi käytetyimmistä aikasarjadatan ennustamisen menetelmistä ovat *ARIMA*-mallit (autoregressive integrated moving avarage-model), joita tämä tutkimus käsittelee. Tutkimuksessa käsitellään ensin aiheeseen liitettävä kirjallisuuskatsaus erityyppisiin *ARIMA*-malleihin, niiden käyttöön ja toimintaan. Kirjallisuuskatsauksen tarkoituksena on kerätä materiaalia tutkimuksen taustalle, jota hyödynnetään toisessa vaiheessa aikasarjamallin sovittamisessa Lappeenrannan lentokentän säähavaintoaseman mittaamaa pitkäketjuiseen ilmanlämpötila dataan aikaväliltä 1960–2020.

ARIMA-mallit kuten muutkin aikasarjamallit ovat tärkeitä, koska ennusteiden tulee olla entistä tarkempia ja luotettavampia, jotta kyetään optimoida mahdollisimman luotettavasti tulevaisuuden tarpeita (Box & Jenkins 1976, ix-xii). Hyöty, mitä onnistuneista tulevaisuuden estimaateista voidaan saada resurssien tehokkaaseen käyttöön, tulee tulevaisuudessa korostumaan entisestään yhteiskunnan kehittyessä kohti resurssien hyödyntämisen tehokkaampaa muotoaan esimerkiksi luonnonvarojen osalta. Tässä tutkimuksessa tutustutaan tarkemmin *ARIMA*-malleihin ja niiden toimintaan, sekä pohditaan millaiselle aikasarja datalle *ARIMA*-mallit soveltuvat parhaiten.

ARIMA-mallit ovat todella yleisiä aikasarjojen pohjalta luotujen ennusteiden estimoinnissa, muodostaen melko yksinkertaisia ja suhteellisen tarkkoja malleja. *ARIMA*-mallit ovat laajalti käytössä ja niihin pohjautuvia tutkimuksia löytyy valtavasti. Monissa tutkimuksissa tutkimusaineistoon on myös sovitettu *ARIMA* pohjaisia hybridimalleja, joissa *ARIMA*-mallia on laajennettu jollakin toisella mahdollisesti monimutkaisemmalla mallilla. Harvemmissa soveltavissa tutkimuksissa kuitenkaan tarkastellaan varsinaisesti mallin muodostamisen tarkempaa matemaattista taustaa, olettaen että lukijalla on tarkempi tietämys mallin taustalla entuudestaan. Tällä tutkimuksella pyritään osoittamaan miten *ARIMA*-mallin kertoimien lukumäärä ja niiden arvot saadaan.

1.1. Tutkimuskysymykset

Tämän työn tarkoitus on sijoittua taustoitukseksi *ARIMA*-metodia käyttäneille tutkimuksille avaten *ARIMA*-mallien taustaa matemaattisesti. Tärkeimmäksi tutkimuskysymykseksi muodostui RQ1:

”Miten ARIMA-mallit toimivat ja millainen on niiden matemaattinen määrittely?”

Kysymystä RQ1 tarkennetaan alakysymyksellä:

”Kuinka ARIMA-mallin matemaattisen esitysmuodon tarvitsemat kertoimet saadaan määriteltyä datasta?”

Toinen tutkimuskysymys RQ2 käsittelee *ARIMA*-mallien soveltamista käytäntöön, jossa esimerkki case-aineistona on lämpötiladata:

”Kuinka ARIMA-mallit soveltuvat lämpötiladatan tutkimiseen ja millainen on Ilmatieteenlaitoksen Lappeenrannan lentokentältä kerätyn ilmanlämpö case-aineiston paras stokastisen sarjan ARIMA-sovite valitulle aikavälille?”,

Toiseen tutkimuskysymykseen on tarkoitus vastata muodostamalla kirjallisuuskatsaus *ARIMA* ja *ARIMA*-hybridimalleja käyttäneisiin tutkimuksiin pitkäketjuisista lämpötila data-aineistosta, ja sovittamalla *ARIMA*-mallin Lappeenrannan lentokentältä kerättyyn ilman lämpötila dataan. Kirjallisuuskatsauksen pohjalta on myös tarkoitus löytää vastaus määritettävän mallin sovitteen rationaaliselle aikavälille.

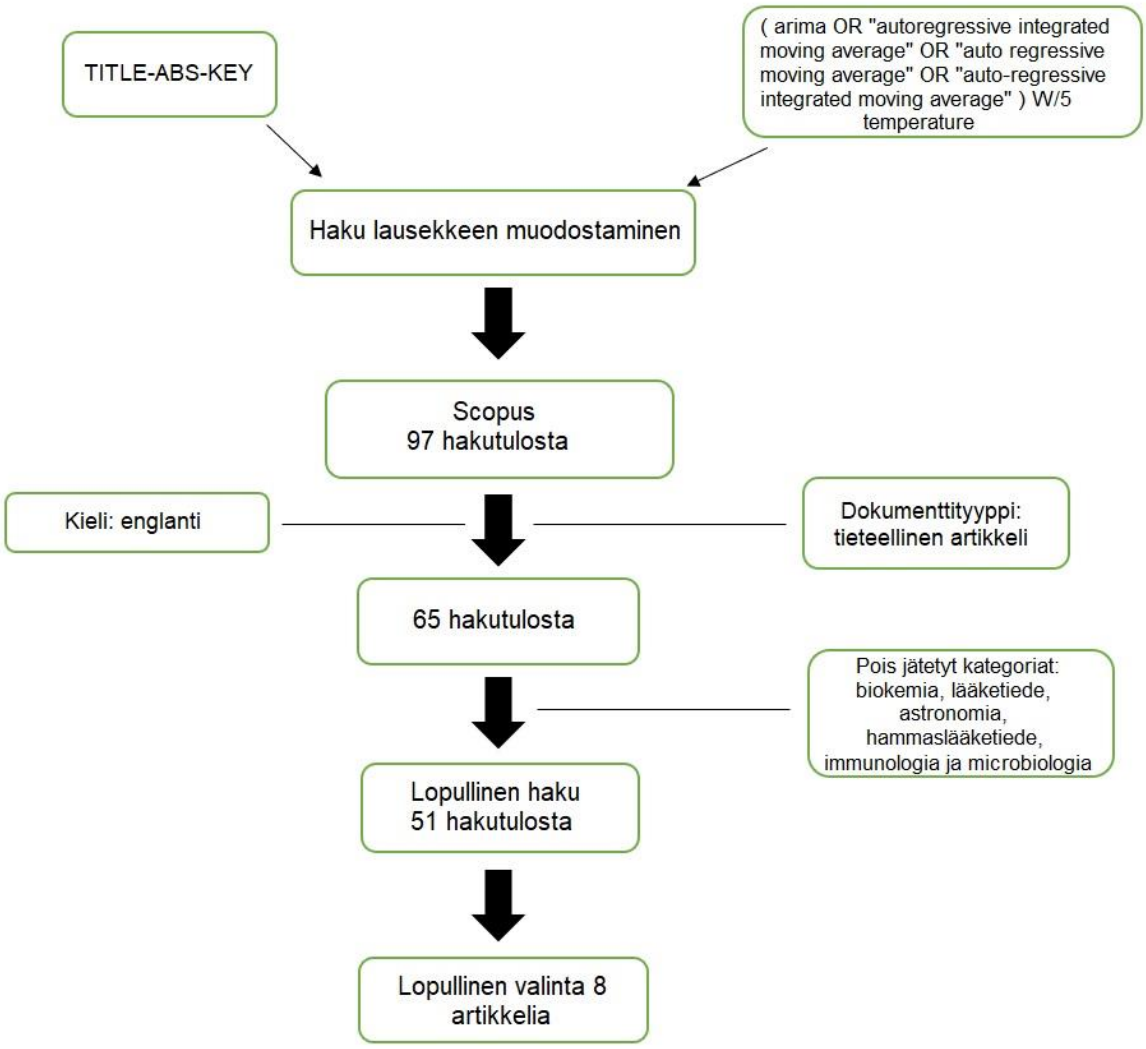
1.2. Rakenne

Tässä työssä tutustutaan *ARIMA*-malleja hyödyntäneiden tieteellisten tutkimusten tutkimusaiheisiin pinta puolisesti, luoden pohjan ymmärrykselle *ARIMA*-mallien monipuolisesta sovellettavuudesta ja niiden tärkeyden monien niistä johdettujen hybridimallien taustalla. *ARIMA*-mallien yleinen muodostaminen sekä ϕ- ja θ-kerrointen määrittämisen teoreettinen matemaattinen tausta käydään tutkimuksessa läpi. ϕ- ja θ-kertoimet kuvaavat aikasarjamallin käyttäytymistä suhteessa jo mitattujen datapisteiden käyttäytymiseen.

Työn viimeisessä osassa *ARIMA*-malli sovelletaan esimerkki datan satunnaisvaihtelun määrittämiseen, aikasarjan deterministisen osan muodostuessa dekomponoinnin (Alsuhail & Kokkinen 2005) (decomposition) seurauksena trendistä, sekä kausivaihtelusta (seasonality) (Bouznad, Guastaldi et al. 2020; Ye, Yang et al. 2013). Sovitettujen mallien pohjalta valitaan parhaimmat parametrit antanut malli ja tehdään tämän mallin pohjalta päätelmät aikasarjan tulevasta jatkuvuudesta, sekä pohditaan, kuinka pitkälle voidaan kyseisellä mallilla antaa luotettavia ennusteita esimerkki datasta.

2. ARIMA-mallien teoriaa

Tutkimuksen taustana on käytetty systemoitua kirjallisuuskatsausta (jamk, 2021), jolla pyrittiin rakentamaan viitekehys tutkimuksen taustalle. Tiedon haku toteutettiin Scopus Elsevierin tietokannassa. Haku tehtiin englanniksi ja aineiston haussa aiheesta jo tehtyjen tutkimusten kieli rajattiin englantiin. Lopullinen haku rajoittui 51 tieteelliseen artikkeliin, kun tulokset rajattiin englannin kielisiin, tieteellisinä artikkeleina julkaistuihin tutkimuksiin ja niistä karsittiin sellaisten aihepiirien tutkimukset, joiden ei uskottu tuovan tämän tutkimuksen kannalta oleelliseen sisältöön minkäänlaista lisäarvoa (Kuva 1). Lopullisesti tutkimuksessa keskityttiin haun 51 artikkelista tarkemmin kahdeksaan, sekä lisäksi Box & Jenkinsin (1976) teokseen ”*Time series analysis forecasting and control”*, joita yhdistämällä ARIMA-mallin parametrit pyritään saada määriteltyä.



