Suomen bruttokansantuotteen sekä toimialoittaisten tuotantojen ennustaminen aikasarja-analyysin keinoin

Pro gradu -tutkielma
Tuomas Hiltunen
242835
Fysiikan ja matematiikan laitos
Itä-Suomen yliopisto
29.1.2021

Tiivistelmä

Aikasarjat ovat jonomuotoisia havaintoaineistoja, jotka ilmaisevat jonkun tietyn muuttujan kehitystä ajan kuluessa. Tässä tutkielmassa tutkitaan Suomen toimialoittaisten tuotantojen ja bruttokansantuotteen arvojen kehitystä aikasarja-analyysin avulla. (Heikosti) stationaarisen aikasarjan muuttujan odotusarvo ja muuttujien väliset varianssit ovat vakioita ajan suhteen, sekä havaintojen väliset korrelaatiot riippuvat vain niiden välisten ajanhetkien etäisyyksistä.

Autoregressiivisen liukuvan keskiarvon mallin $ARMA(p, q)$ on oltava stationaarinen. Siinä aikasarjan tulevaa arvoa ennustetaan sen edellisten havaintoarvojen sekä valkoisen kohinan lineaarikombinaatioiden avulla. Stationaarisuusvaatimuksesta päästään eroon laajentamalla $ARMA(p, q)$-malli autoregressiiviseksi integroiduksi liukuvan keskiarvon malliksi $ARIMA(p, d, q)$, jossa epästationaarinen $ARIMA(p, d, q)$-malli voidaan muuttaa stationaariseksi differoimalla. Koska aikasarjoissa esiintyy usein kausittaisvaihtelua, muodostetaan divisiivinen kausittainen autoregressiivinen integroitu liukuvan keskiarvon malli $ARIMA(p, d, q) \times (P, D, Q)_s$.

Analysoitava data on neljännesvuosidata, jossa esiintyy Suomen bruttokansantuotteen arvojen lisäksi kymmenen eri toimialan tuotantojen arvot vuoden 1990 alusta vuoden 2019 puoliväliin. Datasta poistetaan kunkin aikasarjan kahdeksan viimeistä havaintoarvoa, joita ennustetaan aikasarjojen aiempien havaintoarvojen perusteella käyttäen $ARIMA(p, d, q) \times (P, D, Q)_s$-malleja. Kullekin aikasarjalle valitaan sopivin $ARIMA(p, d, q) \times (P, D, Q)_s$-malli korrelaatiofunktioiden, informaatiokriteerien, standardoitujen residuaalien sekä yksikkökiekkotarkastelujen avulla.

Ennusteiden laatua arvioidaan prosentuaalisen keskivirheiden, absoluuttisten prosentuaalisten keskivirheiden sekä skaalattujen keskineliövirheiden avulla. $ARIMA(p, d, q) \times (P, D, Q)_s$-mallien avulla tutkittaville aikasarjoille luodut ennusteet toimivat pääasiassa hyvin. Mallit ennustavat toimialoista parhaiten kiinteistöalan ja huonoiten rahoituksen ja vakuutusten aikasarjoja. Tarkkoja ennusteita saadaan myös koulutus-, terveys- ja sosiaalipalveluiden sekä taiteiden, viihteen ja virkistyksen aikasarjoille. Epätarkkoja ennusteita saadaan myös kaupan, liikenteen, majoitus- ja ravitsemistoiminnan sekä hallinto- ja tukipalveluiden aikasarjoille.

BKT:n oman aikasarjan ennuste on hieman tarkempi kuin toimialoittaisten ennusteiden summa, joskin molemmilla tavalla ennusteet ovat varsin tarkkoja. Ennusteet heikkenevät ennustettavan ajanjakson edetessä. Tulevien havaintoarvojen ennustaminen on sitä helpompaa, mitä selkeämmät trendit ja kausittaisvaihtelut aikasarjoilla on. Aikasarjojen skaalatut keskihajonnat korreloivat jossain määrin niiden ennustevirheiden kanssa.

Abstract

Time series are sequence-like observation data which indicate improvement of some particular variable in process of time. There is researched improvement of productions of Finnish industries and gross national product with time series analysis in these thesis. Mean of variable and variances of variables of (weakly) stationary time series are constants and correlations of observations depend only on distance of these moments.

Autoregressive moving average model $ARMA(p, q)$ has to be stationary. There is predicted upcoming value of time series based on linear combination of its previous values and its white noise. There is disposed demand of stationarity by extend $ARMA(p, q)$-model to autoregressive integrated moving average model $ARIMA(p, d, q)$, where non-stationary $ARIMA(p, d, q)$-model is modified to stationary by differentiation. Because there is usually seasonal differences in time series, there is defined divisive seasonal autoregressive integrated moving average model $ARIMA(p, d, q) \times (P, D, Q)_s$.

There is quarter data to analyze where is existed production of Finnish GNP besides productions of ten Finnish industries from the beginning of the year 1990 to second quarter of the year 2019. There is deleted eight last productions of each time series, which are predicted based in their previous productions by using $ARIMA(p, d, q) \times (P, D, Q)_s$-models. There is chosen most appropriate $ARIMA(p, d, q) \times (P, D, Q)_s$-model for each time series with correlation functions, information criteria, standardized residuals and unit circle examinations.

Qualities of predictions are analyzed with mean percentage errors, mean absolute percentage errors and scaled mean squared errors. Predictions made with $ARIMA(p, d, q) \times (P, D, Q)_s$-models for time series works mainly well. Models predict most precise time series of properties and most imprecise time series of funding and insurances. There is also precise predictions for time series of education, health and social services as well as time series of arts, entertainment and recreation. There is also imprecise predictions for time series of trade, transport, accommodation and nourishment occupation and for time series of administration and helplines.

Prediction based in GNP's proprietary time series is a bit precise than a sum of predictions for productions of industries when predicting Finnish gross national product, but both ways are quite precise. Precises of predictions decrease during predictable time period. Predicting upcoming productions is easier if time series have explicit trends and seasonal differences. Scaled standard deviations and prediction errors of time series correlate somewhat with each other.

Sisältö

1 Johdanto

Pro gradu -tutkielmani tavoitteena on tutkia Suomen bruttokansantuotetta sekä toimialoittaisia tuotantoja aikasarja-analyysin keinoin.

Tutkielman tarkoituksena on luoda toimialojen sekä bruttokansantuotteen aikasarjoille neljännesvuosittaiset ennusteet kahdeksi vuodeksi $ARIMA$-mallien avulla sekä vertailla ennusteita aikasarjojen toteumiin. Tällöin voidaan arvioida mallien soveltuvuutta eri tavoin käyttäytyvien aikasarjojen ennustamiseen ennustevirheiden tunnuslukuja hyödyntäen. Erityinen mielenkiinto kohdistuu siihen, saadaanko koko bruttokansantuotteelle luotua parempi ennuste sen oman aikasarjan perusteella vai laskemalla toimialoittaiset ennusteet yhteen. Lisäksi on mielenkiintoista nähdä heikkeenekö mallien ennustetarkkuus ennustettavan ajanjakson edetessä.

Tutkimuksessa käytettävää teoriaa esitetään ennen kuin sitä aletaan soveltamaan aikasarjoihin. Luvun 2 alussa esitellään yleisesti aikasarjoja, minkä jälkeen samassa luvussa esitetään $ARIMA$-mallien teoriaa vaadittavine pohjatietoineen sekä tavallisille että kausittaisille malleille. Tämän jälkeen määritellään ennustevirheiden tunnusluvut. Luvussa 3 esitellään tutkielmassa käsiteltävät aikasarjat sekä näytetään vaihe vaiheelta kuinka aikasarjalle valitaan sopiva $ARIMA$-malli. Luvussa 4 luodaan ennusteet sekä bruttokansantuotteen että toimialojen aikasarjoille sekä analysoidaan niitä suhteessa toteumiin. Lopuksi tutkitaan ennusteiden tarkkuutta neljännesvuosittain sekä vedetään tutkimustuloksia yhteen.

Tutkielmassa on käytetty Tilastokeskuksen neljännesvuosidataa [1], jota on muokattu vastaamaan paremmin tutkielman tarkoitusta. Datasta on poistettu päällekkäisyyksiä, jotta jokaisen toimialan tuotanto näkyy datassa tasan yhden kerran koko bruttokansantuotteen lisäksi. Datan muuttujien mittayksikkö on alkuperäinen sarja, miljoonaa euroa, kiintein hinnoin ja niiden viitevuosi on 2010. Tutkielmassa bruttokansantuotetteen aikasarja on muodostettu toimialoittaisten tuotantojen summana eikä kiinteähintaisena, jotta toimialoittaisten ennusteiden summaa voidaan verrata sen oman aikasarjan ennusteeseen. Data oli alun perin Excel-muodossa, josta se on tuotu R-ohjelmaan, jota on käytetty datan analysointiin, laskutoimituksiin sekä kuvien piirtämiseen. Tutkielma on kirjoitettu puhtaaksi LaTeX -ohjelmalla.

2 Aikasarjat ja ARIMA-mallit

Myöhemmin tutkielmassa ennustetaan toimialojen aikasarjoja $ARIMA$-mallien avulla, minkä takia tässä luvussa esitetään aikasarjojen sekä $ARIMA$-mallien teoriaa. Aluksi käsitellään aikasarjoja, minkä jälkeen asetetaan pohjatietoja $ARIMA$-malleja varten. Tämän jälkeen käsitellään $ARIMA$-mallien muodostamisesta ei-kausittaisille aikasarjoille ja luvun lopussa kausittaisille aikasarjoille.

2.1 Aikasarjat

Aikasarjat ovat jonomuotoisia havaintoaineistoja, jotka ilmaisevat muuttujan kehitystä ajan kuluessa. Joskus aikasarjoja esitetään myös taulukoina, jolloin taulukossa on useampi kuin yksi aikasarja ja aikasarjojen havaintoajankohdat ovat samat. Havaintoarvot esiintyvät aikasarjoissa yleensä tasaisin väliajoin, esimerkiksi vuosittain tai neljännesvuosittain. Aikasarjoja voidaan muodostaa lukemattomista eri asioista, kuten esimerkiksi lämpötiloista, matkustajamääristä tai toimialoittaisten tuotantojen arvoista. Aikasarjojen avulla voidaan tehdä ennusteita tutkittavien muuttujien kehitykestä [2, s. 1].

Määritelmä 2.1.1. Olkoon (Ω, Γ, P) todennäköisyysavaruus ja olkoon T indeksijoukko. Reaaliarvoinen aikasarja on joukossa $T \times \Omega$ määritelty reaaliarvoinen funktio $X(t, \omega)$ siten, että jokaisella kiinteällä muuttujan t arvolla $X(t, \omega)$ on satunnaismuuttuja todennäköisyysavaruudessa (Ω, Γ, P). Funktiosta $X(t, \omega)$ käytetään usein merkintöjä $X_t(\Omega)$ tai X_t, jolloin aikasarjat voidaan ilmaista kokoelmina $\{X_t : t \in T\}$.

Kiinteällä muuttujan ω arvolla funktio $X(t, \omega)$ on muuttujan t reaaliarvoinen funktio. Tarkasteltaessa joidenkin tallennettujen aikasarjojen, kuten bruttokansantuotteiden kuvaajia, on tärkeää ymmärtää, että silloin tarkastellaan sellaisten funktioiden $X(t, \omega)$ kuvaajia, joissa ω on kiinnitetty.

Jos indeksijoukko sisältää tasan yhden alkion, kyseessä on yhden satunnaismuuttujan stokastinen prosessi [3, s. 3]. Tällöin satunnaismuuttuja X_t on avaruudessa Ω määritelty reaaliarvoinen funktio siten, että joukko $\{\omega : X(\omega) \leq x\}$ on joukon Γ osajoukko jokaisella $x \in \mathbb{R}$. Tällöin funktiota $F_{X_t}(x) = P(\{\omega : X(\omega) \leq x\})$ kutsutaan satunnaismuuttujan X_t *kertymäfunktioksi*. [3, s. 2]

Useamman kuin yhden muuttujan stokastisille prosesseille täytyy muodostaa *yhteisjakaumafunktio*. Kokoelman $\{X_t : t \in T\}$ äärellisen satunnaismuuttujajoukon $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ yhteisjakaumafunktio määritellään asettamalla

$$F_{X_{t_1},\ldots,X_{t_n}}(x_{t_1}, \ldots, x_{t_n}) = P\{\omega : X(t_1, \omega) \leq x_{t_1}, \ldots, X(t_n, \omega) \leq x_{t_n}\}. \quad (2.1.1)$$

[3, s. 3]

Aikasarjojen analysoinnin kannalta havaintoarvojen on oltava äärellisiä, mutta havaintoarvoja voi olla periaatteessa ääretön määrä [4, s. 6].

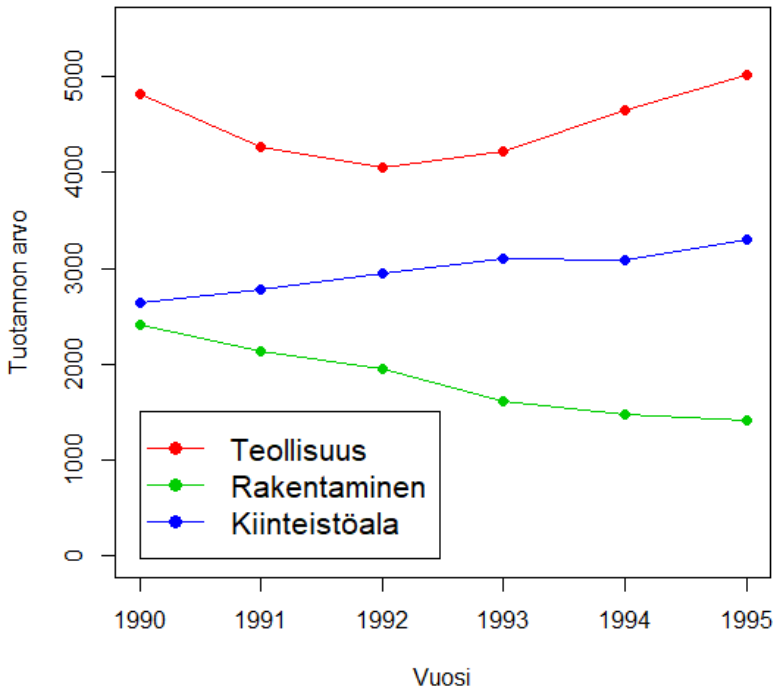
Usein $ARIMA$-malleja sovellettaessa aikasarjoihin havaintoarvoja käytetään jostain tietystä ajanhetkestä alkaen tai johonkin ajanhetkeen asti, jolloin tutkittavat aikasarjat voivat olla joko äärellisen, puoliäärettömän tai äärettömän mittaisia. Seuraavaksi esitän erään pienen esimerkin aikasarjoista.

Esimerkki 2.1.2. Esimerkkinä aikasarjoista toimikoon otos datasta [1], jota analysoin myöhemmin tutkielmassani. Taulukossa 1 on esitetty kolmen eri toimialan: teollisuuden, rakentamisen ja kiinteistöalan, tuotannot vuodesta 1990 vuoteen 1995. Kukin näistä toimialoista muodostaa siis oman aikasarjansa. Havainnot on poimittu kunkin vuoden ensimmäisestä neljänneksestä. Havaintoarvojen mittayksikkö on alkuperäinen sarja, miljoonaa euroa, kiintein hinnoin, ja niiden viitevuosi on 2010. Samat aikasarjat ovat havainnollistettu kuvassa 1.

muuttuja	1990	1991	1992	1993	1994	1995
Teollisuus	4825	4264	4050	4220	4657	5014
Rakentaminen	2417	2133	1954	1613	1470	1409
Kiinteistöala	2642	2783	2947	3107	3093	3306

Taulukko 1: Teollisuuden, rakentamisen ja kiinteistöalan aikasarjat 1990-1995

Aikasarjojen havaintoarvot eivät yleensä muutu kovin nopeasti. Tämän, sekä havaintojen tiheyden, takia peräkkäiset havaintoarvot yleensä korreloivat keskenään. Tällaista peräkkäisten havaintoarvojen välistä korrelaatiota kutsutaan *autokorrelaatioksi*. Autokorreloituneelle datalle monet tavallisista mallintamistavoista ovat usein harhaanjohtavia, koska ne perustuvat oletukseen havaintojen riippumattomuudesta. Siksi autokorreloituneelle datalle on käytettävä vaihtoehtoisia mallintamistapoja, jotka huomioivat havaintojen välisen riippuvuuden. Tämä voidaan toteuttaa soveltamalla aikasarjoille *autoregressiivisia integroituja liukuvan keskiarvon malleja*, eli $ARIMA$*-malleja*. $ARIMA$ on lyhenne sanoista *autoregressive integrated moving average* [2, s. 1].



Kuva 1: Esimerkin 2.1.2 aikasarjojen kuvaajat

2.2 Pohjatietoja ARIMA-malleja varten

Ennen $ARIMA$-malleihin menemistä on syytä asettaa muutamia määritelmiä, jotka tulevat vastaan $ARIMA$-malleja esiteltäessä tai niitä käytettäessä.

Johdatukseksi näihin määritelmiin esitetään neljä todennäköisyyslaskennasta mahdollisesti tuttua määritelmää. Koska aikasarjat koostuvat satunnaismuuttujista ja niiden jakaumat ovat diskreettejä, keskitytään näissä määritelmissä satunnaismuuttujien tunnuslukujen määrittelyyn.

Määritelmä 2.2.1. Satunnaismuuttujan X *odotusarvo* määritetään asettamalla

$$E(X) = \sum_{x_i : S_X} x_i P(X = x_i). \tag{2.2.1}$$

Odotusarvo on olemassa silloin kun summa $\sum_{x_i:S_X} x_i P(X = x_i)$ suppenee itseisesti [5, s. 58]. Toisinaan satunnaismuuttujan X odotusarvoa merkitään myös kirjaimella μ_X.

Määritelmä 2.2.2. Olkoon $E(X) = \mu$. Tällöin satunnaismuuttujan X *varianssi* määritetään asettamalla

$$Var(X) = E([X - E(X)]^2) = \sum_{x_i : S_X} (x_i - \mu)^2 P(X = x_i). \qquad (2.2.2)$$

Varianssi määrittelee kuinka paljon satunnaismuuttujan X arvot vaihtelevat odotusarvon ympärillä. Varianssista käytetään myös merkintää $Var(X) = \sigma_X^2$, missä σ_X on satunnaismuuttujan X *keskihajonta* [5, s. 61].

Määritelmä 2.2.3. Olkoon μ_X satunnaismuuttujan X odotusarvo ja olkoon μ_Y satunnaismuuttujan Y odostusarvo. Tällöin satunnaismuuttujien X ja Y välinen *kovarianssi* määritellään asettamalla

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y). \quad (2.2.3)$$

Kovarianssi on äärellinen vain silloin kun kaikki edellisen rivin odotusarvot ovat äärellisiä. Jos $X = Y$, niin tällöin $Cov(X, Y) = Var(X) = Var(Y)$. Mainittakoon lisäksi, että $Cov(X, Y) = Cov(Y, X)$, ja että $Cov(X, Y) = 0$, kun X ja Y ovat riippumattomia.

