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

Time Series Model for Forecasting the Sales of a Functional Dairy Product

Bachelor's Thesis
20.11.2013

Anton von Schantz

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

ABSTRACT OF
BACHELOR'S THESIS

Author: Anton von Schantz
Title: Time-series Model for Forecasting the Sales of a Functional Dairy Product

Date: November 20, 2013 **Number of pages:** 27 + 5
Professorship: Systems Sciences **Code:** Mat-2
Supervisor: Professor Ahti Salo
Instructor: PhD Antti Vassinen

The effect of marketing and advertising functions on the sales of a large Finnish food company's functional dairy product is measured. In order to model the effects of marketing procedures we have to form a picture of the baseline sales levels. For the estimation of the baseline sales levels product sales and marketing expenditure data is used. Marketing data often includes measures on, for example, sales or marketing mix variables at equally spaced intervals over time. Time series models are uniquely suited to capture the time dependencies in these variables.

An Seasonal AutoRegressive Integrated Moving Average with eXternal inputs (SARIMAX) model is used to model the baseline sales levels. The model is identified, estimated and validated with the Box-Jenkins procedure. In the resulting model the sales is explained by random shocks in the demand and the price promotion of the product. The aforementioned variables also cause a permanent stochastic trend in the sales.

Keywords: SARIMAX, sales, price promotion
Language: English

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

KANDIDAATINTYÖN TIIVISTELMÄ

Tekijä: Anton von Schantz
Työn nimi: Tehojuoman myynnin mallintaminen aikasarjamenetelmällä

Päiväys: 20. marraskuuta 2013 **Sivumäärä:** 27 + 5
Professuuri: Systeemitieteet **Koodi:** Mat-2
Valvoja: Professori Ahti Salo
Ohjaaja: KTT Antti Vassinen

Työssä tutkittiin suuren suomalaisen elintarvikeyhtiön markkinointi- sekä mainostoimenpiteiden vaikutusta myyntituloihin. Jotta markkinointi- ja mainostoimenpiteiden vaikutusta voidaan mitata, täytyy myynnin perustasosta kuitenkin luoda perusteltu arvio. Myynnin perustason estimoimiseksi käytetään myynti- sekä mainostoimenpidedataa. Markkinointidata sisältää usein muuttujia jotka riippuvat ajasta. Tämän takia käytämme aikasarja-analyysiä näiden muuttujien välisten yhteyksien selvittämiseen.

Työssä myynnin perustasoa on mallinnettu SARIMAX mallilla. SARIMAX mallissa myyntituloja selitetään myynnin omalla historialla, satunnaisvaihtelulla sekä ulkoisilla muuttujilla. Tässä tapauksessa markkinointi- ja mainontatoimenpiteet toimivat ulkoisina muuttujina. Mallin tunnistamiseen, estimointiin ja validoimiseen käytetään Box-Jenkins menetelmää. Box-Jenkins menetelmällä saadaan malli, jossa myyntitulot selitetään satunnaisilla shokkeilla sekä tuotteen hintapromootiolla. Edellä mainitut selittäjät aiheuttavat myös pidempiaikaista stokastista trendiä tuotteen myyntiin.

Asiasanat: SARIMAX, myyntitulot, hintapromootio
Kieli: Englanti

Table of Contents

Abbrevations and Acronyms

AR	Auto$R$egressive
MA	Moving Average
$ARMA$	Auto$R$egressive Moving Average
$ARIMA$	Auto$R$egressive Integrated Moving Average
$ARIMAX$	Auto$R$egressive Integrated Moving Average with eXternal variables
$SARIMAX$	Seasonal Auto$R$egressive Integrated Moving Average with eXternal variables
t	discrete time
z_t	the value of a time series at a discrete time t
s	seasonal length
B	backward shift operator defined by $Bz_t = z_{t-1}$ and $B^s z_t = z_{t-s}$
μ	mean level of the process (often $\mu \cong 0$)
ε_t	normally independently distributed white noise residual with mean 0 and variance σ_ε^2 (written as $N(0, \sigma_\varepsilon^2)$)
$\phi(B)$	$= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ nonseasonal autoregressive (AR) operator or polynomial of order p such that the roots of the characteristic equation $\phi(B) = 0$ lie outside the unit circle for nonseasonal stationarity and the $\phi_i,\ i = 1, 2, \dots, p$ are the nonseasonal AR parameters
$(1 - B)^d$	$= \nabla^d$ nonseasonal differencing operator of order d to produce nonseasonal stationarity of the dth differences, usually $d = 0,1,$ or 2
$\phi(B^s)$	$= 1 - \phi_1 B^s - \phi_2 B^{2s} - \cdots - \phi_p B^{Ps}$ seasonal AR operator of order P such that the roots of $\phi(B^s) = 0$ lie outside the unit circle for seasonal stationarity and the $\phi_i,\ i = 1,2, \dots, P$ are the seasonal AR parameters
$(1 - B^s)^D$	$= \nabla^d \nabla_s^D z_t$ stationary series formed by differencing z_t series
$\theta(B)$	$= 1 - \theta_1 B - \theta_2 B^2 - \cdots \phi_q B^q$ nonseasonal moving average (MA) operator or polynomial of order q such that the roots of that the roots of $\theta(B) = 0$ lie outside the unit circle for invertibility and $\theta_i, i = 1,2, \dots, q$ are the nonseasonal MA parameters

$\Theta(B^s)$ $= 1 - \Theta_1 B^s - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$ seasonal MA operator of order Q such that the roots of $\Theta(B^s) = 0$ lie outside the unit circle for invertibility and the Θ_i, $i = 1, 2, \ldots, Q$ are the seasonal MA parameters

The notation $(p, d, q) \times (P, D, Q)_s$ is used to represent the seasonal $ARIMA$ model.

1 Introduction

The justification of marketing spending almost always involves estimating the incremental effects of the marketing or advertising procedure under evaluation. In order to model the effects of marketing procedures we have to form a picture of the baseline sales levels. Estimates of baseline sales levels establish a benchmark for evaluating the incremental sales generated by specific marketing activities. This baseline also helps isolate incremental sales from the effects of other influences, such as seasonality or competitive promotions [1].

The aim of this thesis is to model the baseline sales levels of a large Finnish food company's functional dairy product. The product is a dairy product with added health benefits, thus it bought more during influenza seasons. For the estimation of the baseline sales levels product sales and marketing expenditure data is used. By comparing marketing and advertising activities to the difference between actualized and estimated sales one can form versatile models of marketing functions. In order to model the influence of marketing expenditures, we have to build a model of the sales, where marketing activities constitute as a variable. Marketing data often includes measures on, for example, sales or marketing mix variables at equally spaced intervals over time. Time series models are uniquely suited to capture the time dependencies in these variables [2].

The biggest challenges in modeling sales are the changing dynamics between the variables and the problem of endogeneity. Market responsiveness may not be the same over time. Advertising effectiveness might decline over the life cycle of a product, which means that the dynamics between the variables will change over the life cycle of the product as well [3]. With the problem of endogeneity is meant, that the exogenous variables such as price, promotion, advertising, etc., are themselves endogenous. In statistics an exogenous variable affects a model without being affected by it, and whose qualitative characteristics and method of generation are not specified by the model builder [11]. In turn, an endogenous variable is generated within a model and, therefore, a variable whose value is changed by one of the functional relationships in that model [12]. Marketing managers set the marketing-mix variables based on market information which may be in part unobservable to the researcher but which nevertheless affects consumer choice. This would create a situation where the marketing-mix variables could be correlated with the error terms [2].

An extensively used model in marketing science, to describe effects of marketing activities on sales, is a Vector Auto$R$egressive (VAR) model, as it takes in account the dynamic changes of variables. Because of difficulties in getting the VAR package to function in the statistics software R, a Seasonal Auto$R$egressive Integrated Moving Average with eXternal inputs ($SARIMAX$) model is used instead [4, 5]. The theoretical background of the model is explained in Chapter 2 and the Box-Jenkins methodology for fitting a model to a time series is described in Chapter 3.

In Chapter 4 the case study for the functional dairy product is presented. First the sales and marketing data are described and after that the Box-Jenkins procedure is applied for the data. The goodness of the model is evaluated with Akaike's Information Criterion (AIC) and backtesting

experiments. We will arrive with a baseline sales model for the dairy product. Chapter 5 is for discussion and conclusions.

2 Time Series in Marketing

The following theoretical background about time series in marketing is following closely the presentation by Horváth et al. [2].

