Sähkönkulutuksen ennustaminen toimialoittain

Ville Leinonen

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
1. kesäkuuta 2016

Tiivistelmä: Ville Leinonen: *Sähkönkulutuksen ennustaminen toimialoittain*, Tilastotieteen pro gradu -tutkielma, 65 s. + liitteitä 2 s., Jyväskylän yliopisto, Matematiikan ja tilastotieteen laitos, 1. kesäkuuta 2016.

Tämän tutkielman tavoitteena on kehittää toimialoittaista sähkönkulutusta ennustava tilastollinen malli Jyväskylän Energia -konsernille. Mallin tarkoituksena on ennustaa toimialoittaista sähkönkulutusta kuukausitasolla käyttäen apuna sähkönkulutusta selittäviä muuttujia.

Tarkasteltavat toimialakohtaiaet sähkönkulutukset koostuvat Jyväskylän Energia -konsernin suurten jyväskyläläisten asiakkaiden sähkönkulutusmittauksista. Aineistoon ei kuulu kuluttaja-asiakkaita. Kuukausittaiset sähkönkulutukset ryhmitellään toimialoittain Tilastokeskuksen asiakkaille antamien toimialaluokitusten mukaan. Saman toimialan asiakkaiden kulutuksista muodostetaan toimialan kulutus. Lisäksi sähkönkulutukselle on etsitty mahdollisia selittäviä muuttujia avoimista lähteistä. Selittäviksi muuttujiksi on valittu kansantalouden ja toimialan kehitykseen, säätilaan ja arki- ja vapaapäiviin liittyviä muuttujia.

Malli toteutetaan käyttämällä aikasarjaregressiota. Aikasarjaregressiossa vasteelle käytetään prediktoreina toisia aikasarjoja ja mallin jäännösten oletetaan noudattavan stokastista aikasarjaprosessia. Tarkoituksena on löytää jokaiselle toimialalle prediktorit, jotka parantavat toimialan sähkönkulutuksen ennustettavuutta. Eri prediktorien vaikutusta sähkönkulutuksen ennustettavuuteen tarkastellaan keskimääräisen ennustevirheen perusteella uudessa aineistossa. Mallin valintaa tarkastellaan ennustevirheen ja prediktorien määrän suhteen.

Aikasarjaregressiomallia verrataan multiplikatiiviseen ARIMA-malliin, jossa ei käytetä prediktoreita. Osalla toimialoista aikasarjaregressiomallilla saadaan tarkemmat ennusteet testiaineistossa ja osalla toimialoista aikasarjaregressiomalli ja multiplikatiivinen ARIMA-malli tuottavat likimäärin yhtä tarkat ennusteet.

Avainsanat: aikasarja-analyysi, avoin data, sähkönkulutus, toimialaluokitus

Taulukko 1: Käytetyt merkinnät

Koko tekstin merkinnät

Merkintä	Selitys
$N(\mu, \sigma^2)$	normaalijakauma, joka odotusarvo on μ ja varianssi σ^2
T	aikasarjan pituus
t	aikapiste, $t \in 1, \ldots, T$
τ	viive
y_t	vasteen, sähkönkulutuksen aikasarja, $t = 1, \ldots, T$
y'_t	stationaariseksi perättäisillä erotuksilla ja kausierotuksilla muutettu vaste
x_t	prediktorin aikasarja, $t = 1, \ldots, T$
x'_t	stationaariseksi perättäisillä erotuksilla ja kausierotuksilla muutettu prediktori
z_t	puhdas satunnaisprosessi, $z_t \sim$ i.i.d $N(0, \sigma^2)$ kaikilla t
η_t	ARIMA-aikasarjaprosessi
η'_t	ARMA-aikasarjaprosessi
w	kompleksiluku tai ARMA-aikasarjaprosessi
β	regressiokerroin
s	kauden pituus, sähkönkulutuksen tapauksessa $s = 12$
B	backshift-operaattori, $Bw_t = w_{t-1}$
B^s	backshift-operaattori sovellettuna s kertaa, $B^s w_t = w_{t-s}$
∇	perättäinen erotus, $\nabla = 1 - B$
∇_s	kausierotus, $\nabla_s = 1 - B^s$
p, P	ARMA- ja ARIMA mallien autoregressiivisten kerrointen lukumäärät
q, Q	ARMA- ja ARIMA mallien liukuvan keskiarvon kerrointen lukumäärät
d, D	ARIMA-mallien tarvittavien perättäisten ja kausierotusten lukumäärät
ϕ, Φ	autoregressiiviset kertoimet
θ, Θ	liukuvan keskiarvon kertoimet
i, j, r	indeksimuuttujat
k	prediktorien lukumäärä tai kokonaisluku
n	kokonaisluku
$\mathbb{R}$	Reaalilukujen joukko
$\mathbb{C}$	Kompleksilukujen joukko
$\mathbb{Z}_+$	Positiivisten kokonaislukujen joukko
$\lvert x \rvert$	luvun x itseisarvo, kun $x \in \mathbb{R}$. Kun $x = a + ib \in \mathbb{C}, \lvert x \rvert = \sqrt{a^2 + b^2}$.
	Lihavoitu teksti tarkoittaa joko matriisia tai vektoria
$\boldsymbol{v}'$	vektorin $\boldsymbol{v}$ transpoosi
$\boldsymbol{M}^{-1}$	matriisin $\boldsymbol{M}$ käänteismatriisi

Taulukko 2: Käytetyt merkinnät, osa 2

Aineisto-kappaleeseen liittyvät merkinnät

Merkintä	Selitys
kWh	kilowattitunti
MWh	megawattitunti
A	ampeeri
m_t	aikasarjan trendi
s_t	aikasarjan kausivaihtelu
e_t	aikasarjan ylijäävä osa, kun trendi ja kausivaihtelu poistetaan
$\tilde{s}$	keskimääräinen kausivaihtelu aikavälillä
$\hat{s}_t$	aikavälin skaalattu kausivaihtelu $\hat{s}_t = s_t/\tilde{s}$
d_t	kausivaihtelun estimaatti

Aikasarja-analyysi ja tulokset -kappaleisiin liittyvät merkinnät

Merkintä	Selitys
μ	odotusarvo
σ^2	varianssi
$\gamma(\tau)$	autokovarianssi viiveellä τ
$\boldsymbol{\gamma_\tau}$	autokovarianssivektori, joka sisältää viiveet $1, \ldots, \tau$
$\boldsymbol{\Gamma_\tau}$	autokovarianssimatriisi, joka sisältää viiveet $0, \ldots, \tau - 1$
$\rho(\tau)$	autokorrelaatio viivellä τ
$\alpha(\tau)$	osittaisautokorrelaatio viiveellä τ
π_j	kääntyvän aikasarjaprosessin kertoimet
ψ_j	lineaarisen tai kausaalisen aikasarjaprosessin kertoimet
$\phi^*(z)$	polynomista $\phi(z)$ muutettu polynomi
λ	Box-Cox -muunnoksen parametri
$P_T y_{T+l}$	l askeleen ennuste aikasarjalle y_t käyttäen aikasarjan havaintoja aikahetkeen T asti
r_t	ARMA-prosessin suurimman uskottavuuden estimaatin lausekkeen termi
ν_t	ennusteen keskineliövirhe
θ_{ij}	innovaatioalgoritmin kertoimet
a_j	polynomin j. asteen kerroin
χ_r	keskineliövirheen laskemiseen käytetyt vakiot
$\hat{Q}$	Ljung-Box-Piercen testisuureen arvo
K	Ljung-Box-Piercen testisuureen laskemiseen käytettävien autokorrelaatioiden määrä
df	vapausasteiden lukumäärä
h	ennusteperiodin pituus

Sisältö

1 Johdanto **1**

2 Aineisto **2**

2.1 Sähkönkulutusaineisto 2

2.2 Kulutusten muuttaminen kuukausittaisiksi 6

2.3 Prediktorien aineisto 11

3 Aikasarja-analyysi **15**

3.1 Aikasarja 15

3.2 Stationaarisuus 15

3.3 Aikasarjoihin liittyviä suureita ja operaattoreita 20

3.4 Aikasarjaprosesseja 21

3.5 Aikasarjaregressio ARMA-jäännöksillä 23

3.6 Aikasarjaregression suurimman uskottavuuden estimointi 25

3.7 Ennustaminen 26

3.8 Mallin valinta ja hyvyys 29

4 Mallintaminen **31**

4.1 Mallinnuksen toteutus 31

4.2 Kvartaaliset ja vuosittaiset prediktorit ja perättäiset erotukset 32

4.3 Mallin valinta 32

4.4 Ennustaminen 37

5 Tulokset **38**

6 Pohdinta **60**

Lähdeluettelo **62**

A Prediktorien Internet-osoitteet **65**

1 Johdanto

Tutkielman tavoitteena on ennustaa Jyväskylän toimialoittaista sähkönkäyttöä. Aineistona toimii Jyväskylän Energia -konsernin asiakkaiden osa-aineisto, jossa sähkönkuluttajat on jaettu toimialoihin. Tavoitteena on malli, jota Jyväskylän Energia voi jatkossa käyttää ennustamaan toimialoittaisia sähkönkulutuksia Jyväskylän alueella. Kiinnostuksen kohteena on löytää sähkönkulutukselle prediktoreita, jotka auttaisivat sähkönkulutuksen ennustamisessa. Työn tarkoituksena on etsiä menetelmä, jolla selittävien muuttujien vaikutusta sähkönkäyttöön voidaan mallintaa. Mallin jatkokehitys tapahtuu Jyväskylän Energian ja Jyväskylän yliopiston Informaatioteknologian tiedekunnan teollisen optimoinnin tutkimusryhmän yhteistyöllä.

Prediktoreiksi on ehdolla yleiseen ja toimialan taloudelliseen kehitykseen, säätilaan ja työ- ja arkipäiviin liittyviä muuttujia. Prediktorit etsitään avoimista lähteistä. Prediktorien joukkoa ei rajata ennalta prediktoreihin, joiden tiedetään vaikuttavan sähkönkäyttöön, kuten esimerkiksi lämpötila. Sen sijaan yritetään löytää mukaan paljon erilaisia prediktoreita, joista pyritään löytämään eniten ennustetarkkuutta parantavat prediktorit. Kiinnostavaa on myös, ovatko prediktorien vaikutukset erilaiset eri toimialoilla.

Sähkökulutusta ennustetaan pitkällä aikavälillä, joka tarkoittaa tässä tapauksessa 1-3 vuotta. Hong ja Shu (2016) esittävät useita eri vaihtoehtoja sähkönkulutuksen ennustamiseen pitkälle aikavälille. Tässä tutkielmassa sähkönkulutusta ennustetaan sitä selittävien aikasarjojen avulla. Pitkän aikavälin sähkönkulutusta prediktorien avulla ovat ennustaneet esimerkiksi Hong et al. (2014).

Toimialoittaisten sähkönkulutusten riippuvuutta prediktoreista selvitetään aikasarjaregression avulla. Aikasarjaregressio poikkeaa tavanomaisesta lineaarisesta regressiosta siten, että aikasarjaregressiossa mallin jäännösten oletetaan riippumattomuuden sijasta noudattavan jotain aikasarjaprosessia.

Luvussa 2 käsitellään tarkemmin tutkielmassa käytettyä aineistoa ja esitetään tarvittavat aineiston muokkaukset. Luvussa 3 esitetään ARMA- ja ARIMA-mallien yleistä teoriaa ja käytetty aikasarjaregressio. Luvussa 4 pureudutaan tarkemmin sopivien prediktorien valintaan ja laskentaan R-ympäristössä (R Core Team, 2015). Luvussa 5 esitetään mallinvalintaprosessi eräiden toimialojen tapauksessa ja tarkastellaan käytettyjen mallien ennustekykyä testiaineistossa. Luvussa 6 on yhteenveto ja pohdintaa työstä.

2 Aineisto

Aineistoon kuuluu Jyväskylän Energialta saatu sähkönkulutusaineisto ja prediktorien aineisto. Tässä luvussa esitellään aineistojen rakenne ja käytetyt prediktorit. Lisäksi kerrotaan myös sähkönkulutusaineistoon tehdyistä muokkauksista, jotta aineisto on saatu haluttuun muotoon.

2.1 Sähkönkulutusaineisto

Sähkönkulutusaineisto on osa-aineisto Jyväskylän Energia -konsernin asiakkaiden sähkönkulutusmittauksista. Sähkönkulutusmittaus koostuu mitatusta kulutuksesta ja aikavälistä, jonka aikana mitattu määrä sähköä on kulutettu.

Sähkönkulutusta on mitattu aineistossa sähkönkäyttöpaikoittain eri aikajaksoina. Sähkönkäyttöpaikalla tarkoitetaan paikkaa, jossa sähkö kulutetaan ja joka on yhden sähkömittarin mitattavana. Käytännössä sähkönkäyttöpaikka voi olla esimerkiksi rakennus tai sen osa, kuten esimerkiksi yrityksen toimitila rakennuksen sisällä. Myös omakotitalo ja rivi- tai kerrostalohuoneisto voi muodostaa sähkönkäyttöpaikan, joskin rivi- tai kerrostalo voi kokonaisuudessaankin olla yksi sähkönkäyttöpaikka.

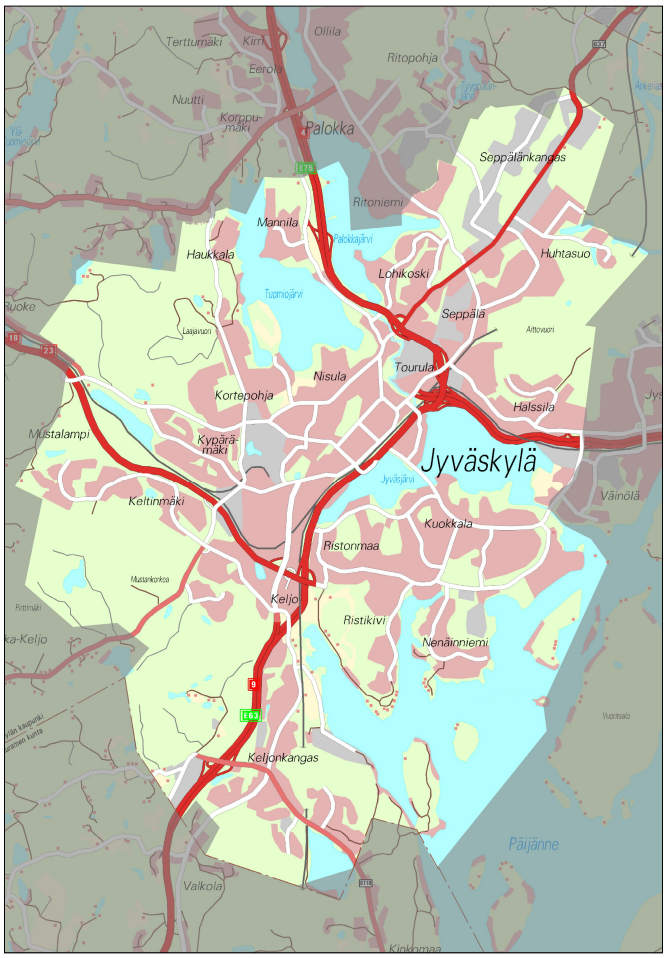
Sähkönkäyttöpaikkaan liittyy asiakas, jolla on sopimus sähkönkäyttöpaikassa kulutettavasta sähköstä. Tämän tutkielman aineistossa asiakas on muu kuin kuluttaja-asiakas. Käyttöpaikan kulutus ei kuitenkaan aina ole suoraan asiakkaan koko kulutus, sillä asiakkaalla voi olla useita toimipisteitä, joissa jokaisessa on oma sähkönkäyttöpaikka. Joillakin asiakkailla myös yhteen toimipisteeseen liittyi useita käyttöpaikkoja. Tässä tapauksessa toimipisteen käyttöpaikoista on tehty aineistoon yksi käyttöpaikka, jonka kulutus on alkuperäisten käyttöpaikkojen kulutusten summa.

Sähkönkäyttöpaikan liittyminen tiettyyn rakennukseen tai sen osaan tarkoittaa sitä, ettei kaikki sähkönkäyttöpaikkaan liittyvä kulutus ole välttämättä saman asiakkaan kulutusta. Sähkönkäyttöpaikan asiakas voi vaihtua, kun asiakkaat muuttavat toimitilasta toiseen. Tällaisten käyttöpaikkojen tapauksissa käyttöpaikan kulutus on jaettu asiakkaittain erillisiin aikajaksoihin sähkönkäyttöpaikkaan liittyvien sopimusten perusteella. Jatkossa, ellei toisin mainita, sähkönkäyttöpaikalla ja sähkönkäyttöpaikan kulutuksella tarkoitetaan yksittäisen asiakkaan yhtä toimipistettä ja asiakkaan siellä kuluttamaa sähköä.

Aineistossa käyttöpaikkojen kulutustiedot ovat saatavilla joulukuusta 2004 marraskuuhun 2015 siltä ajalta kun kyseinen käyttöpaikka on ollut toiminnassa. Käyttöpaikat sijaitsevat Jyväskylän Energian sähkön jakelualueella, joka on esitetty kuvassa 1.

Aineistoa on rajattu vielä ottamalla mukaan vain käyttöpaikat, joiden sähkönkulutus on ollut jonakin vuonna yli 100 MWh. Saadusta aineistosta on poistettu epäoleellisia käyttöpaikkoja, joiden kulutuksista ei olla kiinnostuneita tai joiden kulutuksien ei uskota olevan oleellisia tämän tutkimuksen kannalta.

Jatkossa sähkönkäyttöpaikkojen kulutuksista muodostetaan toimialakohtaiset säh-



Kuva 1: Jyväskylän Energian sähkön jakelualue on kuvassa vaalealla. Kuva: Jyväskylän Energia

könkulutukset. Toimialakohtaisten kulutusten muodostamiseksi tarvitaan kaikki toimialan käyttöpaikat ja niiden kulutukset halutulta aikaväliltä. Toiminnassa olevien käyttöpaikkojen haku onnistui helposti järjestelmästä viimeisen vuoden kulutuksen perusteella. Aineistoon on etsitty toiminnassa olevien käyttöpaikkojen lisäksi käyttöpaikkoja, joiden toiminta on loppunut. Näissä käyttöpaikoissa ei ole enää tällä hetkellä asiakasta. Käyttöpaikan toiminta voi loppua esimerkiksi kiinteistön purkamisen seurauksena. Myös käyttöpaikan jakaminen useampaan käyttöpaikkaan, tai useamman käyttöpaikan yhdistäminen yhdeksi käyttöpaikaksi voi aiheuttaa käyttöpaikan loppumisen. Loppuneista käyttöpaikoista saatiin haettua lista, jossa olivat mukana käyttöpaikat, joissa oli toiminut aiemmin halutun tyyppinen asiakas.

Toiminnassa olevat käyttöpaikat on poimittu järjestelmästä viimeisen vuoden kokonaiskulutuksen perusteella ottamalla mukaan käyttöpaikat, joiden kokonaiskulutus oli yli 100 MWh. Jo poistuneita käyttöpaikkoja ei onnistuttu rajaamaan kulutuksen perusteella. Poistuneista käyttöpaikoista saatiin haettua lista, johon saatiin käyttöpaikasta tietoina mm. käyttöpaikassa viimeisenä käytetyn sähkömittarin sulakekoko.

Käyttöpaikkojen läpikäyminen tehtiin käsin. Yhteensä käyttöpaikkoja, joissa oli päättynyt muun kuin kuluttaja-asiakkaan sopimus, oli yhteensä noin 16 000 kappaletta. Käyttöpaikka ja asiakas otettiin mukaan aineistoon, kun asiakkaan vuosikulutus oli jonkin vuoden aikana yli 100 MWh. Loppuneiden käyttöpaikkojen kulutuksille saatiin suuntaa antava kulutusarvio käyttöpaikan mittarin viimeisimmän sulakekoon perusteella. Käyttöpaikkojen suuren määrän takia kaikkien käyttöpaikkojen läpikäynti käsin olisi vienyt liikaa aikaa, joten läpikäyntiä rajattiin käyttöpaikkoihin, joissa viimeisin käyttöpaikassa ollut sulakekoko oli vähintään 3×40 A tai sulakekoko puuttui. Alkuperäiseen aineistoon oli otettu nykyiset käyttöpaikat, joiden sulakekoko on 3×63 A. Rajaamiseen käytettiin kuitenkin pienempää sulakekokoa, sillä käyttöpaikan sulakekoko voi vaihtua, mikäli käyttöpaikan nykyinen sulakekoko ei ole sopiva käyttäjän tarpeisiin.

Lisäksi käytiin läpi paikkoja, joissa nykyinen kulutus oli alle 100 MWh, ja joissa sopimus ei ole vaihtunut. Koska kulutus vaihtelee vuosittain, voi olla, että asiakas on aiemmin kuluttanut yli 100 MWh, mutta kulutus on pienentynyt vuodelle 2015, johon järjestelmän arvio vuosikulutuksesta perustuu.

Yhteensä käyttöpaikkoja on käyty käsin läpi noin 2000. Näistä käyttöpaikoista on etsitty asiakkaat, joiden käyttö oli jonain vuonna yli 100 MWh. Näiden asiakkaiden kulutukset on otettu mukaan aineistoon koko sopimuksensa ajalta. Yhteensä löydettyjä käyttöpaikkoja, joissa kulutus on yli 100 MWh, on 176. Kun tähän lisätään vielä toiminnassa olevat, viime vuonna yli 100 MWh kuluttaneet käyttöpaikat, aineistoon kuuluvien käyttöpaikkojen kokonaismääräksi tuli 684.

Koska kaikkia käyttöpaikkoja ei ole käyty läpi, on mahdollista, että jokin aiemmin yli 100 MWh kuluttanut asiakas ja käyttöpaikka on jäänyt pois aineistosta. Esimerkiksi käyttöpaikkojen, joiden nykyinen sulakekoko on alle 3×40 A, joukossa voi olla käyttöpaikkoja, joissa on ollut aiemmin suurempi sulakekoko ja sähkönkulutus. Kuitenkin käyttöpaikkojen suuri määrä pakotti käymään läpi vain pienen osan käyttöpaikoista. Lisäksi voitaneen olettaa, että sulakekoon muutoksen todennäköisyys sulakekoosta 3×63 A pienempään kuin 3×40 A on varsin vähäinen. Näin ollen voitaneen olettaa, ettei mahdollisesti aineistosta jääneillä käyttöpaikoilla ole merkittävää vaikutusta toimialan sähkönkulutuksen aikasarjaan.

Sähkönkulutuksen mittaaminen on muuttunut 2000-luvulla täysin, kun etäluettavat sähkömittarit on otettu käyttöön. Nykyiset sähkömittarit mittaavat käyttöpaikan sähkönkulutuksen tunneittain. Täten olisi mahdollista ottaa aineistoon tunneittaiset aikasarjat käyttöpaikkojen kulutuksista etäluettavien mittareiden ajalta. Kuitenkaan tässä tapauksessa, kun kiinnostuksen kohteena on pidemmän aikavälin sähkönkulutuksen riippuvuus prediktoreista, ei tunneittaisella mittaustarkkuudella saada lisähyötyä, mikäli prediktoreista ei ole saatavilla tunneittaista tietoa. Tunneittainen aikasarja pystytään kuitenkin muuttamaan helposti mille tahansa epätarkemmalle aikavälille summaamalla tunneittaiset kulutukset kyseisellä aikavälillä yhteen. Kuukausittainen tarkkuus on valittu tarkasteltavaksi, koska käytetyistä prediktoreista saatavat aikasarjat ovat tarkimmillaan kuukausitasolla. Kuukausittaisesta tarkkuudesta on hyötyä verrattuna kvartaaliseen tai vuosittaiseen tarkkuuteen, sillä prediktorien mahdollinen

vaikutus kulutukseen voidaan havaita paremmin kuukausitasolla, kun havaintoja on enemmän.

Suurin osa kulutuksista oli aineistossa valmiiksi kuukausittain. Kuitenkin aineistossa oli myös kulutuksia, jotka olivat mitattu kuukaudesta poikkeavalla aikavälillä. Osa käyttöpaikkojen kuukausikulutuksista oli jakautunut kahdesta neljään osaan, jotka muodostivat koko kuukauden. Tällöin kuukauden kulutus oli osiensa summa ja kulutus pystyttiin muuntamaan helposti kuukausitasolle. Lisäksi oli mittauksia, joissa osa kuukaudesta tai koko kuukausi on jäänyt jostain syystä aineistosta pois. Tällaiset tilanteet selvitettiin, ja lopulta suurin osa puuttuvista kulutuksista löytyikin järjestelmistä etsinnän jälkeen. Osalle vajaista kuukausista ei kuitenkaan ole löydetty puuttuvan ajan kulutusta. Tapauksissa, joissa tiedetään, että sähkönkulutusta on oikeasti ollut, on kulutusta korjattu kertomalla havaittu kulutus kuukauden päivien määrällä ja jakamalla tulo havaitun aikavälin päivien määrällä. Tällaisia tapauksia aineistossa on yhteensä alle kymmenessä kuukausittaisessa havainnossa, joten korjaavan menetelmän valinnalla ei liene suurta merkitystä.

Sähkönkäyttöpaikoissa on alun perin myös kulutuksia, jotka on mitattu harvemmin kuin kuukausittain. Nämä kulutukset ovat ajalta, jolloin kyseisessä käyttöpaikassa ei ole ollut etäluettavaa sähkömittaria. Valtioneuvoston asetuksessa maaliskuussa 2009 määrättiin, että vuoden 2013 loppuun mennessä 80 %:n verkkoyhtiön asiakkaista oli oltava etäluennan piirissä (Valtioneuvosto, 2009). Tässä aineistossa on yhteensä 134 käyttöpaikkaa, joissa kulutus ei ole ollut kuukausittaista koko ajalta. Suurilla sähkönkäyttäjillä etäluettavat mittarit ovat olleet käytössä jo paljon ennen valtioneuvoston asetusta. Aineistossa käyttöpaikat, joissa kulutusta on mitattu harvemmin kuin kuukausittain, ovat pieniä sähkönkuluttajia, kun verrataan kulutuksia aineiston käyttöpaikkojen kesken.

Etäluettaviin mittareihin siirtymistä tapahtuu aineistossa vuodesta 2005 elokuuhun 2013 asti, suurimman osan mittarinvaihdoista asettuessa vuosille 2010 ja 2011. Ennen etäluettavaa mittausta olevat lukemat on saatu kalenterikuukaudesta poikkeavina aikajaksoina. Aikajaksojen pituus vaihtelee muutamasta päivästä useampaan vuoteen, mediaanin ollessa 356 päivää. Puolet aikajaksojen pituuksista on välillä 247–372 päivää. Yhteen käyttöpaikkaan liittyy yhdestä 13:een aikajaksoa, joissa kuukaudesta poikkeava kulutus on mitattu. Nämä kulutukset on muutettu kuukausittaisiksi käyttäen apuna käyttöpaikan kuukausittaisista kulutuksista estimoitua kausivaihtelua. Menetelmä ja kulutusten muuttaminen esitellään tarkemmin luvussa 2.2.

Koska ollaan kiinnostuneita toimialan vaikutuksista käyttöpaikan kulutukseen, käyttöpaikoille on etsitty toimialaluokitukset. Toimialaluokituksena käytetään tilastollista toimialaluokitusta TOL 2008 (Tilastokeskus, 2008), joka perustuu Euroopan Unionin toimialaluokitukseen NACE Rev. 2. Toimialaluokitus luokittelee asiakkaat niiden päätoimialan mukaan. Asiakkaan päätoimiala on se toimiala, joka tuottaa suurimman arvonlisäyksen. Arvonlisäyksellä tarkoitetaan hyödykkeen hintaan tullutta lisäystä, jonka asiakas on tehnyt. Mikäli asiakkaan toiminta ulottuu monelle toimialalle, näistä yksi valitaan asiakkaan toimialaluokitukseksi määriteltyjen valintakriteereiden mukaan (Tilastokeskus, 2008). Toimialaluokitukset on etsitty Internetistä asiakastieto.fi

-sivustolta (Suomen Asiakastieto Oy, 2016).

Yksittäisten asiakkaiden käyttöpaikan kulutuksiin liittyy haasteita. Käyttöpaikan kulutus saattaa loppua tai muuttua hetkellisesti, jos käyttöpaikassa ei ole toimintaa hetkellisesti esimerkiksi toimipisteeseen liittyvien muutostöiden takia. Lisäksi käyttöpaikkojen kulutuksen taso voi muuttua pysyvästi alas- tai ylöspäin, jos asiakas pienentää tai suurentaa toimintaansa. Myös käyttöpaikkojen välillä tapahtuvat muutokset tuovat haasteita yksittäisten käyttöpaikkojen kulutusten ennustamiseen. Osa käyttöpaikkojen kulutuksista on saatavilla lyhyeltä ajalta, jolloin niiden ennustaminen luotettavasti on haastavaa.

Koska varsinainen mielenkiinto on eri toimialojen kulutusten mallintamisessa ja koska yksittäisten kulutusten luotettava ennustaminen on osassa tapauksista haastavaa, on käyttöpaikkojen kulutuksista muodostettu toimialakohtaiset kulutukset. Käyttöpaikoille etsittyjen toimialaluokitusten perusteella käyttöpaikat on ryhmitelty toimialaryhmiin, ja saman toimialaryhmän käyttöpaikkojen kulutukset on summattu toimialakohtaiseksi kulutukseksi. Muodostetuissa toimialoittaisissa kulutuksissa on käytetty vähimmillään seitsemää käyttöpaikkaa, mediaanin ollessa 23 käyttöpaikkaa. Yhteensä toimialoja on 17.

Käyttäessä toimialakohtaisia kulutuksia yksittäisten käyttöpaikkojen kulutusten sijaan halutut toimialat joudutaan kiinnittämään toimialakohtaisia kulutuksia muodostettaessa. Mikäli ennustetaan yksittäisten käyttöpaikkojen kulutusta, voidaan käyttöpaikkojen kulutuksia ryhmitellä mielivaltaisesti, jolloin myös erilaisia toimialaryhmittelyjä voidaan käyttää. Kuitenkin osan käyttöpaikkojen käyttäjien kulutuksen aikasarjojen lyhyys ei mahdollista näiden kulutusten mallintamista hyvin. Käyttäessä toimialakohtaista mallinnusta, voidaan nämä kulutukset summata osaksi toimialan kulutusta, jolloin ne ovat mukana aineistossa ja niiden lyhyys ei haittaa.

