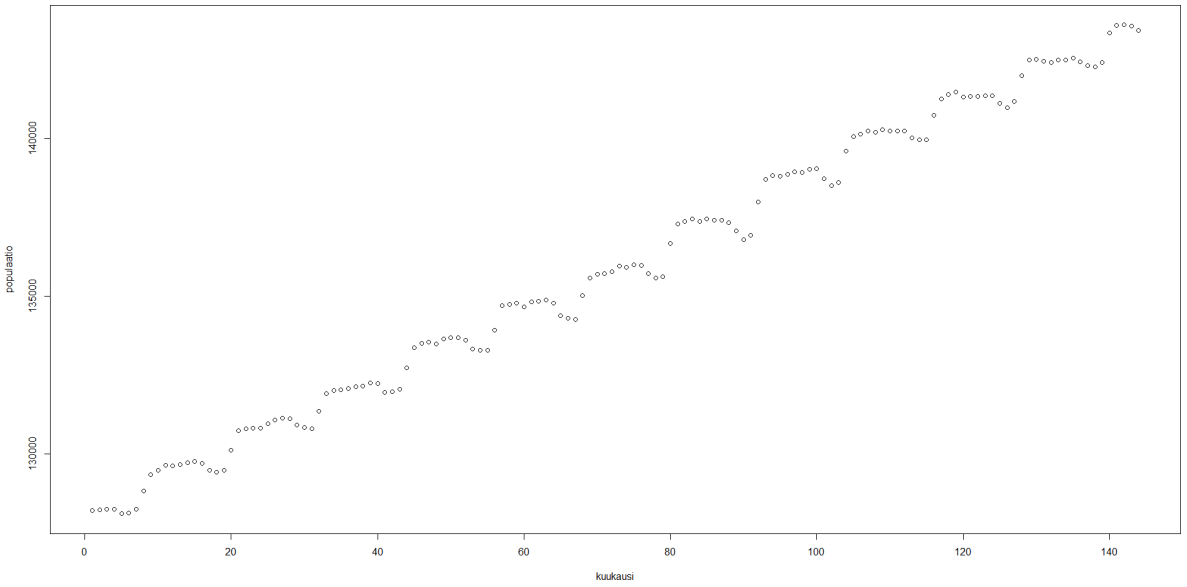
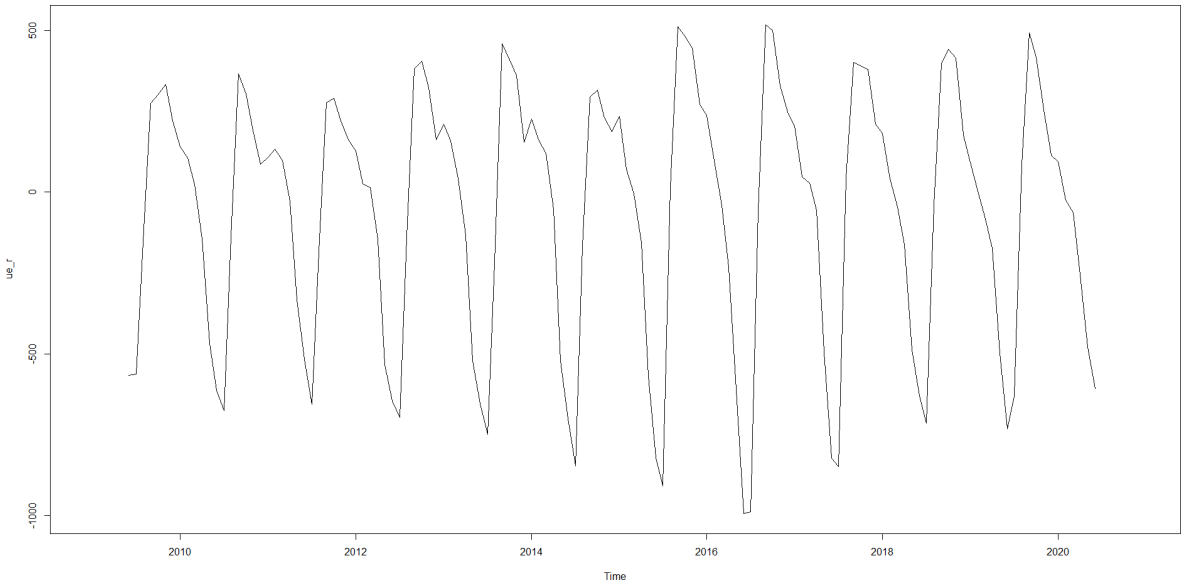
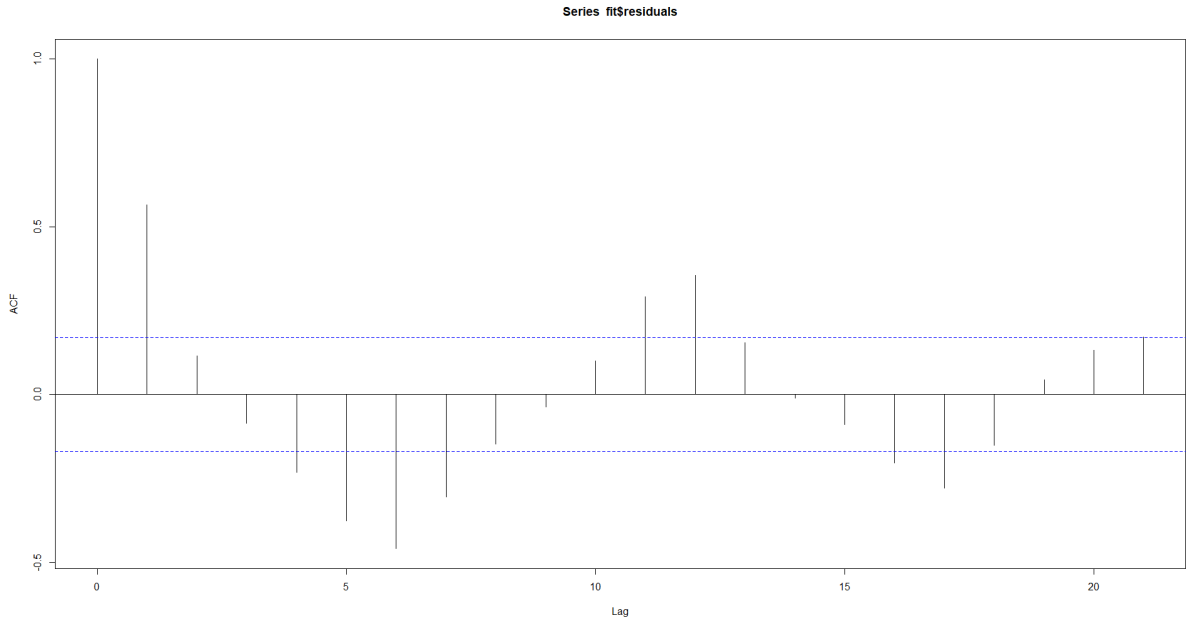
Tutkielman aiheena on yrittää ennustaa Jyväskylän väestönkasvua vuosille 2021 ja 2022. Aineistona on csv tiedosto jossa on jyväskylän asukasluvut vuodelta 1990 vuoteen 2020. Meitä kiinnostaa vain luvut vuoden 2009 tammikuusta eteenpäin, sillä silloin tapahtui viimeisimmät kuntaliitokset. Aloitetaan tarkastelemalla aineistoa.



