Autoregressiiviset mallit sähkön kulutuksen ennustamisessa

Pro gradu -tutkielma
Turun yliopisto
Tulevaisuuden teknologioiden laitos
Data-analytiikka
2019
Altti Tammi

TURUN YLIOPISTO
Tulevaisuuden teknologioiden laitos

ALTTI TAMMI: Autoregressiiviset mallit sähkön kulutuksen ennustamisessa

Pro gradu -tutkielma, 65 s.
Data-analytiikka
Syyskuu 2019

Sähkön kulutuksen tarkka ennustaminen on tärkeää, jotta voidaan turvata sähköverkon toimivuus ja kuluttajille halpa sähkön hinta. Lyhyen aikavälin ennusteet ovat energia-alalle tärkeimpiä päivittäisessä operatiivisessa toiminnassa, sillä sähkön osto tapahtuu yleensä tunneittain seuraavalle päivälle ja väärin arvioitu kulutus aiheuttaa lisäkuluja, jotka ovat sitä suurempia, mitä huonommin kulutus on arvioitu.

Tässä tutkielmassa esitellään aikasarja-analyysia yleisesti, Suomen sähköverkon erityispiirteitä, alan tutkimusta ja Box–Jenkinsin autoregressiivisiä malleja. Tutkielmassa pyritään myös ennustamaan sähkönkulutusta autoregressiivisillä malleilla (ARIMA) käyttäen apuna lämpötilatietoja.

Aineistona on käytetty sähkön kokonaiskulutusta koko Suomessa vuosilta 2004–2011 sekä Hämeenlinnan lämpötilatietoja samalta ajalta. Tulokset osoittavat, että malli sopii ongelmaan kohtalaisesti, mutta ARIMA-malli ei osaa täysin kuvata kaikkia sähkönkulutuksen erityispiirteitä, kuten useaa eri kausittaisuutta, yösähkötariffin alkamista ja työviikon vaihteluita.

Asiasanat: ARIMA-malli, sähköverkot, sähkönkulutus, ennustaminen, sää, lämpötila, STLF

UNIVERSITY OF TURKU
Department of Future Technologies

ALTTI TAMMI: Autoregressiiviset mallit sähkön kulutuksen ennustamisessa

Master of Science Thesis, 65 p.
Data analytics
September 2019

Accurate electric load forecasting is important to guarantee the functionality of the power grid and a low cost of electricity for the consumers. Short term load forecasting is important for the operation of a system operator, because electricity is bought by the hour for the next day and a bad load forecast causes extra costs that are directly proportional to how bad the forecast was.

In this thesis I demonstrate time series analysis in general, special features of the Finnish power grid, research on load forecasting and the Box–Jenkins method of time series analysis. An electric load forecast is built with autoregressive models, using temperature information to help the model.

The data used is the total electricity usage in Finland in the years 2004–2011 and the prevailing temperature in Hämeenlinna. The results show that the model fits the problem moderately, but the ARIMA model has problems with some of the special features of electric load, such as multiple seasonalities, the starting time of the night rate and the fluctuations of the working week.

Keywords: ARIMA, electricity, load forecasting, weather, temperature, STLF

Sisältö

1 Johdanto

Sähkön kulutus on yhteiskunnallisesti sekä taloudellisesti merkittävä suure, jonka ennustaminen on oleellista, jotta kuluttajille voidaan taata halpa, laadukas ja katkeamaton sähkön tuotanto. Tasevastaavien täytyy tasata väärin ennakoitu sähkön kulutus ostamalla kalliimpaa nopeasti saatavaa säätösähköä, jos kulutus on liian suuri, tai myymällä jo ostettua sähköä alihintaan, jos ostoja on tehty liikaa. Sähkön kulutus koostuu miljoonista pienistä kuormista, jotka voivat kytkeytyä päälle ja pois milloin vain. Aurinkosähkö poislukien aina kun joku aktivoi sähkölaitteen, jossain päin maata täytyy pyöriä generaattorissa magneetti, joka tuottaa sähköverkkoon tarvittavan määrän sähköä.

Koko maan mittakaavassa liian suuri sähkön kulutus hidastaa tuotantolaitosten generaattoreita, mikä taas laskee sähkön jännitettä ja pahimmassa tapauksessa aiheuttaa sähkökatkoja ja laiterikkoja. Liian suuri tuotanto taas vastaavasti nostaa jännitettä. Oman lisänsä tuotannon ja kulutuksen tasapainottamiseen tuo se, että tuotantolaitosten päälle- ja poiskytkeytymiseen tarvittava aika vaihtelee. Moderni tietoyhteiskunta on äärimmäisen riippuvainen sähköstä, ja ongelmat sähkön jakelussa pysäyttäisivät palvelut, kirjanpidon, kommunikaation, terveydenhuollon ja maksujärjestelmät.

Nykyaikaisessa ilmastonmuutoksen huomioivassa energiataloudessa tulee olemaan entistä haastavampaa tasapainottaa tuotanto ja kulutus, sillä sähkön käyttö sekä tuotanto on monipuolisempaa. Syinä tähän on muun muassa lisääntyvä aurinkovoiman tuotanto, etäohjattavat ja älykkäät laitteet, talouksien toimiminen pientuottajina, sekä sähköautojen lataus ja mahdollisesti myös näiden akkujen käyttö varavoimalähteenä. Kaikki tämä mo-

nimutkaistaa laitteiden ja ilmiöiden välisiä vuorovaikutuksia ja mahdollisesti heikentää ennustettavuutta.

Sähkön kulutus on tyypillinen *aikasarja*, eli se on ajan mukaan järjestetty sarja tasaisin väliajoin kuten tunneittain, päivittäin, viikoittain tai vuosittain otettuja näytteitä. Aikaisimpia aikasarjoja ovat antiikin ajalla suoritetut väestönlaskut. Nykyään aikasarjoja tuotetaan muun muassa meteorologiaan, markkinoiden tarpeisiin, teollisuuden prosesseihin ja yhteiskunnan tilastoihin. Lämpötilat, pörssikurssit, väkiluku ja työttömyysprosentti ovat useimpien tuntemia esimerkkejä eri tahojen tuottamista aikasarjoista.

Aikasarjat ovat hyvä työkalu muutoksen tutkimiseen ja laadukas katkeamaton aikasarja mahdollistaa myös muutoksen tutkimista tilastollisin menetelmin. Erityisen tärkeä käyttökohde aikasarjoille ovat ennusteet, joita voidaan tuottaa kun aikasarjan pohjalta on saatu luotua *malli*, joka siis on valistunut veikkaus siitä, minkälainen prosessi aikasarjan on tuottanut ja miten prosessiin vaikuttavat muuttujat vaikuttavat toisiinsa.

Aikasarjoja on pyritty selittämään muun muassa ulkoisilla tekijöillä. Jäätelön myyntiä esimerkiksi voitaisiin yrittää selittää kaavalla, joka laskisi myynnin ottamalla huomioon lämpötilan, sateen, viikonpäivän, lomasesongin ja ihmisten ostovoiman yhteisvaikutuksen. Tällaisessa lähestymistavassa on kuitenkin vaarana, että jokin tärkeä selittävä tekijä unohtuu, tai yhteys tekijöiden välillä on vaikea mallintaa.

Toinen tapa analysoida aikasarjoja on havaita, että aikasarjan peräkkäiset arvot eivät ole toisistaan riippumattomia. Sanotaan, että aikasarjassa on *autokorrelaatioita* (kreikan sanasta autós l. itse), eli sarjan arvoilla on vastaavuuksia muiden saman sarjan arvojen kanssa. Arvot riippuvat usein jonkin verran edellisistä arvoista tiettyyn pisteeseen saakka, jota kauempana satunnainen kohina on vahvempi ja peittää alleen autokorrelaatiot. Ottaen lämpötilan esimerkiksi: jos lämpötila on edellisenä tuntina tippunut, on hyvin mahdollista, että lämpötila tippuu myös seuraavanakin tuntina, sillä aurinko aiheuttaa lämpötilaan 24 tunnin jaksollisuuden, jossa lämpötila vuorotellen nousee ja laskee.

Mallia kutsutaan *autoregressiiviseksi*, jos siinä yritetään laskea ennuste sarjan tuleville

arvoille edellisistä arvoista. Metodia, jossa tällaista mallia optimoidaan, kutsutaan *Box–Jenkins*-metodiksi keksijöidensä mukaan. Tässä työssä pyritään luomaan malli ja mallin avulla ennuste sähkön kulutukselle kyseisellä metodilla.

Luvussa 2 tutustutaan sähkön kulutukseen aikasarjanäkökulmasta ja alasta tehtyyn tutkimukseen. Luvussa 3 esitetään autoregressiivisia malleja. Tämän jälkeen luvussa 4 sovitetaan ARIMA-malli Suomen sähkönkulutushistoriaan vuosilta 2004–2011 ja pyritään luomaan ennuste tehtyä mallia ja lämpötilahistoriaa käyttämällä. Lopuksi luvussa 5 esitetään lyhyt yhteenveto tutkimuksen tuloksista.

2 Yleistä sähkönkulutuksen ennustamisesta

2.1 Sähkömarkkinat Suomessa

Kaikki kulutettava sähkö tulee tuottaa sähköverkkoon samalla hetkellä kuin sähkö kulutetaan. Sähkön kulutus vaihtelee arvaamattomasti, sillä sähkölaitteita laitetaan päälle vaihtelevan paljon vaihtelevaan aikaan. Kulutuksen ja tuotannon ollessa epätasapainossa sähkön laadussa ilmenee erilaisia häiriöitä, kuten verkon taajuuden vaihtelua, jännitteen alenemista sekä pahimmassa tapauksessa sähkökatkoja ja laiterikkoja, joten on tärkeää ylläpitää tasapaino kulutuksen ja tuotannon välillä.

Sähkömarkkinoilla toimiminen edellyttää tuotetun/ostetun ja kulutetun/myydyn sähkön tasapainon ylläpitämistä. Käytännössä tähän ei päästä, vaan kullakin taholla täytyy olla avoin toimittaja, joka tasapainottaa sähkötaseen. Suomen valtakunnallisesta sähkötaseen tasapainottamisesta vastaa Fingrid Oyj.

Suomen sähkömarkkinat uudistettiin ja avattiin kilpailulle vuonna 1995. Sähköä tuottavilla yrityksillä ei nykyään enää saa olla sähkönsiirtotoimintoja.

Sähkön kulutustehoa mitataan wateissa (W). Hetkellisen kulutuksen integraali ajan suhteen kertoo tietyssä ajassa käytetyn energian määrän, jonka SI-järjestelmän mukainen mittayksikkö on joule (J), joka on kuluneen energian määrä kun yhden watin teholla työskennellään yhden sekunnin ajan. Sähköenergiaa mitataan kuitenkin yleensä käytännön

syistä kilowattitunteina (kWh), sillä joule on pieni energiamäärä ja kulutusta on järkevämpi mitata tunneittain kuin sekunneittain. Esimerkiksi tunnin saunominen sähkösaunalla, joka keskimäärin käy teholla 5kW, kuluttaa sähköä 5 kilowattituntia, mikä jouleina olisi $1.8 \times 10^7 J$.

Sähköä ostetaan joko Nord Poolin sähköpörssistä tai kahdenkeskisillä sopimuksilla. Nord Pool (ent. Nord Pool Spot) on pohjoismaisten kantaverkko-operaattorien omistama sähkön markkinapaikka. Lisäksi Baltian maiden kantaverkko-operaattorit omistavat Nord Poolista pienet osuudet. Nord Poolin johdannaiskauppa on myyty Yhdysvaltalaiselle Nasdaq-yritykselle ja kulkee nykyään nimellä Nasdaq OMX Commodities Europe. Vuonna 2015 Suomessa kulutettiin Fingridin mukaan n. 81.3 TWh sähköä [1], josta noin 56 TWh eli lähes 70% ostettiin Nord Poolin kautta. [1], [2]. Valtaosa muusta tuotannosta on teollisuuden omistamaa, ja sen tuottama sähkö menee teollisuuden omaan käyttöön. Noin puolet Suomessa kulutetusta sähköstä menee teollisuuden käyttöön. Suurimpia teollisuuden kuluttajia ovat Tornion terästehdas sekä metsä- ja paperiteollisuuspaikkakunnat Lappeenranta, Oulu, Kouvola, Jämsä ja Rauma [3].

Sähkön tuotannosta maksetaan energiaveroa sekä huoltovarmuusmaksua, joista yhdessä arkikielessä käytetään ilmausta sähkövero. Kotitalouksien arvonlisäverollinen sähkövero on 2.79372 snt/KWh ja teollisuudelle vero on 0.87172 snt/KWh. [4]

2.2 Sähkön kulutus

Sähkön kulutuksessa on suuria säännöllisiä sekä epäsäännöllisiä vaihteluja johtuen sekä säästä että ihmisen toiminnasta. Vuotuinen kulutus vaihtelee lämmitys- ja jäähdytystarpeiden mukaan. Kuumilla alueilla sisäilman ilmastointilaitteet kuluttavat paljon sähköä kun taas pohjoisessa kulutus on lämmityksen takia talvella suurempi kuin kesällä. Säätiedon lisääminen auttaa saamaan luotettavampia ennustuksia.

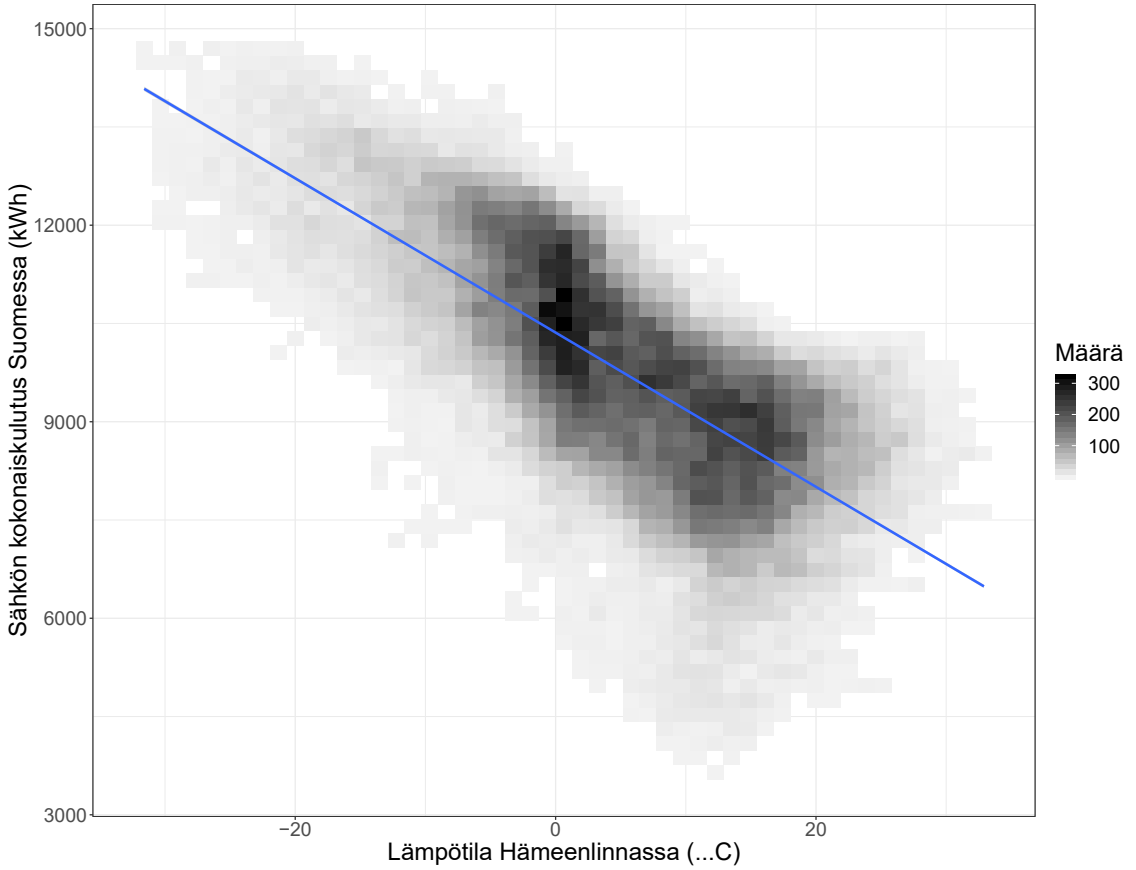
[1] Laskettu vuoden keskiarvokulutuksesta 9282 MWh/h

Lämpötilan lisäksi sähkönkulutusta ennustettaessa hyödyllisiä suureita ovat tuulen nopeus, joka vaikuttaa lämpöhävikkiin, sekä ilman kosteus, josta voidaan laskea pakkasen purevuus (WCI, wind chill index) ja helteen tukaluus (THI, temperature-humidity index). Nämä kuvaavat pelkkää lämpötilaa paremmin lämpötilan vaikutusta ihmiseen. [5], [6]

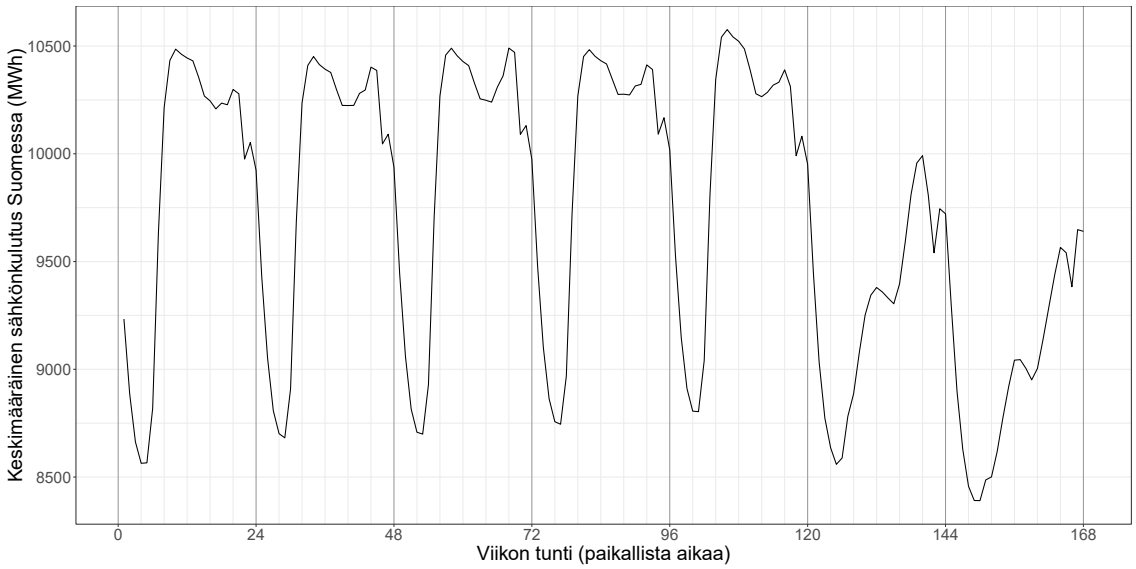
Kuvassa 2.1 esitetään sähkön kulutuksen ja lämpötilan välinen riippuvuus Suomessa. Fingridin ilmoittamaa Suomen kokonaiskulutusta vuosilta 2004-2011 on verrattu Ilmatieteen laitokselta saatuun Hämeenlinnassa vallinneeseen lämpötilaan, ja dataan on sovitettu suora. Kuvasta havaitaan kylmille alueille tyypillinen negatiivinen korrelaatio, missä ilman kylmeneminen lisää sähkön kulutusta lisääntyneen lämmitystarpeen takia. Kuitenkin kuvasta huomataan myös, että tarpeeksi lämpimällä säällä kulutus lähtee hienoiseen nousuun lisääntyneen ilmastoinnin tarpeen vuoksi. Etelä-Euroopan maissa kulutuskäyrä onkin enemmän V:n mallinen ja suurin kulutuspiikki on kesällä. Ilmastonmuutoksen myötä lisääntyvä sisätilan jäähdytyksen tarve saattaa Suomessakin kasvattaa entisestään hellepiikkien sähkönkulutusta.

Vuotuisen eli noin 365.25 päivän jaksollisuuden lisäksi kulutuksessa on viikoittaista ja päivittäistä jaksollisuutta. Arkipäivisin kulutus on suurempaa kuin viikonloppuina, eli kulutuksessa on seitsemän päivän jaksollisuutta. Päivisin kulutus on suurempaa kuin öisin, ja lisäksi ihmisten herääminen ja töistä pääsy aiheuttaa kulutuspiikin, eli kulutuksessa on myös vahva 24 tunnin jaksollisuus. Yön matalan kulutuksen laakso on kestoltaan lyhyempi kuin päivän kulutus, eli päiväsyklisyys ei ole täysin sinimuotoista. Tästä syystä voidaan olettaa, että päiväsyklisyyden kuvaamiseen tarvitaan $1/24$ tunnin signaalin lisäksi tämän harmonisia taajuuksia $1/12h, 1/8h, 1/6h$ jne. Iltaisin on myös kymmenen aikaan päivittäin pieni kulutuspiikki, joka saattaa johtua kotien varaavista lämmityslaitteista, jotka kytkeytyvät päälle yösähkötariffin vaihtuessa päälle (Kuva 2.2). Kansalliset vapaapäivät muuttavat arkipäivän kulutuksen enemmän viikonloppupäivän kaltaiseksi. Kulutus ei yleensä muutu radikaalisti edelliseen vastaavaan päivään nähden.

