Aalto University
School of Science
Degree programme in Engineering Physics and Mathematics

On ARIMAX modelling of frauds and household wealth

Bachelor's thesis
11.8.2018

Otto Saikkonen

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author Otto Saikkonen

Title of thesis On ARIMAX modelling of frauds and household wealth

Degree programme Engineering Physics and Mathematics

Major Mathematics and System Sciences **Code of major** SCI3029

Supervisor Prof. Pauliina Ilmonen (Ph.D.)

Thesis advisor(s) Lauri Viitasaari (D.Sc. (Tech.))

Date 11.08.2018 **Number of pages** 20+8 **Language** English

Abstract

Almost half of the 7200 global organizations interviewed by PwC reported that they had been victim of a fraud. Due to advancing technology, financial crimes are becoming harder to notice. To address the need for financial crime models, we build time series models, which in future use would be able to predict future fraud and payment fraud rates in Finland. In addition to the time series of the crimes, we use fraction of Finnish households who are getting into debt and the fraction of the households that are using their savings as exogenous variables to provide additional information to our models. With the Box-Jenkins method, we are producing six different time series models - one for each time series and two models with exogenous variables. Diagnostics will tell us if the models are considered as good models and the results will show the impact that the financial situation of Finnish households make to the models.

Keywords Financial crime, fraud, household wealth, time series, stochastic models, ARIMAX

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi

Tekniikan kandidaatintyön tiivistelmä

Tekijä Otto Saikkonen		
Työn nimi Petosten ja kotitalouksien taloudellisen tilanteen ARIMAX-mallintaminen		
Koulutusohjelma Teknillinen fysiikka ja matematiikka		
Pääaine Matematiikka ja systeemitieteet		**Pääaineen koodi** SCI3029
Vastuuopettaja Prof. Pauliina Ilmonen (Ph.D.)		
Työn ohjaaja(t) Lauri Viitasaari (TkT)		
Päivämäärä 14.08.2018	**Sivumäärä** 20+8	**Kieli** Englanti

Tiivistelmä

Työntekijä ostaa itselleen tuotteita yrityksen luottokortilla. Rakennuttaja väärentää kirjanpitoa, jolloin yrityksen kustannukset nousevat. Kassatyöntekijä ohjaa yrityksen rahavirtoja väärennetyille tileille.

Lähes puolet PwC:n haastattelemasta 7200 kansainvälisestä yrityksestä raportoi joutuneensa petoksen uhriksi. Toisen puolikkaan on selvitettävä, ovatko he välttyneet petoksilta vai yksinkertaisesta eivät vain tiedä organisaatiossa tapahtuvista petoksista. Kaikkia petoksia ei voida estää, mutta niiden riskiä voidaan vähentää tehokkaalla ja toimivalla sisäisellä valvonnalla. Kuitenkin, kehittyvän teknologian takia talousrikoksia on entistä vaikeampi huomata, jolloin matemaattisten talousrikosmallien tarve kasvaa entisestään. Hyvät matemaattiset mallit pystyvät käsittelemään huomattavasti laajempia aineistoja kuin ihmiset ja selvittämään monimutkaisia riippuvuussuhteita, jotka jäisivät ihmisiltä muuten huomiotta.

Tässä työssä rakennamme aikasarjamalleja, jotka mallintavat petosten ja maksuvälinepetosten määriä Suomessa. Aikasarjamallit sopivat työhön hyvin, koska käsiteltävä aineisto koostuu aikasarjoista, eli havainnoista, jotka ovat liitetty aina tiettyyn päivämäärään. Aikasarjamallien avulla saamme selville petosten riippuvuudet eri aikajaksoilta. Rikosten lukumäärien lisäksi käytämme kotitalouksien taloudellista tilannetta tuomaan lisää informaatiota malleihin. Taloudellisesti huonossa asemassa olevat kotitaloudet saattavat helpommin joutua esimerkiksi internetissä tehtävien petoksien uhreiksi, missä ihmisiltä usein pyydetään sijoituksia lupaamalla suuria tuottoja tulevaisuudessa. On myös mahdollista, että samat kotitaloudet saattavat tehdä enemmän petoksia päästäkseen pois huonosta taloudellisesta asemastaan.

Luomme kuusi aikasarjamallia Box-Jenkins-menetelmällä – yksi jokaista neljää aikasarjaa kohden ja kaksi ulkoisilla muuttujilla. Box-Jenkins-menetelmä koostuu kolmesta vaiheesta. Näistä ensimmäisessä vaiheessa määritellään mallin peruspiirteet, toisessa vaiheessa mallin parametrejä sovitetan sopivaksi ja viimeisessä vaiheessa tarkastellaan mallin jäännöstermejä. Mallin jäännöstermien tarkastelulla selvitetään, kuinka hyvä malli on, ja niiden avulla voimme myös päätellä, toiko kotitalouksien taloudellinen tilanne malleihin lisää informaatiota. Jos mallit eivät ole tarpeeksi hyviä, toistetaan Box-Jenkins-menetelmän kolme vaihetta uudestaan, kunnes tarpeeksi hyvä malli on luotu.

Avainsanat Talousrikokset, petos, kotitaloudet, taloudellinen tilanne, Suomi, aikasarja, ARIMAX

Contents

1 Introduction

In PwC's 2018 Global Economic Crime and Fraud Survey, 49% of 7200 global organizations had been a victim of a fraud [18]. It is unlikely that the other half have had complete success in preventing frauds, since the same report points out that too few companies are fully aware of the fraud risk that they face. According to EY's 15th Global Fraud Survey, "the transformation of business models due to the rapid evolution of digital technology is making the landscape of fraud, bribery and corruption risk even more complex." [5] More complex risks may need more complex solutions to forecast and prevent financial crimes, such as frauds. For example, machine learning algorithms have been introduced succesfully to enhance anti money laundering detection [11]. It is also worth to mention that financial crimes seldom occur at random, i.e., there is always an external causality linked to it. Therefore, it is important to consider more variables than the crimes themselves when analysing financial crimes.

A time series is a set of naturally ordered and equally spaced data points, that describes a quantitative phenomenon. Time series analysis comprehends methods to understand time series data. While time series analysis have wide range of applications, one of the main goals is to forecast future values based on previously observed values [10]. Various fields of research, such as business, engineering, environometrics, economics, medicine, politics and social sciences, have lots of possible applications for time series analysis.

In this thesis we will analyse the time series of frauds committed in Finland and build an *Autoregressive Integrated Moving Average with Explanatory Variable* model, that in further use, would be able to forecast future fraud rates. A time series analysis will not only focus on particular numbers of frauds, but also in the order they appear [13] to unveil complex structures and dependencies the observations have.

Individual's wealth may affect his or hers behaviour when it comes to rationalizing shady decision for personal gain - whether it was becoming a victim of internet scam or committing a payroll fraud. Thus, the status of Finnish households' wealth is added to the model as an exogenous variable, in order to examine if distribution of wealth had an impact on frauds.

2 On financial crimes and household wealth

2.1 Financial crimes

The two financial crimes to be studied in this thesis are 1) fraud & petty fraud and 2) payment frauds, petty payment frauds & preparation of payment frauds. For simplicity, from this point on the concept of fraud will include both fraud and petty fraud and the concept of payment fraud will include payment frauds, petty payment frauds and preparation of payment frauds.

The data is gathered from Statistics Finland PX-Web-database [14]. Both reports of offences and already convicted felonies are included in the data set and there are no distinctions made between these two attributes. It is important to keep in mind that, naturally, neither all reported offences or convicted crimes happen within the same month that they are reported to or by the police. In this thesis, the time difference is assumed to be constant. In Section 5, we will discuss the effect this has on the model.

Fraud and payment fraud are chosen to be studied because they represent the majority of financial crimes committed in Finland from 01/2009 to 09/2017 (Figure 1). In addition, the amount of both crimes has been increasing in Finland which makes them an interesting subject to study.

Both time series show a linear trend (Figure 2). Moreover, the time series of payment frauds had a spike from mid-2015 to late-2016.