2.2 Kulutusten muuttaminen kuukausittaisiksi

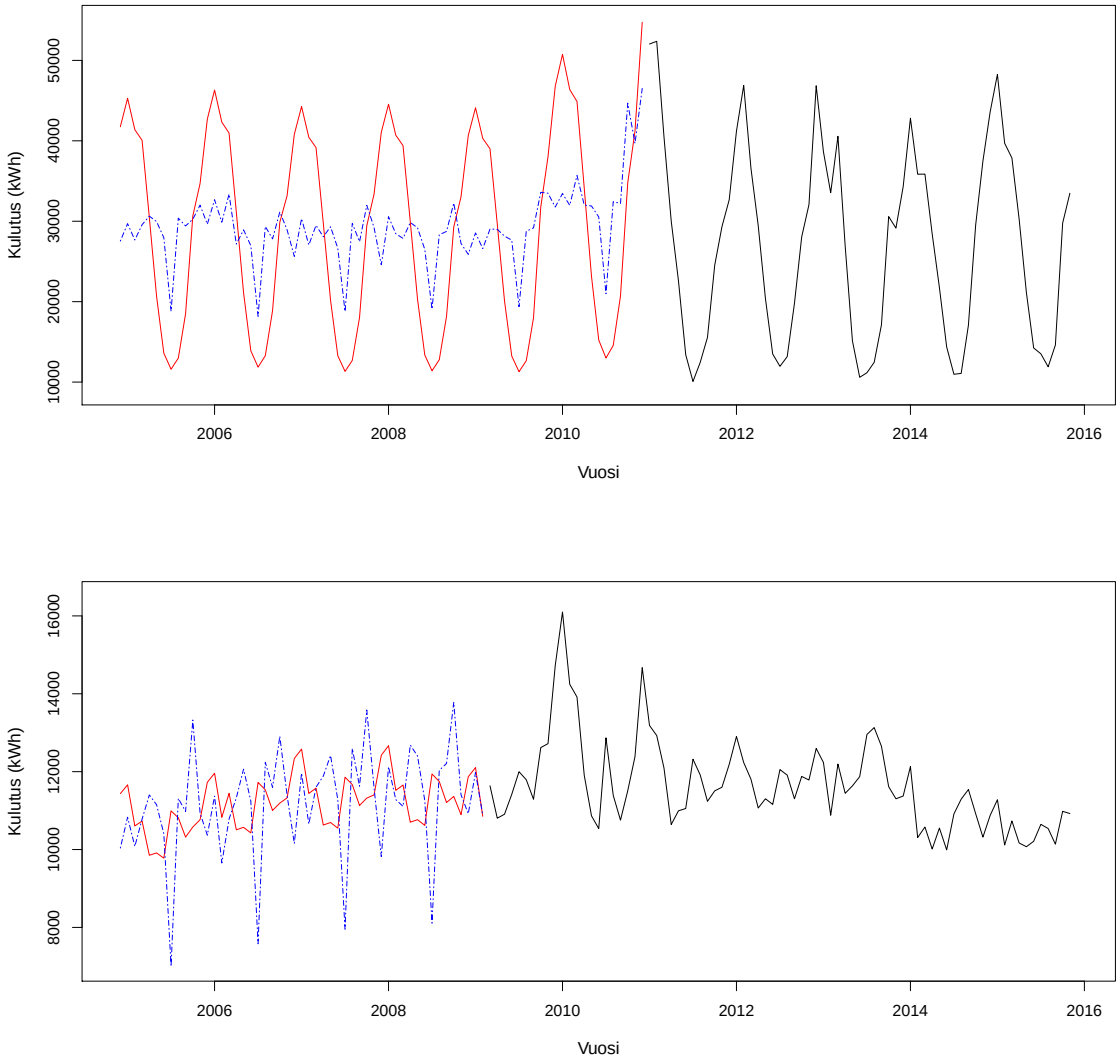
Jotta koko toimialan kulutus saadaan esitettyä kuukausitasolla, on kaikkien toimialan käyttöpaikkojen kulutus oltava kuukausitasolla. Koska osa käyttöpaikoista sisältää myös ei-kuukausittaista kulutusta, on ei-kuukausittaiset kulutukset pystyttävä muuttamaan kuukausittaisiksi kulutuksiksi. Käytetään menetelmää, joka estimoi kuukausittaiset kulutukset havaituista, ei-kuukausittaisista kulutuksista käyttöpaikan havaitun kuukausittaisen kulutuksen perusteella.

Lähtökohtana menetelmälle on, että aikavälillä havaittu kulutus tulee jakaa aikavälin eri kuukausille. Kuukausien yhteiskulutuksen tulee olla havaittu kulutus. Aikaisemmassa aineiston tarkastelussa on huomattu, että käyttöpaikkojen sähkönkäyttöön liittyy kausittaista vaihtelua, joka täytyy ottaa huomioon kulutuksen jakamisessa eri kuukausille.

Ennen etäluettavia mittareita kulutuksen muuntamiseen kuukausi- tai tuntitasolle on käytetty valtakunnallisia, yleisiä tyyppikuormituskäyriä. Tyyppikuormituskäyrien avulla aikavälin kokonaiskulutus on jaettu eri kuukausille ja näin saatu arvio

käyttöpaikan kulutuksesta eri kuukausina.

Tyyppikuormituskäyrä on ollut yhteinen isolle joukolle käyttöpaikkoja, joten sen soveltuvuus yksittäisen käyttöpaikan kulutuksen muuntamiseen kuukausitasolle on kyseenalaista. Koska samaa kuormituskäyrää on käytetty usealla toimialalla, ei käyrä pysty ottamaan huomioon yksittäisen toimialan käyttöpaikkojen tyypillistä kausivaihtelua, eikä varsinkaan yksittäisen käyttöpaikan kulutukseen liittyvää kausivaihtelua kovinkaan hyvin. Kuvassa 2 on esitetty tyyppikuormituskäyrän perusteella jaetut kuukausittaiset kulutukset kahdella ei käyttöpaikalla. Vertailuna on esitetty myös käytetyn menetelmän perusteella jaetut kuukausittaiset kulutukset.



Kuva 2: Esimerkit tyyppikuormituskäyrän (sininen katkoviiva) toimivuudesta kahdella käyttöpaikalla. Vertailuna käytetyn menetelmän antamat kuukausittaiset kulutukset (punainen yhtenäinen viiva).

Lisäksi Jyväskylän Energialla on arvoitu kuukausikulutuksia myös sähkön kulutustuotteisiin perustuvilla kuormituskäyrillä. Kulutustuotteet perustuvat sulakekokoihin ja käytetyn sähkön määrään. Näin ollen kulutustuotteisiinkaan perustuvassa menetelmässä ei oteta huomioon toimialojen välistä vaihtelua kulutuksessa.

Tässä tapauksessa käytetään muuntamiseen apuna käyttöpaikan historiatietoa etäluettavan mittauksen ajalta. Ajalta jolloin etäluettava mittaus on ollut käytössä, saadaan useimmissa tapauksissa arvioitua käyttöpaikan kausittainen vaihtelu kulutuksissa. Tämän havaitun kausivaihtelun oletetaan olleen samanlaista ennen kuukausimittauksen aloittamista. Käyttämällä havaittua kausivaihtelua, saadaan ei-kuukausittaiset kulutukset jaettua kuukausille.

Mikäli käyttöpaikalta ei ole saatavilla kuukausittaisia kulutusmittauksia, tai kuukausittainen aikasarja on liian lyhyt kausivaihtelun estimoimiseksi, etsitään käyttöpaikkojen joukosta sähkönkulutukseltaan mahdollisimman samankokoinen, mahdollisimman saman toimialan käyttöpaikka, jolta kausivaihtelu voidaan estimoida. Tätä estimoitua kausivaihtelua käytetään käyttöpaikan kausivaihtelun estimaattina.

Seuraavana esitetään käytetty menetelmä kulutusten muuttamiseksi kuukausittaisiksi. Menetelmässä oletetaan multiplikatiivinen malli kulutukselle, eli sähkönkulutus hetkellä t voidaan esittää muodossa $y_t = m_t s_t e_t$, missä m_t on estimoitu trendi, s_t estimoitu kausivaihtelu ja e_t jäännöstermi kuukautena t, $e_t = y_t/(m_t s_t)$.

Käytännössä kuukausittaisten kulutusten laskemien tehtiin seuraavalla tavalla.

1. Estimoidaan kausikomponentti s_t havaitusta kuukausisarjasta.

2. Lasketaan kaikille kuukausille suhteellinen osuus kuukaudesta, joka osuu mitattuun kulutukseen liittyvälle aikavälille. Merkitään osuutta muuttujalla p_t.

3. Lasketaan kulutusten määrät eri kuukausille, mikäli kulutus olisi tasaista koko aikavälillä.

4. Lasketaan kertoimille s_t kuukausien suhteellisten osuuksien p_t suhteen painotettu keskiarvo $\tilde{s} = (\sum_t p_t s_t)/(\sum_t p_t)$, jonka avulla lasketaan skaalatut kausivaihtelun kertoimet $\hat{s}_t = s_t/\tilde{s}$.

5. Lasketaan kausivaihtelun sisältämä kuukausittainen kulutus kertomalla kohdassa 3. saatu kulutus kuukauteen liittyvällä kertoimella $\hat{s}_t$.

Kohdat 4. ja 5. on tehty, jotta kuukausien kulutusten summa vastaisi aikavälin havaittua arvoa. Mikäli kulutus ei ole mitattu tasan vuoden mittaiselta ajalta, ei kerrointen keskiarvo $\tilde{s_t}$ ole välttämättä tasan 1. Täten ei voida suoraan kertoa keskimääräistä kulutusta kausivaihtelulla s_t, sillä keskiarvon poiketessa luvusta 1, ei myöskään välin havaittu kulutus ole sama kuin estimoitujen kulutusten summa. Kohta 4. korjaakin kausivaihtelukertoimia s_t niin, että niiden painotettu keskiarvo on 1.

Taulukko 3: Esimerkki ei-kuukausittaisen kulutuksen muuttamisesta kuukausikulutukseksi

Kuukausi	s_t	p_t	tasainen kulutus	$\hat{s}_t$	estimoitu kulutus
6/2010	1.1	0.5	10000	0.9649	9649
7/2010	1.3	1	20000	1.1403	22807
8/2010	1	1	20000	0.8772	17544

Aikasarjojen kausikomponentit s_t estimoitiin R-ympäristön paketin `stats` funktiolla `decompose`. Funktio estimoi ensin aikasarjan trendin liukuvalla keskiarvolla

$$\hat{m}_t = \frac{\frac{1}{2}y_{t-6} + y_{t-5} + y_{t-4} + \cdots + y_{t+4} + y_{t+5} + \frac{1}{2}y_{t+6}}{12}.$$

Trendin estimaatti poistetaan aikasarjasta, jolloin saadaan trendipuhdistettu aikasarja $y_t/\hat{m}_t$. Tämän jälkeen saadaan kausivaihtelun estimaatit d_t trendipuhdistetun aikasarjan kuukausittaisina keskiarvoina. Estimaatit skaalataan vielä keskimäärin lukuun 1 jakamalla ne kausivaihtelun estimaattien keskiarvolla, $s_t = d_t/(\sum_{i=1}^{12} d_t/12)$.

Käydään läpi yksinkertainen esimerkki tilanteesta, jossa kulutusta ei ole mitattu kuukausittain, ja kulutus on muutettava kuukausittaiseksi kausivaihtelun avulla. Olkoon havaittu kulutus 50000 kWh välillä 16.6.2010 – 31.8.2010. Olkoot lisäksi estimoidut kausivaihtelukomponentit $s_6 = 1.1$, $s_7 = 1.3$ ja $s_8 = 1$. Tällöin kulutuksen estimointi voidaan esittää taulukon 3 avulla.

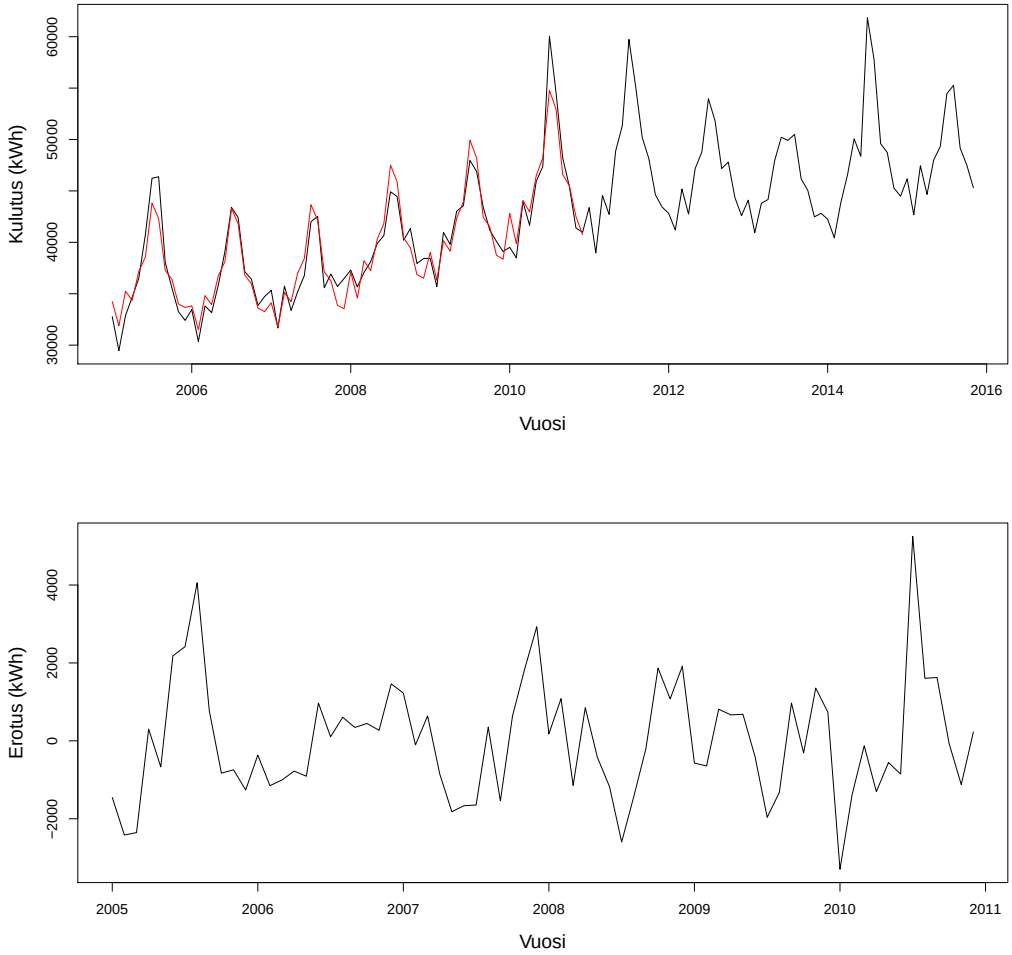
Kuten huomataan, kuukausille jaettu kokonaiskulutus on 50000 kWh niin kuin pitääkin. Koska kesäkuu on mukana vain osittain, on kesäkuuhun liittyvä estimoitu kulutus pienempi kuin heinä- ja elokuun kulutukset, mutta suhteelliset kulutukset ovat oikein.

Mallina olisi voitu käyttää myös muunlaista kausivaihtelun sisältävää mallia. Additiivista mallia $y_t = m_t + s_t + e_t$ kokeiltiin, mutta osassa käyttöpaikoista kulutuksen taso muuttui ajassa niin paljon, että osa estimoiduista kuukausittaisista kulutuksista oli alle nollan, mikä on mahdotonta. Lisäksi multiplikatiivinen ja additiivinen malli eivät eronneet muissa käyttöpaikoissa paljon toisistaan, joten multiplikatiivinen malli valittiin, koska se sopi kaikkiin tapauksiin.

Kuvassa 3 on esitetty mallin toiminta erään käyttöpaikan tapauksessa. Tältä käyttöpaikalta kuukausittainen kulutusaineisto on saatavilla koko aikaväliltä, joten mallin antamia kuukausittaisia arvoja voidaan verrata oikeisiin arvoihin. Tässä tapauksessa pidempänä aikavälinä on käytetty kalenterivuosia 2005, 2006, ..., 2010. Vuosille on laskettu kokonaiskulutukset summaamalla yhteen kuukausittaiset kulutukset kultakin vuodelta. Tämän jälkeen menetelmää on käytetty muuttamaan kulutukset takaisin kuukausittaisiksi käyttämällä käyttöpaikan vuosilta 2011-2015 estimoitua kausivaihtelua kulutuksen jakamiseksi kuukausille.

Ylhäällä olevassa kuvassa on piirretty mustalla käyttöpaikan kulutuksen aikasarja. Punaisella on piirretty vuosille 2005-2010 kuukausikulutuksiksi muutetut kulutukset. Alhaalla olevassa kuvassa on oikean kulutuksen ja mallinnetun kulutuksen erotus. Tehdyn virheen keskihajonta on tässä tapauksessa noin 1500 kWh ja keskimääräinen

virhe suhteessa kuukauden kulutukseen on noin 3 %. Tässä tapauksessa on huomioitava, että kausittainen vaihtelu on melko samankaltaista koko havaintojaksolla, mikä auttaa merkittävästi mallinnuksen onnistumista.



Kuva 3: Yläkuva: Havaittu aikasarja (musta viiva) ja vuosittaisista kulutuksista muutetut kuukausittaiset kulutukset (punainen viiva). Alakuva: Oikean kulutuksen ja muutetun kulutuksen erotus.

Koska kulutusten muuttaminen kuukausittaisiksi perustuu pelkästään käyttöpaikan myöhempiin kuukausittaisiin kulutuksiin tai mahdollisimman samankokoisen käyttöpaikan kuukausittaisiin kulutuksiin ja näiden kulutusten kausittaiseen vaihteluun, on pohdittava, vaikuttaako käytettävä menetelmä toimialakohtaisiin aikasarjoihin ja täten myös mahdollisesti saataviin tuloksiin. Käyttöpaikkojen määrässä ei-kuukausittaista kulutusta sisältävät käyttöpaikat ovat merkittävä osa koko aineistosta, koska noin 25 % käyttöpaikoista sisältää ei-kuukausittaista kulutusta. Kuitenkin suuri osa ei-kuukausittaista kulutusta sisältävistä käyttöpaikoista on pieniä sähkönkuluttajia.

Toimialakohtaisesti nämä käyttöpaikat jakautuvat varsin tasaisesti. Poikkeuksena

yhdellä toimialalla noin puolet käyttöpaikoista sisältää ei-kuukausittaisia kulutuksia ja toisella toimialalla osuus on noin 35 %. Muilla toimialoilla ei-kuukausittaisten käyttöpaikkojen osuus on korkeintaan 25 %. Näin ollen kahta toimialaa lukuun ottamatta ei-kuukausittaisten käyttöpaikkojen osuus toimialan kaikista käyttöpaikoista on siedettävällä tasolla.

Yksi vaihtoehto olisi jättää muuta kuin kuukausittaista kulutusta sisältävät käyttöpaikat kokonaan tarkastelun ulkopuolelle. Kuitenkin tällöin toimialan kulutuksesta saattaisi poistua tarkasteltavaan aikaväliin nähden epätasaisesti kulutusta. Koska ei-kuukausittaiset kulutukset on havaittu tarkasteltavan aikavälin alkupuolella, voisi ei-kuukausittaista kulutusta sisältävien käyttöpaikkojen poistaminen vaikuttaa toimialoittaiseen kulutukseen enemmän aikasarjan alku- kuin loppupäässä, mikäli poistettavista käyttöpaikoista osa loppuu ennen nykyhetkeä. Tällöin toimialan kulutus ei enää vastaisi todellista toimialan kulutusta koko aikavälillä. Täten, vaikkakin ei-kuukausittaisten kulutusten muuttaminen perustuu pelkästään myöhemmän kuukausittaisen kulutuksen kausivaihteluun, tuo sen käyttö paremman arvion toimialakohtaisen aikasarjan kehityksestä kuin käyttöpaikkojen poistaminen.

2.3 Prediktorien aineisto

Toinen osa aineistoa on prediktorien aineisto. Prediktoreiksi on valittu mahdollisesti sähkönkulutukseen vaikuttavia tekijöitä. Prediktoreina käytetään säätilaan, taloudelliseen kehitykseen, toimialaan ja työ- ja vapaapäiviin liittyviä muuttujia.

Sähkönkulutus on samalla tavalla riippuvainen sekä toimijan koosta että markkinaosuudesta kuin minkä tahansa muun kulutustuotteen kulutus; isomman markkinaosuuden toimija tarvitsee myös enemmän resursseja. Myös markkinatilanne toimialalla voi heijastua saatuihin toimeksiantoihin ja tätä kautta toimijan valmistamiin ja valmistamisen resursseina kuluttamiin tuotteisiin. Yleinen markkinatilanne heijastuu toimialojen tilanteeseen ja näin myös yleinen markkinatilanne voi vaikuttaa sähkönkulutukseen. Yleiseen markkinatilanteeseen ja toimialan markkinatilanteeseen liittyviä muuttujia on kerätty, jotta voidaan arvioida niiden vaikutusta sähkönkulutukseen.

Toimialaluokitus auttaa myös prediktorien valitsemisessa. Tilastokeskus käyttää toimialaluokitusta sen tuottamissa tilastojulkaisuissa. Tämän ansiosta Tilastokeskuksen toimialakohtaiset tilastot ovat suoraan käytettävissä prediktoreina. Koska toimialaluokitus täsmää käyttöpaikkojen luokitteluun käytettyyn toimialaluokitukseen, vastaavat toimialakohtaisten prediktorien arvot kyseistä toimialaa tarkasti Suomen tasolla. Toki Suomen tasolla mitatut prediktorit eivät vastaa tarkasti toimialan tilannetta Jyväskylässä, mikäli toimialan tilanne Jyväskylässä poikkeaa toimialan tilanteesta koko Suomessa.

Asiakkaiden toiminta voi riippua viikkorytmistä tai yleisistä loma-ajoista. Täten on luontevaa tutkia viikonloppujen ja muiden vapaiden vaikutusta sähkönkäyttöön. Myös pääsiäinen ja helatorstai liikkuvina pyhinä voivat vaikuttaa sähkönkäyttöön. Viikkorytmi voi olla erilainen eri toimialoilla, joten edellä mainittujen tekijöiden vaiku-

tukset kulutukseen eri toimialoilla voivat olla myös erikokoisia ja erisuuntaisia. Arki- ja vapaapäivistä on tehty erilliset muuttujat, jotka sisältävät arki- ja vapaapäivien lukumäärät eri kuukausina. Arkipäivillä tarkoitetaan tässä päiviä maanantaista perjantaihin, lukuun ottamatta arkipyhiä. Vapaapäiviä ovat muut kuukauden päivät, kuin arkipäivät.

Edellä mainittujen prediktorien lisäksi on etsitty pörssi-indeksejä, jotka kertoisivat talouden tilanteesta. Pörssi-indekseistä on otettu mukaan sekä yleisindeksejä, että toimialakohtaisia indeksejä Pohjoismaista. Toimialakohtaiset pörssi-indeksit löytyvät myös Suomen tasolla, mutta osassa Suomen toimialaindekseistä on mukana vain muutama yritys. Näin ollen on käytetty pohjoismaista indeksiä, jotta yrityksiä tulee mukaan useampia. Pörssi-indeksit kuvaavat sijoittajien näkemyksiä indeksiin kuuluvien yritysten tilanteesta. Yleisindeksejä on valittu Suomesta ja Pohjoismaista kuvaamaan sijoittajien näkemystä Suomen taloudellisesta tilanteesta ja maailmalta kuvaamaan näkemystä globaalista taloudellisesta tilanteesta. Toimialakohtaisten pörssi-indeksien voidaan ajatella kuvaavan sijoittajien näkemystä toimialan tilanneesta Pohjoismaissa.

Käytössä olevien prediktorien mittausväli vaihtelee. Prediktorien arvot ovat saatavilla joko kuukausittain, kvartaaleittain tai vuosittain. Kvartaaleittain ja vuosittain saatavien prediktorien kuukausittaiseksi arvoksi on asetettu kyseisen kvartaalin tai vuoden havaittu arvo.

Kaikki käytetyt prediktorit on listattu taulukossa 4. Taulukossa on lisäksi kerrottu, mistä prediktori on etsitty (julkaisija). Lähdetiedot ja Internet-osoitteet prediktoreille löytyvät lähdeluettelosta ja liitteestä A. Taulukossa 4 on kerrottu, miten usein prediktorista on havaintoja ja mahdollisia tarkennuksia, mikäli prediktoria mitataan useammalla tavalla. Esimerkiksi bruttokansantuote on saatavilla Tilastokeskuksen sivuilla ainakin viitevuoden hintoihin, käypiin hintoihin, muutoksena edelliseen neljännekseen ja muutoksena edelliseen vuoteen. Lisäksi vaihtoehtoina on valitusta aikasarjasta alkuperäinen versio, kausitasoitettu, trendi ja työpäiväkorjattu aikasarja. Tässä tapauksessa on käytetty alkuperäistä aikasarjaa bruttokansantuotteesta, joka on mitattu viitevuoden 2010 hinnoilla. Näin ollen hintojen muutos ei vaikuta käytettyyn bruttokansantuotteeseen.

Käytetyistä prediktoreista on saatavilla arvot joulukuusta 2004 marraskuuhun 2015. Tutkielmassa käytetty aikasarjaregressiomalli edellyttää, että prediktorien arvot ovat saatavilla koko käytetyn kulutuksen ajalta. Mikäli halutaan käyttää jotain prediktoria, joka ei ole saatavilla koko ajalta, joudutaan sähkönkulutuksen aikasarjasta käyttämään vain se osa, jolta prediktorin arvot ovat saatavilla. Kaikilta toimialoilta ei ole saatavilla aivan kaikkia toimialoihin liittyviä prediktoreita.

Muuttujien arvot on haettu lähteistään maaliskuussa 2016. Kaikkien prediktorien arvot eivät ole saatavilla heti kuukauden päätyttyä. Tässä tapauksessa kaikkien käytettävien prediktorien marraskuun 2015 arvot ovat olleet saatavilla maaliskuussa 2016. Esimerkiksi Tilastokeskuksen julkaisemien tilastojen luvut julkaistaan pienellä viiveellä, eli marraskuun 2015 luku ei välttämättä ole saatavilla heti joulukuun alussa.

Prediktorien arvoihin voi liittyä tarkentumista, kun julkaistavasta prediktorista on saa-

tu lisää tietoa. Tällöin prediktorin arvot saattavat muuttua ensimmäisen julkistuksen arvoista. Yksi esimerkki tarkentuvasta prediktoreista on Ansiotasoindeksi (Suomen virallinen tilasto (SVT), 2016b). Osassa prediktoreista maaliskuussa haetut arvot olivat muuttuneet hieman aiemmin haetuista arvoista. Näissä tapauksissa on käytetty uudempaa, tarkentunutta arvoa kuukauden, kvartaalin tai vuoden arvona.

Lisäksi pörssi-indekseistä ja osasta kansantalouden kehitykseen liittyvistä muuttujista on tehty prediktoreita, joiden vaikutus sähkönkulutukseen on viiveellistä. Prediktorista käytetään vuoden, kaksi tai kolme vuotta vanhoja arvoja selittämään sähkönkulutusta nykyhetkellä. Nämä prediktorit on merkitty listassa prediktorin nimen perässä olevalla tähdellä. Viiveellisiksi prediktoreiksi on pyritty valitsemaan sellaisia prediktoreita, joiden vaikutus voisi todellisuudessa olla viiveellistä. Pörssi-indeksien voidaan ajatella kuvaavan sijoittajien näkemystä eri yhtiöiden tilasta ja tulevaisuudennäkymistä. Julkinen ja yksityinen kulutus ja investoinnit sekä luottamusindikaattorit kertovat myös kuluttajien ja julkisen sektorin tulevaisuudennäkymistä, ja voivat vaikuttaa viiveellisesti. Bruttokansantuote on yksi yleisistä mittareista kansantalouden tilalle ja korkojen tavoin ne kertovat kansantalouden tilasta. Sähkön hinta kertoo tarkemmin energiantuotannon tilasta ja korkea sähkön hinta voi vaikuttaa tulevaisuuteen esimerkiksi energiatehokkuuteen tähtäävinä toimenpiteinä. Muutkin prediktorit voivat vaikuttaa viiveellisesti, mutta niitä ei ole alun perin otettu mukaan viiveellisinä mukaan, joten niiden mahdollisten viiveellisten vaikutusten tarkastelu rajataan tutkielman ulkopuolelle.

Osalle prediktoreista on tehty logaritmimuunnos, mikäli prediktorin arvojen vaihtelu on ollut suurempaa aikoina, jolloin prediktori on saanut suurempia arvoja. Tämä on tehty, jotta prediktori saadaan muutettua stationaariseksi myöhemmin. Logaritmiseksi muutetut prediktorit on merkitty taulukkoon x:llä sarakkeessa logaritmi.