Datasta nähdään selvä väestönkasvu vuodesta 2009 vuoteen 2020. keväisin nähdään väestönlasku. Elo-ja syyskuussa nähdään korkea väestönkasvu. Tähän voidaan antaa mahdollisena selityksenä se, että Jyväskylä on opiskelijakaupunki ja opiskelijat muuttavat pois valmistuessaan keväällä ja uudet opiskelijat muuttavat kaupunkiin elo-ja syyskuussa, kun uusi lukukausi alkaa. Tarkastellaan vielä miltä kuvaaja näyttää kun aineistosta vähennetään keskiarvo.

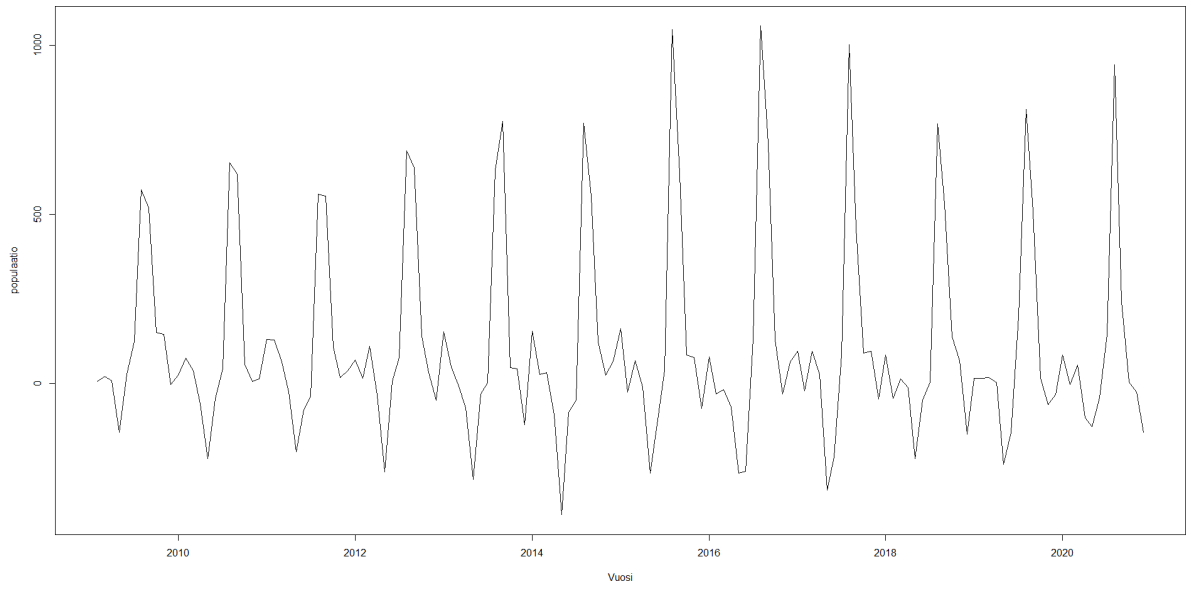


ja lopuksi aineiston residuaalin autokorrelaatio



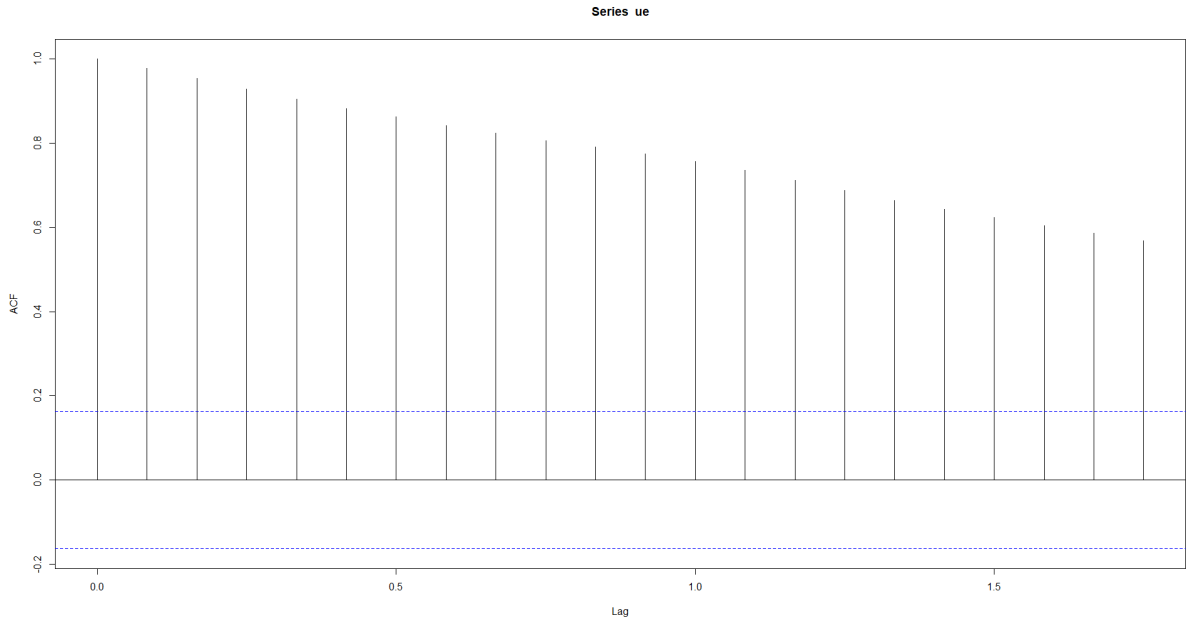
Voidaan kuvaajasta nähdä selvää kausivaihtelua.