2.1 ARMA Processes

Let z_t be the sales of a product in a period *t.* A common and fairly simple way to describe fluctuations in sales over time is with a first-order autoregressive process. In this process, it is assumed that sales at *t-1,* z_{t-1}, affect sales at period *t*

$$z_t = \mu + \varphi z_{t-1} + \varepsilon_t, t = 1, \dots, T, \quad (1)$$

where μ =a constant, ε_t =the error term, φ =a parameter, the starting condition is $z_0 = 0$, and T is the sample length. The model is indicated as $AR(1)$, which means 'autoregressive process of order 1'. It is possible that there is a correlation between higher order lags as well. Values of variables occurring prior to the current observation are called lag values [13]. The order p of an $AR(p)$ process is the highest lag of z_t that appears in the model. The general p-order AR process is

$$\varphi_p(B) z_t = \mu + \varepsilon_t, t = 1, \dots, T \quad (2)$$

where $\varphi_p(B) = \left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right)$. Here B denotes the backshift operator defined by $B^k y_t = y_{t-k}$.

A first-order moving average process assumes that a random shock at $t-1$, ε_{t-1}, affects sales levels at time t

$$z_t = \mu + \varepsilon_t - \theta \varepsilon_{t-1}, t = 1, \dots, T. \quad (3)$$

This model is indicated as $MA(1)$. Notice that from here on the error term ε_t will be called a random shock. Note also that the past shock does not come from past sales (past values of z_t) as in the $AR(1)$ model, but it stems from the random component of ε_{t-1}. The order q of an $MA(q)$ process is the highest lag of ε_t that appears in the model. The general q-order MA process is

$$z_t = \mu + \theta_q(B) \varepsilon_t, t = 1, \dots, T, \quad (4)$$

where $\theta_q(B) = \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\right)$.

AR and MA processes can be combined into a single model to reflect the idea that both past sales and past random shocks affect z_t. For example, the $ARMA(1,1)$ process

$$z_t = \mu + \varphi z_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}, t = 1, \dots, T. \quad (5)$$

The orders (p, q) of a $ARMA(p, q)$ process are the highest lags of z_t and ε_t respectively that appear in the model. For example, for an $ARMA(1,1)$ process, $p = 1$ and $q = 1$. The general $ARMA(p, q)$ process is

$$\varphi_p(B) z_t = \mu + \theta_q(B) \varepsilon_t, t = 1, \dots, T \quad (6)$$

If we are interested in estimating the effects of marketing variables such as price, advertising and competitive behavior on sales when the latter variable is also a subject to other complex patterns, we can include these variables in an $ARMA$ model, and obtain the following $ARMA$ model with eXogenous variables $(ARMAX)$

$$\varphi_p(B) z_t = \mu + \theta_q(B) \varepsilon_t + \eta_b(B) x_t, t = 1, \dots, T. \quad (7)$$

2.2 Stationary and Non-stationary Processes

In order to use statistical tests for an $ARMA$ model selection, we need to estimate the underlying stochastic process. We can do this by deriving the mean, variance, and co-variance of the sample data. However, these quantities are only meaningful, for obtaining the probability distribution and the statistical tests based on it, if they are independent of time. This is the case if the series is stationary. There are different forms of stationarity. The most commonly considered is covariance stationarity. A series z_t is said to be covariance stationary if the following conditions hold:

$$\begin{aligned} &E(z_t) = m, \text{ for all } t = 1, \dots, T \\ &E[(z_t - m)^2] = \gamma_o < \infty, \text{ for all } t = 1, \dots, T \\ &E[(z_t - m)(z_{t-1} - m)] = \gamma_i, \text{ for all } t = 1, \dots, T \text{ and for all } i = \cdots, -2, -1, 0, 1, 2, \dots \end{aligned} \quad (8)$$

where m, γ_0 and γ_i are all finite-valued numbers.

In practice, we often use the requirement that the roots of what are called the characteristic equations, $\varphi_p(.) = 0$, "lie outside the unit circle". For an $AR(1)$ process this requirement implies that the root is larger than one in absolute value. The characteristic equation is $(1 - \varphi B) = 0$. The root equals $\frac{1}{\varphi}$, which is greater than one in absolute sense if $|\varphi| < 1$. The root $\frac{1}{\varphi}$ is called a *unit root* if $|\varphi| = 1$. Thus, the $AR(1)$ process is *stationary* if $|\varphi| < 1$ and *non-stationary* (not stationary) if it has a unit root.

A time series might exhibit a stochastic trend, which implies that the variation is systematic but hardly predictable. If the time series is not stationary, because it has a stochastic trend, the series can be formulated as a stationary $ARMA$ process in differences. The differencing operation removes

the stochastic trend from the data. It may be necessary to take the differences of the series more than once before it becomes stationary. A $ARMA$ model for the differences of z_t is called a $ARIMA$ (Auto Regressive Integrated Moving Average) model. As an example we consider a $ARIMA(1,1,1)$ model

$$\Delta z_t = \mu + \varphi \Delta z_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}, t = 1, \dots, T. \quad (9)$$

2.3 Seasonal Processes

Many time series in marketing display seasonal patterns caused by managerial decisions, weather conditions, events, holidays, etc. A seasonal model may apply, with orders P, D and Q respectively for the AR, I and MA components, denoted by $ARIMA(P, D, Q)_s$, where s is the seasonal lag. To illustrate, suppose there exists a seasonal pattern in monthly data, such that any month's value contains a component that resembles the previous year's value in the same month. Then a purely seasonal $ARIMA(1,0,1)_{12}$ model is

$$z_t = \mu + \varphi z_{t-12} + \varepsilon_t - \theta \varepsilon_{t-12}, \ t = 1, \dots, T. \quad (10)$$

In practice, seasonal and non-seasonal processes usually occur together. The examination of ACF and $PACF$ may suggest patterns in these functions at different lags. This general process is indicated as an $ARIMA(p, d, q)(P, D, Q)_{12}$ process

$$\varphi_p(B^{s)} \varphi_p(B) \Delta^D \Delta^d z_t = \mu + \theta_Q(B^s) \theta_q(B) \varepsilon_t, t = 1, \dots, T. \quad (11)$$

In Eq. 11 the seasonal and non-seasonal AR, MA and differencing operators are multiplied. In practice, the orders p, d, q and P, D, Q are small, ranging from 0 to 2 in most cases.

3 Methodology

When applying an $ARIMA$ model it is recommended to use the three stage Box-Jenkins procedure. The description of the procedure in this chapter widely follows [4]. In order to understand the Box-Jenkins procedure the theory behind correlation functions is first described by following closely the representations in [6] and [7].

3.1 Correlation functions

The cross-correlation between two time series describes the normalized cross-covariance function. Let (X_t, Y_t) represent a pair of stochastic processes that are covariance stationary. Then the cross-covariance is given by

$$\gamma_{xy}(\tau) = E\big[(X_t - \mu_x)\big(Y_{t+\tau} - \mu_y\big)\big], \tag{12}$$

where μ_x and μ_y are the means of X_t and Y_t respectively.

The cross-correlation function ρ_{xy} is the normalized cross-covariance function

$$\rho_{xy}(\tau) = \frac{\gamma_{xy}(\tau)}{\sigma_x \sigma_y}, \tag{13}$$

where σ_x and σ_y are the standard deviations of processes X_t and Y_t respectively. If $X_t = Y_t$, then the cross-correlation function is simply the autocorrelation function [6].

We are often interested in the relationship between two points in time in a time series. One way to measure the linear relationship is with autocorrelation, i.e., to determine the correlation between these two variables. Another way to measure the relationship between variables X_t and $X_{t+\tau}$ is to filter the linear dependency that is caused by the variables $X_{t+1}, \dots, X_{t+\tau-1}$ between the variables X_t ja $X_{t+\tau}$, and after that to calculate the correlation between the random variables. This is called partial autocorrelation, autocorrelation of k:th degree is

$$\phi_{kk} = Corr\big(X_t - P(X_t|X_{t+1}, \dots, X_{t+k-1}), X_{t+k} - P(X_{t+k}|X_{t+1}, \dots, X_{t+k-1})\big), \tag{14}$$

where $P(W|Z)$ is the best linear projection from W to Z, i.e. $P(W|Z) = \Sigma_{WZ}\Sigma_{ZZ}^{-1}Z$ where $\Sigma_{ZZ} = \mathrm{Var}(Z)$ is the covariance matrix of the regressors and $\Sigma_{WZ} = \mathrm{Cov}(W, Z)$ is the variance between W and Z. With the best linear projection we mean a linear projection, which minimizes the square sum of the errors.

An equivalent definition is the solution to the equation system

$$P_k \phi_k = \rho(k), \tag{15}$$

where

$$P_k = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix},$$

and $\phi_k = (\phi_{k1}, \dots, \phi_{kk})^T$ and $\rho_k = (\rho_{k1}, \dots, \rho_{kk})^T$. This equation system is also called the Yule-Walker equations for an AR(k) process. The last coefficient ϕ_{kk} is the autocorrelation of kth degree. Because we are only interested in this coefficient, the equation system can be solved with regards to ϕ_{kk}, using the Cramer rule. We get

$$\phi_{kk} = \frac{|P_k^*|}{|P_k|}, \tag{16}$$

where P_k^* corresponds to the matrix P_k, where the kth column has been replaced by $\rho_{(k)}$. In this context $|\cdot|$ means the determinant. Because this can be applied to different values of k, we can construct a partial autocorrelation function (PACF).