Sähköä kulutetaan kulloinkin tietyllä teholla ja tätä kulutusta pyritään ennustamaan



Kuva 2.1: Lämpötilan vaikutus sähkön kulutukseen



Kuva 2.2: Viikottaisen kulutuksen keskiarvo

jollekin aikavälille. Kirjallisuudessa puhutaan lyhyen, keskipitkän sekä pitkän aikavälin ennustuksista (short term load forecasting l. STLF, medium term load forecasting l. MTLF sekä long term load forecasting l. LTLF). Mitään selkeää rajaa näille aikaväleille ei ole muodostunut, mutta lyhyen aikavälin ennustaminen on yleensä tunnista viikkoon, keskipitkän on viikosta vuoteen ja yli vuoden päähän tehtävät ennustukset ovat pitkän aikavälin ennustuksia. [6]

Kulutusta ennustettaessa voidaan ennustus muodostaa erikseen eri luokan kuluttajille kuten esim. kotitalouksille, liiketaloudelle ja teollisuudelle. Luokan sisällä kuluttajien sähkönkäyttöprofiili on samankaltainen. Kotitalouksien sähkönkulutus koostuu monesta pienestä kuormasta, jotka tasaavat toisiaan. Liiketalouden sähkönkäyttö on samankaltaista kuin kotitalouksien, mutta kulutus keskittyy aamupäivään. Teollisuudessa on suhteessa edellisiin vähemmän sähköä käyttäviä prosesseja, mutta ne voivat olla paljon suurempia kuin muissa luokissa, mikä lisää ennustettamattomuutta. [5], [6]

Lyhyen aikavälin ennustukset muodostetaan korkeintaan muutaman päivän päähän, yleensä tietylle tunnille tai tunneille. Yksi tärkein ennustettava suure on seuraavan tunnin kokonaiskulutus ja toinen on ennuste seuraavan vuorokauden kaikille tunneille. Lyhyen aikavälin ennustuksia tehdään myös päivittäiselle kulutushuipulle ja seuraavan päivän kokonaiskulutukselle. Kulutusennustusten lisäksi energia-alalla on paljon muitakin ennustettavia suureita, kuten uusiutuvan energian tuotanto, kysyntäjouston käytöntarve, sähkökatkojen tapahtuminen ja sähkön hinta. Sähkön pörssihintaan liittyy myös erinäisten johdannaisten hintojen ennustaminen. [7].

Nykyisin eritellään usein vielä omaksi ryhmäkseen erittäin lyhyen aikavälin ennustukset (VSTLF, very short term load forecasting), joissa ennustukset tehdään alle tunnin, jopa muutamien minuuttien mittakaavassa. Näitä tarvitaan lähinnä älykkäissä sähköverkoissa siirron ohjaukseen ja automaattisten laitteiden ohjaukseen. [8]

Lyhyen aikavälin ennustukset ovat tärkeitä ostettaessa sähköä pörssistä. Nord Pool operoi day ahead -markkinoita, joilla sähköä ostetaan kello 13 Suomen aikaa seuraa-

van päivän kullekin tunnille, sekä intraday-markkinoita, jotka täydentävät day ahead -markkinoita niin, että sähköä voi ostaa vielä tuntia ennen toimitusta [9]. Sähkön hinta on sitä korkeampi, mitä myöhemmin sähköä ostaa. Käytännössä kulutusta ja sähkön ostoa tai vastaavasti tuotantoa ja sähkön myyntiä on mahdotonta pitää yhtä suurina, jolloin virheestä on vielä maksettava tasevastaavalle, joka tuottaa tasesähköä. Moderneilla sähkömarkkinoilla on siis ehdottoman tärkeää saada luotettavia ennustuksia, jotta sähköä voidaan ostaa etukäteen tarpeeksi ja mahdollisimman halvalla.

Keskipitkän ja pitkän aikavälin ennustukset ovat hyödyllisiä, jos sähköllä kauppaa tekevä yritys haluaa suunnitella kirjanpitonsa etukäteen pitemmälle ajanvälille kuin mitä tunnittainen kaupankäynti mahdollistaa. Nämä ovat myös tarpeen ostettaessa sähköä etukäteen ja tuotanto- ja siirtokapasiteettia suunniteltaessa ennakoimaan sijoituksen tuottoa. Ennustettavuus on sitä huonompi, mitä kauemmas tulevaisuuteen yritetään ennustaa. [6]

2.3 Aikasarjaennustamisesta

Aikasarjamallit, kuten sähköä ennustavat mallit, voidaan luokitella usealla tavalla. Eräs tapa ryhmitellä menetelmät on erottaa tilastolliset menetelmät sekä koneoppimista käyttävät mallit. Toinen tapa on erottaa lineaariset ja epälineaariset mallit omikse ryhmikseen. Sähkön kulutuksen ennustamiseen käytetään ohjatun oppimisen menetelmiä, sillä todellinen kulutus osataan aina sanoa ja antaa mallille esimerkiksi.

Aikasarjoille on tyypillistä autokorrelaatio, eli näytteet korreloivat aikasarjan edellisten näytteiden kanssa. Tällöin näytteet toki korreloivat myös tulevien näytteiden kanssa, mutta näitä ei voi käyttää apuna ennustuksessa.

2.3.1 Lineaariset mallit

Lineaaristen mallien etuna on usein helposti hahmotettava sisäinen toteutus, joten ihmisen on helppo ymmärtää mallin sisäistä toimintaa. Mallia sanotaan lineaarimalliksi,

jos selitettävä muuttuja voidaan esittää selittävien muuttujien lineaarikombinaationa, eli $x = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_n z_n$, missä x on selitettävä muuttuja ja z_i ovat selittävät muuttujat ja näiden välillä ei ole epälineaarisia riippuvuuksia. Käytännössä tämä tarkoittaa sitä, että selittäviä muuttujia ei saa kertoa keskenään.

Lineaarimallit ovat laskennallisesti yksinkertaisempia kuin epälineaariset mallit. Monille lineaarimalleille virheen neliön minimille on suora kaava, kun taas epälineaarisille malleille (neliö)virheen minimi täytyy etsiä monimutkaisten optimointialgoritmien avulla, mikä on hidasta.

Lineaarinen regressio

Lineaarisessa regressiossa aikasarjan y_t uusin arvo arvioidaan ottamalla lineaarikombinaatio joukosta selittäviä aikasarjoja $x_{i,t}$. ε on ennustamattomissa oleva virhe, jolla on jokin tietty jakauma ja varianssi.

$$y_t = a_0 + a_1 x_{1,t} + a_2 x_{2,t} + \ldots + a_n x_{n,t} + \varepsilon$$

eli matriisimuodossa:

$$y_t = \begin{pmatrix} 1 & x_{1,t} & x_{2,t} & \ldots & x_{n,t} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} + \varepsilon$$

m tällaista yhtälöä voidaan laskea yhtä aikaa kokoamalla muuttujat matriisin riveiksi:

$$\begin{pmatrix} y_{t_1} \\ y_{t_2} \\ \vdots \\ y_{t_m} \end{pmatrix} = \begin{pmatrix} 1 & x_{1,t_1} & x_{2,t_1} & \dots & x_{n,t_1} \\ 1 & x_{1,t_2} & x_{2,t_2} & \dots & x_{n,t_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,t_m} & x_{2,t_m} & \dots & x_{n,t_m} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \quad \text{eli}$$

$$\boldsymbol{y} = X\boldsymbol{a} + \boldsymbol{\varepsilon}, \varepsilon_i \in N(0, \sigma^2)$$

Neliövirheiden summa $\sum_{i=1}^{m} \varepsilon_i^2 = (\boldsymbol{y} - X\boldsymbol{a})^T(\boldsymbol{y} - X\boldsymbol{a})$ saa miniminsä painovektorin a suhteen kun $\boldsymbol{a} = (X^TX)^{-1}X^T\boldsymbol{y}$ [10]. Matriisia $(X^TX)^{-1}X^T$ kutsutaan Mooren-Penrosen käänteismatriisiksi tai näennäiskäänteismatriisiksi.

Autoregressiivinen malli

Autoregressiivinen malli saa nimensä siitä, että mallissa lasketaan regressio, mutta selittävät muuttujat ovat aikasarjan itsensä aikaisempia arvoja.

$$X_t = a_0 + a_1X_{1,t} + a_2X_{2,t} + \ldots + a_nX_{n,t} + \varepsilon$$

Autoregressiivisiä malleja on useita samankaltaisia, jotka yleistävät mallia. Esimerkkejä näistä ovat epästationaariset mallit ($ARIMA$), kausittaiset mallit (SAR, $SARMA$, $SARIMA$), mallit ulkoisilla muuttujilla (ARX, $ARMAX$, $ARIMAX$, $SARIMAX$), ja heteroskedastiset eli muuttuvavarianssiset mallit ($GARCH$), joka on käytössä etenkin finanssialalla.

Esimerkki 2.3.1. Yksinkertainen esimerkki autoregressiivisestä $AR(1)$-mallista on $X_t = 0.6X_{t-1} + Z, \quad Z \in N(0, 1)$

Samankaltainen päivä -malli

Samankaltainen päivä -mallissa historiatiedoista etsitään päivä, jonka kulutukseen vaikuttavat tekijät, kuten esimerkiksi viikonpäiväindeksi, vallitseva säätila ja päivämäärä, ovat

samankaltaiset kuin ennustettavalla päivällä. Ennustettu kulutus on tällöin kyseisen vastaavan päivän kulutus tai useamman vastaavan päivän kulutuksista laskettu yhdistelmä. Tällaiset metodit ovat yksinkertaisia ymmärtää, mutta eivät välttämättä pysty mallintamaan todellisuudessa monimutkaista ongelmaa riittävän tarkasti. [11]

2.3.2 Epälineaariset mallit

Epälineaaristen mallien etuna on hyvä yleistyskyky, itseorganisoituminen, sopeutuva oppimiskyky ja kyky hahmottaa epälineaarisia riippuvuuksia. Monet epälineaariset mallit ovat universaaleja, eli ne voivat oppia approksimoimaan mielivaltaista funktiota.

Epälineaarisilla malleilla on myös haittapuolia lineaarisiin malleihin nähden. Oppiminen on hitaampaa kuin lineaarisilla malleilla, joilla on usein yksinkertainen, suora ratkaisukaava. Epälineaariset mallit sen sijaan vaativat iteratiivisia menetelmiä ja päätyminen optimiin voi olla hidasta. Lisäksi optimointialgoritmi saattaa juuttua paikalliseen ääriarvoon eikä löydä todellista globaalia ääriarvoa. Epälineaarisilla malleilla on haittana myös vaikeasti ymmärrettävä sisäinen toteutus, mikä tekee mallin tulkinnan ihmiselle haastavaksi.

Epälineaaristen mallien oppimiskyky voi olla myös haitaksi, sillä malli saattaa ylioppia, eli oppia prosessin lisäksi havainnoissa olevan kohinan eli todellisuudessa mallinnettamattomissa olevan virheen. Tällöin mallin ennustuskyky kärsii.

Suosittuja epälineaarisia malleja ovat neuroverkot (ANN eli artificial neural network), joita on myös käytetty paljon sähkön kulutuksen ennustamiseen. Nämä pyrkivät simuloimaan aivoissa olevien neuronien toimintaa. Neuroverkot ovat universaalimalleja, jotka kykenevät oppimaan approksimoimaan mielivaltaista funktiota. Neuroverkossa on päällekkäisiä neuronikerroksia, joiden välillä on *synapseja* eli kytköksiä. Syötteet kulkevat yhden tai useamman *neuronin* läpi, joka yhdistää syötteet ja lähettää tuloksen edelleen seuraavan kerroksen neuroneille, kunnes päädytään uloimpaan kerrokseen, joka kokoaa tuloksen. Neuroverkon opettaminen toimii antamalla syötteitä ja muuttamalla neuronien

parametreja suuntaan, joka pienentää virhettä. Virheen suunta lasketaan ensin viimeiselle, tulostuskerrokselle, jotta edelliset kerrokset saavat käytettyä tätä laskiessaan oman virheensä suunnan.

2.4 Kirjallisuuskatsaus

2.4.1 Tutkimus

Ennen vuosituhannen vaihtumista lineaariset mallit olivat laajasti käytössä sähkönkulutuksen ennustuksessa, mutta vuosituhannen vaihteen jälkeen, todennäköisesti kasvaneen laskentakapasiteetin myötä, epälineaariset mallit kuten neuroverkot ovat kasvattaneet suosiotaan.

Erään ensimmäisistä sähkönkulutuksen ennustuksen metodeista esittelee Dryar, joka vuonna 1944 julkaistussa teoksessaan käsittelee työtään Philadelphian alueen sähköyhtiön sähköntuotannon vastuuhenkilönä. Tunnittainen sähkön kulutus, joka tarvittiin järjestelmän ylläpitoon, määriteltiin peruskuormasta kerrottuna painolla, joka saatiin kolmesta säämuuttujasta: lämpötila, tuulen nopeus ja pilvisyys. Sähkön kulutus tällöin koostui pitkälti lämmityksestä, valaistuksesta, radiolähetyksistä sekä sähkökäyttöisistä junista. Peruskuorma oli laskettu käänteisesti samalla kaavalla. Suurin virhe ennusteella oli noin 7% todellisesta kulutuksesta. Tarkemmat virheanalyysit puuttuvat, oletettavasti laskennan työläyden takia, sillä tutkimus on tehty ennen ensimmäisen tietokoneen valmistumista. [12]

Hagan ja Klein käyttivät vuonna 1978 online-ARIMA-mallia Kansasin alueen sähköyhtiön kuluksen ennustamiseen. Ulkoisena tekijänä malli huomioi lämpötilan. Koska lämpötilatiedot kyseisellä alueella olivat saatavilla vain kolmen tunnin välein, malli ennusti myös sähkön kulutuksen kolmen tunnin välein. Mallille opetettiin kuuden viikon kulutustiedot, minkä jälkeen jatkuvasti muuttuvalle mallille annettiin vuoden kulutustiedot. Muuttuvan mallin keskimääräinen prosentuaalinen virhe oli 2.45% ja vertailukoh-

teena olevan muuttumattoman mallin virhe oli 2.72%. Tekijät vetivät johtopäätöksen, että dynaaminen malli pystyy reagoimaan staattista mallia paremmin sähkön kulutuksen kausittaiseen vaihteluun. [13]

Koo et al. (2013) vertailivat kausittaista ARIMA-mallia ja Holt–Winters-mallia. Data ensin normalisoitiin skaalaamalla välille $[0, 1]$ ja luokiteltiin k:n lähimmän naapurin menetelmällä eri tyyppisiin päiviin. Käytetty data oli tunnittainen Etelä-Korean sähköntuotanto vuosilta 2007–2008, ja ennusteet tehtiin päiväksi eteenpäin. Keskimääräinen prosentuaalinen virhe oli ARIMA-mallilla 3.192888% ja Holt–Winters-mallilla 3.645615%. Tekijöidensä mukaan mallit voisivat olla paremmatkin ja hyötyisivät etenkin muista muuttujista, kuten lämpötilasta ja lomapäivätiedoista.

Chen et al. (2010) sovittivat yhteen samankaltainen päivä -mallin aallokeneuroverkkojen kanssa. Ideana yhdistetyssä mallissa *SIWNN* on se, että samankaltaisella päivällä on parempi ennustavuus kuin edellisellä päivällä, ja aallokemuunnoksella voidaan kulutus hajottaa eri taajuuksisiin komponentteihin, jotka yhdistetään neuroverkolla. Mallia testattiin ISO New England -datajoukolla (Uuden Englannin itsenäinen järjestelmäoperaattori) ennustamaan seuraavan päivän kulutus, kun syötteenä mallille annettiin kulutushistoria, lämpötila ja ilmankosteus ja edellisen päivän viimeisen tunnin ennuste. Lämpötila-ilmankosteusindeksi myös prosessoitiin "taittamalla"ottamalla erotus indeksiarvosta, jossa kulutus saavuttaa minimin ja tästä itseisarvon, jolloin kulutuksen ja indeksin vastaavuus, joka tavallisesti on V:n muotoinen, saatiin kutakuinkin lineaariseksi. Malli opetettiin datalla kahden vuoden ajalta ja testattiin seuraavan vuoden tiedoilla. Keskimääräinen prosentuaalinen virhe kuukausittain vaihteli välillä $[1.24\%, 2.22\%]$, keskiarvona 1.635%. Malli pärjäsi paljon paremmin kuin vertailukohteena ollut neuroverkko ilman samankaltainen päivä -mallia. [11]

Ceperic, Ceperic ja Baric (2013) käyttivät tukivektorikonetta (support vector machine) regressoimaan sähkönkulutusta. Mallin syötteinä olivat menneen kulutuksen lisäksi kolme lämpötilamuuttujaa joka tunnilta neljän päivän ajalta (kolme mennyttä ja seuraava

päivä), kosteusindeksi sekä kaksiarvoiset muuttujat kullekin viikonpäivälle, kuukaudelle sekä lomapäiviä varten. Malli loi rinnakkain ennustuksen kullekin seuraavan päivän tunnille. Käytetty data oli ISO New England -datajoukko vuosilta 2003–2006, joista viimeinen vuosi oli testijoukko ja vuodet 2003–2005 mallin validointia ja rakennusta varten. Keskimääräinen prosentuaalinen virhe oli 1.31%, kun käytettiin todellista säätä ja 1.44%, kun käytettiin sääennusteita. Toisella datajoukolla, Pohjois-Amerikan sähköyhtiön datalla vuosilta 1988–1992, keskimääräinen prosentuaalinen virhe oli 1.99% ja tunnin päähän tehtävillä (seuraavan tunnin) ennusteilla 0.72%. Paras piirteenvalinta-algoritmi valitsi 456 muuttujasta 20 parasta kulutusmuuttujaa ja 20 parasta lämpötilamuuttujaa voittaen asiantuntijan manuaalisesti laatiman piirrejoukon jokaisessa datajoukossa. Mallin ennustavuus oli parempi kuin *SIWNN*-mallilla. [14]

Chen et al. esittävät 2018 julkaistussa artikkelissaan syviin jäännösverkkoihin (deep residual network) perustuvan menetelmän, jota he käyttivät luomaan mallin, joka ennustaa seuraavien 24 tunnin sähkönkulutuksen. Neuroverkon syötteenä käytettiin kulutusta samalta viikkotunnilta 1,2,3,4,8,12,16,20 ja 24 viikkoa aiemmin, saman tunnin kulutusta edeltävän seitsemän päivän ajalta, edellisen päivän kaikkien tuntien kulutusta, vallinnutta lämpötilaa kaikilta näiltä tunneilta, ennustettavan tunnin todellista lämpötilaa simuloimassa sääennustetta, sekä muuttujia, jotka kertovat, mikä vuodenaika on ja onko viikonloppu tai lomapäivä. Mallia testattiin tunnetulla kulutusdatalla nimeämättömältä pohjoisamerikkalaiselta sähköyhtiöltä vuosilta 1985–1992, joista malli opetettiin kahden vuoden ajalta ja loput datasta käytettiin testaamiseen. Mallin keskimääräinen prosentuaalinen virhe oli 1.557% ja 1.575%, kun lämpötilatietoihin oli lisätty satunnainen normaalijakautunut virhe yhden fahrenheit-asteen keskihajonnalla. Saadun mallin yleistettävyyttä testattiin myös ISO New England -datajoukolla, ja tekijöidensä mukaan malli on sekä tarkka että helposti yleistettävissä. [15]

Ghareeb ja El Saadany (2013) yrittivät geneettisellä algoritmilla rakentaa matemaattisen lausekkeen, joka kuvaisi sähkönkulutusta. Käytetty data oli 39 viikkoa egyptiläi-

sen sähköoperaattorin alueen sähkönkulutuksesta, joista 38 viikkoa käytettiin opettamaan malli ja viimeistä käytettiin testaamaan malli. Mallille annettiin ulkoisiksi syötteiksi päivittäinen minimi- ja maksimilämpötila. Ennustettavana oli 24 seuraavan tunnin sähkönkulutukset. Tekijät opettivat myös RBF (radial basis function)-neuroverkon samalla datalla. Kymmenestä geneettisellä algoritmilla saadusta mallista huonoimman keskimääräinen prosentuaalinen virhe oli 1.8059%, parhaimman 1.2522% ja keskimäärin tämä oli 1.5716%, kun taas RBF-mallin tulos oli 1.6656%. Malli on siis kilpailukykyinen, mutta altis huonolle tuurille. [16]

Yu et al. (2014) ehdottivat yhteistyökykyisiin oppiviin satunnaisiin partikkeliparviin pohjautuvaa sumean logiikan TSK-mallia (cooperative random learning particle swarm optimization) tunnin päähän tehtäviin ennustuksiin asuin- sekä yliopiston kampusalueen sähkönkulutusdatasta. Mallin keskimääräinen prosentuaalinen virhe oli 3.7892% asuinalueen datalla ja 3.5475% kampusalueen datalla, kun parhaan vertailukohteen (SVR eli tukivektoriregressio) vastaavat luvut olivat 3.8238% ja 3.7239%. Mallin hyvä puoli on, että sumealla logiikalla menetelmästä saadaan selkeät ja ihmisen tulkittavissa olevat jos–sitten–muuten-päättelyketjut. [17]

Sovann, Nallagownden ja Baharudin (2015) sovelsivat aallokemuunnosneuroverkkoa (wavelet neural network) seuraavan tunnin ennustuksiin ISO New England -datajoukolla, ulkoisina muuttujina lämpötila, kastepiste sekä arkipäivätieto. Mallin keskimääräinen prosentuaalinen virhe oli 0.32%, ja lisäksi tekijöiden mukaan mallin virheen keskihajonta oli merkittävästi vertailukohteena olevan neuroverkon virheen keskihajontaa pienempi. [18]

2.4.2 Ohjelmistot

Kuten saattaa olettaa, rahallisesti merkittävään ongelmaan on tarjolla kaupallisia ratkaisuja. Eräitä esimerkkejä nimenomaan sähkönkulutuksen ennustukseen markkinoiduista ohjelmistoista ovat ABB:n ABB Energy Manager, Etapin Load Forecasting Software, GMDH:n Shell for Data Science, SAS:n Energy Forecasting, Enforin Loadfor, Itron sekä

IBM:n alaisen The Weather Companyn WSI Trader.