Seuraavaksi pitäisi valita sopivat mallit joihin aineisto voitaisiin sovittaa. Arma mallit sopivat stationaarisiin aineistoihin mutta, kuten ylempänä nähdään niin aineisto muokkaamattomana ei ole kovin stationaarinen. Voidaan kuitenkin tarkistaa jos aineiston differenssi olisi stationaarisempi.

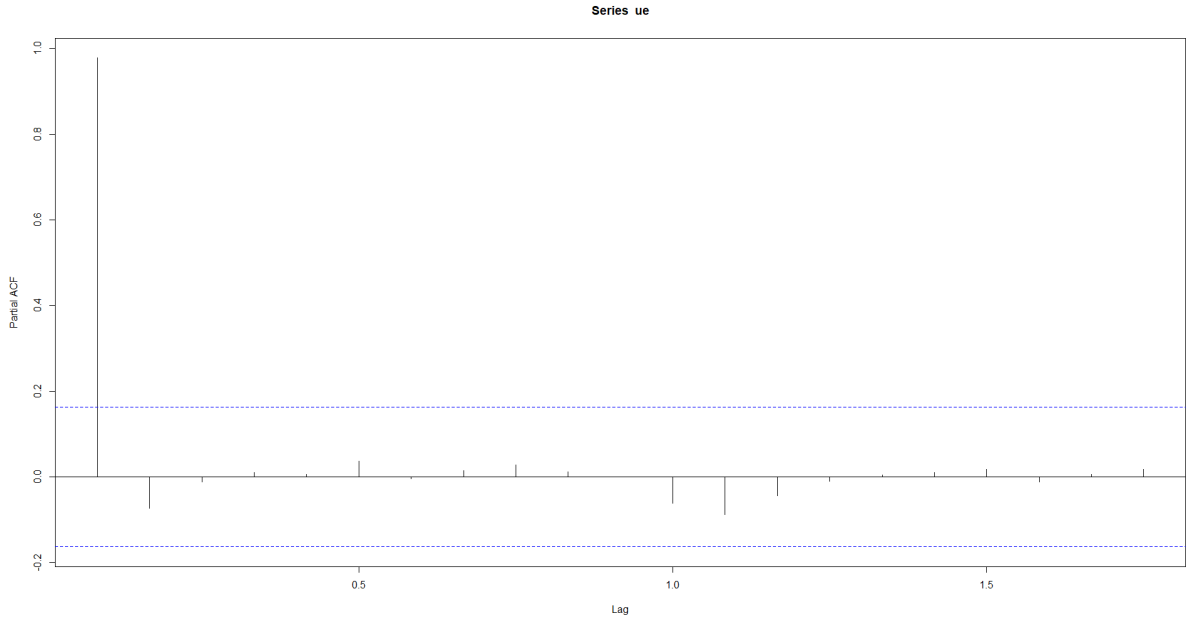


differenssi on stationaarisen näköinen, joten voidaan jatkaa aineiston ja differenssin kanssa. Aloitetaan tarkastelemalla differenssin ja aineiston autokorrelaatiota ja osittaisia autokorrelaatioita joiden avulla voidaan yrittää päätellä hyviä ar ja ma kertoimia.