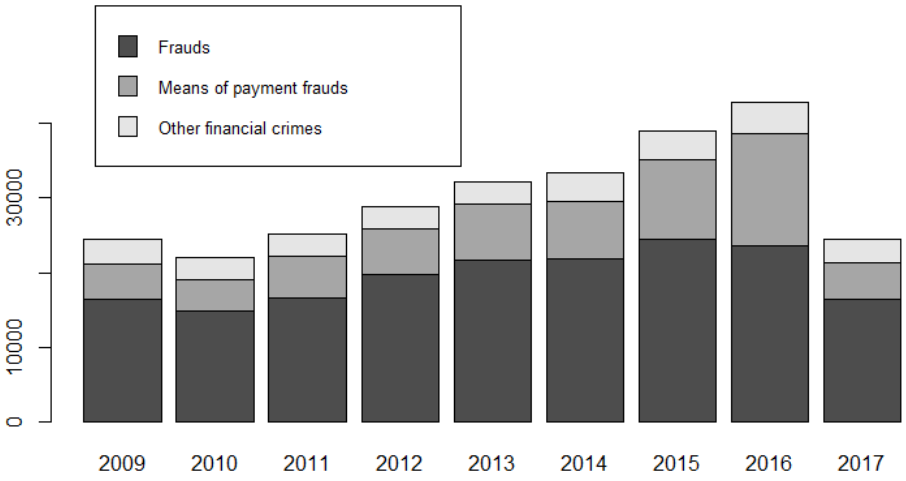

Figure 1: Count of frauds, means of payment frauds and other financial crimes in Finland from 01/2009 to 09/2017.

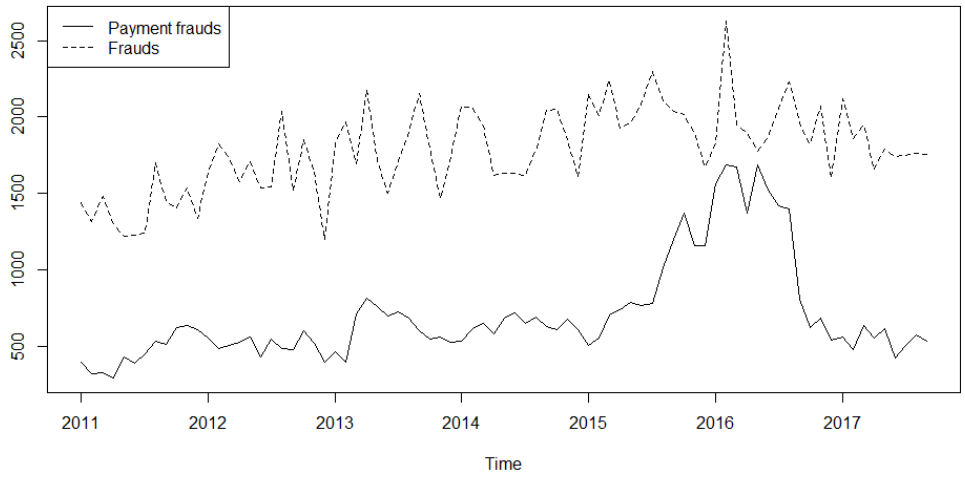

Figure 2: Time series of frauds and payment frauds from 01/2011 to 09/2017.

2.2 Households' wealth

Households' wealth can be determined with respect to many objective variables. We will take a more subjective approach of household wealth by quan-

tifying it according to household's own opinion. The Finnish consumer survey asks individuals about their households financial situation on the time of the survey. The Statistics of Finland reports monthly the fraction of households falling into these categories [15]. The answer options are following:

- Can save a lot
- Can save a little
- Can barely make ends meet
- Have to use savings
- Getting into debt
- Don't know

Analysing suspicious change in the behaviour of people is one method for detecting fraud [6]. Bad financial situation may force individuals to commit financial crimes, such as fraud or means of payment frauds, in order to maintain current living standards. It is also possible that people in strained circumstances are easier targets to scams, which often offer quick financial gains for the capital invested. Thus, the two chosen time series to be studied are the fraction of households who have to use savings and the ones who are getting into debt.

From 01/2011 to 09/2017, the fractions of the savings using & debt acquiring households range from $2,6\%$ to $6,8\%$ and from $1,1\%$ to $3,8\%$, respectively (Figure 3). Both categories represent minority in the survey. However, due to the nature of those categories, fluctuations in chosen categories may provide information to frauds.

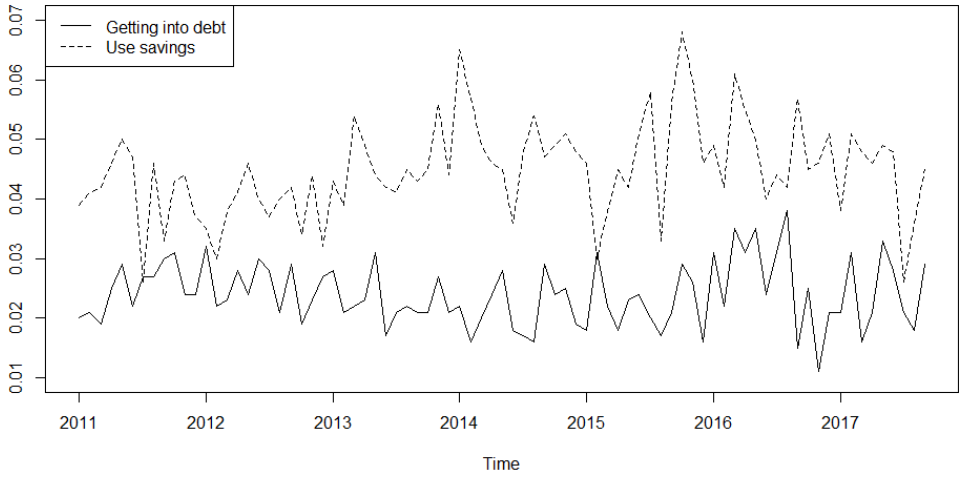

Figure 3: Time series of fractions of the households who are getting into debt and using their savings from 01/2011 to 09/2017.

3 Stationary time series models

3.1 Stationary time series

In time series analysis, stationarity of the time series is important since it implies that statistical properties of the time series remain the same over time [16, Chapter 2.2] and thus, all observations are comparable with each other.

In general, there are two definitions of stationarity: strict stationarity and weak stationarity. A time series x_t is called strictly stationary if and only if the distribution of $(x_{t_1}, ..., x_{t_n})$ and $(x_{t_1+h}, ..., x_{t_n+h})$ is the same for all sets of indices $\{t_1, ..., t_n\}$ and for all integers h [16, Chapter 2.2].

The time series x_t is called weakly stationary if

- $E[x_t]$ is constant,
- $Var(x_t)$ is constant and finite,
- $\exists\ \gamma_k = Cov(x_t, x_{t+|k|})$ for any $t,\ k$.

3.2 Stochastic time series models

In many cases, the values of time series x_t can be generated from series of random variables ϵ_t, ϵ_{t-1}, ϵ_{t-2}..., which are random drawings from a fixed distribution, usually assumed normal and having mean 0 and variance of σ^2. The so called white noise process is transformed to the process x_t by a *linear filter*, which is the weighted sum of past white noise [1, Chapter 1.2.1]:

$$x_t = \epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + ... \tag{1}$$

where $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

From Equation (1) we can derive that, under suitable conditions, x_t can be represented as the weighted sum of past values of x_t plus an instance of white noise:

$$x_t = \epsilon_t + \pi_1 x_{t-1} + \pi_2 x_{t-2} + ... \tag{2}$$

In both cases, the representations of x_t have infinite number of parameters ψ and π. This is not practical, which is why two models are introduced: *moving average (MA) model* and *autoregressive (AR) model*. An MA(q) process is a special case of process defined by Equation (1), where only the first q parameters are non-zero. The model is defined as

$$x_t = \epsilon_t - \sum_{j=1}^{q} \theta_j \epsilon_{t-j}. \tag{3}$$

An AR(p) process is a special case of process defined by Equation (2), where only the first p parameters are non-zero. The model is defined as

$$x_t = \epsilon_t + \sum_{j=1}^{p} \phi_j x_{t-j}. \tag{4}$$

A more common way to represent AR- and MA-models is through *transfer functions*

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_q B^q \tag{5}$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_p B^p, \tag{6}$$

where $Bx_t = x_{t-1}$, $B^2 x_t = x_{t-2}$ and so on.