Yhteistä näille kaikille on se, ettei yksikään yritys kerro, millä metodilla ennusteet luodaan, ellei sitten hyvin yleisellä tasolla, kuten jaolla aikasarja- ja tekoälypohjaisiin menetelmiin, joten vertailu eri ohjelmistojen välillä on hyvin hankalaa ellei mahdotonta. Käytetyt algoritmit saattavat myös vaihtua ilman, että tämä näkyy käyttäjälle mitenkään. [19]–[25]

3 Autoregressiiviset menetelmät sähkönkulutuksen ennustamisessa

3.1 Valintaperusteet autoregressiivisille menetelmille

Autoregressiivisiä malleja on käytetty jo lähes puoli vuosisataa sähkönkulutuksen ennustamiseen ja tekniikka on kypsää. Kuten nimensäkin kertoo, autoregressiivisessa mallissa sarjan aikaisempia arvoja käytetään laskemaan nykyhetki regressiolla. Sähkön kulutukseen tietyllä hetkellä vaikuttaa paljon edellisten tuntien kulutus. Myös ulkoiset tekijät, kuten lämpötila ja vuodenaika voidaan ottaa huomioon autoregressiivisillä malleilla.

3.2 Aikasarjat

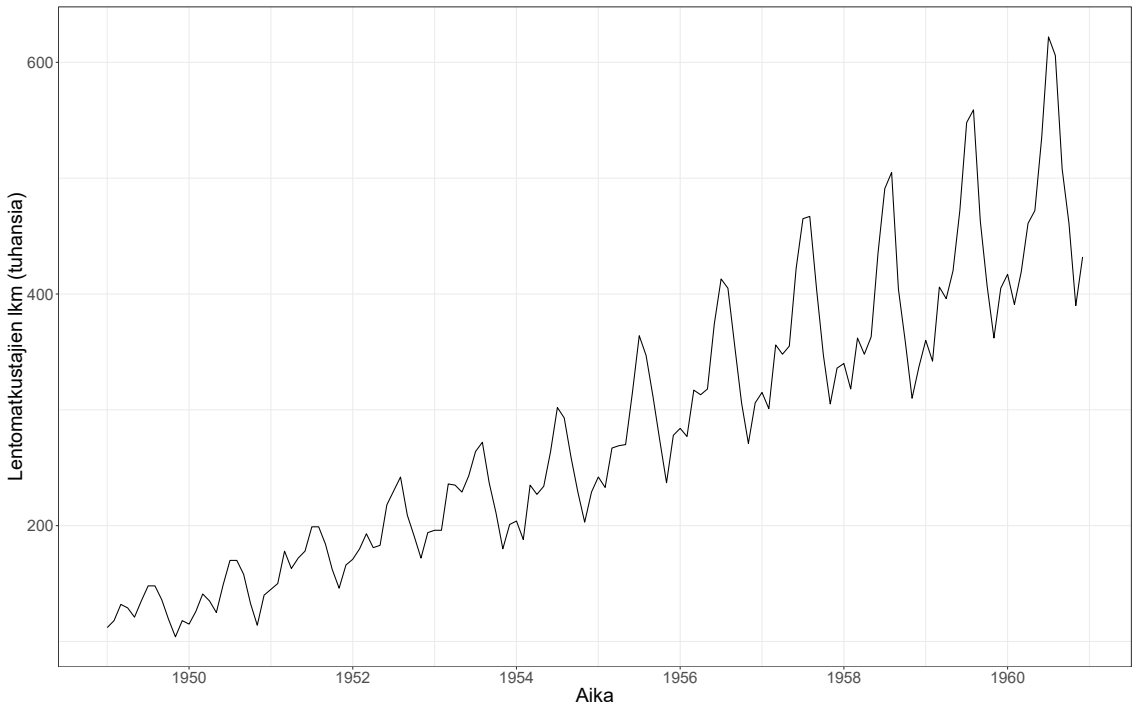
Aikasarjaksi kutsutaan havaintojoukkoa $X_t, t \in T$, jossa jokainen havainto on otetty tietyllä hetkellä t. Joukko T on mahdollisten ajanhetkien joukko. Aikasarjaa kutsutaan diskreetiksi, jos ajanhetket ovat tasaisin väliajoin. Tästä eteenpäin aikasarjasta puhuttaessa tarkoitetaan diskreettiä aikasarjaa. Esimerkkejä aikasarjoista on lukemattomia. Muutamina mainittakoon digitaaliset äänitteet, säätilastot, tehtaan vuotuinen tuotanto ja pörssiosakkeen päivän päätöskurssi. Aikasarjan taustalla on aina jokin prosessi, joka voi olla hyvinkin monimutkainen. Tätä prosessia pyritään mallintamaan ja saatua mallia kutsutaan aikasarjamalliksi. Tämän luvun sisältö pohjautuu teoksiin Timeseries analysis: Forecast and Control (Box, Jenkins & Reinsel, 2008) sekä Time Series Theory and Methods

(Brockwell & Davis, 1991).

Eräs klassinen aikasarjan hajotelma helpommin ymmärrettäviin osiin on additiivinen malli

$$X_t = m_t + s_t + Y_t \tag{3.1}$$

missä M_t on hitaasti muuttuva *trendiksi* kutsuttu aikasarja, S_t on *kausittainen komponentti* – aikasarja jolla on tunnettu jaksollisuus, ja Y_t on satunnainen komponentti eli stationaarinen kohina. Multiplikatiivinen malli on lähes samanlainen, erotuksena vain se, että komponentit kerrotaan keskenään yhteenlaskun sijaan. Myös syklisyys voidaan erottaa omaksi komponentikseen, tässä mallissa syklisyys on yhdistetty trendin kanssa trendi-sykli-komponentiksi. Syklisyyden ja kausittaisuuden ero on se, että kausittaisuuden jaksollisuus on tunnettu ja pysyy samana, kun taas syklisyys on epäsäännöllistä laajaa vaihtelua.



Kuva 3.1: Aikasarja. Kuukausittainen lentomatkustajien lukumäärä

Esimerkki 3.2.1. Kuvassa 3.1 on Box ja Jenkinsin [26] aikasarja kuukausittaisista kansainvälisten lentomatkustajien määristä ajalta 1949 – 1960. Aikasarjassa on havaittavissa epälineaarinen trendi, kausittaisuutta sekä heteroskedastisuutta, eli aikasarjan varianssi ei pysy vakiona.

3.3 Viiveoperaattori

Puhuttaessa autoregressiivisistä malleista on hyödyllistä käyttää viive- ja differointioperaattoria, sillä näin säästytään pitkiltä summalausekkeilta. Viiveoperaattorin B määritelmä on

$$BX_t = X_{t-1}$$

ja tämän avulla voidaan määritellä differointioperaattori ∇

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t$$

Viive- ja differointioperaattoreiden potenssit määritellään seuraavasti:

$$B^n X_t = X_{t-n}$$
$$\nabla^n X_t = \nabla(\nabla^{n-1} X_t) \text{ kun } n \geq 1 \text{ ja } \nabla^0 X_t = X_t$$

3.4 Stationaarinen prosessi

Monet aikasarjamallit ovat stationaarisia eli olettavat mallinnettavan prosessin olevan stationaarinen. Tällöin prosessin täytyy noudattaa tiettyjä rajoituksia. Stationaarisuuden määritelmä on se, että prosessin tilastolliset ominaisuudet kuten keskiarvo ja varianssi eivät muutu ajan myötä.

Usein käytetään toista stationaarisuuden määritelmää, joka on helpompi laskea, ns. heikkoa stationaarisuutta eli kovarianssistationaarisuutta.

Määritelmä 3.4.1. Autokovarianssifunktio on kovarianssin laajennus aikasarjoille. Autokovarianssifunktio aikasarjalle $X_t, t \in \mathbb{Z}$:

$$\gamma_X(r,s) = Cov(X_r, X_s) = E[(X_r - EX_r)(X_s - EX_s)], \quad r,s \in \mathbb{Z}$$

Määritelmä 3.4.2. Sanotaan että aikasarja on *heikosti stationaarinen* jos

$$E|X_t|^2 < \infty$$
$$EX_t = m \quad \forall t \in T$$
$$\gamma_X(r,s) = \gamma_X(r+t, s+t) \quad \forall r,s,t \in \mathbb{Z}$$

Jos aikasarja on heikosti stationaarinen, $\gamma_X(r,s) = \gamma_X(r-s, s-s) = \gamma_X(r-s, 0)$. Tämän takia on hyödyllistä määritellä yhden parametrin autokovarianssifunktio

$$\gamma_X(r) = \gamma_X(r,0) = Cov(X_{r+t}, X_t) \quad \forall r,t \in \mathbb{Z}$$

Lähes vastaavasti määritetään autokorrelaatiofunktio

$$\phi_X(r) = \frac{\gamma_X(r)}{\gamma_X(0)}$$

Osittainen autokorrelaatio on funktio, joka kertoo, kuinka paljon X_t:n ja X_{t+k}:n välillä on autokorrelaatiota, kun $X_{t+1} \ldots X_{t+k-1}$:n vaikutus X_{t+k}:hon on poistettu.

Jos prosessia mallinnetaan stationaarisella mallilla, on huolehdittava siitä, että myös aikasarja saatetaan stationaariseksi. Tähän on monia tapoja, joista pitää tilanteen mukaan osata valita sopivin.

3.5 Trendin ja kausittaisuuden poisto

Jos malli itsessään ei tee aikasarjaa stationaariseksi, on trendi/sykli ja kausittaisuus poistettava esikäsittelyvaiheessa. Pelkän trendin poisto kausittaisuuden puuttuessa voidaan

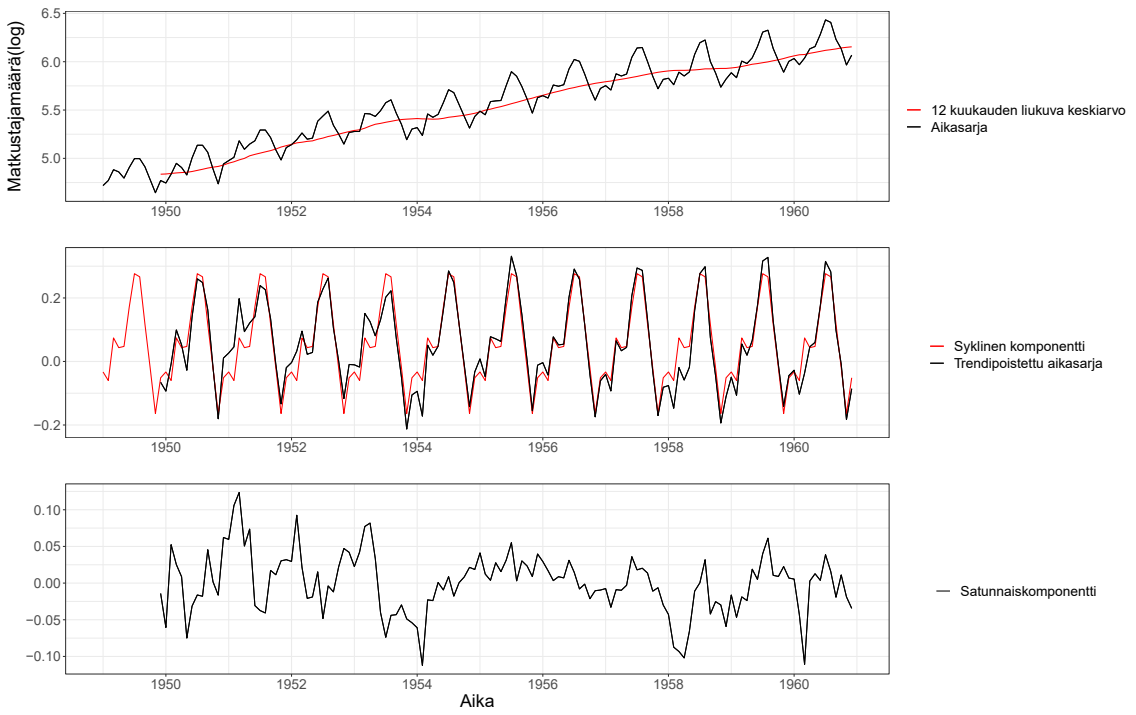
tehdä sovittamalla dataan esim. polynomi pienimmän neliösumman menetelmällä. Toinen vaihtoehto on tasoittaa funktio käyttämällä liukuvaa keskiarvoa tai eksponentiaalista tasoitusta. Liukuvan keskiarvon tapauksessa ei olla kovinkaan kaukana MA- ja $ARMA$-malleista, joista kerrotaan edempänä lisää. Kolmas vaihtoehto trendin poistoon on differoida aikasarja. Jos lineaarinen trendifunktio $X_t = ax + b$ differoidaan, tulos on vakiofunktio $\nabla X_t = a$. Differoimalla aikasarja k kertaa eli operaattorilla ∇^k, saadaan poistettua asteen k polynominen trendi. $ARIMA$-malleissa tämä differointi on osa mallia.

Jos aikasarjassa on trendin lisäksi kausittaisuutta, voidaan kausittainen komponentti estimoida esim. ottamalla keskiarvo kaikista jaksoista, joista data on saatavilla. Liukuvalla keskiarvolla voidaan myös suodattaa pois kausittaisuus ottamalla aikaikkunan pituudeksi kausittaisuuden jakson pituus. Viiveoperaattorilla voidaan myös poistaa kausittaisuus, jos viiveenä käytetään jakson pituutta:

$$\nabla_k = X_t - X_{t-k} = (1 - B^k)X_t$$

Tämä toimii vain, jos jaksollisuus on aikasarjan havaintotaajuuden kokonaislukumoninkerta. Tällöin $\nabla_k X_t = m_t - m_{t-k} + Y_t - Y_{t-k}$ eli $\nabla_k X_t$ saadaan esitettyä kahden uuden aikasarjan, trendin $(m_t - m_{t-k})$ ja kohinan $(Y_t + Y_{t-k})$ summana. Tämä trendi voidaan jälleen poistaa edellä mainittuja keinoja käyttäen.

Esimerkki 3.5.1. Poistetaan trendi ja kausittaisuus kuvan 3.1 lentomatkustaja-aikasarjasta. Ottamalla sarjasta aluksi logaritmi saadaan heteroskedastisuus poistettua. Seuraavaksi eritellään trendi–sykli-komponentti liukuvalla keskiarvolla 12 näytteen ikkunalla. Kausittaisuus on tämän jälkeen arvioitu kuukausittaisista keskiarvoista ja poistettu aikasarjasta, jolloin lopputuloksena on stationaarinen satunnaiskomponentti (Kuva 3.2).



Kuva 3.2: Trendin ja kausittaisuuden poisto esimerkin 3.2.1 aikasarjasta

3.6 AR(p)-prosessi

Autoregressiivinen prosessi eli AR-prosessi on aikasarjaprosessi, jossa aikasarjan seuraava arvo riippuu edellisistä arvoista sekä satunnaisesta tekijästä.

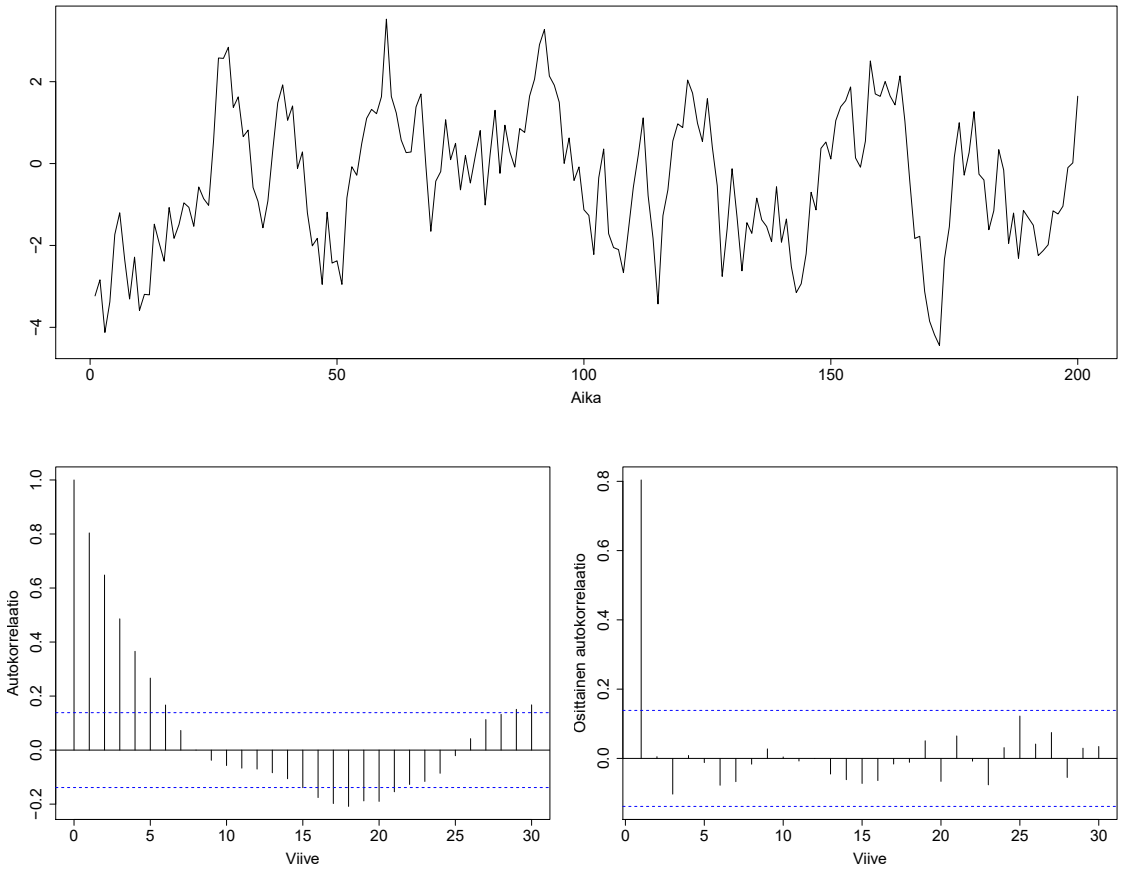
$$X_t = \sum_{i=1}^{n} \phi_i X_{t-1} + Z_t, \quad Z_t \sim N(0, \sigma^2) \tag{3.2}$$

Sama viiveoperaattoria käyttäen:

$$X_t = \sum_{i=1}^{n} \phi_i B^i X_t + Z_t$$

Kun summalauseke siirretään vasemmalle puolelle, saadaan satunnaistekijä Z_t esitettyä B:n polynomina, jota kutsutaan AR-prosessin karakteristiseksi polynomiksi, ja jonka nollakohdat ovat merkityksellisiä prosessin ominaisuuksien kannalta.

$$\phi(B)X_t = Z_t \tag{3.3}$$



Kuva 3.3: AR(1)-prosessi $X_t = 0.8X_{t-1} + Z, Z \in N(0,1)$

Esimerkki 3.6.1. Kuvassa 3.3 on esitetty 200 havaintoa AR(1)-prosessista

$$X_t = 0.8X_{t-1} + Z_t, Z_t \sim N(0,1)$$

sekä

Yhtälöstä 3.2 voidaan X_{t+1} korvata rekursiivisesti $X_{t+2} \dots X_{t+p+1}$:n summalla ja niin edelleen ad infinitum, jolloin X_t saadaan esitettyä aikaisempien virheiden äärettömänä painotettuna summana:

$$X_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i} = \boldsymbol{\psi}(B) Z_t \tag{3.4}$$

AR-prosessi on stationaarinen, jos kertoimet $\psi_i, i \in [1..]$ muodostavat suppenevan sarjan. Tämä ehto toteutuu, jos polynomin $\boldsymbol{\phi}(B)$ (3.3) kaikki (mahdollisesti komplek-

siset) nollakohdat ovat yksikköympyrän ulkopuolella. [1] AR-prosessi voidaan kirjoittaa stationaariseksi vaikkei tämä ehto toteutuisikaan, mutta tällöin joudutaan käyttämään tulevaisuudessa olevia arvoja X_{t+k}. Tällaista prosessia kutsutaan *ei-kausaaliseksi* ja vastaavasti prosessia, joka voidaan kirjoittaa stationaariseksi käyttäen vain menneitä arvoja, *kausaaliseksi*.