Kuva 1: *ARIMA*-mallien kirjallisuuskatsauksen prosessi

Hakulauseke (Kuva 1) muodostui toisen tutkimuskysymyksen pohjalta, sillä toisen tutkimuskysymyksen rajaus oli paljon tarkempi kuin ensimmäisen. Näin lopullisen esimerkki dataan rakennetun mallin sovitteen vertaaminen aiempien rinnastettavien aihepiirien tutkimusten sovitteisiin on helpompaa, jolloin sovitettavan mallin oikeellisuus on helpompi havainnoida. Jos haku olisi muotoiltu ensimmäisen tutkimuskysymyksen mukaan hakemaan kaikkia mahdollisia Scopus Elsevierin kannan tutkimuksia, jossa on käytetty *ARIMA*-mallinnusta, olisi ensisijaisia hakutuloksia löytynyt yli 13000 kappaletta.

2.1 ϕ- ja θ-kerrointen astelukujen määrittäminen

ARIMA-mallit koostuvat nimensä mukaan autoregressiivisestä osasta (*AR*), integroidusta osasta (I), sekä liukuvan keskiarvon osasta (*MA*). *ARIMA*-malli ilmoitetaan yleensä muodossa *ARIMA(p,d,q)*, jossa *p* kertoo autoregressiivinen osan asteen, *d* integroinnin asteen ja *q* monennen asteen liukuvan keskiarvon osa malliin on sisällytetty, (Box & Jenkins 1976; Kesavan, Muthian et al. 2021; Islam, A. R. M. T., Karim et al. 2021; Wang, Huang et al. 2019). Data-aineiston ollessa kausittaista, *ARIMA* malliin sisällytetään myös kausivaihtelua kuvaavat komponentit *P*, *D* ja *Q* alaindeksillä *m*, joka tarkoittaa data-aineistossa olevan kausivaihtelun jakson pituutta havaintoaineiston mittaindeksin asteikolla, $ARIMA(p,d,q)(P,D,Q)_m$, (Box & Jenkins 1976). Esimerkiksi kvartaalisesti kausittain vaihtelevan datan pohjalta m saisi arvon 4, koska kausittainen vaihtelu toteutuu aina 4 aika periodin kuluessa alkaen alusta periodien toteuduttua. Kausittaisesta *ARIMA*-mallista saatetaan käyttää myös nimitystä *SARIMA* (seasonal autoregressive integrated moving avarage).

Box & Jenkins (1976, 18) suosivat *ARIMA*-mallien kerrointen määrittämistä numeerisesti iteroiden. Kertoimet kannattaa määrittää teoreettisesti perustuen autokorrelaatioarvoihin. Niiden teoreettiset arvot on hyvä asettaa iteraation alkuarvaukseksi, koska ne ovat monesti varsin lähellä mallin parhaita kertoimien arvoja, ellei täysin samat (Box & Jenkins 1976, 19). Varsinaiset parhaat mahdolliset mallin kertoimet saadaan iteraation seurauksena, joko minimoiden mallin ja varsinaisen datan välistä pienintä neliösummaa tai maksimoiden todennäköisyysfunktiota (likelihood function). Tämän työn lähestymistapa perustuu pienimmän neliösumman minimoimiseen.

2.2 Viiveoperaattori

ARIMA-mallien kohdalla, malli on monesti esitetty muodossa, jossa on käytetty viiveoperaattoria *B* (backshift operator), tunnettu myös *L* (lag operator). Viiveoperaattori helpottaa huomattavasti mallin esittämistä kaava muodossa ja on ylipäätään välttämätön kaavaa johdettaessa. Viiveoperaattorin tarkoituksena on että, mittapistettä yksi ajanhetki sitten voidaan merkitä viivytettyinä mittapisteenä,

$$S_{t-1} = BS_t \tag{1}$$

jolloin yleisesti ottaen mittapiste *n* ajanhetkeä sitten on,

$$S_{t-n} = B^n S_t \tag{2}$$

koska jos ajatellaan että viiveoperaattori *B* siirtyy yhden mittapisteen taaksepäin, *n* kappaletta viiveoperaattoreita siirtyy *n* mittapistettä taaksepäin, jolloin kun *n* kappaletta viiveoperaattoreita kerrotaan keskenään, saadaan luonnollisesti viiveoperaattorin *B* n:s potenssi B^n.

Viiveoperaattoria voidaan myös hyödyntää vektoreille. Kun ajatellaan, että vektori *S* sisältää kaikki jo mitatut datapisteet, tällöin viiveoperaattorin käyttö muodostaa kertaalleen viivästetyn datavektorin *BS*, joka siirtää kaikki datavektorin *S* mittapisteet yhden ajanhetken taaksepäin, siten että vektori *BS* alkaa vasta datavektorin *S* toisesta mittapisteestä ja saa viimeiseksi alkion arvokseen 0.

$$S = \begin{bmatrix} s_t \\ s_{t-1} \\ s_{t-2} \end{bmatrix}, \; niin \qquad BS = \begin{bmatrix} s_{t-1} \\ s_{t-2} \\ 0 \end{bmatrix} \tag{3}$$

Vektorin *S* ensimmäinen alkio s_t vastaa vektorin *BS* ensimmäistä alkiota s_{t-1}, s_{t-1} vastaa s_{t-2}, jolloin vektorin *S* alkiolle s_{t-2} vastine vektorista *BS* on 0. Jos vektoreiden välistä vaihtelua halutaan tutkia, joudutaan ne tasaamaan. Tasauksessa vektoreiden lopusta joudutaan jättämään alkioita tarkastelun ulkopuolelle, siten että vain alkiot, joilta löytyy vastinpari, joka vastaa jotakin alkuperäisen datavektorin mittapistettä pidetään tarkastelussa. Eli, jos vastinpariksi jäisi ainoastaan luotu 0 arvo, alkiot jätetään tarkastelussa huomiotta. Edellä olleessa tapauksessa, halutessa tasata vektorit niiden välistä tarkastelua varten, jouduttaisiin molemmista vektoreista hylätä viimeiset alkiot,

$$S = \begin{bmatrix} s_t \\ s_{t-1} \end{bmatrix}, \; BS = \begin{bmatrix} s_{t-1} \\ s_{t-2} \end{bmatrix} \tag{4}$$

Toinen viiveoperaattorista saatava hyöty on mallin määrittelyssä mallin kaavaa johdettaessa. Jos otetaan esimerkiksi käsittelyyn $SARIMA(1,0,0)(1,0,0)_{12}$, olisi mallin *AR(1)* ja *SAR(1)* välinen vaikutus,

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})S_t = \varepsilon_t \quad (5)$$

niiden keskinäiseen viivästettyyn datapisteeseen haastavaa havainnoida ilman *B* potenssien laskusääntöjä,

$$(1 - \phi_1 B - \Phi_1 B^{12} - \phi_1 \Phi_1 B^{1+12})S_t = \varepsilon_t$$

jolloin

$$S_t - \phi_1 S_{t-1} - \Phi_1 S_{t-12} - \phi_1 \Phi_1 S_{t-13} = \varepsilon_t \quad (6)$$

2.3 AR-malli

Autoregressiivisessä mallissa selitetään tulevaisuuden estimoitavia havaintoarvoja menneisyyden jo tapahtuneilla havaintoarvoilla. Autoregressiivisen mallin huomioon otettavat mahdolliset eri asteluvut *p*, sekä kausittaisen komponentin asteluku *P,* saadaan selville osittaiskorrelaatiokuvaajista (PACF; partial autocorrelation function), (Box & Jenkins 1976, 185; Shirvani, Nazemosadat et al. 2015) jotka pohjautuvat aikasarjan mittapisteiden välisiin osittaiskorrelaatioihin, eli siihen kuinka hyvin yksi mittapiste selittää toista mittapistettä suoraan, jättäen huomiotta kaiken muista mittapisteistä johtuvan välillisen vaihtelun.

Laskettavan periodin mittapiste voidaan määritellä edellisten periodien mittapisteiden avulla kaavalla (Box & Jenkins 1976, 53),

$$S_t = \phi_0 + \phi_1 S_{t-1} + \cdots + \phi_p S_{t-p} + \varepsilon_t \quad (7)$$

jossa ϕ ovat vakio kertoimia, S_t on mittapiste ajanhetkellä *t*, jolloin S_{t-p} on mittapiste *p* ajanhetkeä sitten, jotka malliin halutaan sisällyttää mukaan. ε_t on virhetermi ajanhetkellä *t*, eli kyseisen ajanhetken virhetermi. Virhetermin ε_t, mallin ollessa onnistunut, tulisi sen käsittää ainoastaan mittauksista johtuva kohina, eikä sitä täten pystytä laskennallisesti määrittämään ennen kuin kyseinen ajanhetki on jo tapahtunut.

2.3.1 ϕ-kertoimien määrittäminen

Autoregressiivistä mallia pystytään ajattelemaan viivästettyjen havaintoarvojen lineaarikombinaationa (Zarei, Moghimi 2019). Kun autoregressiivisen osan lausekkeen järjestää matriisi esitys muotoon,

$$[O \quad BS \quad \cdots \quad B^pS] \times \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \varepsilon = [S] \tag{8}$$

jossa *O* (ones) on pystyvektori, jossa on *n-p* kappaletta ykkösiä, jossa *n* vastaa datan mittapisteiden määrää ja *p* autoregressiivisen osan järjestyslukua, *BS* on pystyvektori kertaalleen viivästetystä datavektorista *S*, jolloin pystyvektori B^pS muodostuu *p* kertaa viivästetystä datavektorista *S*. Kun vektorit ovat tasattu siten että vektori *BS* alkaa datavektorin *S* toisesta, vektori B^pS alkaa vektorin *S* *p*+1 alkiosta, datavektori *S* alkaa sen ensimmäisestä alkiosta ja kaikista vektoreista on poistettu *p* kappaletta arvoja lopusta, jolloin kaikki luodut nolla-arvot poistuvat, pystytään ϕ-arvot approksimoimaan ortogonaaliprojektion avulla (Van Le, Nishio 2015),

$$x = (A'A)^{-1} \times A'b$$

$$x = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}, \; A = [O \quad BS \quad \cdots \quad B^pS], \; b = [S] \tag{9}$$

kun virhetermin ε_t ajatellaan olevan mahdollisimman pieni lähestyen nolla vektoria, muodostaen pienimmän neliösumman henkisen ratkaisun ϕ-arvoille. Näin ollen mallin koostuessa pelkästä autoregressiivisestä osasta, ortogonaaliprojektiolla määritetyt ϕ-arvot ovat parhaat mahdolliset ϕ-arvot. ϕ-arvoja ei pystytä laskemaan ortogonaaliprojektiolla, jos $(A'A)^{-1}$ on singulaarinen, (Van Le, Nishio 2015). Tämä kuitenkin tarkoittaisi sitä, että kahden tai useamman eri viiveillä viivästettyjen mittapisteiden vektoreiden täytyisi olla täysin samat, jolloin suurin osa mitta-arvoista olisivat samoja arvoja.