This allows us to represent AR-model as

$$\phi(B)x_t = \epsilon_t \tag{7}$$

and MA-model as

$$x_t = \theta(B)\epsilon_t \tag{8}$$

Unlike Equation (1) and (2), finite MA(q) process cannot be represented as finite AR(p) process and correspondingly, finite AR(p) process cannot be represented as finite MA(q) process [1, Chapter 3.1]. Sometimes it is necessary to include both MA and AR models to an *autoregressive moving average (ARMA)* model. An ARMA(p,q) process is defined as

$$\phi(B)x_t = \theta(B)\epsilon_t. \tag{9}$$

It is desired that the time series are stationary for MA, AR and ARMA models (see Subsection 3.1). If the time series is not stationary, e.g., shows trend, it is likely that one cannot obtain satisfying results with these three models. Differencing may be needed, since the dth difference of the time series can be stationary. In such case, an *autoregressive Integrated Moving Average (ARIMA)* model is needed. An ARIMA(p,d,q) process is defined as

$$\theta(B)\nabla^d x_t = \phi(B)\epsilon_t, \tag{10}$$

where $\nabla = 1 - B$ is the differencing operator [1, Chapter 4.1].

In cases where percentage changes of time series show non-stationary stability, logarithm operations may be needed to obtain stationary time series, since

$$\nabla \log(x_t) = \log(x_t) - \log(x_{t-1}) = \log\left(\frac{x_t}{x_{t-1}}\right) \approx \frac{x_t - x_{t-1}}{x_{t-1}}. \tag{11}$$

3.2.1 Autocorrelation and Ljung-Box test

The stationarity of time series x_t assumes that covariance between observations x_t and x_{t+k} must be the same for any t. This covariance is called *autocovarince* with lag k and is defined as

$$\gamma_k = Cov(x_t, x_{t+k}). \tag{12}$$

Furthermore, the *autocorrelation* is standardized measure of the dependence of two observations x_t and x_{t+k}, corresponding to the autocovariance function divided by the variance of the process [16, Chapter 2.9]:

$$\rho_k = \frac{\gamma_k}{\gamma_0}. \tag{13}$$

Autocorrelation function (ACF) can indicate which ARMA model fits best for the chosen time series, since different ARMA models show different behaviour of autocorrelation function. For AR models ACF decreases exponentially as the lag grows, whereas for MA models, cuts out when the lag grows. ARMA models have a mixture of both, depending on the values of parameters p and q.

In addition to the autocorrelation function, partial autocorrelation function shows information of the dependency structure of a time series. The partial autocorrelation function α_k at lag k can be seen as the correlation between x_t and x_{t-k}, adjusted with the observations in between the two variables. Moreover, it is the correlation of the two residuals obtained after regressing x_t and x_{t-k} on the intermediate observations [2, Chapter 3.4]. Partial autocorrelation function is defined by

$$\alpha_k = Corr(x_t, x_{t-k}|x_{t-1}, ..., x_{t-k+1}). \tag{14}$$

Ljung-Box test can be used to test if the autocorrelations are significant. The test is defined as

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\rho_k^2}{n-k}, \tag{15}$$

where n is the sample size, ρ_k is the sample autocorrelation at lag k and h is the number of lags being tested [12]. The null hypothesis is that data points in the time series are independently distributed, i.e. the value Q satisfies the inequality

$$Q > \chi^2_{1-\alpha,h}, \tag{16}$$

where $\chi^2_{1-\alpha,h}$ is the α quantile of the χ^2-distribution with h degrees of freedom [4]. Ljung-Box test tests for white noise. Thus χ^2-distribution is used, since it is the distribution of a sum of squares of independent standard normal random variables.

Ljung-Box test is commonly used in ARMA modeling. The test is applied to the residuals of a fitted ARMA model, and therefore the null hypothesis is that the residuals from the ARMA model have no autocorrelation. If the correlations of the residuals are zero for significance level α, the model can be seen as valid.

3.2.2 Augmented Dickey-Fuller test

The Augmented Dickey-Fuller test (ADF) is used to test for the presence of a unit root in the time series sample. The Augmented Dickey-Fuller test

incorporates three types of linear regression models: with no drift or linear trend (17), with drift but no linear trend (18) and with both drift and trend (19) [3, Chapter 4.4]:

$$Dx_t = \alpha x_{t-1} + \sum_{i=1}^{k} \beta_i Dx_{t-i} + \epsilon_t \tag{17}$$

$$Dx_t = \alpha x_{t-1} + \sum_{i=1}^{k} \beta_i Dx_{t-i} + \epsilon_t + \mu \tag{18}$$

$$Dx_t = \alpha x_{t-1} + \sum_{i=1}^{k} \beta_i Dx_{t-i} + \epsilon_t + \mu + \delta_t \tag{19}$$

where D is an operator of first order difference, ϵ_t is an error term, μ is a drift term and δ_t is a linear trend term.

The Augmented Dickey-Fuller test statistic is defined as

$$ADF = \frac{\hat{\alpha}}{S.E(\hat{\alpha})}, \tag{20}$$

where $\hat{\alpha}$ is an generalized least squares estimate for linear regression's α and S.E(α) its standard error [8].

The null hypothesis in the test is that $\alpha = 0$, i.e. x_{t-1} does not provide any information to the change in x_t besides the information in the past terms. Under the null hypothesis, we conclude that there is an unit root presence and thus, we have a non-stationary times series. Therefore, in order to have a stationary time series, the p-value has to be less than the chosen significance level. The p-value is calculated by interpolating the test statistics from the corresponding critical values tables [7]. The null hypothesis of a unit root is valid under a very general set of assumptions that goes far beyond the linear AR(∞) process assumption typically imposed [17].

4 Modelling of frauds

In order to find the models that fit best to the four time series, we used the Box-Jenkins method [1]. The method is an iterative approach to build ARIMA models, which is based on *identification, estimation* and *diagnostics.*

The first step (*identification*) was to assess whether time series is stationary, and if not, obtain stationary time series through various operation (See Section 3.2). After that, we identified the parameters q and p of an ARMA model. In the second step (*estimation*), we estimated values for the parameters of the transfer functions. This was done by using the `forecast` package in R. Finally, we did *diagnostics* to how good the obtained models were.

For clarity, the notation of the four time series were

$X_{1,t}$ = Time series of fraud,
$X_{2,t}$ = Time series of payment fraud,
$X_{3,t}$ = Time series of the fraction of households who have to use savings,
$X_{4,t}$ = Time series of the fraction of households who are getting into debt.

We built one ARIMA model for each time series and one ARIMAX model for both $X_{1,t}$ and $X_{2,t}$, where the exogenous variables were $X_{3,t}$ and $X_{4,t}$. The ARIMA models for $X_{1,t}$ and $X_{2,t}$ gave us a baseline, to which the ARIMAX models were compared.

4.1 Identification of ARMA models

As discussed in Section 3.1, stationarizing the time series was crucial, since it made the observations comparable with each other due to constrains stationarity enforces on them.

$X_{1,t}$, $X_{3,t}$ and $X_{4,t}$ (Figures 2 and 3) showed a linear trend, indicating that the time series were not stationary. First order difference represented the time series to *look* stationary, i.e., the time series had constant means & finite and constant variances (Figures 15, 17 and 18). An ADF test was ran on both the original and once differenced time series to see if the time series were stationary and how they compared with the original time series. The test results (Tables 4, 5 and 6) showed that, for all three time series, the original time series were not stationary, but once differenced were. Therefore, the time series to be used in the models were

$$w_{1,t} = \nabla X_{1,t}, \tag{21}$$
$$w_{3,t} = \nabla X_{3,t}, \tag{22}$$
$$w_{4,t} = \nabla X_{4,t}. \tag{23}$$

$X_{2,t}$ was a more challenging time series to stationarize, since number of payment frauds committed in Finland had a huge spike during mid-2015 to late-2016 , where number of crimes were twice as much compared to other examined time periods (Figure 2). A first order difference was not simply enough to smooth the spike, but second order difference represented the time series stationary (Figure 4). For the original time series, the p-values from ADF test ranged from 0.30 to 0.71, while for the second order differenced time series, all p-values were 0.01 (Table 1), i.e., the twice differenced $X_{2,t}$ was stationary.