Esimerkki 3.6.2. $AR(1)$-prosessi

$$X_t = \phi_1 X_{t-1} + Z_t$$

eli $(1 - \phi_1 B)X_t = Z_t$ on kausaalinen, jos polynomin $(1 - \phi_1 x)$ nollakohta $x = 1/\phi_1$ on kompleksitason yksikköympyrän ulkopuolella eli $|1/\phi_1| > 1$. Tämä ehto on yhtäläinen ehdon $\phi_1 < 1$ kanssa. Kuvassa 3.4 on 20 arvoa aikasarjasta $X_t = 1.25X_{t-1} + Z, Z \in N(0, 1)$. Huomataan, että aikasarja ei ole stationaarinen, sillä ylläoleva stationaarisuusehto ei täyty.

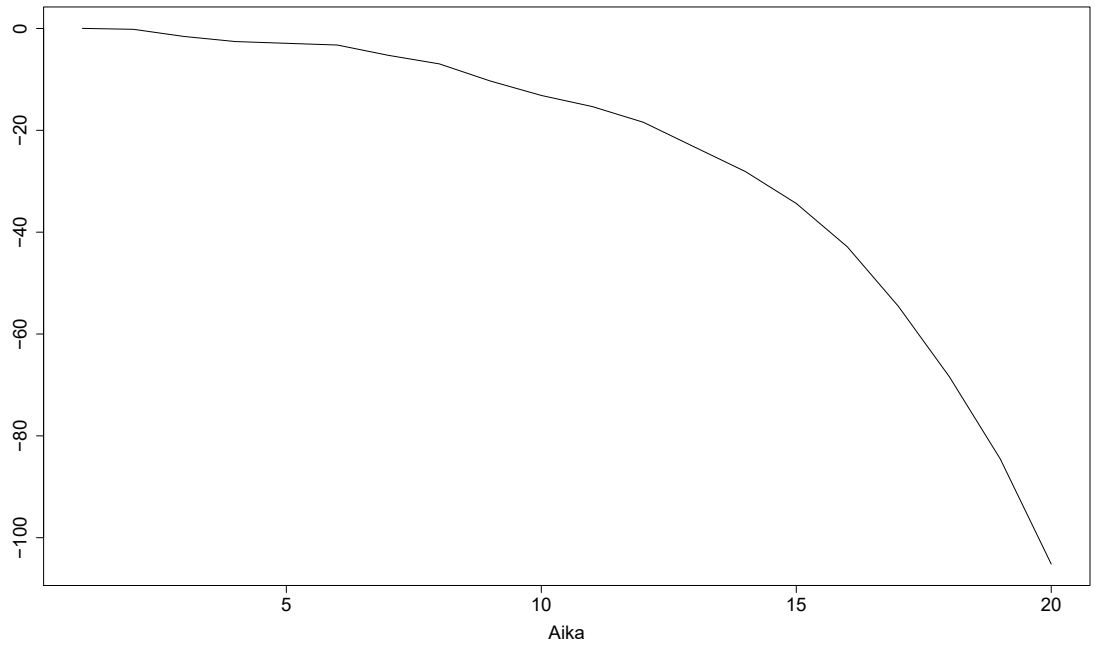
Esimerkki 3.6.3. Tutkitaan $AR(3)$-prosessia

$$\begin{aligned} X_t &= -\frac{1}{12}X_{t-1} - \frac{1}{16}X_{t-2} + \frac{3}{8}X_{t-3} + Z_t \\ X_t &+ \frac{1}{12}X_{t-1} + \frac{1}{16}X_{t-2} - \frac{3}{8}X_{t-3} = Z_t \\ (1 &+ \frac{1}{12}B + \frac{1}{16}B^2 - \frac{3}{8}B^3)X_t = \boldsymbol{\phi}(B)X_t = Z_t \end{aligned}$$

Karakteristisella polynomilla $\boldsymbol{\phi}(x) = 1+\frac{1}{12}x+\frac{1}{16}x^2-\frac{3}{8}x^3 = \frac{1}{48}(2x-3)(9x^2+12x+16)$ on nollakohdat $x = \frac{3}{2}$ ja $x = -\frac{2}{3}(1 \pm i\sqrt{3})$, jotka kaikki ovat yksikköympyrän ulkopuolella (Kuva 3.5). Prosessi on siis kausaalinen.

$AR(p)$-prosessin autokorrelaatiofunktio on eksponentiaalisesti vaimeneva muttei saavuta nollaa, kun taas osittainen autokorrelaatio menee nollaan viiveen p jälkeen. $AR(p)$-prosessi on stationaarinen ja kausaalinen, jos polynomin $\boldsymbol{\phi}(B)$ nollakohdat ovat yksikköympyrän ulkopuolella.

[1] *Algebran peruslauseen* mukaan astetta $k \geq 1$ olevalla polynomilla on kompleksilukujen joukossa k juurta eli nollakohtaa, joista jotkut tosin voivat olla keskenään samoja.



Kuva 3.4: Epästationaarinen AR(1)-prosessi

3.7 MA-malli

Liukuvan keskiarvon malli (MA eli moving average) on AR-mallin kaltainen malli, jossa seuraava arvo lasketaan edellisten kohinatermien Z_t lineaarikombinaationa.

$$X_t = Z_t - \theta_1 Z_{t-1} - \ldots - \theta_q Z_{t-q} \tag{3.5}$$

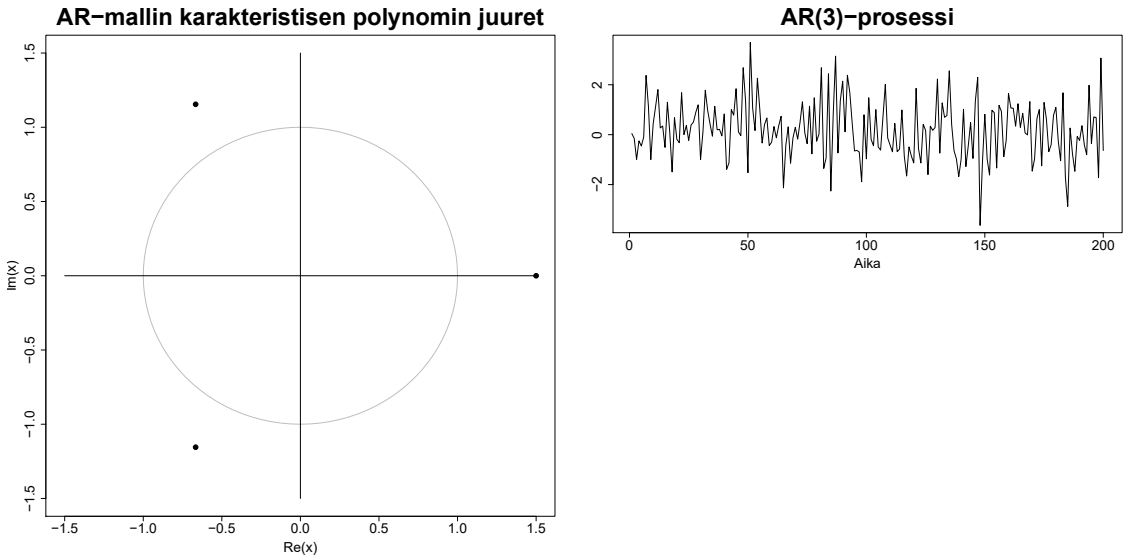
$$= (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) Z_t \tag{3.6}$$

$$= \boldsymbol{\theta}(B) Z_t \tag{3.7}$$

Nimi on siinä mielessä harhaanjohtava, että kertointen ei ole pakko olla yhtä suuria eikä edes positiivisia.

Esimerkki 3.7.1. Kuvassa 3.6 on 200 näytettä $MA(3)$-prosessista

$$X_t = Z_t + \frac{1}{3} Z_{t-1} + \frac{1}{3} Z_{t-2} + \frac{1}{3} Z_{t-3}, Z \in N(0, 1).$$



Kuva 3.5: AR(3)-prosessin $X_t = -\frac{1}{12}X_{t-1} - \frac{1}{16}X_{t-2} + \frac{3}{8}X_{t-3} + Z_t$ karakteristisen polynomin nollakohdat ja kuvaaja 200 näytteestä

Edellä huomasimme (3.4), että $AR(p)$-malli voidaan esittää $MA(\infty)$-mallina. Vastaavasti $MA(q)$-malli voidaan esittää $AR(\infty)$-mallina:

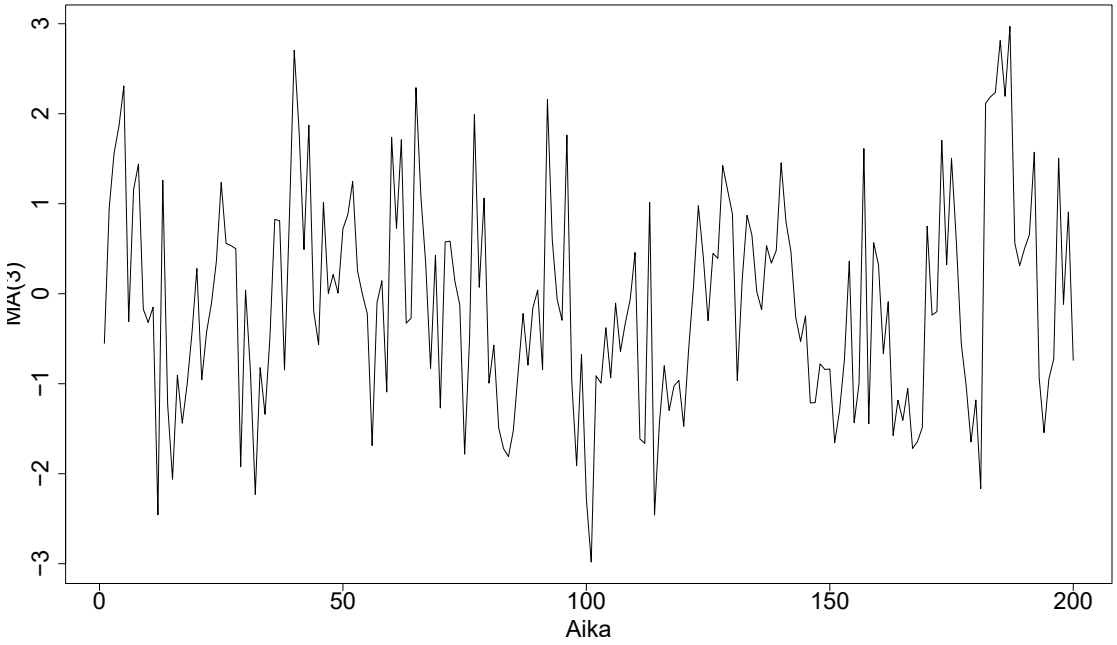
$$
\begin{aligned}
X_t &= Z_t - \theta_1 Z_{t-1} - \ldots - \theta_q Z_{t-q} \\
Z_t &= X_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q} \\
Z_t &= X_t + \theta_1 (X_{t-1} + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}) \\
&\vdots \\
Z_t &= \sum_{i=0}^{\infty} \hat{\theta}_i X_t \\
Z_t &= \boldsymbol{\theta}(B) X_t
\end{aligned}
$$

MA-malli ei ole yksikäsitteinen, sillä esim. mallit

$$X_t = Z_t + 2Z_{t-1}, Z \in N(0,1)$$

$$X_t = Z_t + 0.5Z_{t-1}, Z \in N(0,4)$$

esittävät samaa prosessia. Samaa prosessia esittävistä malleista valitaan sellainen, jonka



Kuva 3.6: MA(3)-prosessi

pystyy esittämään äärettömänä AR-prosessina. Esimerkkinä MA(1)-malli:

$$
\begin{aligned}
X_t &= Z_t + \theta Z_{t-1} \\
Z_t &= -\theta Z_{t-1} + X_t \\
&\vdots \\
Z_t &= \sum_{i=0}^{\infty} (-\theta)^i x_{t-i}
\end{aligned}
$$

joka suppenee jos ja vain jos $|\theta| < 1$. Tällöin sanotaan, että MA-malli on *kääntyvä*. Voidaan osoittaa, että MA-malli on kääntyvä, jos polynomin $\boldsymbol{\theta}$ (3.7) nollakohdat ovat yksikköympyrän ulkopuolella.

Päinvastoin kuin AR-prosessilla, $MA(q)$-prosessin autokorrelaatio menee nollaan viiveen q jälkeen, kun taas osittainen autokorrelaatio on eksponentiaalisesti vaimeneva muttei saavuta nollaa.

3.8 ARMA(p,q)-prosessi

Kuten edellä havaittiin, AR- ja MA-mallit ovat tiivisti yhteydessä toisiinsa. Jos AR-mallin voi esittää äärettömänä MA-mallina ja toisinpäin, $ARMA(p,q)$-malli yhdistää $AR(p)$-mallin ja $MA(q)$-mallin minimoiden yhdistetyn mallin tarvitsemien parametrien lukumäärän. $ARMA$-malli siis huomioi sekä aikasarjan että kohinan edelliset arvot seuraavaa arvoa laskiessa.

Jos aikasarja on $ARMA(p,q)$-prosessi, sarjan seuraava arvo riippuu p:stä edellisestä arvosta sekä q:sta edellisestä kohinan arvosta. Prosessi $X_t, t \in [0, \pm 1, \pm 2 \ldots]$ on $ARMA(p,q)$-prosessi jos X_t on stationaarinen ja jos kaikille t pätee

$$X_t - \phi_1 X_{t-1} - \ldots - \phi p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_q Z_{t-q}, Z \in N(0, \sigma^2)$$

eli

$$\boldsymbol{\phi}(B) X_t = \boldsymbol{\theta}(B) Z_t \tag{3.8}$$

ARMA-malli on kausaalinen, jos polynomin $\boldsymbol{\phi}$ nollakohdat ovat yksikköympyrän ulkopuolella ja kääntyvä, jos polynomin $\boldsymbol{\theta}$ nollakohdat ovat yksikköympyrän ulkopuolella. Näillä kahdella polynomilla ei myöskään tule olla samoja (tai liian lähekkäisiä) nollakohtia, vaan näin käydessä nämä tulee jakaa pois.

Esimerkki 3.8.1. Otetaan esimerkiksi ARMA-prosessi

$$X_t = \frac{7}{6} X_{t-1} - \frac{1}{3} X_{t-2} + Z_t - \frac{22}{15} Z_{t-1} + \frac{8}{15} Z_{t-2}$$

Polynomimuodossa prosessi voidaan esittää muodossa

$$(1 - \frac{7}{6} B + \frac{1}{3} B^2) X_t = (1 - \frac{22}{15} B + \frac{8}{15} B^2) Z_t$$

Autoregressiivisen osan polynomilla $\phi(x) = 1 - 7/6x + 1/3x^2$ on nollakohdat kohdissa $x = 2$ ja $x = 3/2$, ja liukuvan keskiarvon polynomilla $\theta(x) = 1 - 22/15x + 8/15x^2$

on nollakohdat kohdissa $x = 5/4$ ja $x = 3/2$. Koska polynomeilla on yhteinen nollakohta $x = 3/2$, tämä tulee jakaa pois, jolloin prosessi todellisuudessa onkin ARMA(1,1)-prosessi

$$(1 - \frac{1}{2}L)X_t = (1 - \frac{4}{5}L)Z_t \Leftrightarrow \tag{3.9}$$

$$X_t - 0.5X_{t-1} = Z_t - 0.8Z_{t-1} \Leftrightarrow \tag{3.10}$$

$$X_t = 0.5X_{t-1} + Z_t - 0.8Z_{t-1} \tag{3.11}$$

Prosessi on kausaalinen, sillä autoregressiivisen polynomin jäljelle jäänyt nollakohta $x = 2$ on yksikköympyrän ulkopuolella ($|2| > 1$), ja kääntyvä, sillä MA-polynomin jäljelle jäänyt nollakohta $x = 5/4$ on myös yksikköympyrän ulkopuolella ($|5/4| > 1$).

3.9 ARIMA(p,d,q)-prosessi

ARMA-malli voidaan laajentaa käsittelemään tietynlaisia epästationaarisiä malleja differoimalla sopivasti, jolloin saadaan ARIMA-malli (autoregressive integrated moving average).

Jos d on ei-negatiivinen kokonaisluku, X_t on ARIMA(p,d,q)-prosessi, jos d kertaa differoitu sarja $(1 - B)^d X_t$ on kausaalinen ARMA(p,q)-prosessi. [27] Malli voidaan tällöin esittää muodossa

$$\boldsymbol{\phi}^*(B) = \phi(B)X_t = \boldsymbol{\theta}(B)(1 - B)^d Z_t \tag{3.12}$$

eli polynomilla $\phi^*(B)$ on d-kertainen nollakohta ykkösessä. Prosessi on stationaarinen jos ja vain jos $d = 0$. [27]

Erityisesti aikasarjaan voi lisätä mielivaltaisen polynomisen trendin astetta $(d - 1)$, joten ARIMA-malli sopii aikasarjoihin, joissa on polynominen trendi.

3.10 SARIMA

$SARIMA(p, d, q) \times (P, D, Q)_s$-malliksi kutsutaan mallia

$$\boldsymbol{\phi}(B)\boldsymbol{\Phi}(B^s)(1-B)^d(1-B^s)^D X_t = \boldsymbol{\theta}(B)\boldsymbol{\Theta}(B^s)Z_t, Z_t \in WN(0,\sigma^2) \tag{3.13}$$

missä ϕ on AR-polynomi astetta p, Φ on kausittainen AR-polynomi astetta P, θ on MA-polynomi astetta q, Θ on kausittainen MA-polynomi astetta Q, s on kausittaisen vaihtelun syklin pituus ja d ja D ovat tavallisen ja kausittaisen differoinnin kertaluvut, $d, D \geq 1$.

3.11 Ulkoisten tekijöiden vaikutus aikasarjaan

Edellä läpikäydyt autoregressiomenetelmät käyttävät vain aikasarjan ja mahdollisesti kohinan edellisiä arvoja, mutta usein on hyödyllistä käyttää muita aikasarjoja, kuten esimerkiksi lämpötilaa, parantamaan mallin ennustuskykyä. Tällöin malliin täytyy lisätä ulkoisia tekijöitä.

Autoregressiiviseen malliin voidaan lisätä ulkoisia tekijöitä. Autoregressiivinen liukuvan keskiarvon malli ulkoisilla tekijöillä eli ARMAX-malli voidaan kirjoittaa muodossa

$$\begin{aligned} X_t &= \sum_{i=1}^{r} \beta_i Y_{i,t} + \boldsymbol{\phi}(B)X_t + \boldsymbol{\theta}(B)Z_t \quad \text{eli} \\ \boldsymbol{\phi}(B)X_t &= \sum_{i=1}^{r} \beta_i Y_{i,t} + \boldsymbol{\theta}(B)Z_t \\ X_t &= \frac{\sum_{i=1}^{r} \beta_i Y_{i,t}}{\boldsymbol{\phi}(B)} + \frac{\boldsymbol{\theta}(B)}{\boldsymbol{\phi}(B)} Z_t \end{aligned}$$

3.12 Regressio ARMA-virheillä

ARMAX-mallissa polynomin $\boldsymbol{\phi}$ kertoimet sekoittuvat ulkoisten tekijöiden kertoimiin, mistä syystä monet ohjelmistot, kuten R/forecast ja Python/StatsModels, käyttävät ulkoisten tekijöiden läsnäollessa sen sijaan vaihtoehtoista mallia, regressiota ARMA-virheillä:

$$X_t = \sum_{i=1}^{r} \beta_i + \frac{\boldsymbol{\theta}}{\boldsymbol{\phi}} Z_t$$

Taulukko 3.1: AR-, MA- ja ARMA-mallin autokorrelaatiofunktioiden käyttäytyminen

Malli	Autokorrelaatio	Osittainen autokorrelaatio
AR(p)	Eksponentiaalisesti vaimeneva	Menee nollaan viiven p jälkeen
MA(q)	Menee nollaan viiveen q jälkeen	Eksponentiaalisesti vaimeneva
ARMA(p,q)	Eksponentiaalisesti vaimeneva viiveen $(q-p)$ jälkeen	Eksponentiaalisesti vaimeneva viiveen $(p-q)$ jälkeen

Tässä mallissa ulkoisten tekijöiden vaikutus on helpompi analysoida, sillä näiden vaikutus lopputulokseen tulee suoraan regressiosta ilman, että viivetermien monimutkaistaisivat laskuja.