Autoregressiivisen mallin ϕ-kertoimet pystytään laskemaan Yule-Walker:in yhtälön matriisiesitys muodossa myös autokorrelaatio matriisista (Box & Jenkins 1976, 189-190)

$$x = R^{-1}r$$

$$x = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}, \quad R = \begin{bmatrix} 1 & r_1 & \cdots & r_{p-1} \\ r_1 & 1 & \cdots & r_1 r_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1} & r_{p-1} r_1 & \cdots & 1 \end{bmatrix}, \quad r = \begin{bmatrix} r_1 \\ \vdots \\ r_p \end{bmatrix} \tag{10}$$

jossa r_n on datavektorin S ja viivästetyn vektorin B^nS autokovarianssin kerroin (Box & Jenkins 1976, 190, 243), eli

$$r_n = S'B^nS \tag{11}$$

Autoregressiivisen osan malliin pystytään lisäämään myös kausittaisen vaihtelun autoregressiivinen osa $SAR(P)_m$ (seasonal autoregressive), asettamalla kausittaisen autoregressiivisen osan viivästetyt pystyvektorit B^mS … $B^{Pm}S$ osaksi mallia kausivaihtelun periodin pituudella m,

$$x = (A'A)^{-1} \times A'b$$

$$x = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \\ \Phi_1 \\ \vdots \\ \Phi_P \end{bmatrix}, \; A = \begin{bmatrix} O & BS & \cdots & B^pS & B^mS & \cdots & B^{Pm}S \end{bmatrix}, \; b = [S] \tag{12}$$

jolloin *SAR*-osan Φ-kertoimet asetetaan malliin normaalin *AR*-osan ϕ-kerrointen loppuun. Vektorin B^mS täytyy alkaa alkuperäisten datan mittapisteiden sisältävän vektorin S m+1 alkiosta, jolloin vektori $B^{Pm}S$ alkaa datavektorin S Pm+1 mittapisteestä. Kaikki mallissa mukana olevat pystyvektorit täytyy tasoittaa poistamalla niiden viimeiset Pm alkiota.

2.4 MA-malli

Liukuvan keskiarvon mallissa tulevaisuuden havaintoarvot muodostetaan lisäämällä aikaisempien datapisteiden keskiarvoon viivästettyjen virhetermien painotetut arvot, jolloin lähtökohtaisena ajatuksena on, että tulevaisuuden arvot tulevat jatkossakin vaihtelemaan keskiarvon

molemmin puolin θ-painokertoimien mukaisesti. Laskettavan periodin mittapiste S_t voidaan määritellä edellisten periodien mittapisteiden avulla kaavalla,

$$S_t = \theta_0 + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} + \varepsilon_t \quad (13)$$

jossa θ ovat vakiokertoimia, ja ε_t on virhetermi ajanhetkellä t, jolloin ε_{t-q} on virhetermi q ajanhetkeä sitten.

2.4.1 Virhetermien ε muodostaminen

Virhetermi on yksinkertaista määrittää mallin rakennusvaiheessa, koska tiedetään varmuudella datapisteiden jo toteutuneet havaintoarvot. Virhetermi muodostetaan tapahtuneen datapisteen ja estimoidun mittapisteen välisenä erotuksena samalta ajanhetkeltä

$$\varepsilon_{t-q} = S_{t-q} - \hat{S}_{t-q} \quad (14)$$

jossa S_t on mitattu datapiste ja $\hat{S}_t$ on estimoitu datapiste samalta ajanhetkeltä millä tahansa viiveen q arvolla. Virhetermi siis kertoo, kuinka paljon estimoitu mitta-arvo eroaa todellisesta mitta-arvosta kyseisellä ajanhetkellä.

2.4.2 θ-kertoimien määrittäminen

Teoreettiset θ-kertoimet saadaan *MA(q)* prosessissa määriteltyä autokorrelaatiofunktion (ACF) arvoista kaavalla,

$$r_n = \frac{-\theta_n + \sum_{j=1}^{q-n}\theta_j\theta_{j+n}}{1+\sum_{j=1}^{q}\theta_j^2} \qquad | \; n \le q, \; -1 < \theta_n < 1 \quad (14)$$

jossa r_h on autokorrelaatiofunktion h:s arvo (11) ja

$$r_n = 0 \qquad | \; n > q$$

ja

$$r_1 = \frac{-\theta_1}{1+\theta_1} \qquad | \; -1 < \theta_n < 1 \quad (15)$$

kun kyse on *MA(1)*- tai *SMA(1)$_m$*-prosessista (seasonal moving average), jossa r_1 korvataan r_m (Box & Jenkins 1976, 57, 69–71, 187, 314–315).

2.5 Differenssi mallissa

Alkuperäisen datan epäonnistuessa täyttämään mallin vaatimaa stationaarisuusehtoa (Box & Jenkins 1976) täytyy joko malliin sisällyttää stationaarisuuden kumoava komponentti tai alkuperäisestä datasta on otettava jonkin asteen differenssi stationaarisuuden saavuttamiseksi jo ennen mallin muodostamista. *SARIMA*-malleihin voidaan sisällyttää joko tavallinen differenssi tai kausittainen differenssi. Tavallinen differenssi saadaan, kun malliin sisällytetään

$$(1 - B)^d S_t \tag{16}$$

komponentti, ja kausittainen differenssi sisällyttämällä

$$(1 - B)^{Dm} S_t \tag{17}$$

Komponentti (Box & Jenkins 1976, 88–105, 304–320), joissa *d* on laskettavan differenssin asteluku, *D* kausittaisen differenssin asteluku, *m* kausivaihtelun jaksonpituus, *B* merkitsee viiveoperaattoria ja S_t on laskettava mittapiste ajanhetkellä *t*. Jos malliin on täytynyt sisällyttää differenssi tai alkuperäisestä datasta on otettu jonkin asteen differenssi, täytyy ennustetut mittapisteet muistaa integroida kumoamalla muodostettu differenssi tämän käänteistoiminnolla, jotta lopullinen ennuste vastaisi todellisia alkuperäisiä mitta-arvoja eikä niiden välisiä muutoksia.

2.6 ARIMA-mallit

Peruskaava kaikille *ARIMA(p,d,q)(P,D,Q)$_m$* malleille voidaan kirjoittaa muotoon (Box & Jenkins 1976, 305):

$$\begin{aligned}&\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(1 - B)^d (1 - \Phi_1 B^m - \cdots - \Phi_{Pm} B^{Pm})(1 - B)^{Dm} S_t \\ &\quad = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)\left(1 - \Theta_1 B^m - \cdots - \Theta_{Qm} B^{Qm}\right)\varepsilon_t \qquad | \, \phi_0, \theta_0, \Phi_0, \Theta_0 = 0\end{aligned} \tag{18}$$

jossa yleisesti $B^n S_t = S_{t-n}$ ja $B^n \varepsilon_t = \varepsilon_{t-n}$, eli B^n siirtää mittapistettä S_t tai virhettä ε_t n-periodia taaksepäin. Peruskaavasta pystytään johtamaan mikä tahansa kausittainen tai kaudeton *ARIMA*-malli merkitsemällä ylimääräiset komponentit nollalla, jolloin malli automaattisesti sievenee itsestään jättäen ei tahdotut mallin ulkopuolelle.

2.7 Iterointi

Iteraatiossa halutut parhaat mahdolliset mallin kerrointen arvot pyritään löytämään approksimoimalla kokeellisesti mallin sovitetta eri kerrointen arvoilla yhden ja miinus yhden väliltä. Iteroinnissa pyritään minimoimaan varsinaisten datapisteiden, sekä mallin välisten approksimoitujen data pisteiden erotuksen välistä pienintä neliösummaa (RSS) (Zaiontz 2021),

$$\sum_{i=1}^{n} (S_i - \hat{S}_i)^2 \tag{19}$$

jolloin malli olisi mahdollisimman lähellä alkuperäisiä datapisteitä (Box & Jenkins 1976, 210–223), jolloin mallin sovite olisi optimaalinen ja kerrointen arvot samat kuin optimaalisessa mallissa. Iteroinnin alku arvaukseksi kannattaa asettaa mallin kerrointen teoreettiset arvot, jotka ovat monesti hyvin lähellä todellisia parhaimman sovitteen antavia *ARIMA*-mallin kerrointen arvoja (Box & Jenkins 1976, 19, 210–223).