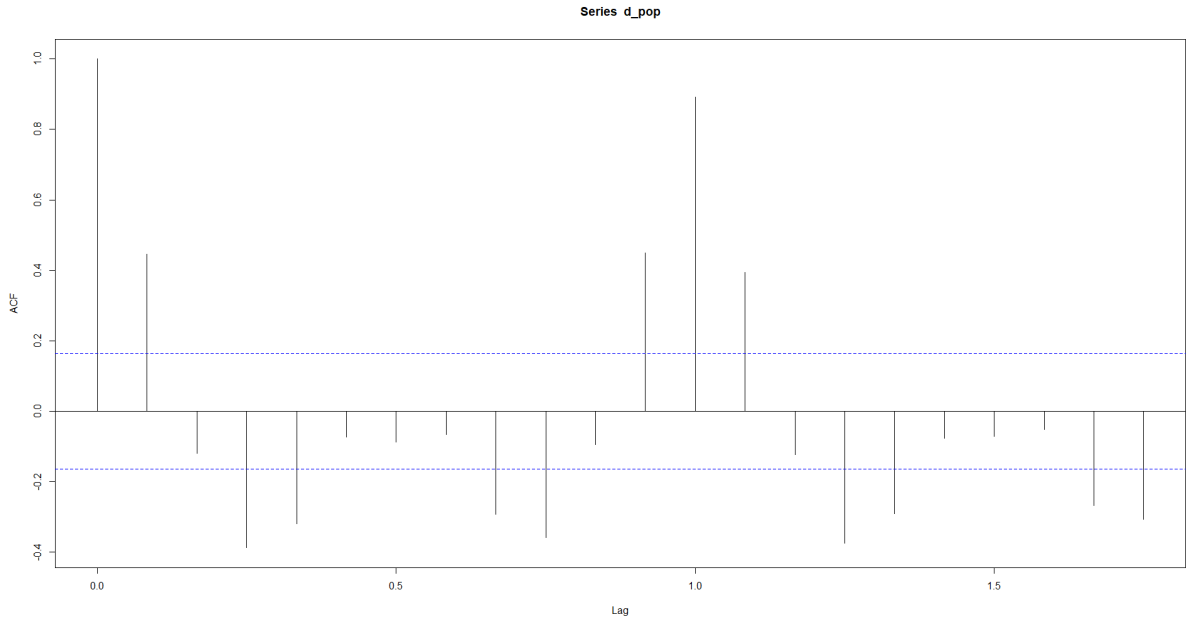
aineiston autokorrelaatiosta on vaikea sanoa mitään



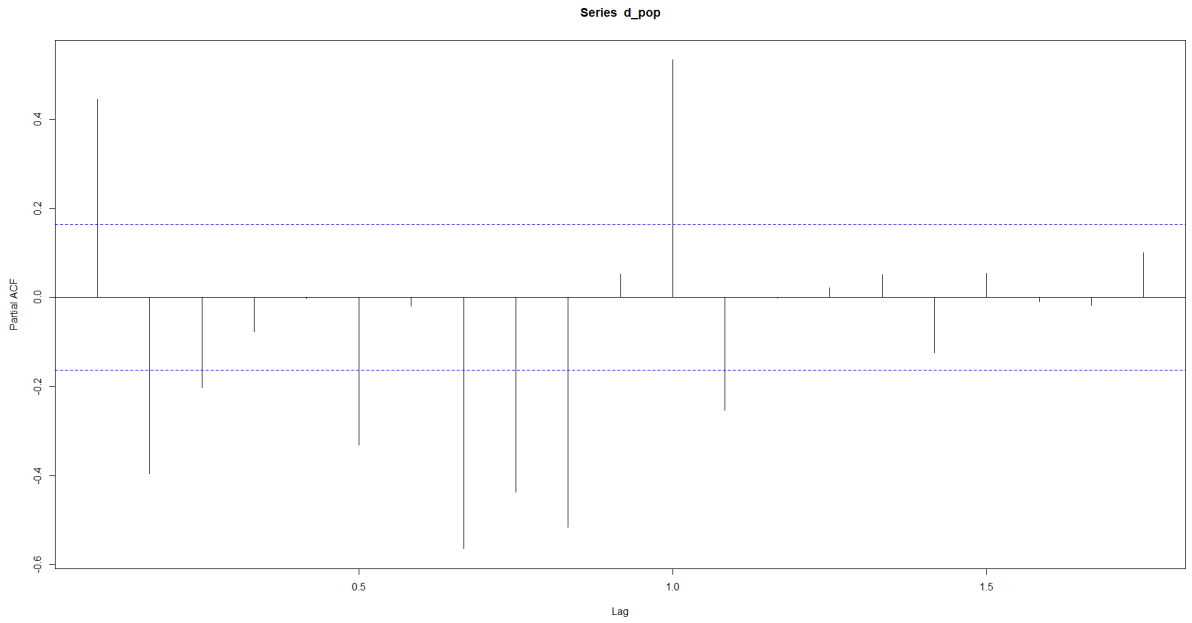
osittaisesta taas saa jotain irti



AR kertoimelle voisi antaa arvoksi 1 sillä arvot ovat lähellä nollaa viiveestä 1 lähtien ja MA voisi myös olla 1 samasta syystä. tarkastellaan kuitenkin vielä differenssin autokorrelaatioita.

