Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan tutkinto-ohjelma

Työvoiman tarpeen ennustaminen SARIMA-aikasarjamallilla

Kandidaatintyö
27.5.2015

Touko Väänänen

Työn saa tallentaa ja julkistaa Aalto-yliopiston avoimilla verkkosivuilla.
Muilta osin kaikki oikeudet pidätetään.

Aalto-yliopisto

AALTO-YLIOPISTO PERUSTIETEIDEN KORKEAKOULU PL 11000, 00076 Aalto http://www.aalto.fi	KANDIDAATINTYÖN TIIVISTELMÄ
Tekijä: Touko Väänänen	
Työn nimi: Työvoiman tarpeen ennustaminen SARIMA-aikasarjamallilla	
Tutkinto-ohjelma: Teknillisen fysiikan ja matematiikan tutkinto-ohjelma	
Pääaine: Systeemitieteet	Pääaineen koodi: F3010
Vastuuopettaja(t): Prof. Harri Ehtamo Ohjaaja(t): Ville Juvonen	

Tiivistelmä:
Tässä kandidaatintyössä luodaan vuositason ennustemalli erään suomalaisen palvelualan yrityksen työvoiman tarpeelle. Ennustemallin luomisessa käytetään dataa työvoiman tarpeesta vuosilta 2008-2014. Datassa on huonoja datapisteitä joita käsitellään puuttuvina. Datassa on myös yksi poikkeava havainto, jonka käsittelemiseksi käytetään additiivisen poikkeavan havainnon (additive outlier) mallia. Työssä pyritään löytämään ennustava ulkoinen muuttuja työvoiman tarpeelle mallin ennustetarkkuuden parantamiseksi. Tätä varten kokeillaan useita taloudellisia aikasarjoja jotka voisivat korreloida työvoiman tarpeen kanssa, mutta hyvää ennustavaa muuttujaa ei löydetä.

Aikasarjamallin tekemisessä käytetään Boxin ja Jenkinsin menetelmää aikasarjamallien tekemiseksi. Työssä sijoitetaan dataan erilaisia SARIMA-malleja ja testataan mallien hyvyyttä diagnostisilla testeillä. Työvoiman tarvetta päädytään ennustamaan SARMA$(1,0,1)\times(1,0,1)_{12}$ mallilla. Mallin antamien ennusteiden tarkkuutta mitattiin suhteellisella keskivirheellä (MAPE) ja tämän perusteella ehdotettiin myös käytettävän ennusteen pituutta.

Tulosten tarkastelussa pohditaan mallin hyvyyttä ja jatkokehitysmahdollisuuksia.

Päivämäärä:	21.11.2014	Kieli:	Suomi	Sivumäärä:	21+4
Avainsanat:	ARMA, työvoiman tarve, ennustaminen, aikasarjamallit				

Sisältö

Merkinnät ja lyhenteet

(S)ARIMA = (Seasonal) Autoregressive Integrated Moving Average

AR = Autoregressive

MA = Moving Average

MAPE = Mean Absolute Percentage Error (Suhteellinen keskivirhe)

Z_t, aikasarja

B, viiveoperaattori s.e. $Bz_t = z_{t-1}$

Δ, differenssioperaattori s.e. $\Delta z_t = z_t - z_{t-1}$

1 Johdanto

Työvoiman suunnittelun yhtenä tavoitteena on saada työvoiman tarve ja tarjonta kohtaamaan. Ernst et al. [2004] mukaan suunnittelu on tavallista jakaa (i) työvoiman tarpeen mallintamiseen, (ii) vapaapäivien suunnitteluun, (iii) vuorojen suunnitteluun, (iv) työlinjaston suunnitteluun, (v) työtehtävien jakamiseen ja (vi) työntekijöiden asettamiseen. Tässä työssä keksitytään kohtaan (i) eli työvoiman tarpeen mallintamiseen, jonka tehtävänä on selvittää kuinka monta työntekijää tarvitaan tietyn ajanjakson aikana työtehtävien suorittamiseen.
Työssä pyritään löytämään ennustemalli palvelualan yrityksen työvoimantarpeelle. Alalla työvoiman tarve vaihtelee runsaasti vuoden ympäri. Jos päivittäistä työvoiman tarvetta ei kyetä täyttämään, joutuu yritys käyttämään kalliimpaa vuokratyövoimaa ja hyvä vuositason ennustemalli helpottaisi valmistautumista kiireellisiin aikoihin esimerkiksi työvoimaa rekrytoimalla.
Ennustemalli tehdään käyttämällä Boxin ja Jenkinsin menetelmää [Box et al., 2008] SARIMA-aikasarjamallien tekemiseen ja malliin pyritään löytämään ulkoinen selittäjä aikasarjamallin ennustetarkkuuden parantamiseksi. Työssä kokeillaan useita ulkoisia selittäjiä jotka voisivat korreloida aikasarjan kanssa, mutta merkitsevää ulkoista selittäjää jota on tilastoitu tarpeeksi pitkältä ajalta ei löydetä.
Työvoiman tarve-aikasarjassa on neljä huonoa datapistettä joita käsitellään tässä työssä puuttuvina. Lisäksi aikasarjassa on yksi poikkeava havainto, joka otetaan mallissa huomioon dummy-muuttujalla.
Työssä päädytään suosittelemaan SARMA$(1,0,1)$x$(1,0,1)_{12}$ mallia työvoiman tarpeen ennustamiseen.

2 Tutkimusongelma ja -menetelmät

2.1 Tutkimusongelma

Tutkimusongelma työssä on löytää vuositason ennustemalli suomalaisen palvelualan yrityksen työvoiman tarpeelle. Työvoiman tarve alalla vaihtelee kuukausien välillä ja tämä aiheuttaa haasteita työvoimasuunnittelulle. Aikasarjamalli tehdään Boxin ja Jenkinsin menetelmällä SARIMA-aikasarjamallien tekemiseen. Työssä pyritään myös löytämään ulkoinen selittävä muuttuja työvoiman tarpeen ennustamiseen aikasarjamallin ennustetarkkuuden parantamiseksi.
Malli validoidaan mallin oletuksia testaamalla ja mallin antaman ennusteen tarkkuutta arvioidaan suhteellisen keskivirheen (MAPE) perusteella.

2.2 Aikasarjat

Aikasarja on sarja havaintoja, jotka on havainnoitu yleensä tasavälisin ajanhetkin. Aikasarjaa Z jossa on t havaintoa merkitään $Z = \{z_1, z_2, ..., z_t\}$.
Aikasarjan sanotaan olevan stationaarinen, jos sen tilastolliset ominaisuudet eivät muutu vaihtamalla tarkasteltavaa ajanjaksoa. Stationaarisella aikasarjalla on siis pysyvä odotusarvo

$$\mu = E[z_t] = \int_{-\infty}^{\infty} zp(z)dz \tag{1}$$

sekä varianssi

$$\sigma_z^2 = E[(z_t - \mu)^2] = \int_{-\infty}^{\infty} (z - \mu)^2 p(z)dz \tag{2}$$

2.3 ARMA-prosessit

Selostus ARMA- ja SARIMA-prosesseista seuraa esitystä kirjassa Box et al. [2008]. Prosessien kuvaamista varten esitellään siirto-operaattori B, sekä aidosti satunnainen prosessi ϵ, jota kutsutaan myös valkoiseksi kohinaksi.