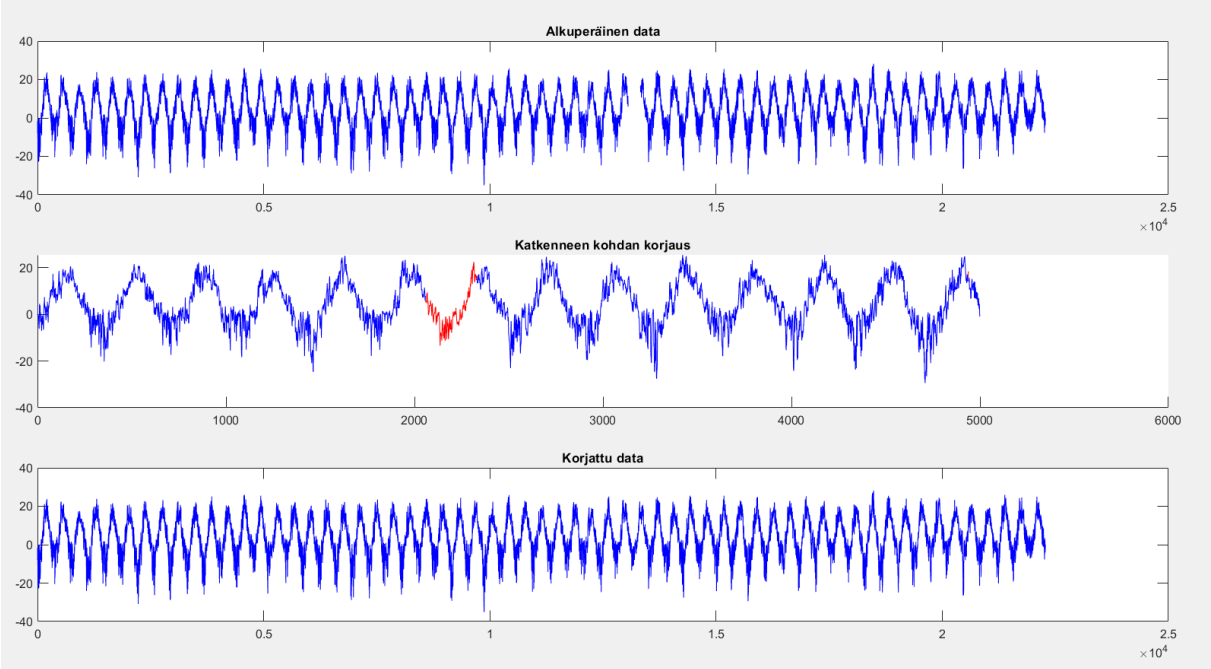
3. Data ja metodologia

Tässä tutkimuksessa esimerkkidatana aikasarja mallien rakentamisessa ja testaamisessa käytettiin Lappeenrannan lentokentän säähavaintoaseman mittaamaa ilmanlämpötila dataa (Ilmatieteenlaitos 2021). Tutkimusaineiston data kattaa päivittäin kello 00:00 kerätyt mittahavaintoarvot 61 vuoden jaksolta aikavälillä 1.1.1960-31.12.2020. Tutkimusaineiston muokkaamiseen ja matemaattiseen tarkasteluun on käytetty MATLAB-ohjelmistoa. Mallin rakentamiseen käytettiin ensimmäistä 59 vuotta data ketjusta, jolloin varsinaisen ennusteen validiteetin tarkasteluun jäi 2 vuoden verran mittapisteitä. Seuraavassa esitetyn ARIMA-mallin koodi sekä käytetty lämpötiladata ovat saatavilla GitHubissa nimellä ”arima-code-for-temperaturedataset” (Liite 1).

3.1. Datan esikäsittely

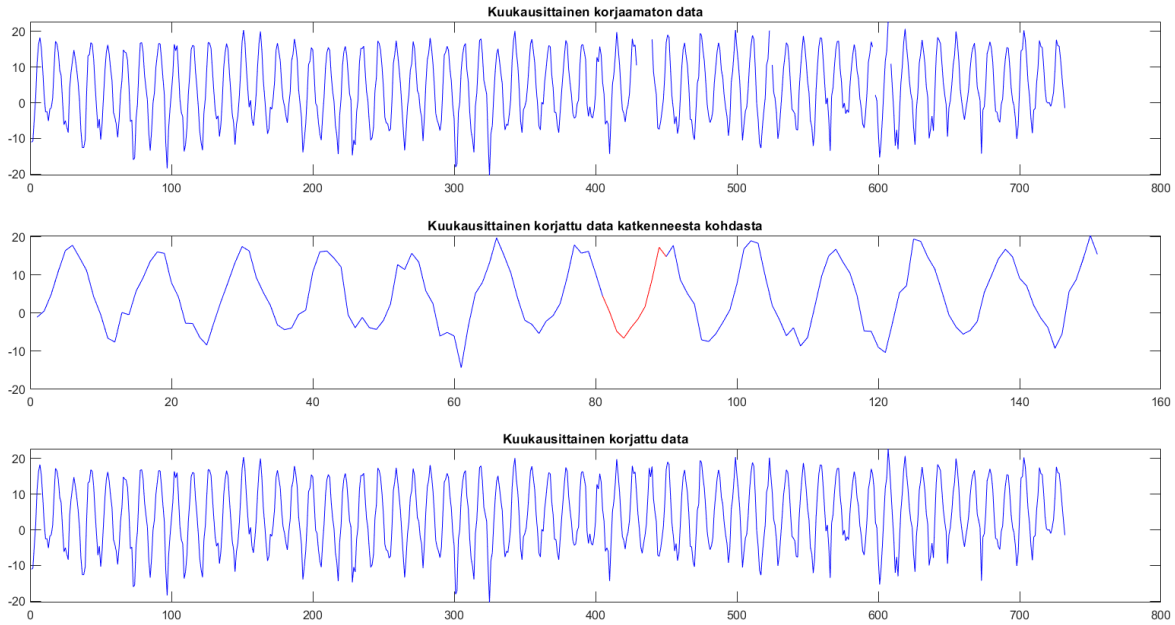
Datasta puuttui joitakin yksittäisiä havaintoarvoja, sekä vuosien 1995 ja 1996 väliltä dataa puuttuu noin puolen vuoden edestä (Kuva 2). Koska *ARIMA*-mallit vaativat täydellisen yhtäjaksoisen dataketjun, puuttuvat havaintoarvot paikattiin datasta. Yksittäisten tai muutamien puuttuvien mittapisteiden kohdat olisi voinut luoda muodostamalla paikallisen ortogonaaliprojektion puuttuvien mittapisteiden tuntumaan ja täten määrittää puuttuville mittapisteille estimoidut korvaavat mitta-arvot. (Van Le H. & Nishio M. 2015) Suurempana ongelmana oli kuitenkin edellä mainittu noin puolenvuoden puuttuvien mittapisteiden periodi (Kuva 2), jolloin jostain syystä säähavaintoasema ei selkeästikään ole ollut käytössä. Tämän ajanjakson puuttuvat mittapisteet, kuten kaikki muutkin yksittäiset puuttuvat mittapisteet luotiin ottamalla edellisen ja seuraavan vuoden vastaavien mitta-arvojen keskiarvo ja käyttämällä tätä uutena puutuvan mittapisteen mitta-arvona (Kuva 2).

Rakennettu sovite ei välttämättä ole paras mahdollinen. Koska luodut datapisteet ovat keskiarvoistettuja, niiden välinen vaihtelu on pienempää, jolloin luotujen mittapisteiden varianssi ei vastaa muiden vastaavien periodien varianssia, vaan on pienempi. Tämän ei kuitenkaan pitäisi haitata varsinaista mallin rakentamista, sillä data käsittää mittapisteitä 61 vuoden ajalta, jolloin luotu puolen vuoden periodi ei vaikuta juurikaan estimoitaviin parametreihin.



Kuva 2: Päivittäisen datan esikatselu

Kirjallisuuskatsauksen perusteella, aikaisempien tutkimusten pohjalta, data muokattiin vielä muotoon, jossa yksi datapiste vastaa yhden kuukauden lämpötilojen keskiarvoa. Kuukausittaisen lämpötiladatan käyttäminen on ollut aikaisemmissa tutkimuksissa paljon yleisempää päivittäiseen lämpötiladataan nähden. Kuukausittain vaihtelevasta datasta huomataan, että alun perin puuttuvien mittapisteiden tilalle luodut mittapisteet näyttäisivät sopivan silmämääräisesti oikein hyvin kuukausittaiseen data sarjaan (Kuva 3).



Kuva 3: Kuukausittaisen datan esikatselu

3.2. Mallin rakentaminen

Tutkimusaineistona käytetty aikasarja on ajateltu koostuvan trendistä (trend), kausittaisesta osasta (seasonality) sekä kohinasta (white noise) (Bouznad, Guastaldi et al. 2020; Ye, Yang et al. 2013). Lisäksi datasarja oletetaan olevan additiivinen (additive), jolloin aikasarja pystyttäisiin ajattelemaan trendin, kausittaisen osan sekä kohinan summana,

$$S = trend + seasonality + white\ noise \tag{20}$$

jonka seurauksena aikasarjaan pystytään käyttämään yksinkertaistettua dekomponointia (decomposition) (Alsuhail & Kokkinen 2005), jossa aikasarjasta erotetaan trendi sekä kausivaihtelu ja jäljelle jäävä kohina (Bouznad, Guastaldi et al. 2020; Ye, Yang et al. 2013), jota voidaan

ajatella alkuperäisen aikasarjan stokastisena muotona (Ye, Yang et al. 2013), pyritään mallintamaan *ARIMA* sovitteella.

3.2.1. Trendin poistaminen

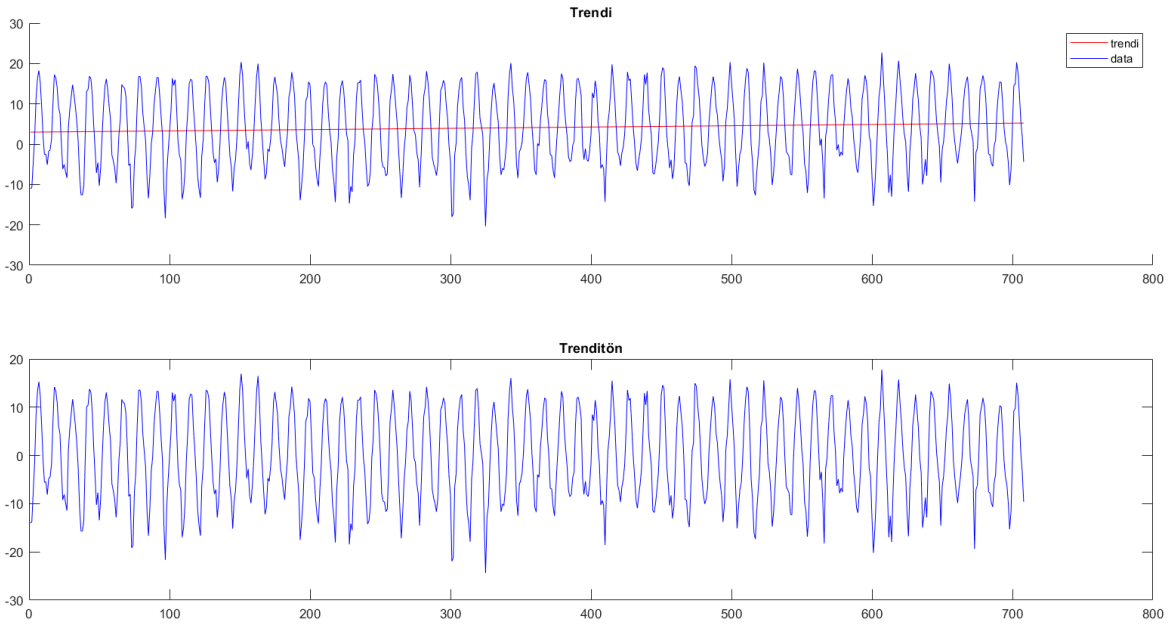
Data-aineistosta pystyttiin määrittelemään loivasti nouseva trendisuora ortogonaaliprojektiolla,