From the definition of PACF we can see that there is no difference between the first degree PACF and ACF:

$$\phi_{11} = \rho_1.$$

3.2 The Box-Jenkins procedure

The first step in the procedure is to identify the form of model that may fit the given data. The second stage is to estimate the model parameters by employing the method of maximum likelihood. The third stage is to employ diagnostic checks to check the model for possible inadequacies

3.2.1 Identification

The purpose of the identification stage is to transform the data so that it becomes stationary and determine the number of parameters in the $ARIMAX$ model. The first stage is to plot the original time series, it's ACF (autocorrelation function, PACF (partial autocorrelation function) and CCF (cross-correlation function) with external inputs, and check for seasonality, trends either in the mean level or in the variance of the series, long-term cycles and extreme values and outliers.

Whether the data is stationary can be checked with the Augmented Dickey-Fuller test. It's a test for a unit root in a time series sample [5]. After the data have been differenced just enough to produce nonseasonal stationarity for a nonseasonal time series and both seasonal and nonseasonal stationarity for seasonal data, then the ACF and PACF of the stationarized series are inspected to determine the number of AR and MA parameters required in the model [4]. The lagged effects of the external inputs in the $ARIMAX$ model can be determined from the CCF.

3.2.2 Estimation

The model parameters are estimated with the maximum likelihood method. Given a sample and a parametric distribution, the method estimates the parameters in the distribution so that the sample is most likely to be drawn from the distribution.

If the distribution of Y_i is $F(y,\theta)$ where F is a known distribution function and $\theta \in \Theta$ is an unknown $m \times 1$ vector, we say that the distribution is parametric and that θ is the parameter of the distribution F. The space Θ is the set of permissible value for θ. In this setting the method of maximum likelihood is the appropriate technique for estimation and inference on θ.

If the distribution F is continuous then the density of Y_i can be written as $f(y,\theta)$ and the joint density of a random sample $\hat{Y} = (Y_1, \dots, Y_n)$ is

$$f_n(\hat{Y},\theta) = \prod_{i=1}^{n} f(Y_i,\theta). \tag{17}$$

The likelihood of the sample is this joint density evaluated at the observed sample values, viewed as a function of θ. The log-likelihood function is its natural log

$$L_n(\theta) = \sum_{i=1}^{n} \ln f(Y_i,\theta). \tag{18}$$

If the distribution F is discrete, the likelihood and log-likelihood are constructed by setting $f(y,\theta) = P(Y = y,\theta)$.

The maximum likelihood estimator or MLE $\hat{\theta}$ is the parameter value which maximizes the likelihood (equivalently, which maximizes the log-likelihood). We can write this as

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L_n(\theta). \tag{19}$$

In some simple cases, we can find an explicit expression for $\hat{\theta}$ as a function of the data, but these cases are rare. More typically, the MLE $\hat{\theta}$ must be found by numerical methods [8].

3.2.3 Diagnostic Checks

Diagnostic checks are designed to test model adequacy. We are mainly interested in testing if including parameters improves the fit and also to test if the residual assumptions hold. The residuals ε_t should be independent, homoscedastic, and normally distributed. Residual estimates are needed for the tests. The estimates for ε_t are automatically calculated at the estimation stage along with the MLE for the parameters.

Transforming the data cannot correct dependence of the residuals because the lack of independence indicates the present model is inadequate. Rather, the identification and estimation stages must be repeated in order to determine a suitable model [4].

a. Overfitting

Overfitting involves fitting a more elaborate model than the one estimated to see if including one or more parameters greatly improves the fit. Extra parameters should be estimated for the more complex model only where it is feared that the simpler model may require more parameters. A

common approach to model selection is a selection criterion. One popular choice is the Akaike Information Criterion (AIC). The AIC for a model *m* is

$$AIC_m = \log(\hat{\sigma}_m^2) + 2\frac{k_m}{n}, \tag{20}$$

where $\hat{\sigma}_m^2$ is the variance estimate for model m, and k_m is the number of coefficients in the model. The AIC can be derived as an estimate of the Kullback-Leibler information distance $\mathrm{K}(\mathcal{M}) = E(\log f(Y|X) - \log f(Y|X,M))$ between the true density and the model density. The rule is to select M_1 if $AIC_1 < AIC_2$, else select M_2.

Another method of testing model adequacy is by calculating the likelihood ratio. Let $\hat{\theta}_i$ denote the maximum likelihood estimator of θ_i for the likelihood function. $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta_1, \dots, \theta_k)$, where the likelihood function is treated as a function of the parameters and the x_i are fixed. Similarly, let $\hat{\theta}_i'$ denote the maximum likelihood estimator of θ_i when H_0 is true; that is, for the likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta_1', \dots \theta_k')$. Now, form the ratio

$$\lambda = \frac{L(\hat{\theta}_i')}{L(\hat{\theta}_i)} \tag{21}$$

This is the ratio of the two likelihood functions $L(\theta')$ and $L(\theta)$ when their parameters have been replaced by their maximum likelihood estimators. Since the maximum likelihood estimators are functions of the random variables $x_1, \dots, x_k$ the ratio λ is a function of $x_1, \dots, x_k$ only and is therefore an observable random variable.

The test statistic is

$$D = -2\ln(\lambda) \tag{22}$$

The model with more parameters will always fit at least as well (have a greater log-likelihood). Whether it fits significantly better and should thus be preferred is determined by deriving the probability or p-value of the difference D. In many cases, the probability distribution of the test statistic is approximately a chi-square distribution with degrees of freedom equal to df2-df1, if the nested model with fewer parameters is correct. Symbols df1 and df2 represent the number of free parameters of models 1 and 2, the null model and alternate model, respectively [8].

b. Test for whiteness of the residuals.

To determine whether the residual ε_t are white noise, an appropriate procedure is to examine the residual autocorrelation function. Another way to check for white noise is to perform the Ljung-Box test.

To test the hypothesis that the residuals are independent and identically distributed random variables we can use the portmanteau test. We consider the statistic

$$Q = n\sum_{j=1}^{h} \hat{\rho}^2(j). \tag{23}$$

If $Y_1, \dots, Y_n$ is a finite-variance iid sequence, Q is approximately distributed as the sum of squares of the independent $N(0,1)$ random variables, $\sqrt{n}\hat{p}(j), j = 1, \dots, h$, i.e., as chi-squared with h degrees of freedom. A large value of Q suggests that the sample autocorrelations of the data are too large for the data to be a sample from i.i.d. sequence. We therefore reject the i.i.d. hypothesis at level α if $Q > \chi^2_{1-\alpha}(h)$, where $\chi^2_{1-\alpha}(h)$ is the $1-\alpha$ quantile of the chi-squared distribution with h degrees of freedom. The program ITSM conducts a refinement of this test, formulated by Ljung and Box (1978), in which Q is replaced by

$$Q_{LB} = \frac{n(n+2)\sum_{j=1}^{h}\hat{\rho}^2_{WW}(k)}{n-k}, \tag{24}$$

whose distribution is better approximated by the chi-squared distribution with h degrees of freedom [8].

c. Homoscedasticity checks of the residuals.

The check for homoscedasticity of residuals is done in this thesis, by only observing the residual plot.

d. Tests for normality of the residuals.

The normality assumption of the residuals is tested to examine the goodness of the fit. The normality assumption is tested by testing for skewness and kurtosis of residuals. Skewness is tested with D'Agostino skewness test and kurtosis with Anscomble-Glynn kurtosis test [9].

4 Case study: Functional Dairy Product

4.1 Description of data

In this thesis the baseline sales of a functional dairy product is modeled, with data from weeks 2/2009-16/2011. There was a vast amount of data available of different marketing expenditures, which could be used as external variables in the model, the advertising of the product in different media, the advertising of substitute products and price promotions. But after examination of the correlation matrix between the sales data and the possible external variables the conclusion was that the price promotion was the only variable with a significant correlation.

The time series of the product's sales is plotted in Figure 1. The sales are the amount sold to a specific retailer. The sales of the product are measured in kilograms. The data has been partitioned into estimation and forecasting intervals. The 104 first weeks are used to estimate the model, the 16 last weeks are used for backtesting of the model. Unfortunately, the time series is relatively short for determining seasonal patterns. Even though, it would be reasonable to assume that the functional

dairy product would be bought more during typical influenza seasons. Thus the time series will not be seasonally differentiated in this case study.