3.13 Mallin tunnistaminen

Jotta mallin parametrit voidaan estimoida, täytyy tietää mitä mallia käytetään. Mallin ollessa $ARIMA(p, d, q)$ tämä tarkoittaa hyperparametrien p, d, q valintaa. Jos aikasarja ei ole stationaarinen, se tulee ensin tehdä stationaariseksi kasvattamalla differoinnin astetta d. Tämän jälkeen parametrit p ja q voidaan arvioida autokorrelaatiosta ja osittaisesta autokorrelaatiosta. Apuna mallin tunnistamisessa voi käyttää taulukkoa 3.1, jossa kerrataan miten eri $ARMA$-prosessien autokorrelaatiot käyttäytyvät. Lisäksi jos aikasarjassa on kausittaisuutta, tämä tulee huomioida mallin valinnassa.

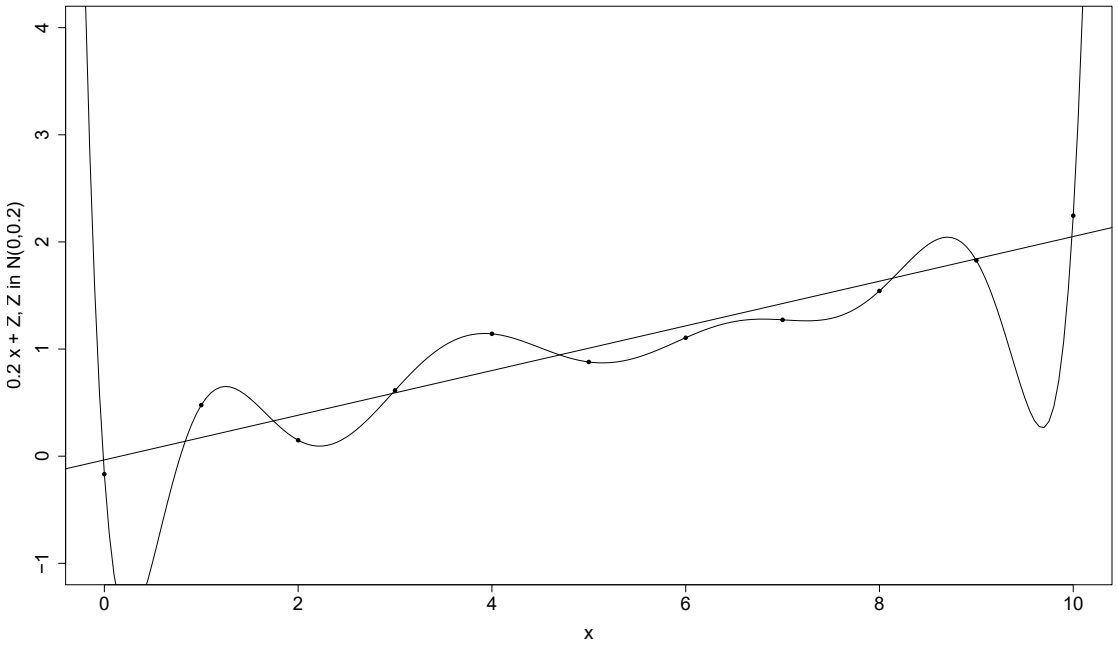
3.14 Mallin kompleksisuuden valinta

Jos käytettävä malli on liian monimutkainen, malli voi oppia testidatassa olevan todellisuudessa mallinnettamattoman kohinan. Tällöin puhutaan ylisovittumisesta. Ylisovittunut malli oppii testijoukon hyvin tai jopa täydellisesti, mutta epäonnistuu, kun yritetään ennustaa testijoukon ulkopuolisia arvoja, joita malli ei ole nähnyt. Klassisena esimerkkinä ylisovittumisesta voidaan ottaa polynomin sovittaminen. n pistettä määrittää yksiselittei-

sesti astetta $n-1$ olevan polynomin, joka kulkee näiden pisteiden kautta. Tällöin malli oppii pisteet täydellisesti, mutta polynomi mitä todennäköisimmin ei vastaa varsinaista prosessia, joka pisteet on tuottanut (Esimerkki 3.14.1).

Occamin partaveitsi -periaatteen mukaan kahdesta yhtä hyvästä mallista tulee valita yksinkertaisempi. Todellisuudessa asia ei ole näin yksinkertainen, vaan valittu malli on aina kompromissi selityskyvyn ja ennustuskyvyn välillä.

Esimerkki 3.14.1. 11:een pisteeseen yhtälöstä $y = 0.2x + Z, \quad Z \in N(0, 0.2)$ voidaan sovittaa täydellisesti asteen 10 polynomi, eli virhe näiden pisteiden joukossa on 0, mutta malli ei etenkään opetusjoukon pisteiden ulkopuolella onnistu kuvaamaan todellista prosessia, jossa y:n odotusarvo on $E(y) = 0.2 * x$. (Kuva 3.7).



Kuva 3.7: Astetta 1 ja 10 olevat polynomit sovitettuna yhtälön $y = 0.2 * x + Z, \quad Z \in N(0, 0.2)$ 11 pisteeseen.

3.14.1 Hyvyyskriteerit

Ylisovittumisen ehkäisemiseksi on mallin kompleksisuuden valitsemiseksi kehitetty eri kriteereitä, jotka rankaisevat liian monimutkaisesta mallista. Eräitä näistä ovat Akaiken informaatiokriteeri (Akaike's Information Criterion) ja tämän biaskorjattu versio AICC; bayesilainen informaatiokriteeri, BIC ja Akaiken lopullinen ennustusvirhe FPE (final prediction error).

$$AICC =$$

$$AICC = ln\hat{\sigma}_k^2 + \frac{n+k}{n-k-2}$$

missä $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$. RSS_k on residuaalien neliösumma k:n parametrin mallissa. [28]

$$BIC = kln(n) - 2ln(L)$$

3.14.2 Ristiinvalidointi

Ristiinvalidoinnissa data jaetaan yhtä suuriin osiin, joista vuorotellen yksi jätetään testijoukoksi, eikä tätä käytetä mallin opettamisessa vaan testaamisessa. Paras malli on se, joka onnistuu ennustamaan testijoukon parhaiten. Etuna hyvyyskriteereihin on se, että ennustuskyky saadaan samoissa yksiköissä ja samoilla metodeilla kuin itse varsinaiset ennusteet.

Stone on todistanut, että AIC on ekvivalentti yksi pois -ristiinvalidoinnin kanssa, jossa malli siis muodostetaan yhtä monta kertaa kuin datassa on arvoja, jättäen aina yhden arvon testiryhmäksi, jolle mallia testataan. [29]

Ristiinvalidointia voidaan mallin kompleksisuuden valinnan lisäksi käyttää mallin hyvyyden testaamiseksi. Ristiinvalidoinnilla saadaan selville, kuinka hyvin malli toimisi uudelle, oppimisessa käyttämättömälle datalle.

3.15 Mallin estimointimenetelmät

Estimaattori $\hat{\theta}(X)$ on kaava, jolla voidaan estimoida jokin suure θ havaintojoukon X perusteella.

Estimaattoria sanotaan harhattomaksi, jos estimaattien keskiarvo vastaa todellista estimoitavan suureen arvoa eli $E(\hat{\theta}(X)) - \theta = 0$. Harhattomien estimaattoreiden joukossa on yksi estimaattori, jolla on pienin varianssi. Tätä kutsutaan minimivarianssiestimaattoriksi. Estimaattori on tarkentuva, jos kasvattamalla tarpeeksi otoksen kokoa estimaatin virhe saadaan mielivaltaisen pieneksi.

3.15.1 Suurimman uskottavuuden estimaatti

Oletetaan, että otoksen $\boldsymbol{x}$ jakauman tiheysfunktio on $f(\boldsymbol{x}|\xi)$, eli tämä on ehdollinen todennäköisyys sille, että $\boldsymbol{x}$ tapahtuu, olettaen että tuntemattomilla parametreillä ξ on tietty arvo. Ennen otoksen näkemistä f antaa kaikille mahdollisille otoksille x jonkin todennäköisyyden tietyillä parametreilla ξ. Kun otos x on nähty, voidaan kysyä, mitkä parametrit ξ olisivat saattaneet olla. Vastinfunktio f:lle, joka vastaa tähän kysymykseen, on *uskottavuusfunktio* $\mathcal{L}(\xi|\boldsymbol{x}) = f(\boldsymbol{x}|\xi)$, jossa $\boldsymbol{x}$ on kiinnitetty ja ξ on muuttuja.

Funktio ja sen logaritmi saavat pienimmän arvonsa samassa kohdassa, joten uskottavuusfunktion sijaan voidaan käyttää uskottavuusfunktion logaritmia $\ell(\xi|\boldsymbol{x}) = ln(\mathcal{L}(\xi|\boldsymbol{x}))$. Parametrien ξ estimaattia, joka maksimoi uskottavuusfunktion (tai sen logaritmin), kutsutaan suurimman uskottavuuden estimaatiksi.

Kun oletetaan, että näytteet x_i ovat riippumattomia, saadaan uskottavuusfunktio ja sen logaritmi muotoon

$$\begin{aligned}\mathcal{L}(\xi|\boldsymbol{x}) &= f(\boldsymbol{x}|\xi) = f(x_1, x_2, \ldots, x_n|\xi) \\ &= f(x_1|\xi) \times f(x_2|\xi) \times \ldots \times f(x_n)|\xi) = \prod_{i=1}^{n} f(x_i|\xi) \\ \ell(\xi|\boldsymbol{x}) &= ln\mathcal{L}(\xi|\boldsymbol{x}) = \sum_{i=1}^{n} lnf(x_i|\xi)\end{aligned}$$

Olettaen, että virheet ovat normaalijakautuneet, saadaan tarpeeksi suurilla otoksilla $AR(p)$-mallin log-uskottavuusfunktioksi likimain

$$\ell(\boldsymbol{\phi}, \sigma^2|\boldsymbol{x}) \simeq -\frac{n}{2}ln(\sigma^2) - \frac{S(\boldsymbol{\phi})}{2\sigma^2}eli$$

missä $S(\boldsymbol{\phi}) = (\boldsymbol{y} - X\boldsymbol{\phi})^T(\boldsymbol{y} - X\boldsymbol{\phi})$, residuaalien neliösumma mallin parametrien ollessa $\boldsymbol{\phi}$. Kun virheet ovat normaalijakautuneet, pienimmän neliösumman estimaatti on suurimman uskottavuuden estimaatti. [26]

$ARMA$-, $ARIMA$- ja $ARIMAX$-mallien parametreja ei pääsääntöisesti saa laskettua pienimmän neliösumman menetelmällä, vaan estimaatit saadaan käyttämällä epälineaarisia optimointialgoritmeja, esim. etsimällä mallin (log-)uskottavuusfunktion globaali maksimi [26].

3.16 Residuaalianalyysi

Kun malli on valittu ja mallin parametrit on estimoitu, tulee tarkastaa mallin sopivuus. Mallin jäännösvirheen eli residuaalin, joka on todellisen virheen estimaatti, tulee käyttäytyä kuten mallin perusteella sopisi olettaa todellisen virheen käyttäytyvän. Usein virheiden oletetaan olevan riippumattomia ja identtisesti jakautuneita. Virheiden jakaumastakin täytyy yleensä tehdä oletuksia, kuten esimerkiksi että virheet ovat normaalisti jakautuneet. Pelkästään kuvaajankin perusteella voidaan havaita poikkeavuuksia.

Mallin ollessa $X_t = f(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) + e_t$ ja löydettyä estimaatit $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2$ parametreille $\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2$, residuaali $\hat{e}_t$ on

$$\hat{e}_t = X_t - f(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$$

Residuaalin autokorrelaatioita voidaan verrata valkoisen kohinan autokorrelaatioon. Residuaaleissa ei tule olla autokorrelaatioita, sillä se olisi vastoin riippumattomuuden oletusta. Noin 95% valkoisen kohinan autokorrelaatioista on välillä $\pm 1.96/\sqrt{n}$ missä n on näytteen koko [27]. Useissa tilastollisissa ohjelmistoissa nämä rajat piirtyvät automaattisesti autokorrelaation kuvaajaan.

3.16.1 Ljung–Box-testi

Ljung–Box-testillä voidaan tarkastaa yksittäisten autokorrelaatioiden sijaan K:ta ensimmäistä autokorrelaatiota yhtä aikaa, jotta voidaan varmistua, että nämä yhdessä käyttäytyvät kuin kohinan autokorrelaatiot. Testin nollahypoteesinä on, että residuaalit ovat valkoista kohinaa. Koska kohinakin voi sattumalta näyttää siltä, että siinä on autokorrelaatioita, yksittäiset autokorrelaatiot voivat olla korkeita. Testi ei voi todistaa, että residuaalit todella olisivat kohinaa, mutta testillä havaitaan jos residuaalit ovat riippuvaisia toisistaan eli autokorreloivat.

Ljung–Box-testin testattava suure on

$$Q = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r_k(\hat{a}) \tag{3.14}$$

missä $r_k(\hat{a})$ on näytteen autokorrelaatio viiveellä k ja K on testattavien viiveiden lukumäärä. $ARMA(p,q)$-mallille Q noudattaa likimain χ^2-jakaumaa $K-p-q$ vapausasteella. [26]

3.17 Virhemitat

Kun mallin sopivuudesta on varmistuttu, voidaan yrittää saada selville, kuinka hyvä malli todellisuudessa on, eli kuinka suuria mallin antamien ennustusten virheet ovat.

Muutamia yksinkertaisimpia käytettyjä mittoja ovat

- Keskimääräinen absoluuttinen prosentuaalinen virhe
  $MAPE = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i-\hat{y}_i}{y_i}| * 100\%$
- Keskimääräinen absoluuttinen virhe $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Keskineliövirhe $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Keskineliövirheen neliöjuuri $RMSE = \sqrt{MSE}$,

missä y_i on todellinen arvo ja $\hat{y}_i$ on ennuste y:lle.

Kaikissa virhemitoissa on hyvät ja huonot puolensa. Keskimääräinen prosentuaalinen virhe on yleisesti käytetty mitta ennustettaessa luonnollisia suureita, mutta tämä ei ole edes määritelty, jos nolla esiintyy havaintoaineistossa. Lisäksi, jos havainnot ja ennustukset ovat luonnostaan positiivisia lukuja kuten esim. kappalemääriä, prosentuaalinen virhe voi olla alapäässä suurimmillaan -100%, mutta ylärajaa ei ole. Keskimääräinen absoluuttinen virhe ja keskineliövirhe (ja sen neliöjuuri) riippuvat mitattavan aineiston skaalasta, joten eri aineistojen väliset mittatulokset eivät ole vertailukelpoisia. Keskineliövirheellä on eri mittayksikkö kuin varsinaisella aineistolla, joten neliöjuuren ottaminen on suositeltavaa. Tämä usein teoreettisesti tärkeä virhemitta on kuitenkin neliöinnin takia altis huomattavasti poikkeaville havainnoille.

Muista vaihtoehtoista esimerkkeinä ovat absoluuttisen prosentuaalisen virheen mediaani, symmetrinen keskimääräinen prosentuaalinen virhe, symmetrinen prosentuaalisen virheen mediaani ja suhteellisen absoluuttisen virheen geometrinen keskiarvo. Erityisesti aikasarjoille kehitetty virhemitta keskimääräinen absoluuttinen skaalattu virhe (MASE) on

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|X_t - \hat{X}_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|X_i - X_{i-1}|}\right),$$

missä X_t on aikasarjan arvo ajankohtana t ja $\hat{X}_t$ on ennuste kyseiselle arvolle.

Absoluuttinen skaalattu virhe suhteuttaa virheet naiiviin ennustukseen, jonka mukaan aikasarja noudattaisi satunnaiskulkua $X_t = X_{t-1} + Z$, missä virheen Z:n odotusarvo on 0 ja siis X_t:n odotusarvo on X_{t-1}. Virhemitta saa arvon 1, jos ennustus on yhtä hyvä kuin naiivilla ennusteella, ja pienemmän kuin 1 jos ennustus on parempi. Tämä virhemitta ei kuitenkaan ole määritelty yksittäiselle näytteelle. [30]

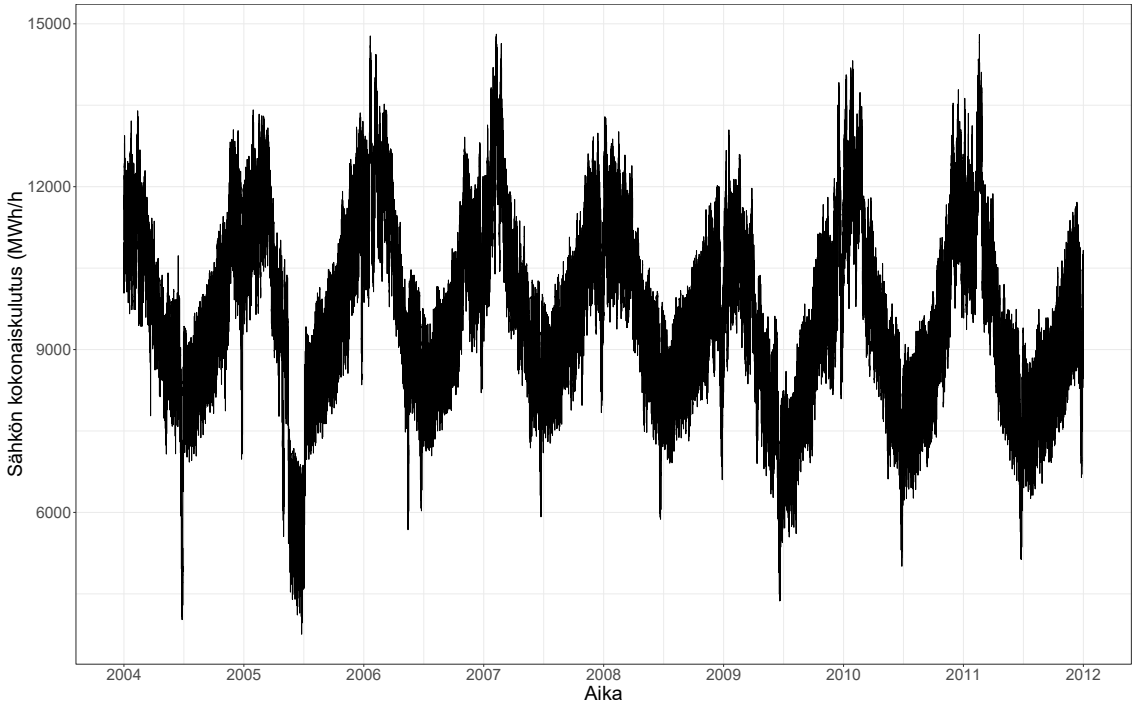
4 Suomen sähkönkulutuksen ennustaminen

4.1 Fingrid

Fingrid on Suomen kantaverkkoyhtiö, joka toimii Suomen tasevastaavana eli viime kädessä huolehtii siitä, että sähköä tuotetaan yhtä paljon kuin kulutetaan. Fingrid perustettiin vuonna 1996 nimellä Suomen Kantaverkko Oyj ja toimintansa se aloitti vuonna 1997. Fingrid omistaa Suomen kantaverkon ja 18.8% yhteispohjoismaisesta sähköpörssistä Nord Poolista. Fingrid ylläpitää käyttövarmuutta omilla varavoimalaitoksillaan sekä yksityisillä varavoimalaitoksilla, joihin Fingridillä on käyttöoikeussopimus, ja jotka eivät saa muuten osallistua sähkömarkkinoille. Järjestelmä kestää minkä tahansa yksittäisen komponentin vikaantumisen. Erityisesti tämä tarkoittaa sitä, että käyttövarmuusreservejä tulee olla yhtä paljon kuin mitä suurin tuotantolaitos tuottaa.

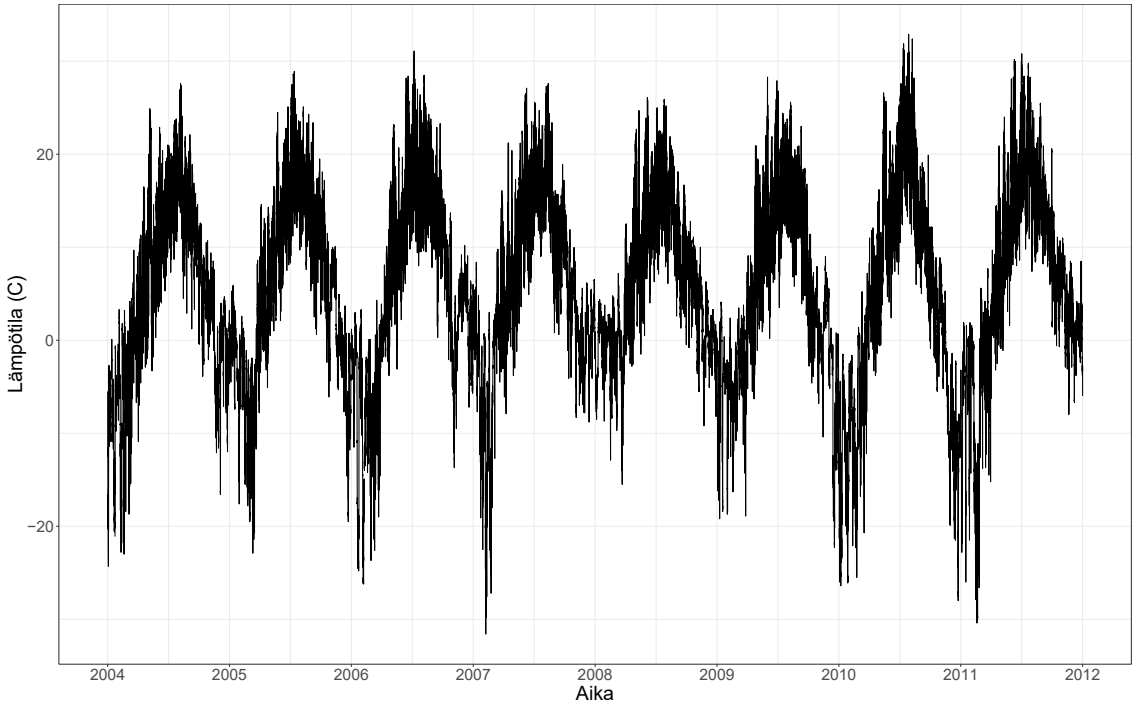
4.2 Datan kuvaus

Mallinnettava aikasarja on Fingridin ilmoittama tunnittainen sähkön kokonaiskulutus vuosilta 2004–2011 (Kuva 4.1). Lisäksi ulkoisena tekijänä on Ilmatieteen laitoksen säähavainnot Hämeenlinnasta samalta aikaväliltä (Kuva 4.2), joita käytetään simuloimaan sääennustuksia ennustettavalle tunnille. Hämeenlinna valikoitui paikaksi aineiston hyvyyden ja Suomen asukaskeskipistettä lähellä olevan sijaintinsa vuoksi.