Taulukko 4: Käytetyt prediktorit, niiden julkaisijat ja tarkempia tietoja prediktoreista

Prediktori	Julkaisija	Havaintoja / vuosi	Lisätieto	logaritmi
Kansantaloudelliset prediktorit				
ansiotasoindeksi	Tilastokeskus	4		
bruttokansantuote*	Tilastokeskus	4	alkuperäinen, viitevuoden 2010 hinnoin	
elinkustannusindeksi	Tilastokeskus	12		
julkiset investoinnit*	Tilastokeskus	4	alkuperäinen, viitevuoden 2010 hinnoin	x
julkinen kulutus*	Tilastokeskus	4	alkuperäinen, viitevuoden 2010 hinnoin	x
kuluttajien luottamusindikaattori*	Tilastokeskus	12		
lyhyen aikavälin korko*	OECD	12		
makrotalouden indikaattori*	Tilastokeskus	12		
pitkän aikavälin korko*	OECD	12		
raakaöljyn hinta*	Öljy- ja biopolttoaineala ry	12		x
sähkön hinta*	Nord Pool	12	elspot-hinta, Suomi	x
tavaroiden tuonti	Tilastokeskus	12		x
tavaroiden vienti	Tilastokeskus	12		x
työllisyysaste Keski-Suomessa	Tilastokeskus	1	15-64v. sukupuolet yhteensä	
työllisyysaste Suomessa	Tilastokeskus	12	15-64v. sukupuolet yhteensä	
väkiluku Jyväskylässä	Tilastokeskus	1		
yksityiset investoinnit*	Tilastokeskus	12	alkuperäinen, viitevuoden 2010 hinnoin	x
yksityinen kulutus*	Tilastokeskus	12	alkuperäinen, viitevuoden 2010 hinnoin	x
Toimialakohtaiset prediktorit				
toimialan arvonlisäys	Tilastokeskus	4	alkuperäinen, viitevuoden 2010 hinnoin	x
toimialan liikevaihto	Tilastokeskus	12	liikevaihto ja määräindeksi, alkuperäinen	x
toimialan luottamusindikaattori*	Tilastokeskus	12		
toimialan työllisten lukumäärä Suomessa	Tilastokeskus	1	työllisten määrä	
toimialan palkkasummakuvaaja	Tilastokeskus	12	alkuperäinen	x
toimialan volyymi	Tilastokeskus	12	alkuperäinen	x
yöpymiset Keski-Suomessa (majoitus-toimiala)	Tilastokeskus	12	yöpymiset, Keski-Suomi, yhteensä	x
Pörssi-indeksit				
Dow Jones Industrial Average-indeksi*	Yahoo!	12	alkuperäinen	x
Helsingin pörssin yleisindeksi*	Nasdaq	12	gross index	x
Pohjoismaiden pörssin OMX Nordic 40 -indeksi*	Nasdaq	12	gross index	x
Standard & Poor's 500 -indeksi*	Yahoo!	12	alkuperäinen	x
Toimialakohtaiset pörssi-indeksit	Nasdaq	12	gross index	x
Muut preditorit				
kuukauden keskilämpötila Jyväskylässä	Ilmatieteen laitos	12		
arkipäivien lukumäärä kuukaudessa		12	ma-pe, arkipyhät kuuluvat vapaapäiviin	
vapaapäivien lukumäärä kuukaudessa		12	muut kuin arkipäivät	

3 Aikasarja-analyysi

Tässä luvussa esitetään mallinnuksessa käytettävä teoria, lähtien liikeelle yksikertaisemmista aikasarjamalleista. Lisäksi käydään läpi mallin ennustaminen, mallin valintakriteerit ja malliin liittyvät oletukset.

Tavallisiin aikasarjamalleihin ja niiden estimointiin liittyvä osa on peräisin kirjoista Box et al. (2008), Brockwell ja Davis (1991) ja Brockwell ja Davis (2002). Myös mallin jäännöksiin liittyvät testit ovat peräisin edellä mainituista kirjoista. Luvussa käydään läpi myös käytetty regressiomalli ARMA-jäännöksillä. Mallin teoria on peräisin kirjoista Brockwell ja Davis (2002) ja Hyndman ja Athanasopoulos (2013).

3.1 Aikasarja

Aikasarjalla tarkoitetaan joukkoa perättäin ajassa mitattuja havaintoja (Box et al., 2008, s. 21). Aikasarjaa sanotaan diskreetiksi, kun saatavilla on vain diskreetti joukko havaintoja. Usein diskreetin aikasarjan havainnot ovat mitattu tietyllä, kiinnitetyllä aikavälillä toisistaan, esimerkiksi kuukauden välein. Mikäli havaintoja saadaan jatkuvasti, sanotaan aikasarjan olevan jatkuva.

Aikasarja voi olla deterministinen, jolloin sen arvot määräytyvät joko aikasarjan aikaisemmista arvoista tai jonkin ulkopuolisen tekijän funktiona. Vastakohta deterministiselle aikasarjalle on stokastinen aikasarja. Stokastisen aikasarjan arvot eivät määräydy suoraan aikasarjan aikaisemmista arvoista tai muista tekijöistä. Stokastisen aikasarjan voi ajatella olevan reaalisaatio prosessista, jolla on todennäköisyysjakauma. Aikasarja-analyysissa kiinnostuksen kohteena ovat stokastiset aikasarjat. Kiinnostuksen kohteena on mallintaa prosessia, joka on tuottanut kyseisen aikasarjan. (Box et al., 2008, s. 22-23)

Tässä työssä käsiteltävät aikasarjat ovat stokastisia ja diskreettejä. Aineiston aikasarjat y_t koostuvat havainnoista y_1, y_2, $\ldots$, y_T. Aikasarjan havaintojen määrää merkitään jatkossa muuttujalla T.

3.2 Stationaarisuus

Useat aikasarjamallit olettavat aikasarjalta stationaarisuuden. Aikasarja on stationaarinen, jos aikasarjan ominaisuudet eivät riipu ajasta. Stationaarisen aikasarjan havaintojen y_{t_1}, y_{t_2}, $\ldots, y_{t_n}$ yhteisjakauma on sama kuin havaintojen $y_{t_{1+\tau}}, y_{t_{2+\tau}}$, $\ldots, y_{t_{n+\tau}}$ riippumatta aikapisteistä $t_1, t_2, \ldots, t_n, n = 1, 2, \ldots$ ja viiveestä τ.

Aikasarjan stationaarisuutta heikompi käsite on niin kutsuttu 2. asteen heikko stationaarisuus, joka olettaa aikasarjan odotusarvon $\mathbb{E}(y_t)$ olevan riippumaton ajanhetkestä t ja havaintojen autokovarianssin $\gamma(t, t+\tau)$ riippuvan vain viiveestä τ. Toisen asteen heikosta stationaarisuudesta ja aikasarjan havaintojen normaalijakautuneisuudesta seuraa stationaarisuus (Box et al., 2008, s. 29).

Kaikki aikasarjat eivät ole stationaarisia. Aikasarjan vaihtelu saattaa riippua voimakkaasti aikasarjan tasosta. Vaihtelu voi olla tällöin suurempaa aikasarjan saadessa suuria arvoja. Aikasarjassa voi myös esiintyä esimerkiksi trendiä tai kausivaihtelua, jolloin aikasarja ei ole stationaarinen. Edellä mainituissa tilanteissa aikasarja pystytään usein muuttamaan stationaariseksi (Brockwell ja Davis, 1991, kpl. 1.4 ja 9.2).

Mikäli aikasarjan vaihtelu kasvaa trendin kasvaessa, pyritään aluksi muuttamaan vaihtelu riippumattomaksi aikasarjan tasosta (Brockwell ja Davis, 1991, s. 284). Tämä voidaan tehdä esimerkiksi logaritmimuunnoksella, jota käytetään tässä tutkielmassa. On mahdollista käyttää myös esimerkiksi Box-Cox -muunnosta (Box ja Cox, 1964) $y_t = \frac{y_t^{\lambda}-1}{\lambda}$, missä $\lambda > 0$. Logaritmimuunnos saadaan Box-Cox -muunnoksen rajaarvona, kun $\lambda \to 0$.

Brockwell ja Davis (1991) esittävät kaksi tapaa käsitellä trendiä ja kausivaihtelua. Ensimmäisessä tavassa aikasarja y_t jaetaan trendiin m_t, kausivaihteluun s_t ja satunnaiseen komponenttiin e_t, esimerkiksi $y_t = m_t + s_t + e_t$. Trendi- ja kausivaihtelukomponentti oletetaan deterministisiksi. Satunnaisen komponentin e_t oletetaan noudattavan stationaarista prosessia, jolloin stationaarisiin prosesseihin liittyvä teoria on käytössä mallinnettaessa satunnaiskomponenttia. Halutut tulokset aikasarjalle y_t saadaan yhdistämällä tulokset satunnaiskomponentilla e_t deterministisiin trendi- ja kausivaihtelukomponentteihin.

Toinen tapa on käyttää tarvittava määrä perättäisiä erotuksia ja kausierotuksia. Perättäisellä erotuksella tarkoitetaan aikasarjan y_t muutosta aikapisteiden $t-1$ ja t välillä, $y_t - y_{t-1}$. Kausierotus taas on aikasarjan muutos aikapisteiden $t-s$ ja t välillä, $y_t - y_{t-s}$, missä s on kauden pituus. Alkuperäistä aikasarjaa muutetaan perättäisten ja kausierotusten avulla. Esimerkiksi aikasarjassa y_t, jossa on havaittavissa selkeä kausivaihtelu, voidaan lähteä liikkeelle ottamalla kausierotukset, ja tarkastelemalla onko aikasarja $y_t - y_{t-s}$ stationaarinen. Mikäli aikasarja $y_t - y_{t-s}$ ei ole stationaarinen, voidaan ottaa perättäiset erotukset aikasarjasta $y_t - y_{t-s}$ ja tarkastella tämän sarjan stationaarisuutta. Tarvittaessa voidaan käyttää myös useampaa kappaletta perättäisiä ja kausierotuksia.

Stationaarisuutta voidaan tarkastella sekä aikasarjan kuvaajan, että autokorrelaatiofunktion avulla. Autokorrelaatiofunktio esitetään luvussa 3.3. Mikäli aikasarjassa ei esiinny kuvaajan perusteella selvää trendiä tai kausivaihtelua ja autokorrelaatio laskee nopeasti lähelle nollaa viiveen τ kasvaessa, voidaan aikasarjan ajatella olevan stationaarinen (Brockwell ja Davis, 1991, kpl. 9.2). Jos joko kuvaajassa on selkeää trendiä tai kausivaihtelua tai autokorrelaatiofunktio ei laske nopeasti lähelle nollaa, voidaan käyttää uudestaan perättäisiä tai kausierotuksia ja tarkastella stationaarisuusoletuksia uudestaan tämän sarjan osalta. Yleensä perättäisiä erotuksia tarvitaan korkeintaan kolme ja kausierotuksia yksi (Brockwell ja Davis, 1991, s. 323).

Määritellään muutamia jatkossa käytettäviä operaattoreita helpottamaan perättäisten erotusten ja kausierotusten merkitsemistä. Operaattori B antaa aikasarjasta y_t edellisen havainnon, $By_t = y_{t-1}$. Mikäli operaattoria käytetään s kertaa peräkkäin, merkitään $B^s y_t = y_{t-s}$.

Operaattorilla ∇ voidaan merkitä muutosta aikasarjan y_t aikapisteiden $t-1$ ja t välillä: $\nabla y_t = (1-B)y_t = y_t - y_{t-1}$. Vastaavasti muutosta aikapisteiden $t-s$ ja t välillä merkitään $\nabla_s y_t = (1-B^s)y_t = y_t - y_{t-s}$. Mikäli operaattoria ∇ käytetään s' kertaa peräkkäin, merkitään $\nabla^{s'}$. On huomattava, että yleisesti $\nabla_{s'} \neq \nabla^{s'}$.

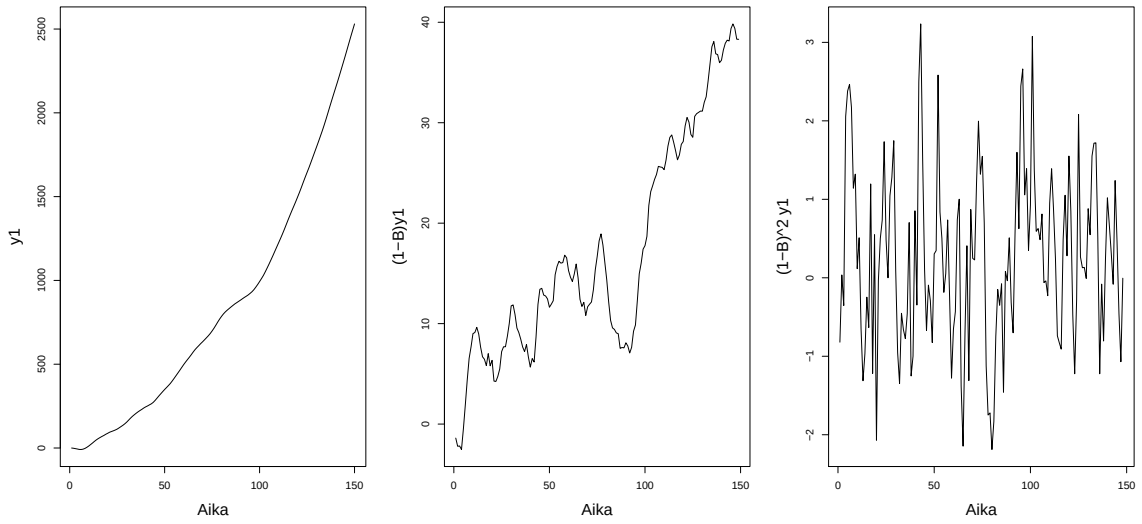
Aikasarjaa y'_t, joka on saatu aikasarjasta y_t ottamalla d kappaletta perättäisiä ja D kappaletta kausittaisia erotuksia, voidaan merkitä $y'_t = \nabla_s^D \nabla^d y_t = (1-B^s)^D(1-B)^d y_t$.

Tarkastellaan kahta esimerkkiä aikasarjan muuttamisesta stationaariseksi perättäisten erotusten ja kausierotusten avulla. (Hyndman ja Athanasopoulos, 2013, kpl. 8.1) toteavat, että perättäisten erotusten ja kausierotusten määrään valinta ei ole aina selvää, ja eri tutkija voi päätyä erilaisiin valintoihin tarkastellessa samaa aikasarjaa.

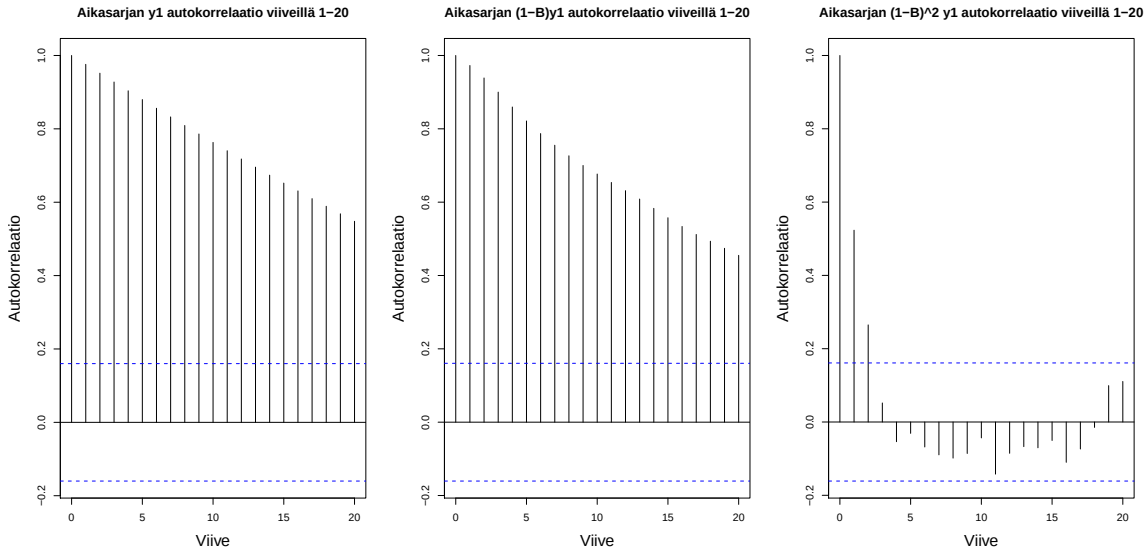
Kuvassa 4 on esitetty vasemmanpuoleisessa kuvassa alkuperäinen aikasarja $y_{1,t}, t = 1, \ldots, 150$. Aikasarjassa on selvä trendi, joten aikasarja ei ole itsessään stationaarinen. Koska aikasarjassa ei ole kausittaista vaihtelua, käytetään perättäisiä erotuksia, eli tarkastellaan aikasarjaa $(1-B)y_{1,t} = y_{1,t} - y_{1,t-1}, t = 2, \ldots 150$, joka on esitetty kuvassa keskellä. Myös tässä aikasarjassa on selkeä ylöspäin menevä trendi. Lisäksi aikasarjassa on hienoista ylös- ja alaspäin vaihtelua trendin lisäksi. Tämä ei kuitenkaan ole kausittaista, joten käytetään stationaariseksi muuttamiseen toisen kerran perättäisiä erotuksia.

Aikasarja $(1-B)^2 y_{1,t} = y_{1,t} - 2y_{1,t-1} + y_{1,t-2}$ ei sisällä selkeää trendiä eikä kausivaihtelua, joten tarkastellaan vielä autokorrelaatiofunktiota, joka on esitetty kuvassa 5. Kuvassa on esitetty autokorrelaatiofunktio myös aikasarjoille $y_{1,t}$ ja $(1-B)y_{1,t}$. Nähdään, että autokorrelaatio menee nopeasti nollaan aikasarjan $(1-B)^2 y_{1,t}$ tapauksessa, joten aikasarjan voidaan sanoa olevan muutettu stationaariseksi käyttämällä kahta perättäistä erotusta. Aikasarjojen $y_{1,t}$ ja $(1-B)y_{1,t}$ autokorrelaatiotkin ovat korkeita myös suurilla viiveillä, joten myös autokorrelaatio kertoo, etteivät nämä aikasarjat ole stationaarisia.

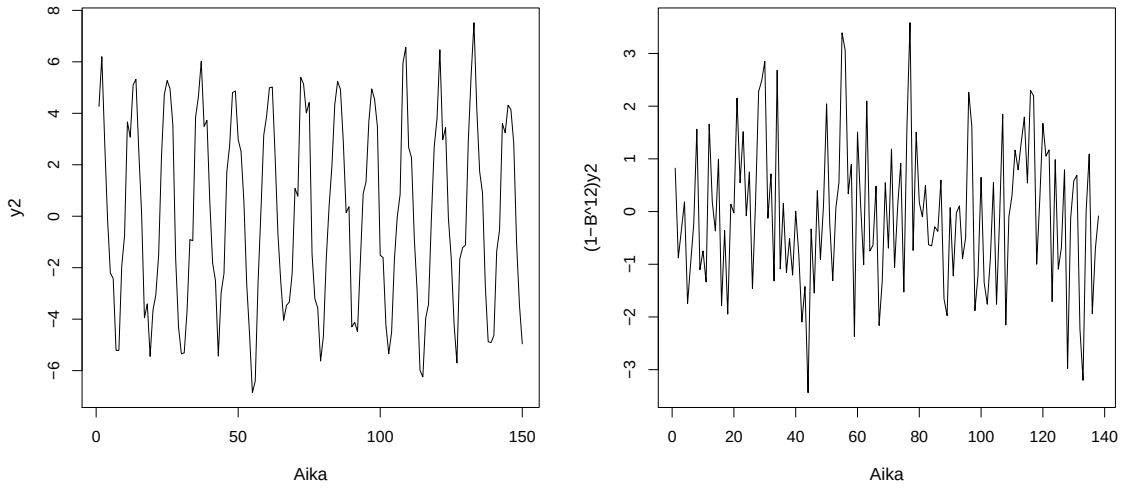
Kuvassa 6 vasemmalla on esitetty toinen ei-stationaarinen aikasarja $y_{2,t}$. Tässä aikasarjassa ei ole trendiä, mutta selkeä kausivaihtelu on havaittavissa. Täten aikasarja ei ole suoraan stationaarinen. Käytetään aikasarjaan kausierotuksia, eli tarkastellaan aikasarjan $(1-B^{12})y_{2,t} = y_{2,t} - y_{2,t-12}$ stationaarisuutta. Aikasarja $(1-B^{12})y_{2,t}$ on esitettu kuvassa 6 oikealla. Kausierotusten aikasarjassa ei esiinny selkeää trendiä eikä kausivaihtelua, joten tarkastellaan vielä autokorrelaatiofunktiota, joka on molemmille aikasarjoille kuvassa 7. Autokorrelaatiofunktion ainoa suurempi piikki on 12 havainnon viiveellä, eli aikasarjassa on negatiivista kausittaista riippuvuutta. Pelkkä kausittainen riippuvuus ei kuitenkaan ole este stationaarisuudelle, jos aikasarjassa ei ole kausivaihtelua. Täten aikasarjan $y_{2,t} - y_{2,t-12}$ voidaan sanoa olevan stationaarinen, ja alkuperäinen aikasarja saatiin muutettua stationaariseksi käyttämällä yhtä kausierotusta.



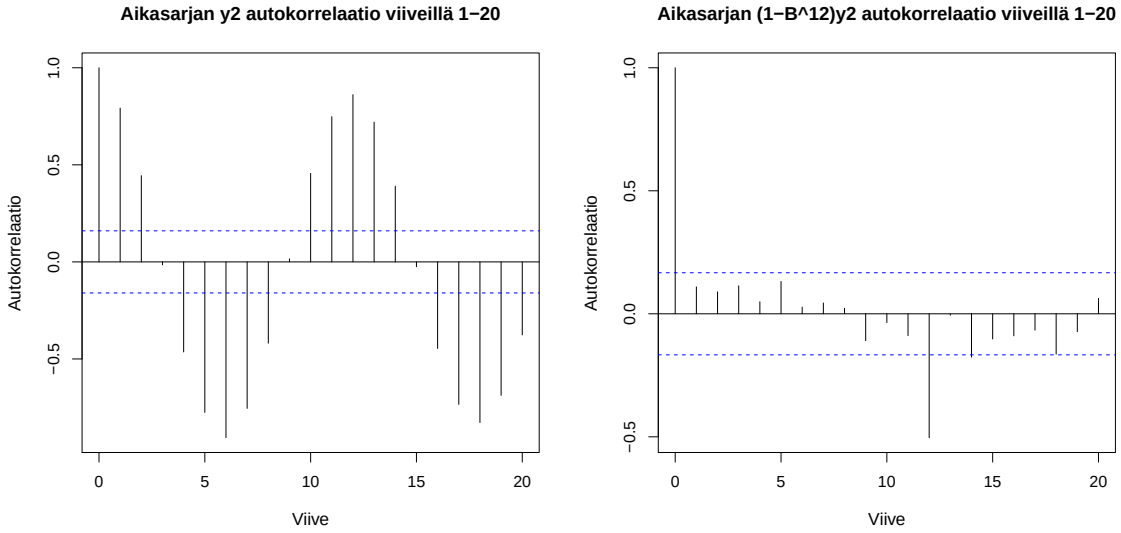
Kuva 4: Esimerkki ei-stationaarisesta aikasarjasta (vasen kuva). Aikasarjasta on otettu perättäiset erotukset (keskimmäinen kuva), mutta aikasarja ei ole vielä stationaarinen. Käyttämällä perättäisiä erotuksia toisen kerran, aikasarja on saatu muutettua stationaariseksi (oikea kuva).



Kuva 5: Autokorrelaatiofunktiot aikasarjoille $y_{1,t}$, $(1-B)y_{1,t}$ ja $(1-B)^2y_{1,t}$, jotka on esitetty kuvassa 4. Autokorrelaatiofunktioiden perusteella aikasarjat $y_{1,t}$ ja $(1-B)y_{1,t}$ eivät ole stationaarisia. Aikasarjan $(1-B)^2y_{1,t}$ stationaarisuus on mahdollista autokorrelaatiofunktion perusteella, mikäli aikasarjassa ei ole trendiä tai kausivaihtelua.



Kuva 6: Esimerkki ei-stationaarisesta aikasarjasta (vasen kuva). Aikasarja on muutettu stationaariseksi käyttämällä kausierotuksia (oikea kuva).



Kuva 7: Autokorrelaatiofunktiot aikasarjoille $y_{1,t}$ ja $(1 - B^{12})y_{1,t}$, jotka on esitetty kuvassa 6. Autokorrelaatiofunktioiden perusteella aikasarja $y_{1,t}$ ei ole stationaarinen. Aikasarjan $(1 - B^{12})y_{1,t}$ stationaarisuus on mahdollista autokorrelaatiofunktion perusteella, mikäli aikasarjassa ei ole trendiä tai kausivaihtelua.

3.3 Aikasarjoihin liittyviä suureita ja operaattoreita

Odotusarvon μ ja varianssin σ^2 estimaatit määritellään tavalliseen tapaan,

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t,$$
$$\hat{\sigma^2} = \frac{1}{T-1}\sum_{i=1}^{T}(y_t - \hat{\mu})^2.$$

Aikasarjan havaintojen y_t ja $y_{t+\tau}$ välistä kovarianssia kutsutaan autokovarianssiksi ja se määritellään

$$\gamma(t, t+\tau) = \mathrm{cov}(y_t, y_{t+\tau}) = \mathbb{E}\left[(y_t - \mu_t)(y_{t+\tau} - \mu_{t+\tau})\right].$$

Kun aikasarja y_t on stationaarinen, autokovarianssi on sama riippumatta aikahetkestä t ja riippuu näin ollen vain viiveestä τ

$$\gamma(t, t+\tau) = \gamma(\tau) = \mathrm{cov}(y_t, y_{t+\tau}) = \mathbb{E}\left[(y_t - \mu)(y_{t+\tau} - \mu)\right].$$

Otoksesta laskettava autokovarianssin estimaatti viiveellä τ on

$$\widehat{\gamma(\tau)} = \frac{1}{T}\sum_{t=1}^{T-\tau}(y_t - \hat{\mu})(y_{t+\tau} - \hat{\mu}).$$

Stationaarisen aikasarjan autokorrelaatio voidaan laskea autokovarianssin ja varianssin avulla, $\rho(\tau) = \gamma(\tau)/\sigma^2$, $\widehat{\rho(\tau)} = \widehat{\gamma(\tau)}/\hat{\sigma^2}$.

Merkitään autokovarianssimatriisia, joka sisältää viiveet $0, 1, \ldots, \tau - 1$

$$\boldsymbol{\Gamma_\tau} = \begin{bmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(\tau-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(\tau-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(\tau-1) & \gamma(\tau-2) & \ldots & \gamma(0) \end{bmatrix}$$

ja autokovarianssivektoria $\boldsymbol{\gamma_\tau} = \begin{bmatrix}\gamma(1) & \gamma(2) & \ldots & \gamma(\tau)\end{bmatrix}'$.

Osittaisautokorrelaatio $\alpha(\tau)$ voidaan laskea autokovarianssin avulla $\alpha(\tau) = \phi_{\tau\tau}$, missä $\phi_{\tau\tau}$ on viimeinen alkio vektorissa $\boldsymbol{\phi_\tau} = \boldsymbol{\Gamma_\tau}^{-1}\boldsymbol{\gamma_\tau}$ (Brockwell ja Davis, 2002, kpl. 3.2.3). Otososittaisautokorrelaatiot $\widehat{\alpha(\tau)}$ voidaan laskea vastaavasti otosautokovarianssimatriisin $\widehat{\boldsymbol{\Gamma_\tau}}$ ja -vektorin $\widehat{\boldsymbol{\gamma_\tau}}$ avulla. Osittaisautokorrelaation voi ajatella havaintojen y_t ja $y_{t+\tau}$ väliseksi korrelaatioksi, josta on poistettu havaintojen $y_{t+1}, \ldots, y_{t+\tau-1}$ vaikutus (Box et al., 2008, s. 68).

3.4 Aikasarjaprosesseja

Puhdas satunnaisprosessi z_t on yksinkertainen aikasarjaprosessi, johon muut jatkossa esitettävät prosessit perustuvat. Puhtaalle satunnaisprosessille pätee $\mathbb{E}(z_t) = 0$ ja $\gamma(0) = \sigma^2, \gamma(\tau) = 0$, kun $\tau \neq 0$.

Jatkossa oletetaan lisäksi, että puhdas satunnaisprosessi z_t on normaalijakautunut. Normaalijakautuneisuudesta ja korreloimattomuudesta seuraa myös satunnaisprosessin arvojen riippumattomuus toisistaan. Normaalisuusoletus helpottaa jatkossa esimerkiksi ennusteiden ennustevälien laskentaa (Hyndman ja Athanasopoulos, 2013, kpl. 2.6). Puhdas satunnaisprosessi on z_t, missä $z_t \sim$ i.i.d. $N(0, \sigma^2)$. Puhdas satunnaisprosessi on stationaarinen aikasarja.

Kertalukua p oleva autoregressiivinen prosessi on muotoa

$$y_t - \sum_{i=1}^{p} \phi_i y_{t-i} = z_t,$$

missä $\phi_i \in \mathbb{R}, i = 1, \ldots, p$ ja z_t on puhdas satunnaisprosessi. Autoregressiivinen prosessi on tavallaan regressiomalli, jossa prediktoreina ovat aikasarjan p edellisestä havaintoa. Kertaluvun p autoregressiivinen prosessi on stationaarinen täsmälleen silloin, kun polynomin

$$\phi(w) = 1 - \phi_1 w - \phi_2 w^2 - \cdots - \phi_p w^p$$

kaikki juuret $w \in \mathbb{C}$ ovat kompleksitason yksikköympyrän ulkopuolella. Todistus löytyy esimerkiksi lähteestä (Box et al., 2008, kpl. 3.2.1). Kertaluvun p autoregressiivistä prosessia voidaan merkitä lyhenteellä AR(p).

Kertalukua q oleva liukuvan keskiarvon prosessi y_t on muotoa

$$y_t = \sum_{i=0}^{q} \theta_i z_{t-i},$$

missä $\theta_0 = 1$, $\theta_i \in \mathbb{R}, i = 1, \ldots, q$ ja z_t on puhdas satunnaisprosessi. Liukuvan keskiarvon prosessi on painotettu summa satunnaisprosessin z_t nykyisestä ja q edellisestä arvosta. Kertaluvun q liukuvan keskiarvon prosessi on stationaarinen, koska se on erikoistapaus lineaarisesta prosessista $x_t = \mu + \sum_{i=0}^{\infty} \psi_i z_{t-i}$, kun asetetaan $\mu = 0, \psi_i = \theta_i, i = 0, \ldots, q$ ja $\psi_i = 0, i > q$. Lineaarinen prosessi on stationaarinen, kun $\sum_{i=0}^{\infty} |\psi_i| < \infty$ (Box et al., 2008, s. 9). Kertaluvun q liukuvan keskiarvon prosessille käytetään lyhennettä MA(q).

Autoregressiivisen ja liukuvan keskiarvon yhdistelmä on ARMA(p,q)-prosessi (Brockwell ja Davis, 1991, kpl 3.1), joka on muotoa

$$\phi(B) y_t = \theta(B) z_t$$

missä z_t on puhdas satunnaisprosessi, $\phi(B)$ määritellään kuten AR-prosessin tapauksessa ja $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$. Lisäksi oletetaan, että y_t on stationaarinen.

Tarkastellaan tarkemmin, mitä polynomeilta $\phi(B)$ ja $\theta(B)$ vaaditaan, jotta y_t on yksikäsitteisesti määritelty ja stationaarinen.