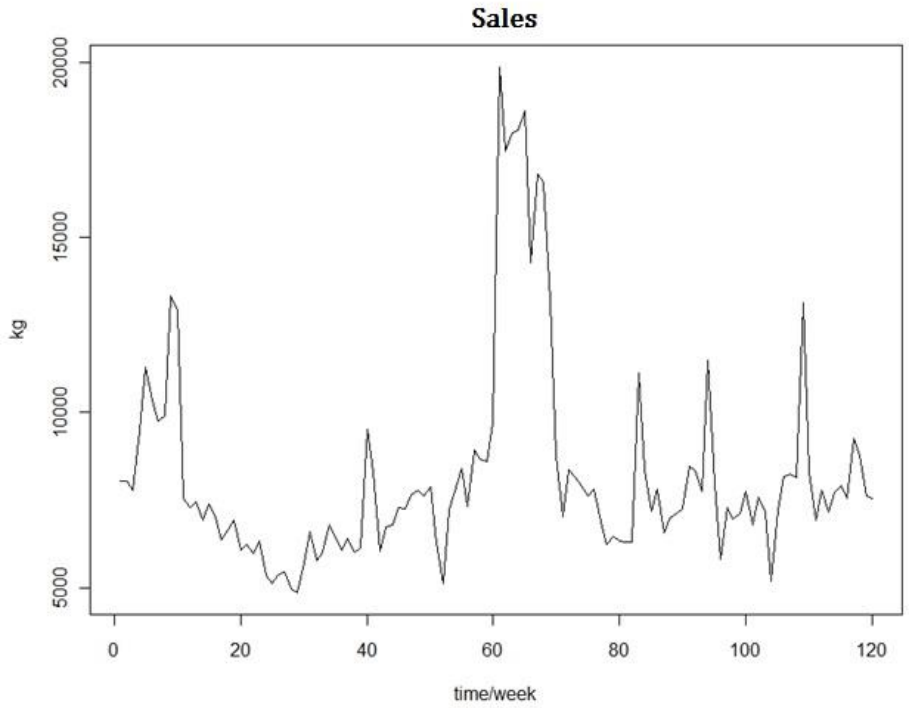

Figure 1: The sales of the functional dairy product to a retailer during weeks 2/2009-16/2011 measured in kilograms.

The products price promotion time series in the retailer chain is plotted in Figure 2. Price promotion is a variable, assuming the value 4 if the prices are at highest and the value 0 if the prices are at lowest.

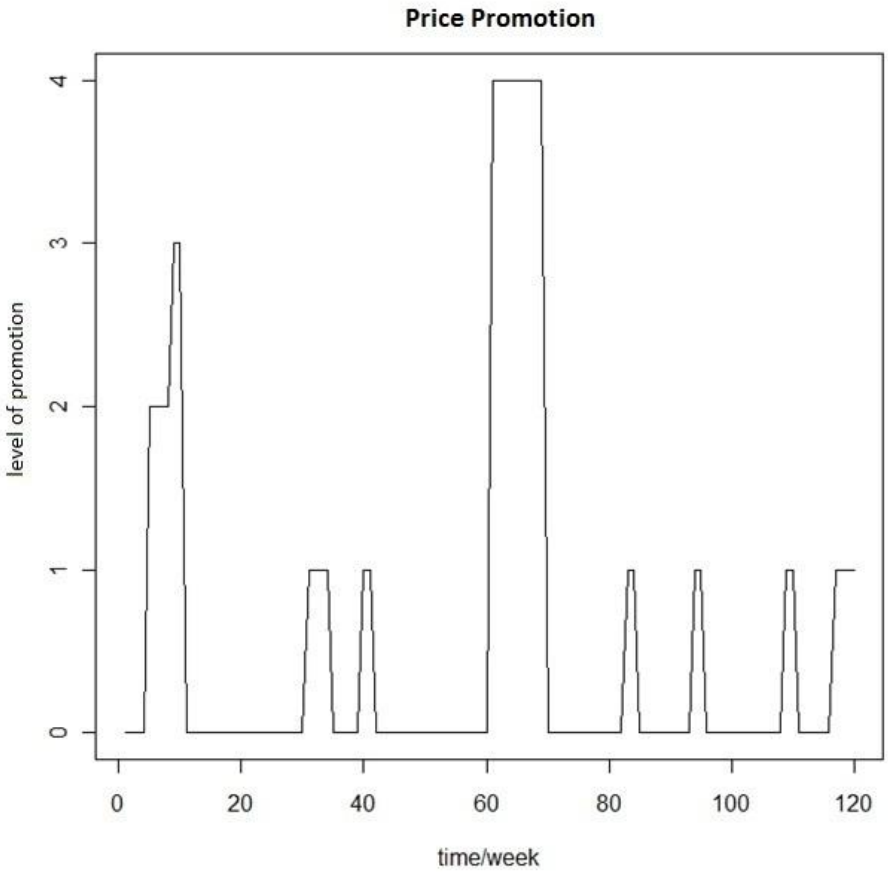

Figure 2: The product's price promotions in the retail chain during weeks 2/2009-16/2011.

4.2 Results

Identification

First a logarithmic transformation of the data was used to stabilize the variance in the data [10]. The log-transformed sales data is plotted in Figure 3.

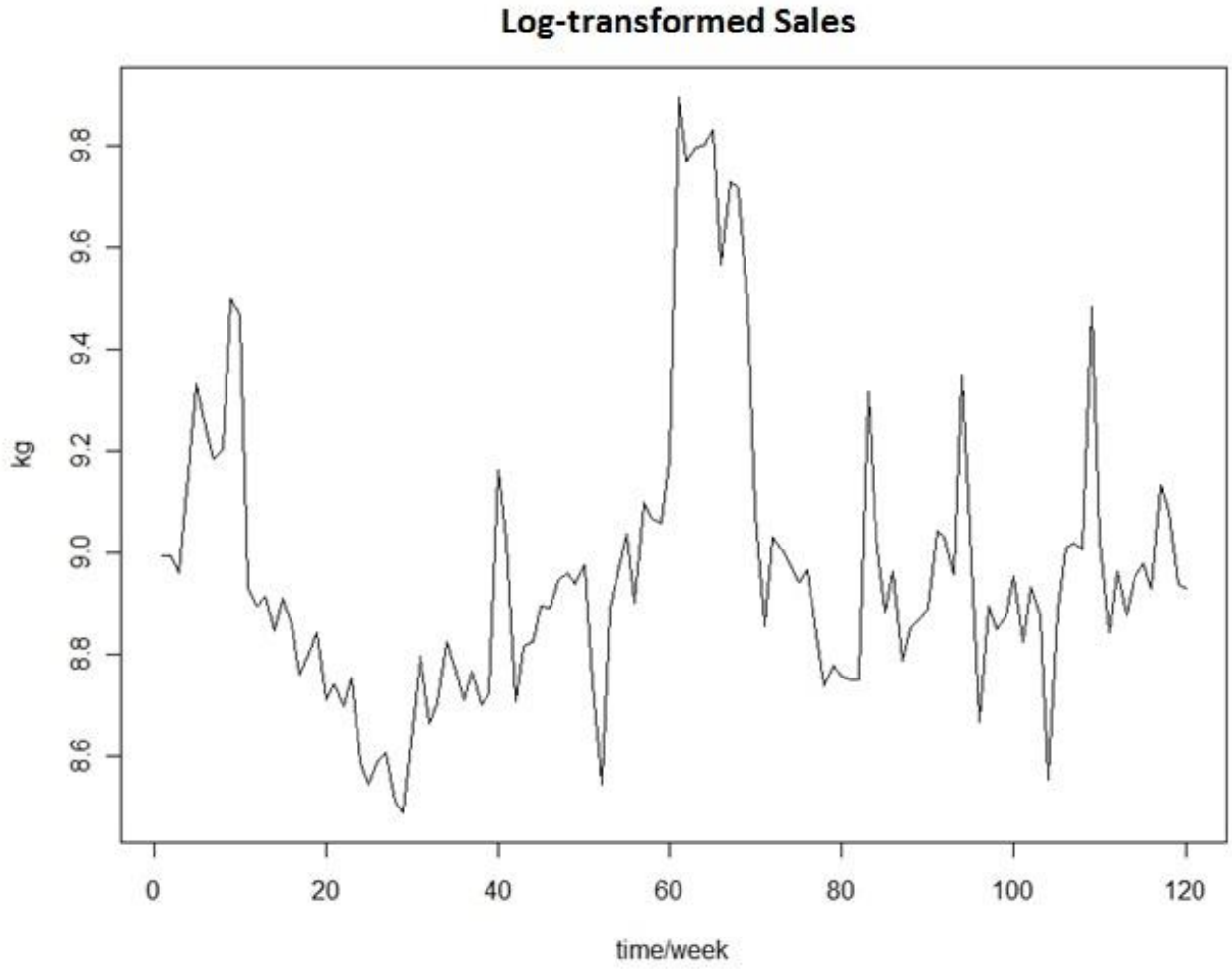

Figure 3: Log-transformed sales of the product in the retailer chain during weeks 2/2009-16/2011 measured in kilograms.