Kuva 4.1: Fingridin ilmoittama sähkön kokonaiskulutus Suomessa vuosilta 2004-2011

Sähkönkulutusdatan periodogrammista (Kuva 4.3) huomataan kolme selvempää piikkiä spektritiheydessä. Ensimmäinen on taajuudella $1/9000h$, joka vastannee vuosittaista sykliä, jonka todellinen kesto on $8766h$. Toinen piikki on taajuudella $1/24h$, joka vastaa päivittäistä sykliä ja kolmas on taajuudella $1/12h$, joka on päivittäisen syklin harmoninen taajuus, mikä johtuu siitä, että päivittäinen sykli ei ole sinimuotoinen. Taulukosta 4.1 havaitaan, että viikoittainen 168 tunnin sykli ei erotu periodogrammista juurikaan ja on vasta kahdeksanneksi suurin tiheyspiikki, edellään luultavasti virheelliset kohinasta tai näytteenotosta johtuvat 10286, 8000, 7200 ja 72000 tunnin spektrihuiput. Ensimmäinen muuttuja on spektrihuipun sijaintikohdan taajuuden käänteisluku eli syklin kestoaika, ja toinen luku on spektrin tiheys eli miten voimakas signaali kyseisellä taajuudella on.



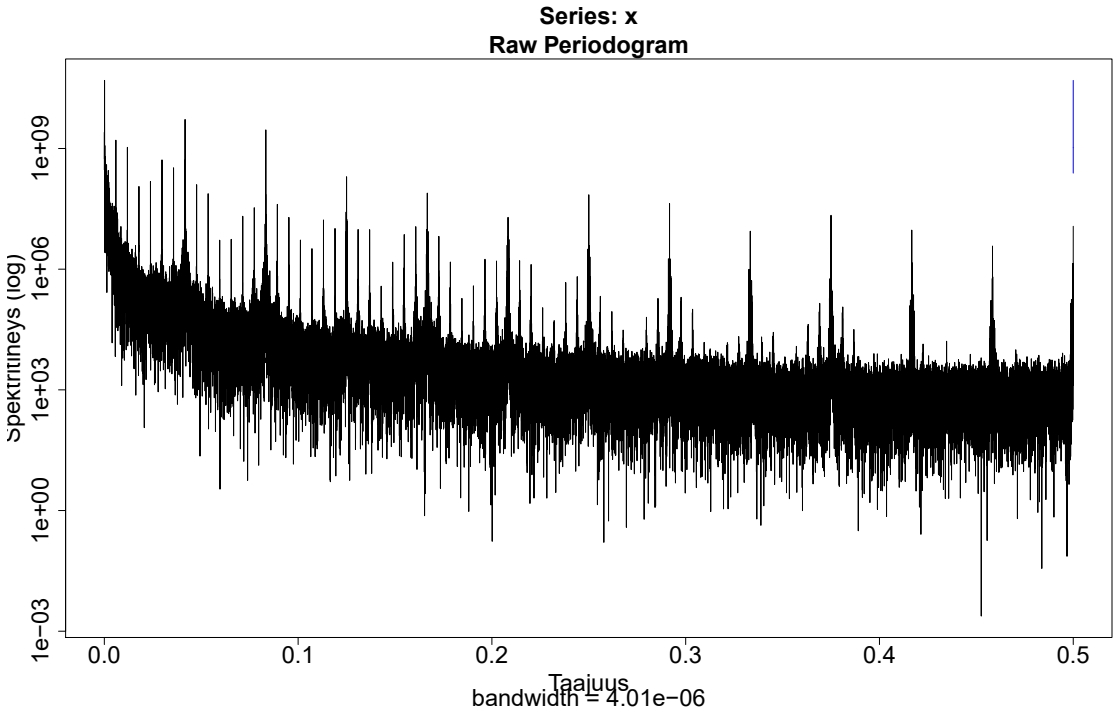
Kuva 4.2: Ilmatieteen laitoksen lämpötilahavainto Hämeenlinnasta

4.3 Datan hyvyys

Sähkönkulutusaineistossa on vain 8 ja lämpötila-aineistossa on 195 puuttuvaa arvoa, jotka esiintyvät enimmäkseen maksimissaan kolmen puuttuvan arvon ryhmissä (Kuva 4.4). Nämä puuttuvat arvot imputoitiin lineaarisella interpolaatiolla.

4.4 Mallinnustehtävän tavoite

Mallinnustehtävänä on ennustaa sähkön kokonaiskulutus seuraavalle tunnille. Mallille annetaan edellisten tuntien kulutuksen lisäksi ulkoisina syötteinä tieto ilman lämpötilasta käyttäen lämpötilahavaintoa ennusteena, eli käyttäen lämpötilan tulevaa arvoa, sekä sijainti päivä-, viikko-, ja vuosisyklissä. Kutakin sykliä kohti luotiin kaksi muuttujaa ottamalla sini ja kosini syklin vaiheesta. Näin saadaan muutettua epäjatkuva laskurimuuttuja yksikäsitteiseksi pariksi jatkuvia muuttujia. Sähkönkulutuksen ja ilman lämpötilan välinen Pearsonin korrelaatiokerroin on -0.7070186 ja sovitetun lineaarimallin R^2-luku on



Kuva 4.3: Sähkönkulutusdatan periodogrammi

0.499999 (Kuva 2.1), mistä voidaan päätellä, että ilman viileneminen nostaa sähkön kulutusta, ja tieto ilman lämpötilasta saattaa parantaa sähkönkulutuksen ennustettavuutta (noin puolet sähkönkulutuksen varianssista voidaan selittää lämpötilalla).

4.5 Mallin rakentaminen

Mallien virhemittoina käytettiin sekä absoluuttista prosentuaalista virhettä yksittäisille tunneille että MASE (keskimääräinen absoluuttinen skaalattu virhe) koko aineistolle. Yksittäisten tuntien virheiden vertailu on luontevaa prosentuaalisella virheellä, kun taas koko aineistolle skaalattu virhe antaa hyvän vertailukohteen. Yksittäiselle tunnille MASE olisi sama kuin absoluuttinen virhe, sillä skaalaaminen tapahtuu koko aineistolle kerrallaan.

Taulukko 4.1: Jaksollisuusanalyysi spektritiheydestä

Jakson kestoaika (h)	Spektritiheys
9000.0000	100839609585
24.0000	11431925678
12.0000	6261272992
10285.7143	6061244300
8000.0000	5781432478
7200.0000	4948674589
72000.0000	4031056844
167.8322	3079678417
4500.0000	2524812039
84.0140	2279586527

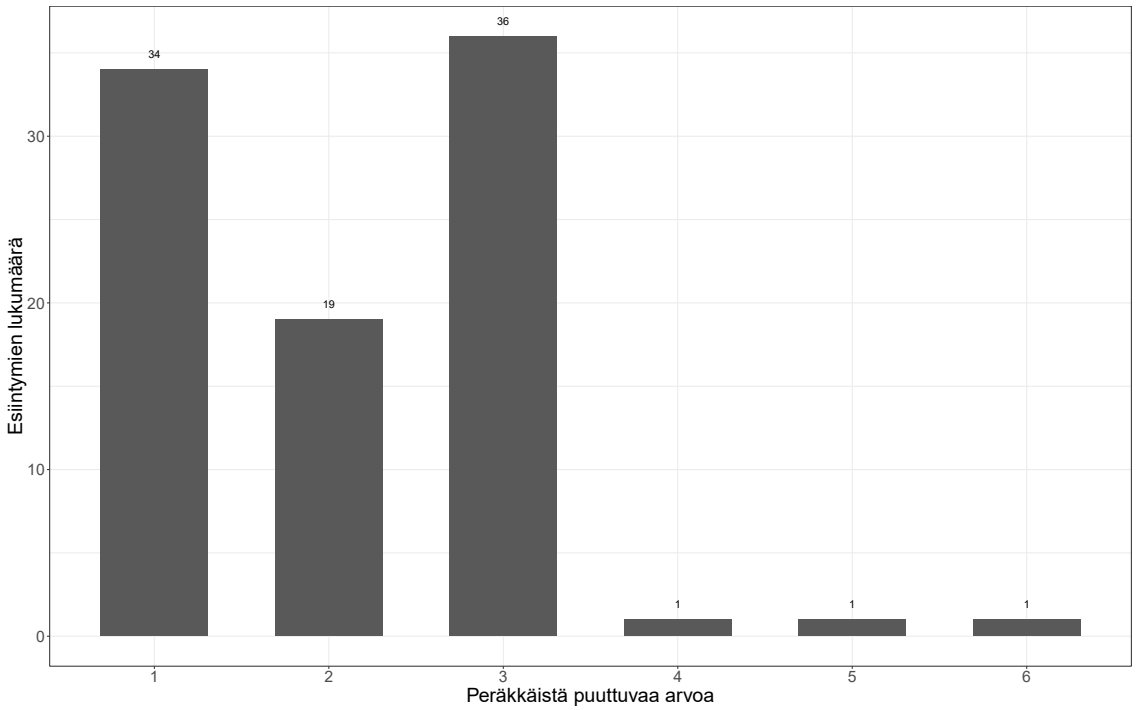
4.5.1 Malli 1

Malli valittiin ja estimoitiin R-ohjelmaympäristön *forecast*-paketin [31] *auto.arima*-funktiolla. Tämä käy malliavaruutta läpi askeleittain ja vaihtaa viereiseen malliin (p:n, d:n ja q:n suhteen) jos tällä on parempi AIC-arvo, kunnes parempaa mallia ei löydy. Mallin parametrien estimointi tapahtuu suurimman uskottavuuden menetelmällä.

Mallille annettiin ulkoisiksi muuttujiksi Hämeenlinnan lämpötilatiedot seuraavalle tunnille simuloimaan sääennustusta, sekä kausittaisuutta mallintamaan päivän, puolen päivän, kolmasosapäivän, puolen viikon, viikon ja vuoden sykleissä kulkevat jaksolliset termit $sin(2\pi x/T), cos(2\pi x/T)$, missä $T \in [8, 12, 24, 84, 168, 365.25 * 24]$, kuvaamaan kausittaista vaihtelua.

4.5.2 Malli 2

Malli 2 tehtiin myös R/forecast-ohjelmalla, mutta tällä kertaa päiväkausittaisuus otettiin malliin mukaan, eli malli on $SARIMA(p, d, q)(P, D, Q)_{24}$-tyyppiä. Ulkoisina muuttuji-



Kuva 4.4: Säähavaintojen peräkkäisten puuttuvien arvojen jakauma

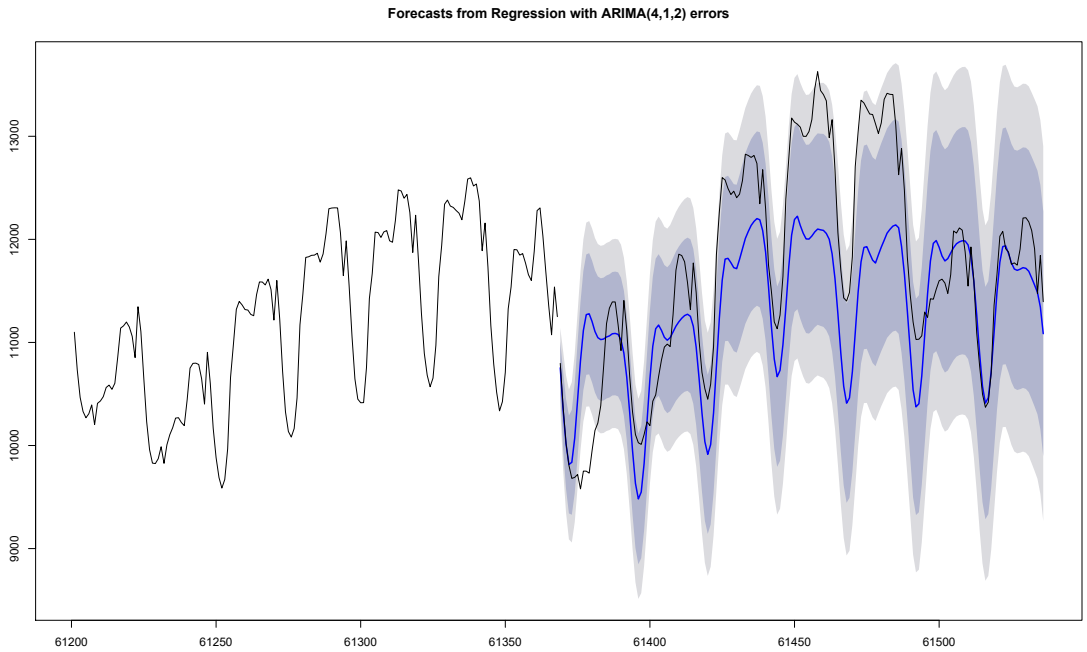
na ovat lämpötila, vuosisykli, viikkosykli ja puoliviikkosykli. Päivän ja sitä lyhyemmän syklin muuttujat jätettiin pois, sillä kausittainen 24 tunnin differointi hävittää nämä täysin ja mallin havaittiin huonontuvan.

4.6 Testitulokset

4.6.1 Malli 1

Ohjelman ehdottama malli on regressio $ARIMA(4, 1, 2)$-virheillä. Hyvyyskriteereiksi saatiin testijoukossa $AIC = 115840.6, BIC = 115847.7$ Mallin parametrit on luetelu taulukossa 4.2. Ensimmäisen sarake kertoo muuttujan nimen. Autoregressiiviset (AR) sekä liukuvan keskiarvon (MA) termit ovat numeroidut. *wx* on ulkoinen tekijä eli lämpötila ja kausittaisen vaihtelun sini- ja kosinitermit ovat muotoa λa_b, missä λ on vuosi, viikko tai päivä, a on taajuuden kerroin ja b erottelee kaksi komponenttia keskenään, eli

esimerkiksi $W2\_2$ on puoliviikon kosinitermi. Kerroinsarake kertoo mallissa luvulle asetetun kertoimen ja standardivirhesarake kertoo muuttujan virheen suuruuden. Kuvassa 4.5 esitetään mallin antama ennuste vuoden ensimmäiselle viikolle.



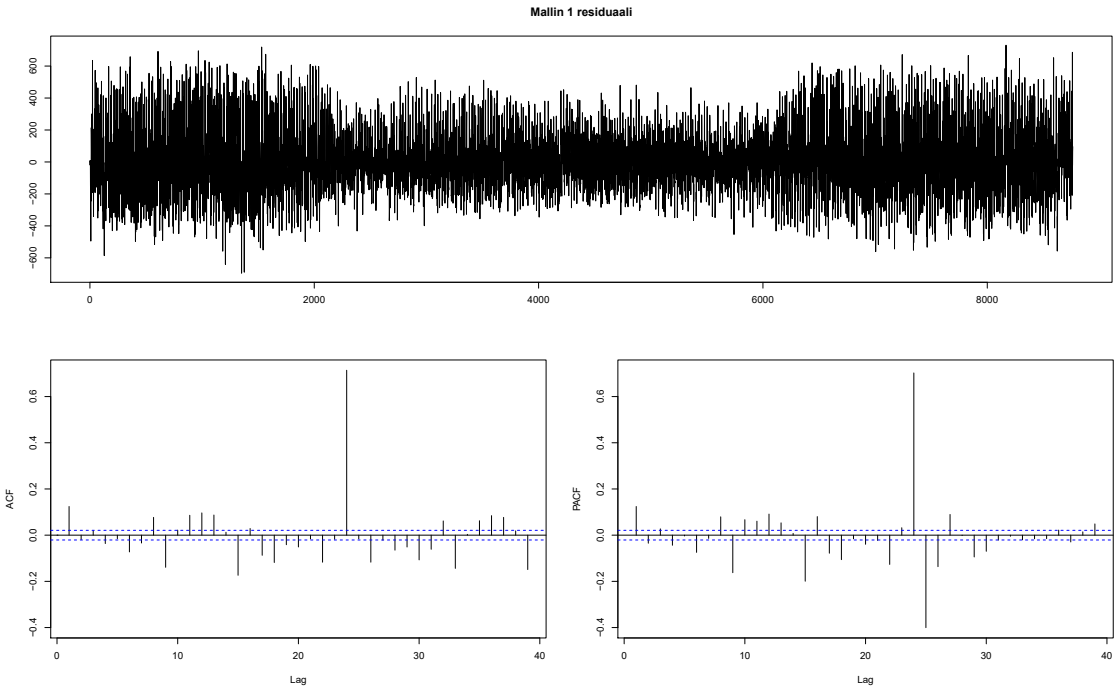
Kuva 4.5: Mallin 1 ennuste viikoksi eteenpäin

Mallin residuaalit eivät kuitenkaan läpäise Ljung–Box-testiä edes viiveellä 7, eli jäännöksissä on jäljellä autokorrelaatioita, joita malli ei pystynyt esittämään. Tämä on selvästi havaittavissa kuvasta 4.6, missä on katkoviivalla merkitty autokorrelaation ja osittaisen autokorrelaation 95% luottamusväli, joiden sisällä siis 95% autokorrelaatioista pitäisi olla, olettaen että jäännökset ovat identtisesti normaalijakautuneet ja riippumattomat. Sen sijaan huomaamme voimakkaan piikin 24 tunnin kohdalla, mikä tarkoittaa sitä, että päiväsyklin muuttuja ei ole tarpeeksi voimakas kuvaamaan kausittaisuutta. Residuaalit ovat myös voimakkaasti heteroskedastisia, sillä kesällä residuaalin varianssi on selvästi pienempi kuin talvella.

Mallin 1 keskimääräinen prosentuaalinen virhe (MAPE) testijoukossa oli 1.412076% ja keskimääräinen skaalattu virhe (MASE) oli 0.6507736, eli malli oli kuitenkin parempi

Taulukko 4.2: Mallin 1 parametrit

Muuttuja	Kerroin	standardivirhe
ar1	1.9638	0.0062
ar2	-1.3159	0.0099
ar3	0.4937	0.0087
ar4	-0.1838	0.0042
ma1	-1.8599	0.0049
ma2	0.8739	0.0048
wx	-23.7806	1.0326
Y1_1	1627.9553	522.1991
Y1_2	411.8931	521.5104
W1_1	-309.9117	9.9305
W1_2	271.0276	9.9260
W2_1	-3.1151	5.1300
W2_2	267.1732	5.1303
D1_1	-439.9532	6.4327
D1_2	-459.2407	6.1357
D2_1	9.0479	2.4068
D2_2	-407.3536	2.3984
D3_1	114.5301	1.6696
D3_2	-37.1295	1.6710



Kuva 4.6: Mallin 1 residuaali, autokorrelaatio ja osittainen autokorrelaatio

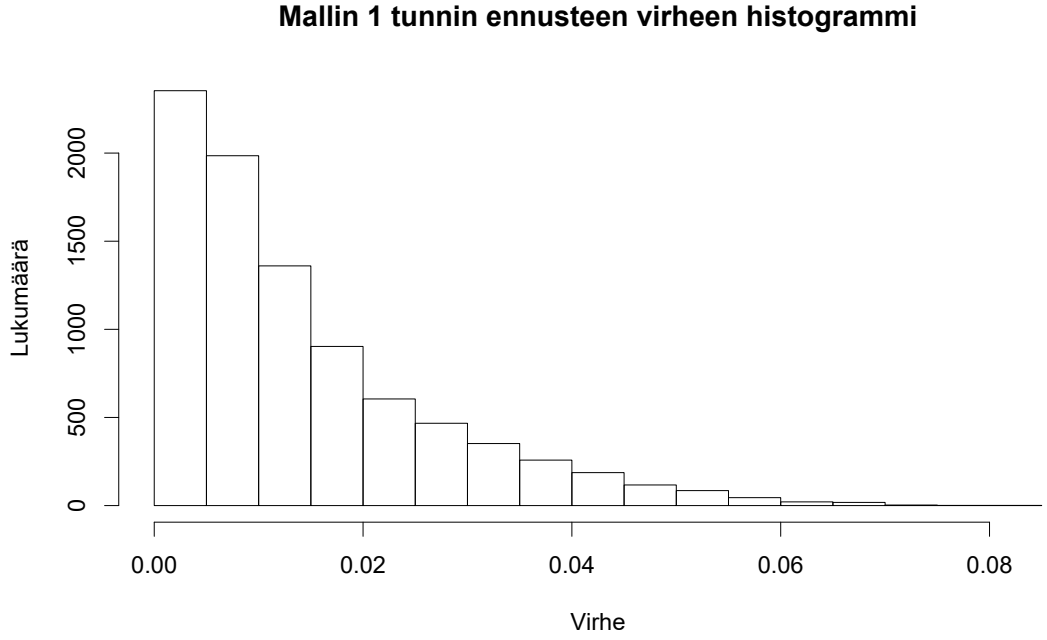
kuin naiivi edellisen tunnin estimaatti. Mallin virheen jakaumasta esitetään histogrammi kuvassa 4.7.