[5, s. 134]

Satunnaismuuttujien X ja Y keskihajonnat σ_X ja σ_Y vaikuttavat niiden välisen kovarianssin suuruuteen. Siksi eri tapauksia vertailtaessa kovarianssi ei ole yksikäsitteinen. Siksi kannattaa vertailla mieluummin normeerattujen satunnaismuuttujien $\frac{X-\mu_X}{\sigma_X}$ ja $\frac{Y-\mu_Y}{\sigma_Y}$ välistä kovarianssia, jota kutsutaan *korrelaatiokertoimeksi* [5, s. 135]. Satunnaismuuttujien välistä lineaarista riippuvuutta mitataan korrelaation avulla. [7, s. 5]

Määritelmä 2.2.4. Korrelaatiokerroin ρ määritetään asettamalla

$$\rho = Cov\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}, \qquad (2.2.4)$$

missä kovarianssin ja varianssien täytyy olla äärellisiä ja varianssien on oltava erisuuria kuin nolla. Jos puolestaan $Cov(X, Y) = 0$, niin tällöin myös $\rho = 0$. Muissa tapauksissa kovarianssi ja korrelaatiokerroin ovat samanmerkkisiä. Jos satunnaismuuttujien X ja Y välinen korrelaatiokerroin $\rho = 0$, niin tällöin X ja Y ovat *korreloimattomia*. Korrelaatiokertoimelle ρ pätee aina
$-1 \leq \rho \leq 1$, sekä lisäksi $|\rho| = 1$, jos on olemassa vakiot a ja b, $a \neq 0$, siten että $Y = aX + b$. [5, ss. 135-136]

Stationaariset aikasarjat muodostavat aikasarja-analyysin perustan. Pääasiassa stationaarisuuden ansiosta aikasarjoista voidaan rakentaa malleja ja ennusteita. [2, s. 48]

Määritelmä 2.2.5. Aikasarja x_t on *heikosti stationaarinen*, jos sen muuttujan x_t odostusarvo ja muuttujien x_t ja x_{t-h} väliset varianssit ovat vakioita ajan suhteen, sekä havaintojen x_t ja x_{t-h} väliset korrelaatiot riippuvat ainoastaan ajanhetkien etäisyyksistä, eli mielivaltaisesta luvusta $h \in \mathbb{Z}$. Toisin sanoen aikasarjan oletetaan olevan heikosti stationaarinen, jos sen odotusarvo $E(x_t) = \mu$ riippuu ainoastaan ajanhetkestä $t \in \mathbb{Z}$ ja kovarianssi $Cov(x_t, x_{t-h})$ riippuu ainoastaan luvusta h jokaiselle luvulle t. [2, s. 49]

Vahvasti stationaarisille aikasarjoille on voimakkaammat vaatimukset, mutta ne sivuutetaan, koska niitä ei tarvita myöhemmin tässä tutkielmassa. Ne löytyvät viitteen [2] sivulta 49. Myöhemmin tässä tutkielmassa stationaarisuudesta puhuttaessa tarkoitetaan heikkoa stationaarisuutta.

Sovellettaessa $ARIMA$-malleja dataan täytyy valita kullekin aikasarjalle sopivin $ARIMA$-malli. Sopivimman $ARIMA$-mallin löytämiseksi täytyy tarkastella mallien *autokorrelaatio-* ja *osittaisautokorrelaatiofunktioita* ACF ja $PACF$. [2, s. 56] [7, s. 21]

Lemma 2.2.6. *Olkoon sarja x_t stationaarinen. Tällöin*

$$Cov(x_t, x_{t-h}) = Cov(x_0, x_{-h}) = Cov(x_0, x_h) \tag{2.2.5}$$

sekä

$$\begin{aligned} Var(x_t) = Var(x_{t-h}) &\Rightarrow \sqrt{Var(x_t)Var(x_{t-h})} = \sqrt{Var(x_t)^2} \\ &= Var(x_t) = Cov(x_t, x_t) = Cov(x_t, x_{t-0}) = Cov(x_0, x_0) \\ &= Var(x_0). \end{aligned} \tag{2.2.6}$$

[7, ss. 5-6]

Tällöin sarjan x_t autokorrelaatiofunktio ACF määritetään asettamalla

$$\rho_h = \frac{Cov(x_t, x_{t-h})}{\sqrt{Var(x_t)Var(x_{t-h})}} = \frac{Cov(x_0, x_h)}{Var(x_0)}, \tag{2.2.7}$$

missä ρ_h on havaintojen x_t ja x_{t-h} välinen korrelaatiokerroin [6, s. 21]. Huomataan, että stationaarisuuden seurauksena korrelaatiokerroin ρ_h riippuu ainoastaan luvusta h, ja että $\rho_h = 1$, kun $h = 0$ [7, ss. 5-6]. Cauchy-Schwarzin epäyhtälön [6, s. 18] nojalla $-1 \leq \rho_h \leq 1$ jokaiselle $h \in \mathbb{Z}$. [6, s. 21]

Määritelmä 2.2.7. Kun sarja x_t on stationaarinen, sen osittaisautokorrelaatiofunktiosta PACF käytetään merkitää ϕ_{hh}, jossa $h \in \mathbb{N}$. Olkoon $h = 1$. Tällöin

$$\phi_{11} = Corr(x_1, x_0) = \rho(1). \tag{2.2.8}$$

Kun $h \geq 2$,

$$\phi_{hh} = Corr(x_h - x_h^{h-1}, x_0 - x_0^{h-1}), \tag{2.2.9}$$

missä $x_h^{h-1} = E(x_h|x_1, \ldots, x_{h-1})$, $x_0^{h-1} = E(x_0|x_1, \ldots, x_{h-1})$ ja merkintä $Corr(a, b)$ on korrelaatiokerroin $\rho_{a,b}$. Nyt $x_h - x_h^{h-1}$ ja $x_0 - x_0^{h-1}$ ovat riippumattomia sarjan $\{x_1, \ldots, x_{h-1}\}$ kanssa. Stationaarisuuden nojalla ϕ_{hh} on ehdollinen korrelaatio arvojen x_t ja x_{t-h} välillä jokaisella $t \in \mathbb{Z}$.

[6, s. 99] [7, s. 6]

Aikasarjoissa esiintyy usein jonkinlaista satunnaista heilahtelua, joka ei riipu muista tekijöistä. Jotta sellaiset saataisiin huomioitua mahdollisimman hyvin, määritellään *valkoinen kohina*. [2, s. 54]

Määritelmä 2.2.8. Olkoon w_t sarja, jonka havaintoarvot ovat korreloimattomia satunnaismuuttujia, joiden odotusarvo on 0 ja varianssi σ_w^2 on äärellinen. Tällöin sarjaa w_t kutsutaan valkoiseksi kohinaksi. Tällöin merkitään $w_t \sim wn(0, \sigma_w^2)$.

Jos sarjan w_t havaintoarvot koostuvat samoin jakautuneista riippumattomista satunnaismuuttujista odotusarvonaan 0 ja varianssinaan σ_w^2, sarjaa kutsutaan *valkoiseksi riippumattomaksi kohinaksi*. Tällöin merkitään $w_t \sim iid(0, \sigma_w^2)$.

Lisäksi jos w_t on normaalijakautunut odotusarvonaan 0 ja varianssinaan σ_w^2, merkitään $w_t \sim iidN(0, \sigma_w^2)$. Tätä kutsutaan *normaaliseksi valkoiseksi kohinaksi*. [6, s. 9]

Kuhunkin aineistoon sopivinta $ARIMA$-mallia valittaessa on syytä vertailla malleja toisiinsa muun muassa *informaatiokriteerien* avulla. Ennen itse informaatiokriteereihin menoa asetetaan seuraava määritelmä.

Määritelmä 2.2.9. Varianssin $\hat{\sigma}_k^2$ *suurimman uskottavuuden estimaattori* MLE kirjoitetaan muodossa

$$\hat{\sigma}_k^2 = \frac{SSE_k}{n}, \tag{2.2.10}$$

jossa SSE_k tarkoittaa residuaalien neliöiden summaa, k on mallin parametrien lukumäärä, ja n on koko aineiston havaintojen lukumäärä.

Akaike esitti, että mallin hyvyyttä voitaisiin mitata etsimällä tasapainoa sopivan virhetermin ja mallin parametrien lukumäärän välillä.

Määritelmä 2.2.10. *Akaiken informaatiokriteeri* AIC määritellään seuraavasti:

$$AIC = \log \hat{\sigma}_k^2 + \frac{n + 2k}{n}, \tag{2.2.11}$$

jossa $\hat{\sigma}_k^2$, k ja n ovat kuten määritelmässä 2.2.9. [6, s. 51] [7, s. 7]

Tässä määritelmässä sopivin malli löydetään sillä asteluvulla k, jolla AIC saa pienimmän arvonsa. Akaiken informaatiokriteeri saattaa kuitenkin käyttää liian suurta parametrien lukumäärää k, minkä takia on olemassa *Akaiken korjattu informaatiokriteeri* $AICC$. [7, s. 8]

Määritelmä 2.2.11. Akaiken korjattu informaatiokriteeri $AICC$ määritellään seuraavasti:

$$AICC = \log \hat{\sigma}_k^2 + \frac{n + k}{n - k - 2}, \tag{2.2.12}$$

jossa $\hat{\sigma}_k^2$, k ja n ovat kuten määritelmässä 2.2.9. [6, s. 51]

$AICC$ toimii erinomaisesti pienemmille aineistoille, mutta suurempia aineistoja varten määritelään vielä *Bayesin informaatiokriteeri* BIC.

Määritelmä 2.2.12. Bayesin informaatiokriteeri BIC määritetään asettamalla

$$BIC = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}, \tag{2.2.13}$$

jossa $\hat{\sigma}_k^2$, k ja n ovat kuten määritelmässä 2.2.9.

Joissain yhteyksissä Bayesin informaatiokriteeriä kutsutaan *Schwarzin informaatiokriteeriksi*, mutta tässä tutkielmassa siitä puhutaan Bayesin informaatiokriteerinä tai käytetään lyhennettä BIC. [6, s. 52]

2.3 ARIMA-mallin muodostaminen

$ARIMA$-mallin muodostamiseen tarvitaan *autoregressiivista mallia* $AR(p)$, *liukuvan keskiarvon mallia* $MA(q)$ sekä niiden yhdistelmää *autoregressiivista liukuvan keskiarvon mallia* $ARMA(p, q)$ [2, s.84][2, s. 59]. Siksi nämä mallit esitellään ennen varsinaista $ARIMA$-mallia.

Autoregressiivisessa mallissa $AR(p)$ aikasarjan x_t nykyistä arvoa selitetään sen edellisten havaintoarvojen $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ funktiona, missä $p \in \mathbb{N}$ määrittelee montako edellistä havaintoarvoa tarvitaan nykyisen arvon x_t ennustamiseksi. Tällöin x_t saadaan määriteltyä sen edellisten havaintoarvojen lineaarikombinaation sekä virhetermin summana. [6, ss. 77-78]

Määritelmä 2.3.1. Autoregerssiivisen mallin $AR(p)$ yleinen astetta p oleva yhtälö on muotoa

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{2.3.1}$$

missä x_t on stationaarinen, $w_t \sim wn(0, \sigma_w^2)$ ja $\phi_1, \phi_2, \ldots, \phi_p$ ovat vakioita, $p \in \mathbb{N}$, ja $\phi_p \neq 0$. Yhtälössä (2.3.1) muuttujan x_t odotusarvon oletetaan olevan 0. Mikäli muuttujan x_t odotusarvo $\mu \neq 0$, sijoitetaan tällöin yhtälöön (2.3.1) muuttujan x_t tilalle $x_t - \mu$, jolloin saadaan

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t, \tag{2.3.2}$$

tai

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{2.3.3}$$

missä $\alpha = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$ [6, s. 78].

Autoregressiivinen operaattori on muotoa

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \tag{2.3.4}$$

missä B on viiveoperaattori, jolle pätee $Bx_t = x_{t-1}$, ja sen potensseille B^n pätee $B^n x_t = x_{t-n}$. [6, s. 79] [10, s. 20]

Liukuvan keskiarvon mallissa $MA(q)$ sarjan x_t nykyistä arvoa selitetään valkoisen kohinan $w_t, w_{t-1}, \ldots, w_{t-q}$ lineaarikombinaationa, missä $q \in \mathbb{N}$ kuvaa liukuvan keskiarvon viivettä. Liukuvan keskiarvon malli on stationaarinen kaikilla parametrien $\theta_1, \theta_2, \ldots, \theta_q$ arvoilla. [6, s. 83] [7, s. 13]

Määritelmä 2.3.2. Liukuvan keskiarvon mallin $MA(q)$ astetta q oleva yhtälö on muotoa

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \tag{2.3.5}$$

missä $w_t \sim wn(0, \sigma_w^2)$ ja $\theta_1, \theta_2, \ldots, \theta_q$ ovat parametreja siten, että $\theta_q \neq 0$.

Liukuvan keskiarvon operaattori on muotoa

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q. \tag{2.3.6}$$

[6, s. 83]

Seuraavaksi muodostetaan **autoregressiivinen liukuvan keskiarvon malli** $ARMA(p, q)$, jotta saadaan malli, joka huomioi sekä aiempien havaintoarvojen että valkoisen kohinan lineaarikombinaatiot. $ARMA(p, q)$ on autoregressiivisen ja liukuvan keskiarvon mallien yhdistelmä.

Määritelmä 2.3.3. Aikasarja x_t on $ARMA(p, q)$-malli, jos se on stationaarinen, ja jos se voidaan kirjoittaa muodossa

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \qquad (2.3.7)$$

missä p ja q ovat mallin asteluvut, $\phi_p \neq 0$, $\theta_q \neq 0$ ja $\sigma_w^2 > 0$. Jos sarjan x_t odotusarvo $\mu \neq 0$, niin tällöin voidaan määrittää $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$, jolloin saadaan

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \quad (2.3.8)$$

missä $w_t \sim wn(0, \sigma_w^2)$.

$ARMA(p, 0)$-malli on sama asia kuin $AR(p)$-malli ja vastaavasti $ARMA(0, q)$-malli tarkoittaa $MA(q)$-mallia. [6, ss. 85-86]

Huomautus 2.3.4. $ARMA(p, q)$-mallin yleisessä määritelmässä saattaa esiintyä seuraavia ongelmia:

- Mallissa saattaa olla tarpeettomia parametreja.
- Stationaarinen AR-malli saattaa riippua sen tulevista arvoista.
- MA-mallit eivät ole välttämättä yksikäsitteisiä.

Jotta näistä ongelmista päästään yli, asetetaan mallin parametreille muutamia lisäehtoja seuraavien määritelmien avulla.

Määritelmä 2.3.5. AR- ja MA-mallien polynomit $\phi(z)$ ja $\theta(z)$ määritetään asettamalla

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad \phi_p \neq 0, \qquad (2.3.9)$$

sekä

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \quad \theta_q \neq 0, \qquad (2.3.10)$$

missä z on kompleksiluku ja $\phi_1, \ldots, \phi_p$ sekä $\theta_1, \ldots, \theta_q$ ovat kuten aiemmin. Tarpeettomien parametrien välttämiseksi $ARMA(p, q)$-mallin on aina oltava yksinkertaisimmassa muodossaan, eli polynomeilla $\phi(z)$ ja $\theta(z)$ ei saa olla yhteisiä tekijöitä. [6, s. 87]

Jotta vältytään siltä, että malli riippuu sen tulevista arvoista, määrittellään $ARMA(p,q)$-mallin *kausaalisuuden* käsite.

Määritelmä 2.3.6. $ARMA$-mallia kutsutaan kausaaliseksi, jos aikasarjat x_t, joissa $t \in \mathbb{Z}$, voidaan kirjoittaa yksipuolisena lineaarisena prosessina

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B) w_t, \tag{2.3.11}$$

missä $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ ja $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Olkoon lisäksi $\psi_0 = 1$. Prosessi on kausaalinen vain jos polynomin $\phi(z) = 1 - \phi z$ ratkaisun $z_0 = \frac{1}{\phi}$ moduli $|z_0| > 1$, koska $AR(1)$-prosessi $x_t = \phi x_{t-1} + w_t$ on kausaalinen ainoastaan silloin kun $|\phi| < 1$. [6, s. 87]

Yleisesti kausaalisilla $ARMA(p,q)$ -malleilla on seuraava ominaisuus.

Seuraus 2.3.7. *$ARMA(p,q)$-malli on kausaalinen jos ja vain jos $\phi(z) \neq 0$ jokaiselle $|z| \leq 1$. Määritelmän 2.3.6 summalausekkeen kertoimet ψ_j voidaan määrittää ratkaisemalla*

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1. \tag{2.3.12}$$

Määritelmä 2.3.6 voidaan ilmaista myös siten, että $ARMA(p,q)$-prosessi on kausaalinen ainoastaan silloin kun polynomin $\phi(z)$ juuret sijaitsevat yksikkökiekon ulkopuolella, koska $\phi(z) = 0$ vain silloin kun $|z| > 1$.
[6, ss. 87-88]

Jotta varmistutaan $ARMA(p,q)$-mallin yksikäsitteisyydestä, valitaan malli, joka mahdollistaa päättymättömän autoregressiivisen esityksen.

Määritelmä 2.3.8. $ARMA(p,q)$-mallia kutsutaan *kääntyväksi*, jos aikasarjat x_t, joissa $t \in \mathbb{Z}$, voidaan kirjoittaa muodossa

$$\pi(B) x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t, \tag{2.3.13}$$

missä $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ ja $\sum_{j=0}^{\infty} |\pi_j| < \infty$. Olkoon lisäksi $\pi_0 = 1$.

Vastaavasti kuin kausaalisuuden tapauksessa, myös kääntyvyydellä on seurauksen 2.3.7 kaltainen seuraus.

Seuraus 2.3.9. *$ARMA(p,q)$-malli on kääntyvä jos ja vain jos $\theta(z) \neq 0$ jokaiselle $|z| \leq 1$. Määritelmän 2.3.8 summalausekkeen $\pi(B)$ kertoimet π_j voidaan määrittää ratkaisemalla*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1. \tag{2.3.14}$$

Seuraus 2.3.9 voidaan ilmaista myös siten, että $ARMA(p,q)$-prosessi on käänteinen vain silloin kun polynomin $\theta(z)$ juuret sijaitsevat yksikkökiekon ulkopuolella, eli $\theta(z) = 0$ vain silloin kun $|z| > 1$. [6, s. 88]

Seuraava määritelmä lukeutuu periaatteessa $ARIMA$-mallin valinnassa tarvittaviin pohjatietoihin, mutta se esitetään vasta nyt, jotta $ARMA(p,q)$-mallia voidaan pitää tunnettuna.

Määritelmä 2.3.10. Olkoon x_t $ARMA(p,q)$-prosessi, jossa $w_t \sim iidN(0, \sigma_w^2)$, ja olkoon $\beta = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ mallin parametreista muodostettu $(p+q+1)$-vektori. Tällöin mallin *uskottavuusfunktio* on

$$L(\beta, \sigma_w^2) = (2\pi\sigma_w^2)^{-\frac{n}{2}} (r_1(\beta) r_2(\beta) \cdots r_n(\beta))^{-\frac{1}{2}} e^{-\frac{S(\beta)}{2\sigma_w^2}}, \tag{2.3.15}$$

missä

$$S(\beta) = \sum_{t=1}^{n} \frac{(x_t - x_t^{t-1}(\beta))^2}{r_t(\beta)}. \tag{2.3.16}$$

Tässä x_t^{t-1} ja r_t^{t-1} ovat vektorin β funktioita, $x_t^{t-1} = E(x_t | x_1, \ldots, x_{t-1})$ ja $r_t^{t-1} = Var(x_t | x_1, \ldots, x_{t-1})$.