$$x = (A'A)^{-1} \times A'b \tag{21}$$

jossa pystyvektori *b* on varsinaiset data pisteet, jolle projektio määritetään, matriisissa *A* ensimmäinen vektorin *b* pituinen pystyvektori on täynnä ykkösiä ja toinen pystyvektori käsittää vektorin *b* indeksin, (Zarei, Moghimi 2019). Varsinainen projektio trendin määrittämiseksi muodostuu, kun matriisi *A* kerrotaan määritetyillä kertoimilla *x*,

$$Ax = trend \tag{22}$$

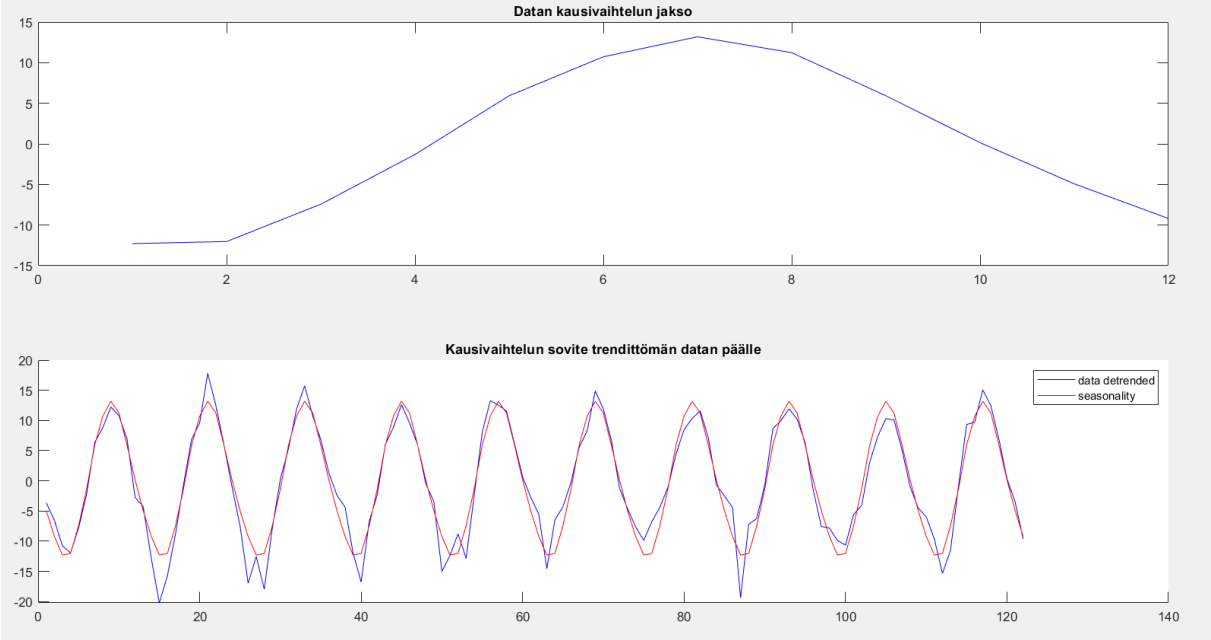
Suora näyttäisi sopivan silmämäärisesti melko hyvin dataan (Kuva 4), vaikka todellisuudessa Lappeenrannan keskilämpötilan nousu tuskin noudattaa täysin lineaarista suoraa, vaan hieman eksponentiaalisesti kasvavaa käyrää. Suoran *x*-akselin leikkauspisteeksi saatiin noin 3 celsius astetta, ja kulmakertoimeksi 0.0032.



Kuva 4: Trendin poistaminen

3.2.2. Kausivaihtelun poistaminen

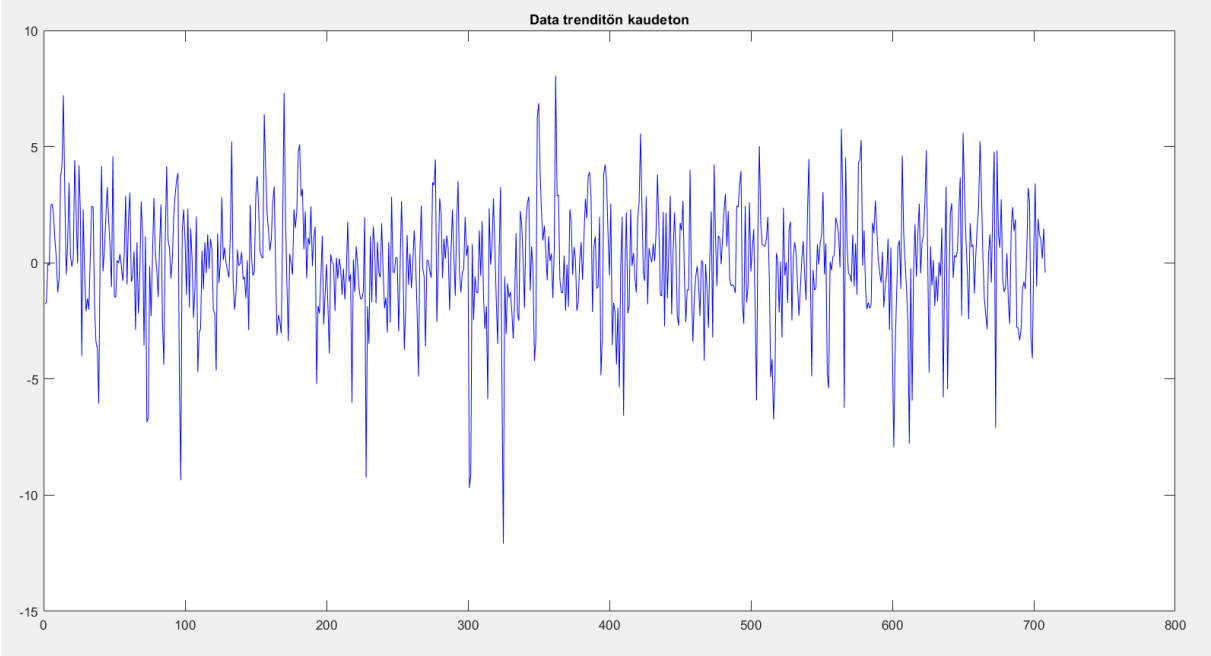
Kausivaihtelun jakson on odotettu olevan 12 kuukautta data aineiston muokatun version käsittäessä kuukausien keskiarvoistettujen lämpötilojen mittapisteet, (Kuva 5). Keskimääräisen kausivaihtelun periodi on saatu laskemalla kaikkien vastaavien kuukausien mittapisteiden keskiarvo, eli esimerkiksi tammikuun mitta-arvo periodissa on saatu laskemalla kaikkien data-aineiston tammikuiden mittapisteiden keskiarvo. Tavan ongelmana on, ettei se ota huomioon sitä, että hyvin todennäköisesti pitkä ketjuisen datan periodin kausivaihtelu ei ole sama datan alkupäässä verrattuna loppuun, koska data käsittää mittapisteitä niin pitkältä aikaväliltä, jolloin mahdollisten muutosten todennäköisyys lisääntyy.



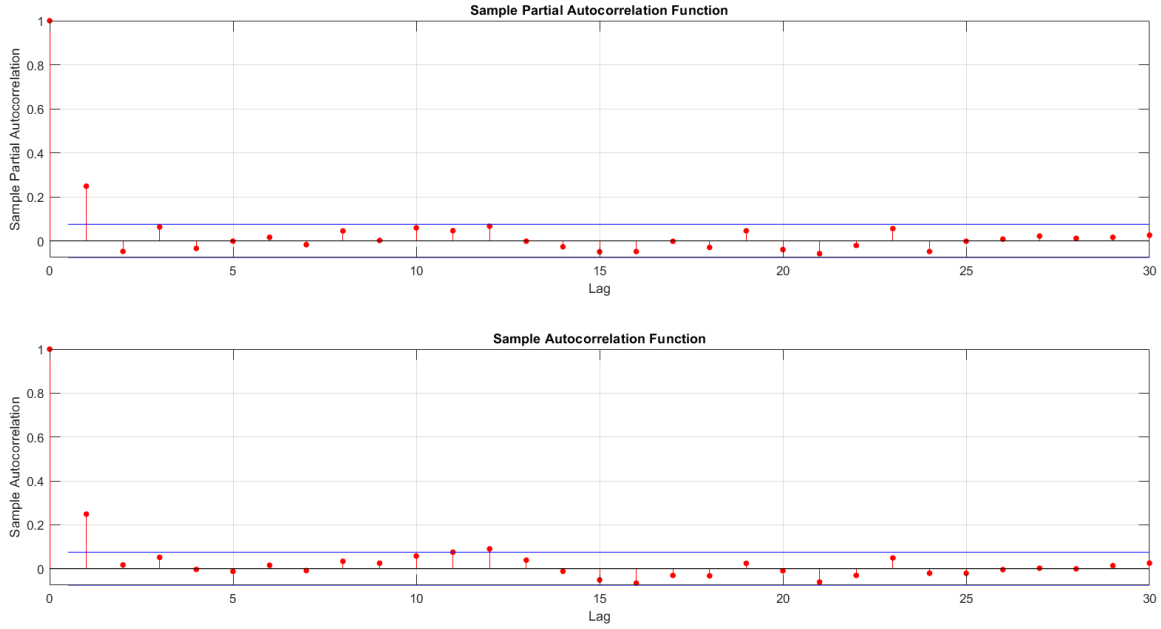
Kuva 5: Kausivaihtelun poisto

3.2.3. ARIMA(p,d,q)(P,D,Q)$_m$ sovite jäännöstermeihin

Kausivaihtelun poiston jälkeiset jäännöstermit (residuals) (Kuva 6), joihin *ARIMA*-malli rakennetaan, täyttivät *ARIMA*-mallin vaatiman stationaarisuus, että käännettävyys ehdot, (Box & Jenkins 1976) MATLAB:in sisäänrakennetun Dickey-Fuller-testin (Augmented Dickey-Fuller-test) hylätessä yksikköjuuren olemassaolon vaihtoehtoisen mallin hyväksi (MathWorks 2009). Samaa tulosta puoltavat myös ACF sekä PACF kuvaajat, (Kuva 7), joissa sekä autokorrelaatiot että osittaisautokorrelaatiot laskevat hyvin nopeasti 0.05 merkitsevyysrajatason alle ja jatkavat vaihtelua negatiivisten sekä positiivisten arvojen välillä, (Box & Jenkins 1976, 179-187).