Määritellään kaksi käsitettä, joita tarvitaan jatkossa. ARMA-prosessi $\phi(B)y_t = \theta(B)z_t$ on kausaalinen (Brockwell ja Davis, 1991, Määr. 3.1.3), kun on olemassa vakiot $\psi_j, j = 0, 1, \ldots,$ siten, että $\sum_{j=0}^{\infty} |\psi_j| < \infty$ ja

$$y_t = \sum_{j=0}^{\infty} \psi_j z_{t-j} \text{ kaikilla } t.$$

Toisin sanoen mikäli ARMA-prosessi y_t on kausaalinen, se voidaan esittää lineaarisena prosessina. ARMA-prosessi on kausaalinen täsmälleen silloin, kun polynomeilla $\phi(w)$ ja $\theta(w)$ ei ole yhteisiä juuria $w \in \mathbb{C}$ ja polynomin $\phi(w)$ juuret ovat kompleksitason yksikköympyrän ulkopuolella. Tällöin y_t on myös stationaarinen (Brockwell ja Davis, 1991, Lause 3.1.1).

Lisäksi ARMA-prosessin sanotaan olevan kääntyvä (Brockwell ja Davis, 1991, Määr. 3.1.4), kun on olemassa vakiot $\pi_j, j = 0, 1, \ldots$ siten, että $\sum_{j=0}^{\infty} \pi_j < \infty$ ja

$$z_t = \sum_{j=0}^{\infty} \pi_j y_{t-j} \text{ kaikilla } t.$$

ARMA-prosessi on kääntyvä täsmälleen silloin, kun polynomeilla $\phi(w)$ ja $\theta(w)$ ei ole yhteisiä juuria $w \in \mathbb{C}$ ja polynomin $\theta(w)$ juuret ovat kompleksitason yksikköympyrän ulkopuolella (Brockwell ja Davis, 1991, Lause 3.1.2).

Jatkossa käsitellään pelkästään ARMA-prosesseja, jotka ovat kausaalisia ja kääntyviä. Brockwell ja Davis (1991, s. 88) toteavat, että mikäli ARMA-prosessin $\phi(B)y_t = \theta(B)z_t$ polynomeilla $\phi(w)$ ja $\theta(w)$ ei ole juuria $w \in \mathbb{C}$, joille $|w| = 1$, voidaan löytää polynomit $\tilde{\phi}(B)$ ja $\tilde{\theta}(B)$ siten, että prosessi $\tilde{\phi}(B)y_t = \tilde{\theta}(B)\tilde{z}_t$ on kausaalinen ja kääntyvä ja $\tilde{z}_t \sim$ i.i.d. $N(0, \sigma^2)$.

Kertaluvun p' autoregressiivinen prosessi saadaan ARMA-prosessista asettamalla $p = p', q = 0$ ja kertaluvun q' liukuvan keskiarvon prosessi saadaan asettamalla $p = 0, q = q'$.

Määritellään malli, joka ottaa huomioon eri vuosien saman kauden välisen riippuvuuden. Merkitään kauden pituutta muuttujalla $s, s > 1$. Multiplikatiivinen ARMA$(p, q) \times (P, Q)_s$-malli on muotoa

$$\Phi(B^s)\phi(B)y_t = \Theta(B^s)\theta(B)z_t.$$

Polynomit $\phi(B)$ ja $\theta(B)$ määritellään kuten edellä, $\Phi(B^s) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{sP}$ ja $\Theta(B^s) = 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{sQ}$, missä $\Phi_i, \Theta_j \in \mathbb{R}, i = 1, \ldots, P, j = 1, \ldots, Q$.

On syytä huomata, että ARMA$(p, q) \times (P, Q)_s$ -prosessi voidaan esittää ARMA$(p + sP, q + sQ)$-prosessina, jossa osa kertoimista voi olla nollia (Brockwell ja Davis, 2002, s. 206).

Esimerkiksi ARMA$(1,0) \times (1,0)_s$ -prosessi, jossa $\phi_1 = 0.5$ ja $\Phi_1 = 0.7$, on muotoa

$$\begin{aligned}(1 - 0.7B^s)(1 - 0.5B)y_t &= z_t \\ \Leftrightarrow (1 - 0.5B - 0.7B^s + 0.35B^{1+s})y_t &= z_t.\end{aligned}$$

Tällöin valitsemalla kertoimet $\phi_1^* = 0.5, \phi_s^* = 0.7, \phi_{1+s}^* = -0.35$ ja $\phi_i^* = 0$, kun $i = 3, 4, \ldots, s-1$, saadaan ARMA$(1,0) \times (1,0)_s$-prosessi esitettyä ARMA$(1+s,0)$-prosessina.

Multiplikatiivinen ARMA$(p,q) \times (P,Q)_s$-prosessi on stationaarinen täsmälleen silloin kun polynomien $\phi(w)$ ja $\Phi(w)$ juuret ovat kompleksitason yksikköympyrän ulkopuolella (Brockwell ja Davis, 2002, s. 203).

Edelliset mallit olettavat aikasarjan y_t olevan stationaarinen. Tällöin aikaisempien mallien käyttö ei ole järkevää ei-stationaarisen aikasarjan tapauksessa. Edellisiä malleja voidaan kuitenkin käyttää, mikäli aikasarja muutetaan ensin stationaariseksi. Oletetaan, että stationaariseksi muuttaminen voidaan tehdä, kuten luvussa 3.2. Oletetaan lisäksi, että stationaariseksi muuttamiseen on tarvittu d kappaletta perättäisiä erotuksia ja D kappaletta kausierotuksia.

Näihin perustuen määritellään multiplikatiivinen ARIMA$(p,d,q) \times (P,D,Q)_s$ -malli:

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d y_t = \Theta(B^s)\theta(B)z_t.$$

Multiplikatiivinen ARIMA$(p,d,q) \times (P,D,Q)_s$ -malli on sama malli kuin multiplikatiivinen kausivaihtelumalli ARMA$(p,q) \times (P,Q)_s$, jos asetetaan $d, D = 0$.

3.5 Aikasarjaregressio ARMA-jäännöksillä

Sähkönkulutusta y_t pyritään selittämään prediktoreilla $x_{i,t}, i = 1, \ldots, k, k \in \mathbb{Z}_+$. Mallinnus tehdään aikasarjaregressiolla. Aikasarjaregression teoria esitetään stationaarisille kulutukselle y'_t ja prediktoreille $x'_{i,t}, i = 1, \ldots, k$, joten etsitään ensin perättäisten erotusten ja kausierotusten määrät d ja D siten, että kulutus $y'_t = \nabla_{12}^D \nabla^d y_t$ ja prediktorit $x'_{i,t} = \nabla_{12}^D \nabla^d x_{i,t}, i = 1, 2, \ldots, k$ ovat likimain stationaarisia, ks. luku 3.2. Hyndman ja Athanasopoulos (2013) suosittelevat käyttämään perättäisiä erotuksia ja kausierotuksia sekä sähkönkulutukselle että prediktoreille, siten että sähkönkulutuksen ja prediktorit suhtautuvat toisiinsa järkevästi. Esimerkiksi jos käytetään sähkönkulutukselle yhtä kausierotusta, mittaa muunnettu aikasarja sähkönkulutuksen muutosta vuoden takaiseen sähkönkulutukseen. Tällöin on järkevä käyttää prediktorina prediktorin muutosta vuoden takaiseen prediktorin arvoon. Mikäli prediktori on mitattu kuukautena liittyvänä arvona, on syytä käyttää kausierotuksia prediktorin muuttamiseksi mittaamaan vuosittaista muutosta. Sen sijaan mikäli prediktori on mitattu suoraan vuosittaisena muutoksena, ei kausierotusta tarvitse käyttää.

Tällöin sähkönkulutusta y'_t voidaan kuvata mallilla

$$y'_t = \sum_{j=1}^{k} \beta_j x'_{j,t} + \eta'_t,$$

missä $x'_{i,t}$ on prediktorin i arvo hetkellä t, $i = 1, \ldots, k$. Mallissa siis oletetaan prediktorien vaikutuksen vasteeseen olevan lineaarista. Lisäksi jäännösten η'_t oletetaan noudattavan multiplikatiivista ARMA$(p, q) \times (P, Q)_s$ -mallia, eli

$$\Phi(B^s)\phi(B)\eta'_t = \Theta(B^s)\theta(B)z_t.$$

Jäännösten $z_t, t = 1, \ldots, T$ oletetaan olevan keskenään korreloimattomia. Lisäksi niiden oletetaan noudattavan normaalijakaumaa $N(0, \sigma^2)$.

Tämä poikkeaa tavallisesta lineaarisesta regressiomallista vain jäännösten osalta. Koska aikasarjan viiveellä τ havaittujen arvojen välillä on usein korrelaatiota, on luonnollista olettaa myös jäännösten sarjan termien välinen korreloituneisuus. Lisäksi asettamalla mallin parametrit p, q, P, Q nolliksi, saadaan tavallinen regressiomalli erikoistapauksena kyseisestä mallista.

Tässä vaiheessa on syytä huomata, mikäli käytettäisiin tavallista regressiomallia aikasarjaregression sijasta, voivat tavallisen regressiomallin estimoidut kertoimet $\hat{\beta}_1, \ldots, \hat{\beta}_k$ poiketa aikasarjaregression estimoiduista kertoimista. Kertoimet poikkeavat, jos tavallisen regressiomallin oletukset jäännösten riippumattomuudesta ja normaalijakautuneisuudesta eivät ole voimassa. Tällöin tavallisen regressiomallin antamat tulkinnat prediktorien muutoksen vaikutuksesta vasteeseen eivät pidä paikkaansa.

Kun malli on saatu tehtyä stationaariselle kulutukselle, täytyy se muuttaa takaisin sähkönkulutuksen aikasarjalle sopivaksi. Olkoon a_n polynomin $(1 - B^{12})^D(1 - B)^d$ n. asteen kerroin, jolloin $(1-B^{12})^D(1-B)^d = a_0 + a_1 B + \cdots + a_{d+12D}B^{d+12D}$. Polynomin vakio $a_0 = 1$.

Tällöin $y'_t = \sum_{i=0}^{d+12D} a_i y_{t-i} = y_t + \sum_{i=1}^{d+12D} a_i y_{t-i}$, joten y_t saadaan ratkaistua vasemman- ja oikeanpuoleisimpia lausekkeita käyttämällä:

$$\begin{aligned} y_t &= y'_t - \sum_{i=1}^{d+12D} a_i y_{t-i} \\ &= \eta'_t + \sum_{j=1}^{k} \beta_j x'_{j,t} - \sum_{i=1}^{d+12D} a_i y_{t-i}, \end{aligned}$$

missä toinen yhtäsuuruus on totta aikasarjaregressio-oletuksen nojalla. Näin ollen sähkönkulutus y_t saadaan esitettyä aikaisempien havaintojen ja stationaarisen aikasarjaregression avulla. Piste-ennusteet sähkönkulutukselle y_t saadaan tehtyä kuten luvussa 3.7.3 käydään läpi, korvaamalla luvussa käytetyt termi η_t termillä y_t ja $-\sum_{j=1}^{d+Ds} a_j \eta_{t-j}$ termillä $\sum_{j=1}^{k} \beta_j x'_{j,t} - \sum_{i=1}^{d+12D} a_i y_{t-i}$ ja ennustamalla jäännösten aikasarjan η'_t arvoja. Lisäksi tarvitaan prediktorien arvot $x'_{i,t}$ halutulle ennustevälille $t = T + 1, \ldots, T + h$.

Aikasarjaregressiossa oletetaan, että prediktorit $x'_{i,t}$ vaikuttavat vasteeseen y'_t, mutta vaikutusta ei ole toiseen suuntaan. Mikäli oletetaan, että myös vasteen nykyarvo voi vaikuttaa prediktorien nykyisiin arvoihin tai tulevaisuuden arvoihin, voidaan käyttää esimerkiksi monimuuttujaista aikasarja-analyysia, ks. esim. Box et al. (2008, kpl. 14) tai

Brockwell ja Davis (2002, kpl. 7). Tässä tutkielmassa oletetaan, että sähkönkulutuksen riippuvuus prediktoreista on yksisuuntaista. Käytössä olevilla prediktoreilla tämä oletus lienee todenmukainen.

3.6 Aikasarjaregression suurimman uskottavuuden estimointi

Parametrien estimointi tehdään R-ohjelmistolla (R Core Team, 2015) stationaarisille kulutuksille ja prediktoreille, joille estimoidaan kertoimet $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \dots & \beta_k\end{bmatrix}'$ ja ARMA$(p,q) \times (P,Q)_s$-prosessin parametrit. Tämä luku perustuu kirjaan Brockwell ja Davis (2002, kpl. 6.6.2). Oletetaan aluksi, että prediktorit ja parametrit p, q, P ja Q on kiinnitetty, jolloin parametrien estimointi perustuu uskottavuusfunktioon.

Esitetään ARMA$(p,q) \times (P,Q)_s$-prosessi ARMA$(p+sP, q+sQ)$-prosessina, jossa osa kertoimista $\boldsymbol{\phi} = \begin{bmatrix}\phi_1 & \phi_2 & \dots & \phi_{p+sP}\end{bmatrix}'$ ja $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2 & \dots & \theta_{q+sQ}\end{bmatrix}'$ voi olla asetettu nollaksi. Lisäksi kertoimia rajoitetaan siten, että kertoimet ovat muunnettavissa takaisin ARMA$(p,q)\times(P,Q)_s$-prosessin kertoimiksi $\phi_1^*,\dots\phi_p^*,\Phi_1^*,\dots\Phi_P^*,\theta_1^*,\dots,\theta_q^*,\Theta_1^*,\dots,\Theta_Q^*$.

Kertoimet $\boldsymbol{\beta}, \boldsymbol{\phi}$ ja $\boldsymbol{\theta}$ saadaan estimoitua maksimoimalla uskottavuusfunktio

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = (2\pi)^{-\frac{T}{2}} (\det \boldsymbol{\Gamma_T})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Gamma_T}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right\},$$

missä

$$\boldsymbol{y} = \begin{bmatrix}y_1 & y_2 & \dots & y_T\end{bmatrix}'$$

ja

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ x_{2,1} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,T} & x_{2,T} & \dots & x_{k,T} \end{bmatrix}.$$

Merkitään regressiomallin ARMA-jäännöksiä $\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{X\beta}$. Tällöin $\mathrm{cov}(\boldsymbol{w}) = \boldsymbol{\Gamma_T}$. Koska $\boldsymbol{w} = \begin{bmatrix}w_1 & \dots & w_T\end{bmatrix}'$ on ARMA$(p+sP, q+sQ)$-prosessi, saadaan parametrit $\boldsymbol{\beta}, \boldsymbol{\phi}$ ja $\boldsymbol{\theta}$ estimoitua (Brockwell ja Davis, 2002, s. 158-160), minimoimalla parametrien $\boldsymbol{\beta}, \boldsymbol{\phi}$ ja $\boldsymbol{\theta}$ suhteen lauseke

$$\log\left(\frac{\sum_{t=1}^{T} \frac{(w_t - \widehat{w_t})^2}{r_{t-1}}}{T}\right) + \frac{\sum_{t=1}^{T} \log r_{t-1}}{T}.$$

Edellisessä lausekkeessa muuttujalla $\widehat{w_t}$ tarkoitetaan muuttujan w_t yhden askeleen ennustetta ehdolla aikaisemmat havainnot,

$$\widehat{w_t} = \mathbb{E}(w_t | w_{t-1}, w_{t-2}, \dots, w_1), \widehat{w_1} = 0, t = 1, \dots, T.$$

Muuttuja $r_t\sigma^2 = \mathbb{E}(w_t - \widehat{w_t})^2$ on yhden askeleen ennusteen keskineliövirhe. Varianssin σ^2 suurimman uskottavuuden estimaattori saadaan estimoitujen parametrien $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\phi}}$ ja $\hat{\boldsymbol{\theta}}$ avulla, $\hat{\sigma^2} = S(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/T$.

3.7 Ennustaminen

Kulutuksen ennuste kuukausille $T+1, T+2, \ldots, T+h$, missä h on halutun ennustevälin pituus, koostuu kahdesta osasta. Regressioparametrien osuus ennusteesta saadaan kuten tavallisessa lineaarisessa regressiossa estimoitujen parametrien arvojen $\hat{\boldsymbol{\beta}}$ ja prediktorien arvojen $x_{i,T+1}$, $x_{i,T+2}$, $\ldots$, $x_{i,T+h}$, $i = 1, 2, \ldots, k$ avulla. Toinen osa ennusteesta on multiplikatiivisen ARMA-prosessin osa ennusteesta. Käydään läpi ensin ARMA-prosessin ennustaminen, jota tarvitaan regressiomallin ennusteen tekemiseen. Lisäksi käydään läpi tuloksissa vertailumallina käytettävän multiplikatiivisen ARIMA-prosessin ennustaminen. Multiplikatiivinen ARIMA-prosessi ei sisällä prediktoreita ja on täten hyvä vertailumalli sille, onko prediktoreista hyötyä.

3.7.1 ARMA-prosessin ennustaminen

Olkoon η'_t ARMA(p, q)-prosessi, $\phi(B)\eta'_t = \theta(B)z_t$. Ennusteet $\eta'_{T+1}, \ldots, \eta'_{T+h}$ voidaan laskea havaintojen η'_1, $\ldots$, η'_T ja näiden yhden askeleen ennusteiden $\hat{\eta}'_1 = 0$, $\hat{\eta}'_2$, $\ldots$, $\hat{\eta}'_T$ avulla.

Brockwell ja Davis (2002, kpl. 3.3) antavat kaavat yhden askeleen ennusteelle $\hat{\eta}'_{t+1}$, $t = 1, \ldots, T-1$:

$$\begin{aligned}
\hat{\eta}'_{t+1} &= \sum_{j=1}^{t} \theta_{tj} \left(\eta'_{t+1-j} - \hat{\eta}'_{t+1-j}\right), && 1 \leq t < \max(p, q) \\
&= \sum_{i=1}^{p} \phi_i \eta'_{t+1-i} + \sum_{j=1}^{q} \theta_{tj} \left(\eta'_{t+1-j} - \hat{\eta}'_{t+1-j}\right), && t \geq \max(p, q).
\end{aligned}$$

Ennusteen keskineliövirhe on $\mathbb{E}\left(\eta'_{t+1} - \hat{\eta}'_{t+1}\right)^2 = \nu_t$. Ennusteen laskemisessa tarvittavat kertoimet θ_{tj} ja keskineliövirheet ν_t saadaan innovaatioalgoritmista (Brockwell ja Davis, 2002, kpl. 2.5.2), joka laskee rekursiivisesti yhden askeleen ennusteet aikasarjalle ehdolla aikaisemmat havainnot. Jotta innovaatioalgoritmia voidaan käyttää, aikasarjan odotusarvon on oltava nolla ja varianssin äärellinen.

ARMA(p, q)-prosessin tapauksessa innovaatioalgoritmia voidaan soveltaa aikasarjaan w_t,

$$\begin{aligned}
w_t &= \sigma^{-1}\eta'_t, && t = 1, \ldots, \max(p, q) \\
&= \sigma^{-1}\phi(B)\eta'_t, && t > \max(p, q).
\end{aligned}$$

Muunnos helpottaa laskemista verrattuna siihen, että sovellettaisiin suoraan sarjaan η'_t (Brockwell ja Davis, 1991, s. 175).

Innovaatioalgoritmia varten määritellään muutamia merkintöjä. Merkitään kovarianssia $\kappa(i, j) = \mathbb{E}(w_i w_j)$, $i, j = 1, \ldots T$, $\nu_0 = \kappa(1, 1) = \mathbb{E}(w_1)^2$. Merkitään lisäksi $\theta_{0,i} = 0$, kaikilla $i = 1, \ldots T$ ja $\nu_j = 0$, kun $j < 0$. Kertoimet saadaan laskettua innovaatioalgoritmilla seuraavista kaavoista järjestyksessä $\theta_{1,1}, \nu_1, \theta_{2,2}, \theta_{2,1}, \nu_2, \ldots$ (Brockwell ja Davis, 2002, s. 73):

$$\theta_{t,t-k} = \nu_k^{-1}\left(\kappa(t+1,k+1) - \sum_{j=0}^{k-1}\theta_{k,k-j}\theta_{t,t-j}\nu_j\right), 0 \le k < t$$

$$\nu_t = \kappa(t+1,t+1) - \sum_{j=0}^{t-1}\theta_{t,t-j}^2\nu_j.$$

Varsinainen l askeleen ennuste $P_T\eta'_{T+l}$ saadaan laskettua alkuperäisten havaintojen ja näiden yhden askeleen ennusteiden avulla:

$$P_T\eta'_{T+l} = \sum_{i=1}^{p}\phi_i P_T\eta'_{T+l-i} + \sum_{j=l}^{q}\theta_{T+l-1,j}\left(\eta'_{T+l-j} - \hat{\eta}'_{T+l-j}\right).$$

Termi P_T viittaa siihen, että ennuste on tehty aikasarjan T havaitun arvon perusteella. Kun $1 \le t \le T$, pätee $P_T\eta'_t = \eta'_t$. Ennusteen keskineliövirhe saadaan ennusteen pituuden l funktiona

$$\sigma_T^2(l) = \mathbb{E}(\eta'_{T+l} - P_T\eta'_{T+l})^2 = \sum_{j=0}^{l-1}\left(\sum_{r=0}^{j}\chi_r\theta_{n+l-r-1,j-r}\right)^2\nu_{n+l-j-1},$$

missä $\chi_0 = 1$ ja $\chi_r = \sum_{i=1}^{\min(p,r)}\phi_i\chi_{r-i}, r = 1,2,\ldots$.

3.7.2 ARIMA-prosessin ennustaminen

Olkoon η_t ARIMA(p,d,q)-prosessi, $(1-B)^d\eta_t = \eta'_t$. Prosessi $\eta'_t, t = d+1,\ldots,T$ on tällöin ARMA(p,q)-prosessi. Tällöin η_t voidaan kirjoittaa aikaisempien havaintojensa ja prosessin η'_t avulla muodossa (Brockwell ja Davis, 2002, s. 198-200)

$$\eta_t = \eta'_t - \sum_{j=1}^{d}\binom{d}{j}(-1)^j\eta_{t-j}.$$

Ennuste ajanhetkelle $T+l$ saadaan edellisen yhtälön avulla

$$P_T\eta_{T+l} = P_T\eta'_{T+l} - \sum_{j=1}^{d}\binom{d}{j}(-1)^j P_T\eta_{T+l-j}.$$

Olettamalla, että havainnot $\eta_1,\ldots\eta_d$ ovat riippumattomia havainnoista $\eta'_t, t = d+1,\ldots,T$, voidaan $P_T\eta'_{T+l}$ laskea kuten ARMA-mallin tapauksessa havainnoista $\eta'_{d+1},\ldots,\eta'_T$. Koska $P_T\eta_t = \eta_t$, kun $1 \le t \le T$, riittää yhden askeleen ennusteeseen laskea $P_T\eta'_{T+1}$. Yhden askeleen ennusteen avulla voidaan laskea kahden askeleen ennuste laskemalla $P_T\eta'_{T+2}$. Vastaavasti kun $P_T\eta_{T+1},\ldots,P_T\eta_{T+l-1}$ on jo laskettu, saadaan $P_T\eta_{T+l}$ laskettua ennusteen $P_T\eta'_{T+l}$ avulla.

Ennusteen keskineliövirhe saadaan laskettua vastaavasti kuin ARMA-mallin tapauksessa,

$$\sigma_T^2(l) = \mathbb{E}(\eta_{T+l} - P_T\eta_{T+l})^2 = \sum_{j=0}^{l-1}\left(\sum_{r=0}^{j}\chi_r\theta_{n+l-r-1,j-r}\right)^2 \nu_{n+l-j-1},$$

missä $\nu_{n+l-j-1} = \mathbb{E}(\eta'_{n+l-j} - \hat{\eta}'_{n+l-j})^2 = \mathbb{E}(\eta_{n+l-j} - \hat{\eta}_{n+l-j})^2$.

Määritellään polynomi $\phi^*(z) = (1-z)^d\phi(z) = 1 - \phi_1^* z - \cdots - \phi_{p+d}^* z^{p+d}$. Kertoimet χ_r saadaan laskettua rekursiivisesti $\chi_r = \sum_{i=1}^{\min(p,r)} \phi_i^*\chi_{r-i}$. Vaihtoehtoisesti kertoimet voidaan laskea Taylorin sarjakehitelmästä pisteessä 0 polynomille $1/\phi^*(z)$, eli kertoimet voidaan esittää muodossa $1/\phi^*(z) = \sum_{r=0}^{\infty}\chi_r z^r$.

3.7.3 Multiplikatiivisen ARIMA-prosessin ennustaminen

Olkoon η_t multiplikatiivinen ARIMA-prosessi siten että

$$\eta'_t = (1-B)^d(1-B^s)^D\eta_t, t = d + Ds + 1, \ldots, T$$

on ARMA-prosessi. Vastaavasti kuin edellä, multiplikatiivinen ARIMA-prosessi voidaan esittää aikaisempien havaintojensa ja ARMA-prosessin viimeisen havainnon avulla.

Olkoon a_j polynomin $(1-B^s)^D(1-B)^d$ j. asteen termin kerroin, $j = 0, \ldots, d + Ds$, jolloin

$$(1-B^s)^D(1-B)^d = 1 + a_1 B + \cdots + a_{d+Ds}B^{d+Ds}.$$

Koska

$$\eta'_t = (1-B^s)^D(1-B)^d\eta_t = \eta_t + a_1\eta_{t-1} + \cdots + a_{d+Ds}\eta_{t-d-Ds}, \text{ kun } t > d + Ds,$$

saadaan vasemman- ja oikeanpuoleisen lausekkeen avulla η_t esitettyä muodossa

$$\eta_t = \eta'_t - \sum_{j=1}^{d+Ds} a_j\eta_{t-j}, \text{ kun } t > d + Ds.$$

Vastaavasti myös l askeleen ennuste voidaan esittää muodossa

$$P_T\eta_{T+l} = P_T\eta'_{T+l} - \sum_{j=1}^{d+Ds} a_j P_T\eta_{T+l-j}.$$

Olettamalla, että havainnot $\eta_1, \ldots \eta_{d+Ds}$ ovat riippumattomia havainnoista η'_t, voidaan myös tässä tapauksessa ennusteet $P_T\eta_{T+l}$ laskea rekursiivisesti ennusteiden $P_T\eta'_{T+l}$ ja $P_T\eta_{T+1}, \ldots, P_T\eta_{T+l-1}$ avulla.

Ennusteen keskineliövirhe on

$$\sigma_T^2(l) = \mathbb{E}(\eta_{T+l} - P_T\eta_{T+l})^2 = \sum_{j=0}^{l-1}\left(\sum_{r=0}^{j}\chi_r\theta_{n+l-r-1,j-r}\right)^2 \nu_{n+l-j-1}.$$

Merkitään $\phi^*(z) = \phi(z)\Phi(z^s)(1-z)^d(1-z^s)^D = 1-\phi_1^* z - \cdots - \phi_{p+d+Ds}^* z^{p+d+Ds}$. Kerroin χ_r saadaan r. asteen kertoimena Taylorin sarjakehitelmästä pisteessä 0 polynomille $1/\phi^*(z)$.

3.8 Mallin valinta ja hyvyys

Edellä oletettiin, että multiplikatiivisen ARMA-prosessin parametrien $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}$ lukumäärät p, q, P ja Q on kiinnitetty. Tällöin voidaan käyttää uskottavuusfunktiota mallin parametrien estimoimiseen. Kuitenkaan tässä tapauksessa parametrien määrä ei ole selvillä ennen mallinnuksen aloitusta, vaan myös p, q, P ja Q valitaan mallinnuksessa. Päätökset parametrien lukumääristä p, q, P ja Q tehdään sovittamalla useita eri malleja jäännöksille eri parametrien määrillä p, q, P ja Q ja valitsemalla parhaat p, q, P ja Q päätetyn mallinvalintakriteerin perusteella.

Akaiken informaatiokriteeri on Akaiken (1974) kehittämä mallinvalintakriteeri, joka perustuu mallin logaritmiseen uskottavuusfunktioon ja mallissa käytettävien parametrien määrään. Akaiken informaatiokriteeri määritellään multiplikatiiviselle ARMA-mallille:

$$AIC = -2l(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \sigma^2) + 2(p + q + P + Q + 1),$$

missä $l(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \sigma^2)$ on multiplikatiivisen ARMA-mallin logaritminen uskottavuusfunktio ehdolla parametrit $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \sigma^2$. Koska myös σ^2 on mallin parametri, tulee parametrien yhteismääräksi $p + q + P + Q + 1$. Malli, jolle Akaiken informaatiokriteeri antaa pienimmän arvon, on informaatiokriteerillä mitattuna paras.

Akaiken informaatiokriteerin on huomattu suosivan malleja, joissa on suuri määrä parametreja. Näin mallin valinnan perustuessa Akaiken informaatiokriteeriin, voi valittu malli olla yliparametrisoitu Hurvich ja Tsai (1989). Hurvich ja Tsai (1989) esittivät korjatun version Akaiken informaatiokriteeristä. Korjattu Akaiken informaatiokriteeri määritellään multiplikatiiviselle ARMA-mallille:

$$AICc = -2l(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \sigma^2) + 2\left(\frac{T(p + q + P + Q + 1)}{T - p - q - P - Q - 2}\right).$$

Koska parametrien määrän kasvattaminen suurentaa korjatun Akaiken informaatiokriteerin parametrien määrästä riippuvaa osaa enemmän kuin Akaiken informaatiokriteerin vastaavaa osaa, suosii korjattu versio kriteeristä vähemmän parametreja käyttävää mallia kuin alkuperäinen kriteeri. Mallin valintaan käytetään korjattua Akaiken informaatiokriteeriä.

Kun aikasarjaregressiomalli on sovitettu, täytyy vielä tarkistaa mallin oletusten voimassaolo, ennen kuin voidaan tehdä päätelmiä mallin toimivuudesta. Esitetyt aikasarjamallit olettavat jäännösten z_t olevan keskenään korreloimattomia, mallin ennusteesta riippumattomia ja normaalijakautuneita odotusarvolla 0 ja varianssilla σ^2.

Jäännösten korreloimattomuutta voidaan testata Ljung-Box-Piercen testisuureen avulla (Box et al., 2008, s. 340). Testisuure lasketaan jäännösten aikasarjan z_t ensimmäisen

K autokorrelaatiokertoimen $\rho_z(\tau), \tau = 1, \ldots, K$ avulla. Testisuure $\tilde{Q}$ määritellään kaavalla

$$\tilde{Q} = n(n+2) \sum_{\tau=1}^{K} \frac{\rho_z^2(\tau)}{T-\tau},$$

missä $n = T - d - D$ on jäännösten lukumäärä. Testisuure noudattaa likimain $\chi^2(K-p-q-P-Q-k)$-jakaumaa, mikäli oletus jäännösten korreloimattomuudesta pitää paikkaansa. Muuttuja k on käytettyjen prediktorien määrä. Muuttujan K arvoksi Hyndman ja Athanasopoulos (2013, kpl. 2.6) ehdottavat, että käytetään $K = 10$, mikäli aikasarjassa ei ole kausivaihtelua ja $K = \min(24, T/5)$, jos aikasarjassa on kausivaihtelua. Suuret testisuureen arvot viittaavat mallin epäsopivuuteen.