[6, ss. 120-121][7, s.9]

Aiemmin esitetty $ARMA(p,q)$-malli edellyttää havaintoaineiston stationaarisuutta. Jotta tästä vaatimuksesta päästään eroon, muodostetaan **autoregressiivinen integroitu liukuvan keskiarvon malli** $ARIMA(p,d,q)$, joka on $ARMA(p,q)$-mallin laajennettu versio. Epästationaarinen $ARIMA$-malli voidaan muuttaa stationaariseksi *differoinnin* avulla. [6, ss. 133-134] $ARIMA(p,d,q)$-mallissa parametri $d \in \mathbb{N}_0$ ilmaisee differointien lukumäärän. [7, s. 19]

Määritelmä 2.3.11. Aikasarjaa x_t kutsutaan $ARIMA(p,d,q)$-malliksi, jos

$$\nabla^d x_t = (1-B)^d x_t \tag{2.3.17}$$

on $ARMA(p,q)$-malli. Yleisesti $ARIMA(p,d,q)$ voidaan kirjoittaa muodossa

$$\phi(B)(1-B)^d x_t = \theta(B)w_t. \tag{2.3.18}$$

Jos $E(\nabla^d x_t) = \mu$, yhtälö 2.3.18 voidaan kirjoittaa muodossa

$$\phi(B)(1-B)^d x_t = \delta + \theta(B)w_t, \tag{2.3.19}$$

missä $\delta = \mu(1-\phi_1-\cdots-\phi_p)$.

[6, s. 134]

Yhtälössä (2.3.17) merkintä ∇ on differenssioperaattori ja B on viiveoperaattori. [7, s. 19]

2.4 Kausittaiset ARIMA-mallit

Aikasarjoissa esiintyy usein *kausittaista vaihtelua*, joka voi olla sidoksissa esimerkiksi vuodenaikaan. Esimerkkejä tällaisista ovat muun muassa lämpötilat sekä erilaiset biologiset, fysikaaliset ja taloudelliset prosessit [6, s. 148]. Koska myöhemmin tutkielmassa käsiteltävässä datassa esiintyy kausittaisia vaihteluita, esitetään seuraavaksi kuinka $ARIMA$-malleja voidaan soveltaa kausittain vaihteleville aikasarjoille.

Määritelmä 2.4.1. *Kausittainen autoregressiivinen liukuvan keskiarvon malli* $ARMA(P,Q)_s$ ilmaistaan muodossa

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \tag{2.4.1}$$

missä

$$\Phi_P(B^s) = 1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \tag{2.4.2}$$

on astetta P oleva kausittainen autoregressiivinen operaattori ja

$$\Theta_Q(B^s) = 1+\Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \tag{2.4.3}$$

on astetta Q oleva kausittainen liukuvan keskiarvon operaattori ja s on kauden pituus aikasarjan havaintoarvoissa mitattuna. [6, s. 148]

Määritelmä 2.4.1 huomioi ainoastaan aikasarjan kausittaisen vaihtelun eikä se ota huomioon aikasarjan edellisiä havaintoarvoja, minkä takia on syytä yhdistää kausittaisen aikasarjan kausittaiset ja ei-kausittaiset operaattorit, jolloin saadaan *divisiivinen kausittainen autoregressiivinen liukuvan keskiarvon malli.*

Määritelmä 2.4.2. Divisiivinen kausittainen autoregressiivinen liukuvan keskiarvon malli $ARMA(p,q) \times (P,Q)_s$ kirjoitetaan muodossa

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t, \tag{2.4.4}$$

jossa merkinnät ovat vastaavat kuin aiemmissa määritelmissä.

[6, s. 150]

Jotta myös kausittaisten aikasarjojen osalta päästään eroon $ARMA$-mallien stationaarisuusvaatimuksesta, määritellään myös kausittaisille aineistoille oma $ARIMA$-malli laajentamalla määritelmä 2.4.2 *divisiiviseksi kausittaiseksi autoregressiiviseksi integroiduksi liukuvan keskiarvon malliksi* $ARIMA(p,d,q)\times(P,D,Q)_s$.

Määritelmä 2.4.3. Divisiivinen kausittainen autoregressiivinen integroitu liukuvan keskiarvon malli $ARIMA(p,d,q) \times (P,D,Q)_s$ määritetään asettamalla

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t, \tag{2.4.5}$$

missä $w_t \sim iidN(0,\sigma_w^2)$, $\phi(B)$ ja $\theta(B)$ ovat tavallisessa $ARMA$-mallissa esiintyvät operaattorit, ja p ja q ovat niiden asteluvut. $\Phi_P(B^s)$ ja $\Theta_Q(B^s)$ ovat kausittaisen autoregressiivisen liukuvan keskiarvon mallin komponentit astelukuinaan P ja Q, sekä $\nabla^d = (1-B)^d$ on tavallinen ja $\nabla_s^D = (1-B^s)^D$ on kausittainen differenssikomponentti, joissa d ja D kertovat differointien lukumäärät, B on viiveoperaattori, ja s on kauden pituus aikasarjan havaintoarvoissa mitattuna. [6, s. 152]

2.5 Ennustevirheiden tunnusluvut

Luvussa 4 analysoidaan $ARIMA$-mallien avulla luotuja ennusteita toimialoittain sekä neljännesvuosittain hyödyntäen muutamia ennustevirheitä mittaavia tunnuslukuja, joiden laskukaavat määritellään seuraavaksi.

Määritelmä 2.5.1. Olkoon $a_t \neq 0$ jokaiselle $t \in \mathbb{N}$. Tällöin *prosentuaalinen keskivirhe (MPE)* määritetään asettamalla

$$MPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{a_t - f_t}{a_t}, \tag{2.5.1}$$

jossa n on ennusteiden lukumäärä, a_t on toteutunut arvo ja f_t on ennusteen arvo kullakin $t \in \mathbb{N}$.

Prosentuaalisen keskivirheen arvo voi olla joko positiivinen tai negatiivinen ja se on sitä lähempänä nollaa mitä parempi ennuste on.

Tässä tutkielmassa toimialoja sekä BKT:a tutkittaessa $n = 8$, ja vuosineljänneksiä tutkittaessa $n = 11$, koska ennusteet luodaan kahdeksalle eri vuosineljännekselle ja tutkittavia aikasarjoja on yksitoista kappaletta.

Prosentuaalisessa keskivirheessä positiiviset ja negatiiviset ennustevirheet kumoavat toisiaan mikäli sen summalausekkeessa esiintyy niitä molempia. Jotta saadaan käyttöön prosentuaalinen tunnusluku, jossa vältytään tältä, määritellään myös *absoluuttinen prosentuaalinen keskivirhe (MAPE)*.

Määritelmä 2.5.2. Olkoon $a_t \neq 0$ jokaiselle $t \in \mathbb{N}$. Tällöin absoluuttinen prosentuaalinen keskivirhe ($MAPE$) määritetään asettamalla

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{a_t - f_t}{a_t} \right|, \tag{2.5.2}$$

jossa n, t, a_t, ja f_t ovat kuten määritelmässä 2.5.1.

Absoluuttisen prosentuaalisen keskivirheen arvo on aina ei-negatiivinen ja se on sitä lähempänä nollaa, mitä tarkempi ennuste on.

Molemmat edellä esitetyt ennustevirheet ovat keskenään melko samanlaisia ja ne perustuvat L^1-normiin, joten määritellään vielä L^2-normiin perustuva *keskineliövirhe (MSE)*.

Määritelmä 2.5.3. Aikasarjan toteumien keskiarvolla $\overline{a_t}$ skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) määritetään asettamalla

$$\frac{MSE}{\overline{a_t}} = \frac{1}{\overline{a_t}} \sum_{t=1}^{n} (a_t - f_t)^2, \tag{2.5.3}$$

jossa $\overline{a_t} = \frac{1}{n} \sum_{t=1}^{n} a_t$ ja n, t, a_t sekä f_t ovat kuten määritelmässä 2.5.1.

Keskineliövirheen arvo on aina ei-negatiivinen ja se on sitä lähempänä nollaa, mitä parempi ennuste on. Se reagoi herkemmin yksittäisiin huonoihin ennusteisiin kuin L^1-normiin perustuvat ennustevirheet. [11]

Koska tutkittavien aikasarjojen keskineliövirheet riippuvat tuotantojen suuruusluokista, luvussa 4 aikasarjojen ennusteita analysoidaan nimenomaan aikasarjojen ennustettavien vuosineljännesten toteumien keskiarvoilla skaalatuilla keskineliövirheillä. Tällöin toimialojen ennusteita voidaan vertailla toi-

siinsa myös keskineliövirheiden avulla. Vastaavasti alaluvussa 4.12 keskineliövirheet skaalataan kunkin vuosineljänneksen aikasarjojen toteumien keskiarvolla, jotta aikasarjojen kausittaisvaihteluiden ja trendien vaikutukset keskineliövirheiden arvoihin saadaan poistettua.

3 Tutkittavat aikasarjat ja ARIMA-mallin valinta

Tämän luvun aluksi esitellään tutkimuksessa käytettävää dataa, johon myöhemmin tässä tutkielmassa sovelletaan edellisessä luvussa esitettyä teoriaa. Edellisessä luvussa esitetyn teorian pohjalta alaluvussa 3.2 näytetään vaihe vaiheelta bruttokansantuotteen aikasarjan avulla kuinka aikasarjalle valitaan sille sopivin $ARIMA$-malli.

3.1 Tutkittavat aikasarjat

Tutkittava data [1] on neljännesvuosidata vuoden 1990 alusta vuoden 2019 toiseen neljännekseen asti. Datassa on kymmenen eri toimialaa, jotka ovat teollisuus; koulutus, terveys- ja sosiaalipalvelut; kauppa, liikenne majoitus ja ravitseminen; kiinteistöala; hallinto- ja tukipalvelut; rakentaminen; informaatio ja viestintä; taiteet, viihde ja virkistys; rahoitus ja vakuutukset, sekä maa-, metsä- ja kalatalous. Lisäksi yhtenä datan muuttujana on bruttokansantuotteen kokonaisarvo. Datan muuttujien mittayksikkö on alkuperäinen sarja, miljoonaa euroa, kiintein hinnoin, ja niiden viitevuosi on 2010.

Määritelmä 3.1.1. Mittayksikön *kiinteähintaisuus* tarkoittaa sitä, että tuotannon arvo V_{t_0} lasketaan käyttäen hintaa p_{t_0}, joka on kiinnitetty jollekin tietylle ajanhetkelle t_0, tässä tutkielmassa vuodelle 2010. Näin ollen kiinteähintaisen sarjan tuotannon arvo saadaan laskettua kaavalla

$$V_{t_0}\left(\frac{€}{v}\right) = p_{t_0}\left(\frac{€}{m}\right) \cdot q_t\left(\frac{m}{v}\right), \tag{3.1.1}$$

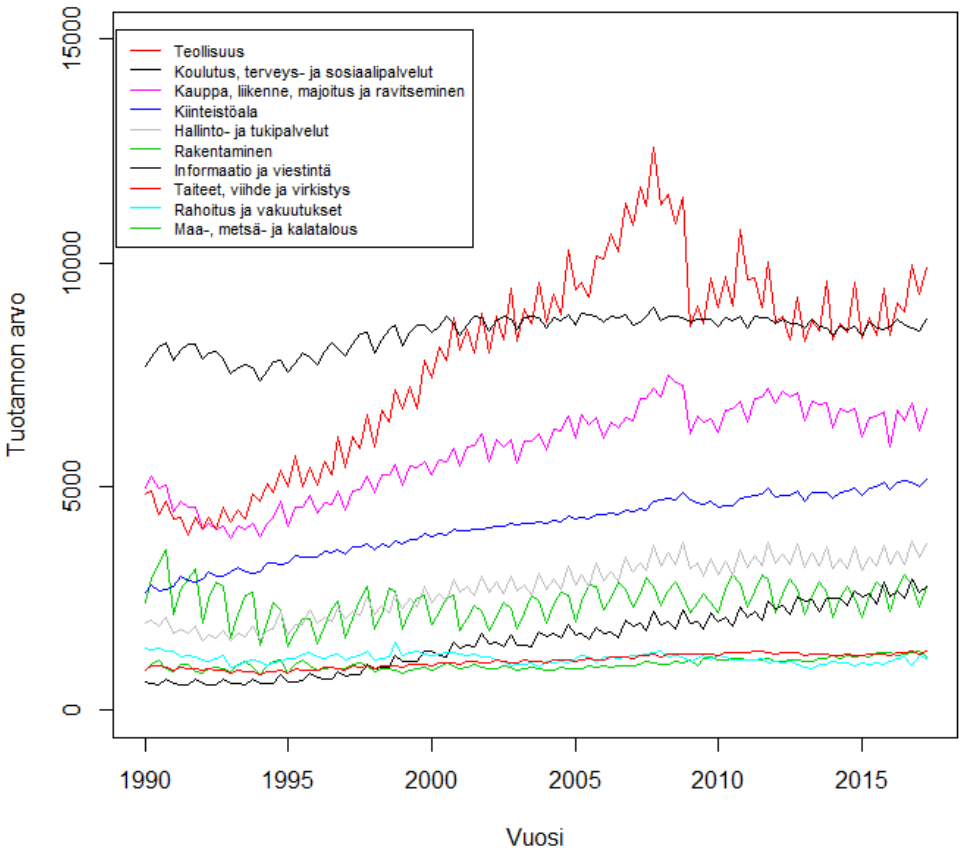
jossa q_t tarkoittaa tuotannon määrää, t aikaa (jota ei ole kiinnitetty), sekä $\left(\frac{€}{v}\right)$, $\left(\frac{€}{m}\right)$ ja $\left(\frac{m}{v}\right)$ ovat tuotannon arvon, hinnan ja tuotantomäärän mittayksiköt. Mittayksiköissä v tarkoittaa vuotta ja m tuotannon määrän mittayksikköä, jonka suure voi vaihdella toimialoittain.

Tuotannon arvo *käyvin hinnoin* voidaan laskea korvaamalla yhtälössä 3.1.1 esiintyvät alaindeksit t_0 alaindekseillä t. [8]

Koska datassa on 118 havaintoarvoa kullekin yhdelletoista muuttujalle, sen esittäminen taulukkomuotoisena tässä tutkielmassa ei ole käytännössä mahdollista saati järkevää. Siksi data on havainnollistettu graafisesti kuvissa 2 ja 3. Kuvassa 2 näkyvät toimialoittaisten tuotantojen aikasarjojen kuvaajat vuoden 1990 alusta vuoden 2017 toiseen neljännekseen asti. Kuvan toimialat ovat kerrottu selitteessä jotakuinkin siinä järjestyksessä, jossa niiden kuvaajat esiintyvät tuotannon arvossa mitattuna. Kuvaajien perusteella voidaan

havaita eri toimialojen tuotantojen kehityksissä merkittäviä eroavaisuuksia. Monilta aloilta on löydettävissä selkeitä trendejä sekä kausittaisvaihteluita tai ainakin jompaakumpaa. Kuvan skaalaus saattaa heikentää pienempien kausittaisvaihteluiden havaittavuutta. 1990-luvun alun ja vuoden 2009 lamat näyttävät vaikuttaneen eri toimialoihin hyvin eri tavoin.

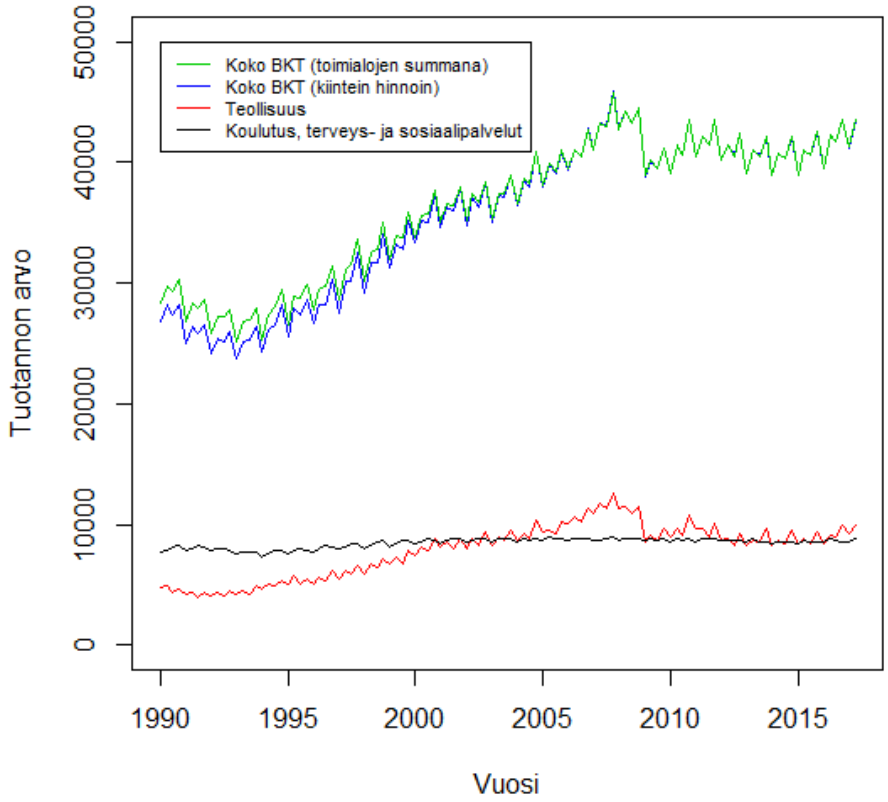
Luvussa 4 esitetään kuvat, joissa näkyy toimialojen sekä bruttokansantuotteen aikasarjojen kuvaajat vuoden 2017 kolmannesta neljänneksestä vuoden 2019 toiseen neljännekseen asti verrattuna ennusteisiin, jotka luodaan $ARIMA$-mallien avulla samassa luvussa.



Kuva 2: Suomen toimialoittaisten tuotantojen aikasarjojen kuvaajat 1/1990-2/2017

Kuvassa 3 näkyy Suomen bruttokansantuotteen kiinteähintaisen aikasar-

jan kuvaaja vuoden 1990 alusta vuoden 2017 toiseen neljännekseen asti. Vertailun vuoksi kuvassa esiintyy myös bruttokansantuotetta kuvaava toimialoittaisten tuotantojen summa, jonka perusteella ennustetaan BKT:a myöhemmin tutkielmassa. Toimialoittaisten tuotantojen summa vaikuttaa olevan tarkastelujakson alkupuolella jonkun verran kiinteähintaista BKT:a suurempi, mutta myöhemmin erittäin lähellä kiinteähintaisen BKT:n arvoa. Tämä johtunee siitä, että datan kiinteähintaisten aikasarjojen viitevuosi on 2010. Kuvassa näkyy myös edellisessä kuvassa esiintyvät teollisuuden sekä koulutuksen, terveys- ja sosiaalipalveluiden kuvaajat, jotta toimialoittaisten tuotantojen ja BKT:n väliset mittasuhteet hahmottuisivat. Kuvat 2 ja 3 ovat siis skaalattu keskenään hyvin eri tavoin tuotannon arvon suhteen.