Using the ADF-test we can check if the log-transformed sales data is stationary. We use the ADF-test with 4 lags. The p-value of the test statistic is 0.1397, so we can't reject the null that a unit root exists and the data is non-stationary. We need to remove a stochastic trend of first degree by differencing the data in order to be able to estimate an ARIMAX model. The differenced and log-transformed sales data is plotted in Figure 4.

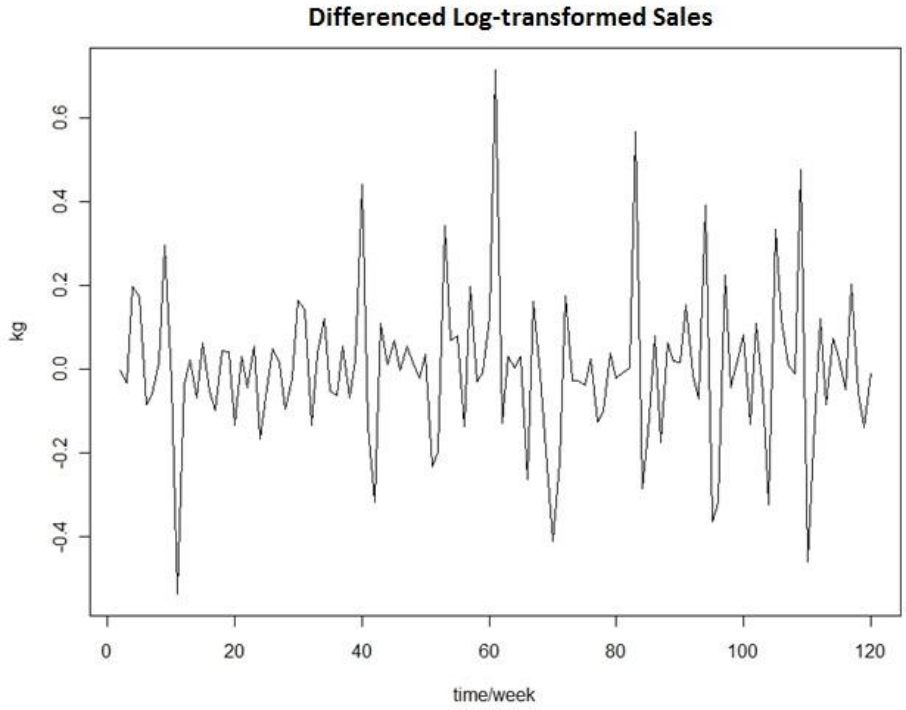

Figure 4: Differenced and log-transformed sales of the product in the retailer chain during weeks 2/2009-16/2011 measured in kilograms.

Now the time series seems relatively stationary. We use the ADF-test again with 4 lags to check the stationarity assumption. The p-value of the test statistic is 0.01, which means that we can reject the null that there exists a unit root and assume that the data is stationary.

From looking at the ACF and PACF functions (Figure 5) for the stationary data, we can identify the model parameters. It's not however that unambiguous to determine the order of AR and MA parts. It looks like the PACF function abates exponentially and that the ACF function cuts at the first lag. This indicates, that one should use a $ARIMAX(0,1,1)$ model.

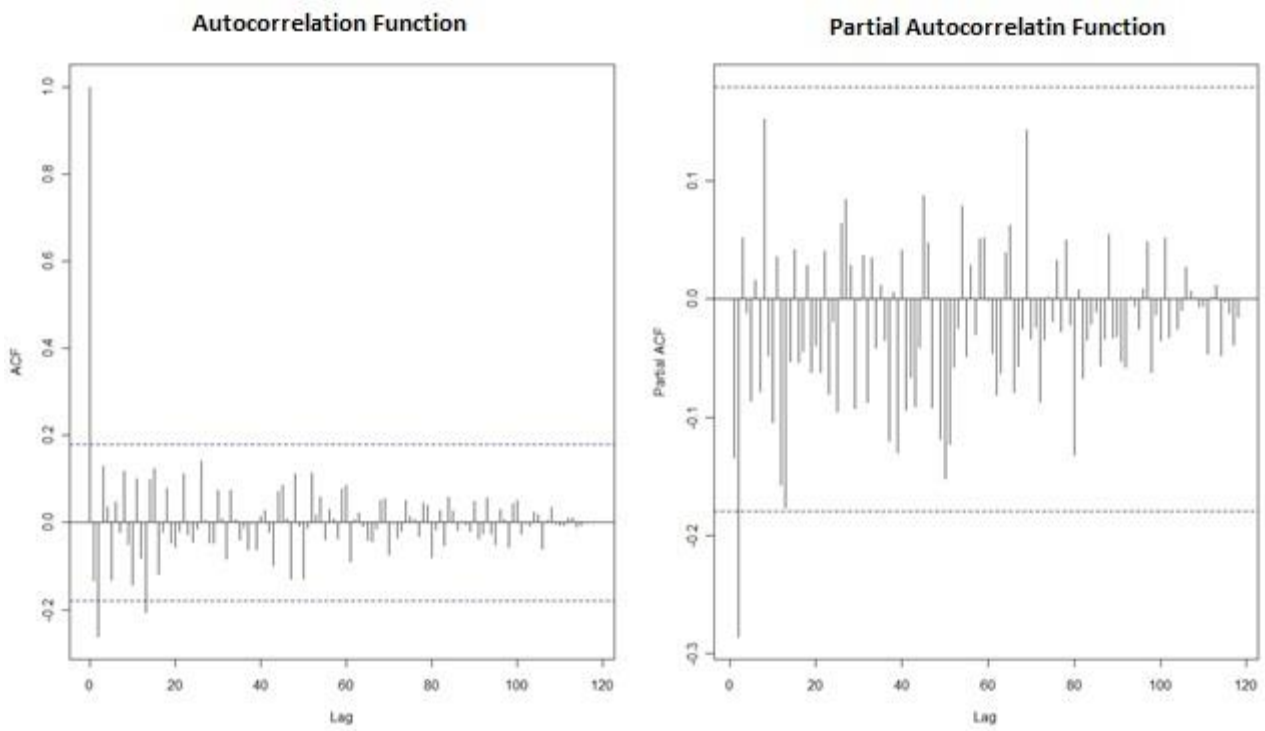

Figure 5: ACF and PACF of the differenced and transformed sales data.

The cross-correlation function (CCF) between the transformed sales data and external variables helps us to determine at which lag the external variable affects sales. The CCF can only be counted for wide-sense stationary series, so we need to differentiate the price promotion time series first. From Figure 6 we see that the correlation between sales and price promotion. Price promotion is highest at the zero lag.

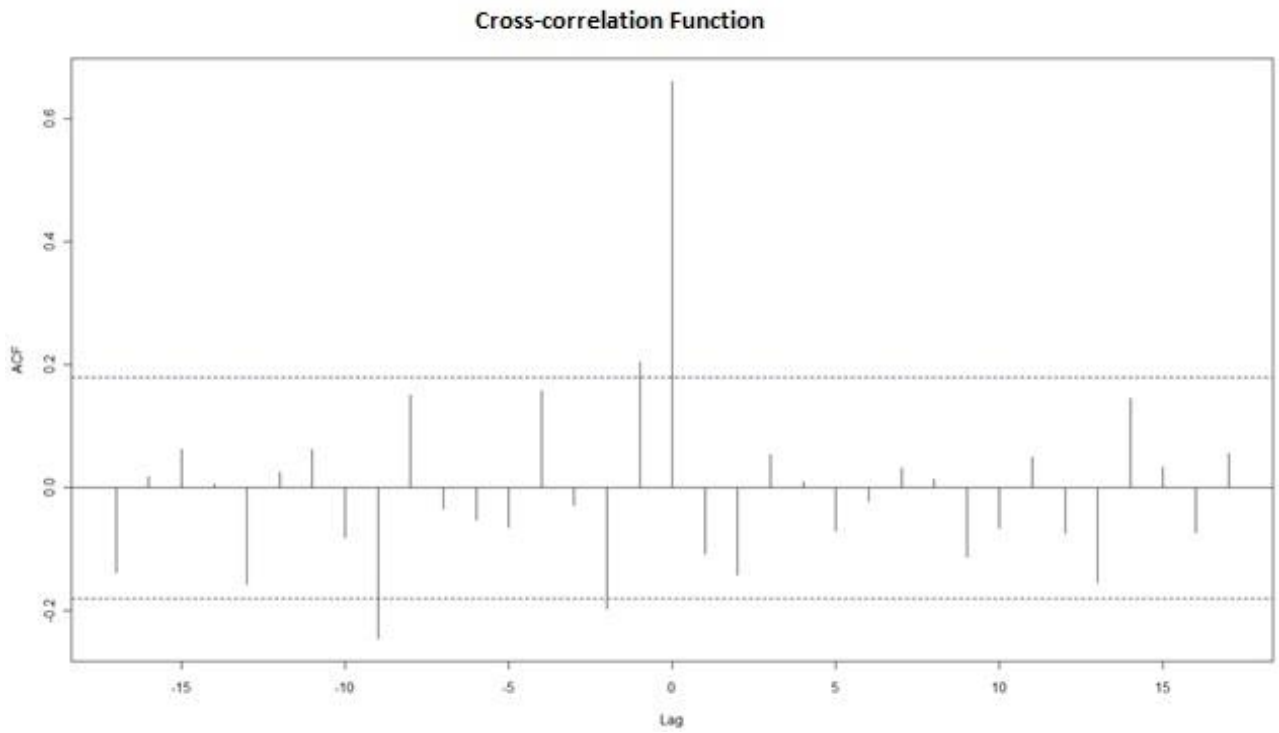

Figure 6: CCF between sales and price promotion data.