Jäännösten z_t normaalisuutta voidaan tarkastella vertaamalla jäännösten arvojen jakaumaa normaalijakaumaan kvantiilikuvion (Quantile-Quantile Plot) avulla (Brockwell ja Davis, 2002, s. 38).

Lopullisen mallin valinta tehdään ennusteen tarkkuuden perusteella. Merkitään muuttujalla $\hat{y}_t$ ennustetta hetkelle t. Tällöin ennustevirhe e_t määritellään $e_t = y_t - \hat{y}_t$. Ennustevirheen avulla ennusteen tarkkuutta voidaan laskea monella eri tavalla. Eri tapoja on esitetty mm. lähteessä Hyndman ja Koehler (2006).

Tässä tutkielmassa ennusteen tarkkuuden mittarina käytetään keskimääräistä ennustevirhettä (MAE):

$$MAE = \frac{1}{h} \sum_{i=1}^{h} |e_i|, \tag{3.1}$$

missä h on ennustevälin pituus. Keskimääräinen ennustevirhe valikoitui käytettäväksi sen helpon ymmärrettävyyden takia.

4 Mallintaminen

Mallintamiseen liittyvät laskelmat on tehty R-ympäristön (R Core Team, 2015) paketista `forecast` (Hyndman, 2016) (Hyndman ja Khandakar, 2008) löytyvällä funktiolla `Arima`. Funktiolla pystytään sovittamaan suoraan aikasarjaregressio eri parametrien p, d, q, P, D, Q arvoilla, jotka annetaan funktiolle parametreina. Funktio ottaa automaattisesti d kappaletta perättäisiä ja D kappaletta kausierotuksia sähkönkulutuksesta ja prediktoreista ennen mallin sovittamista. Käytännössä parametrien p, d, q, P, D, Q valinta toteutetaan siten, että käyttäjä valitsee arvot d ja D. Loput arvot valitaan paketin `forecast` funktion `auto.arima` avulla.

4.1 Mallinnuksen toteutus

Koska sähkönkulutus ja osa prediktoreista sisältävät kausivaihtelua ja trendiä, eivät sähkönkulutukset ja prediktorit ole suoraan stationaarisia. Käytetään aikasarjojen stationaariseksi muuttamiseen logaritmimuunnosta, kausierotuksia ja perättäisiä erotuksia. Jotta sarjojen väliset riippuvuudet pysyvät ennallaan, on kausierotukset ja perättäiset erotukset tehtävä kaikille sarjoille. Merkitään alkuperäistä aikasarjaa muuttujalla y_t. Tällöin stationaarinen aikasarja

$$y'_t = \nabla_s^D \nabla^d y_t = (1 - B^s)^D (1 - B)^d y_t,$$

on aikasarja, johon on käytetty kausierotuksia D kertaa ja perättäisiä erotuksia d kertaa. Muuttujien d ja D valinta voidaan tehdä perustuen graafisiin tarkasteluihin, joita käytiin läpi luvussa 3.2.

Sähkönkulutuksille sovitetaan aikasarjaregressio ehdolla prediktorit. Eri prediktorien joukoista halutaan valita paras malli sähkönkulutukselle. Koska mallia on tarkoitus käyttää ennustamiseen, perustetaan prediktorien valinta mallin ennustekykyyn. Mallin ennustekykyä voidaan tutkia usealla eri tavalla. Hyndman ja Athanasopoulos (2013, kpl. 2.5) esittävät seuraavat kaksi tapaa.

Ensimmäisessä tavassa aineisto jaetaan opetusaineistoon ja testiaineistoon. Opetusaineistossa muodostetaan malli, jonka hyvyys ennustamiseen testataan testiaineistossa. Testiaineiston suositeltava koko on vähintään halutun ennusteen pituus h, tässä tapauksessa käytetään kahta vuotta. Mallin valinta perustuu testiaineistoon laskettavan ennusteen ennustevirheeseen liittyvään tunnuslukuun.

Toinen tapa on käyttää ristiinvalidointia. Koska aikasarjan havaitut arvot riippuvat toisistaan, ei havaintoja voida jakaa esimerkiksi viiteen joukkoon mielivaltaisesti, kuten riippumattomien havaintojen tapauksessa voidaan tehdä. Aikasarjan tapauksessa ristiinvalidointi suoritetaan käyttämällä opetusaineistona eripituisia aineistoja aikasarjan alkupäästä. Seuraava opetusaineisto saadaan edellisestä lisäämällä edelliseen opetusaineistoon aikasarjan seuraava havainto. Testiaineistona voidaan käyttää joko koko loppua aikasarjaa opetusaineiston päättymisen jälkeen tai h:ta seuraavaa havaintoa heti opetusaineiston päättymisen jälkeen. Mikäli käytetään h havaintoa

testiaineistona, ja lyhin opetusaineisto on $n \in \mathbb{N}$ pituinen, voidaan ristiinvalidoinnissa käyttää yhteensä $T - n - h + 1$ opetus- ja testiaineistoa. Mallin valinta tehdään yhdistämällä eri testiaineistoista lasketut ennustevirheiden tunnusluvut ja valitsemalla malli yhdistetyn tunnusluvun perusteella.

Tässä tapauksessa päätettiin käyttää opetus- ja testiaineistoa, sillä aineistoa on riittävästi. Myös varsinainen laskenta on huomattavasti nopeampaa tällä tavalla, kun malleja tarvitsee sovittaa vain yhdelle opetusaineistolle ristiinvalidoinnin useamman opetusaineiston sijaan. Prediktorien suuri lukumäärä aiheuttaa joka tapauksessa suuren määrän eri mallien sovittamisia yhdellekin opetusaineistolle, joten mallin löytämiseen menee aikaa. Mikäli malli sovitettaisiin useammalla opetusaineistolla, olisi mallin sovittamisia huomattavasti enemmän, mikä näkyisi suoraan mallintamiseen kuluvassa ajassa. Opetusaineistona on käytetty aineistoa kulutuksista ja prediktoreista joulukuusta 2004 marraskuuhun 2013 ja testiaineistona vastaavia arvoja joulukuusta 2013 marraskuuhun 2015.

4.2 Kvartaaliset ja vuosittaiset prediktorit ja perättäiset erotukset

Mikäli aikasarja y_t ei ole stationaarinen, tai kausittaisten erotusten aikasarja $y'_t = y_t - y_{t-12}$ ei ole stationaarinen, käytetään stationaarisuuden saavuttamiseksi tarvittava määrä perättäisiä erotuksia $y_t - y_{t-1}$. Jotta kulutuksen ja prediktorien välinen suhde ei muuttuisi, voi tällöin tarvita käyttää perättäisiä erotuksia myös prediktoreille.

Kvartaaleittain ja vuosittain havaituilla prediktoreilla ei voida käyttää suoraan perättäisiä erotuksia. Koska kvartaalisia ja vuosittaisia prediktoreita ei havaita joka kuukautena, on osa prediktorien perättäisistä arvoista samoja. Kvartaalisen prediktorin tapauksessa esimerkiksi tammi-, helmi- ja maaliskuun arvot ovat samat. Tällöin kuukausisarjan perättäisistä erotuksista suurin osa olisi nollia ja näin ollen regressiokertoimet vaikuttaisivat kvartaalisten ja vuosittaisten prediktorien tapauksessa vain neljänä tai yhtenä kuukautena vuodessa.

Kvartaalisten prediktorien tapauksessa perättäisten erotusten sijaan käytetään kvartaalisia erotuksia $y_t - y_{t-3}$, joka kuvaa kvartaalien välistä muutosta prediktorin arvossa. Vastaavasti vuosittaisilla prediktoreilla käytetään kausierotusta $y_t - y_{t-12}$.

4.3 Mallin valinta

Oletetaan, että sähkönkulutuksille ja prediktoreille on tehty tarvittavat muunnokset, jotta aikasarjojen vaihtelu ei riipu aikasarjan arvoista. Sähkönkulutukselle sovitetaan luvussa 3.5 esitetty regressiomalli. Laskentaan on käytetty R-ohjelmistoa (R Core Team, 2015).

Aikasarjaregression tapauksessa multiplikatiivisten ARIMA-jäännösten parametrien p, q, P ja Q valinta tuo samanlaisen haasteen kuin tavallisen multiplikatiivisen ARIMA-

mallin tapauksessa. Koska oikeaa parametrien määrää ei tiedetä, on se valittava jonkin kriteerin perusteella. Kuitenkaan jäännösten ARIMA-mallia ei pystytä valitsemaan, jos ei tunneta regressiokertoimia $\beta_1, \ldots, \beta_k$, jotka määrittelevät prediktorien kanssa jäännökset.

Hyndman ja Athanasopoulos (2013, kpl. 9.1) ehdottavat menetelmää, jossa ensin sovitetaan malli, jolla pyritään estimoimaan regressiokertoimet $\beta_1, \ldots, \beta_k$ mahdollisimman hyvin tietämättä jäännösten ARIMA-mallia. Tämä tehdään olettamalla jäännöksille ARIMA$(2,0,0) \times (1,0,0)$-prosessi. Estimoitujen regressiokerrointen ja prediktorien avulla saadaan jäännösten estimaatit, joille valitaan paras ARIMA-malli halutun kriteerin perusteella. Koko regressiomalli sovitetaan vielä uudelleen ehdolla valitut jäännösten ARIMA-mallin parametrit. Lopuksi tarkistetaan mallin sopivuus.

Menetelmä on tarkemmin esitettynä seuraava:

1. Sovitetaan ARIMA$(2,0,0) \times (1,0,0)_{12}$-regressiomalli kulutukselle y'_t ehdolla prediktorit $x'_1, x'_2, \ldots, x'_k$. Toisin sanoen malli on muotoa

$$\begin{aligned} &y'_t = \beta_1 x'_{1,t} + \cdots \beta_k x'_{k,t} + \eta'_t, \text{ missä} \\ &(1 - \Phi_1 B^{12})(1 - \phi_1 B - \phi_2 B^2)\nabla_{12}^D \nabla^d \eta'_t = z_t, \end{aligned}$$

   missä z_t ovat i.i.d $N(0, \sigma^2)$ -jakautuneita.

2. Sovitetaan ARIMA$(p,0,q) \times (P,0,Q)_{12}$ -malli edellisen mallin jäännöksille $\eta'_t = y'_t - \hat{\boldsymbol{\beta}} \boldsymbol{x'_t}$. Malli on siis muotoa

$$\phi_p(B)\Phi_P(B^{12})\nabla_{12}^D \nabla^d \eta'_t = \theta_q(B)\Theta_Q(B^{12}) z_t,$$

   missä z_t ovat i.i.d $N(0, \sigma^2)$ -jakautuneita. Parametrien määrät p, q, P ja Q valitaan korjatun Akaiken informaatiokriteerin perusteella.

3. Sovitetaan ARIMA$(p,0,q) \times (P,0,Q)_{12}$- regressiomalli kulutukselle ehdolla prediktorit.

4. Tarkistetaan edellisen kohdan mallin jäännösten z_t normaalisuus ja korreloimattomuus.

Vaiheet 1. ja 3. sovitetaan paketin `forecast` funktiolla `Arima`. Koska sovitettavien mallien määrä on suuri, tehdään vaihe 2. saman paketin funktiolla `auto.arima`. Funktio sovittaa mallin useilla parametrien p,q,P,Q arvoilla ja valitsee näistä sopivimman vertaamalla mallien korjattuja Akaiken informaatiokriteerejä. Jäännösten malliksi valitaan se, jolle informaatiokriteeri antaa pienimmän arvon. Funktiolla pystyisi valitsemaan myös arvot d ja D, mutta funktio valitsee usein arvon D olemaan nolla, vaikkakin aikasarjassa olisi selvää kausivaihtelua, joten d ja D valinta tehdään itse.

Funktio `auto.arima` sovittaa kausaalisia ja kääntyviä ARMA-malleja. Kausivaihtelua sisältävät mallit ovat hitaampia sovittaa kuin kausivaihtelua sisältämättömät mallit. Funktio käyttää oletuksena askelittaista mallin sovitusta, joka on kuvattu tarkemmin lähteessä Hyndman ja Khandakar (2008, kpl. 3.2). Askelittaisessa menetelmässä

sovitetaan neljä alkumallia, joista valitaan paras korjatun Akaiken informaatiokriteerin perusteella. Tämän jälkeen valittua mallia lähdetään muuttamaan askelittain muuttamalla parametrien p, q, P ja Q arvoja yhdellä johonkin suuntaan. Mikäli jollain muutetulla mallilla saadaan pienempi korjatun Akaiken informaatiokriteerin arvo, otetaan tämä malli uudeksi valituksi malliksi, jonka parametreja muutetaan. Mikäli parametrien muuttamisella ei saada pienennettyä korjattua Akaiken informaatiokriteeriä, algoritmi pysähtyy ja valittu malli on funktion palauttama malli.

Automaattisessa mallin valinnassa saattaa joskus tulla ongelmia. Joskus funktio löytää mallin, jossa jonkin parametrin estimoitu varianssi on alle nollan. Tämä aiheuttaa optimointiin ongelman. Usein ongelma liittyy malliin, jossa jäännösten mallissa on ollut suuri määrä parametreja, eli $p + q + P + Q$ on ollut vähintään kuusi.

Askelittaisen mallin valinnan sijaan on mahdollista sovittaa jäännöksille myös kaikki mahdolliset ARIMA-mallit, joissa parametrien määrä $p + q + P + Q$ on pienempi kuin annettu luku. Kun askelittaisen mallinvalinnan antaman mallin parametrien määrä on vähintään kuusi, on askelittaisen mallinvalinnan malli hylätty. Sen sijaan on sovitettu ARIMA-malli, jossa parametrien määrän summa saa olla korkeintaan viisi. Tällöin suuri osa ongelmista häviää.

Askelittaisen mallin valinnan etu verrattuna valintaan kaikista mahdollisista riittävän vähäparametrisista ARIMA-malleista on sen nopeus. Askelittainen mallin valinta sovittaa usein huomattavasti vähemmän malleja, joten se on luonnollisesti nopeampi. Kaikkien korkeintaan viisi parametria sisältävien mallien sovitus vie aikaa käytetyllä tietokoneella (Intel i5-prosessori, 3.10 GHz) keskimäärin noin 11.7 sekuntia, kun askelittaisella mallinnuksella käyttäen vain tarvittaessa kaikkien korkeintaan viisi parametria sisältävien mallien sovitusta, aikaa kuluu keskimäärin 4.6 sekuntia.

Askelittainen mallin valinta palauttaa osassa tapauksessa korjatulta Akaiken informaatiokriteeriltään huonomman mallin, koska se ei käy kaikkia mahdollisia malleja läpi. Kuitenkaan ero ei ole usein kovin suuri. Täten askelittaista mallinnusta on käytetty tilanteissa, joissa se toimii, koska sen etu nopeudessa on selkeä.

Malli toteutetaan kaikille halutuille prediktorien joukoille erikseen. Paras prediktorien joukko käytettäväksi valitaan vertaamalla mallien ennustekykyä testiaineistossa. Ennustekyvyn kriteerinä käytetään MAE:a, joka määriteltiin kaavassa (3.1).

Prediktorien määrä tuo haasteen mallinnukseen. Prediktorien määrän kasvu lisää mahdollisten prediktorien kombinaatioiden määrää eksponentiaalisesti. Käsitellään pelkästään prediktorien päävaikutuksia, eli miten yksittäinen prediktori vaikuttaa kulutukseen. Esimerkiksi viidellä prediktorilla mahdollisia prediktorien kombinaatioita on $2^5 - 1 = 31$ kappaletta, käytettäessä 25 prediktoria, kombinaatioita on jo $2^{25} - 1 = 33\ 554\ 431$. Laskentaan kuluvan ajan takia hiemankin suuremmalla prediktorien määrällä kaikkia mahdollisia mallivaihtoehtoja ei pystytä kokeilemaan, vaan joudutaan tyytymään tarkastelemaan jotakin kombinaatioiden osajoukkoa, josta paras kombinaatio valitaan.

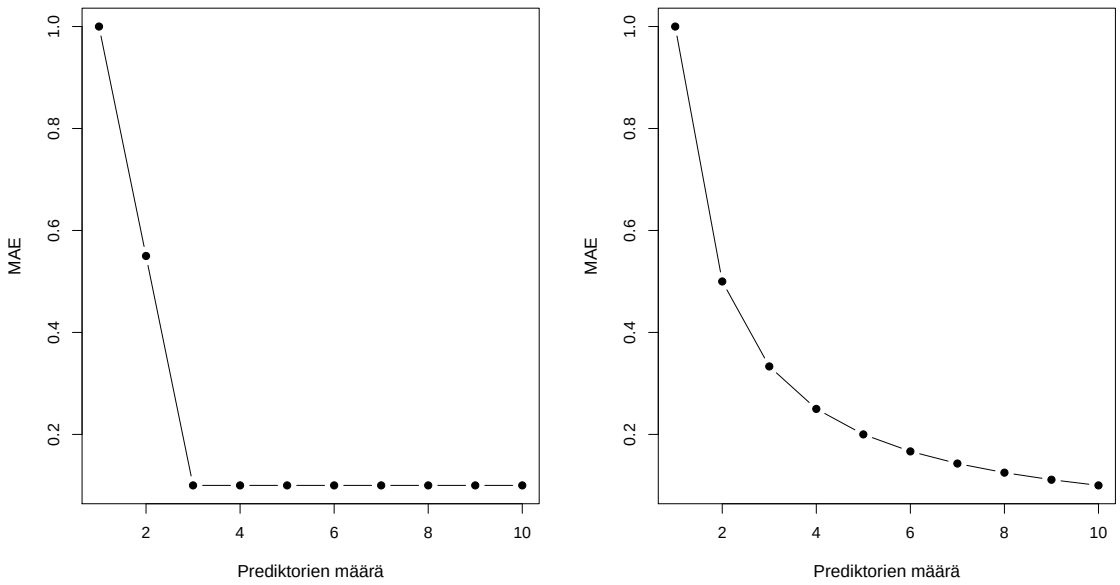
Mikäli prediktoreita on paljon, voivat prediktorien määrä ja prediktorien väliset riippuvuudet aiheuttaa laskentaan ongelmia. Prediktorit voivat olla multikollineaarisia,

jolloin yksi prediktori voidaan esittää muiden prediktorien tai prediktorien osajoukon lineaarikombinaationa. Tämä aiheuttaa kovarianssimatriisin singulaarisuuden, joka johtaa virheeseen laskennassa, koska käänteismatriisia ei voida laskea. Multikollineaarisuudesta päästään eroon esimerkiksi poistamalla yksi lineaarikombinaatioon kuuluvista prediktoreista. Kuitenkin käytettäessä numeerisia menetelmiä käänteismatriisin laskemiseen, saattavat myös lukujen pyöristykset mahdollisesti johtaa matriisin käännettävyysongelmiin, vaikka prediktorit eivät olisikaan keskenään multikollineaarisia.

Matriisin kääntyvyysongelma ilmenee joskus myös varsin pienellä prediktorien määrällä. Ei siis ole täysin varmaa, johtuuko ongelma ainoastaan prediktorien mahdollisesta multikollineaarisuudesta. Ongelmaan saattaisi myös auttaa prediktorien standardisointi keskiarvoon 0 ja varianssiin 1. Tätä ei kuitenkaan ole tehty tässä tutkielmassa, sillä matriisin kääntyvyysongelma on pienillä prediktorien määrällä varsin harvinainen. Standardisoinnin ei pitäisi muuttaa prediktorien riippuvuutta sähkönkulutuksesta. Myös tässä tapauksessa auttaa osassa tilanteista siirtyminen askelittaisesta parametrien p, q, P ja Q valinnasta tilanteeseen, jossa valitaan paras korkeintaan viisi aikasarjaparametria sisältävä malli. Joskus kuitenkin tämäkään ei auta, ja näissä tapauksissa mallia ei ole sovitettu. Virhetilanne on kirjattu ylös, jotta jälkikäteen voidaan katsoa, paljonko sovittamattomia malleja on. Suurimmalla osalla toimialoista kaikki mallit saadaan sovitettua. Joillakin toimialoilla joidenkin prediktorien kombinaatioiden malli jää sovittamatta. Kyse on kuitenkin pienestä osasta kaikista mahdollisista malleista.

Matriisien käännettävyysongelmien takia on kuitenkin helpompi lähteä etsimään prediktorien joukkoa lähtien liikkeelle yhdestä prediktorista ja suurentamalla joukkoa kuin toisinpäin. Käytetään prediktorien valintaan askelittaista prediktorien valintaa lähtien liikkeelle yhdestä prediktorista (Forward-Stepwise Selection) (Hastie et al., 2008, kpl. 3.3.2). Menetelmässä sovitetaan joka vaiheessa malli opetusaineistossa sähkönkulutukselle ehdolla käytettyjen prediktorien arvot. Prediktorien valinnassa lähdetään liikkeelle etsimällä prediktorien joukosta se, joka yksittäin tuo pienimmän ennustevirheen sähkönkulutukselle testiaineistossa, kun ennustamiseen käytetään prediktorien arvoja testiaineistossa. Seuraavassa vaiheessa lisätään vuorotellen vielä mallin ulkopuolella olevat prediktorit ja katsotaan, millä kahden prediktorin kombinaatiolla, joista toinen on 1. vaiheen prediktori, saadaan pienin ennustevirhe. Tätä voidaan jatkaa aina haluttuun k prediktorin kombinaatioon asti.

Menetelmä sovittaa sisäkkäisiä malleja, jossa $n = 1, \ldots, k-1$ prediktorin malliin kuuluneet prediktorit kuuluvat myös $n+1$ prediktorin malliin, lisättynä yhdellä uudella prediktorilla. Todellisuudessa esimerkiksi paras yksittäinen prediktori jonkin kriteerin perusteella ei välttämättä kuulu parhaan kahden prediktorin malliin samalla kriteerillä mitattuna. Kuitenkaan kaikkien mahdollisten k prediktorin kombinaation tutkiminen ei ole mahdollista edes pienillä k:n arvoilla, jos prediktorien määrä on joitakin kymmeniä. Esimerkiksi jos prediktoreita on yhteensä 25 ja $k = 5$, on mahdollisia viiden prediktorin joukkoja yhteensä $\binom{25}{5} = 53130$.



Kuva 8: Esimerkit prediktorien määrän valinnasta MAE:n perusteella. Vasemmanpuoleisessa kuvassa on helppo valita prediktorien määräksi 3. Oikeanpuoleisessa kuvassa prediktorien määrän valinta ei ole yhtä selvä, sillä MAE pienenee koko ajan lisättäessä prediktoreita.

Ennusteen onnistumisesta tietyn kriteerin mielessä voidaan piirtää kuva prediktorien määrän funktiona. Prediktorien sopiva määrä n voidaan valita esimerkiksi kuvasta katsomalla. Mikäli huomataan, ettei ennusteen tarkkuus pienene huomattavasti n prediktorin jälkeen, voisivat kyseiset n prediktoria olla sopivat käytettäväksi ennusteessa.

Esimerkkinä prediktorien määrän valinnasta on kuvan 8 kaksi esimerkkitilannetta. Kuvat eivät ole todellisista tilanteista. Ensimmäisessä tilanteessa MAE on 1, kun käytetään yhtä prediktoria, 0.55 kahdella prediktorilla ja 0.1 kolmella ja sitä useammalla prediktorilla. Kolmen prediktorin jälkeen MAE ei enää laske lisättäessä useampia prediktoreita. Näin ollen kolme on hyvä prediktorien määrä ensimmäisessä tilanteessa.

Toisessa tilanteessa MAE on $1/n$, missä n on prediktorien määrä. Tässä tapauksessa MAE laskee aina lisättäessä prediktoreita. Tässä tapauksessa ei ole aivan yhtä selvää, montaako prediktoria kannattaa käyttää. Valinnan voi tehdä esimerkiksi MAE:n suuruuden perusteella, jolloin valitaan prediktorien määräksi pienin, millä MAE on pienempi kuin haluttu raja. Toisaalta valinnan voi tehdä myös prediktorin tuoman ennustekyvyn paranemisen perusteella. Kun seuraavan prediktorin lisäys ei enää paranna mallin ennustekykyä lisäämättömään malliin verrattuna tarpeeksi, on prediktorien määrä sopiva.

4.4 Ennustaminen

Kun käytettävät prediktorit ja jäännösten aikasarjaprosessin parametrit on valittu ja estimoitu, voidaan mallilla ennustaa tulevaisuuden sähkönkulutusta ehdolla prediktorien arvot. Malli olettaa prediktorien arvot kiinteiksi, joten prediktorien arvoihin liittyvä epävarmuus on tarvittaessa otettava huomioon muilla tavoin. Mikäli prediktorien arvot ennustetaan tulevaisuuteen, tulisi prediktorien arvoihin liittyvä epävarmuus huomioida jotenkin laskettaessa ennusteiden ennustevälejä.

Ennustettava malli on valittu MAE:n perusteella. Malli sovitetaan koko aineistolle ehdolla valitut prediktorit. Lisäksi tarvitaan prediktoreille x_i arvot $x_{i,T+1}, x_{i,T+2}, \ldots x_{i,T+h}$, missä h on ennustevälin pituus. Paketin `forecast` funktiolla `forecast` pystytään laskemaan suoraan ennuste halutulle aikasarjaregressiomallille ja myös ennusteeseen liittyvät ennustevälit ehdolla kiinteät prediktorien arvot.

5 Tulokset

Tarkastellaan tarkemmin prediktorien määrän valintaa kolmella eri toimialalla. Toimialat ovat oikeita toimialoja, mutta niiden nimiä ei mainita tässä työssä, kosja kyseessä on liiketoiminnan kannalta merkittävä tieto. Tuloksista on tehty Jyväskylän Energialle erillinen liite, jota ei esitetä tässä tutkielmassa. Tutkielmassa tarkemmin käsiteltävät toimialat on pyritty valitsemaan siten, että ne vastaisivat mallin toimivuuden suhteen koko toimialojen joukkoa. Taulukossa 5 on verrattu mallien toimintaa MAE:n suhteen myös toimialoilla, jotka eivät ole tarkemman tarkastelun kohteena tässä työssä.

Toimialojen sähkönkulutusta on kerrottu vakiolla siten, että kulutuksen keskiarvoksi joulukuussa 2004 – marraskuussa 2015 on saatu 100. Muuten toimialojen kulutus vastaa kulutuksen todellista käyttäytymistä. Mikäli tarkastellaan logaritmista kulutusta, kuten esimerkkitapausten kuvissa MAE:n kehityksestä, toimenpide ei vaikuta MAE:hen. Toimenpide ei vaikuta myöskään prediktorien määrään, siihen mitkä prediktorit otetaan mukaan malliin tai jäännöksiin liittyviin oletuksiin.

Käydään läpi koko mallinvalintaprosessi tarvittavien perättäisten ja kausierotusten määrien valinnasta aina mallin jäännöksiin liittyviin oletuksiin asti. Mallina käytetään aikasarjaregressiomallia stationaarisille kulutukselle y'_t ja prediktoreille $x'_{i,t}, i = 1, \dots k$,

$$y'_t = \sum_{i=1}^{k} \beta_i x_{i,t} + \eta'_t.$$

Jäännökset η'_t noudattavat multiplikatiivista ARMA$(p, q) \times (P, Q)$ -prosessia.

Alkuperäinen tai logaritminen sähkönkulutus on muunnettu stationaariseksi käyttämällä d kappaletta perättäisiä erotuksia ja D kappaletta kausierotuksia. Aikasarjaregressiomalli saadaan palautettua alkuperäistä kulutusta kuvaavaksi malliksi esittämällä y_t aikaisempien havaintojen ja stationaarisen kulutuksen y'_t ja prediktorien $x_{i,t}, i = 1, \dots, k$ avulla

$$\begin{aligned} y_t &= y'_t - \sum_{i=1}^{d+12D} a_i y_{t-i} \\ &= \eta'_t + \sum_{j=1} \beta_j x'_{j,t} - \sum_{i=1}^{d+12D} a_i y_{t-i}. \end{aligned}$$

Tarkastellaan kahta eri mallia eri prediktoreilla. Mallissa 1 on käytetty prediktoreina prediktorien nykyarvoa ja vuoden takaisia arvoja selittämään sähkönkulutuksen nykyarvoa. Vuoden takaisia arvoja on käytetty vain prediktoreilta, joiden vaikutuksen uskottiin voivan olla viiveellistä. Nämä prediktorit on merkitty taulukossa 4 tähdellä. Mallissa 2 on käytetty prediktoreina vain näiden viiveellisten prediktorien kahden ja kolmen vuoden takaisia arvoja. Näiden lisäksi otetaan vertailumalliksi tavallinen multiplikatiivinen ARIMA-malli kulutukselle y_t

$$\Phi(B^{12})\phi(B)\nabla_{12}^{D}\nabla^{d} y_t = \Theta(B^{12})\theta(B) z_t,$$

missä z_t on puhdas satunnaisprosessi. Multiplikatiivinen malli on vastaava malli kuin aikasarjaregressiomalli, mutta ilman prediktoreita. Näin ollen se on hyvä vertailumalli sille, onko prediktoreista hyötyä sähkönkulutuksen ennustetarkkuuteen.