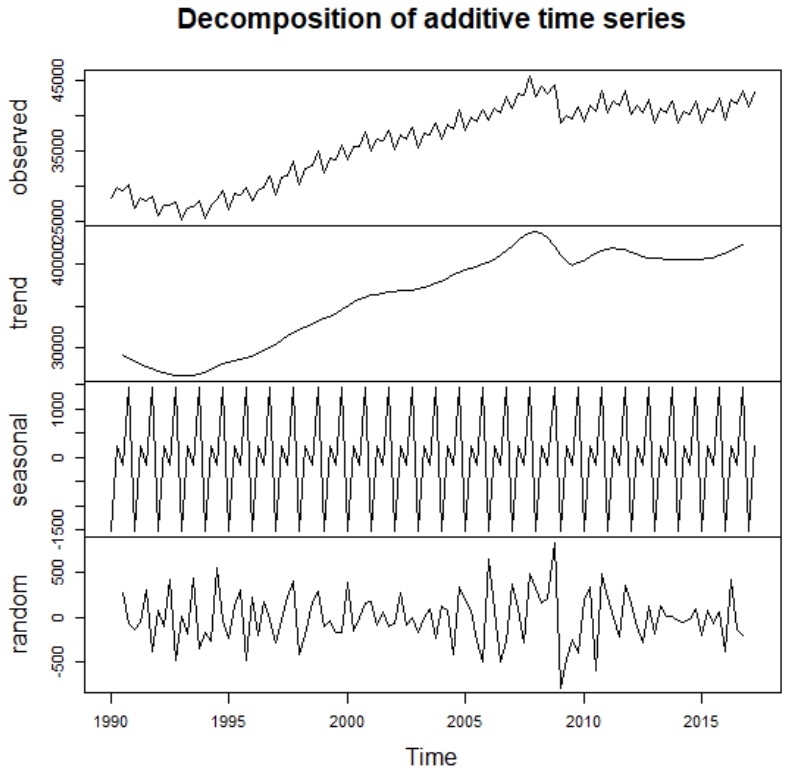


Kuva 3: Suomen bruttokansantuotteen aikasarjan kuvaajat sekä kiintein hinnoin että toimialojen summana 1/1990-2/2017 verrattuna suurimpiin edellisessä kuvassa esiintyviin toimialoittaisiin tuotantoihin

3.2 Sopivan ARIMA-mallin valinta

Myöhemmin tutkielmassa tullaan tutkimaan Suomen toimialoittaisten tuotantojen aikasarjadataa osittain R-ohjelman *auto.arima* -komennon avulla. Kyseinen komento tunnistaa aikasarjan ja pyrkii valitsemaan automaattisesti sen ennustamiseen sopivimman $ARIMA$-mallin. Ennen siihen menoa näytetään edellisessä luvussa esitetyn teorian perusteella vaihe vaiheelta kuinka aikasarjalle valitaan sille sopivin $ARIMA$-malli.

Tähän käytetään bruttokansantuotteen aikasarjaa, joka on laskettu toimialoittaisten tuotantojen summana. Kyseisen aikasarjan kuvaaja esiintyy vihreänä käyränä kuvassa 3. Jotta kyseisestä aikasarjasta voidaan tehdä tarkempia havaintoja, muodostetaan kuva 4, jossa näkyy aikasarjan kuvaajan lisäksi sen trendi, kausittaisvaihtelu sekä satunnaisheilahtelu.

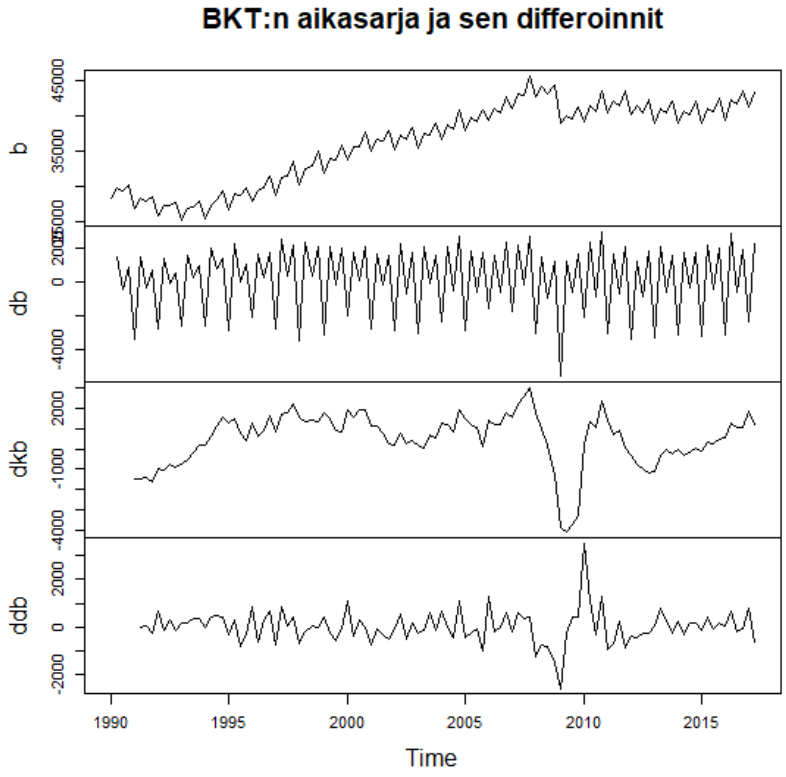


Kuva 4: BKT:n aikasarjan kuvaaja toimialojen summana (observed) sekä sen trendi (trend), kausittaisvaihtelu (seasonal) ja satunnaisheilahtelu (random)

Kuvasta 4 nähdään, että aikasarjassa esiintyy kausittaista vaihtelua ja

kauden pituus näyttäisi olevan yksi vuosi. Kausittaisuuden takia sovelletaan toimialoittaisten tuotantojen summista koostuvan BKT:n aikasarjaa määritelmään 2.4.3. Koska kyseessä on kausittainen neljännesvuosidata kauden pituutena yksi vuosi, sijoitetaan tällöin $s = 4$ yhtälöön 2.4.5.

Aikasarjan trendi on enimmäkseen nouseva, mutta tarkasteluajanjakson loppupuolella nouseva trendi kääntyy ensin hieman laskevaksi ja lopuksi melko tasaiseksi. Aikasarja saadaan muutettua stationaariseksi differoimalla. Kuvasta 5 nähdään aikasarjan tavallisesti differoidun (*db*), kausittaisdifferoidun (*dkb*) sekä molemmilla tavoilla differoidun sarjan (*ddb*) kuvaajat. Tavallisesti differoiduissa aikasarjoissa (*db* ja *ddb*) ei ole trendiä ja kausittaisdifferoiduissa aikasarjoissa (*dkb* ja *ddb*) ei ole kausittaisia vaihteluita, joten ne ovat stationaarisia. Näin ollen valitaan tarkasteltavaksi sekä tavallisesti että kausittain differoitu aikasarja (*ddb*). Koska BKT:n aikasarja on differoitu kerran sekä tavallisesti että kausittain, sijoitetaan $d = D = 1$ yhtälöön 2.4.5.



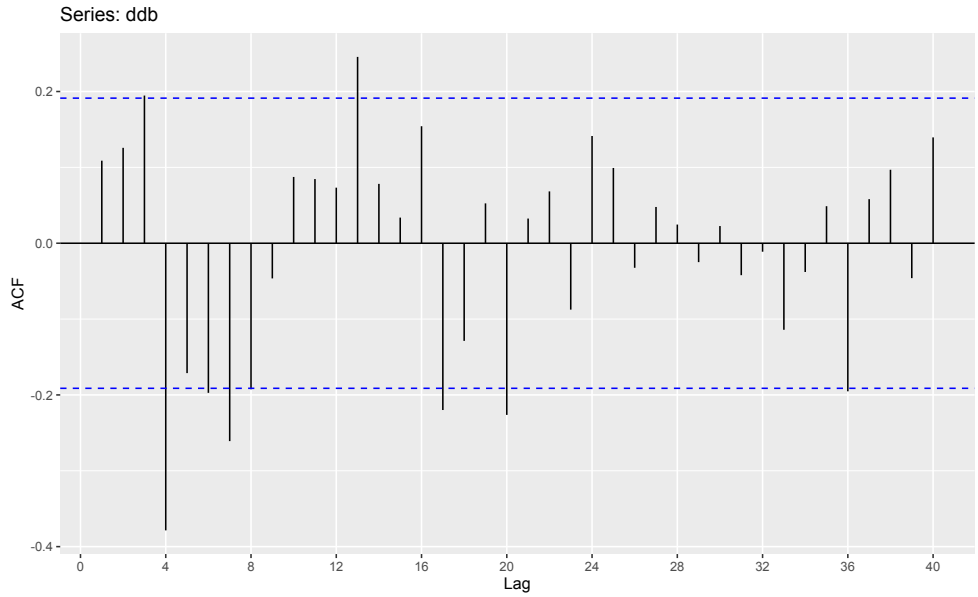
Kuva 5: BKT:n aikasarjan kuvaaja (*b*), sen tavallinen differointi (*db*), sen kausittainen differointi (*dkb*) sekä sen differointi sekä tavallisesti että kausittain (*ddb*)

Tarkastellaan seuraavaksi mallin $ARIMA(p,1,q) \times (P,1,Q)_4$ ACF- ja $PACF$-kuvaajia (kuvat 6 ja 7), jotta mallille voidaan määrittää sopivat asteluvut p, q, P ja Q. Sopivat asteluvut p ja P löydetään tarkastelemalla $PACF$-kuvaajaa ja vastaavasti q ja Q tarkastelemalla ACF-kuvaajaa. Tavallisen mallin asteluvut p ja q löydetään kuvaajista kun $Lag < 4$ ja kausittaisen mallin asteluvut löydetään kohdista $Lag = 4n$, kun $n \in \mathbb{N}$.

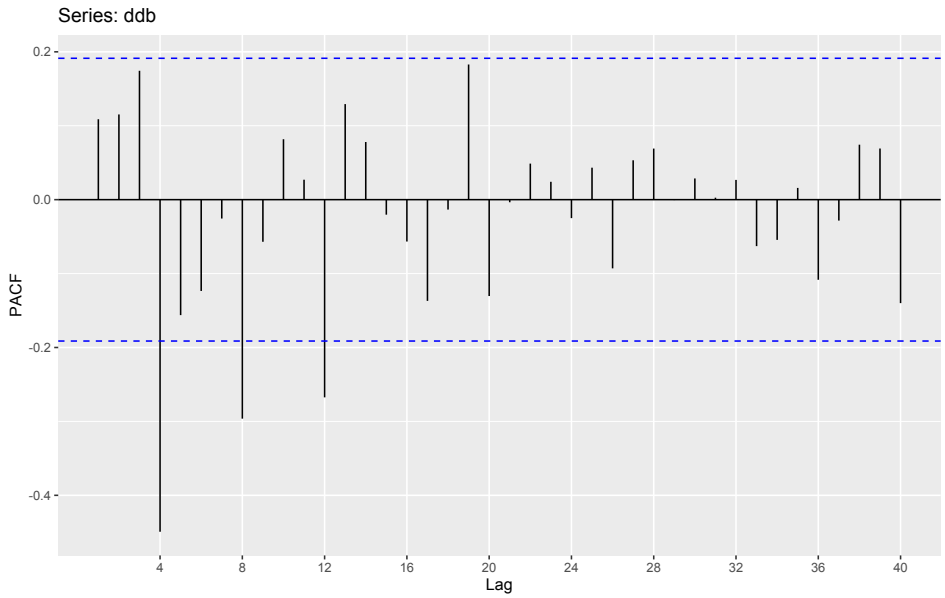
$PACF$-kuvaajasta nähdään parametrin P viimeisen merkittävän piikin olevan kohdassa $Lag = 12$, jolloin $P = 3$, koska $Lag = 12$ tarkoittaa tässä kolmen kauden pituutta. Parametrin p kannalta merkittäviä piikkejä ei näytä olevan, joten valitaan $p = 0$. [9]

ACF-kuvaajassa parametrin q kannalta viimeinen merkittävä piikki vaikuutaa olevan kohdassa $Lag = 3$, joten voidaan valita $q = 3$. Parametrin Q kannalta merkittävät piikit ovat kohdissa $Lag = 36$, $Lag = 20$ ja $Lag = 4$, jolloin $Q = 9$, $Q = 5$ ja $Q = 1$. Mallin ylisovittamisen välttämiseksi valitaan kuitenkin kausittaisen liukuvan keskiarvon parametriksi $Q = 1$ [6, s. 144] [10, ss. 32-33].

Näin ollen ACF- ja $PACF$-kuvaajien perusteella bruttokansantuotteen aikasarjalle sopivista malleista paras olisi $ARIMA(0,1,3) \times (3,1,1)_4$. Verrataan seuraavaksi tätä mallia BKT:n aikasarjan erilaisiin $ARIMA$-malleihin informaatiokriteerien avulla.



Kuva 6: Kerran sekä tavallisesti että kausittain differoidun BKT:n aikasarjan (*ddb*) ACF-kuvaaja



Kuva 7: Kerran sekä tavallisesti että kausittain differoidun BKT:n aikasarjan (*ddb*) $PACF$-kuvaaja

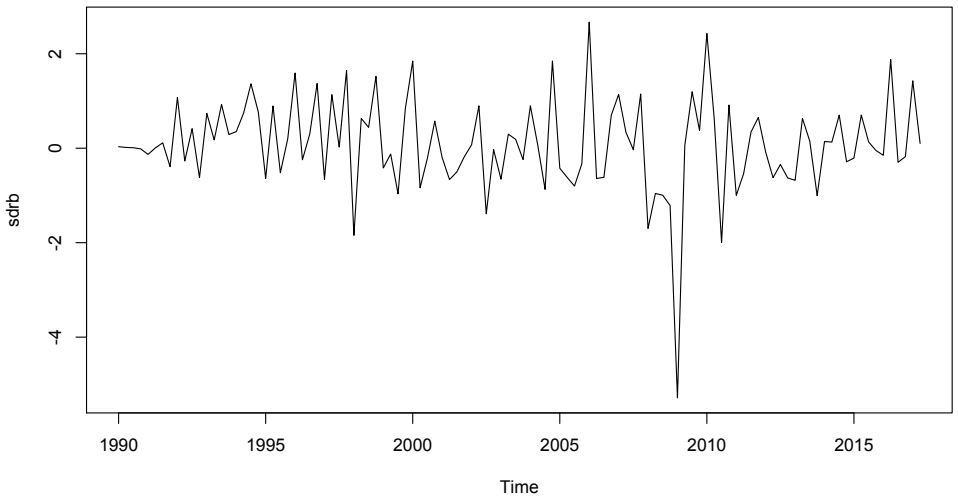
Määritetään aikasarjan $ARIMA$-malleille AIC-, $AICC$- ja BIC-luvut eri parametrien p, q, P ja Q arvoilla, kun $d = D = 1$. Taulukon 2 luvut on laskettu R-ohjelmalla. Taulukkoon on laskettu informaatiokriteerien perusteella BKT:n aikasarjaan parhaiten soveltuvien $ARIMA$-mallien AIC- $AICC$- ja BIC-luvut.

p	q	P	Q	AIC	$AICC$	BIC
0	0	0	1	1632,647	1632,765	1637,955
0	0	0	2	1633,673	1633,910	1641,635
0	0	1	1	1633,997	1634,234	1641,959
0	1	0	1	1632,617	1632,854	1640,579
0	1	0	2	1634,055	1634,455	1644,671
0	2	0	1	1633,121	1633,521	1643,737
0	3	3	1	1632,836	1634,336	1654,068
1	1	0	1	1631,391	1631,791	1642,006
1	2	0	1	1631,390	1631,996	1644,659
1	3	0	1	1630,117	1630,974	1646,041
2	0	0	1	1631,676	1632,076	1642,292
2	1	0	1	1632,345	1632,951	1645,615
2	2	0	1	1627,730	1628,587	1643,654
2	2	0	2	1629,730	1630,884	1648,307
2	2	1	1	1629,730	1630,884	1648,307
3	0	0	1	1630,730	1631,336	1643,999
3	2	0	1	1630,323	1631,478	1648,901

Taulukko 2: $ARIMA(p, 1, q) \times (P, 1, Q)_4$-mallin AIC-, $AICC$- ja BIC-luvut eri parametrien p, q, P ja Q arvoilla

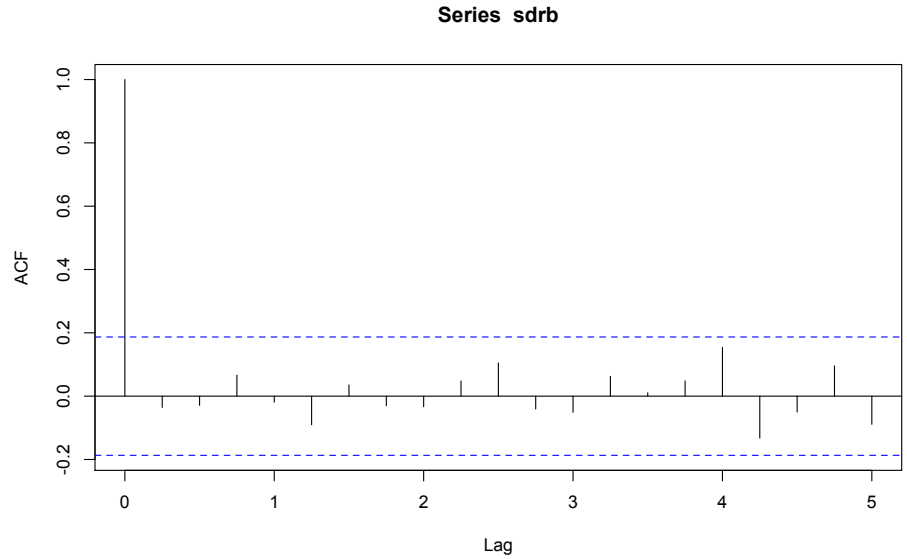
AIC- ja $AICC$-arvojen mukaan paras malli on $ARIMA(2,1,2)\times(0,1,1)_4$. Puolestaan BIC-arvojen mukaan malli on $ARIMA(0,1,0)\times(0,1,1)_4$, mutta priorisoidaan mallin valinnassa kuitenkin AIC- ja $AICC$-arvoja. Näin ollen informaatiokriteerien perusteella BKT:n aikasarjalle sopivimpana voidaan pitää $ARIMA(2,1,2)\times(0,1,1)_4$ -mallia. ACF- ja PACF-kuvaajien perusteella päätelty $ARIMA(0,1,3)\times(3,1,1)_4$-malli ei pärjää vertailtaessa sen informaatiokriteerejä $ARIMA(2,1,2)\times(0,1,1)_4$-mallin vastaaviin, joten valitaan tarkasteltavaksi $ARIMA(2,1,2)\times(0,1,1)_4$-malli.

Tarkastellaan seuraavaksi $ARIMA(2,1,2)\times(0,1,1)_4$-mallin *standardoituja residuaaleja*. Mallin voidaan olettaa sopivan hyvin aineistoon silloin kun standardoitut residuaalit käyttäytyvät valkoisen kohinan tavoin varianssinaan 1 ja odotusarvonaan 0 [7]. Kuvasta 8 nähdään, että standardoitujen residuaalien käyttäytyvän varsin hyvin yhtä poikkeuksellisen pientä arvoa lukuun ottamatta.