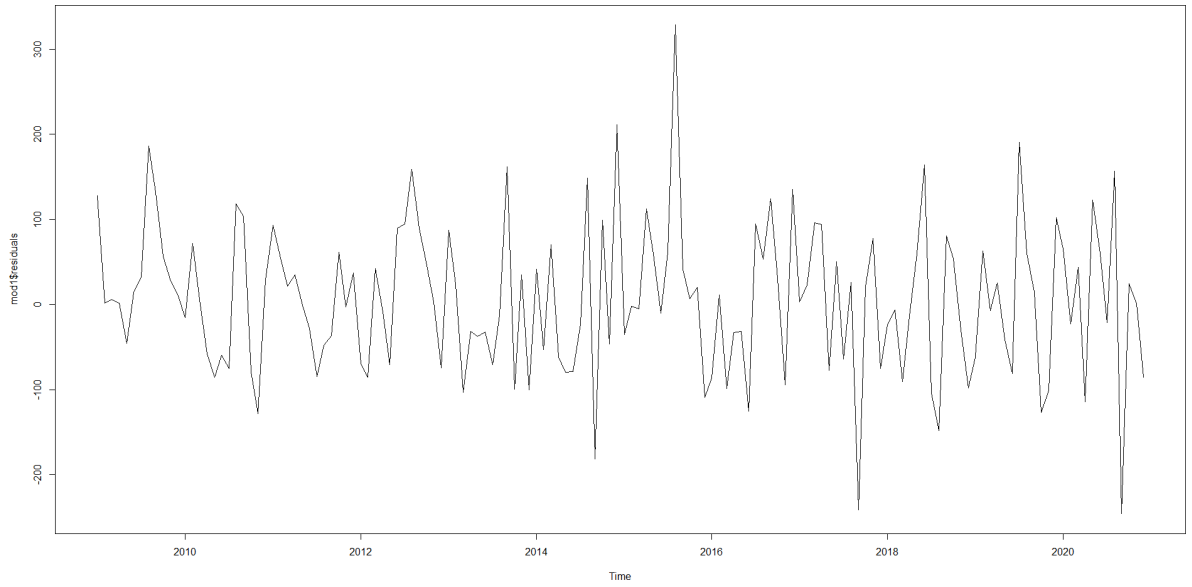


Datassa nähdään selvä trendi mutta en oikein voisi sanoa, että jokin AR tai MA kerroin olisi hyvä datalle. Lopuksi vielä differenssin osittaisautokorrelaatio

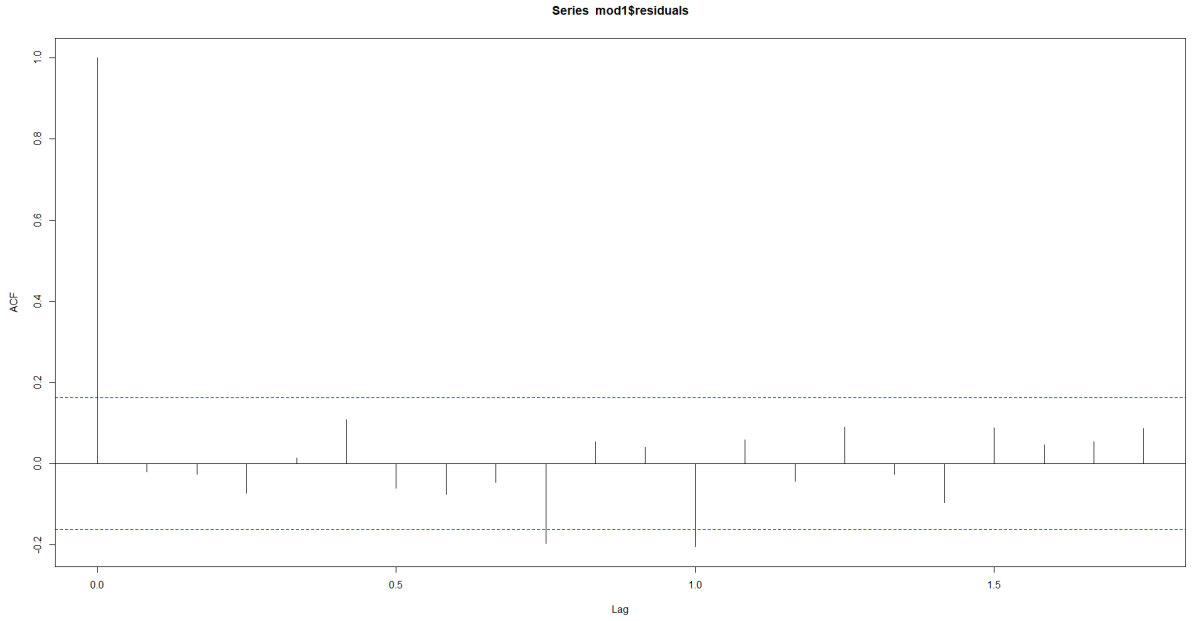


Tästäkään ei oikein saa sanottua jos jokin kerroin olisi hyvä, mutta datassa on kuitenkin jonkinlaista trendiä.

autokorrelaatioissa ja aineistossa näkyy kausivaihtelua, joten sopivaksi malliksi voisi sopia sarima malli, koska siihen on rakennettu valmiiksi mukaan kausivaitelevuus. Laitetaan mallille kausi 12 sillä se sopii vuoden mukaiseksi ja lisäksi laitetaan siihen aiemmin mietityt AR ja MA kertoimet. Lisäksi laitoin argumentiksi, että funktio katsoo edellisen vuoden arvoa.