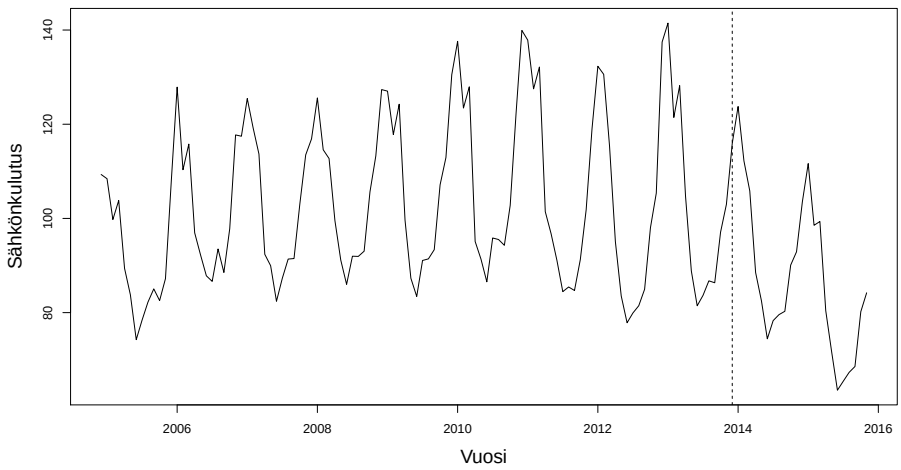
Parhaan multiplikatiivisen ARIMA-mallin valinta suoritetaan lähes vastaavasti kuin prediktoreita sisältävän mallin valinta. Malli sovitetaan eri p, q, P ja Q arvoilla opetusaineistossa ja sillä ennustetaan arvot testiaineistoon. Paras malli on se, jolle MAE on testiaineistossa pienin. Aikasarjaregression tapauksessa parametrien määrät p, q, P ja Q valittiin käyttämällä korjattua Akaiken informaatiokriteeriä, jotta sovitettujen mallien määrä pysyisi laskennallisesti järkevänä. Multiplikatiivisten ARIMA-mallien tapauksessa tätä rajoitusta ei tehdä. Sovitettuja malleja tulee laskennallisesti järkevä määrä, vaikka sovitetaankin kaikki halutut multiplikatiiviset ARIMA-mallit. Tässä on sovitettu kaikki mahdolliset multiplikatiiviset ARIMA-mallit, joille parametrien määrien summa $p + q + P + Q$ on pääasiassa korkeintaan seitsemän. Seitsemän parametrin jälkeen laskenta-aika alkaa kasvaa huomattavasti, eikä parametrien lisäys yhdeksään ainakaan esimerkkitapauksissa enää paranna MAE:ta. Toimialoilla 1,2,3,4 ja 11 on sovitettu multiplikatiivinen malli yhdeksään parametriin asti. Toimialoille 2,4 ja 11, koska seitsemällä parametrilla saatiin pienin MAE, toimialoille 1 ja 3, jotta vertailukuvista saatiin samanlaiset tarkemmin tarkastelluilla toimialoilla. Toimialoilla 4 ja 11 saatiin pienempi MAE käyttämällä yli seitsemää parametria.

Kaikkiaan prediktorien valinnan prosessi kestää 15:een prediktoriin asti valittaessa 50:llä prediktorilla noin tunnin. Tuloksissa vertailumallina käytetyn multiplikatiivisen ARIMA-mallin sovittaminen korkeintaan seitsemällä parametrilla kestää 15-20 minuuttia. Kahdeksas ja yhdeksäs parametri vievät aikaa huomattavasti kauemmin, jolloin yhteisaika yhdeksälle parametrille on noin 1,5-2 tuntia. Multiplikatiivisen ARIMA-mallin tapauksessa valinnan voisi tehdä myös esimerkiksi tarkastelemalla autokovarianssifunktion arvoja (Box et al., 2008, kpl. 9.3.2). Tässä tapauksessa parhaan mallin valinta on pyritty tekemään mahdollisimman samalla tavalla kuin prediktorien malleille käyttämällä opetus- ja testiaineistoa.

Kahden prediktoreita sisältävän mallin valinta ja niiden ero liittyy sähkönkulutuksen ennustamiseen ja mallien käyttöön ennustettaessa. Mallin 1 prediktoreille tarvitaan ennusteet seuraavalle yhdelle tai kahdelle vuodelle, kun halutaan ennustaa sähkönkulutusta seuraavalle kahdelle vuodelle. Mallin 2 prediktorien arvot seuraavalle kahdelle vuodelle saadaan jo havaituista, joulukuun 2012 – marraskuun 2015 arvoista, riippuen onko prediktorin kahden vai kolmen vuoden takaisella arvolla vaikutusta sähkönkulutuksen nykyarvoon. Näin ollen mallin 2 prediktorien arvoihin testiaineistossa ei liity epävarmuutta.

Multiplikatiivista ARIMA-mallia pystytään myös ennustamaan pelkästään siihen mennessä havaitun sähkönkulutuksen aikasarjan y_t perusteella. Multiplikatiivisen ARIMA-mallin ennusteet saadaan laskettua kuten luvussa 3.7.3.

Tarkastellaan ensiksi toimialaa 1. Toimialan sähkönkulutuksen aikasarja on kokonaisuudessaan esitetty kuvassa 9. Kulutus on varsin tasaista koko aikajakson aikana, mutta viimeisen kahden vuoden aikana kulutus on kääntynyt hienoiseen laskuun. Tämä lasku on tapahtunut kokonaan testiaineiston aikana. Onkin mielenkiintoista



Kuva 9: Toimialan 1 sähkönkulutuksen aikasarja. Pystyviiva merkitsee testiaineiston alkua.

selvittää, osaavatko mallien antamat ennusteet ottaa huomioon kulutuksen laskun, ja miten suuret ennustevälit mallien ennusteisiin testiaineistossa liittyvät.

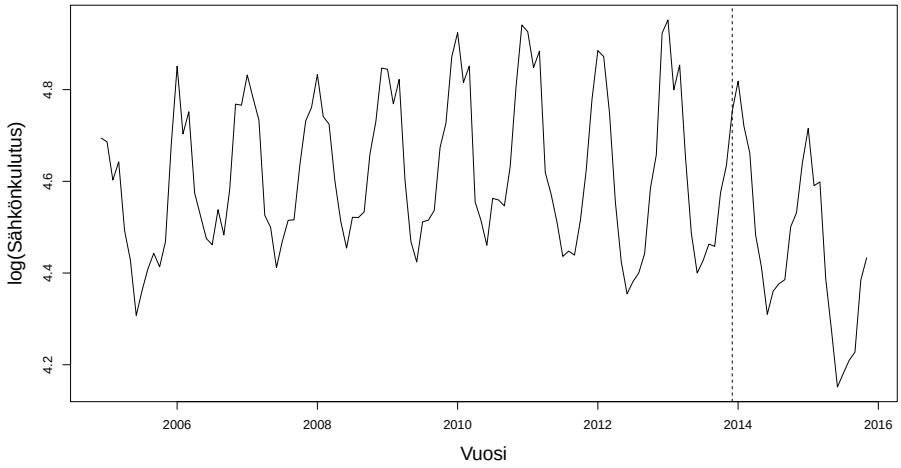
Kulutuksessa on näkyvissä hieman suurempaa kausivaihtelua vuosina 2010-2013, kun kulutuksen trendi on korkeimmillaan. Tämän takia tehdään kulutukselle logaritmimuunnos, joka tasoitaa kausivaihtelua. Kuvassa 10 nähdään, että logaritmisen kulutuksen $\log(y_t)$ kausivaihtelu on varsin tasaista riippumatta aikasarjan tasosta. Kulutuksessa on myös näkyvissä selkeää kausivaihtelua ja trendiä, joten aikasarjan stationaarisuusoletus ei ole suoraan voimassa. Lasketaan kausierotukset, eli tarkastellaan sarjaa $(1 - B^{12}) \log(y_t) = \log(y_t) - \log(y_{t-12})$, joka on esitetty kuvassa 11. Tässä ja jatkossa tarkastellessa tarvittavien perättäisten erotusten ja kausierotusten määrää, perustetaan päätelmät ennen joulukuuta 2013 havaittuun osaan aikasarjasta. Joulukuu 2013 on merkitty katkoviivalla kuviin. Koska mallin valinta perustuu opetusaineistoon, ei testiaineisto ole käytettävissä mallin valintaan edes tarvittavien perättäisten erotusten ja kausierotusten määrien valinnassa. Tämä vastaa myös todellista tilannetta ennustettaessa kulutusta tulevaisuuteen, kun tulevaisuuden havainnoista ei ole tietoa.

Koska kausierotusten aikasarjan arvot laskevat hieman ajan kuluessa, lasketaan kausierotusten sarjalle vielä perättäiset erotukset, eli tarkastellaan aikasarjaa

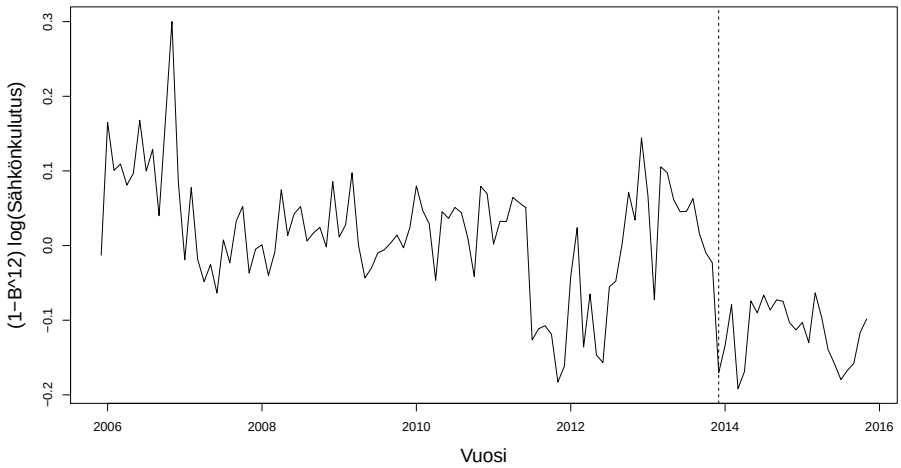
$$(1 - B)(1 - B^{12}) \log(y_t) = \log(y_t) - \log(y_{t-1}) - \log(y_{t-12}) + \log(y_{t-13}).$$

Aikasarja on esitetty kuvassa 12. Tässä aikasarjassa ei ole näkyvissä selvää poikkeavuutta stationaarisuudesta, joten valitaan käytettäväksi perättäisten ja kausierotusten määriksi $d = 1, D = 1$.

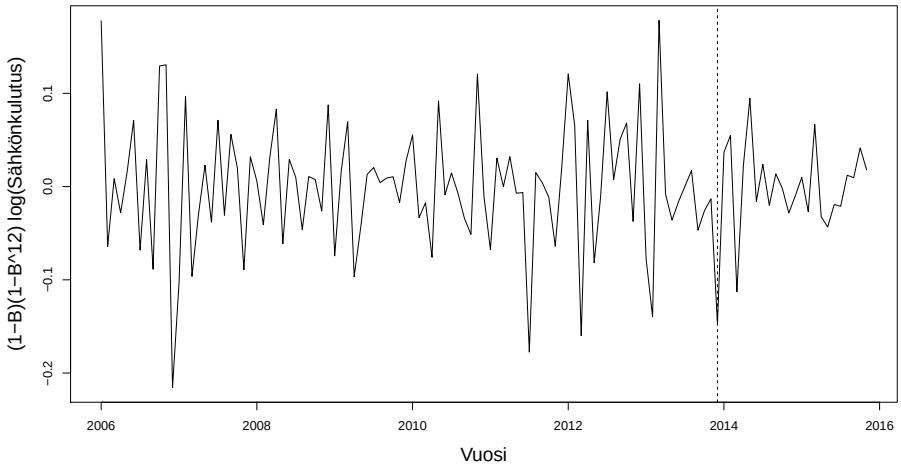
Seuraavaksi tarkastellaan mallin valintaa prediktorien määrän suhteen piirtämällä kuvat MAE:n kehityksestä, kun aikasarjaregressiomalliin lisätään yksi kerrallaan prediktoreita, samaan tapaan kuin esitettiin kappaleessa 4.3. Kuvat MAE:sta prediktorien



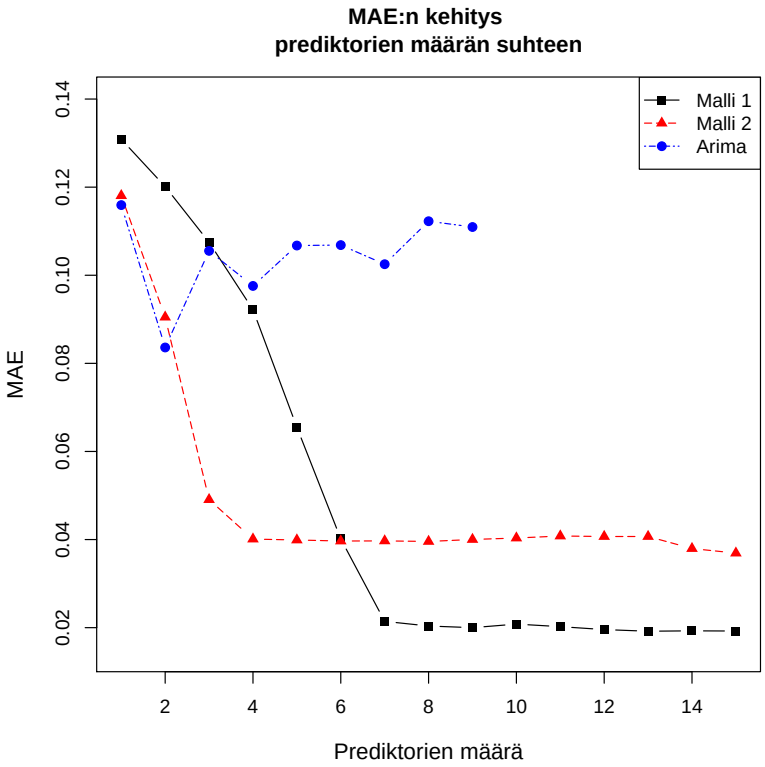
Kuva 10: Toimialan 1 logaritmisen sähkönkulutuksen aikasarja. Pystyviiva merkitsee testiaineiston alkua.



Kuva 11: Toimialan 1 kausierotusten aikasarja $\log(y_t) - \log(y_{t-12})$. Pystyviiva merkitsee testiaineiston alkua.



Kuva 12: Toimialan 1 aikasarja, josta on otettu kausierotukset ja perättäiset erotukset, $\log(y_t) - \log(y_{t-1}) - \log(y_{t-12}) + \log(y_{t-13})$. Pystyviiva merkitsee testiaineiston alkua.



Kuva 13: MAE:n kehitys prediktorien määrän funktiona toimialalla 1, kun käytetään prediktoreina prediktorien nykyarvoja ja vuoden takaisia arvoja (malli 1) ja kahden ja kolmen vuoden takaisia arvoja (malli 2). Lisäksi vertailumallina käytetty multiplikatiivinen ARIMA-malli, jossa prediktorien määrä = käytettyjen parametrien määrä.

määrän suhteen on esitetty kuvassa 13. Multiplikatiivisen ARIMA-mallin tapauksessa kuvassa prediktorien määrällä tarkoitetaan käytettyjen parametrien määrää $p + q + P + Q$.

Kuvassa 13 nähdään, että molemmilla prediktoreita sisältävillä malleilla prediktorien lisääminen pienentää aluksi MAE:ta. Kun prediktoreita on lisätty mallissa 1 seitsemän ja mallissa 2 neljä kappaletta, ei MAE enää tämä jälkeen pienene yhtä selvästi kuin pienemmillä prediktorien määrällä. Tästä eteenpäin prediktorien lisääminen ei auta enää selvästi ennustevirheen pienenemiseen. Näin ollen valitaan käytettäviksi prediktorien määriksi mallissa 1 seitsemän ja mallissa 2 neljä. Multiplikatiivisella ARIMA-mallilla pienin MAE saavutetaan kahdella parametrilla. ARIMA-mallin MAE on tässä tapauksessa noin 0.08.

Nähdään myös, että mallilla 1 saadaan pienempi keskimääräinen ennustevirhe kuin mallilla 2. Mallissa 1 MAE on noin 0.02, kun käytetään seitsemää tai useampaa prediktoria. Mallissa 2 MAE on noin 0.04, kun käytetään neljää tai useampaa prediktoria. Näin ollen mallin 1 ennustetarkkuus testiaineistossa on ollut parempi. Huomataan kuitenkin, että mallissa 1 testiaineistossa prediktorien arvoina on käytetty prediktorien jo havaittuja arvoja joulukuusta 2013 marraskuuhun 2015, jotka eivät todellisuudessa olisi tiedossa ennustetta muodostettaessa. Näin ollen mallin antama ennustetarkkuus vastaa todellisuutta vain siinä tapauksessa, että prediktorit olisi ennustettu täsmälleen oikein kyseiselle aikavälille. Kuitenkin valittaessa parhaita prediktoreita malliin tulee käyttää prediktorien havaittuja arvoja eikä esimerkiksi ennustaa niitä prediktorien opetusaineistossa havaittujen arvojen perusteella. Jos prediktorien arvot testiaineistossa ennustettaisiin, ennusteeseen liittyvät virheet saattaisivat muuttaa prediktorien riippuvuussuhdetta kulutukseen testiaineistossa. Tällöin joissakin tapauksissa prediktoreiksi saatettaisiin valita eri prediktorit, kuin mitkä testiaineistossa havaittujen prediktorien arvojen perusteella on valittu. Näin ollen vaikka mallin 1 ennuste testiaineistossa ei vastaa todellista ennustetta, kelpaa laskettu ennustevirhe prediktorien valintaan käytettäessä mallin 1 prediktoreita.

Muutenkin on tärkeä erottaa toisistaan prediktorien valinta ja valittujen prediktorien mallin antaman ennusteen tarkkuus uudessa aineistossa. Tässä tapauksessa käytettävät prediktorit valitaan niiden testiaineiston ennustekyvyn perusteella. Äkkiseltään voisi ajatella, että menetelmällä saadaan laskettua samalla ennustetarkkuuden estimaatti uudessa aineistossa. Kuitenkaan testiaineistossa lasketut MAE:n arvot eri prediktoreilla muodostetuille malleille eivät ole mallien ennustetarkkuuden estimaatteja uudessa aineistossa. Tämä johtuu siitä, että malleja on sovitettu testiaineistolle useita, ja mallin valinta perustuu ennustetarkkuuteen. Valitun mallin ennustevirhe testiaineistossa on usein alhaisempi, kuin mallin ennustevirhe uudessa aineistossa olisi (Hastie et al., 2008, kpl. 7.4). Tässä työssä keskitytään valitsemaan malli, jolla tulevia kulutuksia voidaan ennustaa, mutta ei oteta kantaa, paljonko valitun mallin keskimääräinen ennustevirhe esimerkiksi seuraavalle kahdelle vuodelle on.

Mikäli halutaan estimoida parhaan mallin keskimääräinen ennustevirhe aineistossa, tulisi aineisto jakaa kolmeen osaan. Ensimmäiset kaksi osaa ovat opetus- ja testiaineisto, jotka vastaavat tässä tutkielmassa käytettyjä opetus- ja testiaineistoa. Kolmas

osa on niin kutsuttu validointiaineisto, jona voisi käyttää kahta seuraavaa vuotta testiaineiston loppumisen jälkeen (Hastie et al., 2008). Kun käytettävä malli on valittu, voidaan keskimääräistä ennustevirhettä estimoida validointiaineistossa tekemällä ennuste validointiaineistolle samalla tapaa kuin testiaineistolle tässä tutkielmassa, käyttäen opetusaineistona yhdistettyä opetus- ja testiaineistoa ja prediktoreina valittuja prediktoreita.

Tässä tapauksessa mielenkiinnon kohde oli kuitenkin parhaan ennustemallin valinta. Tähän opetus- ja testiaineisto riittää hyvin. Kääntöpuolena keskimääräiselle ennustevirheelle ei saada estimaattia. Aineistoa ei kuitenkaan haluta pilkkoa kolmeen osaan, sillä tällöin opetusaineiston pituus olisi vain seitsemän vuotta, joka on varsin lyhyt aika. Nyt käytettäessä vain opetus- ja testiaineistoa, saadaan opetusaineiston pituudeksi yhdeksän vuotta.

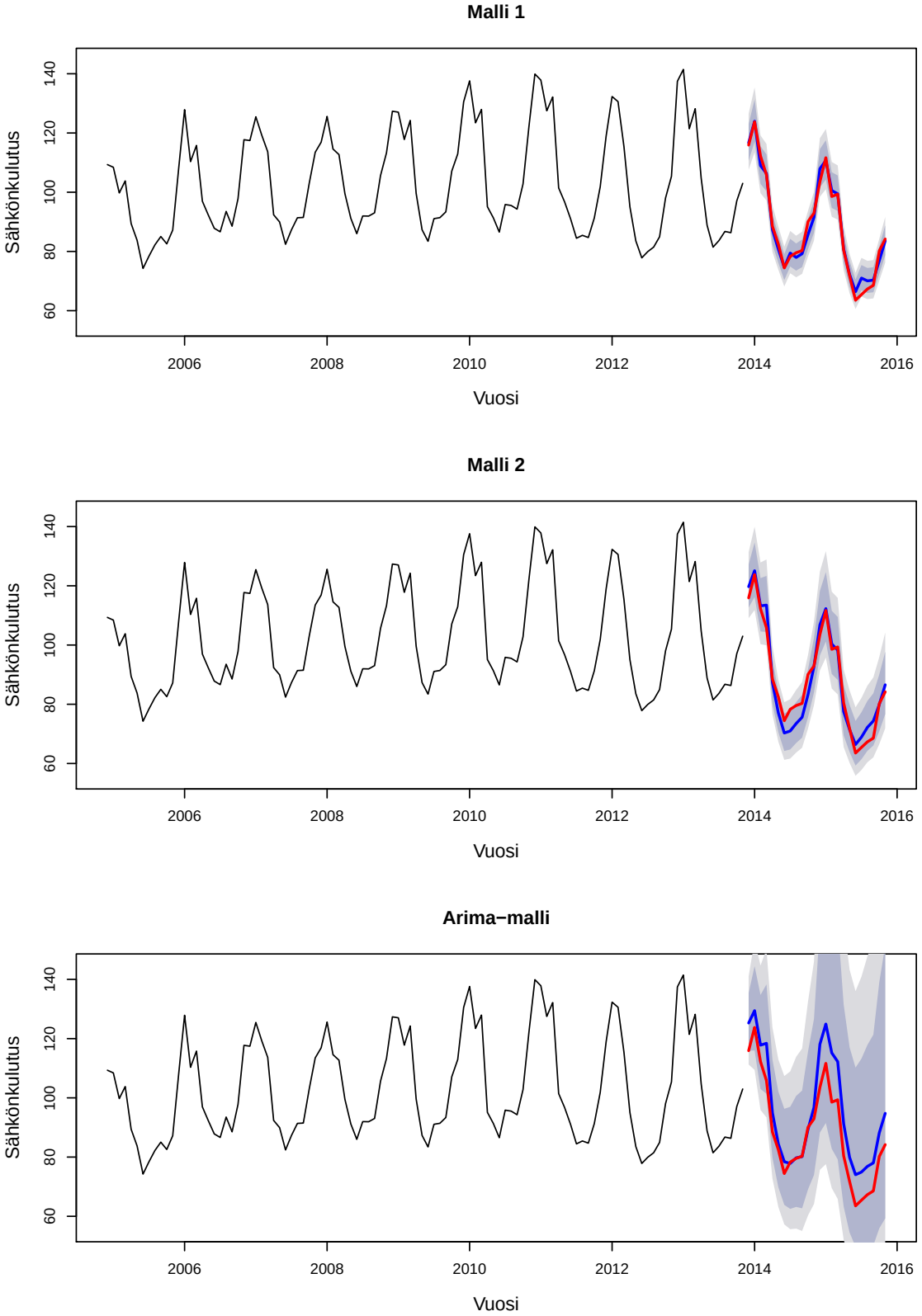
Tarkastellaan seuraavaksi ennustetta testiaineistossa tarkemmin. Prediktorien valintaan käytetty MAE mittaa kulutuksen ennusteen etäisyyttä havaitusta sähkönkulutuksesta. Se ei kuitenkaan esimerkiksi ota huomioon, onko ennusteen poikkeamissa johdonmukaisuutta, kuten ovatko poikkeamat aina ylös- tai alaspäin. MAE ei myöskään ota kantaa mallin sopivuuteen opetusaineistoon.

Kuvassa 14 on piirretty ennusteet testiaineistossa molempien mallien tapauksessa. Nähdään, että mallin 1 ennustevälit ovat huomattavasti pienemmät kuin mallin 2. Näin ollen malli 1 sopii opetusaineistoon paremmin. Tarkasteltaessa ennusteen virhettä testiaineistossa, kummassakaan ennusteessa ei näy systemaattisuutta ennustevirheessä, vaan ennusteet ovat molemmilla puolilla oikeaa kulutusta varsin tasaisesti. Molemmilla ennusteilla vuoden 2015 kesän ennuste on hieman liian alhainen.

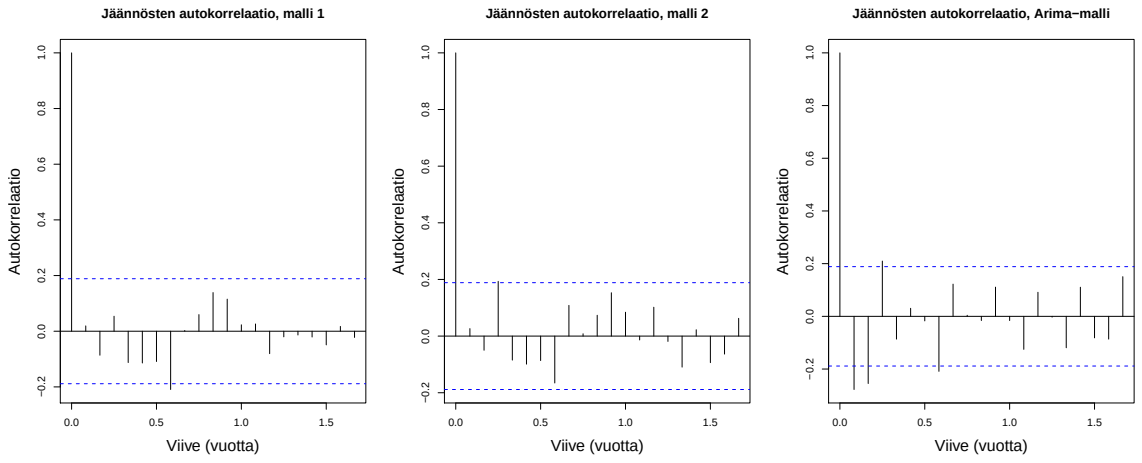
Multiplikatiivisen ARIMA-mallin ennusteet ovat tässä tapauksessa hieman muita huonommat. Lisäksi ennuste on systemaattisesti korkeampi kuin oikea kulutus. Tämä ei kuitenkaan ole yllättävää, sillä kyseisellä toimialalla kulutus on laskenut selvästi kahden viimeisen vuoden aikana ja kulutuksen lasku on kokonaan testiaineistossa. Tämä onkin ARIMA-mallille hankala ennustaa.

Myös ennustevälit ovat huomattavasti suuremmat käytettäessä multiplikatiivista ARIMA-mallia. Tämä johtunee siitä, että valitussa mallissa on käytössä vain kaksi parametria. Täten malli ei sovi opetusaineistoon yhtä hyvin kuin aikasarjaregressiomallit.

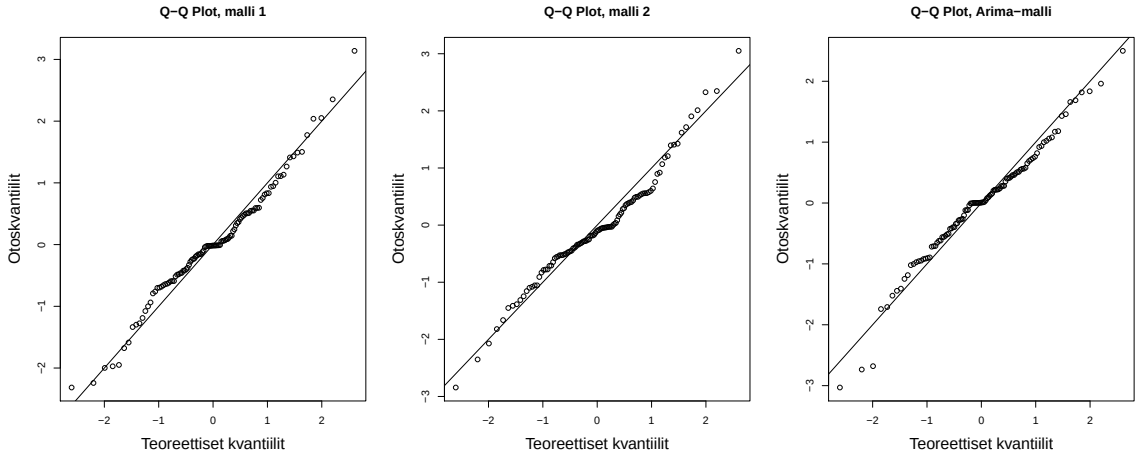
Tarkastellaan vielä jäännösten z_t korreloimattomuus- ja normaalijakautuneisuusoletuksia opetusaineistossa. Jäännösten autokorrelaatiokuvat ja kvantiilikuviot on piirretty kuvissa 15 ja 16. Jäännösten ollessa korreloimattomia, tulisi autokorrelaatioiden $\rho(\tau)$ olla likimain nollia. Kuvaan on piirretty 95 % ennustevälit autokorrelaatioille, mikäli oletus jäännösten korreloimattomuudesta on voimassa. Tässä tapauksessa autokorrelaatiot ovat prediktorien malleissa viiveillä $\tau = 1, \ldots, 20$ ennustevälien sisällä. Ljung-Box-Piercen testin mukaan testisuureen arvot ovat vertailujakauman suuremmasta päästä malleilla 1 ja 2 (malli 1: $\hat{Q} = 16,61, df = 10, p = 0.083$, malli 2: $\hat{Q} = 21.68, df = 13, p = 0.061$). Multiplikatiivisen ARIMA-mallin osalta neljä autokorrelaatiota on suurempia kuin luottamusväli $\pm 1.96\sqrt{n}$, missä n on jäännösten



Kuva 14: Mallien antamat ennusteet toimialalle 1 testiaineistossa. Sininen viiva on piste-ennuste kulutukselle. Lisäksi on piirretty 80 ja 95 % ennustevälit. Punainen viiva on testiaineistossa havaittu kulutus.



Kuva 15: Jäännösten z_t autokorrelaatio viiveillä 0-20 kuukautta toimialalla 1.