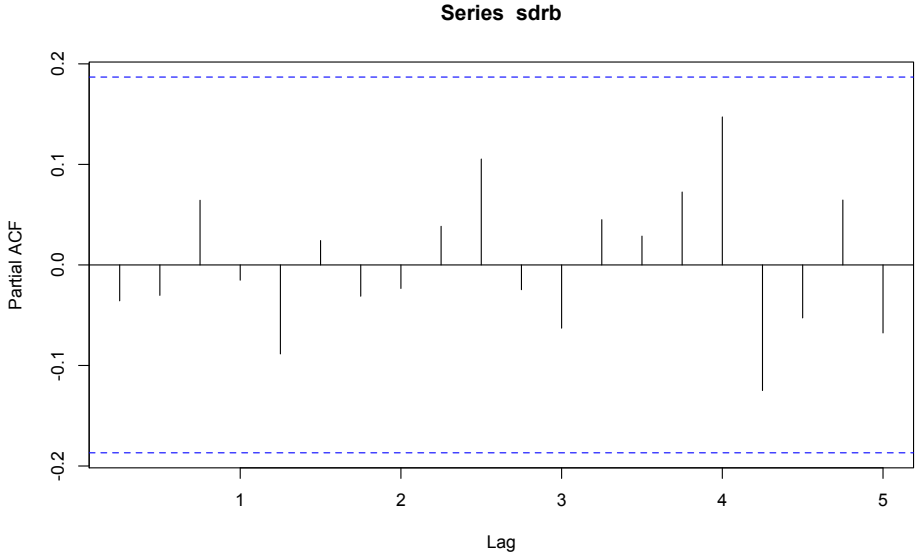


Kuva 8: $ARIMA(2,1,2)\times(0,1,1)_4$-mallin standardoidut residuaalit

Tarkastellaan vielä $ARIMA(2,1,2)\times(0,1,1)_4$-mallin standardoitujen residuaalien ACF- ja $PACF$-kuvaajia. Kuvista 9 ja 10 nähdään, ettei standardoitujen residuaalien välillä ole korrelaatioita. Näin ollen standardoitujen residuaalien perusteella malli vaikuttaisi sopivan aineistoon hyvin.



Kuva 9: $ARIMA(2,1,2)\times(0,1,1)_4$-mallin standardoitujen residuaalien ACF-kuvaaja



Kuva 10: $ARIMA(2,1,2) \times (0,1,1)_4$-mallin standardoitujen residuaalien PACF-kuvaaja

Toimialojen summana esitettävän bruttokansantuotteen aikasarjan $ARIMA(2,1,2)\times(0,1,1)_4$-mallin parametrit ϕ_1, ϕ_2, θ_1, θ_2 ja Θ_1 saadaan tulostettua R-ohjelmasta. Parametrien arvoiksi saadaan

$$\begin{aligned} \phi_1 &= 1,4881, \phi_2 = -0,7946, \\ \theta_1 &= -1,3926, \theta_2 = 0,8234, \text{ ja} \\ \Theta_1 &= -0,7555. \end{aligned} \tag{3.2.1}$$

Näin ollen mallin viiveoperaattorit $\phi(B)$, $\theta(B)$, $\Phi_0(B^4)$ ja $\Theta_1(B^4)$ voidaan kirjoittaa muodoissa

$$\begin{aligned} \phi(B) &= 1 - 1,4881B + 0,7946B^2, \\ \theta(B) &= 1 - 1,3926B + 0,8234B^2, \\ \Phi_0(B^4) &= 1, \text{ ja} \\ \Theta_1(B^4) &= 1 - 0,7555B^4. \end{aligned} \tag{3.2.2}$$

Tässä tapauksessa $ARIMA(2,1,2) \times (0,1,1)_4$-mallin yhtälössä oleva δ määritellään yhtälöllä

$$\delta = (1 - \phi_1 - \phi_2)\mu = (1 - 1,4881 + 0,7916) \cdot 25,2857 = 7,7501, \tag{3.2.3}$$

jossa $\mu = E(\nabla_4^1 \nabla^1 x_t) = 25,2857$ on tulostettu R-ohjelmalla. Näin ollen $ARIMA(2,1,2) \times (0,1,1)_4$-mallin yhtälö saadaan muotoon

$$\begin{aligned} &(1 - 1,4881B + 0,7946B^2)(1 - B^4)(1 - B)x_t \\ =&7,7501 + (1 - 0,7555B^4)(1 - 1,3926B + 0,8234B^2)w_t. \end{aligned} \tag{3.2.4}$$

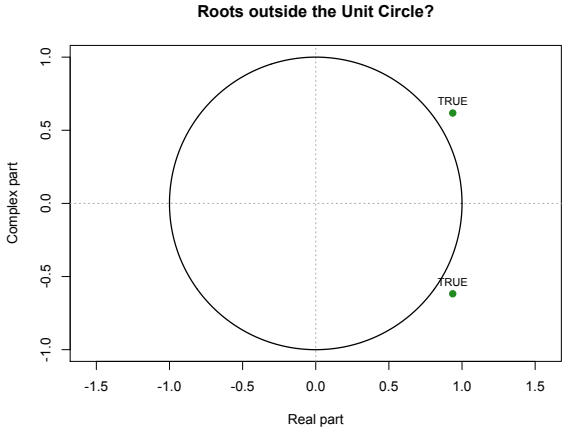
Ennen ennusteen tekemistä tarkastetaan bruttokansantuotteen aikasarjan $ARIMA(2,1,2) \times (0,1,1)_4$-mallin kausaalisuus ja kääntyvyys kyseisen mallin polynomien avulla. Mallin polynomit ovat

$$\begin{aligned} \phi(z) &= 1 - 1,4881z + 0,7946z^2, \\ \theta(z) &= 1 - 1,3926z + 0,8234z^2, \text{ ja} \\ \Theta_1(z^4) &= 1 - 0,7555z^4, \end{aligned} \tag{3.2.5}$$

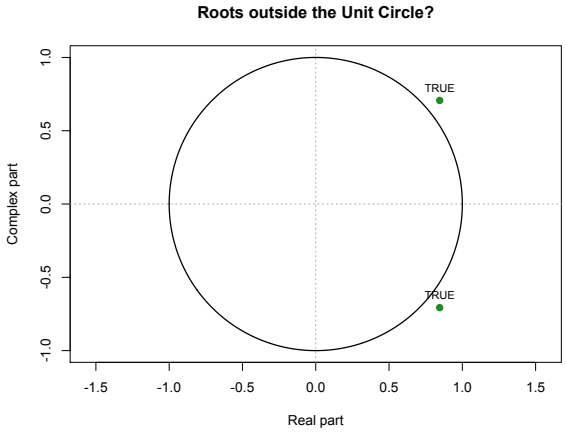
joissa z on kompleksiluku. Jotta malli on kausaalinen, polynomin $\phi(z)$ kompleksijuurten on sijaittava yksikkökiekon ulkopuolella, ja jotta se on kääntyvä, polynomien $\theta(z)$ ja $\Theta_1(z^4)$ kompleksijuurten on sijaittava yksikkökiekon ulkopuolella.

Polynomin $\phi(z)$ kompleksijuuriksi saadaan $z = 0,936383 \pm 0,617804i$. Polynomin $\theta(z)$ kompleksijuuret ovat $z = 0,84564 \pm 0,706661i$, ja polynomin $\Theta_1(z^4)$ kompleksijuuret ovat $z = \pm 1,07261i$ sekä $z = \pm 1,07261$. Tarkastellaan polynomien juurten sijaintia kompleksitasossa kuvien 11, 12 ja 13 avulla. Juuret on merkitty kuviin vihreinä pisteinä.

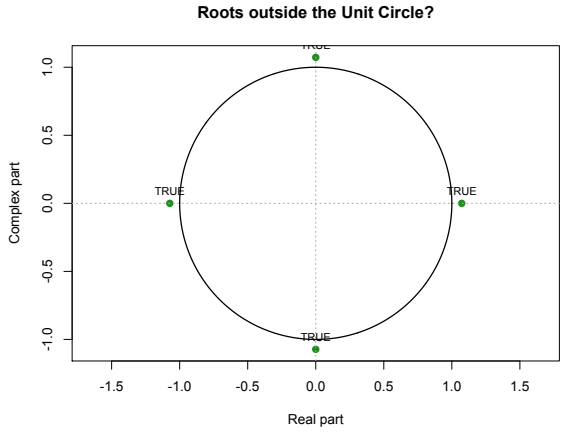
Kuvista 11, 12 ja 13 nähdään, että polynomien $\phi(z)$, $\theta(z)$ ja $\Theta_1(z^4)$ kompleksijuuret ovat yksikkökiekon ulkopuolella, joten $ARIMA(2,1,2) \times (0,1,1)_4$-malli on sekä kausaalinen että kääntyvä. Näin ollen kyseinen malli kelpaa tarkasteltavaksi, ja sen avulla voidaan tehdä ennusteet bruttokansantuotteen arvoille seuraaville kahdeksalle vuosineljännekselle, eli kvartaalille. Tämä tehdään alaluvussa 4.11.



Kuva 11: Polynomin $\phi(z)$ kompleksijuuret suhteessa yksikkökiekkoon



Kuva 12: Polynomin $\theta(z)$ kompleksijuuret suhteessa yksikkökiekkoon



Kuva 13: Polynomin $\Theta_1(z^4)$ kompleksijuuret suhteessa yksikkökiekkoon

4 Toimialojen ja bruttokansantuotteen ARIMA-mallit, ennusteet ja tulokset

Edellisessä luvussa esiteltiin tutkittavaa dataa sekä käsiteltiin $ARIMA$-mallin valintaa käyttäen esimerkkinä toimialoittaisten tuotantojen summana muodostetun bruttokansantuotteen aikasarjaa. Luvussa 2 esitetyn teorian pohjalta tehdään ennusteet Suomen toimialoittaisille tuotannoille sekä BKT:lle kahdeksi vuodeksi. Tämän tekemiseksi poistetaan alkuperäisestä datasta [1] kahden viimeisen vuoden havaintoarvot, joita on yhteensä kahdeksan kappaletta. Sen jälkeen etsitään muokatun datan pohjalta kullekin aikasarjalle sopivin $ARIMA$-malli. Toimialojen $ARIMA$-mallit valitaan samalla periaatteella, jota on käytetty alaluvussa 3.2 hyödyntäen osin myös R-ohjelman auto.arima-komentoa, vaikkei mallien valintaprosesseja kirjoiteta auki tässä luvussa samaan tapaan kuin alaluvussa 3.2.

Valittujen $ARIMA$-mallien avulla luodaan aikasarjoille ennusteet vuoden 2017 kolmannesta neljänneksestä aina vuoden 2019 toiseen neljännekseen asti. Ennusteet luodaan R-ohjelman *forecast*-komennon avulla ja pyöristetään kokonaisluvuiksi toimialakohtaisiin taulukoihin, joissa näkyy sekä tuotantojen ennusteet että toteumat. Tämän jälkeen vertaillaan tälle ajanjaksolle luotuja ennusteita alkuperäisessä datassa esiintyviin toteutuneisiin arvoihin. Alaluvussa 4.11 verrataan toimialoittaisten ennusteiden summia BKT:n oman aikasarjan pohjalta luotuihin ennusteisiin ja tutkitaan kumpiko ennustaa paremmin bruttokansantuotetta.

Kutakin aikasarjaa käsittelevässä alaluvussa ennusteita suhteessa toteumiin on havainnollistettu sekä taulukon että kuvan avulla. Kuvissa aikasarjan todelliset arvot ovat kuvattu mustalla ja ennusteen arvot sinisellä käyrällä, joiden lisäksi niissä tummanharmaa alue kuvaa ennustetta 80 prosentin ja vaaleanharmaa 95 prosentin luottamusvälillä.

Alaluvussa 4.12 tutkitaan ennustevirheiden kehitystä ennustettavan ajanjakson edetessä. Lopuksi alaluvussa 4.13 tehdään yleisiä havaintoja tutkimuksesta sekä vedetään tutkimustuloksia yhteen.

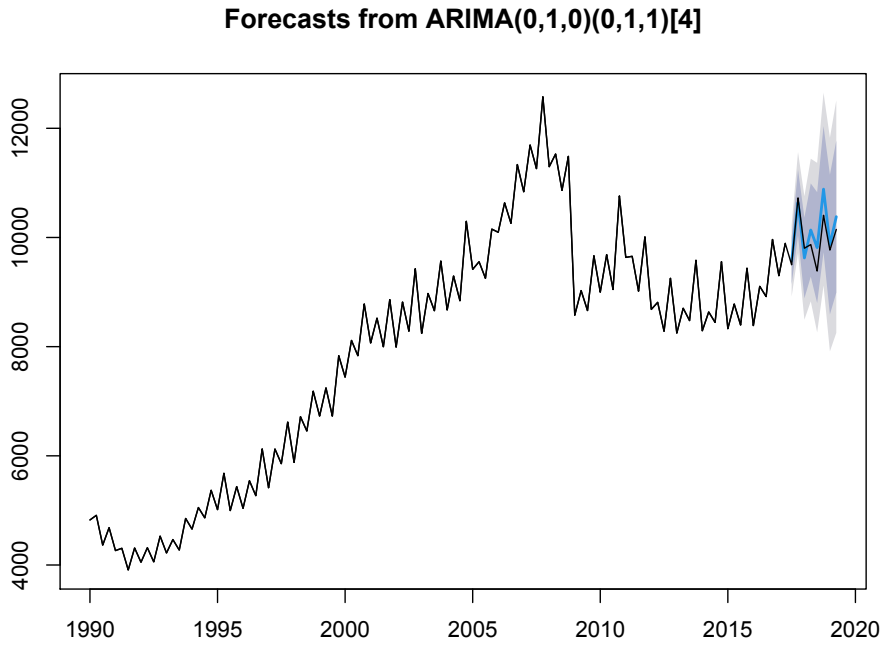
4.1 Teollisuus

Teollisuuden aikasarjan ennusteet luodaan $ARIMA(0,1,0)\times(0,1,1)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 3.

Teollisuuden aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $-1,668$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $2,318$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $7,462$.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	9 569	9 500	$-0,73$
4/2017	10 641	10 725	$0,78$
1/2018	9 625	9 803	$1,82$
2/2018	10 137	9 870	$-2,71$
3/2018	9 814	9 385	$-4,57$
4/2018	10 886	10 406	$-4,61$
1/2019	9 869	9 769	$-1,02$
2/2019	10 381	10 147	$-2,31$

Taulukko 3: Teollisuuden aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 14: Teollisuuden aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Kuvasta 14 nähdään, että teollisuuden aikasarjalla on selkeä nouseva trendi aina 1990-luvun alun lamasta vuoteen 2008 asti, jonka jälkeen laman myötä sen tuotannon arvo on romahtanut nopeasti. Sen jälkeen trendi on sahaillut jonkun verran, mutta tuotanto ei ole noussut lähellekään vuoden 2008 tasoa. Tämä kertoo teollisuuden tuotannon arvon riippuvan paljon taloudellisista suhdanteista.

Teollisuuden aikasarjan ennusteiden kuvaajan muoto vastaa kohtuullisesti sen toteumien kuvaajan muotoa. Ennusteiden kuvaaja mukailee pitkälti

teollisuuden aikasarjan aiempaa käyttäytymistä. Ennusteet ovat tarkimpia vuonna 2017, jonka jälkeen vuoden 2018 ensimmäisessä neljänneksessä ennuste notkahtaa toteumaa enemmän. Vuoden 2018 kolmessa viimeisessä neljänneksessä toteumat jäävät selvästi ennusteiden arvojen alle. Tämä johtuu pitkälti siitä, ettei toteuman arvo nouse juurikaan vuoden 2018 ensimmäisestä toiseen neljännekseen toisin kuin enemmän aiemman kausittaisvaihtelun mukaisesti käyttäytyvä ennuste. Myös vuonna 2019 toteumat ovat ennusteita pienempiä joskaan eivät yhtä paljoa kuin vuoden 2018 loppupuolella.

Tämän seurauksena prosentuaalinen keskivirhe on selvästi negatiivinen. Absoluuttinen prosentuaalinen keskivirhe on vain jonkin verran prosentuaalisen keskivirheen itseisarvoa suurempi, mikä kertoo siitä, että ennusteet ovat suurimmaksi osaksi toteumia suurempia. Suurehko skaalatun keskineliövirheen arvo selittyy pitkälti vuoden 2018 kolmannen ja neljännen vuosineljänneksen suurilla ennustevirheillä.

4.2 Koulutus, terveys- ja sosiaalipalvelut

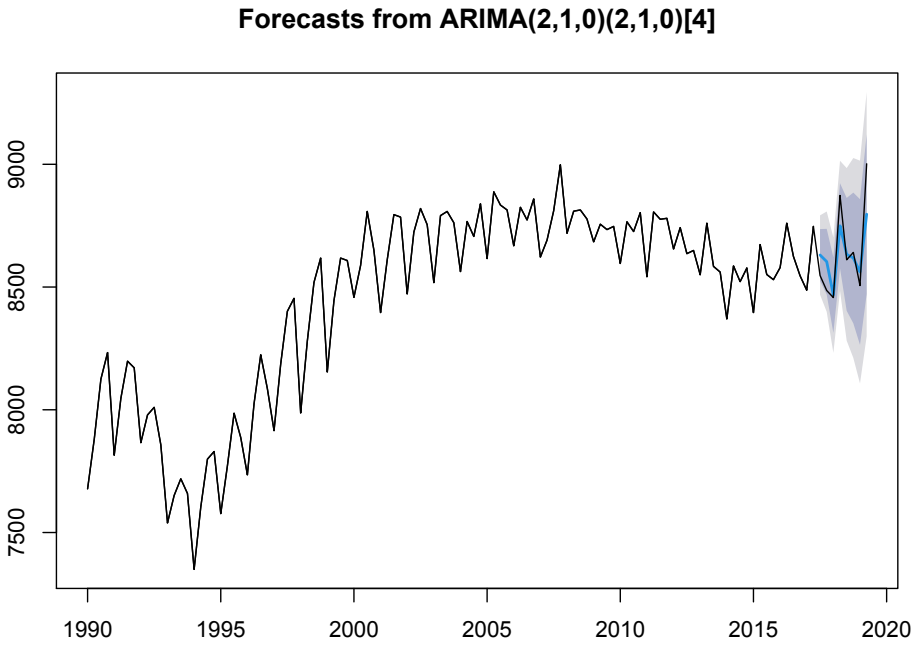
Koulutuksen sekä terveys- ja sosiaalipalveluiden aikasarjan ennusteet luodaan $ARIMA(2,1,0) \times (2,1,0)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 4.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	8 630	8 574	$-0,65$
4/2017	8 604	8 486	$-1,39$
1/2018	8 467	8 457	$-0,12$
2/2018	8 749	8 873	$1,40$
3/2018	8 633	8 611	$-0,26$
4/2018	8 618	8 641	$0,27$
1/2019	8 561	8 506	$-0,65$
2/2019	8 797	9 002	$2,28$

Taulukko 4: Koulutuksen sekä terveys- ja sosiaalipalveluden aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain

Koulutuksen sekä terveys- ja sosiaalipalveluiden aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $0,110$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $0,876$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $1,137$.