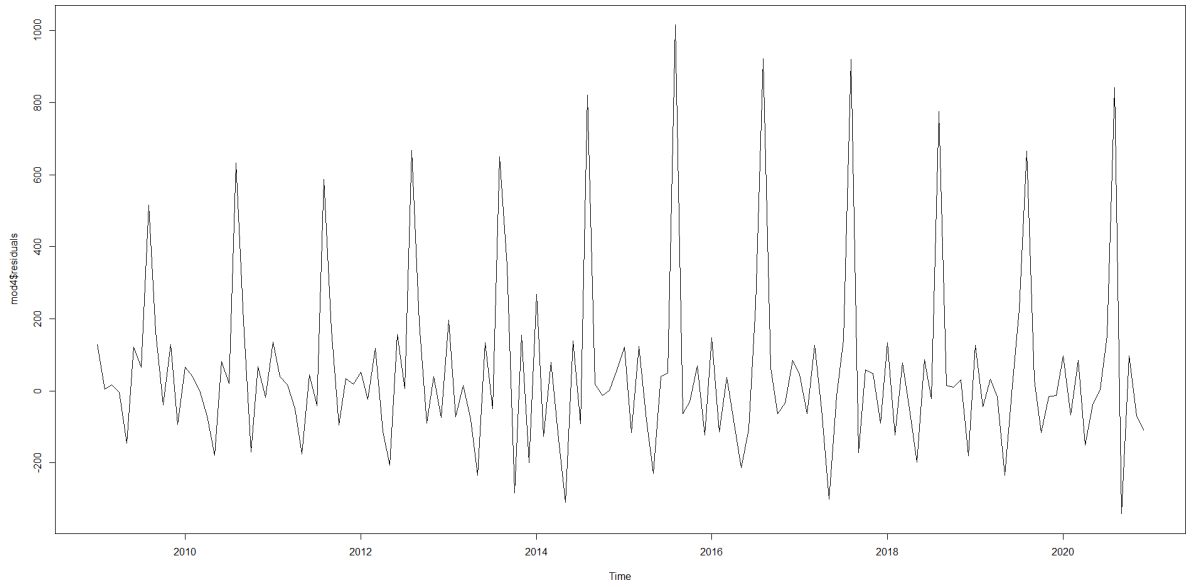


Mallin residuaali

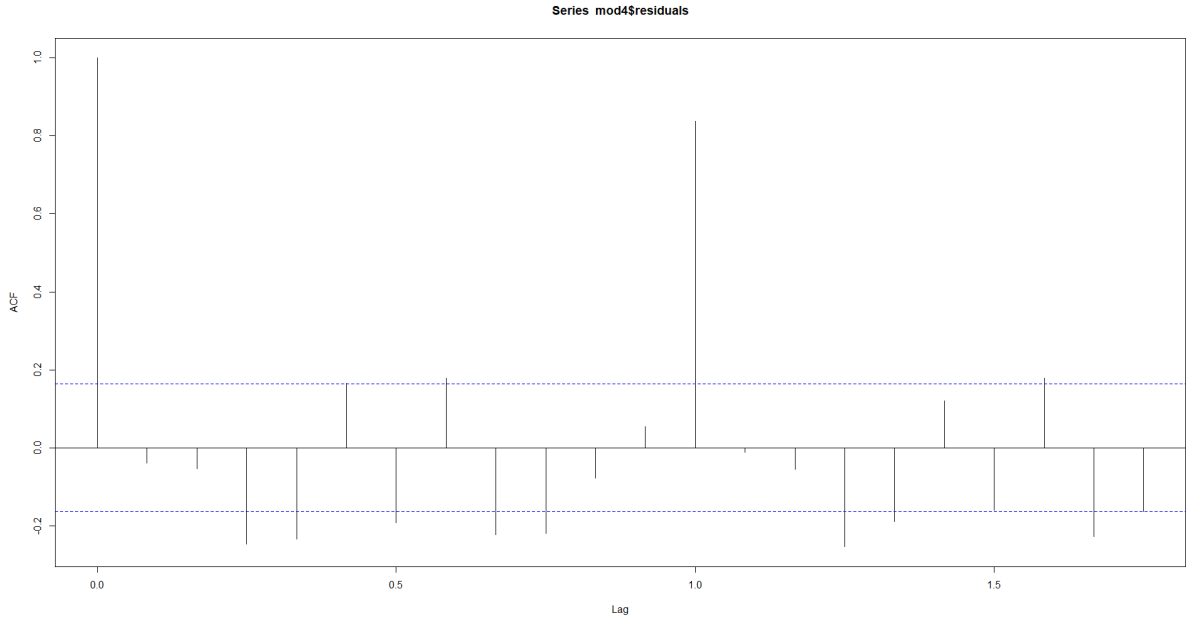


Mallin autokorrelaatio

Sanoisin, että malli näyttäisi hyvältä mutta kuitenkin voisi olla hyvä tarkistaa jos normaali arima malli sopisi aineistoon. Alla arima samoilla arvoilla kuin sarima

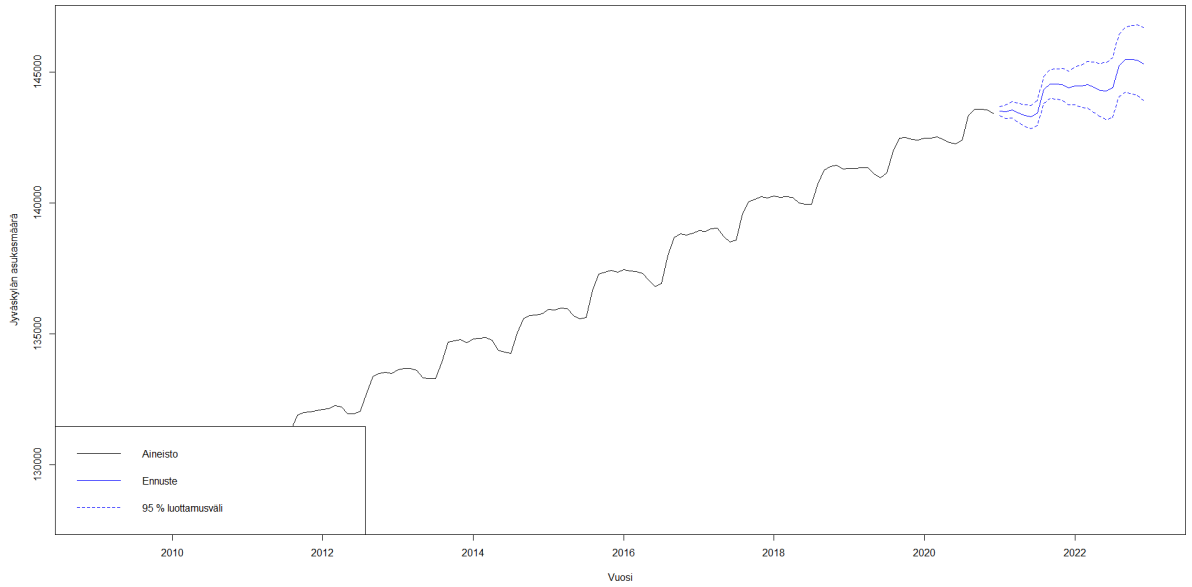


arima mallin differenssi näyttäisi, että siinä on enemmän vaihtelua ja vahvoja piikkejä kausittain.