Kuva 6: Stokastisen aikasarjan esikatselu



Kuva 7: ACF- ja PACF-kuvaajat

ACF sekä PACF kuvaajista pääteltynä, (Kuva 7) autoregressiivisen osan asteluku *p* saa joko arvon 0 tai 1 ja liukuvan keskiarvon asteluku *q* arvon 0 tai 1. ACF-kuvaajasta voidaan myös huomata että 12. viivästetyn autokorrelaation arvo on yli merkitsevyysrajatason, joka tarkoittaisi mahdollisen kausittaisen liukuvan keskiarvon sisällyttämisen malliin 12 kuukauden periodilla.

Tieto, että kuukaudet vaihtelevat 12 kuukauden sykleissä, ja etenkin Suomessa, jossa eri kuukausien lämpötilat voivat erota huomattavasti toisistaan (Kuva 2, Kuva 3), puoltaisi ACF-kuvaajasta saatua päätelmää peräkkäisten vuosien vastaavien kuukausien vaikutuksesta toistensa ennustettavuuteen. Näin ollen kausittaisen liukuvan keskiarvon asteluvun *Q* saa joko arvon 0 tai 1, kun periodin pituus *m* on 12. Sopivin malli kohinan ennustamiseen tässä tapauksessa olisi joko *ARMA(1,1)*, *AR(1)*, *MA(1)*, $SARIMA(1,0,1)(0,0,1)_{12}$, $SARIMA(1,0,0)(0,0,1)_{12}$ tai $SARIMA(0,0,1)(0,0,1)_{12}$ (Box & Jenkins 1976, 186).

Parhaaksi malliksi valikoitui $SARIMA(0,0,1)(0,0,1)_{12}$,

$$(1-\phi_0 B^0)(1-B)^0(1-\Phi_0 B^{0*12})(1-B)^{0*12}a_t = (1-\theta_1 B)(1-\Theta_1 B^{1*12})\varepsilon_t \quad | \varphi_0, \theta_0, \Phi_0, \Theta_0 = 0$$

jossa a_t tarkoitetaan muodostuvan mallin datapistettä, joka merkkaa kohinan mallin mittapistettä, josta kun sijoitetaan ja sievennetään, kaava muuttuu muotoon,

$$a_t = (1-\theta_1 B-\Theta_1 B^{12}+\theta_1\Theta_1 B^{13})\varepsilon_t$$

eli,

$$a_t = \varepsilon_t-\theta_1\varepsilon_{t-1}-\Theta_1\varepsilon_{t-12}+\theta_1\Theta_1\varepsilon_{t-13}$$

jolloin varsinainen malli on,

$$\hat{a}_t = -\theta_1\varepsilon_{t-1}-\Theta_1\varepsilon_{t-12}+\theta_1\Theta_1\varepsilon_{t-13} \quad (23)$$

koska periodin virhetermiä ei pystytä tietämään ennen kuin mittapiste on jo tapahtunut. $SARIMA(0,0,1)(0,0,1)_{12}$ todettiin parhaaksi malliksi, koska malli antoi pienimmän neliösumman, kun approksimoituja jäännöstermejä verrattiin alkuperäisiin jäännöstermejä. Iteraation seurauksena θ sai arvokseen -0.2609 ja Θ -0.0831, kun näiden teoreettisiksi mitta-arvoiksi, eli iteroinnin lähtöarvoiksi saatiin -0.2660 ja -0.0908, kun

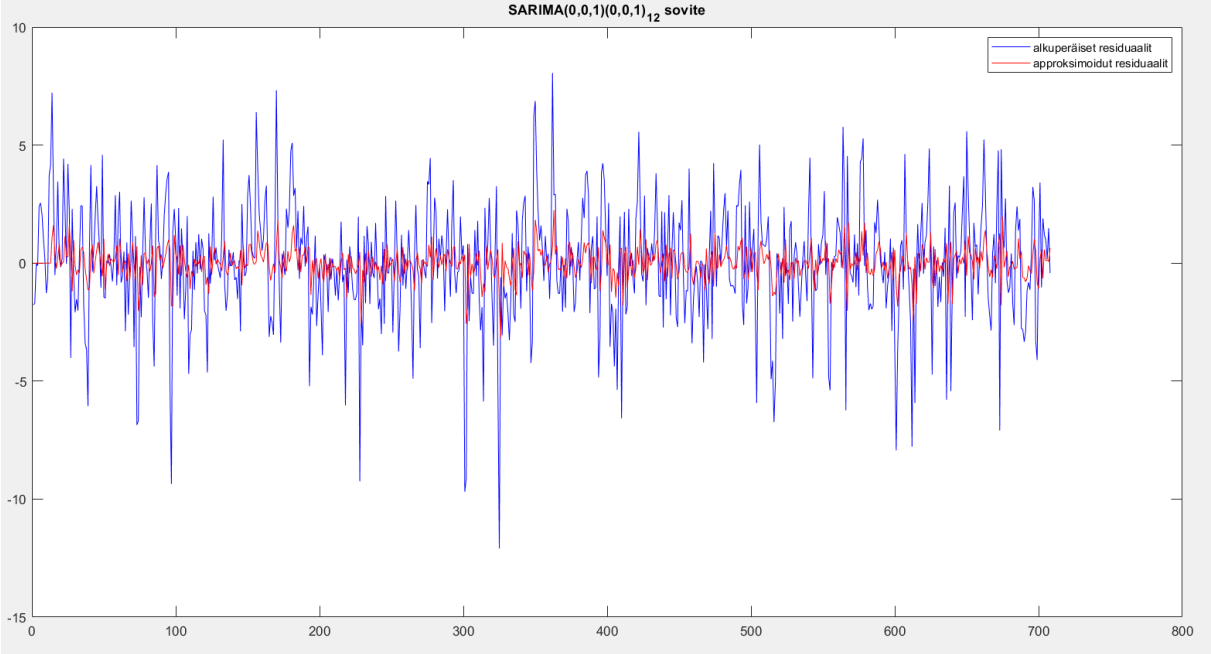
$$r_1 = \frac{-\theta}{1+\theta^2} \quad | \, r_1 = 0.2484, \quad -1 < \theta < 1 \tag{24}$$

$$r_{12} = \frac{-\Theta}{1+\Theta^2} \quad | \, r_{12} = \ 0.0900, \quad -1 < \Theta < 1 \tag{25}$$

Lopullinen malli saa siis muodon,

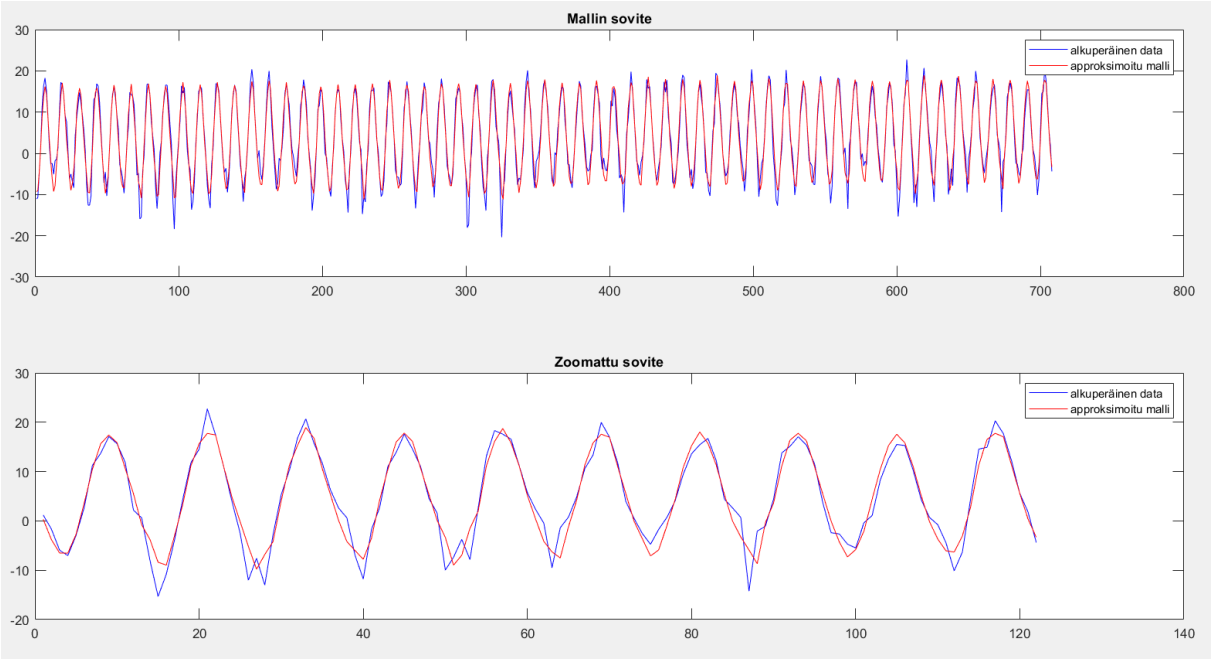
$$\hat{a}_t = 0.2609\varepsilon_{t-1} + 0.0831\varepsilon_{t-12} + 0.0217\varepsilon_{t-13} \tag{26}$$

Juuritettu keskimääräinen neliöity virhe (RMSE, root mean squared error) mallilla on 2.4154, eikä malli näyttäisi osuvan jäännöstermien ääriarvoihin juuri ollenkaan (Kuva 8).