Taulukossa 4.3 esitetään mallin 1 25 suurinta prosentuaalista virhettä. Residuaali on mallin yhden tunnin ennusteen ja todellisen kulutuksen ero. Kulutussarake on todellinen kulutus. Kuukausi- ja päiväsarakkeet kertovat miltä päivältä tieto on. Tunti vaihtelee välillä 0-23. Viikonpäivä vaihtelee välillä 0-6, missä 0 on maanantai ja 6 sunnuntai. Virhesarake on absoluuttinen suhteellinen virhe eli virhe jaettuna todellisella kulutuksella.

Taulukosta havaitaan, että mallilla oli usein vaikeuksia ennustaa aamulla kuudelta tapahtuva nopea kulutuksen kasvu, mikä havaitaan hyvin myös kuvasta 4.8, jossa on esitetty tunneittain keskimääräinen absoluuttinen prosentuaalinen virhe. Samasta kuvasta havaitsemme myös, että iltakymmeneltä tapahtuva pieni kulutuspiikki näkyy suurina virheinä. Malli ei siis ole oppinut ennustamaan näitä kulutuspiikkejäkään kunnolla. Samoin arjen ja viikonlopun vaihtuminen tuotti mallille ongelmia, mikä näkyy viikonpäivien 0 ja 5 eli maanantain ja lauantain runsaudessa taulukossa 4.3. Erikoispäiviä, kuten joulupyhät

Taulukko 4.3: Mallin 1 suurimmat prosentuaaliset virheet

Residuaali	Kulutus	Kuukausi	Päivä	Tunti	Viikonpäivä	Suhteellinen virhe
-557.4387	6941	12	26	7	0	0.08031101
652.0748	8411	12	24	22	5	0.07752643
451.8678	6201	6	25	23	5	0.07287014
619.5149	8585	9	26	6	0	0.07216248
444.6805	6230	6	24	23	4	0.07137728
-453.7016	6535	8	28	6	6	0.06942641
-689.5730	9945	2	27	7	6	0.06933866
730.5564	10566	12	7	6	2	0.06914219
568.2986	8264	9	21	6	2	0.06876798
596.2034	8773	9	29	6	3	0.06795889
671.7071	9950	10	29	18	5	0.06750825
581.7221	8661	10	6	6	3	0.06716570
686.2289	10288	12	31	22	5	0.06670188
601.4801	9028	10	12	6	2	0.06662385
666.1576	10032	11	23	6	2	0.06640327
422.4125	6375	6	27	6	0	0.06626078
570.7763	8617	9	28	6	2	0.06623841
558.5556	8457	10	5	6	2	0.06604654
601.7393	9166	11	2	6	2	0.06564906
-463.3358	7087	9	25	7	6	0.06537828
570.2990	8744	10	7	6	4	0.06522175
582.5333	8954	10	10	6	0	0.06505844
718.9733	11057	3	5	18	5	0.06502427
534.3936	8365	12	25	22	6	0.06388447
-695.8360	10961	2	26	7	5	0.06348289



Kuva 4.7: Mallin 1 prosentuaalisen virheen jakauma

ja uudenvuodenaatto, listassa on myös paljon, mikä on luonnollista, sillä lomapäiviä ei otettu huomioon mallissa. Kolme suurinta yksittäistä virhettä ovat maanantaille osuneena tapaninpäivänä, jouluaattona ja juhannuspäivänä.

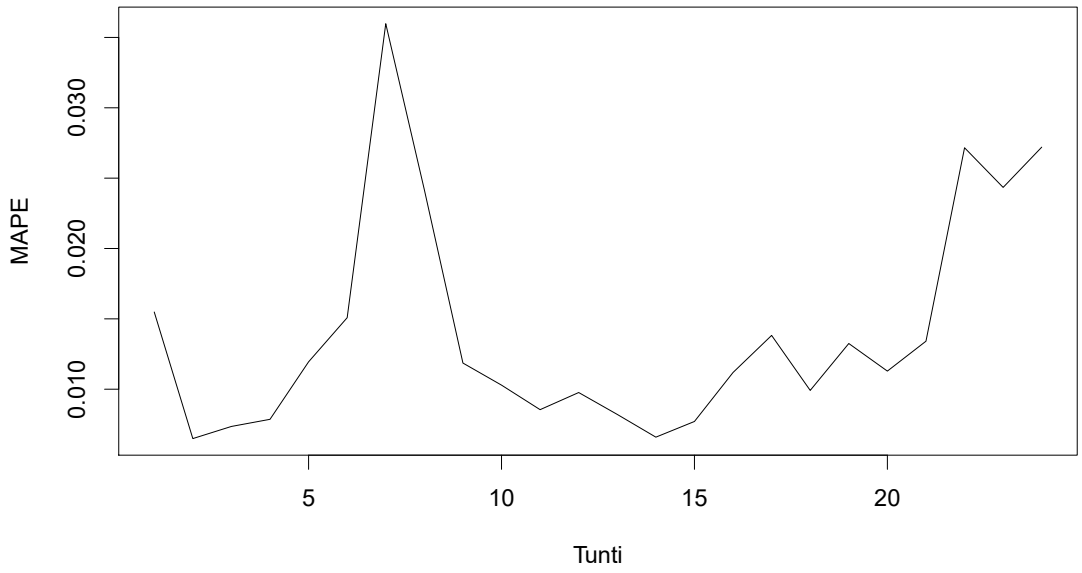
4.6.2 Malli 2

Ohjelman ehdottama malli oli regressio $ARIMA(5,0,0)(2,1,0)_{24}$-virheillä. Mallin parametrit ja niiden kertoimet sekä standardivirheet on esitetty taulukossa 4.4. Muuttujat ovat kuten mallissa 1, paitsi että päiväsyklin ulkoiset muuttujat puuttuvat ja lisäksi on tullut kaksi kausittaista AR-termiä, *sar1* ja *sar2*. Mallin luoma ennuste testivuoden ensimmäiselle viikolle esitetään kuvassa 4.9.

Hyvyyskriteereiksi testijoukossa saatiin $AIC = 105979.8, BIC = 105986.8$. Keskimääräinen prosentuaalinen virhe MAPE seuraavan tunnin ennusteille oli 0.7969549% ja keskimääräinen skaalattu virhe MASE oli 0.3593386. Virheen histogrammi on esitetty kuvassa 4.11.

Taulukko 4.4: Mallin 2 parametrit

Muuttuja	Kerroin	standardivirhe
ar1	1.4649	0.0042
ar2	-0.5573	0.0073
ar3	-0.0290	0.0075
ar4	0.1131	0.0071
ar5	-0.0438	0.0040
sar1	-0.4023	0.0040
sar2	-0.2842	0.0039
wx	-4.0116	0.6148
Y1_1	1694.8784	438.5024
Y1_2	404.0711	446.0568
W1_1	-313.5292	10.3659
W1_2	270.7995	10.3637
W2_1	-3.3809	9.3700
W2_2	266.3715	9.3734



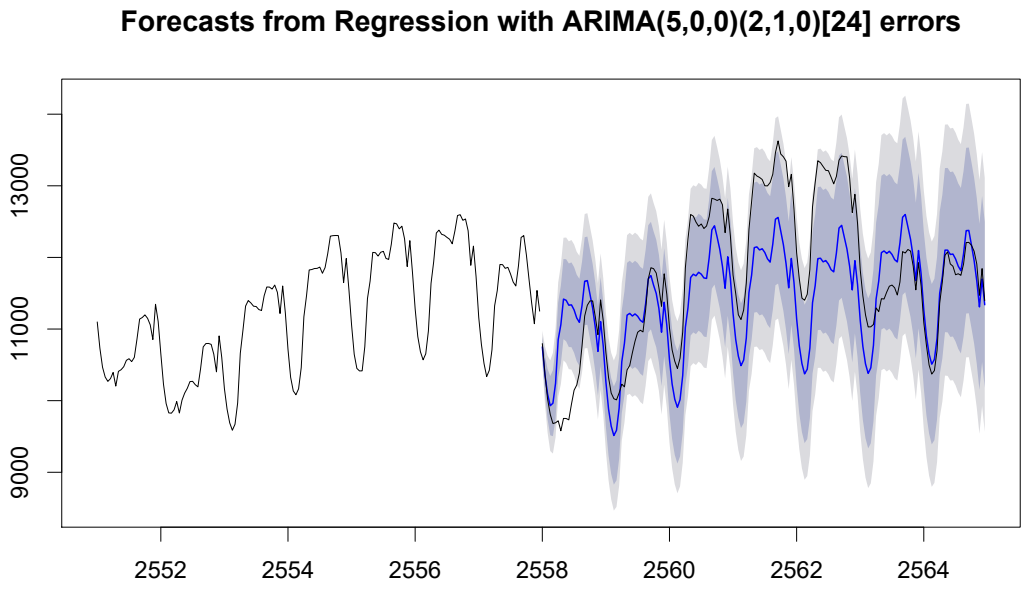
Kuva 4.8: Mallin 1 keskimääräinen prosentuaalinen virhe tunneittain

Kuvasta 4.10 havaitaan, että residuaaleissa on viikon välein suurempi piikki, eli malli ei oppinut viikottaista jaksollisuutta täysin. Malli 2 ei myöskään läpäissyt Ljung–Box-testiä. Kuten kuvasta 4.10 havaitaan, residuaalin autokorrelaatiot eivät ole pysyneet 95% luottamusrajojen sisällä.

Taulukossa 4.5 esitetään mallin 2 25 suurinta prosentuaalista virhettä. Sarakkeet ovat identtiset mallin 1 vastaavan taulukon kanssa. Mallilla 2 oli erityisen suuria vaikeuksia lauantaiaamun arkipäivää hitaamman kulutusnousun kanssa, sillä melkein kaikki suurimmat virheet esiintyivät lauantaina tunnilla 6. Muutenkin kaikki suurimmat virheet olivat tunnilla 6. Kuvassa 4.12 on esitetty virheen jakautuminen eri tunneille, mistä nähdään selkeästi mallin ongelma aamun kulutusnousun ennustamisessa. Malli 2 kuitenkin oppi mallia 1 paremmin illan kulutuspiikin, sillä kello 22 esiintyvä virhepiikki ei ole yhtä suuri kuin mallin 1 vastaava piikki (kuva 4.8). Vaikka malli 2 keskimäärin paransi ennustuksia joka tunnilla malliin 1 nähden, aamukuuden virhepiikki erottuu jopa paremmin muiden virheiden ollessa verrattain pieniä.

Taulukko 4.5: Mallin 2 suurimmat virheet

Residuaali	Kulutus	Kuukausi	Päivä	Tunti	Viikonpäivä	Suhteellinen virhe
541.8360	6375	6	27	6	0	0.08499388
-441.6696	5631	6	24	6	4	0.07843538
-566.6974	7341	9	17	6	5	0.07719622
-534.0844	7005	8	27	6	5	0.07624332
-568.6512	7507	9	24	6	5	0.07574946
-588.7294	7821	10	1	6	5	0.07527546
-596.3406	8041	10	8	6	5	0.07416249
-537.2247	7334	9	3	6	5	0.07325126
616.0170	8420	4	26	6	1	0.07316116
-583.8572	8185	11	5	6	5	0.07133259
-502.2939	7064	8	20	6	5	0.07110616
-546.5239	7757	4	30	6	5	0.07045557
-507.5631	7372	5	28	6	5	0.06885012
-506.3193	7384	12	24	6	5	0.06856979
-594.4316	8671	11	12	6	5	0.06855399
-678.3577	9912	1	29	6	5	0.06843803
-580.9865	8575	10	15	6	5	0.06775353
704.0158	10566	12	7	6	2	0.06663030
-602.4991	9064	12	10	6	5	0.06647165
-536.2424	8071	4	22	6	4	0.06644064
-535.8994	8099	10	22	6	5	0.06616858
-483.1752	7486	5	21	6	5	0.06454384
-489.3931	7617	5	14	6	5	0.06425012
-471.7960	7402	6	2	6	3	0.06373899
-562.6090	8887	11	26	6	5	0.06330696

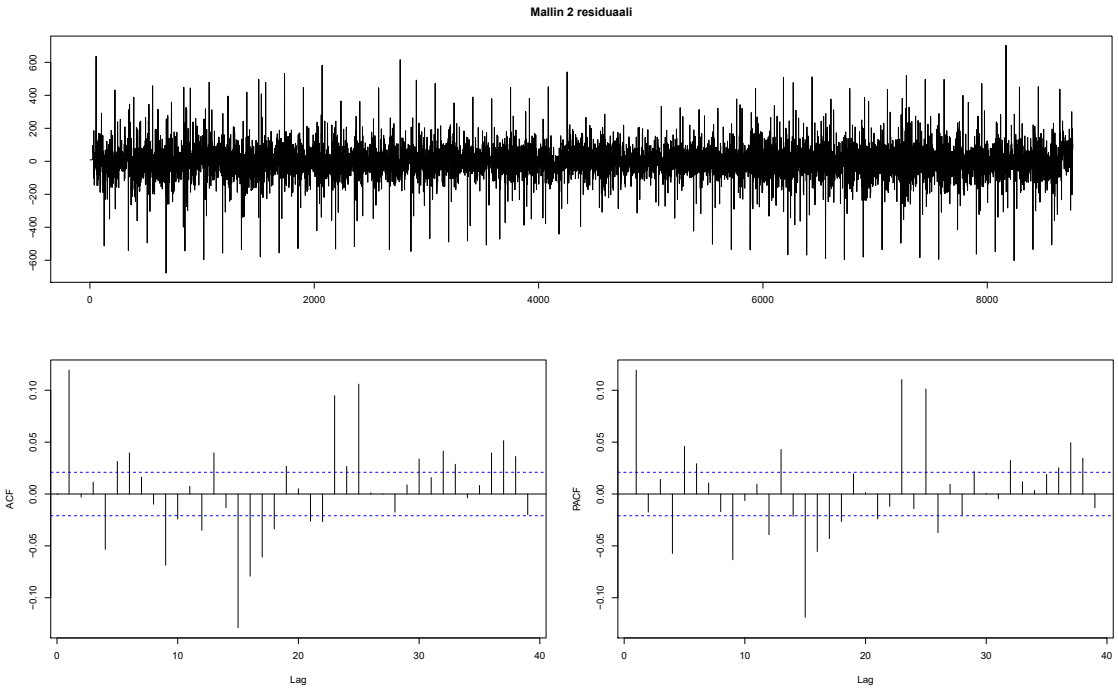


Kuva 4.9: Mallin 2 ennuste viikoksi eteenpäin

4.7 Mallien 1 ja 2 vertailu

Kuvassa 4.13 on esimerkki mallien 1 ja 2 seuraavan tunnin ennusteista kahdelle vuorokaudelle.

Seuraavan tunnin ennusteiden lisäksi malleista otettiin 24 tunnin ennusteet seuraavalle päivälle, joista laskettiin päivittäinen virheen keskiarvo. Mallin 1 keskimääräinen virhe koko päivän ennustuksille oli 3.491221% ja mallin 2 virhe oli 3.394646%. Kuvissa 4.14 ja 4.15 esitetään molempien mallien 24 tunnin ennusteen virheen jakauma. Huomataan, että jopa 24 tunnin ennusteessa aamun kulutusnousun ennustaminen on vaikeinta molemmille malleille, vaikka myöhempien tuntien ennustuksiin vaikuttaa edellisten tuntien kumulatiivinen virhe. Taulukossa 4.6 on vielä lopuksi kerrattu mallien 1 ja 2 tärkeimmät tulosluvut.



Kuva 4.10: Mallin 2 residuaali, autokorrelaatio ja osittainen autokorrelaatio

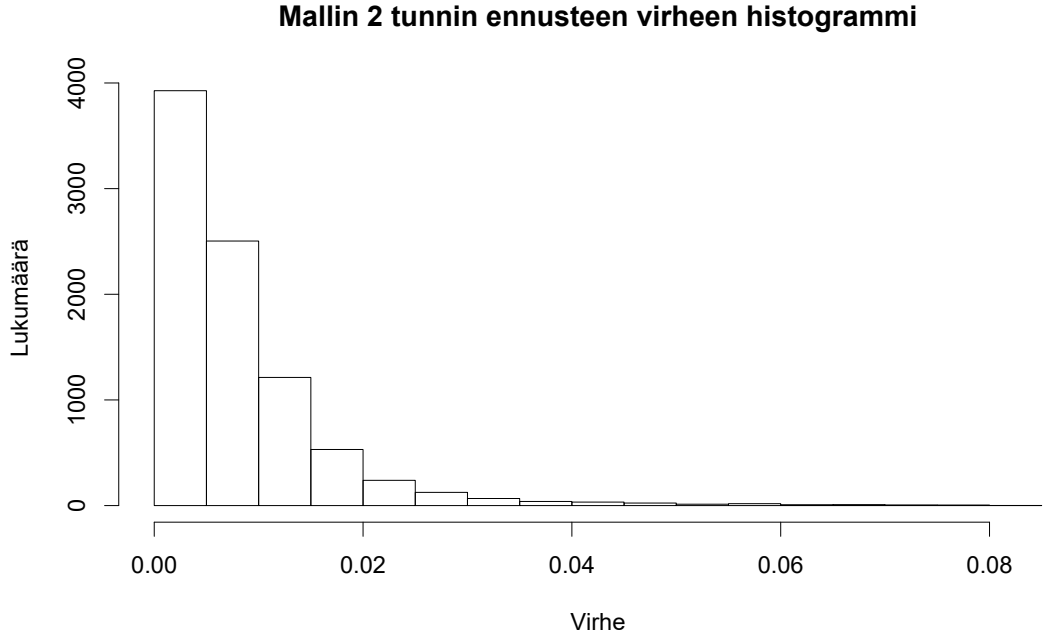
Taulukko 4.6: Tärkeimmät tulosluvut

Tulos	Malli 1	Malli 2
Seuraavan tunnin ennusteen MAPE	1.412076%	0.7969549%
Seuraavan tunnin ennusteen MASE	0.6507736	0.3593386
24 tunnin ennusteen virhe	3.491221%	3.394646%

4.8 Piirteenvalinta

Dataa käytettiin myös testaamaan piirteenvalintaa ahnaalla eteenpäinvalinnalla, missä yhtä ulkoista muuttujaa kerrallaan yritetään lisätä malliin, mutta vain paras muuttuja (se, joka laskee mallin AIC-arvoa eniten) lisätään malliin. Tätä jatketaan, kunnes kaikki muuttujat on saatu lisättyä. Koska ulkoisina muuttujina olleet syklit oli jaettu kahteen komponenttimuuttujaan, nämä lisättiin aina pareittain, sillä näiden kertoimien ero riippuu vain siitä, mihin vaiheeseen sykliä kulutuspiikki sattuu osumaan.