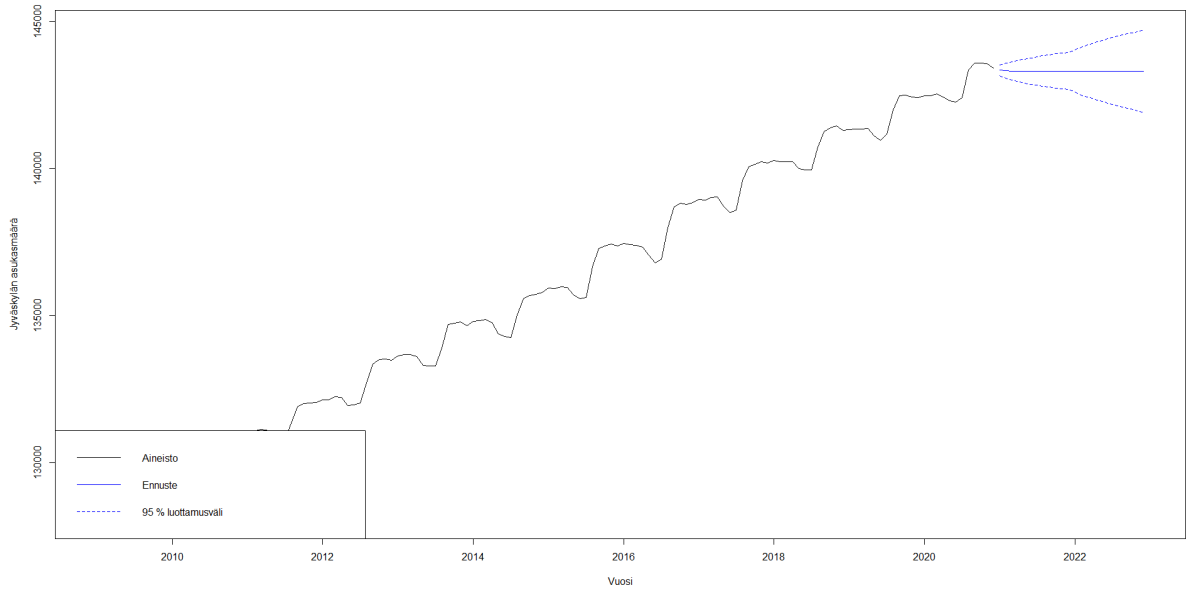


Toisin kuin sarima mallissa arima mallin autokorrelaatiosta ei saatu arvoja lähelle nollaa ja kuuvasata nähdään, että autokorrelaatiota on mallissa edelleen.

Sanoisin, että on järkevää siirtyä ennustukseen sarima mallilla.



Sarima mallilla saatu ennustus. Sanoisin, että ennustus on fiksun näköinen. ennustuksessa näkyy jatkuvana trendi, joka näkyy aiemmassa datassa. Lisäksi siinä näkyy vuoden sisällä näkyvät vaihtelut. Vaikka sanoisin tämän olevan hyvä on kuitenkin viisasta katsoa minkälaisen ennustuksen aikaisempi arima malli antaisi.



Tämä ennustus ei minusta vaikuta järkevältä, itse asiassa se näyttää melkein suoralta viivalta. Voisi melkein sanoa, että tämä ennustus on arvoton.

Jos kaksi vuotta, joita pyritään ennustamaan olisivat normaaleja vuosia niin sanoisin, että ennustus on järkevä. Vaikka data jatkuu vuoden 2020 loppuun, joka oli korona vuosi niin ennustaminen ei välttämättä ole täysin tarkkaa sillä datasta yritetään ennustaa kaksi korona vuotta (2021-2022), jotka eivät ole omasta mielestäni ole normaaleja vuosia ja siksi eivät ole välttämättä verrattavissa normaaleihin vuosiin joita data ennustaa. Voi olla että korona

on voinut vaikuttaa muuttamiseen, sillä etätyöskentelystä on tullut normaalia ja oppilaat eivät välttämättä ole muuttaneet Jyväskylään vaan ovat jääneet kotikuntaansa koronan vuoksi. Metodeista joita käytin niin sanoisin, että omasta mielestä sarima malli, jonka sain tehtyä on hyvä tai ainakin vaikuttaa hyvältä ja ennustus joka sen perusteella tuli vaikutti hyvin samannäköiseltä luin aikaisemmat vuodet. Arima malli taas ei tuottanut minusta järkeviä tuloksia joten voi olla, että jotain meni pieleen sen kertoimien kanssa vaikka kertoimet ovat päätelty datasta. Kokeilin myös muita kertoimia mutta mikään niistä ei tuntunut toimivan, joten päätin mennä noilla kertoimilla, joita “löysin” aineistosta.