We also tried logarithmic transformations on the time series, but concluded that the second order difference performed better. Thus, the process to be used in the model was

$$w_{2,t} = \nabla^2 X_{2,t}. \tag{24}$$

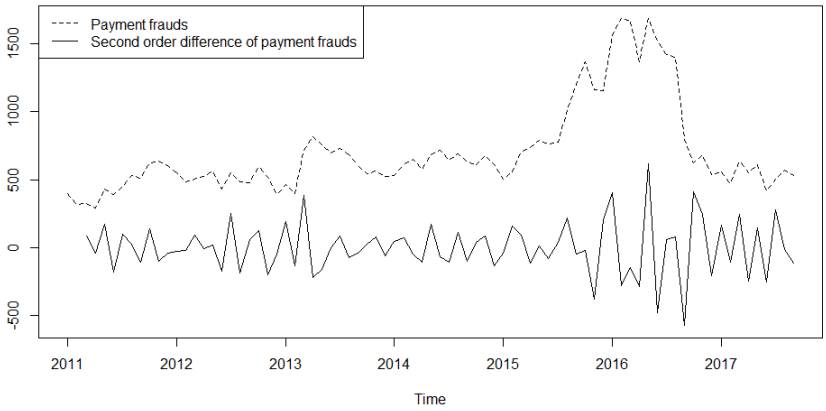

Figure 4: $X_{2,t}$ and $\nabla^2 X_{2,t}$.

Table 1: ADF-test results of original and stationarized time series of means of payment frauds.

	Original			Stationarized		
	Lag	ADF	P-value	Lag	ADF	P-value
No drift no trend	0	-0.70	0.43	0	-10.28	0.01
	1	-0.64	0.45	1	-6.46	0.01
	2	-0.61	0.46	2	-5.26	0.01
	3	-0.64	0.45	3	-4.36	0.01
With drift no trend	0	-1.89	0.37	0	-10.23	0.01
	1	-1.95	0.35	1	-6.43	0.01
	2	-1.90	0.37	2	-5.24	0.01
	3	-2.08	0.30	3	-4.33	0.01
With drift and trend	0	-1.76	0.67	0	-10.31	0.01
	1	-1.75	0.68	1	-6.55	0.01
	2	-1.67	0.71	2	-5.44	0.01
	3	-1.88	0.62	3	-4.41	0.01

With stationary processes, next step was to analyse autocorrelation and partial autocorrelation functions for clues about the orders of p and q for our ARMA models (Subsection 3.2.1). For $w_{2,t}$, the autocorrelation and partial autocorrelation functions indicated MA(1) model due to a spike at lag = 1 in autocorrelation function and exponential decrease in partial autocorrelation function (Figure 5). Since $w_{2,t} = \nabla^2 X_{2,t}$ (Equation 24), the final ARIMA model for payment frauds is IMA(2,1).

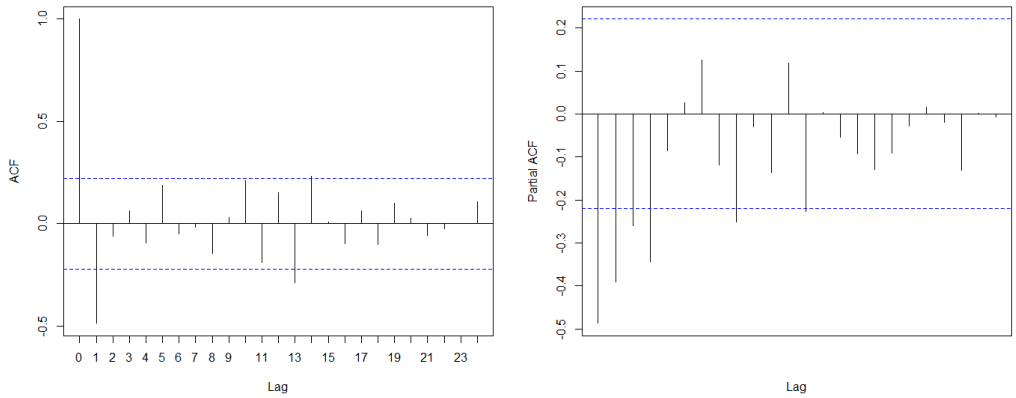

Figure 5: Autocorrelation and partial autocorrelation functions of $\nabla^2 X_{2,t}$.

For the three remaining time series, p and q were obtained by analysing autocorrelation functions and partial autocorrelation functions. The iterative Box-Jenkins method finally gave us sufficient integers for p and q, which can be found in Table 2.

Table 2: ARIMA models for the time series

Time series	ARIMA model
$X_{1,t}$	ARIMA(2,1,2)
$X_{2,t}$	IMA(2,1)
$X_{3,t}$	ARIMA(1,1,1)
$X_{4,t}$	ARIMA(1,1,2)

4.2 Estimation of the parameters

The model parameters (Table 3) were estimated with `Arima` function, which was included in `forecast` package. With the estimated parameters, we were able to construct our ARIMA models:

$$X_{t,1} = X_{t-1,1} + 0,7490(X_{t-1,1} - X_{t-2,1}) - 0,4314(X_{t-2,1} - X_{t-3,1}) + 1.5086\ \epsilon_{t-1} - 0.7434\ \epsilon_{t-2} + \epsilon_t$$
$$X_{t,2} = 2X_{t-1,2} - X_{t-2,2} + \epsilon_{t-1} + \epsilon_t$$
$$X_{t,3} = X_{t-1,3} + 0,2421(X_{t-1,3} - X_{t-2,3}) + 0,999\ \epsilon_{t-1} + \epsilon_t$$
$$X_{t,4} = X_{t-1,4} + 0,8285(X_{t-1,4} - X_{t-2,4}) + 1,7458\ \epsilon_{t-1} - 0,7458\ \epsilon_{t-1} + \epsilon_t$$

Table 3: Model parameters.

	$X_{1,t}$	$X_{1,t}$	$X_{2,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$
Model	ARIMA(2,1,2)	ARIMAX(2,1,2)	IMA(2,1)	IMAX(1,2)	ARIMA(1,1,1)	ARIMA(1,1,2)
ϕ_1	0,7490	0,7145	-	-	0,2421	0,8285
ϕ_2	-0,4314	-0,4833	-	-	-	-
θ_1	-1.5086	-1,4224	-1,000	-1,000	-0,999	-1,7458
θ_2	0.7434	0,6731	-	-	-	0,7458
β_1	-	1359,34	-	-1090,61	-	-
β_2	-	-8010,72	-	-4972,94	-	-

It was hard to interpret the parameters due to complexity of the models. It seemed that, on all models, the difference between lagged values at *t-1* and *t-2* was proportional to the value at t. For the ARIMAX models, exogenous variable Z_t with coefficient vector β was added to the existing ARIMA models to see if it made the model better. Unlike in regression, the value of β is *not* the effect on X_t when Z_t is increased by one due to the presence of lagged values, and β can be only interpreted as conditional on the value of previous values of the response variable, which is hardly intuitive [9]:

$$\begin{aligned} \phi(B)X_t &= \beta Z_t + \theta(B)\epsilon_t \\ \Leftrightarrow X_t &= \frac{\beta}{\phi(B)} Z_t + \theta(B)\epsilon_t. \end{aligned}$$

At this point, all fitted models were plotted against the stationary time series to approximate the goodness of the models and to see if the exogenous variable Z_t made a significant difference (Figures 15 and 16). For $X_{t,2}$, both IMA and IMAX models seemed to model payment frauds well and there were no clear distinctions to be made (Figures 6 and 7). The same applied to ARIMA and ARIMAX models of frauds (Figures 15 and 16). ARIMA models of the status of the household wealth had some trouble to model the spikes of the time series, but overall, they performed well (Figures 17 and 18).

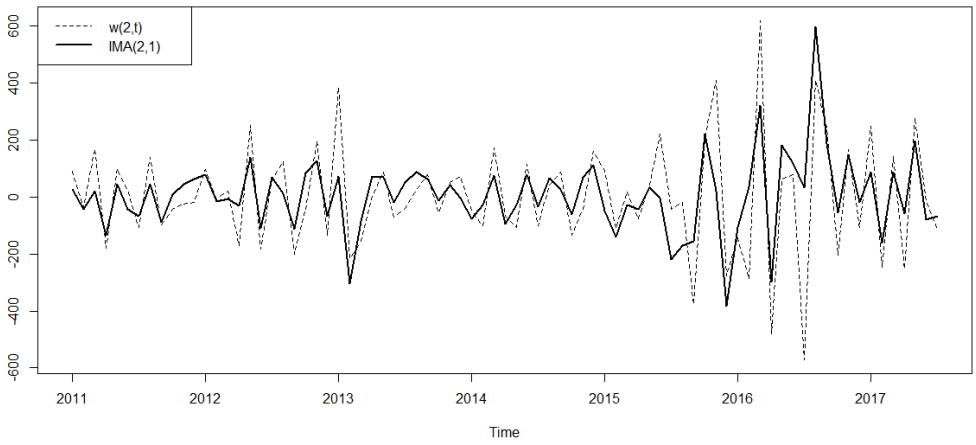

Figure 6: $w_{2,t}$ and the IMA(2,1)-model.