Siirto-operaattori B operoi aikasarjan yhteen havaintoon z_t siten, että $Bz_t = z_{t-1}$.

Valkoinen kohina $\epsilon_t, t \in \{T\}$ on satunnaisprosessi jolla on seuraavat ominaisuudet

$$\begin{array}{ll} E[\epsilon_t] = 0, & \forall t \in \{T\} \\ Var[\epsilon_t] = \sigma^2, & \forall t \in \{T\} \\ Cov[\epsilon_t, \epsilon_s] = 0, & t \neq s \end{array} \tag{3}$$

Autoregressiivinen prosessi astetta p eli AR(p)-prosessi voidaan kirjoittaa

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} + a_t \tag{4}$$

jossa z_t on mallinnettava aikasarja, ϕ_p merkitsee painoa kullekin aikasarjan realisaatiolle ajanhetkellä z_{t-p} ja a_t on valkoista kohinaa.

(4) voidaan myös kirjoittaa siirto-operaattorin B avulla

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) z_t = a_t \tag{5}$$

jota merkitään

$$\phi(B) z_t = a_t \tag{6}$$

ja jossa siirto-operaattorin funktiota ϕ kutsutaan viivepolynomiksi.

Liikkuvan keskiarvon prosessi astetta q eli MA(q)-prosessi voidaan kirjoittaa

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \tag{7}$$

Kuten AR(p)-prosessin tapauksessa, myös 7 voidaan kirjoittaa viivepolynomin avulla

$$z_t = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_p B^q) a_t \tag{8}$$

jota merkitään

$$z_t = \theta(B) a_t \tag{9}$$

Yhdistämällä AR(p)- ja MA(q)-prosessit saadaan autoregressiivinen liikkuvan keskiarvon prosessi astetta (p,q), eli ARMA(p,q)-prosessi joka voidaan kirjoittaa

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q} \quad (10)$$

tai

$$\phi(B) z_t = \theta(B) a_t \quad (11)$$

2.4 SARIMA-prosessit

Monet aikasarjat eivät ole luonnostaan stationaarisia, mutta usein aikasarjat pystytään stationarisoimaan differoimalla aikasarja tarpeeksi monta kertaa. Tätä varten esitellään differenssioperaattori $\Delta = (1 - B)$ joka operoi aikasarjaan z_t siten, että $\Delta z_t = (1 - B) z_t = z_t - z_{t-1}$. Jos aikasarja z_t stationarisoituu yhdellä differenssillä, niin tälle voidaan kirjoittaa ARIMA-malli

$$\phi(B) \Delta z_t = \theta(B) a_t \quad (12)$$

vastaavasti yleinen ARIMA-malli voidaan kirjoittaa

$$\phi(B) \Delta^d z_t = \theta(B) a_t \quad (13)$$

jossa $\Delta^d = (1 - B)^d$ ja d kuvaa aikasarjan z_t stationarisoimiseksi tarvittavien differenssien määrää.
Prosessia voidaan myös kuvata seuraavilla kahdella yhtälöllä

$$\phi(B) w_t = \theta(B) a_t \quad (14)$$

$$w_t = \Delta^d z_t \quad (15)$$

Kausittaista trendiä aikasarjassa pystytään myös kuvaamaan esiteltyjen malliluokkien avulla. Kausittaisen trendin kuvaamiseen käytettävä differenssioperaattori on $\Delta_s = (1 - B^s)$ joka operoi aikasarjaan siten, että $\Delta_s z_t =$

$(1 - B^s)z_t = z_t - z_{t-s}$. Tämän avulla aikasarjaa, jossa esimerkiksi tammikuun havainnot muistuttavat edellisen tammikuun havaintoja voidaan kuvata mallilla

$$\Phi(B^{12})\Delta_{12}^D z_t = \Theta(B^{12})\alpha_t \tag{16}$$

jossa $\Phi(B^{12})$ ja $\Theta(B^{12})$ ovat B^{12}:ta polynomeja astetta P ja Q. Tämänlaisessa mallissa kuitenkin virhetermit α_t eivät olisi korreloimattomia, joten esittelemme mallin

$$\phi(B)\Delta^d\alpha_t = \theta(B)a_t \tag{17}$$

jossa a_t on valkoista kohinaa.
Yhdistämällä mallit 16 ja 17 saamme yleisen multiplikatiivisen mallin

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta_s^D z_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{18}$$

jota kutsutaan SARIMA-prosessiksi astetta (p,d,q)x(P,D,Q)$_s$.

2.5 Korrelaatiofunktiot

Selostus korrelaatiofunktioista seuraa esitystä kirjassa Box et al. [2008]. Aikasarjoja analysoitaessa ollaan usein kiinnostuneita siitä, kuinka aikasarjan eri ajanhetkien realisaatiot korreloivat keskenään. Stationaarisen aikasarjan tapauksessa k:n aika-askeleen pässä olevien havaintojen z_t ja z_{t+k} kovarianssi on sama kaikille t. Tätä kovarianssia kutsutaan autokovarianssiksi viiveellä k ja se määritellään

$$\gamma_k = cov[z_t, z_t + k] = E[(z_t - \mu)(z_{t+k} - \mu)] \tag{19}$$

Vastaavasti viiveen k autokorrelaatio on

$$\rho_k = \frac{E[(z_t - \mu)(z_{t+k} - \mu)]}{\sqrt{E[(z_t - \mu)^2]E[(z_{t+k} - \mu)^2]}} = \frac{[(z_t - \mu)(z_{t+k} - \mu)]}{\sigma_z^2} \tag{20}$$

Koska stationaariselle aikasarjalle realisaation t varianssi $\sigma_z^2 = \gamma_0$ on sama kaikille ajanhetkille, niin z_t:n ja z_{t+k}:n välinen autokorrelaatio on

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{21}$$

Autokorrelaatiofunktiota voidaan käyttää mallin MA-osan identifioinnissa, sillä MA(q)-prosessissa autokorrelaatiot ρ_k poikkeavat nollasta kaikilla k jotka ovat pienempää tai yhtä suurta kuin q ja ovat nolla kaikilla k jotka ovat suurempaa kuin q. Tämän tiedon avulla voidaan arvioida aikasarjan MA-osan asteluku.

AR(p)-prosessin osittaisautokorrelaatiofunktio saadaan laskettua Yule-Walkerin yhtälöistä. Olkoon ϕ_{kj} j:s kerroin AR(k)-prosessissa jolloin ϕ_{kk} on prosessin viimeinen kerroin. Tällöin ϕ_{kj} on ratkaisu yhtälöille

$$\phi_j = \phi_{k1}\rho_{j-1} + ... + \phi_{k(k-1)}\rho_{j-k+1} + \phi_{kk}\rho_{j-k} \quad j = 1, 2, ..., k \tag{22}$$

joista seuraa Yule-Walkerin yhtälöt

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} \tag{23}$$

Nämä yhtälöt ratkaisemalla saadaan prosessin osittaisautokorrelaatiofunktio. AR(p)-prosessille osittaisautokorrelaatiofunktio ϕ_{kk} poikkeaa nollasta kaikilla k jotka ovat pienempiä tai yhtä suurta kuin p ja on nolla kaikilla k jotka ovat suurempia kuin p.