Tässä lopuksi vielä harjoitustyön koodi

```
#osa 1
#meitä kiinnostaa vain arvot vuodesta 2009 eteenpäin joten, sitä edeltävät arvot
#voidaan pudottaa pois
#D:/kurssit/TILS619/aikasarjaharjoitustyö/
populaatio <- read.csv2(file = "jkl.csv",header = TRUE)
populaatio$vuosi <- as.numeric(substr(populaatio$Kuukausi, 1, 4))
populaatio$Kuukausi <- rep(1:12,times=length(populaatio$Kuukausi)/12)
populaatio <- subset(populaatio, populaatio$vuosi >= 2009)
#tarkastellaan miltä data näyttää ennen kuin sille tehdään mitään
plot(populaatio$JyvÃ.skylÃ., ylab="populaatio",xlab="kuukausi")
ue <- ts(populaatio$JyvÃ.skylÃ.,start=2009, frequency=12)
plot(ue)
#dataan voidaan lisätä liukuva keskiarvo mutta data on jo aika selvää niin
#se ei tuo hirveästi uutta tietoa
ma_smooth <- function(x, window_size)
{
  win <- rep(1/window_size, window_size) # painot
  stats::filter(x, win, sides=2)
}

ue_smooth <- ma_smooth(ue, window_size=12)
ts.plot(ue, ue_smooth, col=c(1,2))
#nähdään kuinka paljon väestönmuutos on verrattuna liukuvaan keskiarvoon
ue_r <- ue - ue_smooth
plot(ue_r)

m <- month.abb[cycle(ue)]
fit <- lm(ue_r ~ m)
#autokorrelaatio
acf(fit$residuals)



par(mfrow=c(4, 1), mar=c(1,1,1,1))
plot(ue)
plot(ue_smooth)
ue_seasonal <- ts(fit$fitted.values, start=2009, frequency=12); plot(ue_seasonal)
ue_residual <- ts(fit$residuals, start=2009, frequency=12); plot(ue_residual, type='l')
```

```
par(mfrow=c(1, 1),mar=c(1,1,1,1))
#osa 2

d_pop <- diff(ue)
ts.plot(d_pop, xlab = "Vuosi", ylab = "populaatio")

#differenssi on stationaarisen näköinen
#joten voidaan jatkaa differenssin kanssa
#Aloitetaan tarkastelemalla differenssin ja aineiston autokorreaatioita ja osittaisia
#autokorrelaatioita joiden avulla voidaan yrittää päätellä hyviä ar ja ma kertoi-
#mia

#aineiston auokorrelaatiosta on vaikea sanoa mitään
acf(ue)
#osittaisesta taas saa jotain irti
pacf(ue)
#tämän perusteella voitaisiin käyttää ma(1) mallia sillä viiveen yksi jälkeen ei
#ole nollasta eroavia arvoja ja samoin ar(1) sopisi sillä arvot ovat pieniä ei-
#vätkä kasva sen myöhemmin
acf(d_pop)
#ar malli ei oikein sovi sillä autokorrelaatio
#ei laske koko ajan vaan tulee takaisin jaksollisesti
#ma ei myöskään sovi sillä arvot kasvavat merkittäviksi useasti
pacf(d_pop)
#sama kuin edellisessä vaikea sanoa mitään lukuja

#vaikuttaisi siltä, että muuttamattomasta aineistosta saatiin joitain arvoja
#joita voitaisiin mahdollisesti käyttää

#kokeillaan ensin sarima mallia
mod1 <- arima(ue, order = c(1, 1, 1), seasonal = list(order = c(1, 0, 0), period = 12))
plot(mod1$residuals)
acf(mod1$residuals)
#ei näy suurta muutosta joten sarma(1,1,1)(1,0,0)s12 malli näyttäisi sopivalta
#tässä vaiheessa kiinnostaa kokeilla jos pelkkä arima toimisi,
#joten kokeillaan vielä sitä
mod4 <- arima(ue, order = c(1, 1, 1))
plot(mod4$residuals)
pacf(mod4$residuals)
acf(mod4$residuals)
#alkulukemilla ainakin vaikuttaisi, että arima olisi huonompi mutta kokeillaan
#kuitenkin jos tätä saataaisiin parannettua
mod5 <- arima(ue, order = c(6, 1, 2))
plot(mod5$residuals)
acf(mod5$residuals)
#sanoisin edelleen, että sarima malli olisi paras malli tälle datalle

#osa4
```

```
#nyt kun olemme löytäneet sopivan mallin voimme kokeilla ennustaa Jyväskylän
#väestön kasvua seuraavalle 24 kuukaudelle
#mod1 on paras malli joten se sijoitetaan ennustus algoritmiin
pred <- predict(mod1, n.ahead = 2 * 12)
prediction <- ts(pred$pred, start = c(2021, 1), frequency = 12)
upper <- ts(pred$pred + qnorm(0.975) * pred$se, start = c(2021, 1), frequency = 12)
lower <- ts(pred$pred - qnorm(0.975) * pred$se, start = c(2021, 1), frequency = 12)
ts.plot(ue, prediction, lower, upper, col = c("black", rep("blue", 3)),
        lty = c(1, 1, rep(2, 2)),  ylab = "Jyväskylän asukasmäärä", xlab = "Vuosi")
legend("bottomleft", legend = c("Aineisto", "Ennuste", "95 % luottamusväli"), col = c("black",
"blue", "blue"),
        lty = c(1, 1, 2))

#vaikka tulin tulokseen, että sarima oli parempi niin voi olla hyvä kuitenkin
#hyvä tarkistaa millainen ennustus tulisi arima mallilla
pred2 <- predict(mod4, n.ahead = 2 * 12)
prediction <- ts(pred2$pred, start = c(2021, 1), frequency = 12)
upper <- ts(pred2$pred + qnorm(0.975) * pred$se, start = c(2021, 1), frequency = 12)
lower <- ts(pred2$pred - qnorm(0.975) * pred$se, start = c(2021, 1), frequency = 12)
ts.plot(ue, prediction, lower, upper, col = c("black", rep("blue", 3)),
        lty = c(1, 1, rep(2, 2)),  ylab = "Jyväskylän asukasmäärä", xlab = "Vuosi")
legend("bottomleft", legend = c("Aineisto", "Ennuste", "95 % luottamusväli"), col = c("black",
"blue", "blue"),
        lty = c(1, 1, 2))
```