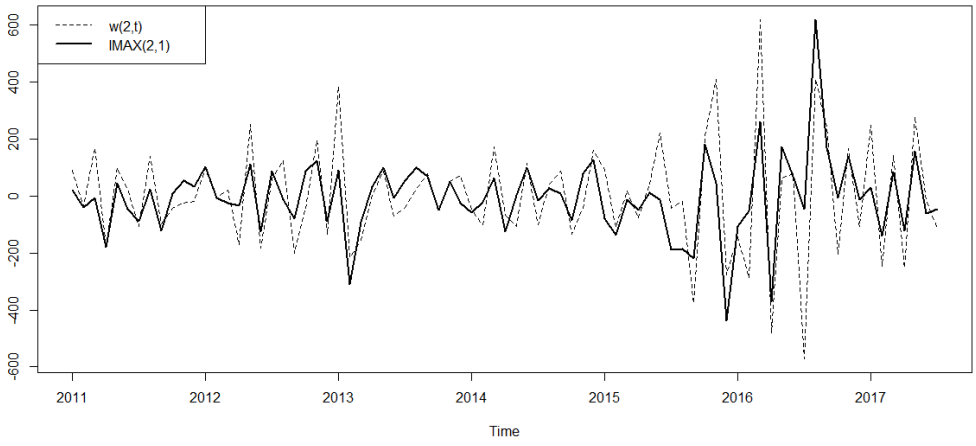

Figure 7: $w_{2,t}$ and the IMAX(2,1)-model.

4.3 Diagnostics

The last step of Box-Jenkins methods was to check if residuals were normally distributed with zero mean and had a zero correlation. All residuals looked more or less normally distributed, except for ARIMA(1,1,2) model of $X_{t,4}$ (Figure 8, 19 and 20). However, the p-values of Ljung-Box test for the

ARIMA(1,1,2) model were all above the significance level (5%), and therefore we kept our null hypothesis that the correlation of model residuals was zero. The null hypothesis of Ljung-Box test, which was that residuals had zero correlation, was also kept with the other models, as almost every p-value was above the chosen significance level (Figures 9, 21 and 22).

Compared to the basic ARIMA and IMA models, p-values of Ljung-Box test were generally higher with ARIMAX and IMAX models. This indicates that the exogenous variable made the models better, since we could more certainly say that the residuals had a zero correlation. The models with exogenous variable also had more normally distributed residuals (Figure 8 and 19). Therefore, it is safe to say that the subjective information about household wealth made the models better, in terms of diagnostics.

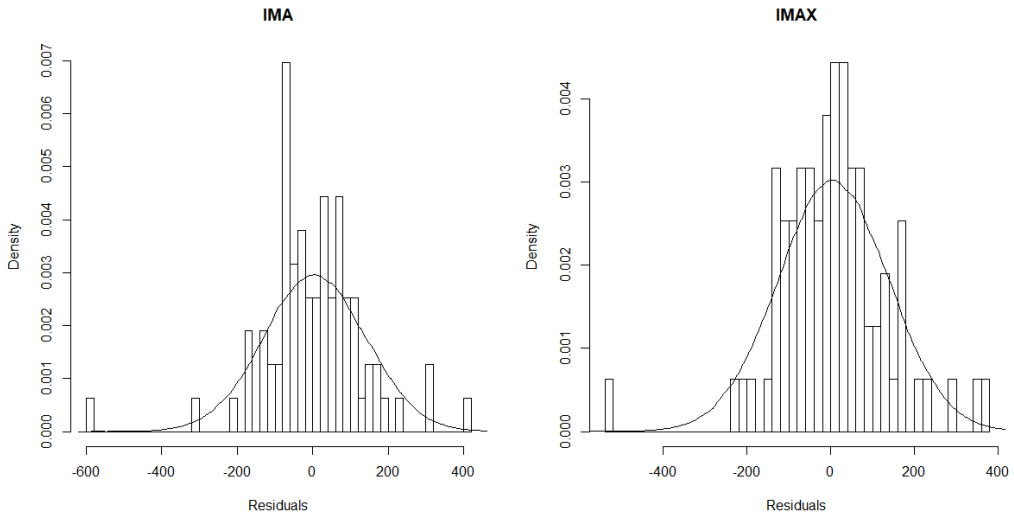

Figure 8: Histogram of density of residuals of IMA(2,1) process and IMAX(2,1) process.

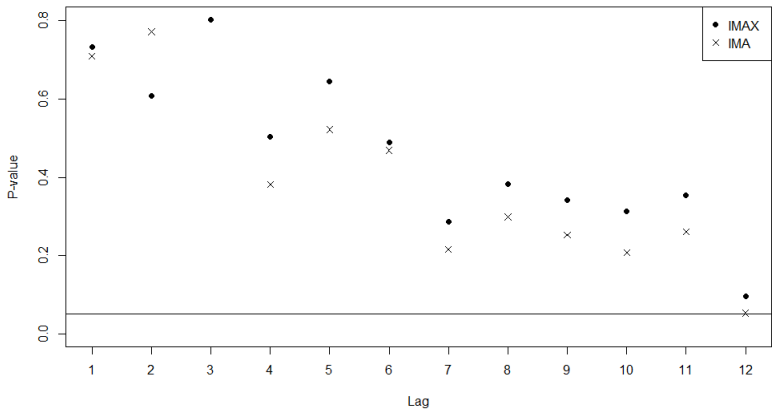

Figure 9: P-values from Ljung-Box test from the IMA and the IMAX process.

5 Conclusions

With an ARIMAX model, relationship between frauds and household wealth can be examined and quantity of future frauds can be forecasted with the historical data. Both the ARIMAX and ARIMA models we built in this thesis performed well and they were able to model the quantity of frauds and payment frauds. The diagnostic tests showed that the residuals of the models were white noise and had zero correlation, which is in-line with the assumption that residuals are homoscedastic over time. However, the status of households' wealth did not make the model significantly better. This does not conclude that there would not be any other predictors to the time series of frauds and payment frauds, but it shows that households' subjective opinion of their wealth is not one of them.

One fundamental problem with crimes and time series analysis is the time difference between when crimes are actually committed and when they are reported to the authorities. Assuming a constant time difference, one can interpret the impact the household wealth has on frauds and payment frauds, while being aware that the results may not reflect the real world situation. For further research, the distributions of previously mentioned time difference should be examined.

This paper focused on building the stochastic models, but for future research, *ex post* testing, i.e,. testing forecasting performance with historical data, is needed for all ARIMA and ARIMAX models to examine how well the mod-

els would predict future crime rates. No clear diagnostic distinction of the ARIMAX and the ARIMA models could be made, but that does not imply equal forecasting performance. Furthermore, the forecasting performance of the ARIMA models play an important role, since in order to forecast with the ARIMAX model, we will first need forecasts of the exogenous variables representing the fraction of households who are using savings and the ones who are getting into debt.

To improve the model, one should introduce more exogenous variables. The fluctuations of all six categories in the Finnish consumer survey may provide information to the models. In addition to the consumer survey, more objective exogenous variables, such as interest rates, strength of police forces and employment rate, should be examined and added into the models if they seem to influence crimes studied. That being said, the current models still performed very well and it would be interesting to see if more objective variables would made the models even better.

References

[1] G. Box & G. Jenkins & G. Reinsel,
Time series analysis: forecasting and control,
John Wiley & Sons 4th (2008).

[2] P. Brockwell & R. Davis, *Time Series: Theory and Methods*,
Springer New York (1991).

[3] S. Chatterjee & N. Singh & D. Goyal & N. Gupta,
Managing in Recovering Markets,
Springer (2015).

[4] R. Di Lorenzo,
Trading systems: theory and immediate practice,
Springer (2013).

[5] EY,
Integrity in the spotlight,
15th Global Fraud Survey.

[6] T. Fawcett & F. Provost,
Adaptive fraud detection,
Data mining and knowledge Discovery, 1997, Volume 1, Issue 3, Pages 291–316.

[7] W. Fuller,
Introduction to Statistical Time Series,
John Wiley & Sons (1976).