2.6 Suhteellinen keskivirhe

Suhteellinen keskivirhe (MAPE) on mitta mallin ennusteen tarkkuudelle. Havainnoille z_t ja havaintojen ennustetuille arvoille y_t MAPE määritellään

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \mid \frac{y_t - z_t}{y_t} \mid \tag{24}$$

MAPE:n heikkous on se, että jos aikasarjan taso on lähellä nollaa niin sen jakauma on hyvin vääristynyt ja voi antaa harhaanjohtavia tuloksia [Hyndman and Koehler, October - December 2006]. Tämä ei kuitenkaan ole ongelma

kyseessä olevan työnvoiman tarve-aikasarjan kanssa. MAPE:a hyödynnetään mittana käytettävän ennusteen pituuden suositukseen.

2.7 Box-Jenkinsin menetelmä

Ennustemallin tekemiseen käytettiin Box-Jenkinsin menetelmää [Box et al., 2008]. Menetelmä koostuu kolmesta vaiheesta jotka ovat

1. Mallin identifiointi
2. Mallin parametrien estimointi
3. Mallin validointi diagnostisin testein

Mallin identifioinnissa aikasarja tarkastetaan onko aikasarjassa havaittavissa trendiä ja stationarisoidaan aikasarja jos siihen on tarvetta. Tämän jälkeen tutkitaan mallin autokorrelaatio- ja osittaisautokorrelaatiofunktioita ja päätellään niistä minkä asteiset AR ja MA termit malliin otetaan. Lisäksi selvitetään onko aikasarjassa kausittaisuutta jota pitäisi mallintaa.

Kun dataan sovitettava malli on identifioitu, parametrien estimoinnissa estimoidaan parametrit valittuun malliin siten, että malli sopii annettuun dataan mahdollisimman hyvin. Parametrien estimoinnissa käytettiin suurimman uskottavuuden menetelmää.

Mallin validoinnissa tarkastetaan, että mallin virhetermejä koskevat oletukset pätevät. Tähän on käytössä useita diagnostisia testejä. Testattavia oletuksia ovat residuaalien korreloimattomuus, residuaalien normaalijakautuneisuus sekä residuaalien homoskedastisuus. Residuaalien korreloimattomuus testattiin Ljung-Box testillä ja residuaalien autokorrelaatio- ja osittaisautokorrelaatiokuvaajien perusteella. Normaalijakautuneisuutta testattiin Shapiro-Wilk testillä, sekä residuaalien histogrammin avulla. Residuaalien homoskedastisuutta arvioitiin residuaalikuvaajasta.

Ljung-Box testisuure on [Ljung and Box, Aug., 1978]

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{(n-k)} \tag{25}$$

jossa n on havaintojen määrä, r on residuaalien autokorrelaatiofunktio ja m on testattavien autokorrelaatioiden määrä. Tämä testisuure noudattaa ap-

proksimatiivisesti $\chi^2(m-p-q)$-jakaumaa, jossa p ja q ovat mallin viivepolynomien asteet. Jos testisuure poikkeaa paljon odotetusta on mallin nollahypoteesi residuaalien korreloimattomuudesta hylättävä.

Shapiro-Wilk testisuure on [Shapiro and Wilk, Dec., 1965]

$$W = \frac{(\sum_{i=1}^{n} a_i y_{(i)})^2}{\sum_{i=1}^{n} y_i - \bar{y}} \tag{26}$$

jossa n on havaintojen määrä, $y_{(i)}$ i:s järjestystunnusluku eli i:nneksi pienin havaintojen y arvo, $\bar{y}$ on havaintojen aritmeettinen keskiarvo ja vakiot a_i saadaan kaavasta

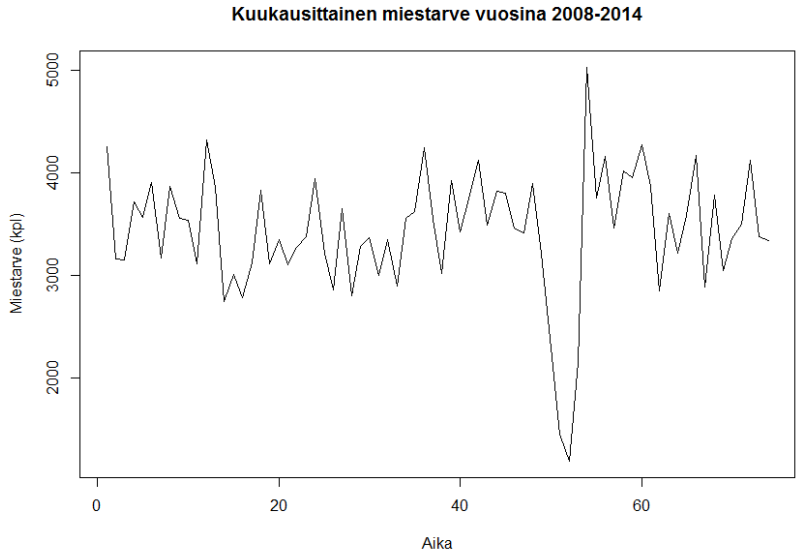
$$(a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

jossa $m = (m_1, ..., m_n)^T$ ja $m_1, ..., m_n$ ovat normaalijakautuneiden satunnaismuutujien järjestystunnuslukujen odotusarvot ja V on näiden järjestystunnuslukujen kovarianssimatriisi. Testisuure W ei noudata mitään tavanomaista jakaumaa, mutta testisuureen arvojen p-arvoja löytyy taulukoista. Lisäksi aikasarjan analysointiin käytetty R-ohjelmisto ilmoittaa Shapiro-Wilk testin p-arvon. Pienet testisuureen arvot johtavat normaalisuushypoteesin hylkäämiseen.