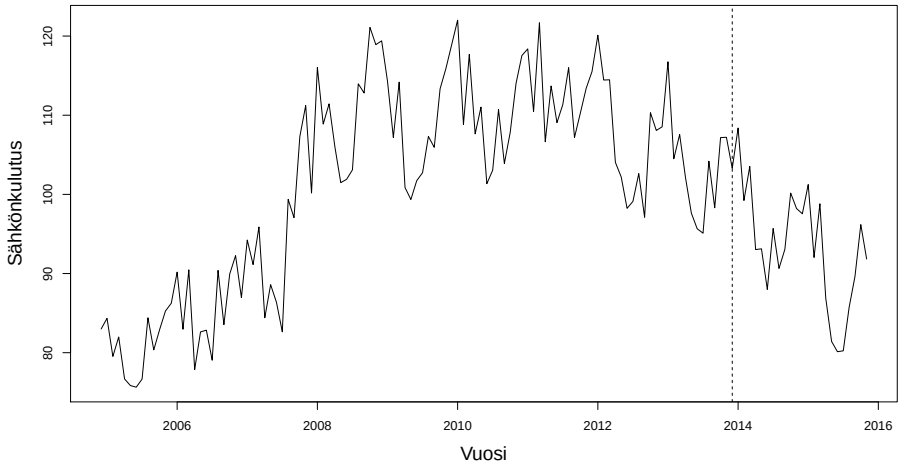


Kuva 16: Standardoitujen jäännösten kvantiilikuviot eri malleilla toimialalla 1.

lukumäärä (Brockwell ja Davis, 2002, kpl. 1.6). Tämä viittaa siihen, että mallin jäännökset eivät ole korreloimattomia. Myös Ljung-Box-Piercen testin mukaan testisuureen arvo on varsin suuri jakaumaoletukseen nähden ($\hat{Q} = 41.69, df = 18, p = 0.001$). Näin ollen mallin sopivuus aineistoon ei ole selvää.

Jäännösten normaalijakautuneisuuden tutkimiseksi piirretään kvantiilikuviot standardoiduista jäännöksistä $z_t/\hat{\sigma}_z$. Oletuksen perusteella järjestettyjen jäännösten $z_{(1)}, \ldots, z_{(n)}$ ja jäännösten teoreettisten odotusarvojen avulla piirretyn hajontakuvion pisteiden tulisi olla likimain suoralla $y = x$. Kuvassa 16 suora $y = x$ on piirretty havainnollistamaan jäännösten eroa suorasta. Kaikissa tapauksissa jäännökset asettuvat varsin hyvin suoralle. Näin ollen myöskään oletusta mallien jäännösten normaalijakautuneisuudesta ei tarvitse kyseenalaistaa.

Tarkastellaan seuraavaksi toimialaa 2. Toimialan 2 aikasarja on piirretty kuvassa 17. Myös tässä tapauksessa näyttää siltä, että kulutuksen vaihtelu on suurempaa vuosina



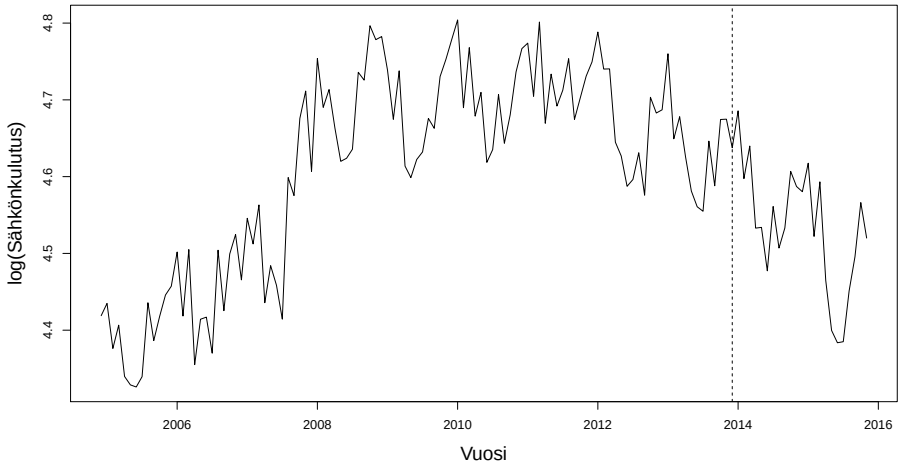
Kuva 17: Toimialan 2 sähkönkulutuksen aikasarja. Pystyviiva merkitsee testiaineiston alkua.

2008–2012, kun kulutus on muutenkin suurempaa. Näin ollen käytetään analyysissä logaritmista kulutusta, joka on esitetty kuvassa 18.

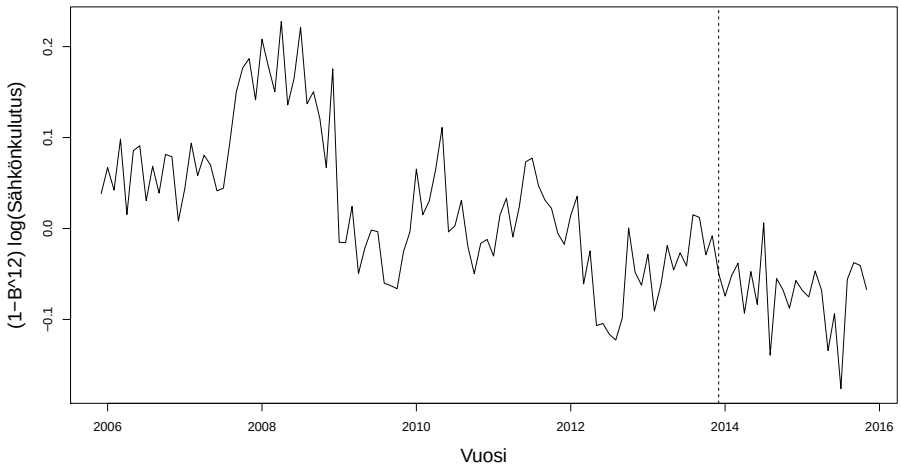
Kulutuksessa on havaittavissa selkeähkö kausittainen vaihtelu ja trendi, joten sähkönkulutuksen aikasarja ei itsessään ole stationaarinen. Koska kulutuksessa on kausivaihtelua, tarkastellaan ensin kausierotusten aikasarjaa $\log(y_t) - \log(y_{t-12})$, joka on esitetty kuvassa 19. Kausierotusten sarjassa on nähtävissä selkeä trendi, joten otetaan tässäkin tapauksessa perättäiset erotukset. Aikasarja, josta on otettu sekä perättäiset että kausierotukset, on esitetty kuvassa 20. Perättäisten ja kausierotusten aikasarja $\log(y_t) - \log(y_{t-1}) - \log(y_{t-12}) + \log(y_{t-13})$ ei osoita selkeää poikkeamaa stationaarisuudesta trendin tai kausivaihtelun suhteen. Näin ollen valitaan käytetäviksi perättäisten ja kausierotusten määriksi toimialalla 2 $d = 1$ ja $D = 1$.

Toimialalla 2 prediktorien määrän valinta ei ole aivan yhtä selkeä kuin toimialalla 1. Mallin 1 tapauksessa valittiin käytettäväksi viittä ja mallin 2 tapauksessa kuutta prediktoria. Kuten kuvasta 21 nähdään, mallin 2 tapauksessa MAE pienenee koko ajan prediktorien määrän kasvaessa. Ei ole selvää, tulisiko valita kuuden sijasta esimerkiksi kymmenen tai 15 prediktoria, koska MAE laskee koko ajan prediktorien määrää lisättäessä. Tällä kertaa päädyttiin kuuteen, koska kuuden prediktorin jälkeen MAE:n laskee tämän jälkeen yhden prediktorin lisäämisellä vähemmän kuin tätä aiemmin. ARIMA-mallin tapauksessa valitaan parametrien määräksi seitsemän, sillä MAE vähenee suhteellisen tasaisesti seitsemään parametriin asti.

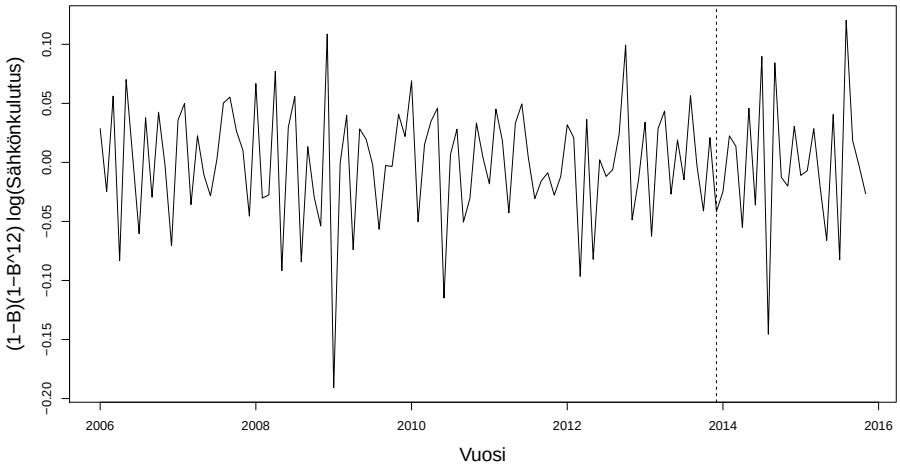
Ennusteet testiaineistossa on esitetty kuvassa 22. Prediktorien mallien ennusteet osuvat jälleen hyvin kohdalleen, mallissa 2 ennusteet ovat systemaattisesti hieman liian suuria. Myös multiplikatiivisen ARIMA-mallin ennusteet ovat tässäkin tapauksessa hieman liian suuria. Myös tässä tapauksessa kulutus on laskenut, mutta erona toimialaan 1 laskua on ollut myös opetusaineiston lopussa. Tässä tapauksessa multiplikatiivisen



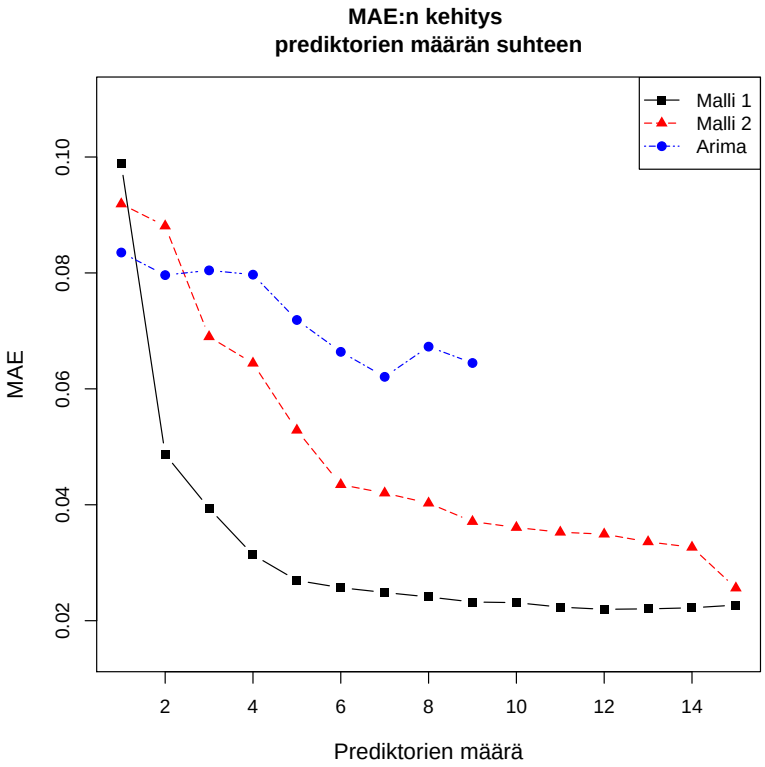
Kuva 18: Toimialan 2 logaritmisen sähkönkulutuksen aikasarja. Pystyviiva merkitsee testiaineiston alkua.



Kuva 19: Toimialan 2 kausierotusten aikasarja $\log(y_t) - \log(y_{t-12})$. Pystyviiva merkitsee testiaineiston alkua.



Kuva 20: Toimialan 2 aikasarja, josta on otettu kausierotukset ja perättäiset erotukset, $\log(y_t) - \log(y_{t-1}) - \log(y_{t-12}) + \log(y_{t-13})$. Pystyviiva merkitsee testiaineiston alkua.



Kuva 21: MAE:n kehitys prediktorien määrän funktiona toimialalla 2, kun käytetään prediktoreina prediktorien nykyarvoja ja vuoden takaisia arvoja (malli 1) ja kahden ja kolmen vuoden takaisia arvoja (malli 2). Lisäksi vertailumallina käytetty multiplikatiivinen ARIMA-malli, jossa prediktorien määrä = käytettyjen parametrien määrä.

ARIMA-mallin ennustevälit ovat jo lähes samankokoiset kuin mallissa 2. Mallin 1 ennustevälit ovat odotusten mukaisesti hieman kapeammat.

Jäännöksissä on nähtävissä hieman poikkeavuutta normaalijakautuneisuudesta. Kuvassa 23 on piirretty jäännösten kvantiilikuviot. Varsinkin mallissa 1 ja ARIMA-mallissa absoluuttisesti suurimmat jäännökset eivät osu aivan suoralle $y = x$. Molemmissa malleissa varsinkin suurimmat ja positiiviset jäännökset ovat hieman liian suuria normaalijakautuneisuusoletukseen nähden. Kuitenkin muuten jäännökset asettuvat melko hyvin suoralle.

Kuvassa 24 nähdään mallien jäännösten autokorrelaatiot viiveillä 0-20. Mallissa 1 kolme autokorrelaatiota on suurempaa kuin 95 % luottamusväli autokorrelaatioille on. Mallin 1 jäännösten korreloimattomuus ei ole selvää tarkasteltaessa autokorrelaatioita. Mallissa 2 ja ARIMA-mallissa autokorrelaatiot näyttäisivät asettuvan hyvin luottamusvälien sisään.

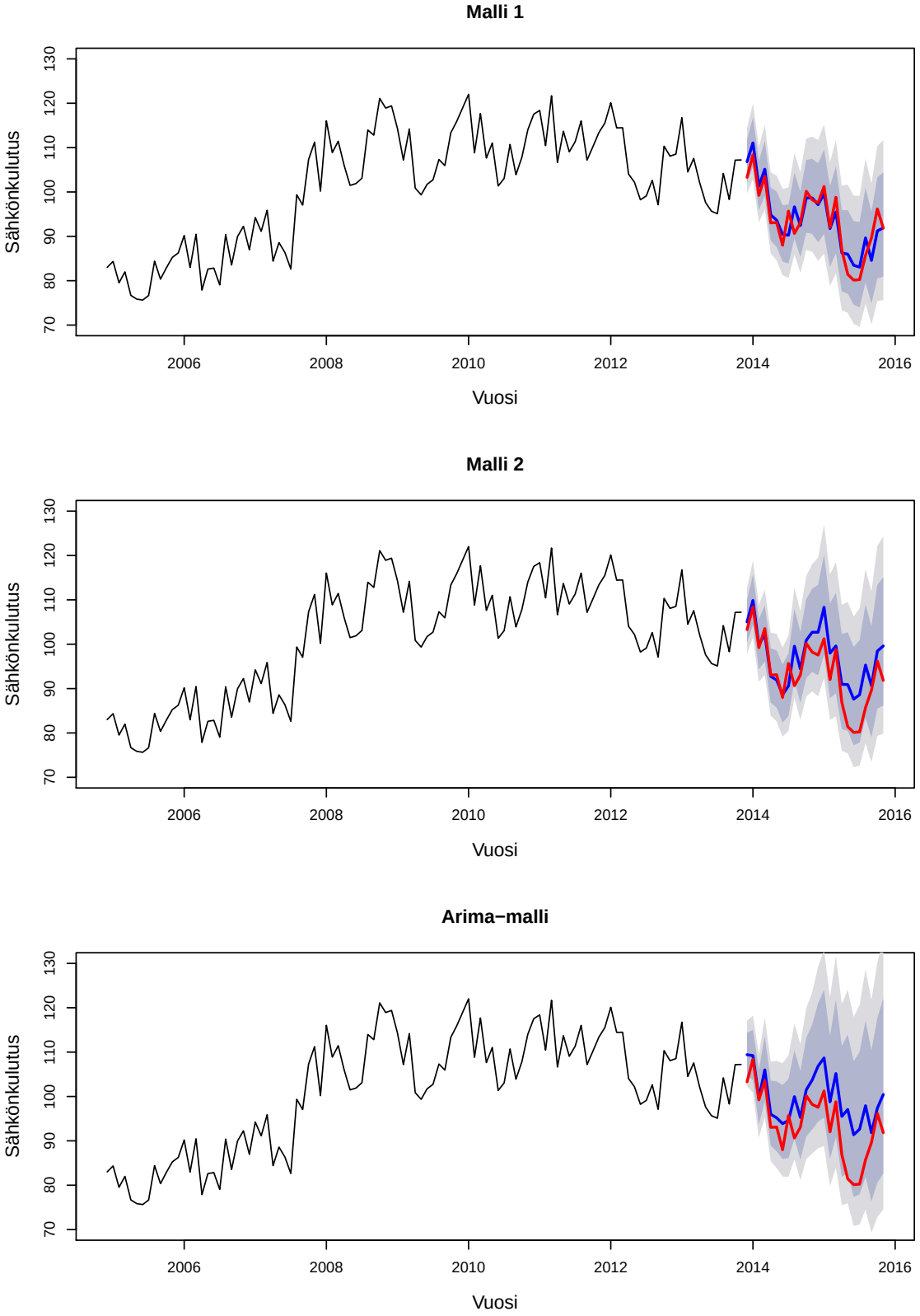
Ljung-Box-Piercen testin mukaan mallin 1 jäännösten autokorrelaatiot poikkeavat selvästi riippumattomuusoletuksesta ($\hat{Q} = 31.02, df = 13, p = 0.003$). Mallin 2 ja ARIMA-mallin tapauksessa testisuureen arvot ovat todennäköisemmin jakaumastaan (malli 2: $\hat{Q} = 16.02, df = 12, p = 0.19$, ARIMA-malli: $\hat{Q} = 22.00, df = 13, p = 0.055$).

Tarkastellaan vielä kolmatta toimialaa. Toimialan 3 sähkönkulutus on esitetty kuvassa 25. Tässäkin tapauksessa analyysiin on käytetty logaritmista sähkönkulutusta, mutta logaritmimuunnoksen tarve ei ole yhtä selkeä kuin aikaisemmissa tapauksissa. Alkuperäinenkin kulutus voisi olla analyysiin sopiva. Logaritminen kulutus on esitettu kuvassa 26.

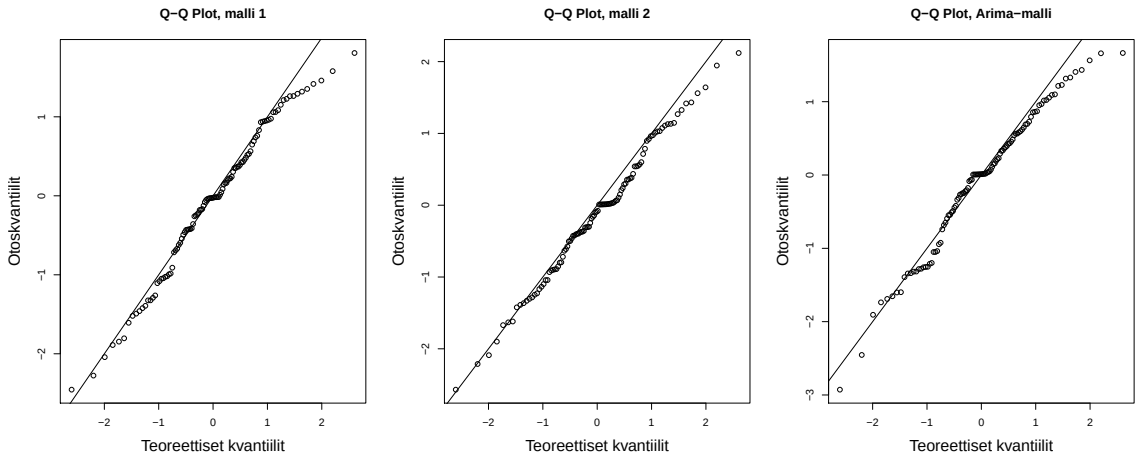
Myöskään kulutuksen kausittainen vaihtelu ei ole yhtä selkeää kuin kahdessa aikaisemmassa tapauksessa. Kuitenkin sarjassa on hienoinen kasvu ajan kuluessa, ja jonkinlaista kausivaihteluakin on, joten aikasarja ei ole suoraan stationaarinen. Tässäkin tapauksessa päädytään perättäisille ja kausierotuksille valitsemaan $d = 1, D = 1$. Osa prediktoreista sisältää selvää kausivaihtelua, joten prediktorien saamiseksi likimain stationaarisiksi, on otettava kausierotukset. Koska prediktorien ja sähkönkulutuksen välinen suhde halutaan pitää samana, otetaan myös sähkönkulutuksen aikasarjasta kausierotukset. Mikäli käytetään pelkkää sähkönkulutuksen aikasarjaa, voisi $d = 1, D = 0$ olla myös toimiva valinta. Kuvassa 29 esitetty aikasarja $\log(y_t) - \log(y_{t-1})$ ei poikkea selvästi stationaarisuudesta. Kuitenkin tässä tapauksessa myös ARIMA-mallin tapauksessa on käytetty perättäisille ja kausierotuksille $d = 1, D = 1$.

Tarkastellessa MAE:a prediktorien määrän suhteen, huomataan että prediktorien määrän lisäyksellä ei saada suurta parannusta MAE:en. Mallissa 2 valitaan käytettäväksi vain kahta prediktoria ja mallissa 1 viittä. Mallien erot MAE:n suhteen ovat hyvin pienet. Myös multiplikatiivisella ARIMA-mallilla saadaan lähes samansuuruinen MAE. Näin ollen voi sanoa, että tässä tapauksessa prediktoreilla ei saada hyötyä kulutuksen ennustamiseen ainakaan MAE:n suhteen.

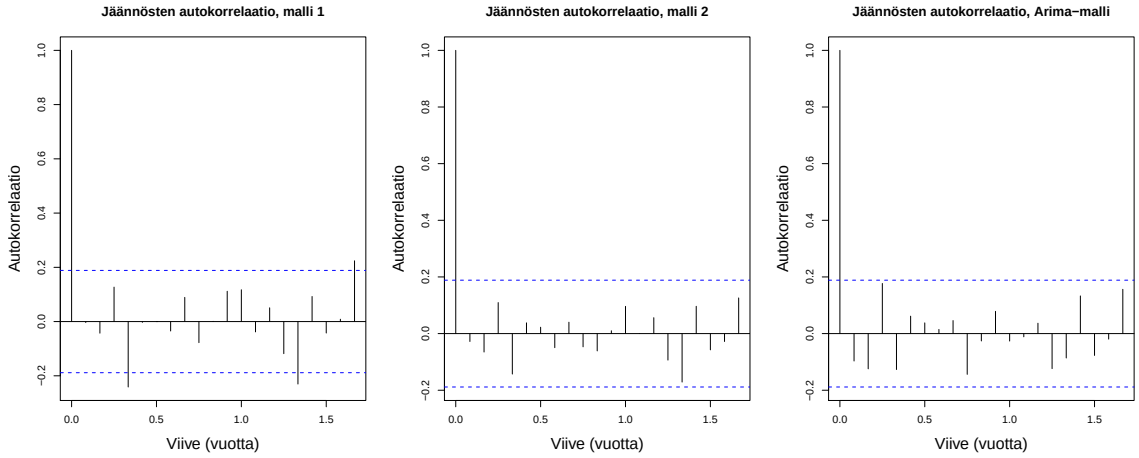
Ennustekuvasta 31 nähdään, että molemmat ennusteet testiaineistossa osuvat jälleen varsin hyvin kohdalleen. Tässä tapauksessa kuitenkin malliin 2 liittyvät ennustevälit ovat kapeammat, joten malli 2 sopisi opetusaineistoon paremmin kuin malli 1. Myös



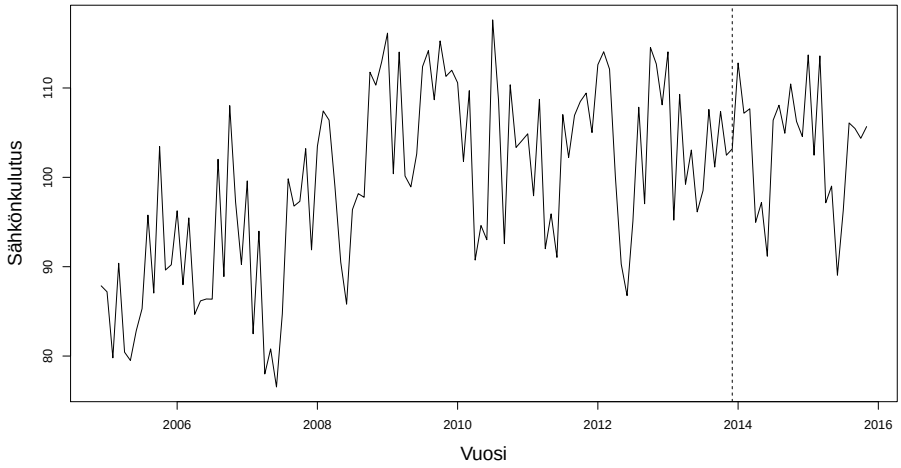
Kuva 22: Mallien antamat ennusteet toimialalle 2 testiaineistossa. Sininen viiva on piste-ennuste kulutukselle. Lisäksi on piirretty 80 ja 95 % ennustevälit. Punainen viiva on testiaineistossa havaittu kulutus.



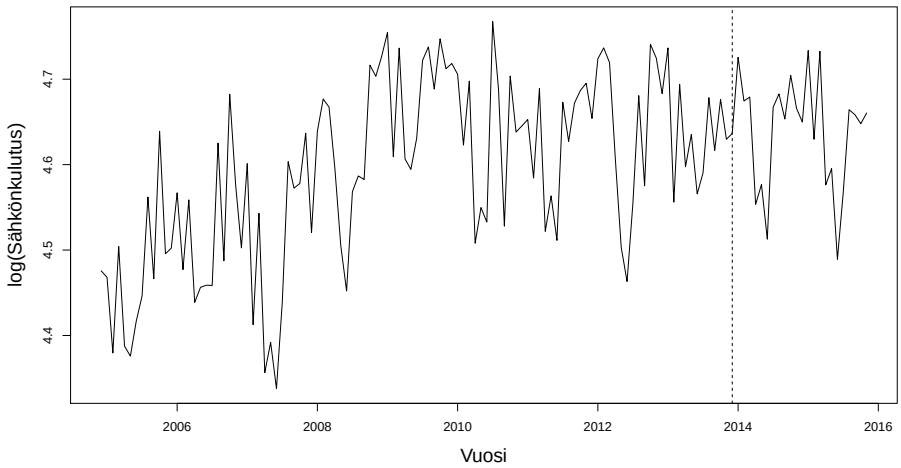
Kuva 23: Standardoitujen jäännösten kvantiilikuviot toimialalla 2.



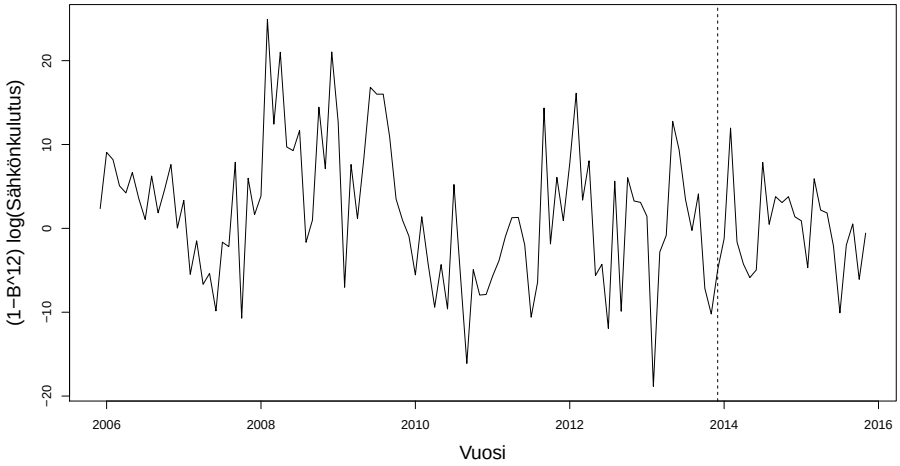
Kuva 24: Jäännösten z_t autokorrelaatio viiveillä 0-20 kuukautta toimialalla 2.



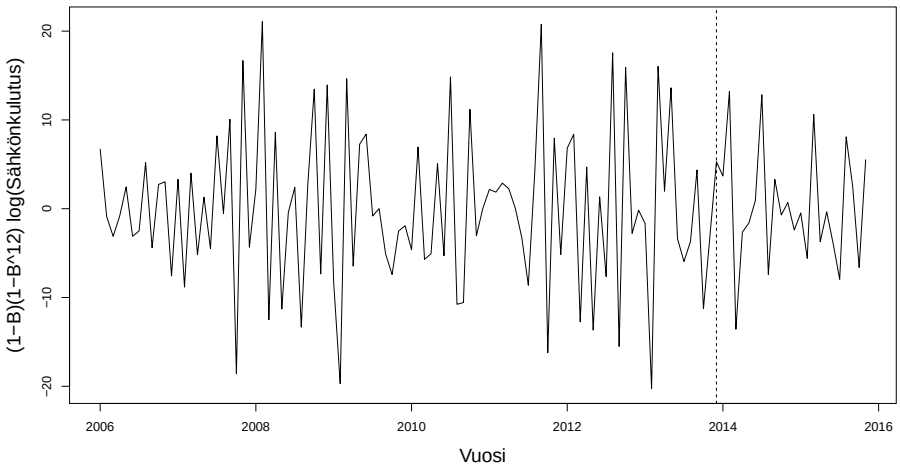
Kuva 25: Toimialan 3 sähkönkulutuksen aikasarja. Pystyviiva merkitsee testiaineiston alkua.



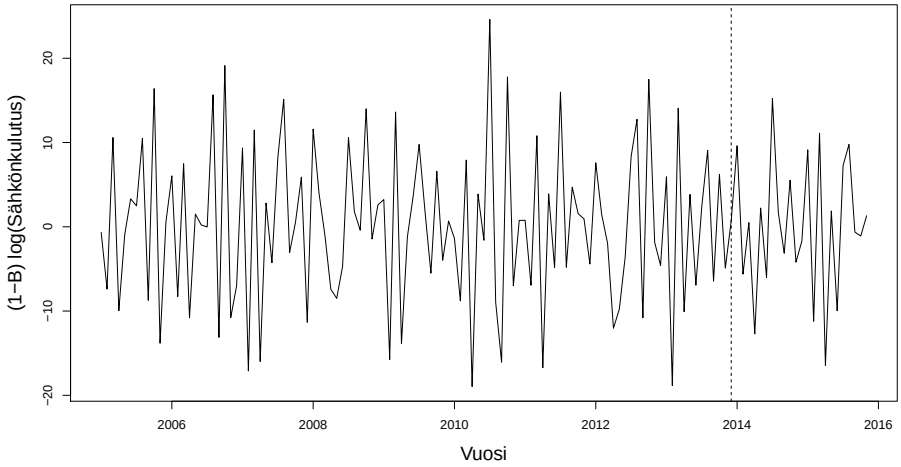
Kuva 26: Toimialan 3 logaritmisen sähkönkulutuksen aikasarja. Pystyviiva merkitsee testiaineiston alkua.