[8] R. Harris,
Testing for unit roots using the augmented Dickey-Fuller test,
Economics Letters, 1992, Volume 38, Issue 4, Pages 381-386.

[9] R. Hyndman,
The ARIMAX model muddle,
https://robjhyndman.com/hyndsight/arimax/,
(accessed 3.7.2018).

[10] M. Ivanovic & V. Kurbalija,
Time series analysis and possible applications,
Published in: Information and Communication Technology, Electronics and Microelectronics (MIPRO),
2016 39th International Convention (30.6.2016), Croatia.

[11] R. Kanth,
Enhanced AML fraud detection solutions with Azure Machine Learning,
https://www.youtube.com/watch?v=KdVwmph7HDk&feature=youtu.be
(accessed 25.7.2018).

[12] G. Ljung & G. Box,
On a Measure of a Lack of Fit in Time Series Models,
Biometrika, 1978, Volume 65, Issue 2, Pages 297-303.

[13] C. Ostrom,
Time Series Analysis: Regression Techniques ,
SAGE Publications 2nd Edition (1990).

[14] Official Statistics of Finland (OSF),
Monthly crimes in Finland 2009-2017,
http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__oik__rpk/statfin_rpk_pxt_001.px/?rxid=28176528-bf30-4cf0-8980-0f01e7b7d3b7,
(accessed 21.2.2018).

[15] Official Statistics of Finland (OSF),
Consumer survey [e-publication],
ISSN=1799-1382
http://www.stat.fi/til/kbar/kas_en.html,
(accessed 21.2.2018).

[16] W. Palma,

Time Series Analysis,
John Wiley & Sons (2016).

[17] E. Paparoditis & D. Politis,
The Asymptotic Size and Power of the Augmented Dickey-Fuller Test for a Unit Root,
Econometric Reviews, 2018, Volume 37, Issue 9, Pages 955-973.

[18] PwC,
Pulling fraud out of the shadows,
Global Economic Crime and Fraud Survey 2018.

A Figures

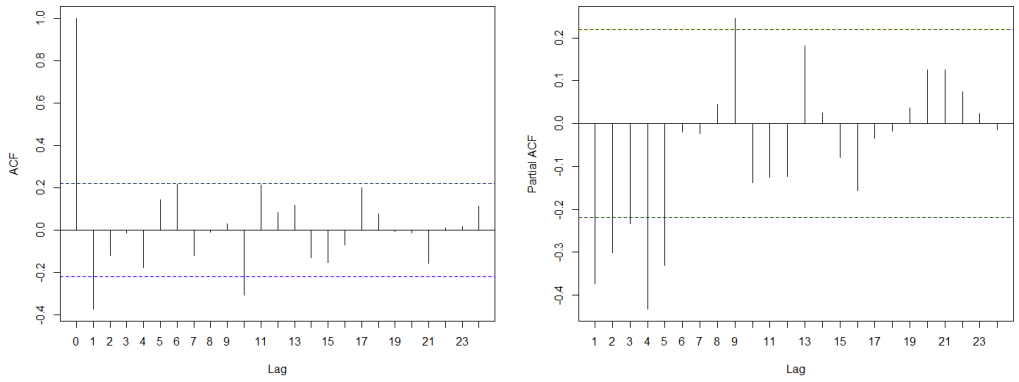

Figure 10: Autocorrelation and partial autocorrelation functions of $\nabla X_{1,t}$.

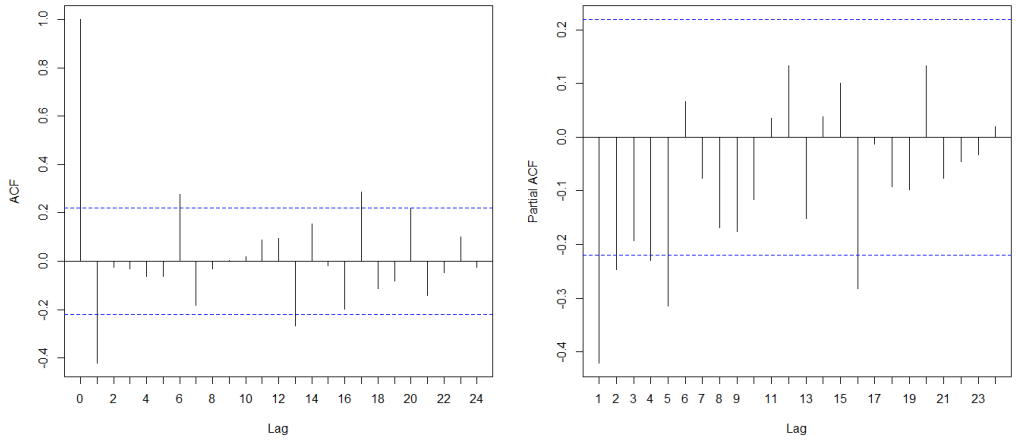

Figure 11: Autocorrelation and partial autocorrelation functions of $\nabla X_{3,t}$.

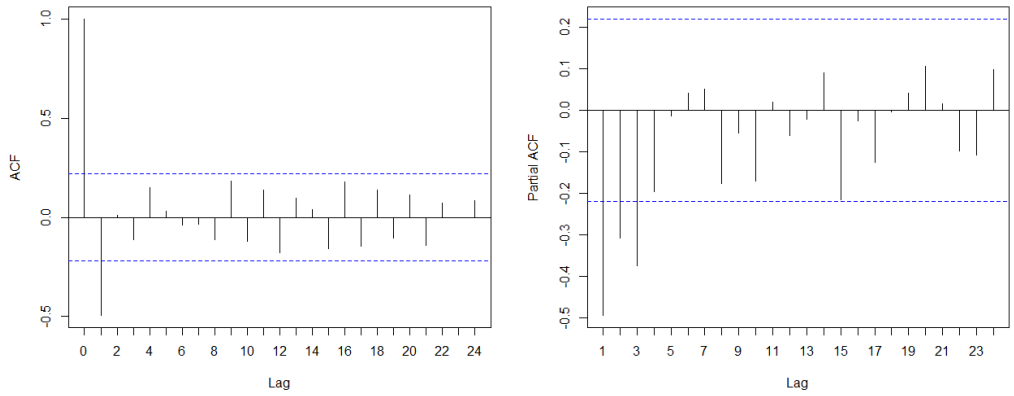

Figure 12: Autocorrelation and partial autocorrelation functions of $\nabla X_{4,t}$.

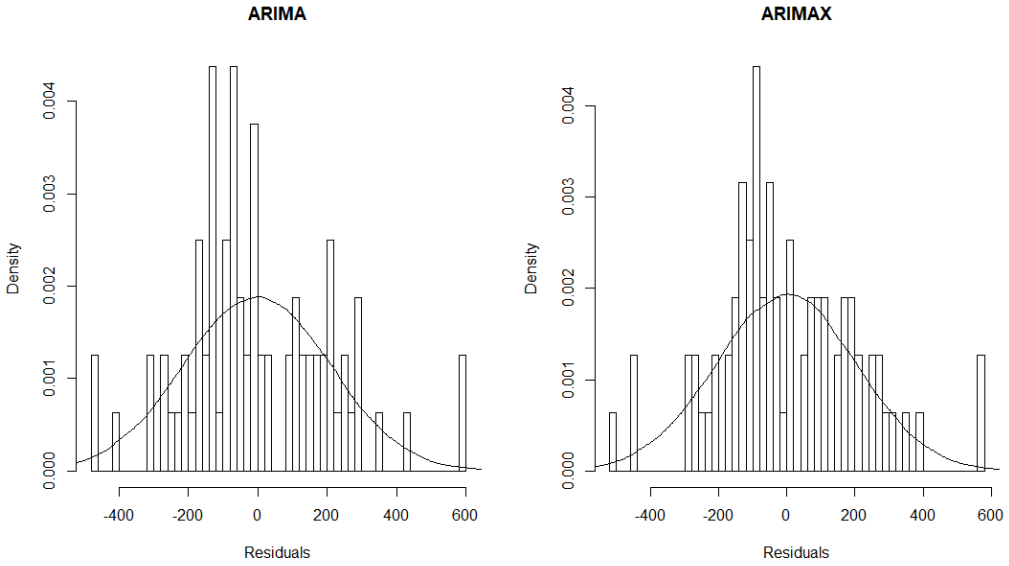

Figure 13: Histogram of density of residuals of ARIMA(2,1,2) process and ARIMAX(2,1,2) process.