3 Tulokset

3.1 Aikasarja

Ennustettava aikasarja on esitetty kuvassa 1. Aikasarjassa on esitetty yrityksen kuukausittainen työvoiman tarve välillä tammikuu 2008, helmikuu 2014. Aikasarjassa ei näy trendiä ja aikasarjan vaihteluväli pysyy tasaisena, joten aikasarja on stationaarinen. Ajanhetkillä 50-53 olevat työvoiman tarpeet ovat selkeästi muita alhaisempia. Nämä kuukaudet ovat vuoden 2011 joulukuu, sekä vuoden 2012 tammikuu, helmikuu ja maaliskuu. Näinä kuukausina yrityksessä otettiin käyttöön uusi työvoiman tarpeen seurantaohjelmisto. Tällöin ohjelmistoon ei vielä tallentunut kaikki työvoiman tarpeet joten näitä neljää datapistettä käsiteltiin puuttuvina. Puuttuvien datapisteiden estimointiin käytettiin R:n arima-funktion sisäänrakennettua puuttuvien datapisteiden käsittelyä. R:n käyttämä puuttuvien datapisteiden estimointialgoritmi

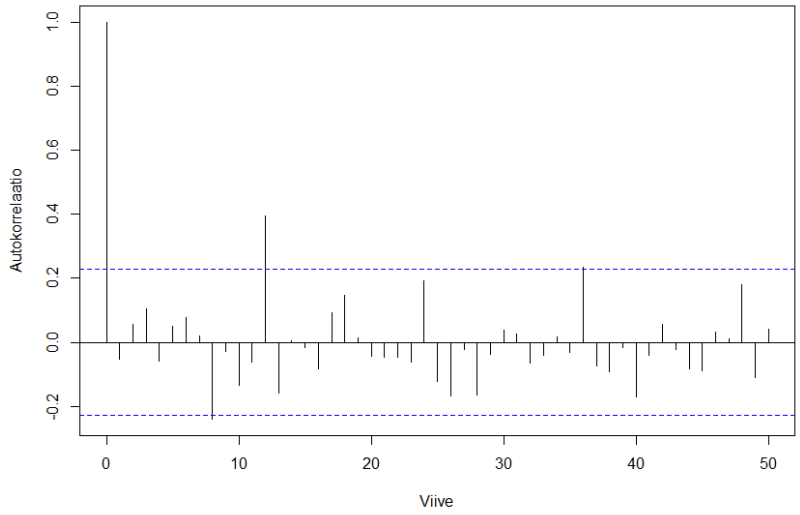


Kuva 1: Ennustettava aikasarja

on esitetty kirjassa Durbin and Koopman [2001]. Heti näiden kuukausien jälkeisenä kuukautena työvoiman tarve oli huomattavan suurta. Tähän yksittäiseen suureen tarpeeseen ei löytynyt mitään selitystä, mutta piste on selkeästi poikkeava. Piste otettiin mallissa huomioon dummy muuttujalla P(t) siten että P(t) = 1 kun t=54 ja P(t)=0 kun t≠54. Mallina käytetään siis additiivisen poikkeavan havainnon mallia [Tsay, No. 393 (Mar., 1986].

3.2 Ulkoinen selittäjä

Ennen kuin aikasarjalle estimoitiin malli, pyrittiin löytämään aikasarjalle ulkoinen selittäjä. Kokeiltuja ulkoisia selittäjiä olivat asuntojen hinta, asuntojen hintaindeksi, korkotaso, kuluttajien luottamus talouteen, mikrotalouden indikaattori ja kestotavaroiden ostoaikomus. Kaikki tilastot haettiin tilastokeskuksen sivuilta ja tilastot haettiin ajanjaksolta tammikuu 2008 - helmikuu 2014, paitsi asuntoihin liittyvät indeksit joita oli saatavilla vasta tammikuusta 2010 alkaen. Kokeiltujen ulkoisten selittäjien ja työvoiman tarpeen väliset ristikorrelaatiot on esitetty liitteessä A. Jotta ulkoistan selittäjää voidaan käyttää ennustamisessa pitää sen ja ennustettavan aikasarjan välillä olla ristikorrelaatiota selittäjän menneen arvon ja aikasarjan nykyisen arvon välillä.

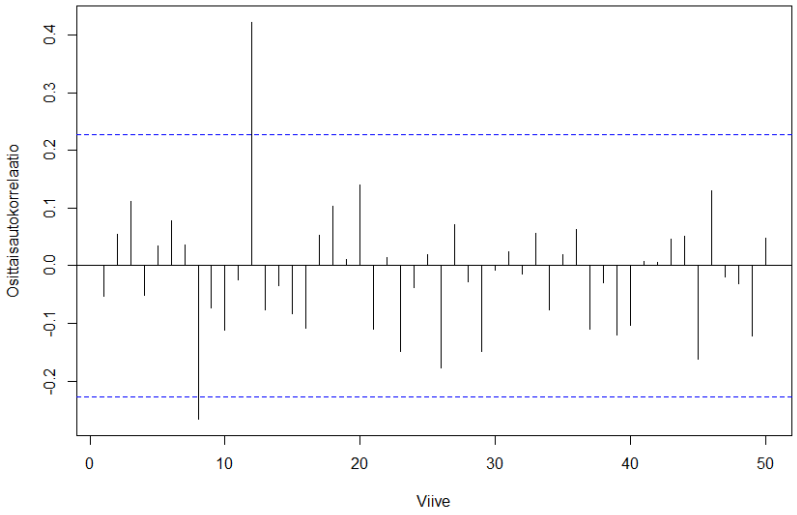


Kuva 2: Työvoiman tarve-aikasarjan autokorrelaatiofunktio. Ensimmäinen piirretty autokorrelaatio on havainnon autokorrelaatio itsensä kanssa.

Työssä ei päädytty käyttämään ulkoista selittäjää aikasarjamallissa. Yksikään tutkittujen ulkoisten selittäjien ja työvoiman tarpeen välisistä ristikorrelaatioista ei ole paljoa yli 5%:n luottamustason, joten vahvaa korrelaatiota kokeiltujen ulkoisten selittäjien ja työvoiman tarve-aikasarjan välillä ei löytynyt. Suurin ristikorrelaatio positiivisella viiveellä havaitaan asuntojen hintaindeksin aikasarjassa. Kuitenkin asuntojen hintaindeksiä on tilastoitu vasta tammikuusta 2010 alkaen, ja suurin ristikorrelaation arvo havaitaan viiveellä k=6. Tämä jättäisi aikasarjan estimointiin vain 32 datapistettä ja jos aikasarjan ennustevoimaa testataan esimerkiksi kuuden kuukauden päähän, jäisi aikasarjan estimointiin enää 26 datapistettä mitä ei koettu riittävän pitkäksi ajanjaksoksi aikasarjamallin luotettavaan estimoimiseen.

3.3 Aikasarjamallin tekeminen

Työvoiman tarve-aikasarjan autokorrelaatio ja osittaisautokorrelaatiofunktiot on esitetty kuvissa 2 ja 3. Autokorrelaatiofunktion kuvassa ensimmäinen piirretty autokorrelaatio on havainnon autokorrelaatio viiveellä 0, eli autokorrelaatio itsensä kanssa. Autokorrelaatio ja osittaisautokorrelaatiofunktioiden perusteella aikasarjaan voisi kokeilla mallia $(0,0,0)\mathrm{x}(1,0,1)_{12}$. Kun tämä malli sijoitetaan aikasarjaan ja estimoidaan parametrit havaitaan, että mallin