Estimation

The next step is to estimate the $ARIMAX(0,1,1)$ model parameters with maximum likelihood method. The time series model was estimated with several parameter combinations because of the ambiguity of the model parameters. The log-likelihood and AIC was calculated for all these models in Table 1.

Table 1: Model statistics.

(p,d,q)	Loglikelihood	AIC
(0,1,1)	73,83	-141,65
(1,1,0)	69,35	-132,7
(1,1,1)	74,15	-140,29
(0,1,2)	74,27	-140,55
(2,1,0)	74,7	-141,4
(1,1,2)	75,89	-141,79
(2,1,1)	74,71	-139,42
(2,1,2)	75,91	-139,83
(0,1,3)	74,84	-139,68
(3,1,0)	74,7	-139,41
(1,1,3)	75,91	-139,82
(3,1,1)	75,24	-138,48
(2,1,3)	75,92	-137,83
(3,1,2)	75,88	-137,77
(3,1,3)	75,9	-135,79
(0,1,4)	75,98	-139,96
(1,1,4)	76,03	-138,06
(2,1,4)	76,1	-136,2
(4,1,4)	78,15	-136,3
(4,1,0)	75,52	-139,03
(4,1,1)	75,92	-137,85
(4,1,2)	75,93	-135,86
(4,1,3)	75,93	-133,86
(3,1,4)	77,67	-137,34

It seems that the $ARIMAX(0,1,1)$ has the lowest AIC, so it's considered as the best model.

Diagnostic Checks

After the most suitable model is chosen, we have to check for model adequacy. The residuals should be white noise. In Figure 7 we see that the residuals have a constant mean, the variance is also relatively constant except for few spikes.

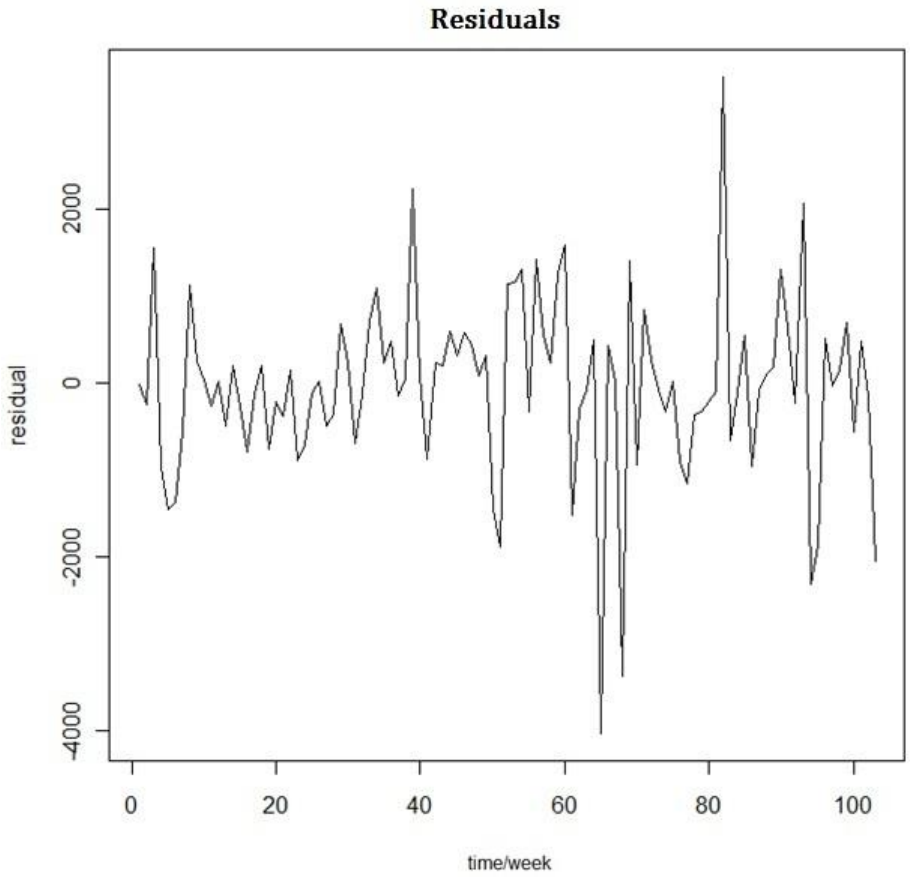

Figure 7: The residual time series.

For the residuals to be white noise, we also need to make sure that they are not autocorrelated. From both the residual ACF (Figure 8) and Ljung-Box test (Figure 9) we see that the residuals aren't autocorrelated.

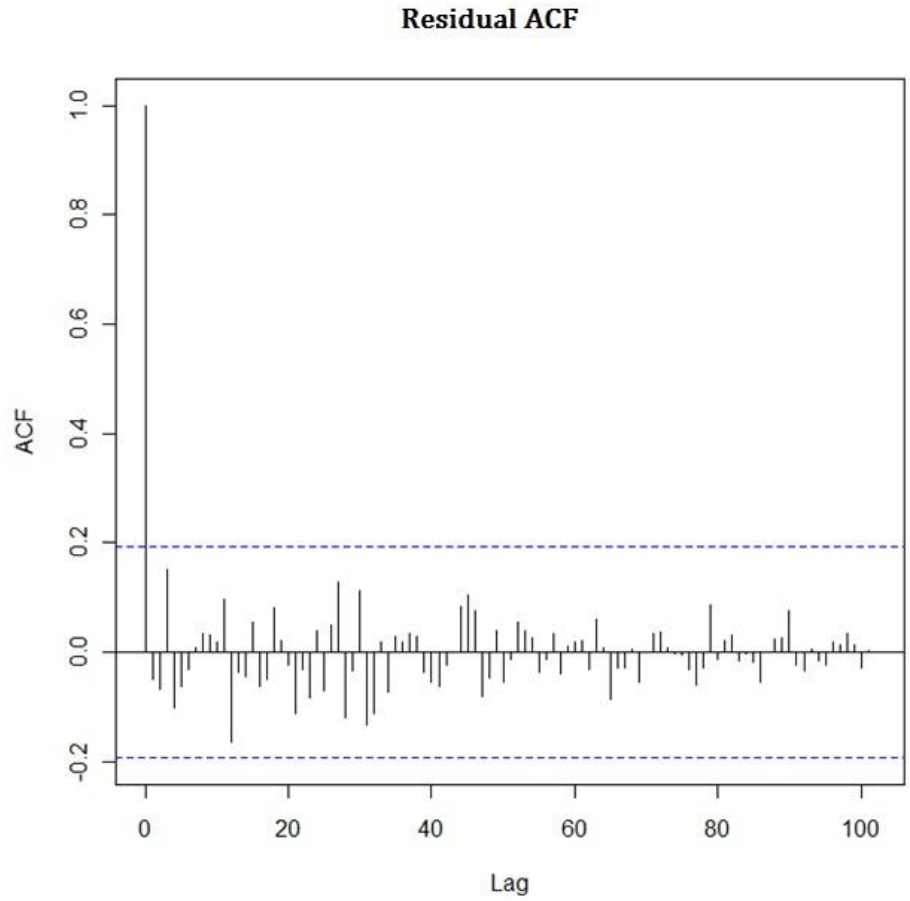

Figure 8: Residual ACF.