Kuva 8: Kohinan sovite

Varsinainen malli, (20) soveltuu melko hyvin alkuperäiseen dataan silmämääräisesti (Kuva 9), mutta se ei tavoita alkuperäisen datan ääriarvoja, koska *SARIMA(0,0,1)(0,0,1)*$_{12}$-malli ei pystynyt tavoittamaan jäännöstermien ääriarvoja riittävällä tarkkuudella.

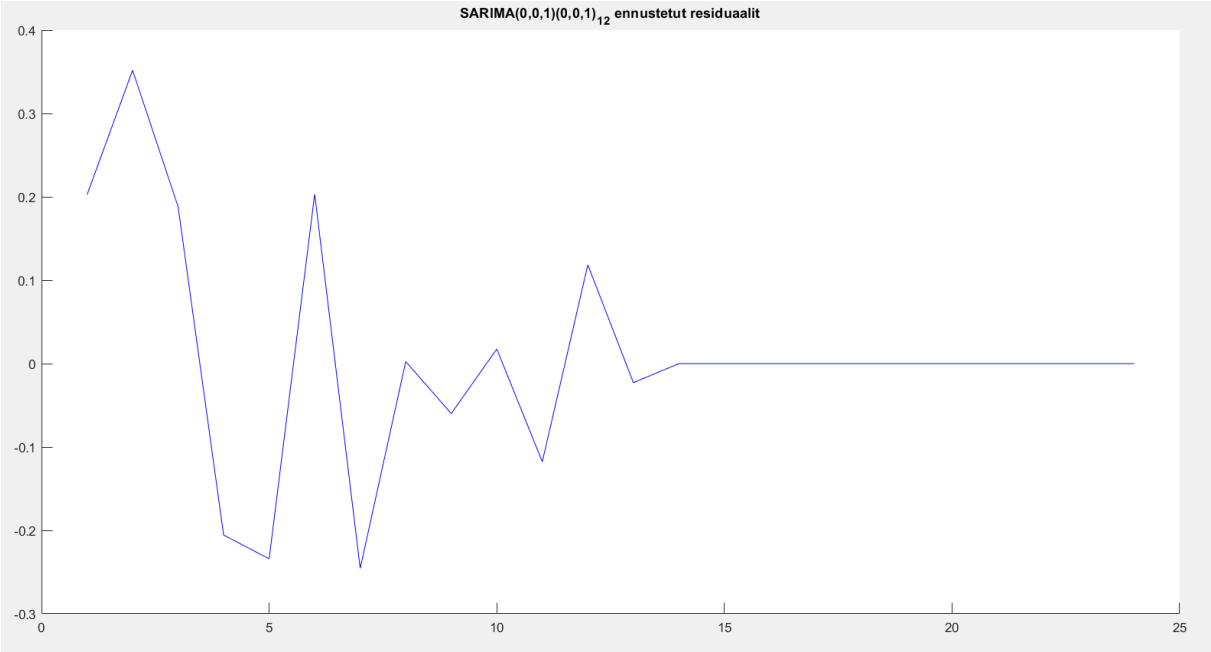


Kuva 9: Mallin sovite

3.3 Ennusteen rakentaminen

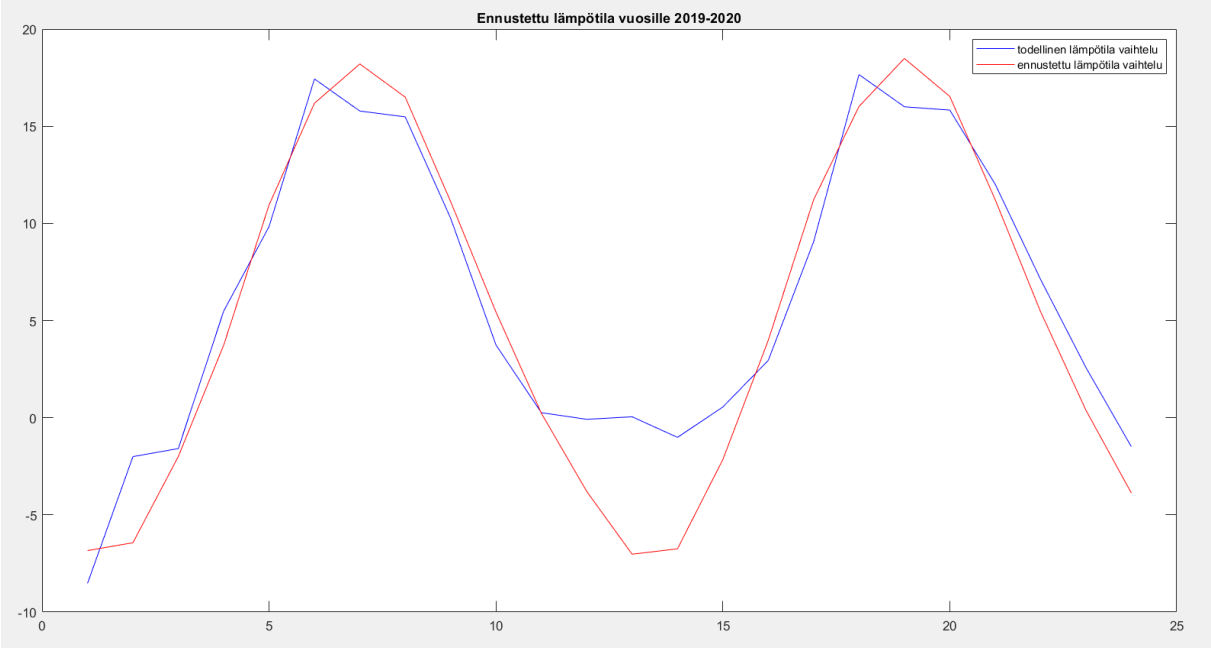
Mallin rakentamisessa on käytetty yksinkertaistettua dekomponointia (Bouznad, Guastaldi et al. 2020; Ye, Yang et al. 2013) (20), ja mallin on ajateltu olevan additiivinen. Tämän seurauksena tulevaisuuden ennuste on rakennettu käänteisesti, liittämällä jo estimoidut mallin palat (Bouznad, Guastaldi et al. 2020; Ye, Yang et al. 2013), ennusteessa yhteen, sovittaen trendisuora jatkumaan indeksiltään myös tulevaisuuteen (22), sekä ennustaessa jäännöstermejä *ARIMA(0,0,1)(0,0,1)*$_{12}$-mallilla (26).

Box & Jenkins (1976, 309) mukaan *ARIMA* ennusteita voidaan pitää teoreettisesti luotettavina aivan maksimissaan niiden parametrien lukumäärän verran. *SARIMA(0,0,1)(0,0,1)*$_{12}$ tapauksessa malli tarjoaisi luotettavan ennusteen maksimissaan 13 ennustetun jäännöstermin pisteen verran, mutta kuten kohinan *SARIMA(0,0,1)(0,0,1)*$_{12}$-sovitteesta nähdään, malli kuolee (dampening) melko nopeasti. Kun malli ei pääse päivittämään itseään, virhetermit, joista sekä liukuva keskiarvo että kausittainen liukuva keskiarvo lasketaan ovat nollia, jolloin voidaan huomata myös ennustettavien jäännöstermien lähestyvän nollaa, (Kuva 10). Ennusteessa ensimmäisenä 13 tarvittavana virheterminä on käytetty alkuperäisen datan, sekä luodun mallin välisiä 13 viimeistä virhetermiä.

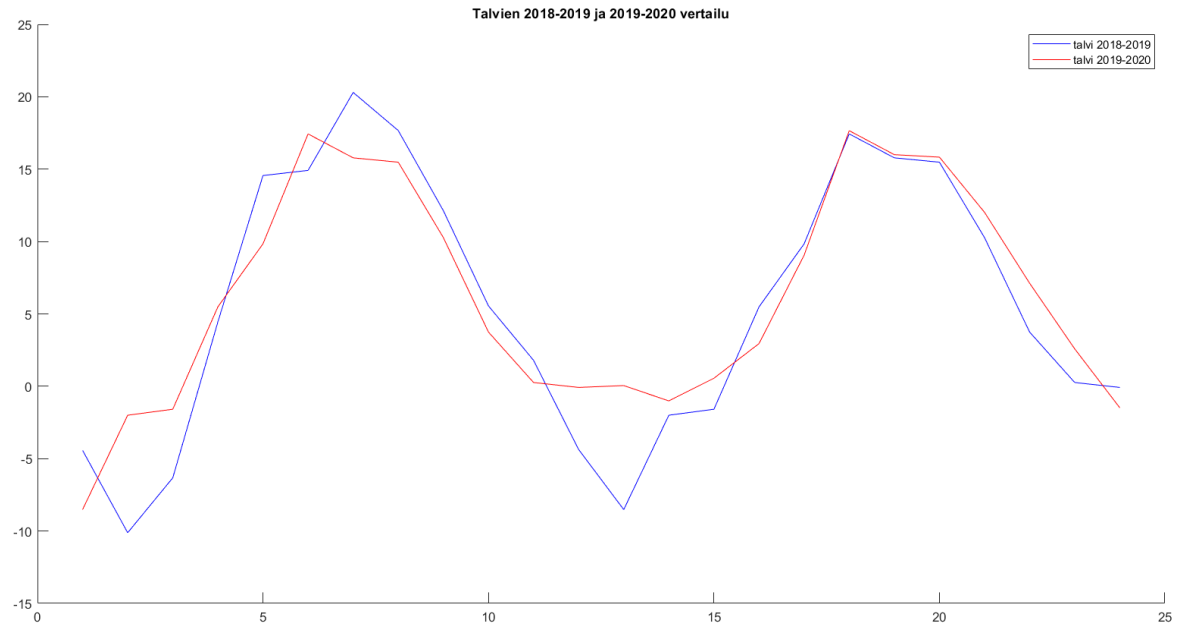


Kuva 10: Kohinan ennuste

Varsinaisen ennuste ei osu kovin hyvin vuosien 2019 ja 2020 datapisteisiin, (Kuva 11). Ennusteen juuritettu keskimääräinen neliöity virhe (RMSE) saadaan 2.6767, joka on noin 11 % suurempi kuin mallin juuritettu keskimääräinen neliöity virhe. Kun verrataan vuosien 2018 ja 2019 välistä talvea, voidaan todeta, että vuosien 2019 ja 2020 välinen talvi on ollut harvinaisen leuto (Kuva 12), joka voidaan todeta myös ilmatieteenlaitoksen jäätalvi-raportista (Vainio, 2021), ja on näin ollen ollut omiaan aiheuttamaan selkeän mitta eron ennusteen ja tapahtuneen välillä.