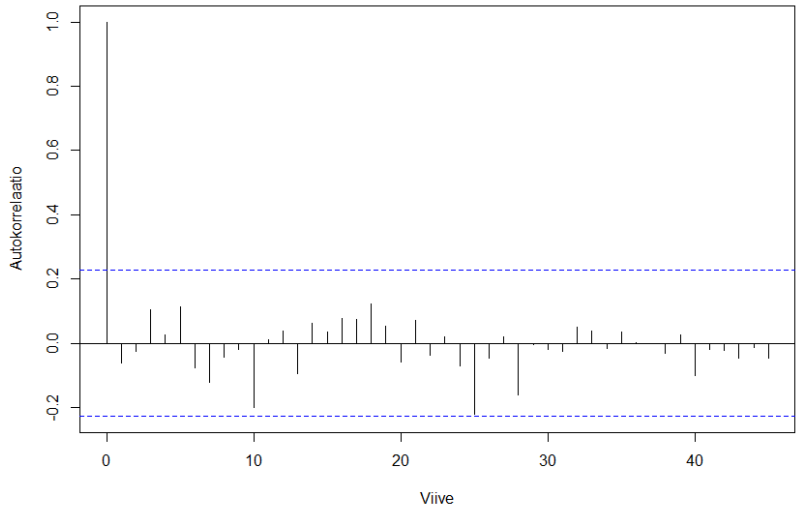


Kuva 3: Työvoiman tarve-aikasarjan osittaisautokorrelaatiofunktio.

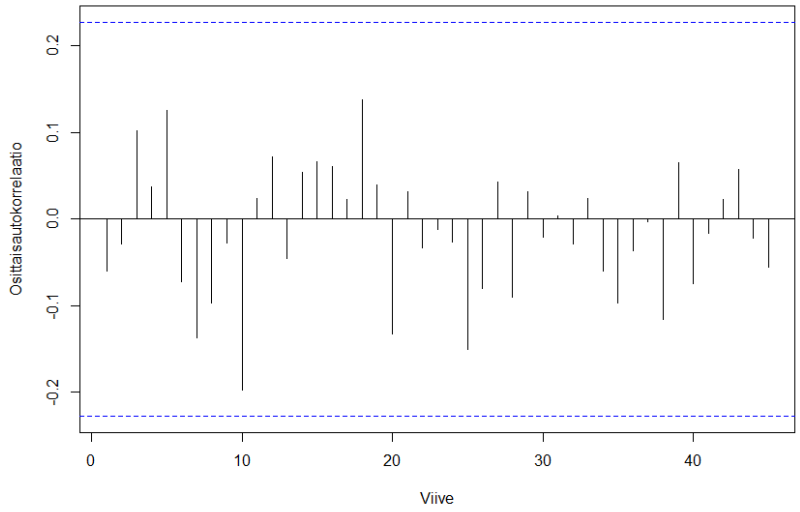
residuaaleissa on merkitsevää auto- ja osittaisautokorrelaatiota. Residuaalit eivät myöskään läpäise Box-Ljungin testiä residuaalien autokorreloituneisuudelle 5% merkitsevyystasolla useilla viiveillä. Residuaalien ensimmäinen osittaisautokorrelaatio sekä ensimmäinen ja kolmas autokorrelaatio ylittävät merkitsevyysrajan.

Näiden havaintojen perusteella kokeiltiin sueraavaksi mallia $(1,0,1)\mathrm{x}(1,0,1)_{12}$. Nyt residuaalien autokorrelaatiot ja osittaisautokorrelaatiot eivät olleet merkitseviä (kuvat 4 ja 5) ja residuaalit läpäisivät Box-Ljungin testin 5% merkitsevyystasolla kaikilla viiveillä (kuva 6). Residuaalien histogrammi on esitetty kuvassa 7. Histogrammi muistuttaa normaalijakauman todennäköisyystiheysfunktiota. Lisäksi Shapiro-Wilkin testi normaalisuudelle antaa residuaaleille p-arvon 0.1083, eli testin nollahypoteesia residuaalien normaalijakautuneisuudesta ei voida hylätä 5%:n luottamustasolla. Residuaalit on piirretty aikaa vastaan kuvassa 8. Kuvasta nähdään, että residuaalipilvi on tasalevyinen eli residuaalit ovat homoskedastisia. Malli $(1,0,1)\mathrm{x}(1,0,1)_{12}$ läpäisee siis kaikki sille asetetut diagnostiset testit.

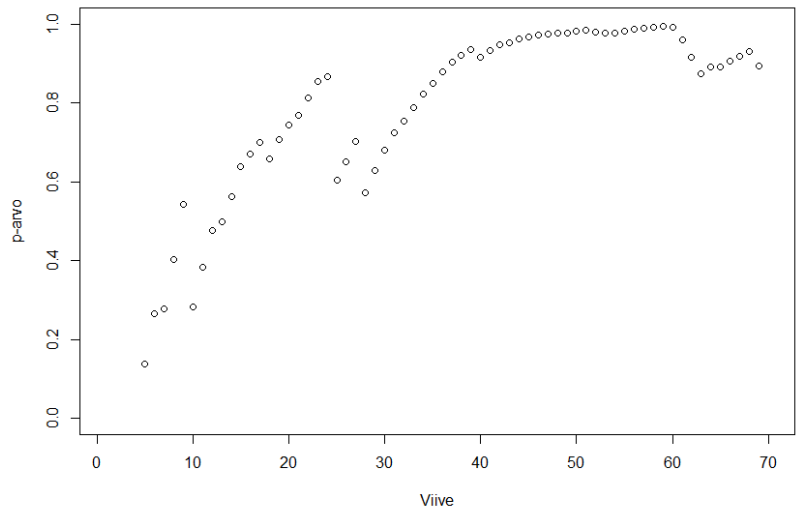
Ennustettavan ajanjakson pituutta ehdotettiin perustuen suhteelliseen keskivirheeseen (MAPE). Tätä varten aikasarjasta piilotettiin aina ennustettava määrä kuukausia, estimoitiin aikasarja näiden kuukausien avulla ja ennus-



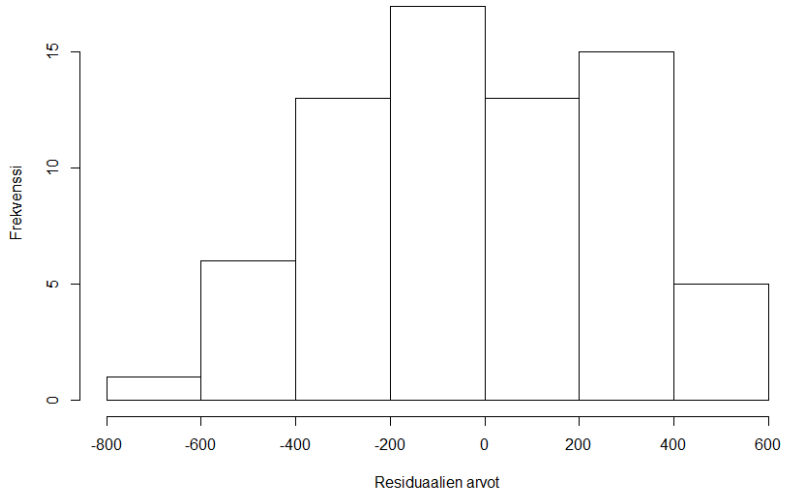
Kuva 4: Lopullisen aikasarjamallin residuaalien autokorrelaatiofunktio.