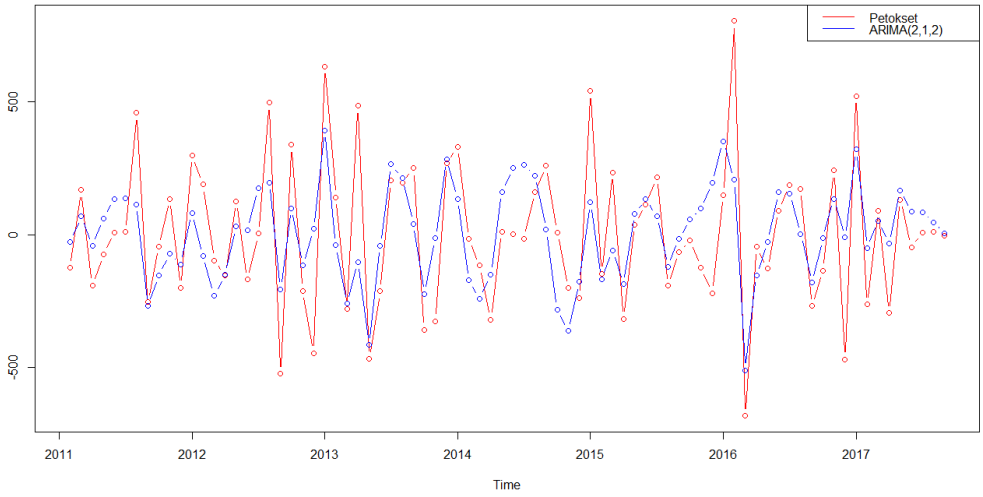

Figure 14: $X_{1,t}$ and corresponding ARIMA(2,1,2) process.

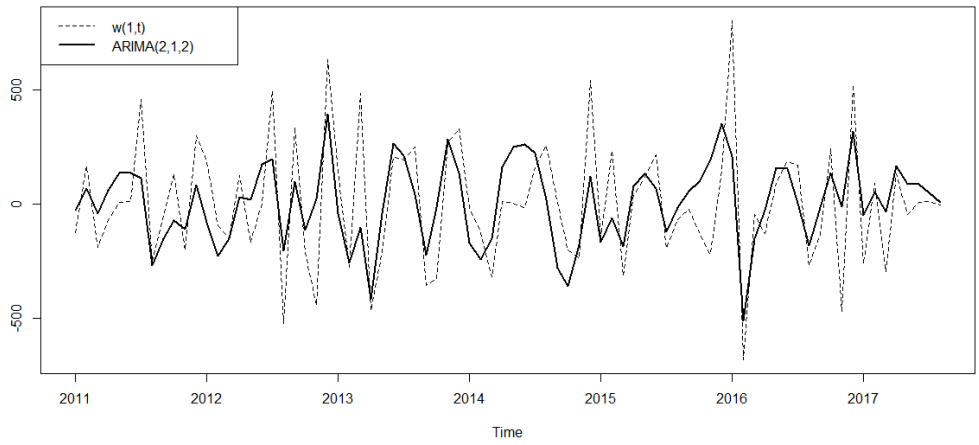

Figure 15: $w_{1,t}$ and the ARIMA(2,1,2) model.

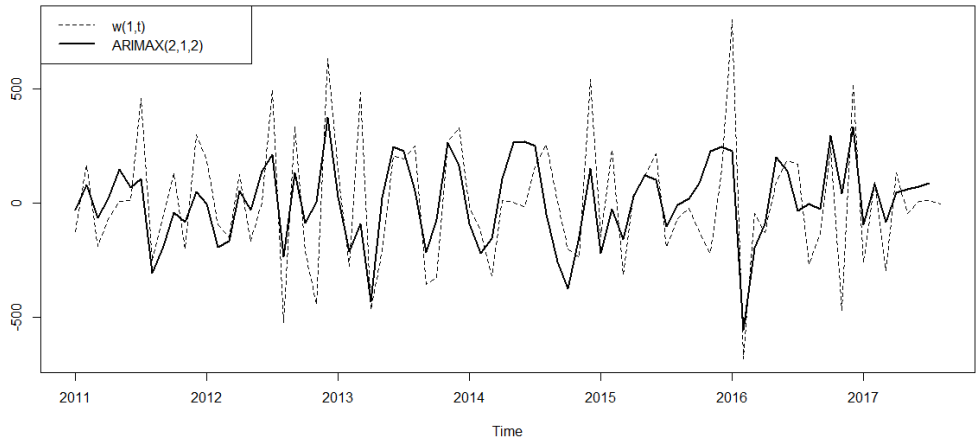

Figure 16: $w_{1,t}$ and the ARIMAX(2,1,2) model.

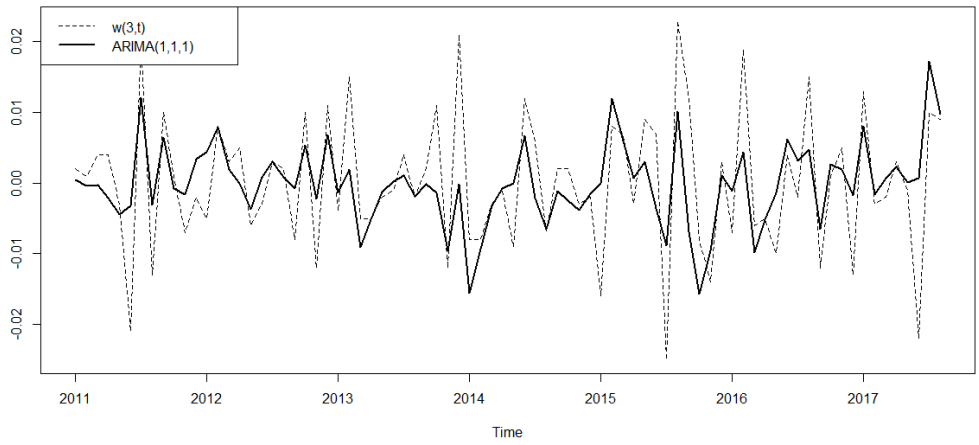

Figure 17: $w_{3,t}$ and the ARIMA(1,1,1) model.

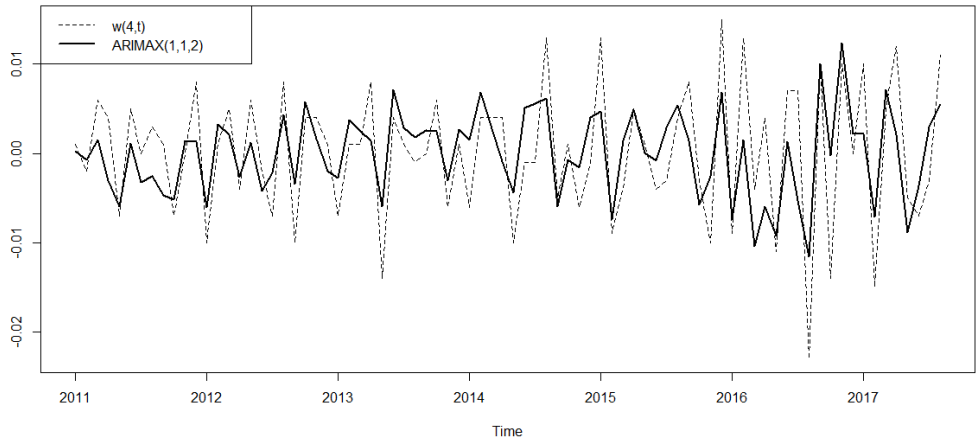

Figure 18: $w_{4,t}$ and the ARIMA(1,1,2) model.