Kuvan 15 nojalla koulutuksen sekä terveys- ja sosiaalipalveluiden aikasarjan ennusteiden käyrä mukailee varsin hyvin sen toteumien käyrää. Vuoden 2017 viimeisessä neljänneksessä ennusteet laskevat toteumia hitaammin sekä vuosien 2018 ja 2019 toisissa neljänneksissä ennusteiden piikit eivät ole



Kuva 15: Koulutuksen sekä terveys- ja sosiaalipalveluiden aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

yhtä suuret kuin toteumien vastaavat. Tämä näkyy myös kyseisten vuosineljännesten muita suurempina prosentuaalisina virheinä taulukossa 4. Muissa kvartaaleissa ennusteet ovat varsin tarkkoja. Vaikka ennusteet ovat viidessä kvartaalissa toteumia suurempia, prosentuaalinen keskivirhe on niukasti positiivinen, koska ennustevirheet ovat suurempia toteumien ollessa ennusteita suurempi. Prosentuaalinen keskivirhe on noin lähellä nollaa siksi, että positiiviset ja negatiiviset ennustevirheet kumoavat toisiaan.

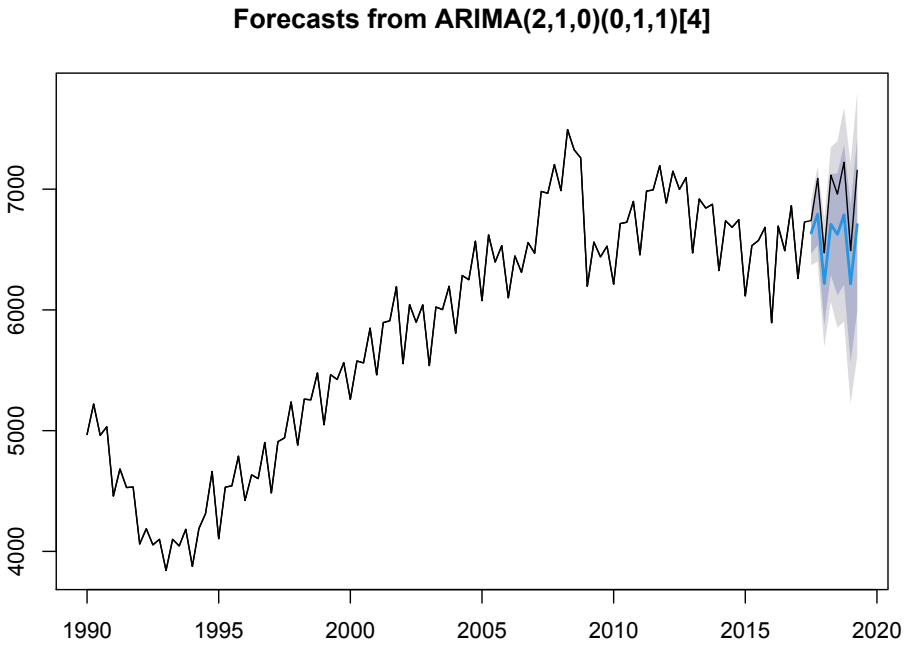
Muutamista hieman poikkeavista arvioista huolimatta malli ennustaa aikasarjan käyttäytymistä varsin hyvin, koska sekä absoluuttisen prosentuaalisen keskivirheen että skaalatun keskineliövirheen arvot ovat pieniä. Koulutuksen sekä terveys- ja sosiaalipalveluiden aikasarjan tietty vakaus ja ennustettavuus saattaa selittyä sillä, että kyseessä on pitkälti julkisen sektorin vastuulla olevasta toimialasta, jolloin suhdanteiden vaihtelut eivät osu siihen samalla tavalla kuin moniin muihin toimialoihin.

4.3 Kauppa, liikenne, majoitus ja ravitseminen

Kaupan, liikenteen, majoituksen ja ravitsemisen aikasarjan ennusteet luodaan $ARIMA(2,1,0) \times (0,1,1)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 5.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	6 636	6 740	1,54
4/2017	6 794	7 089	4,16
1/2018	6 219	6 473	3,92
2/2018	6 709	7 116	5,72
3/2018	6 625	6 959	4,80
4/2018	6 788	7 222	6,01
1/2019	6 216	6 491	4,24
2/2019	6 707	7 156	6,27

Taulukko 5: Kaupan, liikenteen, majoituksen ja ravitsemisen aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 16: Kaupan, liikenteen, majoituksen ja ravitsemisen aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Kaupan, liikenteen, majoituksen ja ravitsemisen aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $4,853$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on sama $4,853$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $16,384$.

Kuvan 16 nojalla kaupan, liikenteen, majoituksen ja ravitsemisen aikasarjan käyttäytyminen muistuttaa paljon teollisuuden aikasarjan käyttäytymistä. Se nousee tasaisesti aina 1990-luvun alusta vuoteen 2008, jonka jälkeen se tippuu nopeasti ja sahailee sen jälkeen vuoden 2008 tasoa alempana. Kaupan, liikenteen, majoituksen ja ravitsemisen ennusteiden ja toteumien käyrät ovat keskenään hyvinkin saman muotoisia, joten malli ennustaa aikasarjan kausittaisvahtelun hyvin.

Sen sijaan samasta kuvasta nähdään, ettei malli tunnista aikasarjan 2010-luvun lopun hieman nousevaa trendiä vaan kaikki ennusteiden arvot ovat selvästi toteumien arvoja pienempiä. Ennusteiden ja toteumien välinen ero muodostuu pitkälti vuonna 2017, jolloin kolmannessa neljänneksessä toteuma ei laske, toisin kuin malli ennustaa, ja viimeisessä neljänneksessä toteuma nousee selvästi ennustetta enemmän. Tästä eteenpäin toteumat ovat melko tasaisesti ennusteita suurempia.

Näin ollen ennusteiden prosentuaalinen keskivirhe ja absoluuttinen prosentuaalinen keskivirhe ovat yhtä suuret ja niiden perusteella nämä tämän aikasarjan ennusteet ovat tämän tutkielman toiseksi huonoimmat. Skaalattu keskineliövirhe on suurempi kun millään muulla tutkielmassa käsiteltävällä aikasarjalla. Niinpä kyseinen malli ei ennusta aikasarjan toteumia kovinkaan hyvin. Tämä lienee selitettävissä ainakin osin aikasarjan trendin vaihtelevuudella 2010-luvun aikana.

4.4 Kiinteistöala

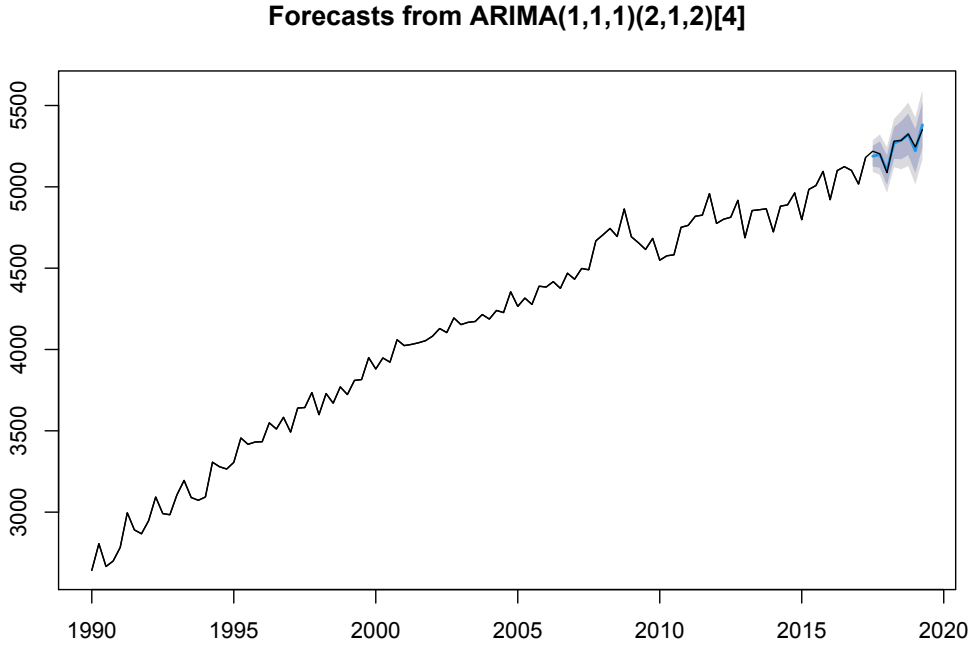
Kiinteistöalan aikasarjan ennusteet luodaan $ARIMA(1,1,1) \times (2,1,2)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 6.

Kiinteistöalan aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $0,070$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $0,284$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $0,066$.

Kuvan 17 perusteella kiinteistöalan aikasarjalla on selkeä nouseva trendi ja selkeähkö kausittaisvaihtelu, jotka malli ennustaa erittäin hyvin sekä kuvaajien että ennustevirheiden tunnuslukujen perusteella. Ennusteiden prosentuaaliset virheet ovat muutamaa lukuunottamatta erittäin lähellä nollaa eivätkä ne muutamatkaan poikkea kovin paljoa nollasta. Suurimmatkin prosentuaaliset virheet ovat vain puolen prosentin luokkaa, mikä kertoo ennusteen erinomaisesta tarkkuudesta. Ennustevirheiden etumerkit vaihtelevat,

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	5 188	5 219	0,59
4/2017	5 198	5 203	0,10
1/2018	5 102	5 087	−0,29
2/2018	5 269	5 281	0,23
3/2018	5 286	5 285	−0,02
4/2018	5 323	5 325	0,04
1/2019	5 221	5 245	0,46
2/2019	5 381	5 352	−0,54

Taulukko 6: Kiinteistöalan aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 17: Kiinteistöalan aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

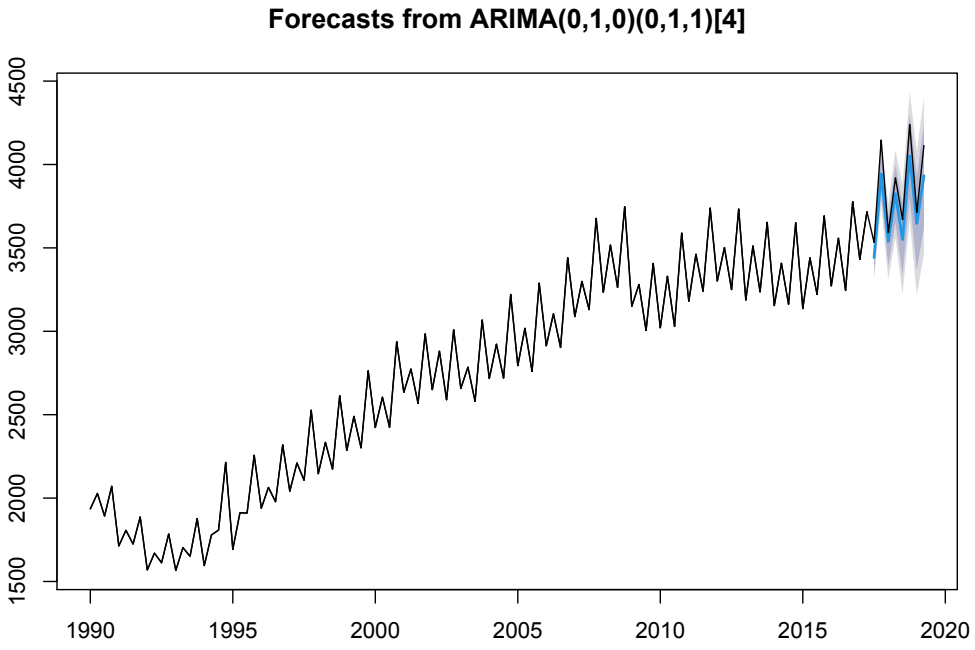
minkä seurauksena kiinteistöalan prosentuaalinen keskivirhe on vielä lähempänä nollaa kuin sen absoluuttinen prosentuaalinen keskivirhe, joka sekin on pieni. Kiinteistöalan aikasarjan ennusteet ovat kaikilla tunnusluvuilla mitattuna paremmat kuin minkään muun tässä tutkielmassa tutkittavan aikasarjan vastaavat.

4.5 Hallinto- ja tukipalvelut

Hallinto- ja tukipalveluiden aikasarjan ennusteet luodaan $ARIMA(0,1,0) \times (0,1,1)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 7.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	3 442	3 533	2,58
4/2017	3 943	4 147	4,92
1/2018	3 539	3 592	1,48
2/2018	3 825	3 920	2,42
3/2018	3 550	3 669	3,24
4/2018	4 051	4 239	4,44
1/2019	3 647	3 713	1,78
2/2019	3 932	4 114	4,42

Taulukko 7: Hallinto- ja tukipalveluiden aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 18: Hallinto- ja tukipalveluiden aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Hallinto- ja tukipalveluiden aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $3,159$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on sama $3,159$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $4,809$.

Kuvan 18 nojalla hallinto- ja tukipalveluiden aikasarjan trendi on pääosin nouseva eivätkä lamat ole vaikuttaneet siihen yhtä paljoa kuin useisiin muihin toimialoihin. Sen kausittaisvaihtelu on selkeää. Hallinto- ja tukipalveluiden aikasarjan ennusteiden kuvaajan muoto mukailee sen toteumien kuvaajan muotoa, mutta kaupan, liikenteen, majoituksen ja ravitsemisen aikasarjan tavoin toteumilla on jyrkempi trendi kuin ennusteilla.

Tällöin kaikki ennusteiden arvot ovat toteumien arvoja pienempiä. Siksi absoluuttinen prosentuaalinen keskivirhe ja prosentuaalinen keskivirhe ovat yhtä suuret. Suurimmat erot ennusteiden ja toteumien välillä ovat silloin kuin toteumien arvot ovat suurimmillaan, eli vuosien 2017 ja 2018 viimeisissä neljänneksissä sekä vuoden 2019 toisessa neljänneksessä. Hallinto- ja tukipalveluiden aikasarjan ennuste tarkasteltavan ajanjakson ensimmäiselle kvartaalille on selvästi toteumaa pienempi. Tämä selittyy kenties sillä, että toteuman arvo laskee siinä kohdassa vähemmän kuin vastaavissa kvartaaleissa aiemmin. Siitä eteenpäin sekä ennusteet että toteumat käyttäytyvät jotakuinkin aikasarjan aiemman trendin mukaisesti.

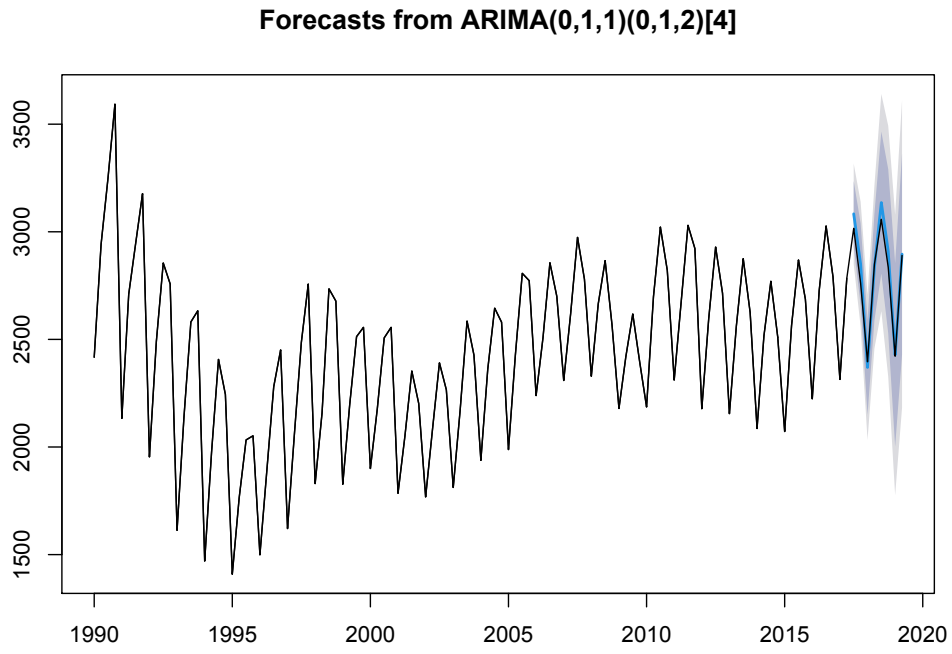
Kaikki ennustevirheiden tunnusluvut ovat suurehkoja, joten malli ei ennusta aikasarjaa kovin hyvin kausittaisvaihtelun hyvästä tunnistamisesta huolimatta. Tämä selittynee aikasarjan trendin kasvun nopeutumisella 2010-luvun loppupuolella.

4.6 Rakentaminen

Rakentamisen aikasarjan ennusteet luodaan $ARIMA(0,1,1)\times(0,1,2)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 8.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	3 083	3 016	$-2,22$
4/2017	2 850	2 761	$-3,22$
1/2018	2 370	2 396	$1,09$
2/2018	2 836	2 849	$0,46$
3/2018	3 136	3 058	$-2,55$
4/2018	2 909	2 835	$-2,61$
1/2019	2 428	2 422	$-0,25$
2/2019	2 896	2 890	$-0,21$

Taulukko 8: Rakentamisen aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 19: Rakentamisen aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Rakentamisen aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $-1,190$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $1,575$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $1,120$.

Kuvan 19 nojalla rakentamisen aikasarjalla on selkeä kausittaisvaihtelu, mutta aikasarjan trendi aaltoilee jonkun verran. Rakentamisen tuotannon arvo ei pudonnut kovin pahasti vuoden 2009 lamassa toisin kuin 1990-luvun alun lamassa. Usein lama-aikoina julkinen sektori pyrkii elvyttämään taloutta matalakorkoisilla lainoilla rahoitetuilla rakennusinvestoinneilla, mikä kompensoi yksityisen kysynnän vähenemistä rakennusalalla. Tämä saattaa selittää sitä, että rakentamisen tuotannon arvo notkahti vain vähän vuoden 2009 laman aikana.

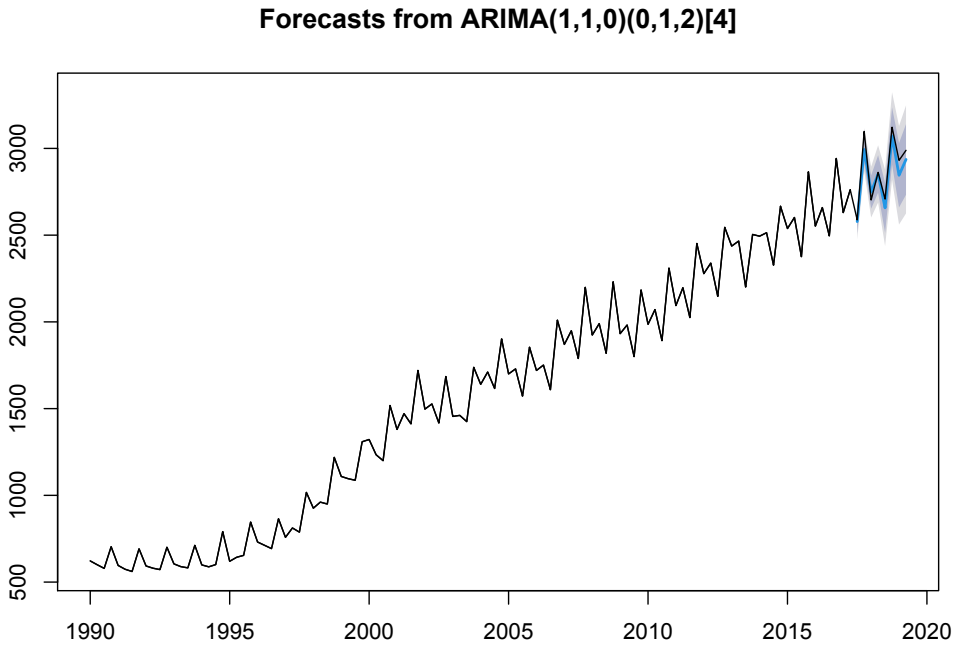
Malli huomioi aikasarjan kausittaisvaihtelun, mutta vuosien kolmansissa ja neljänsissä kvartaaleissa se yliarvioi kausittaisvaihtelun suuruutta. Tällöin ennusteiden arvot ovat toteumia suurempia ja ennusteiden prosentuaalinen keskivirhe on negatiivinen. Tämä voi johtua siitä, että kuvan perusteella kausittaisvaihtelun suuruus näyttää hivenen pienenevän 2010-luvun loppupuolella. Toisaalta vuosien ensimmäisten ja toisten neljännesten ennusteet ovat paljon tarkempia, mistä osaltaan kertoo se, että prosentuaalisen keskivirheen itseisarvo ei ole kovin paljoa absoluuttista prosentuaalista keskivirhettä pienempi. Kuitenkin kaikkien ennustevirheiden tunnuslukujen perusteella malli näyttää ennustavan aikasarjan käyttäytymistä varsin hyvin.