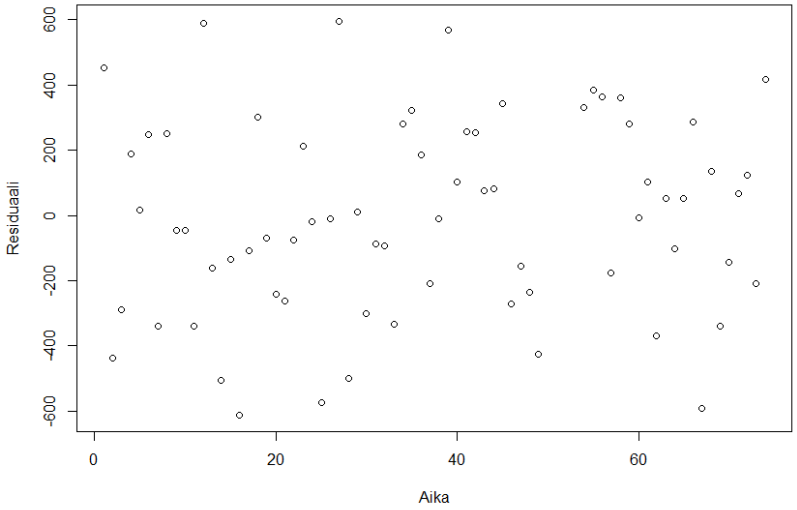
Kuva 5: Lopullisen aikasarjamallin residuaalien osittaisautokorrelaatiofunktio.



Kuva 6: Lopullisen aikasarjamallin Box-Ljung testin tulokset viiveille 5-69.



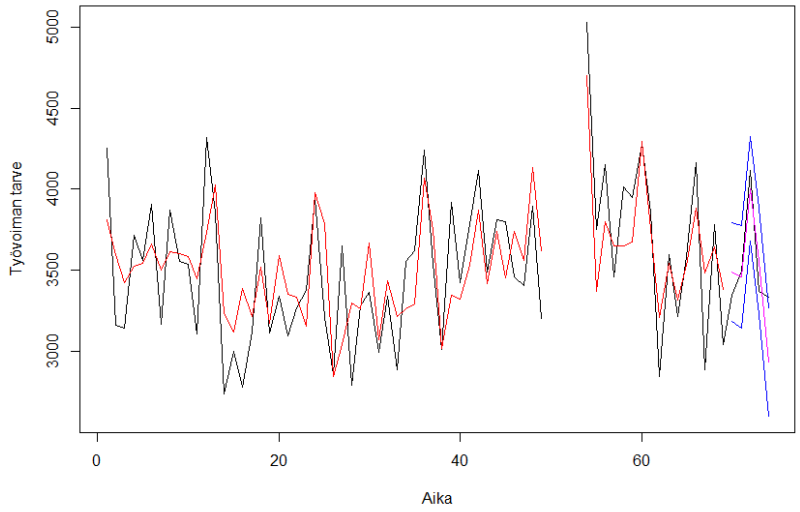
Kuva 7: Lopullisen aikasarjamallin residuaalien histogrammi.



Kuva 8: Lopullisen aikasarjamallin residuaalit.

Taulukko 1: Lopullisen aikasarjamallin ennusteiden MAPE:t.

Ennusteen pituus (kk)	MAPE
1	0.1562
2	0.0971
3	0.0732
4	0.0625
5	0.0534
6	0.0628
7	0.0563
8	0.0877
9	0.0788



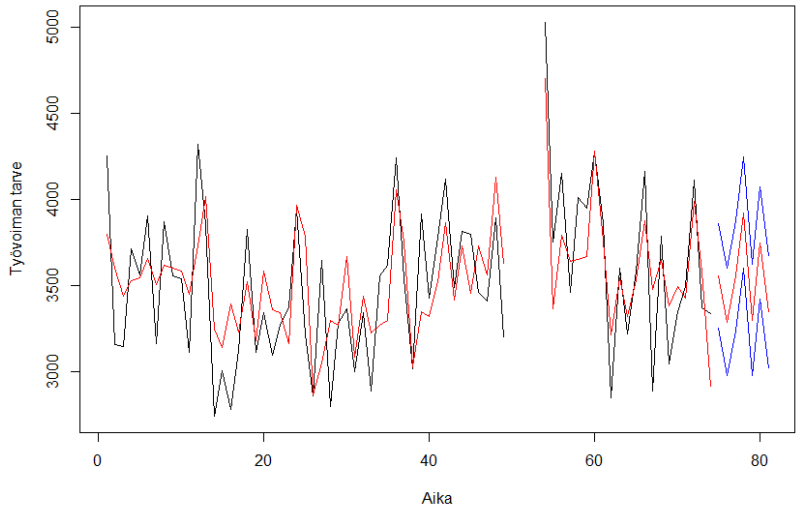
Kuva 9: Lopullisen aikasarjamallin viiden kuukauden ex-post ennuste ennustettavaan aikasarjaan. Ennusteen luottamusvälit on kuvattu sinisellä.

Taulukko 2: Lopullisen mallin antama ennuste seuraavalle viidelle kuukaudelle, sekä ennusteen 68% luottamusvälit.

Kuukausi	Työvoiman tarve	Luottamusväli
Maaliskuu	3557.746	[3254.202, 3861.290]
Huhtikuu	3289.624	[2975.858, 3603.390]
Toukokuu	3555.342	[3235.191, 3875.493]
Kesäkuu	3924.121	[3603.647, 4244.595]
Heinäkuu	3299.434	[2976.221, 3622.648]

tettiin piilotettujen kuukausien työvoiman tarve. Tämän jälkeen laskettiin ennusteen MAPE. MAPE:n arvot eri ennustusjaksoilla on esitetty taulukossa 1. MAPE:n perusteella paras ennusteen pituus on viisi kuukautta, jolloin suhteellinen keskivirhe oli pienin. Viiden kuukauden ex-post ennuste on esitetty kuvassa 9.

Kuvassa 10 on esitetty lopullisen aikasarjamallin sovite ennustettavaan aikasarjaan. Kuvassa on lisäksi ennustettavan aikasarjan viiden kuukauden ennuste työvoiman tarpeelle ja ennusteen 68%:n luottamusvälit. Taulukossa 2 on esitetty mallin antamat ennusteet viidelle seuraavalle kuukaudelle, sekä ennusteen 68% luottamusvälit.



Kuva 10: Lopullisen aikasarjamallin sovite ennustettavaan aikasarjaan ja viiden kuukauden ennuste. Ennustettava aikasarja on piirretty mustalla, sovite punaisella ja ennusteen 68% luottamusvälit sinisellä.

4 Tarkastelu

Lopulliseksi aikasarjamalliksi työvoiman tarpeen ennustamiseen saatiin malli

$$z_t = 3528.8028 + P(t) + 0.8117z_{t-1} + 0.9830z_{t-12} - 0.5475a_{t-1} - 0.7959a_{t-12}$$

jossa P(t) = 946.5832 kun t = 54 ja P(t) = 0 kun t≠54.