Kuva 11: Mallin ennuste



Kuva 12: Talvien 2018–2019 ja 2019–2020 vertailu

4. Tulokset

Työ kävi läpi *ARIMA*-mallien yleisen muodostamisen pintapuolisesti, määritti tarvittavat kertoimet, jotka laskettiin auki esimerkkidatan avulla ja pohti lasketun mallin pohjalta muodostetun ennusteen tarkkuutta ja sen luotettavuuden pituutta, niin kuin alun perin oli tarkoituskin. Tutkimus onnistui löytämään vastauksen kumpaankin esitettyyn tutkimuskysymykseen, vaikka varsinainen malli ei onnistunut ennustamaan vuosien 2019 ja 2020 keskilämpötiloja niin hyvin kuin olisi ehkä toivottu.

4.1 Tutkimuskysymyksiin vastaaminen

RQ1: *”Kuinka ARIMA-mallin matemaattisen esitysmuodon tarvitsemat kertoimet saadaan määriteltyä datasta?”*

ARIMA-mallien kertoimien ratkaisemisen osalta yleisesti järkevin ratkaisu on iterointi, jota voidaan nopeuttaa laskemalla kertoimille teoreettiset arvot, jotka ovat monesti melko lähellä iteroinnilla saaduista parhaista arvoista. Mallin itsensä toimintaperiaate perustuu aikasarjan stokastisen muodon mallintamiseen ja ennustamiseen, sen aikaisempien arvojen (*AR*), sekä

mallin ja varsinaisten datapisteiden välisen jäännöstermien aikaisempien arvojen (*MA*) pohjalta. Malliin saadaan sisällytettyä tarvittaessa samalla periaatteella myös kausittaiset komponentit, jotka määrittelevät mallin ennustettavuutta aikaisempien periodien tapahtumien perusteella. Ennusteen voidaan teoreettisesti olettaa luotettavaksi sen parametrien lukumäärän verran, jolloin teoreettisesti *ARIMA*-mallilla ennustettavia mittapisteitä voidaan pitää luotettavana p + q + Pm + Qm kappaleen verran (Box & Jenkins 1976, 309), joka ei yleensä vastaa ihan täysin todellisuutta.

5.2 RQ2: *”Kuinka ARIMA-mallit soveltuvat lämpötiladatan tutkimiseen ja millainen on Ilmatieteenlaitoksen Lappeenrannan lentokentältä kerätyn ilmanlämpö case-aineiston paras stokastisen sarjan ARIMA-sovite valitulle aikavälille?”*

Kirjallisuuskatsauksen pohjalta voidaan sanoa, että *ARIMA*-sovitteita käytetään yleisesti myös lämpötiladatan tutkimiseen monien muiden aikasarjojen mallintamisen ohella. Siihen onko *ARIMA*-malli paras käytetty malli lämpötiladatan tutkimisessa, tässä tutkimuksessa ei oteta kantaa muuten kuin toteamalla, että ainakaan tutkimuksessa käytettyyn case-aineistoon rakennettu aikasarja sovite ei varmastikaan ole paras mahdollinen. Yleisesti, kuten myös tämän tutkimuksen data muokattiin, *ARIMA*-malleja ja näistä rakennettuja hybridi malleja, on lähtökohtaisesti käytetty kuukausien keskilämpötilojen ennustamiseen kiitettävin tuloksin, mutta löytyy myös tutkimuksia, jossa *ARIMA* pohjaisia malleja on käytetty vuorokauden keskilämpötilojen ennusteina.

Case-aineiston parhaaksi stokastiseksi *ARIMA* sovitteeksi saatiin pienintäneliösummaa minimoimalla $SARIMA(0,0,1)(0,0,1)_{12}$-malli. Sovite ei valitettavasti kyennyt ennustamaan poikkeuksellisen leutoa talvea, jonka seurauksena sovitetta ei kyetä mieltämään parhaaksi mahdolliseksi malliksi. Alkuperäiseen datasarjaan olisi saattanut sopia paremmin suora *SARIMA*-sovite jättäen pois dekomponointia hyödyntäneen deterministisen osan, määrittäen myös tämän yhdessä *SARIMA*-mallissa.

4.2 Tutkimuksen rajoitteet ja jatkotutkimusaiheita

Tutkimuksen suurimpana rajoitteena voidaan pitää aikasarjoille tyypillisen tilastollisen merkittävyyden tarkastelua. Tämä tutkimus ei ota kantaa siihen, onko jokin määritetyistä parametreistä tilastollisesti merkitsevä, eli eroaako se merkittävästi nollasta, vai ei. Toisena rajoitteena mallin matemaattisessa taustassa voidaan pitää stationaarisuusehdon tarkastelua. Esimerkki mallin sovitteessa data on stationaarinen, mutta sitä ei ole todistettu teoreettisesti, eikä laskettu itse auki, vaan tässä tutkimuksessa on tyydytty kuittaamaan stationaarisuusehdon täyttyminen viittaamalla MATLAB:in sisään rakennettuun Dickey-Fuller-testiin (MathWorks 2009). Tutkimuksessa stationaarisuus päädyttiin todistamaan melko kevyesti soveltavien *ARIMA*-malleja käyttävien tutkimusten selittäessä stationaarisuusehdon täyttymistä teoreettisesti kiitettävästi.

Tämän tutkimuksen tarkastellessa *ARIMA*-mallien parametrien määrittämistä niiden teoreettisen muodostamisen ja iteroinnin pohjalta, voisi jatko tutkimus mallin matemaattisen taustoittamisen osalta keskittyä parametrien ratkaisuun differentiaaliyhtälöiden pohjalta. Myös luottamusvälien (confidence interval), sekä stationaarisuuden osoittaminen ja laskeminen voisivat olla matemaattisen taustan osalta jatkotutkimusaiheita. Ilmanlämpötiladataan rakennettavan mallin osalta jatkotutkimuskohteita voisi olla Suomen ilmanlämpötila dataan sovitettavien mallien rakentaminen pidemmällä aikavälillä, sillä kaikissa tämän tutkimuksen pohjana käytetyissä tutkimuksissa lämpötiladata näytti olevan melko tasajakautunutta, ilman suuria periodien välisiä ääriarvojen eroja, jotka taas Lappeenrannan lentokentän tuottamissa sää havaintoarvoissa olivat merkittäviä pahimmillaan jopa liki parikymmentä astetta.

Lähdeluettelo

ALSUHAIL, F. & KOKKINEN, A., 2005. Aikasarjan ARIMA-pohjaisesta kausitasoituksesta. *Kansantaloudellinen aikakauskirja*, **101**, pp. 469–483

BOUZNAD, I.-., GUASTALDI, E., ZIRULIA, A., BRANCALE, M., BARBAGLI, A. and BENGUSMIA, D., 2020. Trend analysis and spatiotemporal prediction of precipitation, temperature, and evapotranspiration values using the ARIMA models: case of the Algerian Highlands. *Arabian Journal of Geosciences,* **13**(24).

BOX, G. & JENKINS, G., 1976 TIME SERIES ANALYSIS forecasting and control. San Fransisco, Holden-Day Inc.

ILMATIETEENLAITOS, 2021. Havaintojen lataus. [Verkkodokumentti]. [Viitattu 16.9.2021]. Saatavilla: https://www.ilmatieteenlaitos.fi/havaintojen-lataus

ISLAM, A. R. M. T., KARIM, M.R. and MONDOL, M.A.H., 2021. Appraising trends and forecasting of hydroclimatic variables in the north and northeast regions of Bangladesh. *Theoretical and Applied Climatology,* **143**(1–2), pp. 33–50.

JAMK, 2021. Opinnäytetyön ohjaajan käsikirja. [verkkodokumentti]. [Viitattu 12.9.2021]. Saatavilla: https://oppimateriaalit.jamk.fi/yamk-kasikirja/kirjallisuuskatsaukset/

KESAVAN, R., MUTHIAN, M., SUDALAIMUTHU, K., SUNDARSINGH, S. and KRISHNAN, S., 2021. ARIMA modeling for forecasting land surface temperature and determination of urban heat island using remote sensing techniques for Chennai city, India. *Arabian Journal of Geosciences,* **14**(11).

MATHWORKS, 2009. adftest. [verkkodokumentti]. [Viitattu 16.9.2021]. Saatavilla: https://se.mathworks.com/help/econ/adftest.html

SHIRVANI, A., NAZEMOSADAT, S.M.J. and KAHYA, E., 2015. Analyses of the Persian Gulf sea surface temperature: prediction and detection of climate change signals. *Arabian Journal of Geosciences,* **8**(4), pp. 2121-2130.

VAINIO, J., 2021. Kaikkien aikojen leudoin jäätalvi 2019–2020. Ilmatieteenlaitos. [verkkodokumentti]. [Viitattu 12.9.2021]. Saatavilla: https://www.ilmatieteenlaitos.fi/jaatalvi-2019-2020

VAN LE, H. and NISHIO, M., 2015. Time-series analysis of GPS monitoring data from a long-span bridge considering the global deformation due to air temperature changes. *Journal of Civil Structural Health Monitoring,* **5**(4), pp. 415-425.

WANG, H., HUANG, J., ZHOU, H., ZHAO, L. and YUAN, Y., 2019. An integrated variational mode decomposition and ARIMA model to forecast air temperature. *Sustainability (Switzerland),* **11**(15).

YE, L., YANG, G., VAN RANST, E. and TANG, H., 2013. Time-series modeling and prediction of global monthly absolute temperature for environmental decision making. *Advances in Atmospheric Sciences,* **30**(2), pp. 382-396.

ZAIONTZ, C., 2021. Calculate ARMA(p,q) coefficients using Solver. REAL STATISTICS USING EXCEL. [verkkodokumentti]. [Viitattu 16.9.2021]. Saatavilla: ARMA coefficients via Solver | Real Statistics Using Excel (real-statistics.com)

ZAREI, A.R. and MOGHIMI, M.M., 2019. Environmental assessment of semi-humid and humid regions based on modeling and forecasting of changes in monthly temperature. *International Journal of Environmental Science and Technology,* **16**(3), pp. 1457-1470.

Liitteet

Liite 1: Työssä rakennettu malli sekä rakentamiseen käytetty data

https://github.com/AndroidAPa/ARIMA-CODE-FOR-TEMPERATUREDATASET.git