4.7 Informaatio ja viestintä

Informaation ja viestinnän aikasarjan ennusteet luodaan $ARIMA(1,1,0) \times (0,1,2)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 9.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	2 578	2 588	0,39
4/2017	2 994	3 098	3,36
1/2018	2 748	2 703	−1,66
2/2018	2 853	2 862	0,31
3/2018	2 657	2 708	1,88
4/2018	3 072	3 122	1,60
1/2019	2 845	2 931	2,93
2/2019	2 936	2 989	1,77

Taulukko 9: Informaation ja viestinnän aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 20: Informaation ja viestinnän aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Informaation ja viestinnän aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $1,323$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $1,739$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $1,232$.

Kuvan 20 perusteella informaation ja viestinnän aikasarjalla on selkeä nouseva trendi ja selkeähkö kausittaisvaihtelu. Aikasarjan ennusteiden kuvaaja käyttäytyy aika lailla aikasarjan aiemman trendin ja kausittaisvaihtelun mukaisesti. Toteumien käyrän muoto on melko samanlainen kuin ennusteiden käyrän vastaava.

Tarkasteltavan ajanjakson ensimmäisen kvartaalin ennuste on hyvinkin tarkka. Kuitenkin vuoden 2017 viimeisessä kvartaalissa toteumien käyrä tekee ennusteita selvästi korkeamman piikin. Ennuste on ainoan kerran toteumaa suurempi vuoden 2018 ensimmäisessä neljänneksessä, jonka jälkeen toteumien arvojen trendi nousee selvästi ennusteiden arvojen vastaavaa nopeammin. Siksi ennusteet ovat systemaattisesti merkittävästi toteumia pienempiä vuoden 2018 puolivälistä alkaen.

Niinpä prosentuaalinen keskivirhe on selvästi positiivinen eikä edes paljoa absoluuttista prosentuaalista keskivirhettä pienempi. Ennustevirheiden tunnusluvut ovat kuitenkin pienehköjä, joten mallin voidaan sanoa ennustavan aikasarjan käyttäytymistä kohtuullisen hyvin.

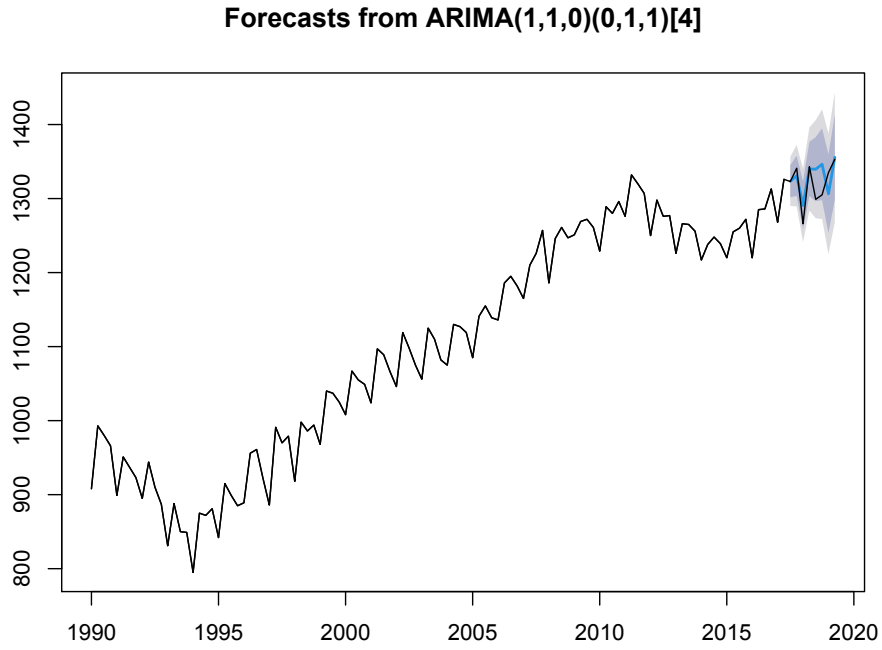
4.8 Taiteet, viihde ja virkistys

Taiteiden, viihteen ja virkistyksen aikasarjan ennusteet luodaan $ARIMA(1,1,0) \times (0,1,1)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 10.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	1 323	1 323	0,00
4/2017	1 331	1 341	0,75
1/2018	1 291	1 266	−1,97
2/2018	1 340	1 343	0,22
3/2018	1 340	1 299	−3,16
4/2018	1 346	1 305	−3,14
1/2019	1 306	1 335	2,17
2/2019	1 356	1 353	−0,22

Taulukko 10: Taiteiden, viihteen ja virkistyksen aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain

Taiteiden, viihteen ja virkistyksen aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $-0,669$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $1,454$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $0,468$.



Kuva 21: Taiteiden, viihteen ja virkistyksen aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Kuvan 21 nojalla taiteiden, viihteen ja virkistyksen aikasarjan ennusteiden käyrä mukailee varsin hyvin aikasarjan aiempaa käyttäytymistä. Toteumat poikkeavat kuitenkin osin niiden aiemmasta käyttäytymisestä. Vuoden 2018 kahdessa viimeisessä neljänneksessä toteumat ovat merkittävästi ennusteita pienemmät, koska vastoin aikasarjan aiempaa kausittaisvahtelua vuonna 2018 toteuman arvo laskee merkittävästi toisesta kolmanteen neljännekseen. Myöskin vastoin aiempaa kausittaisvaihtelua toteuman arvo nousee vuoden 2018 viimeisestä neljänneksestä vuoden 2019 ensimmäiseen neljännekseen, jolloin toteuma on jo ennustetta suurempi. Vuoden 2018 ensimmäisessä neljänneksessä toteumien käyrä tekee selvästi ennusteiden käyrää suuremman notkahduksen. Tämän sekä vuoden 2018 lopun ennustevirheiden takia prosentuaalinen keskivirhe on negatiivinen. Koska ajoittain kuitenkin toteuma on ennustetta suurempi, prosentuaalisessa virheessä positiiviset ja negatiiviset ennustevirheet kumoavat toisiaan ja siten absoluuttinen prosentuaalinen keskivirhe on selvästi prosentuaalisen keskiarvon itseisarvoa suurempi.

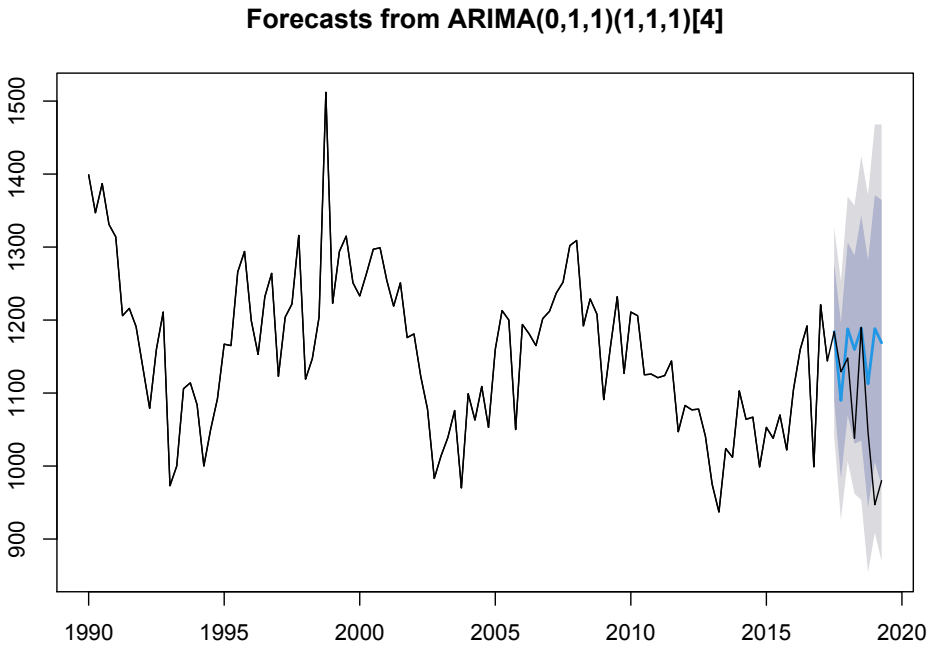
Kaikki taiteiden, viihteen ja virkistyksen aikasarjan ennustevirheiden tunnusluvut ovat pienehköjä, joista skaalattu keskineliövirhe on erityisen pieni. Tämä selittynee sillä, ettei mikään yksittäinen ennustevirhe ole niin iso, että sen neliö kasvaittaisi tunnuslukua suureksi. Kokonaisuutena mallin voidaan sanoa ennustavan aikasarjaa kohtuullisen hyvin.

4.9 Rahoitus ja vakuutukset

Rahoituksen ja vakuutusten aikasarjan ennusteet luodaan $ARIMA(0,1,1) \times (1,1,1)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 11.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	1 184	1 184	0,00
4/2017	1 090	1 129	3,45
1/2018	1 188	1 148	−5,23
2/2018	1 160	1 038	−2,12
3/2018	1 189	1 190	0,08
4/2018	1 113	1 046	−6,41
1/2019	1 188	947	−25,45
2/2019	1 169	980	−19,29

Taulukko 11: Rahoituksen ja vakuutusten aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 22: Rahoituksen ja vakuutusten aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Rahoituksen ja vakuutusten aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $-6,868$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $7,753$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $11,995$.

Kuvan 22 perusteella rahoituksen ja vakuutusten aikasarjan kuvaaja käyttäytyy hyvin ennalta arvaamattomasti. Siitä on vaikeaa löytää selkeää trendiä, jonka lisäksi satunnaisheilahtelu on suurehkoa ja siten kausittaisvaihtelua on vaikeahkoa tunnistaa. Tämä asettaa heikohkot lähtökohdat aikasarjan mallintamiselle, mikä näkyy myös ennusteiden laaduissa. Rahoituksen ja vakuutusten aikasarjan epävakaa käyttäytyminen voi johtua rahoitusmarkkinoiden ailahtelevuudesta sekä niiden taipumuksesta reagoida herkästi erilaisiin ulkoisiin ärsykkeisiin.

Malli ennustaa vuosien 2017 ja 2018 kolmansien vuosineljännesten toteumat likimain täydellisesti, mutta muissa ennusteissa virheet ovat joko suurehkoja tai suuria. Vuoden 2017 viimeisessä neljänneksessä ennuste notkahtaa toteumaa enemmän, mutta sen jälkeen vuoden 2018 alussa ennusteet ovat jo toteumia korkeammalla tasolla. Erityisesti vuoden 2018 toisessa neljänneksessä toteuma notkahtaa merkittävästi ennustetta enemmän, mutta sitä seuraavassa neljänneksessä toteuma nousee vielä ennusteen tasolle.

Vuoden 2018 viimeisessä neljänneksessä toteuman arvo laskee selvästi ennusteen arvoa enemmän ja seuraavan vuoden ensimmäisessä neljänneksessä toteuma jatkaa romahdustaan ennusteen kääntyessä kasvuun. Vastaavasti vuoden 2019 toisessa neljänneksessä ennusteen arvo laskee hieman ja toteuman vastaava nousee hieman, mutta ennusteiden käyrä jää siitä huolimatta reilusti toteumien käyrää korkeammalle tasolle.

Toteumien arvojen raju lasku suhteessa ennusteisiin erityisesti tarkasteluajanjakson loppupuolella selittää sen, että absoluuttinen prosentuaalinen keskivirhe on hyvin suuri ja että prosentuaalinen keskivirhe on melkein yhtä paljon negatiivinen. Myös skaalattu keskineliövirhe on suuri. Näin ollen tunnuslukujen valossa rahoituksen ja vakuutusten aikasarjan ennusteet ovat epätarkemmat kuin minkään muun tässä tutkielmassa tutkittavan aikasarjan vastaavat.

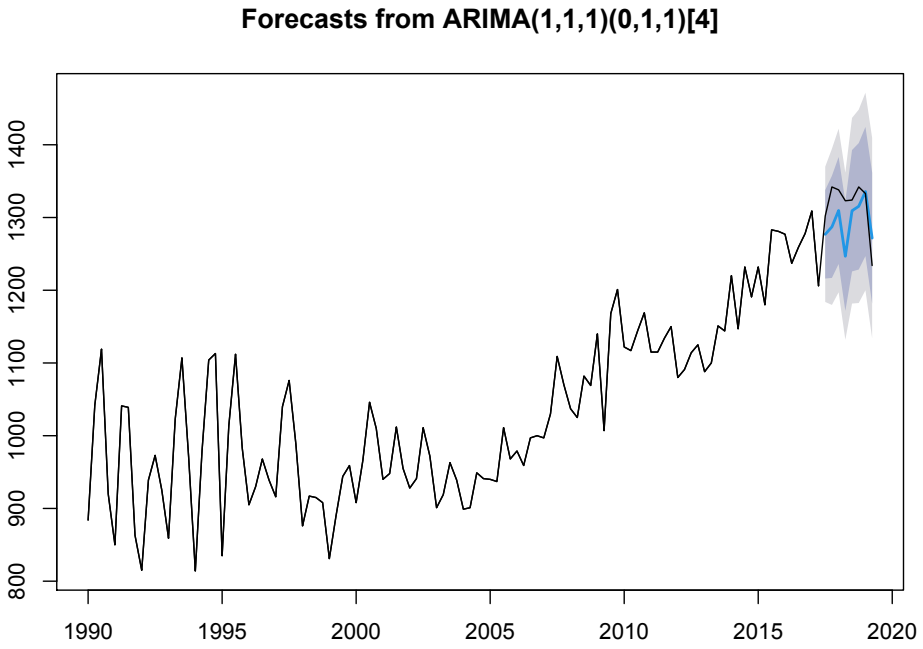
4.10 Maa-, metsä- ja kalatalous

Maa- metsä- ja kalatalouden aikasarjan ennusteet luodaan $ARIMA(1,1,1) \times (0,1,1)_4$-mallilla. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 12.

Maa-, metsä- ja kalatalouden aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on $1,712$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $2,538$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $1,197$.

vuosineljännes	ennuste	toteuma	%-virhe
3/2017	1 277	1 302	1, 92
4/2017	1 287	1 342	4, 10
1/2018	1 310	1 338	2, 09
2/2018	1 247	1 323	5, 74
3/2018	1 309	1 324	1, 13
4/2018	1 315	1 342	2, 01
1/2019	1 336	1 333	−0, 23
2/2019	1 272	1 234	−3, 08

Taulukko 12: Maa-, metsä- ja kalatalouden aikasarjan ennusteet, toteumat ja prosentuaaliset ennustevirheet neljännesvuosittain



Kuva 23: Maa-, metsä-, ja kalatalouden aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

Kuvan 23 nojalla maa-, metsä- ja kalatalouden aikasarjan ennusteet mukailevat pitkälti sen aiksarjan aiempaa trendiä ja kausittaisvaihteluita. Ennusteiden ja toteumien käyrien muodot muistuttavat jonkin verran toisiaan siinä mielessä, että ne nousevat tarkasteltavan ajanjakson alussa, notkahtavat puolivälin maissa, nousevat sen jälkeen hiukan ja laskevat merkittävästi lopussa.

Kuitenkin tarkasteltavan ajanjakson alkupuolella toteumien arvot nousevat selvästi ennusteiden arvoja korkeammalle ja pysyvät siellä vuoden 2018

loppuun asti. Merkittävimmät erot löytyvät vuoden 2017 viimeisestä neljänneksestä sekä vuoden 2018 toisesta neljänneksestä, jossa ennuste notkahtaa selvästi toteumaa enemmän. Sen jälkeen ennusteiden arvot alkavat kiriä toteumia kiinni siten, että vuoden 2019 ensimmäisessä neljänneksessä ennuste on jo hivenen toteumaa suurempi. Tarkasteluajanjakson viimeisessä vuosineljänneksessä toteuman arvo laskee reilusti ennustetta enemmän, jolloin sen prosentuaalinen virhe on selvästi negatiivinen.

Maa-, metsä- ja kalatalouden aikasarjan ennusteiden prosentuaalinen keskivirhe on selvästi positiivinen, koska aivan loppua lukuun ottamatta toteumat ovat ennusteita suurempia. Viimeisen kvartaalin ansiosta kuitenkin absoluuttinen prosentuaalinen keskivirhe on jonkun verran prosentuaalista keskivirhettä suurempi. Skaalatun keskineliövirheen arvo on kohtuullisen pieni. Kokonaisuutena malli ennustaa aikasarjan käyttäytymistä kohtalaisesti.