Mallista tarkasteltavan ennusteen pituudeksi ehdotetaan viittä kuukautta. Mallista huomataan, että kuukauden työvoiman tarve muistuttaa paljon edellisen vuoden saman kuukauden työvoiman tarvetta. Lisäksi edellisen kuukauden työvoiman tarpeella on suuri vaikutus tämän kuukauden työvoiman tarpeeseen.
Satunnaisvaihtelun termien merkit ovat negatiivisia. Usein taloudellisissa aikasarjoissa SARMA-mallien satunnaistermit tulkitaan markkinashokeiksi. Shokkien merkin negatiivisuus viittaisi siihen, että alalla jolla yritys toimii työn tarve on jossain määrin vakio. Jos edellisenä kuukautena tai vuotena on tehty vähemmän töitä kuin on ennustettu niin töitä on tulevaisuudessa enemmän. Toisaalta jos töitä on tehty enemmän kuin mitä ennuste sanoo,

tulee seuraavana kuukautena olemaan vähemmän töitä tarjolla.

Mallin antaman ennusteen 68%:n luottamusvälit ovat isot. Lisäksi aikasarjan viimeinen realisaatio kuvassa 9 ei ole ennusteen luottamusvälien sisällä, joka toisaalta on odotettavaa sillä ennustettavia kuukausia oli viisi. Mallin viiden kuukauden ex-post ennusteen suhteellinen keskivirhe oli kohtuullinen 5.34%, minkä perusteella malli on käyttökelpoinen työvoiman tarpeen ennustamiseen. Mallin jatkokehitys tulisi kohdistaa mallin tarkkuuden parantamiseen ja luottamusvälien pienentämiseen, mikä voisi onnistua esimerkiksi ulkoisen muuttujan lisäämisellä malliin. Työssä havaittu ristikorrelaatio asuntojen hintaindeksin kanssa antaa lupaavan lähtökohdan ulkoisen selittäjän sisältävän SARMAX-mallin kehittämiseen. Tähän kuitenkin tarvitaan vielä useampia havaintoja asuntojen hintaindeksistä. Jatkossa ennustemallia tulee myös päivittää kuukausittain estimoimalla uudet parametrit malliin kun lisähavaintoja työvoiman tarpeesta saadaan. Lisäksi malliin liittyvät oletukset olisi hyvä testata aina uusien parametrien estimoinnin yhteydessä.

Viitteet

George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis, Fourth Edition.* John Wiley & Sons, Inc., 2008.

J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods.* Oxford University Press, 2001.

A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153:3–27, 2004.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22, Issue 4:679–688, October - December 2006.

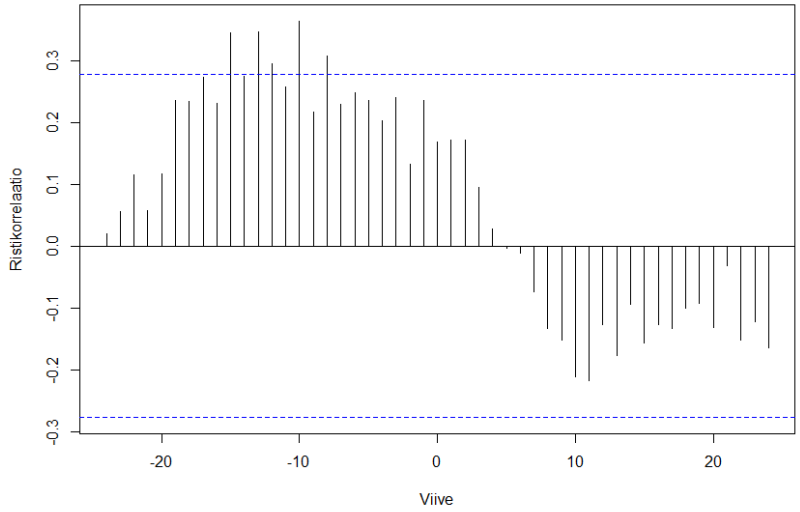
G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65, No.2:297–303, Aug., 1978.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52, No. 3/4:591–611, Dec., 1965.

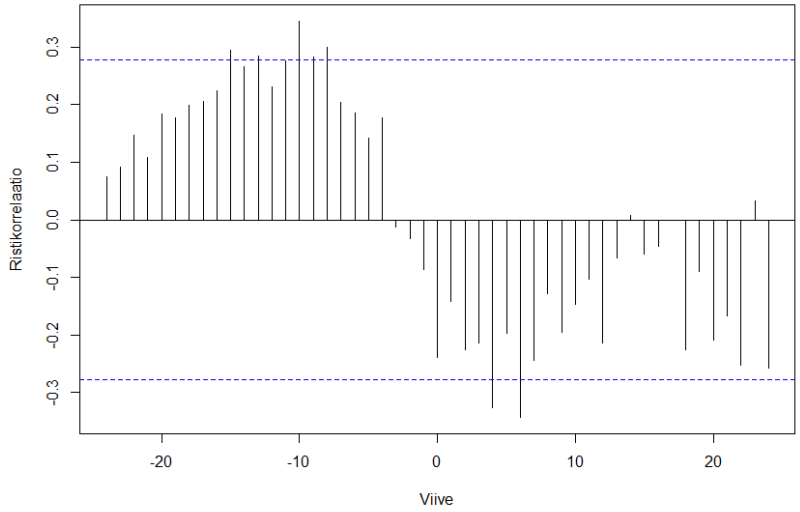
Ruey S. Tsay. Time series model specification in the precence of outliers. *Journal of the American Statistical Association*, Vol. 81:132–141, No. 393 (Mar., 1986).

A Ulkoisten selittäjien ja aikasarjan väliset ristikorrelaatiot

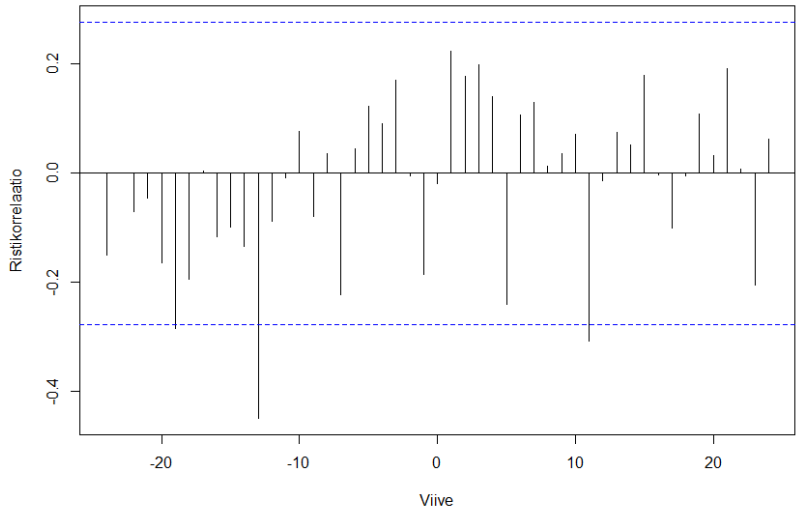
Tässä liitteessä on kuvaajat työssä tutkittujen mahdollisten ulkoisten selittäjien ja ennustettavan aikasarjan välisistä ristikorrelaatioista. Ristikorrelaatiot on esitetty siten, että viiveen k ristikorrelaatio tarkoittaa ristikorrelaatiota z_{t+k}:n ja y_t:n välillä, missä y_t on ulkoinen selittäjä ajanhetkellä t ja z_{t+k} aikasarjan arvo ajanhetkellä t+k. Jotta ulkoista selittäjää voitaisiin käyttää aikasarjan ennustamisessa, pitää työvoiman tarpeen ja ulkoisen selittäjän välillä olla merkitsevää ristikorrelaatiota viivellä k>0. Siniset viivat kuvissa näyttävät 5%:n luottamustason.