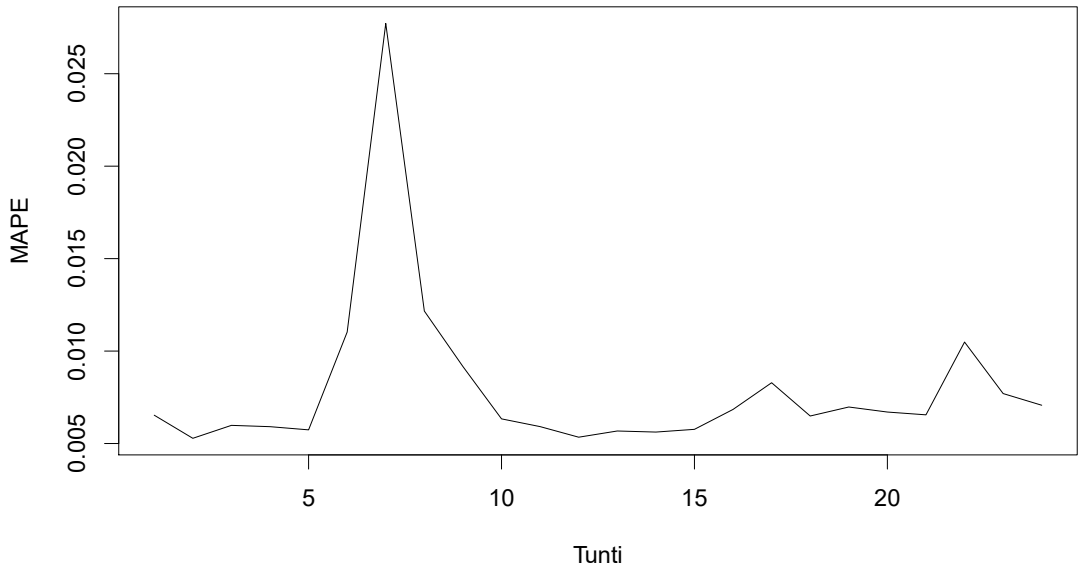
Mallina käytettiin regressiota kausittaisilla ARIMA-virheillä, kuten mallissa 2. Käy-



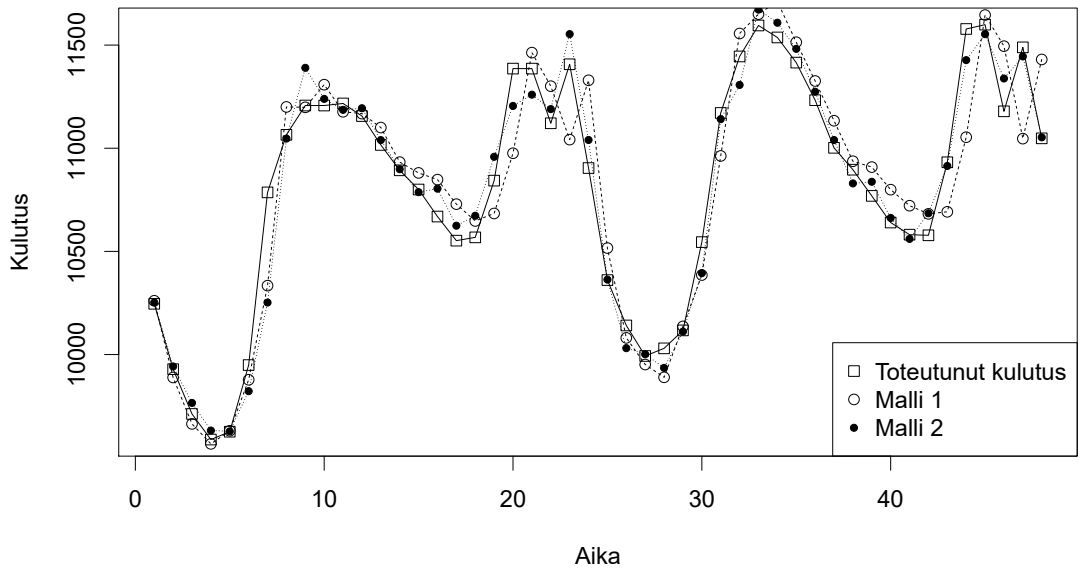
Kuva 4.11: Mallin 2 prosentuaalisen virheen jakauma

tetyt ulkoiset tekijät olivat lämpötila *wx*, vuosisykli *Year*, viikkosykli *Week1* ja puoliviikkosykli *Week2*. Kuten mallissa 2, päiväsykliä ei otettu mukaan, sillä tämä kausittaisuus otetaan huomioon mallissa.

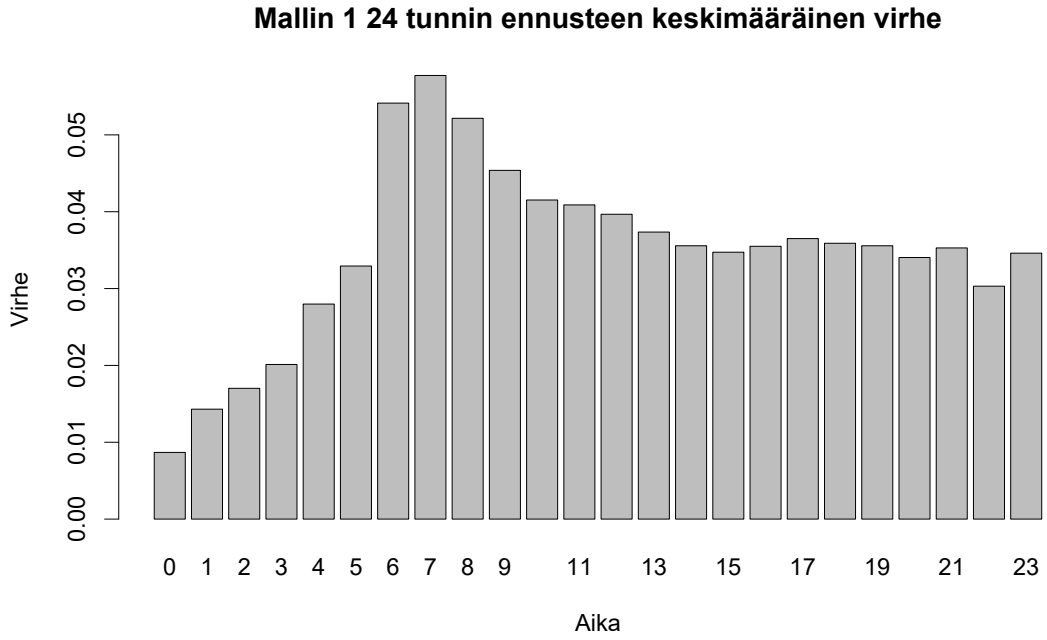
Kuten kuvasta 4.16 havaitaan, lyhyemmän aikavälin muuttujat valikoituvat ennen pitemmän aikavälin muuttujia. Säämuuttuja ei myöskään ole niin merkityksellinen kuin voisi olettaa, eikä nähtävästi paranna mallin AIC-arvoa.



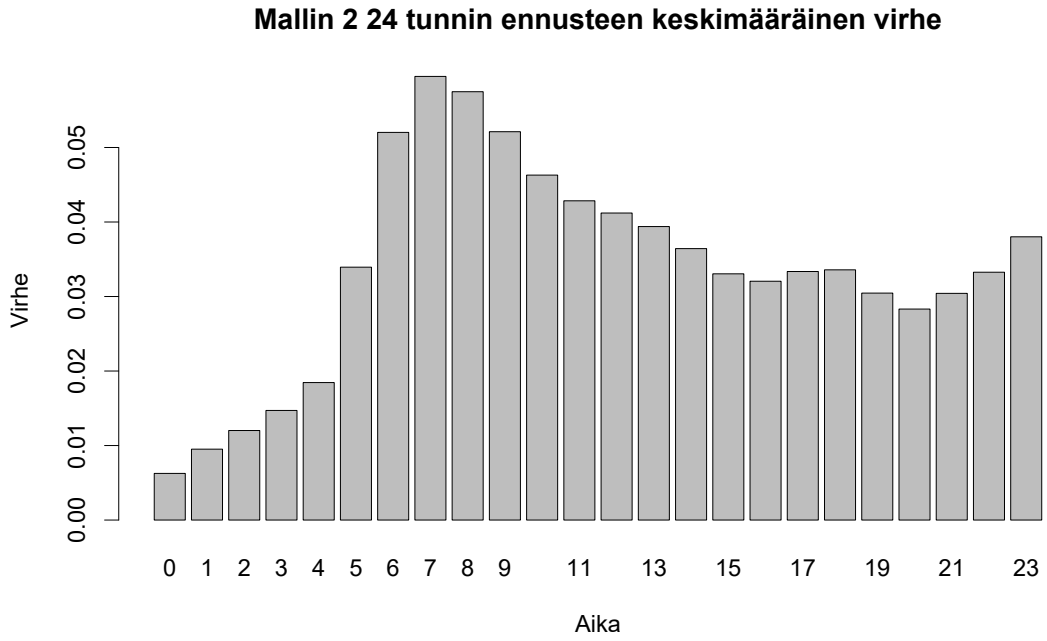
Kuva 4.12: Mallin 2 keskimääräinen prosentuaalinen virhe tunneittain



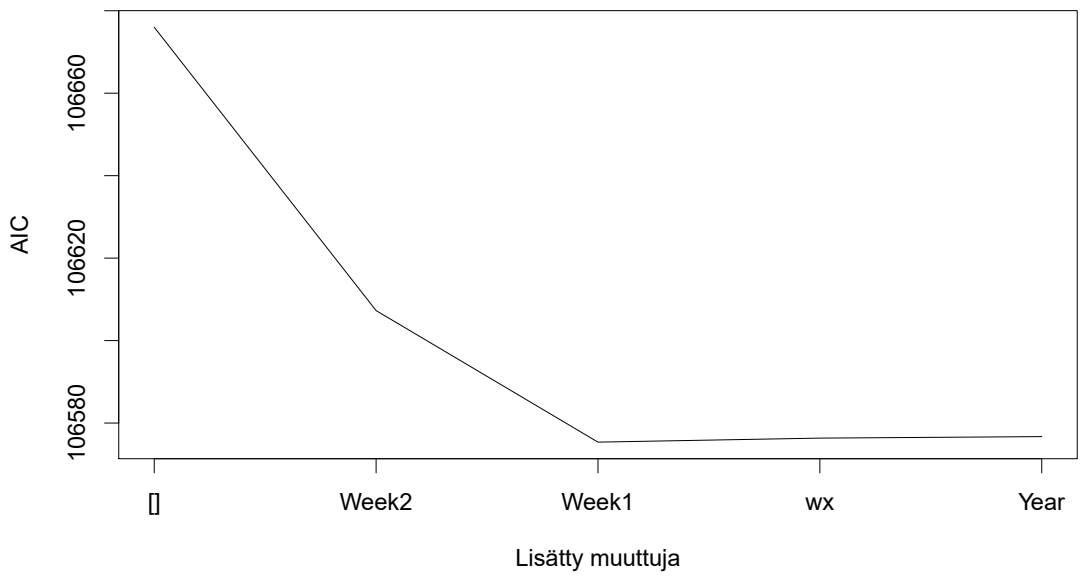
Kuva 4.13: Esimerkki mallien seuraavan tunnin ennusteista



Kuva 4.14: Mallin 1 24 tunnin ennusteen keskimääräinen tunnittainen virhe



Kuva 4.15: Mallin 2 24 tunnin ennusteen keskimääräinen tunnittainen virhe



Kuva 4.16: Piirteenvalinta ahneella eteenpäinvalinnalla

5 Yhteenveto

Tässä tutkielmassa tutkittiin, miten Suomen sähkönkulutusta saa mallinnettua ARIMA-mallilla lämpötilatietoja apuna käyttäen ja miten hyviä ennusteita mallista saadaan seuraavalle tunnille ja seuraaville 24 tunnille.

Työssä testattiin kahta autoregressiivistä mallia, regressiota ARIMA-virheillä ja regressiota SARIMA-virheillä. Kumpikaan malleista ei saanut täysin mallinnettua sähkönkulutuksen voimakkaita syklisiä autokorrelaatioita, sillä kumpikaan ei läpäissyt Ljung–Box-testiä. Mallit ovat siis stokastisessa mielessä epäonnistuneita. Sähkön kulutus on vaikea mallintaa autoregressiivisellä mallilla, sillä eri kausittaisuuksia on useita eikä mikään näistä ole täysin sinimuotoinen vaan voimakkaasti ihmisten toiminnallaan muokkaama ja kausittaisissa autoregressiivisissa malleissa differointi onnistuu vain yhdelle kausittaisuudelle. Autoregressiivinen malli ei välttämättä ole paras malli oppimaan sähkönkulutusta.

Etenkin aamukuuden kulutuksen nousu, yösähköstä johtuva kulutuspiikki ja lauantain laiska kuluksen nousu tuottivat malleille vaikeuksia. Lisäksi vapaapäivät ja arkipyhät tuottavat omat ongelmansa, jotka saattaisivat olla ratkaistavissa lisäämällä malliin tieto erikoispäivistä.

Pitkät syklit kuten vuosisykli eivät havaittavasti auta autoregressiivistä mallia, mikä on loogista, sillä autoregressio lasketaan vain muutamasta edellisestä tunnista, eikä vuosisykli ehdi tässä ajassa muuttua juuri yhtään. Etenkin differoinnin jälkeen vuosisyklitermi on likimain nolla. Vuosisykli ja ulkolämpötila myös korreloivat jokseenkin vahvasti (li-

neaarimallin $R^2 = 0.6312$), ja säätieto on näistä kahdesta oleellisempi autoregressiivisen mallin kannalta ollen nopeampi vaihteluissaan.

Suurin ongelma autoregressiivisillä malleilla sähkönkulutuksen kanssa oli oppia kulutuksen monikausittaisuus, sillä kulutus nousee tai laskee voimakkaasti, mutta säännöllisesti, kuten arkena tai viikonloppuna. Kulutuskäyrän muoto voi myös vaihdella eri aikoina vuodesta, mitä malli ei välttämättä opi. Kausittaisuus voisikin mallintua paremmin, jos se esitettäisiin sinikäyrien sijaan vuodenajan mukaan vaihtelevana keskiarvoprofiilina tai yhdistettynä hybridinä esim. samankaltainen päivä -mallin kanssa. Myös tieto yösähkön alkutunnista voitaisiin antaa mallille tietona.

Vaikka malli stokastisessa mielessä ei ollut täysin onnistunut, onnistui etenkin malli 2 kuitenkin tekemään kohtalaisen hyviä ennustuksia. Parantamisen varaa on vielä etenkin lomapäivissä ja aamukuuden ennustuksissa.

Lähdeluettelo

[1] Fingrid, *Sähkön kulutus ja tuotanto*, http://www.fingrid.fi/fi/sahkomarkkinat/kulutus-ja-tuotanto/Sivut/default.aspx, Viitattu 2019-08-28.

[2] Nord Pool, *Market data*, Viitattu 2019-08-28. url: https://www.nordpoolgroup.com/historical-market-data/.

[3] Energiateollisuus ry, *Sähkönkäyttö kunnittain 2007-2017*, https://energia.fi/ajankohtaista_ja_materiaalipankki/materiaalipankki/sahkonkaytto_kunnittain_2007-2017.html, Viitattu 2019-08-28.

[4] Oikeusministeriö, *Laki sähkön ja eräiden polttoaineiden valmisteverosta*, Viitattu 2019-08-28. url: http://www.finlex.fi/fi/laki/ajantasa/1996/19961260.

[5] R. Weron, *Modeling and forecasting electricity loads and prices: a statistical approach*, sarja Wiley finance series. John Wiley & Sons, 2006, ISBN: 9780470057537. url: http://books.google.com.pe/books?id=dBl%5C_37mLv6EC.

[6] E. A. Feinberg ja D. Genethliou, "Applied Mathematics for Restructured Electric Power Systems: Optimization, Control, and Computational Intelligence", teoksessa, J. H. Chow, F. F. Wu ja J. Momoh, toim. Boston, MA: Springer US, 2005, luku Load Forecasting, s. 269–285, ISBN: 978-0-387-23471-7. DOI:

10.1007/0-387-23471-3_12. url:
http://dx.doi.org/10.1007/0-387-23471-3_12.

[7] G. Gross ja F. D. Galiana, ”Short-term load forecasting”, *Proceedings of the IEEE*, vol. 75, nro 12, s. 1558–1573, joulukuu 1987, ISSN: 0018-9219. DOI: 10.1109/PROC.1987.13927.

[8] L. Hernandez, C. Baladron, J. M. Aguiar, B. Carro, A. J. Sanchez-Esguevillas, J. Lloret ja J. Massana, ”A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings”, *IEEE Communications Surveys Tutorials*, vol. 16, nro 3, s. 1460–1495, ThirdThird 2014, ISSN: 1553-877X. DOI: 10.1109/SURV.2014.032014.00094.

[9] Nord Pool, *Intraday market*, Viitattu 2019-08-28. url: https://www.nordpoolgroup.com/trading/intraday-trading/.

[10] M. R. Berthold, C. Borgelt, F. Hppner ja F. Klawonn, *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*, 1st. Springer Publishing Company, Incorporated, 2010, ISBN: 1848822596, 9781848822597.

[11] Y. Chen, P. B. Luh, C. Guan, Y. Zhao, L. D. Michel, M. A. Coolbeth, P. B. Friedland ja S. J. Rourke, ”Short-Term Load Forecasting: Similar Day-Based Wavelet Neural Networks”, *IEEE Transactions on Power Systems*, vol. 25, nro 1, s. 322–330, helmikuu 2010, ISSN: 0885-8950. DOI: 10.1109/TPWRS.2009.2030426.

[12] H. A. Dryar, ”The Effect of Weather on the System Load”, *Transactions of the American Institute of Electrical Engineers*, vol. 63, nro 12, s. 1006–1013, joulukuu 1944, ISSN: 0096-3860. DOI: 10.1109/T-AIEE.1944.5058843.

[13] M. Hagan ja R. Klein, ”On-Line Maximum Likelihood Estimation for Load Forecasting”, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, nro 9,

s. 711–715, syyskuu 1978, ISSN: 0018-9472. DOI: 10.1109/TSMC.1978.4310059.

[14] E. Ceperic, V. Ceperic ja A. Baric, ”A Strategy for Short-Term Load Forecasting by Support Vector Regression Machines”, *IEEE Transactions on Power Systems*, vol. 28, nro 4, s. 4356–4364, marraskuu 2013, ISSN: 0885-8950. DOI: 10.1109/TPWRS.2013.2269803.

[15] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu ja J. He, ”Short-Term Load Forecasting With Deep Residual Networks”, *IEEE Transactions on Smart Grid*, vol. 10, nro 4, s. 3943–3952, heinäkuu 2019, ISSN: 1949-3053. DOI: 10.1109/TSG.2018.2844307.

[16] W. T. Ghareeb ja E. F. El Saadany, ”Multi-Gene Genetic Programming for Short Term Load Forecasting”, teoksessa *2013 3rd International Conference on Electric Power and Energy Conversion Systems*, lokakuu 2013, s. 1–5. DOI: 10.1109/EPECS.2013.6713061.

[17] J. Yu, H. Lee, Y. Jeong, E. K. Kim ja S. Kim, ”CRPSO-based automatic TSK fuzzy model extraction for one hour ahead load forecasting”, teoksessa *2014 International Conference on Fuzzy Theory and Its Applications (iFUZZY2014)*, marraskuu 2014, s. 148–152. DOI: 10.1109/iFUZZY.2014.7091249.

[18] N. Sovann, P. Nallagownden ja Z. Baharudin, ”One-Hour Ahead Load Forecasting Based on Wavelet Neural Networks”, English, *Applied Mechanics and Materials*, vol. 785, s. 53–57, elokuu 2015, Copyright - Copyright Trans Tech Publications Ltd. Aug 2015; Last updated - 2018-10-06. url: https://search-proquest-com.ezproxy.utu.fi/docview/1707910519?accountid=14774.

[19] ABB Asea Brown Boveri Ltd, *ABB Industrial Energy Load Forecasting and Planning software - ABB Energy Manager software solution for industrial plants*,

https://new.abb.com/cpm/energy-manager/industrial-energy-load-planning-forecasting-scheduling, Viitattu 2019-08-28.

[20] ETAP, *Load Forecasting Software | Load Forecasting Analysis*, https://etap.com/product/load-forecasting-software, Viitattu 2019-08-28.

[21] GMDH, *Best Electric Load Forecasting Software 2019*, https://gmdhsoftware.com/electricity-load-forecasting-software, Viitattu 2019-08-28.

[22] SAS Institute Inc., *SAS Energy Forecasting*, https://www.sas.com/en_us/software/energy-forecasting.html, Viitattu 2019-08-28.

[23] Itron Inc., *Energy Forecasting | Analytics*, https://www.itron.com/na/solutions/what-we-enable/analytics/forecasting, Viitattu 2019-08-28.

[24] The Weather Company, *Load Forecasting - Short & Long Term*, https://business.weather.com/products/load-forecasts, Viitattu 2019-08-28.

[25] Enfor A/S, *Accurate electricity demand forecasting for electricity companies*, https://enfor.dk/services/loadfor/, Viitattu 2019-08-28.

[26] G. Box, G. Jenkins ja G. Reinsel, *Time Series Analysis: Forecasting and Control*, sarja Wiley Series in Probability and Statistics. Wiley, 2008, ISBN: 9780470272848.

[27] P. J. Brockwell ja R. A. Davis, *Time Series: Theory and Methods*, 1991. painos. New York, NY: Springer-Verlag, 1991.

[28] R. Shumway ja D. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, sarja Springer Texts in Statistics. Springer New York, 2010, ISBN: 9781441978653. url: https://books.google.fi/books?id=NIhXa6UeF2cC.

[29] M. Stone, "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, nro 1, s. 44–47, 1977, ISSN: 00359246. url: http://www.jstor.org/stable/2984877.

[30] R. J. Hyndman ja A. B. Koehler, "Another look at measures of forecast accuracy", *International Journal of Forecasting*, vol. 22, nro 4, s. 679–688, 2006, ISSN: 0169-2070. DOI: http://dx.doi.org/10.1016/j.ijforecast.2006.03.001. url: http://www.sciencedirect.com/science/article/pii/S0169207006000239.

[31] R. Hyndman ja Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R", *Journal of Statistical Software, Articles*, vol. 27, nro 3, s. 1–22, 2008, ISSN: 1548-7660. DOI: 10.18637/jss.v027.i03. url: https://www.jstatsoft.org/v027/i03.