4.11 Bruttokansantuote

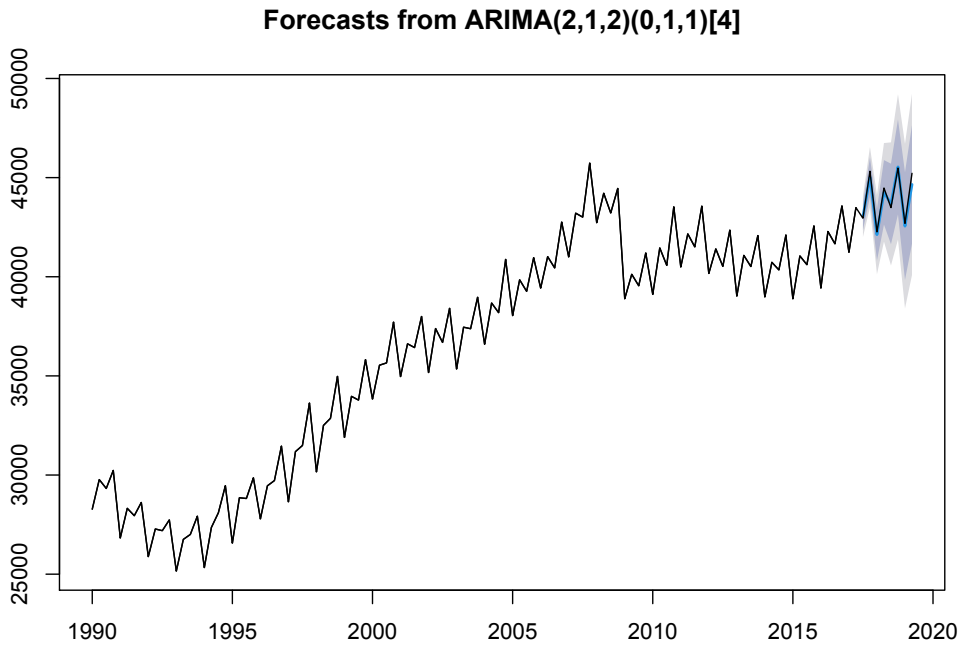
Alaluvun 3.2 tarkastelun nojalla toimialojen summana muodostetun BKT:n aikasarjalle sopivin $ARIMA$-malli on $ARIMA(2,1,2) \times (0,1,1)_4$. Kyseisen mallin pohjalta tehtyjen ennusteiden arvot eri vuosineljänneksille löytyvät taulukosta 13. Muodostetaan samaan taulukkoon toimialojen ennusteiden summa, jota verrataan bruttokansantuotteen oman aikasarjan ennusteisiin.

vuosineljännes	toim. summa	%-virhe	ennuste	%-virhe	BKT:n arvo
3/2017	42 913	0,09	43 029	−0,18	42 952
4/2017	44 731	1,30	45 023	0,66	45 321
1/2018	41 858	0,96	42 125	0,33	42 263
2/2018	44 124	0,79	44 258	0,49	44 475
3/2018	43 540	−0,12	43 674	−0,43	43 488
4/2018	45 421	0,14	45 540	−0,13	45 483
1/2019	42 618	0,17	42 553	0,33	42 692
2/2019	44 828	0,86	44 655	1,24	45 217

Taulukko 13: BKT:n aikasarjan ennusteet, toimialojen ennusteiden summat sekä molempien prosentuaaliset ennustevirheet ja BKT:n toteumat neljännesvuosittain

Bruttokansantuotteen oman aikasarjan ennusteiden prosentuaalinen keskivirhe (MPE) on 0,289, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on 0,472 ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on 1,517.

Kuva 24 osoittaa bruttokansantuotteen omalla aikasarjalla olevan sekä selkeä kausittaisvaihtelu että melko selkeä trendi. Kuvan 24 nojalla malli en-



Kuva 24: Bruttokansantuotteen aikasarjan ennusteet (sininen käyrä) ja toteumat (musta käyrä)

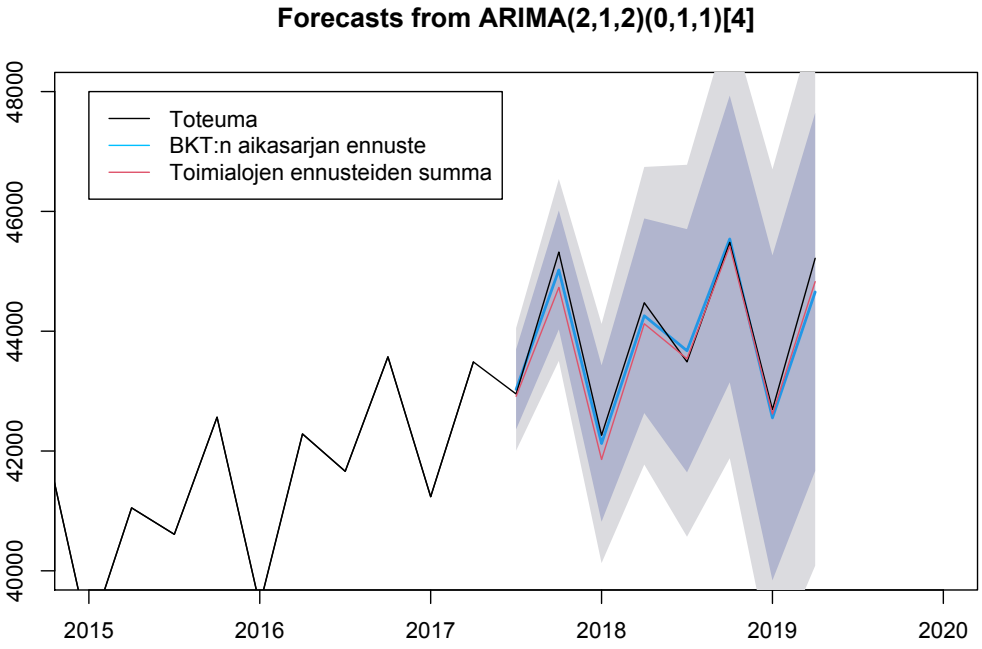
nustaa BKT:n aikasarjaa hyvin, ja taulukosta 13 nähdään ennusteiden prosentuaalisten virheiden olevan varsin pieniä.

Tarkasteluajanjakson alussa toteumat ovat vuoden 2017 kolmatta neljännestä lukuun ottamatta hiukan ennusteita suurempia, mutta vuoden 2018 kahdessa viimeisessä neljänneksessä ennusteet ovat niukasti toteumia suurempia. Vuoden 2019 puolella taas toteumien käyrä nousee ennusteiden käyrän yläpuolelle.

Myös prosentuaalisen keskivirheen positiivinen arvo kertoo siitä, että bruttokansantuotteen toteumat ovat pääasiassa sen oman aikasarjan perusteella tehtyjä ennusteita suuremmat. BKT:n oman aikasarjan ennustevirheiden tunnusluvut ovat pieniä, joten mallin voidaan sanoa ennustavan aikasarjaa hyvin.

Toimialoittaisten ennusteiden summien prosentuaalinen keskivirhe (MPE) suhteessa bruttokansantuotteen toteumiin on $0,524$, absoluuttinen prosentuaalinen keskivirhe ($MAPE$) on $0,554$ ja skaalattu keskineliövirhe ($\frac{MSE}{\overline{a_t}}$) on $2,274$.

Kuvasta 25 ja taulukosta 13 nähdään kuinka toimialoittaisten ennusteiden summat käyttäytyvät verrattuna bruttokansantuotteen aikasarjan ennusteisiin ja toteumiin. Tarkasteluajanjakson ensimmäisen puolikkaan ajan toimialoittaisten ennusteiden summat ovat hieman BKT:n aikasarjan ennusteita



Kuva 25: Tarkennettu versio kuvasta 24, johon on lisätty BKT:n aikasarjan ennusteiden (sininen käyrä) ja toteumien (musta käyrä) lisäksi toimialoittaisten tuotantojen ennusteiden summat (punainen käyrä).

sekä toteumia pienempiä. Näin ollen vuoden 2017 viimeisestä neljänneksestä vuoden 2018 toiseen neljännekseen asti BKT:n oman aikasarjan ennusteet ovat toimialoittaisten ennusteiden summia tarkempia. Siitä eteenpäin toimialoittaisten ennusteiden summat ovat lähempänä bruttokansantuotteen toteumia kuin sen oman aikasarjan ennusteet, joskin ennustevirheiden suuruudet ja niiden väliset erot ovat silloin pääasiassa pienempiä kuin tarkasteltavan ajanjakson alkupuolella.

Koska toimialoittaisten ennusteiden summa on vain yhdessä vuosineljänneksessä bruttokansantuotteen toteumaa suurempi, ja silloinkin vain niukasti, niin prosentuaalinen keskivirhe on lähes samansuuruinen kuin absoluuttinen prosentuaalinen keskivirhe. Kuitenkin myös toimialoittaisten ennusteiden summien ennustevirheiden kaikki tunnusluvut ovat varsin pieniä, mutta kuitenkin suurempia kuin BKT:n oman aikasarjan ennustevirheiden vastaavat.

Toimialojen ennusteiden summana lasketussa ennusteessa bruttokansantuote on pilkottu pienempiin osiin, jolloin toimialojen erilainen käyttäytyminen on tullut huomioiduksi, mutta toisaalta toimialojen ennusteiden summaan vaikuttaa yhteensä kymmenen eri toimialan ennusteiden ennustevirheet. Tämä vaikuttanee siihen, että tässä tutkimuksessa bruttokansantuot-

teen oman aikasarjan perusteella saatiin ennustettua sen käyttäytymistä hieman tarkemmin.

Näin ollen tässä tutkimuksessa käytettyjen mallien ja aineiston perusteella bruttokansantuotetta saadaan ennustettua tarkemmin sen oman aikasarjan perusteella kuin toimialoittaisten ennusteiden summana. Täytyy kuitenkin muistaa, että tämän tutkimuksen perusteella ero näiden tapojen välillä on varsin pieni, ja että niistä molemmat tuottavat varsin tarkkoja ennusteita BKT:lle.

On myös syytä huomioida, että tässä on kyse vain Suomen bruttokansantuotteesta ja että monissa muissa maissa toimialoittaisten tuotantojen aikasarjat saattavat olla hyvin eri tavoin käyttäytyviä ja eri suuruisia suhteessa bruttokansantuotteeseen. Niinpä pelkästään tämän tutkimuksen perusteella ei voida yleistää, että BKT:a kannattaa ennustaa mieluummin sen oman aikasarjan perusteella kuin toimialoittaisten tuotantojen ennusteiden summana. Jotta tällainen yleistys voitaisiin tehdä, tulisi tarkastella vastaavasti monien muidenkin valtioiden bruttokansantuotteen aikasarjoja sekä sellaisenaan että toimialoittain pilkottuna.

4.12 Ennustevirheet neljännesvuosittain

Tarkastellaan seuraavaksi ennustevirheitä neljännesvuosittain tavoitteena selvittää miten ennustetarkkuudelle käy ennustettavan ajanjakson edetessä. Taulukkoon 14 on laskettu aikasarjojen ennustevirheiden keskiarvot neljännesvuosittain vastaavasti kuin aiemmin tässä luvussa laskettiin toimialoittain. Niihin on huomioitu toimialojen aikasarjojen ennustevirheiden lisäksi bruttokansantuotteen oman aikasarjan ennustevirheet, muttei toimialoittaisten tuotantojen summien vastaavia, jotta toimialoittaiset tuotannot eivät tule huomioiduksi kahteen kertaan. Siis esimerkiksi taulukossa 14 esiintyvä kvartaalin 3/2017 prosentuaalinen keskivirhe on aikasarjojen prosentuaalisten virheiden keskiarvo kyseisessä kvartaalissa.

Taulukossa 14 prosentuaalinen keskivirhe sahailee aluksi siten, että parittomissa kvartaaleissa sen arvot ovat lähellä nollaa ja parillisissa ne ovat ykköstä suurempia. Kuitenkaan sama trendi ei jatku enää vuoden 2018 neljännessä kvartaalissa, vaan prosentuaalisen keskivirheen arvo menee ensimmäistä kertaa nollan alapuolelle, jossa se myös pysyy tarkasteluajanjakson loppuun asti. Prosentuaalinen keskivirhe saa pienimmän arvonsa vuoden 2019 ensimmäisessä neljänneksessä, mutta on selvästi negatiivinen myös toisessa neljänneksessä. Kuitenkin prosentuaalisen keskivirheen arvot ovat jokaisessa kvartaalissa varsin pieniä. Tämä selittyy osin sillä, että positiiviset ja negatiiviset ennustevirheet kumoavat toisiaan. Kahden viimeisen vuosineljänneksen selvästi negatiiviset arvot selittynevät pitkälti rahoituksen ja vakuutusten

vuosineljännes	MPE	$MAPE$	$\frac{MSE}{\overline{a_t}}$
3/2017	$0,294$	$0,982$	$0,455$
4/2017	$1,605$	$2,444$	$2,889$
1/2018	$0,131$	$1,818$	$1,492$
2/2018	$1,106$	$1,984$	$3,542$
3/2018	$0,015$	$2,011$	$4,087$
4/2018	$-0,230$	$2,842$	$5,197$
1/2019	$-1,426$	$3,591$	$2,100$
2/2019	$-0,877$	$3,785$	$7,610$

Taulukko 14: Neljännesvuosittaisten ennustevirheiden prosentuaaliset keskivirheet (MPE), absoluuttiset prosentuaaliset keskivirheet ($MAPE$) sekä skaalatut keskineliövirheet ($\frac{MSE}{\overline{a_t}}$)

huomattavan suurilla prosentuaalisilla virheillä kyseisissä kvartaaleissa.

Ennustetarkkuuden arvioinnissa prosentuaalista keskivirhettä parempia mittareita ovat absoluuttinen prosentuaalinen keskivirhe sekä skaalattu keskineliövirhe, koska niissä positiiviset ja negatiiviset ennustevirheet eivät kumoa toisiaan.

Neljännesvuosittaisten ennusteiden absoluuttiset prosentuaaliset keskivirheet kasvavat vuoden 2018 ensimmäistä neljännestä lukuun ottamatta koko tarkasteluajanjakson ajan. Vuoden 2017 kolmannessa neljänneksessä absoluuttinen prosentuaalinen keskivirhe on vielä varsin pieni, mutta sitä seuraavan vuoden ajan sen arvot ovat kakkosen luokkaa kunnes vuoden 2018 viimeisessä neljänneksessä absoluuttisen prosentuaalisen keskivirheen arvo nousee lähemmäs kolmosta ja vuoden 2019 puolella yli kolmen ja puolen, joka kuvastaa jo ihan merkittäviä ennustevirheitä. Rahoituksen ja vakuutusten huomattavan suuret prosentuaaliset virheet saattavat osaltaan selittää myös vuoden 2019 suurehkoja absoluuttisia prosentuaalisia keskivirheitä.

Myös neljännesvuosittaisten ennusteiden skaalatut keskineliövirheet kasvavat ennustettavan ajanjakson edetessä. Ainoat poikkeukset tähän trendiin tekevät vuosien 2018 ja 2019 ensimmäiset neljännekset, joissa skaalattu keskineliövirhe saa selvästi pienemmät arvot kuin niitä edeltävissä ja seuraavissa kvartaaleissa. Tarkasteluajanjakson ensimmäisen kvartaalin skaalattu keskineliövirhe on vielä hyvinkin pieni, mutta viimeisen kvartaalin vastaava arvo on jo suurehko ja paria mainittua poikkeusta lukuun ottamatta niiden kasvu on tasaista.

Absoluuttisten prosentuaalisten keskivirheiden ja skaalattujen keskineliövirheiden arvojen kasvu ennustettavan ajanjakson edetessä kertoo siitä, että tutkimuksessa käytettyjen mallien ennustetarkkuus heikkenee tarkasteltavan ajanjakson edetessä.

4.13 Yhteenveto

Tämän tutkielman perusteella kokonaisuutena kausittaisten $ARIMA$-mallien voidaan sanoa sopivan hyvin toimialoittaisten tuotantojen sekä bruttokansantuotteen aikasarjojen ennustamiseen. Mallit tunnistivat aikasarjoista selkeät trendit sekä kausittaisvaihtelut ja niiden ennusteet olivat parhaita niillä toimialoilla, joiden toteumat käyttäytyivät niiden aiempien trendien ja kausittaisvaihtelun mukaisesti. Ennusteet olivat epätarkimpia sellaisilla toimialoilla, joilla ei ollut selkeää trendiä tai kausittaisvahtelua ja joilla satunnaisheilahtelu oli suurta.

Ennustetarkkuuden ja aikasarjoissa esiintyvän hajonnan välistä yhteyttä voidaan mitata esimerkiksi toimialojen tutkittavien vuosineljännesten toteumien skaalattujen keskihajontojen ja ennustevirheiden välisillä korrelaatiokertoimilla. Toimialojen tutkittavien vuosineljännesten toteumien skaalattujen keskihajontojen ja absoluuttisten prosentuaalisten keskivirheiden välinen korrelaatiokerroin on $0,567$, mikä kertoo keskivirheiden ja -hajontojen korreloivan merkittävästi. Toteumien skaalattujen keskihajontojen ja skaalattujen keskineliövirheiden välinen vastaava korrelaatiokerroin on $0,329$, mikä tukee jossain määrin edellisen korrelaatiokertoimen perusteella esitettyä tulosta. Toteumien skaalattujen keskihajontojen ja prosentuaalisten keskivirheiden välinen korrelaatiokerroin on $-0,319$ mikä selittyy sillä, että toimialoilla, joilla skaalatut keskihajonnat ovat suurimmat, prosentuaaliset keskivirheet ovat negatiivisia.

Tämän pienen tarkastelun perusteella ennustaminen on helpompaa aikasarjoille, joissa esiintyy vähän hajontaa. On kuitenkin huomattava, että keskihajontaan vaikuttaa sekä trendi, kausittaisvaihtelu että satunnaisheilahtelu. Tällainen tarkastelu voisi olla mielekkäämpi, jos tarkasteltaisiin esimerkiksi toimialojen valkoisten kohinoiden skaalattujen keskihajontojen ja ennustevirheiden välisiä korrelaatiokertoimia.

Taloudessa voi kuitenkin tapahtua paljon sellaisia asioita, joita hyvätkään matemaattiset ja tilastotieteelliset mallit eivät kykene ennustamaan. Esimerkiksi voidaan olettaa, ettei tutkielmassa käytetyt mallit olisi kyenneet ennustamaan toimialoittaisten tuotantojen eikä bruttokansantuotteen arvoa kovinkaan hyvin vuoden 2020 kolmelle viimeiselle neljännekselle kiitos koronapandemian aiheuttaman taloudellisen romahduksen.

Lähteet

[1] Suomen virallinen tilasto (SVT): *Neljännesvuositilinpito* [verkkojulkaisu]. ISSN=1797-9749. Helsinki: Tilastokeskus [viitattu: 21.10.2019]. Saantitapa: http://www.stat.fi/til/ntp/index.html

[2] Bisgaard, S. ja Kulahci, M. *Time Series Analysis and Forecasting by Example*. John Wiley& Sons, Inc. Hoboken, New Jersey, 2011.

[3] Fuller, W.A. *Introduction to Statistical Time Series, Second Edition*. John Wiley& Sons, Inc. New York, 1996.

[4] Brockwell, P.J. ja Davis, R.A. *Introduction to Time Series and Forecasting, Third Edition*. Springer Text in Statistics, 2016.

[5] Laininen, P. *Todennäköisyys ja sen tilastollinen soveltaminen*. Tekijä ja Oy Yliopistokustannus, Helsinki, 1998.

[6] Shumway, R.H. ja Stoffer, D.S. *Time Series Analysis and Its Applications With R Examples, Fourth Edition*. Springer Text in Statistics, 2016.

[7] Lizana Bister, L. *ARIMA- ja GARCH-mallit sekä mallin sovittaminen osakeaineistoon*. Tampereen yliopisto, Informaatiotieteiden laitos, Tampere, 2011.

[8] *Kansantalouden tilinpidon verkkosivut*. Helsinki: Tilastokeskus [viitattu: 11.2.2020]. Saantitapa: https://www.stat.fi/meta/kas/index.html?K

[9] *STAT 510: Applied Time Series Analysis*. The Pennsylvania State University, 2020 [viitattu: 25.2.2020]. Saantitapa: https://online.stat.psu.edu/stat510/lesson/4/4.1

[10] Tammi, A. *Autoregressiiviset mallit sähkön kulutuksen ennustamisessa*. Turun yliopisto, Tulevaisuuden teknologioiden laitos, Turku, 2019.

[11] Sellaiah, S. *Statistical Forecast Errors*. Coventry University Technology Park, Olivehorse Consulting Services Ltd, Coventry 2017 [viitattu: 24.11.2020]. Saantitapa: https://blog.olivehorse.com/statistical-forecast-errors