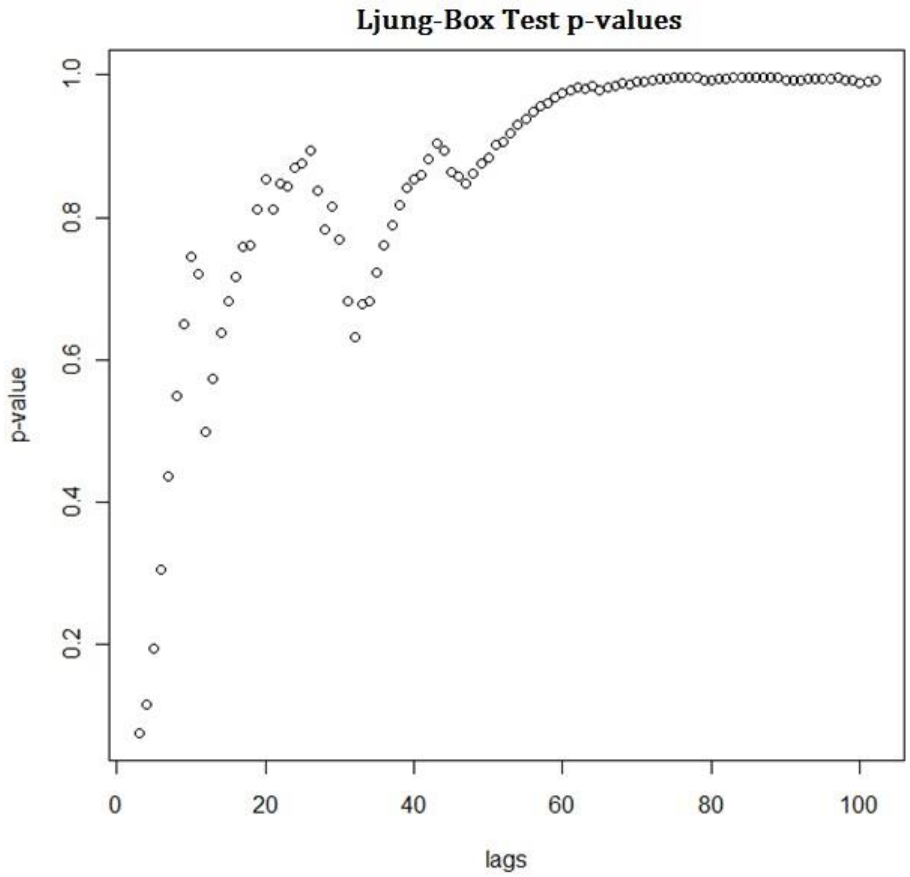

Figure 9: The p-values corresponding to the Ljung-Box test statistic plotted against number of lags.

If the model fit is good the residuals should be normally distributed. The residual histogram is plotted in Figure 10. The residuals seem to be relatively normally distributed, but we check the normality assumption by performing D'Agostino skewness test and Anscomble-Glynn kurtosis test. The p-value of the skewness test is 0.1978 meaning that there doesn't exist skewness in the data. But the kurtosis test p-value is equal to 0.0003801, which means that the kurtosis is not equal to that of a normal distribution. But when the Anscomble-Glynn kurtosis test was performed to other model parameter combinations, they didn't either have the kurtosis of a normal distribution. Thus, the model is not going to be rejected based on this test.

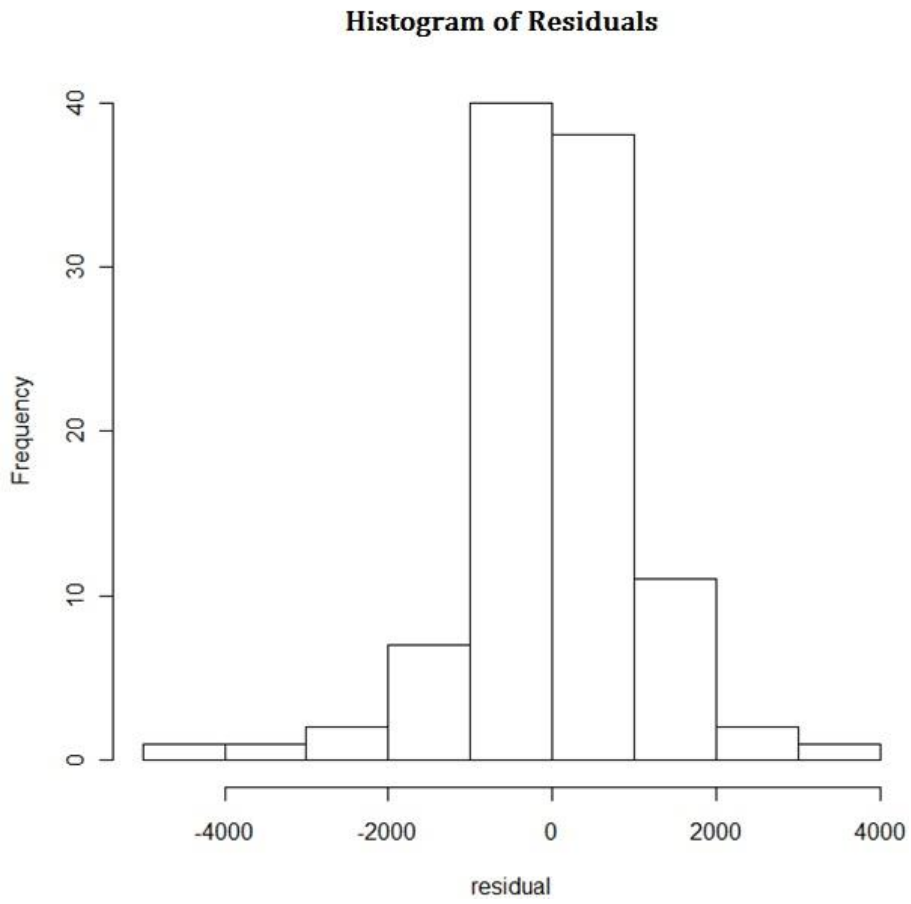

Figure 10: Residual histogram.

The value of the parameter coefficients, standard errors and variance are listed in Table 2.

Table 2: Parameter statistics.

	MA(1)	ppvaliokeskotehojuoma
coefficient	0.5471	0.177
standard error	0.0936	0.014
variance	0.00876096	0.000196

Thus, the final model takes the form

$$\Delta z_t = -0.5471\varepsilon_{t-1} + 0.177x_t + \varepsilon_t, t = 1, \dots, T,$$

or

$$z_t = z_{t-1} - 0.5471\varepsilon_{t-1} + 0.177x_t + \varepsilon_t, t = 1, \dots, T. \quad (25)$$

The original and fitted time series are plotted in Figure 11 to demonstrate the goodness of the fit. The pseudo- adjusted R^2 statistic was also calculated for the goodness of fit [14]. The statistic value is 0.89 which means that the model describes the sales data relatively well.

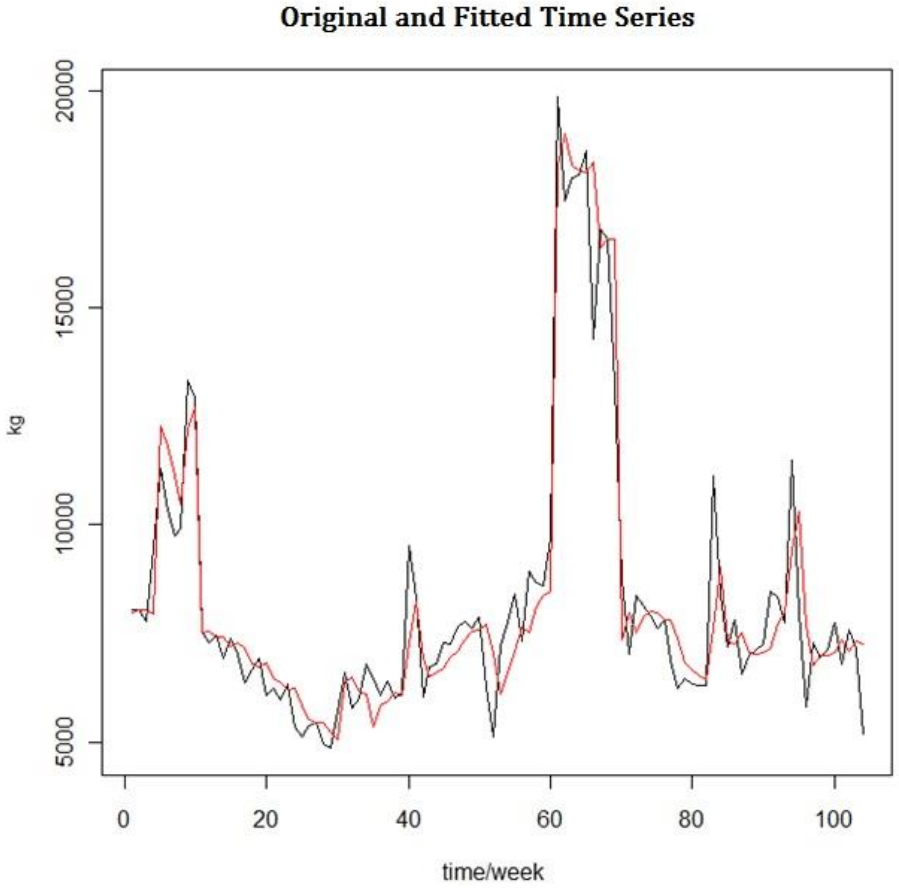

Figure 11: Original (black) and fitted (red) time series.

Backtesting

After fitting the model to the data, the model was used to predict the sales of the 16 next weeks. The forecast was compared to the actualized sales of the 16 next weeks. The ex-post forecast and 95% confidence intervals are plotted with the actualized sales in Figure 12. It can be noticed that the forecast is quit poor since the actualized sales doesn't even stay inside the 95% confidence intervals.