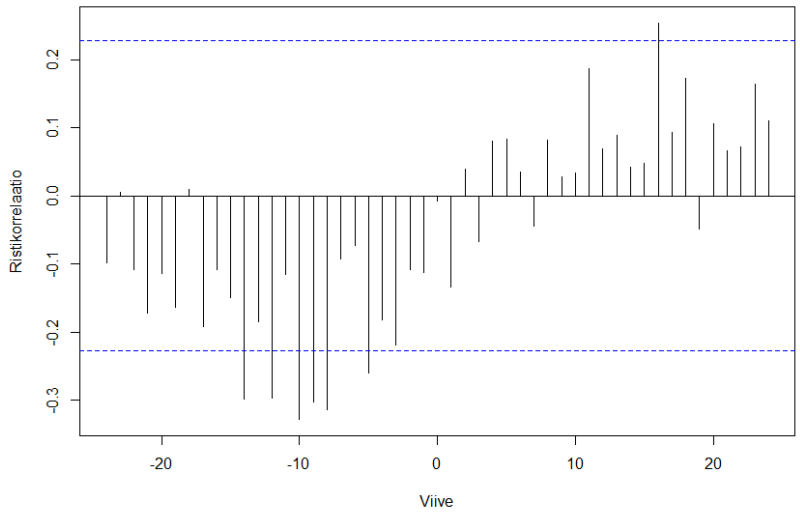
Kuva 11: Asuntojen hinnan ja työvoiman tarpeen välinen ristikorrelaatio.



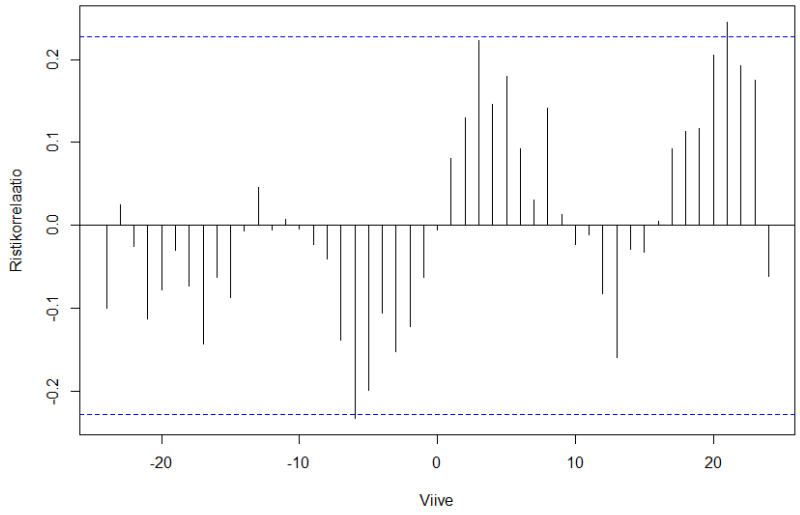
Kuva 12: Asuntojen hintaindeksin ja työvoiman tarpeen välinen ristikorrelaatio.



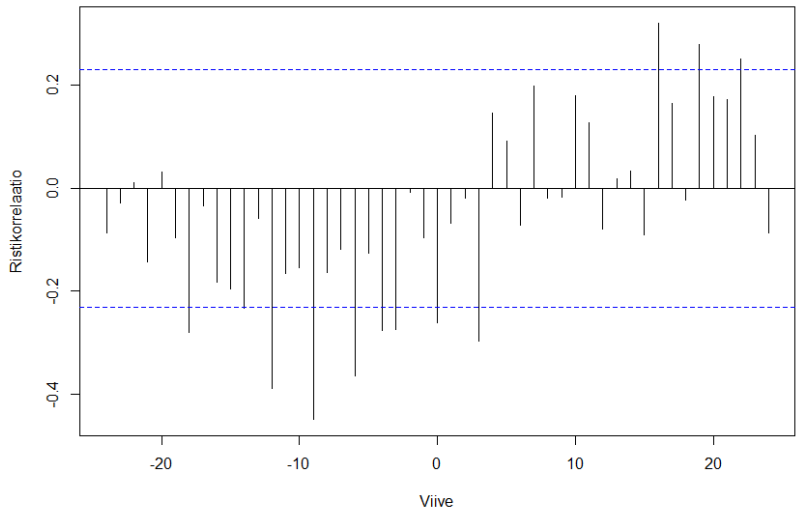
Kuva 13: Asuntokauppojen määrän ja työvoiman tarpeen välinen ristikorrelaatio.



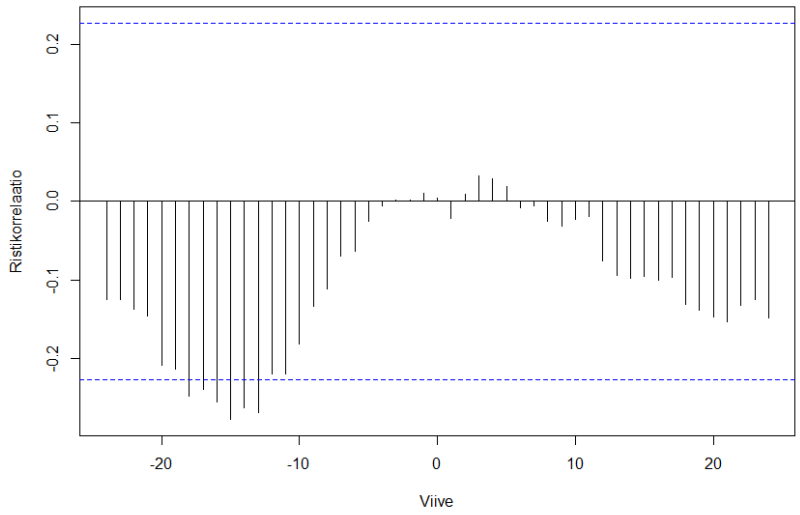
Kuva 14: Kestotavaroiden ostoaikomuksen ja työvoiman tarpeen välinen ristikorrelaatio.



Kuva 15: Kuluttajien luottamuksen talouteen ja työvoiman tarpeen välinen ristikorrelaatio.



Kuva 16: Mikrotalouden indeksin ja työvoiman tarpeen välinen ristikorrelaatio.



Kuva 17: Korkotason ja työvoiman tarpeen välinen ristikorrelaatio.