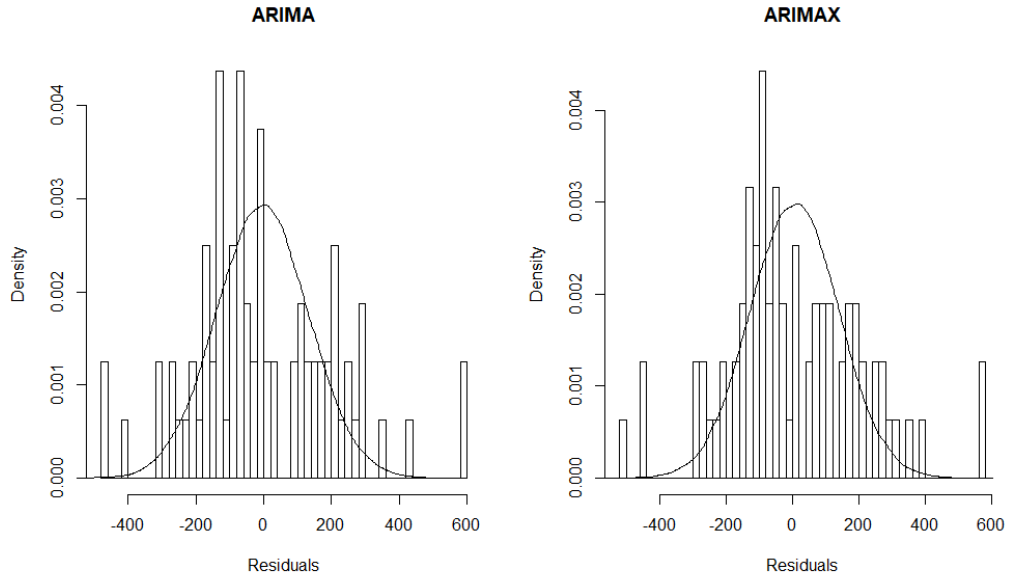

Figure 19: Histogram of density of residuals of IMA(2,1) process and IMAX(2,1) process.

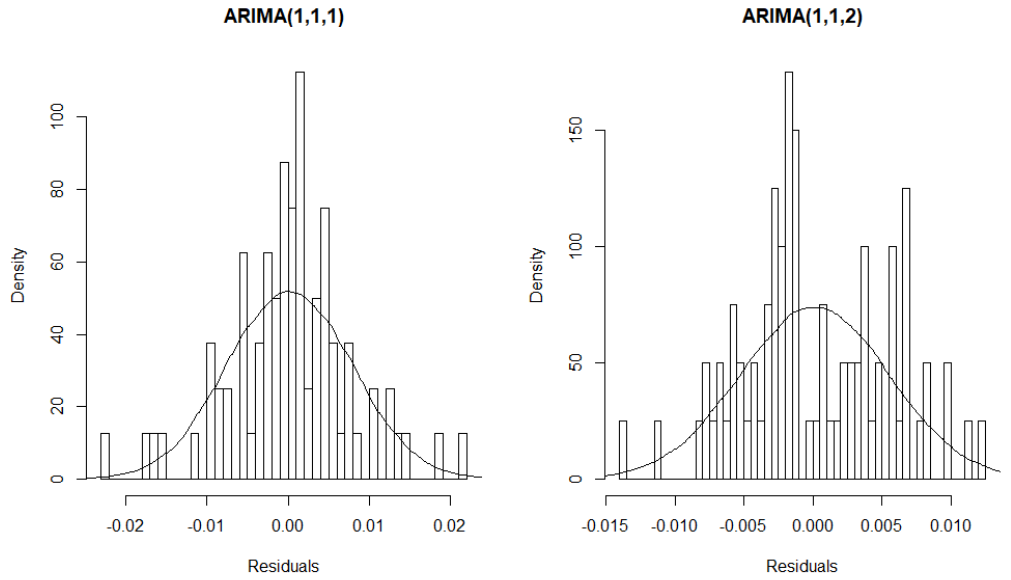

Figure 20: Histogram of density of residuals of ARIMA(1,1,1) process and ARIMA(1,1,2) process.

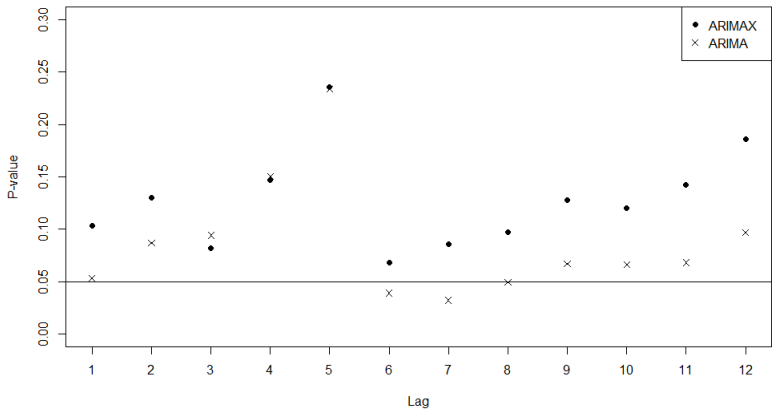

Figure 21: P-values from Ljung-Box test from the ARIMA(2,1,2) and the ARIMAX(2,1,2) process.

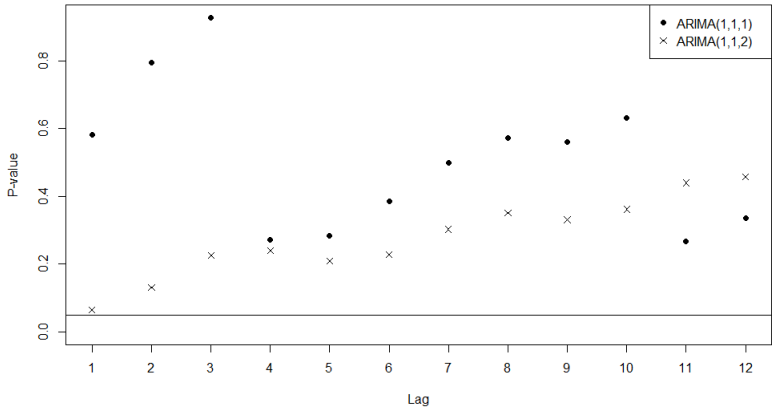

Figure 22: P-values from Ljung-Box test from the ARIMA(1,1,1) and the ARIMA(1,1,2) process.

B Tables

Table 4: ADF-test results of original and stationarized time series of frauds.

	Original			Stationarized		
	Lag	ADF	P-value	Lag	ADF	P-value
No drift no trend	0	-0.56	0.48	0	-13.00	0.01
	1	-0.24	0.57	1	-10.04	0.01
	2	-0.10	0.61	2	-7.81	0.01
	3	0.09	0.67	3	-9.44	0.01
With drift no trend	0	-5.15	0.01	0	-12.93	0.01
	1	-3.95	0.01	1	-9.98	0.01
	2	-3.09	0.03	2	-7.78	0.01
	3	-2.80	0.07	3	-9.45	0.01
With drift and trend	0	-6.50	0.01	0	-12.88	0.01
	1	-5.05	0.01	1	-9.95	0.01
	2	-3.99	0.01	2	-7.81	0.01
	3	-3.36	0.07	3	-9.67	0.01

Table 5: ADF-test results of original and stationarized time series of households who have to use savings.

	Original			Stationarized		
	Lag	ADF	P-value	Lag	ADF	P-value
No drift no trend	0	-0.945	0.339	0	-13.99	0.01
	1	-0.617	0.457	1	-9.52	0.01
	2	-0.454	0.512	2	-8.28	0.01
	3	-0.365	0.538	3	-7.64	0.01
With drift no trend	0	-6.78	0.01	0	-13.9	0.01
	1	-4.93	0.01	1	-9.46	0.01
	2	-4.21	0.01	2	-8.23	0.01
	3	-3.44	0.0141	3	-7.59	0.01
With drift and trend	0	-7.1	0.01	0	-14.28	0.01
	1	-5.23	0.01	1	-9.35	0.01
	2	-4.53	0.01	2	-8.23	0.01
	3	-3.78	0.0238	3	-7.68	0.01

Table 6: ADF-test results of original and stationarized time series of households who are getting into debt.

	Original			Stationarized		
	Lag	ADF	P-value	Lag	ADF	P-value
No drift no trend	0	-1.434	0.163	0	-16.37	0.01
	1	-0.773	0.401	1	-10.72	0.01
	2	-0.55	0.481	2	-9.92	0.01
	3	-0.408	0.525	3	-7.91	0.01
With drift no trend	0	-8.8	0.01	0	-16.26	0.01
	1	-5.67	0.01	1	-10.66	0.01
	2	-4.66	0.01	2	-9.86	0.01
	3	-3.3	0.0201	3	-7.86	0.01
With drift and trend	0	-8.81	0.01	0	-16.17	0.01
	1	-5.72	0.01	1	-10.62	0.01
	2	-4.75	0.01	2	-9.82	0.01
	3	-3.4	0.0609	3	-7.81	0.01