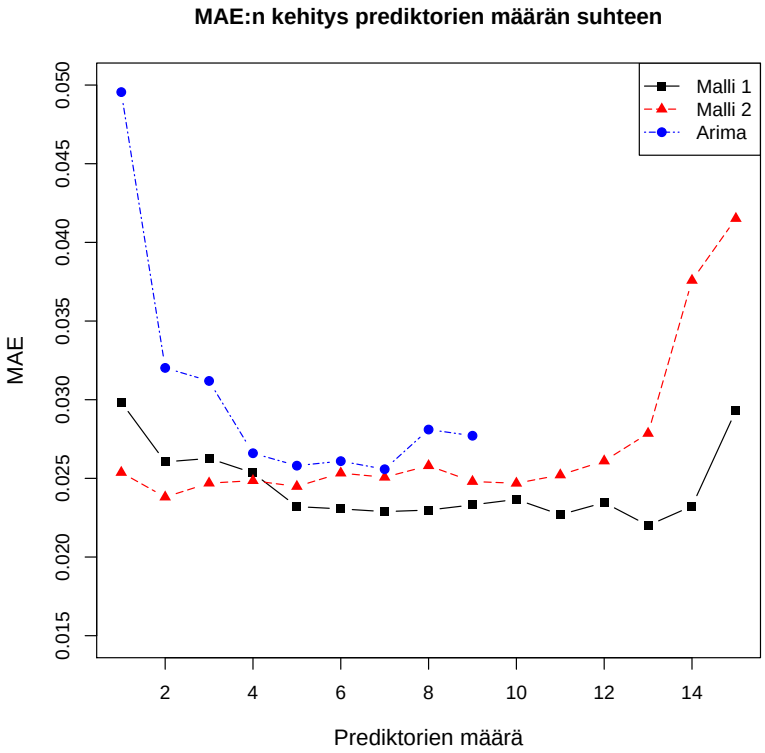
Kuva 27: Toimialan 3 kausierotusten aikasarja $\log(y_t) - \log(y_{t-12})$. Pystyviiva merkitsee testiaineiston alkua.



Kuva 28: Toimialan 3 aikasarja, josta on otettu kausierotukset ja perättäiset erotukset, eli $\log(y_t) - \log(y_{t-1}) - \log(y_{t-12}) + \log(y_{t-13})$. Pystyviiva merkitsee testiaineiston alkua.



Kuva 29: Toimialan 3 perättäisten erotusten aikasarja $\log(y_t) - \log(y_{t-1})$ Pystyviiva merkitsee testiaineiston alkua.



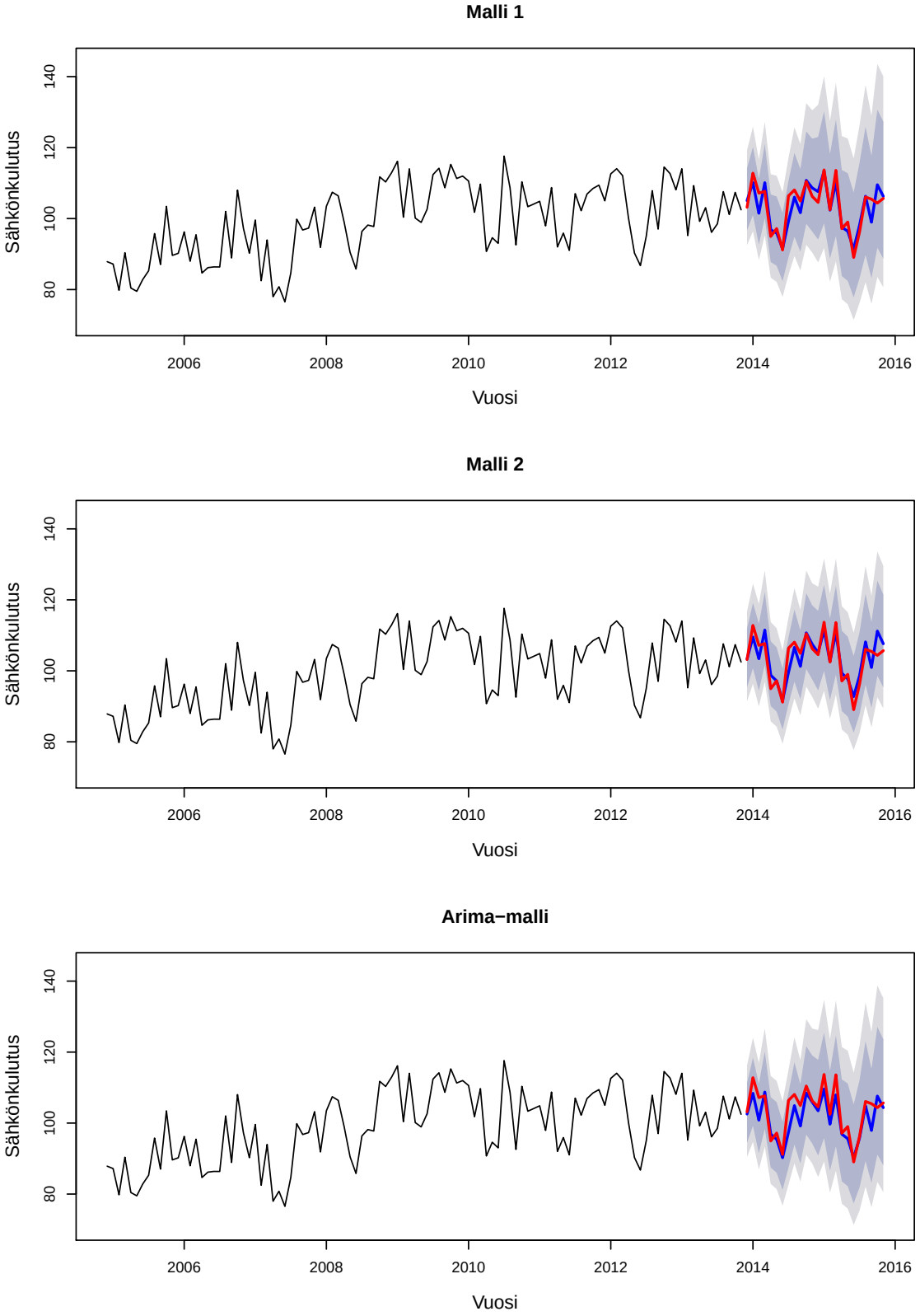
Kuva 30: MAE:n kehitys prediktorien määrän funktiona toimialalla 3, kun käytetään prediktoreina prediktorien nykyarvoja ja vuoden takaisia arvoja (malli 1) ja kahden ja kolmen vuoden takaisia arvoja (malli 2). Lisäksi vertailumallina käytetty multiplikatiivinen ARIMA-malli, jossa prediktorien määrä = käytettyjen parametrien määrä.

multiplikatiivisen ARIMA-mallin ennuste on hyvin samankaltainen mallin 1 ja 2 ennusteiden kanssa sekä piste-ennusteen että ennustevälien suhteen tarkasteltuna.

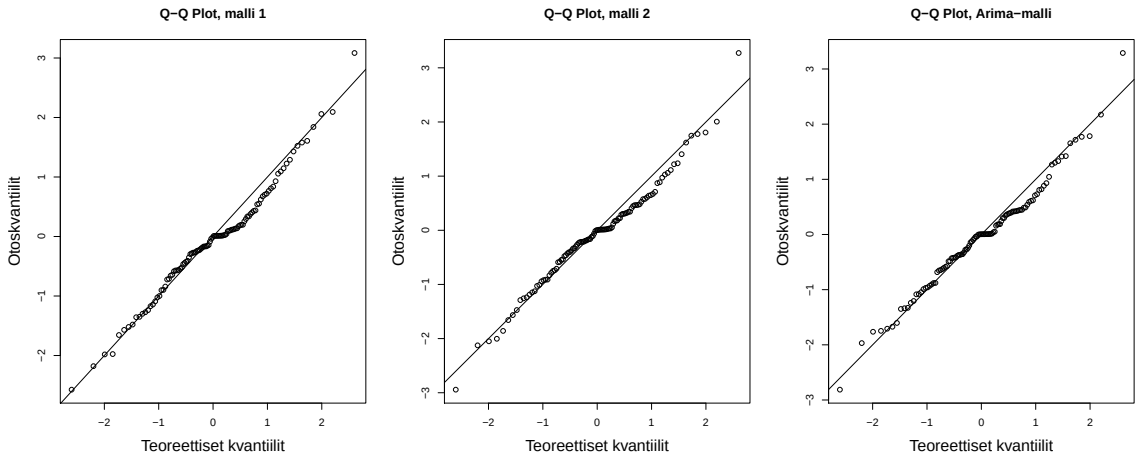
Jäännösten kvantiilikuviot kuvassa 32 näyttävät kaikkien mallien tapauksessa melko hyviltä. Myös jäännösten autokorrelaatiot kuvassa 33 osuvat yhtä mallin 1 ja mallin 2 piikkiä lukuun ottamatta hyvin ennustevälien sisäpuolelle. Myöskään Ljung-Box-Piercen testi ei anna aihetta kyseenalaistaa jäännösten riippumattomuutta, (malli 1: $\hat{Q} = 22.32, df = 14, p = 0.072$, malli 2: $\hat{Q} = 14.04, df = 15, p = 0.52$, ARIMA-malli: $\hat{Q} = 15.29, df = 13, p = 0.289$).

Tarkastellaan vielä yleisesti muitakin toimialoja ja mallien eroja ennustetarkkuudessa. Taulukossa 5 on esitetty mallien ennustekyky eri toimialoilla MAE:n suhteen. Joka toimialalla pienin MAE on lihavoitu. Toimialoilla 15-17 on MAE laskettu oikealle kulutukselle, jolle keskiarvo on 100. Muilla toimialoilla on käytetty logaritmista kulutusta MAE:n laskentaan. Koska logaritmiset arvot ovat selvästi pienempiä, on MAE pienempi toimialoilla, joissa MAE on laskettu logaritmiselle kulutukselle.

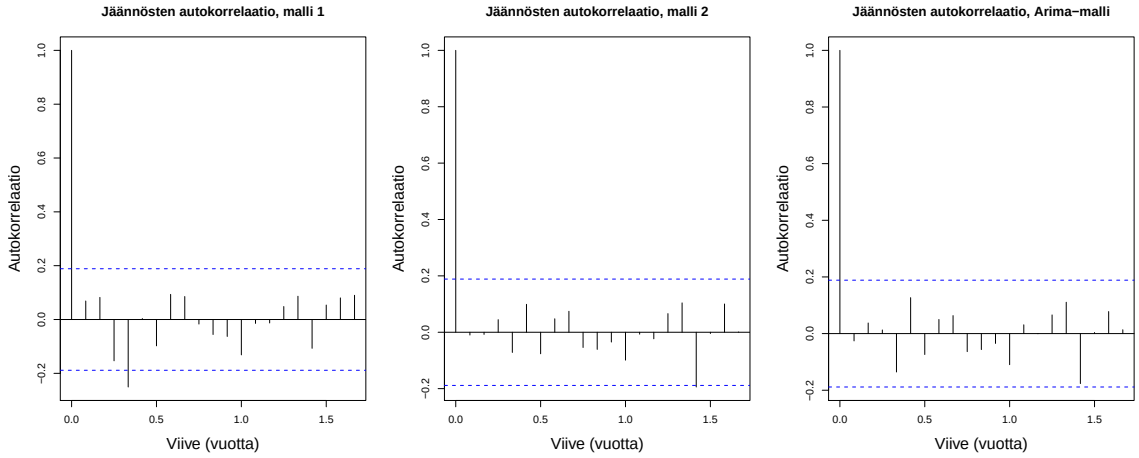
Huomataan, että MAE:n mielessä malli 1 on suurimmalla osalla toimialoista paras. Kuitenkin viittä (toimialat 1,2,6,8 ja 12) toimialaa lukuunottamatta mallit 1 ja 2 ovat lähes yhtä hyviä MAE:n suhteen. Malli 1 on mallia 2 parempi näistä neljällä toimialalla. Arima-malli on jonkin verran huonompi kuin molemmat prediktorien mallit toimialoilla 1,2,8,11,13,16 ja 17. Vastaavasti ARIMA-malli ja malli 2 ovat mallia 1 parempia toimialalla 12. Toimialalla 8 malli 1 on muita malleja parempi. Muilla toimialoilla mallit ovat lähes yhtä hyviä. Tämän tarkastelun perusteella osalla toimialoista prediktorien käytöstä voisi olla hyötyä ennustettaessa toimialoittaista sähkönkulutusta. Suurta haittaa prediktoreista ei ole yhdelläkään toimialalla.



Kuva 31: Mallien antamat ennusteet toimialalle 3 testiaineistossa. Sininen viiva on piste-ennuste kulutukselle. Lisäksi on piirretty 80 ja 95 % ennustevälit. Punainen viiva on testiaineistossa havaittu kulutus.



Kuva 32: Standardoitujen jäännösten kvantiilikuviot toimialalla 3.



Kuva 33: Jäännösten z_t autokorrelaatio viiveillä 0-20 kuukautta toimialalla 3.

Taulukko 5: Eri mallien vertailu toimialoittain. Jokaisella toimialalla pienin MAE on lihavoitu.

Toimiala	Malli 1		Malli 2		Arima-malli	
	MAE	Prediktoreja	MAE	Prediktoreja	MAE	Parametreja
1	**0.0214**	7	0.0401	4	0.0836	2
2	**0.0269**	5	0.0434	6	0.0621	7
3	**0.0232**	5	0.0238	2	0.0265	4
4	0.0283	7	**0.0225**	7	0.0269	8
5	**0.0516**	3	0.0532	5	0.0549	3
6	**0.0089**	7	0.0171	2	0.0149	6
7	**0.0166**	6	0.0185	6	0.0213	3
8	**0.0188**	3	0.0284	2	0.0358	5
9	0.0224	6	**0.0200**	5	0.0214	5
10	**0.0125**	5	0.0139	9	0.0159	5
11	0.0343	5	**0.0337**	4	0.0473	9
12	0.0174	2	0.0139	2	**0.0124**	4
13	**0.0411**	7	0.0418	8	0.0541	3
14	**0.0134**	4	0.0138	4	0.0179	3
15	2.1732	6	**2.0936**	9	2.4642	3
16	**1.8683**	6	2.1442	5	2.7491	4
17	**3.5054**	7	3.5593	8	4.5461	3

6 Pohdinta

Työssä esitettiin eräs tapa ennustaa toimialojen sähkönkulutusta käyttämällä apuna muita aikasarjoja. Aikasarjojen vaikutuksia sähkönkulutukseen mallinnettiin aikasarjaregression avulla. Tavoitteena oli valita parhaat selittävät aikasarjat ja niiden määrä suuresta joukosta siten, että ennuste olisi tarkka. Selittävien prediktoreiden valinta toteutettiin jakamalla aineisto opetus- ja testiaineistoon ja vertaamalla prediktorien mallien ennustekykyä testiaineistossa.

Tutkielmassa huomattiin, ettei sopivan prediktorien määrän valinta ole aina selvää. Joskus prediktorien määrän lisääminen auttaa aina ennusteen tarkkuuden paranemiseen. Käyttäessä prediktoreita, joiden arvot tulee ennustaa kulutusennusteen muodostamiseksi, prediktorien määrän lisääminen lisää sähkönkulutuksen mallin epävarmuutta, kun tarvitaan ennusteet useammalle prediktorille. Tällaisissa tilanteissa olisi löydettävä prediktorien määrä, jolla ennusteen tarkkuus on hyvä, mutta prediktorien määrä pysyy vielä pienenä. Tutkielmassa ei ole otettu tarkemmin kantaa sopivaan määrään prediktoreita. Tähän tulisi jatkossa kehittää tarkoituksenmukainen ratkaisu.

Sähkönkulutusta mallinnettiin käyttämällä selittäjinä kansantalouden ja toimialan taloudelliseen kehitykseen, lämpötilaan ja viikonpäiviin liittyviä prediktoreita. Osasta prediktoreista käytettiin myös vuoden tai useamman vuoden takaisia arvoja. Näiden prediktorien vaikutuksen ajateltiin voivan olla viiveellistä. Jälkikäteen ajateltuna rajausta viiveellisten prediktorien valintaan ei ole syytä tehdä tässä vaiheessa, vaan viiveellisinä prediktoreina voisi käyttää lämpötilaa ja viikonpäivämuuttujia lukuun ottamatta kaikkia käytössä olevia prediktoreita.

Kokonaisuudessa aikasarjaregressiomalli toimi pääosin hyvin ennustettaessa toimialojen kulutusta testiaineistossa. Prediktorit, joiden arvoja joudutaan ennustamaan, tuovat ennusteeseen epävarmuutta. Mikäli mallia varten ennustetaan useita prediktoreita, prediktorien ennustamisessa tulisi myös ottaa huomioon prediktorien väliset riippuvuussuhteet. Sähkönkulutusta ennustettaessa on helpompi käyttää prediktoreita, joita ei tarvitse ennustaa, mikäli niiden avulla tehdyn mallin ennustekyky testiaineistossa on lähellä mallia, jossa prediktorien arvoja tulisi ennustaa. Tällaisia ovat muun muassa tässä tutkielmassa käytetyt viiveelliset prediktorit ja viikonpäivämuuttujat. Näiden arvoihin ennustevälillä ei liity epävarmuutta, joten ennusteen ennustevälit ovat todenmukaiset.

Täytyy muistaa, että opetus- ja testiaineiston perusteella tehty mallin valinta ei tuo tietoa mallin keskimääräisestä ennustevirheestä. Kuitenkin ennustemallin valintaan tähän asti olemassa olevan tiedon perusteella menetelmä sopii hyvin.

Mallin jatkokehitys on mahdollista monella eri tapaa. Automaattisuutta prediktorien määrän valintaan ja malliin liittyvien oletusten tarkasteluun voi tehdä. Prediktorien joukkoa voi muuttaa tarpeen mukaan. Myös esimerkiksi mallinvalintakriteeriä voidaan tarkentaa ennustemallin käyttötarkoituksen perusteella. Kuukausittaiseen kulutukseen liittyvä kuukausien pituuden vaihtelusta johtuvaa kausivaihtelua on mahdollista hallita skaalaamalla jakamalla kulutukset ennen analyysia kuukauden pituudella.

Kiitokset

Kiitos Jyväskylän Energialle opinnäytetyöpaikasta ja positiivisesta työilmapiiristä. Haluan kiittää tutkielmani ohjaajia professori Juha Karvasta, professori Kaisa Miettistä, tutkijatohtori Markus Hartikaista ja Jyväskylän Energian kehityspäällikkö Risto Ryyminiä kannustavasta ohjauksesta ja hyvistä kommenteista työhön liittyen. Haluan myös kiittää Jyväskylän Energian Reijo Laapottia suuresta avusta aineiston hankinnassa.

Lähdeluettelo

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716--723.

Box, G. E. P. ja Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 26(2):211--252.

Box, G. E. P., Jenkins, G. M. ja Reinsel, G. C. (2008). *Time Series Analysis*. Wiley, New Jersey, 4. edition.

Brockwell, P. J. ja Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, New York, 2. edition.

Brockwell, P. J. ja Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer, New York, 2. edition.

Hastie, T., Tibshirani, R. ja Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, New York, 2. edition.

Hong, T. ja Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*. `http://dx.doi.org/10.1016/j.ijforecast.2015.11.011`.

Hong, T., Wilson, J. ja Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, 5(1):456--462.

Hurvich, C. M. ja Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297--307.

Hyndman, R. J. (2016). *forecast: Forecasting functions for time series and linear models*. R package version 7.0.

Hyndman, R. J. ja Athanasopoulos, G. (2013). *Forecasting: Principles and Practice*. Otexts. `http://www.otexts.org/fpp`.

Hyndman, R. J. ja Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1--22.

Hyndman, R. J. ja Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(2):679--688.

Ilmatieteen laitos (2016). Lämpötila ja sademääräaineistoja vuodesta 1961. Elektroninen aineisto. Helsinki: [viitattu: 4.3.2016]. Saantitapa: `http://ilmatieteenlaitos.fi/tilastoja-vuodesta-1961`.

Nasdaq Inc. (2016). Pörssi-indeksit. Elektroninen aineisto. [viitattu: 4.3.2016]. Saantitapa: `http://www.nasdaqomxnordic.com/indeksit`.

Nord Pool (2016). Sähkön tukkuhinta. Elektroninen aineisto. Lysaker,Norja: [viitattu 4.3.2016]. Saantitapa: http://www.nordpoolspot.com/.

OECD (2016). Long-term interest rates (indicator). Elektroninen aineisto. Pariisi: doi: 10.1787/662d712c-en. [viitattu: 4.3.2016]. Saantitapa: https://data.oecd.org/.

OECD (2016). Short-term interest rates (indicator). Elektroninen aineisto. Pariisi: doi: 10.1787/2cc37d77-en. [viitattu: 4.3.2016]. Saantitapa: https://data.oecd.org/.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Suomen Asiakastieto Oy (2016). http://www.asiakastieto.fi/web/fi. Viitattu 11.5.2016.

Suomen virallinen tilasto (SVT) (2016a). Aluetilinpito. Verkkojulkaisu. ISSN=1799-3393. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/altp/index.html.

Suomen virallinen tilasto (SVT) (2016b). Ansiotasoindeksi. Verkkojulkaisu. ISSN=1796-3737. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/ati/index.html.

Suomen virallinen tilasto (SVT) (2016a). Kansantalouden tilinpito. Verkkojulkaisu. ISSN=1795-8881. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/vtp/index.html.

Suomen virallinen tilasto (SVT) (2016b). Kaupan liikevaihtokuvaaja. Verkkojulkaisu. ISSN=1799-0912. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/klv/index.html.

Suomen virallinen tilasto (SVT) (2016c). Kuluttajahintaindeksi. Verkkojulkaisu. ISSN=1796-3524. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/khi/index.html.

Suomen virallinen tilasto (SVT) (2016a). Majoitustilasto. Verkkojulkaisu. ISSN=1799-6309. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/matk/index.html.

Suomen virallinen tilasto (SVT) (2016b). Maksutase ja ulkomainen varallisuusasema. Verkkojulkaisu. ISSN=2342-3463. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/mata/index.html.

Suomen virallinen tilasto (SVT) (2016c). Neljännesvuositilinpito. Verkkojulkaisu. ISSN=1797-9749. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/ntp/index.html.

Suomen virallinen tilasto (SVT) (2016a). Palkkasummakuvaajat. Verkkojulkaisu. ISSN=1798-6109. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/ktps/index.html.

Suomen virallinen tilasto (SVT) (2016b). Palvelualojen liikevaihtokuvaaja. Verkkojulkaisu. ISSN=1799-0971. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/plv/index.html.

Suomen virallinen tilasto (SVT) (2016a). Teollisuuden liikevaihtokuvaaja. Verkkojulkaisu. ISSN=1798-5943. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/tlv/index.html.

Suomen virallinen tilasto (SVT) (2016b). Teollisuuden uudet tilaukset. Verkkojulkaisu. ISSN=1798-6710. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/teul/index.html.

Suomen virallinen tilasto (SVT) (2016c). Teollisuustuotannon volyymi-indeksi. Verkkojulkaisu. ISSN=1796-3788. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/ttvi/index.html.

Suomen virallinen tilasto (SVT) (2016). Tuotannon suhdannekuvaaja. Verkkojulkaisu. ISSN=1798-5404. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/ktkk/index.html.

Suomen virallinen tilasto (SVT) (2016a). Työvoimakustannusindeksi. Verkkojulkaisu. ISSN=1798-3568. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/tvki/index.html.

Suomen virallinen tilasto (SVT) (2016b). Työvoimatutkimus. Verkkojulkaisu. ISSN=1798-7830. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/tyti/index.html.

Suomen virallinen tilasto (SVT) (2016). Väestörakenne. Verkkojulkaisu. ISSN=1797-5379. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/vaerak/index.html.

Tilastokeskus (2016). Rakentamisen liikevaihtokuvaaja. Verkkojulkaisu. ISSN=1798-5919. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/rlv/index.html.

Tilastokeskus (2008). Toimialaluokitus TOL 2008. Helsinki. https://www.stat.fi/meta/luokitukset/toimiala/001-2008/kasikirja.pdf.

Tilastokeskus (2016). Tuottavuustutkimukset. Verkkojulkaisu. Helsinki: Tilastokeskus [viitattu: 23.3.2016]. Saantitapa: http://www.stat.fi/til/ttut/index.html.

Tulli (2016). Uljas - tavaroiden ulkomaankauppatilastot. Elektroninen aineisto. Helsinki: [viitattu: 4.3.2016]. Saantitapa: http://uljas.tulli.fi/.

Työ- ja elinkeinoministeriö (2016). Toimiala online tietopalvelu. Elektroninen aineisto. Helsinki: [viitattu: 4.3.2016]. Saantitapa: http://www2.toimialaonline.fi/.

Valtioneuvosto (2009). Valtioneuvoston asetus sähköntoimitusten selvityksestä ja mittauksesta. http://energia.fi/sites/default/files/vn_sahkontoimitusten_selvitykset_ja_mittaukset.pdf.

Yahoo Inc. (2016). Pörssi-indeksit. Elektroninen aineisto. [viitattu: 4.3.2016]. Saantitapa: http://finance.yahoo.com.

A Prediktorien Internet-osoitteet

Taulukko 6: Prediktorien Internet-osoitteet

Prediktori	Linkki tilastoon
Kansantaloudelliset prediktorit	
ansiotasoindeksi	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__pal__ati/020_ati_tau_102_fi.px/?rxid=11bacf59-3ae6-466c-aacd-1c8a4b04312e
bruttokansantuote*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kan__ntp__ntp_1/120_ntp_tau_102.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
elinkustannusindeksi	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__hin__khi/050_khi_tau_105.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
julkiset investoinnit*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kan__ntp__ntp_1/120_ntp_tau_102.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
julkinen kulutus*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kan__ntp__ntp_1/120_ntp_tau_102.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
kuluttajien luottamusindikaattori*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__tul__kbar/010_kbar_tau_101.px/?rxid=386f590b-644f-4daa-b902-211f4c90e6a8
lyhyen aikavälin korko*	https://data.oecd.org/interest/short-term-interest-rates.htm
makrotalouden indikaattori*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__tul__kbar/010_kbar_tau_101.px/?rxid=386f590b-644f-4daa-b902-211f4c90e6a8
pitkän aikavälin korko*	https://data.oecd.org/interest/long-term-interest-rates.htm
raakaöljyn hinta*	http://www.oil.fi/fi/tilastot-1-hinnat-ja-verot/13-raakaoljyn-hintakehitys
sähkön hinta*	http://www.nordpoolspot.com/Market-data1/Elspot/Area-Prices/FI/Monthly/?dd=FI&view=table
tavaroiden tuonti	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kau__taukyv/010_taukyv_tau_101.px/?rxid=f1b53b88-8e3a-491f-84d3-f9d85e44040f
tavaroiden vienti	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kau__taukyv/010_taukyv_tau_101.px/?rxid=f1b53b88-8e3a-491f-84d3-f9d85e44040f
työllisyysaste Keski-Suomessa	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__tym__tyti/060_tyti_tau_109.px/?rxid=4523480c-0d75-4c09-b205-5ae2a6144504
työllisyysaste Suomessa	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__tym__tyti/010_tyti_tau_102.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
väkiluku Jyväskylässä	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__vrm__vaerak/058_vaerak_tau_104.px/?rxid=6a04ac30-8de6-47d9-b2e0-5458352e76e4
yksityiset investoinnit*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kan__ntp__ntp_1/120_ntp_tau_102.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
yksityinen kulutus*	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kan__ntp__ntp_1/120_ntp_tau_102.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
Toimialakohtaiset prediktorit	
toimialan arvonlisäys	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kan__ntp__ntp_1/110_ntp_tau_101.px/?rxid=f892d4ad-33be-4c45-a952-1898710a06ba
toimialan liikevaihto + yöpymiset	
kaupan liikevaihtokuvaaja	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__kau__klv/009_klv_tau_103_fi.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
palvelujen liikevaihtokuvaaja	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__pav__plv/004_plv_tau_103_fi.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
rakentamisen liikevaihtokuvaaaja	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__rak__rlv/004_rlv_tau_108_fi.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
teollisuuden liikevaihtokuvaaja	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__teo__tlv/004_tlv_tau_108_fi.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
hotellien yöpymiset (majoitus-toimialalla)	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__lii__matk/020_matk_tau_102_fi.px/?rxid=7ba9d74d-9c4b-45e4-9af2-bbd478726b57
toimialan luottamusindikaattori*	http://www2.toimialaonline.fi/
toimialan työllisten lukumäärä Suomessa	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__tym__tyti/065_tyti_tau_108.px/?rxid=76f63318-4589-4c79-b782-7637cbd9b110
toimialan palkkasummakuvaaja	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__pal__ktps/009_ktps_tau_112_fi.px/?rxid=a2a9fdcc-f3fb-4994-808f-5cbf405c21dc
toimialan volyymi	http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__pal__ktps/009_ktps_tau_112_fi.px/?rxid=a2a9fdcc-f3fb-4994-808f-5cbf405c21dc

Taulukko 7: Prediktorien Internet-osoitteet

Prediktori	Linkki tilastoon
Pörssi-indeksit	
Dow Jones Industrial Average-indeksi*	http://finance.yahoo.com/q?s=^DJI
Helsingin pörssin yleisindeksi*	http://www.nasdaqomxnordic.com/index/index_info?Instrument=FI0008900220
Pohjoismaiden pörssin OMX Nordic 40 -indeksi*	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0001809476
Standard & Poor's 500 -indeksi*	http://finance.yahoo.com/q?s=%5EGSPC
Toimialakohtaiset pörssi-indeksit	
N Industrial Goods & Services GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384584
N Food & Beverage GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384618
N Consumer Goods GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384592
N Consumer Services GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384659
N Retail GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384667
N Media GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384675
N Travel & Leisure GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384683
N Telecommunications GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384691
N Financials GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384733
N Real Estate GI	http://www.nasdaqomxnordic.com/index/index_info?Instrument=SE0004384766
Muut preditorit	
kuukauden keskilämpötila Jyväskylässä	http://ilmatieteenlaitos.fi/tilastoja-vuodesta-1961
arkipäivien lukumäärä kuukaudessa	
vapaapäivien lukumäärä kuukaudessa	