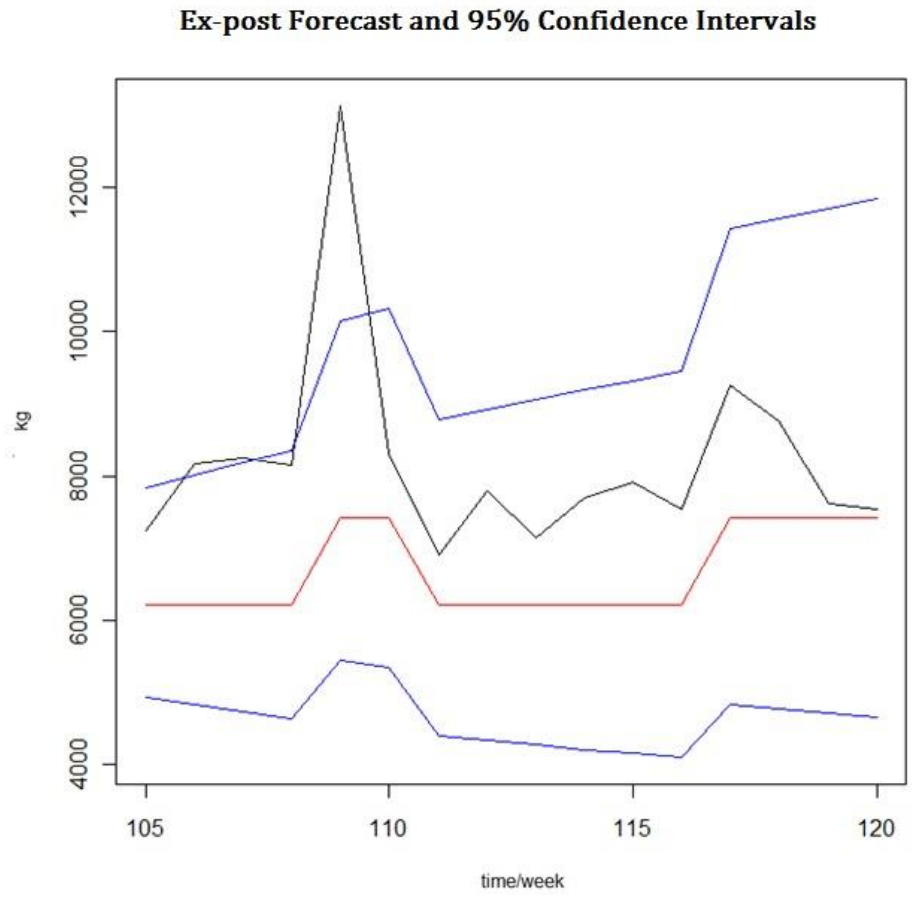

Figure 12: Ex-post forecast (red) with 95% confidence intervals (blue) and the actualized sales (black).

5 Discussion

The aim of this thesis was to model the baseline sales of a large Finnish food company's functional dairy product. The baseline sales level was needed for measuring the effect of marketing and advertising functions. Sales data from the manufacturer to a retail chain and marketing and advertising expenditures for end-customers in the retail chain was available for estimation of the model. Because the variables were equally spaced in time, a time series analysis was used to model the baseline sales level.

The preferable time series model to model baseline sales would have been a VAR model, as it would have been able to take product life cycle into account. That is, the fact that the effect of advertising on sales changes through product life cycle. But, there were difficulties in getting the package in R statistics software to function. Thus a $ARIMAX$ model was used instead. A $SARIMAX$ model would have been more preferable, because of the seasonal nature of the product, but the time series was too short for seasonal differencing of the model.

The Box-Jenkins procedure was used to identify, estimate and validate the model. The best model derived from the Box-Jenkins procedure was $\Delta z_t = -0.5471\varepsilon_{t-1} + 0.177x_t + \varepsilon_t$. The model is a $ARIMAX(0,1,1)$ model, meaning that it is a pure first order Moving Average model that has a first order integrated part. There was no remarkable difference in how well the different $ARIMAX$ models described the data and the model with smallest AIC was chosen. The model described the data relatively well, with a pseudo- adjusted R^2 statistic of 0.89.

After finding the best model, the sales for the next 16 weeks were forecasted and compared to the actualized sales. The forecast was pretty poor, since the actualized sales didn't even fit in the 95% confidence interval of the forecast. The difference in the forecasted and actualized sales can however be explained by the marketing and advertising activities that occurred during the forecast interval. During the forecast interval there was an advertising campaign, which could not be explained by our model, since price promotion is the only external variable in the model.

The interpretation of the model is that a random shock at time $t-1$ affects the difference in sales between t and $t-1$ negatively. This means that the random shock causes a descending stochastic trend in the sales. It can be difficult to state what the random shock represents, but it could be the product life cycle or some competing product that is eating up the market share of this functional dairy product. Obviously, the price promotion of the dairy product affects the difference in sales positively. This means that price promotions create an ascending stochastic trend in the sales. The reason for this could be that customers buy more of the product during price promotion and this simultaneously creates a brand awareness, which makes the customers want to buy of the product in the future, even though the product is then not price promoted.

As can be seen from the ex-post forecast the model is not that reliable in forecasting, but can be used to model baseline sales. When there is a noticeable difference in the actualized sales and estimated sales, the effect can be attributed to a marketing or advertising function. However, it

would be advisable to perform the Box-Jenkins procedure with a longer time series if possible. That way the seasonal effect might also be captured.

6 References

[1] Rust, Roland T., Tim Ambler, Gregory S. Carpenter, V. Kumar, and Rajendra K. Srivastava. "Measuring marketing productivity: current knowledge and future directions." *Journal of Marketing* (2004): 76-89.

[2] Csilla, Horváth, Marcel Kornelis, and Peter SH Leeflang. *What marketing scholars should know about time series analysis: time series applications in marketing*. No. 02F17. University of Groningen, Research Institute SOM (Systems, Organisations and Management), 2002.

[3] Pauwels, Koen, Imran Currim, Marnik G. Dekimpe, Dominique M. Hanssens, Natalie Mizik, Eric Ghysels, and Prasad Naik. "Modeling marketing dynamics by time series econometrics." *Marketing Letters* 15, no. 4 (2004): 167-183.

[4] Hipel, Keith William, Angus Ian McLeod, and William C. Lennox. "Advances in Box-Jenkins modeling: 1. Model construction." *Water Resources Research* 13, no. 3 (1977): 567-575.

[5] Said, Said E., and David A. Dickey. "Testing for unit roots in autoregressive-moving average models of unknown order." *Biometrika* 71, no. 3 (1984): 599-607.

[6] http://www.staff.ncl.ac.uk/oliver.hinton/eee305/Chapter6.pdf (retrieved 1.7.2011)

[7] Box, George EP, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time series analysis: forecasting and control*. 3rd ed. Prentice Hall, Englewood Cliffs, NJ, 1994.

[8] http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics2006.pdf (retrieved 1.7.2011)

[9] Pearson, Egon S. "I. Note on Tests for Normality." *Biometrika* 22, no. 3-4 (1931): 423-424.

[10] Lütkepohl, Helmut, and Fang Xu. "The role of the log transformation in forecasting economic variables." *Empirical Economics* 42, no. 3 (2012): 619-638.

[11] http://www.businessdictionary.com/definition/exogenous-variable.html (retrieved 1.10.2013)

[12] http://www.businessdictionary.com/definition/endogenous-variable.html (retrieved 1.10.2013)

[13] http://www.dtreg.com/TimeSeries.htm (retrieved 1.10.2013)

[14] Long, J. Scott, and Jeremy Freese. "Regression models for categorical dependent variables using Stata." *Stata Press Books* (2006).

7 Appendix

7.1 Summation of the Thesis in Finnish

Markkinointitoimenpiteiden perustelu sisältää niiden myynnin lisäämisen arvioimista. Markkinointitoimenpiteiden vaikutuksen arvioimiseksi täytyy ensimmäiseksi muodostaa estimaatti perusmyynnin tasosta, jonka perusteella luodaan vertailuindeksi myynnille. Tämän estimaatin avulla on myös mahdollista eristää muiden myyntiin vaikuttavien seikkojen, kuten kausivaihtelun ja kilpailijoiden hintakampanjoiden vaikutukset [1].

Tämän työn tarkoitus on mallintaa suuren suomalaisen elintarvikeyhtiön tehojuoman perusmyynnin taso. Tuote on maitotuote, johon on lisätty terveysvaikutuksia. Myönteisten terveysvaikutuksien takia tuotetta ostetaan selkeästi enemmän influenssakausina. Perusmyynnin mallintamiseksi käytetään sekä tuotteen myynti- että mainospanosdataa. Mainospanosdata kuvaa paljonko varoja eri markkinointitoimenpiteisiin on käytetty. Varojen käytön tehokkuutta voidaan mallintaa vertailemalla markkinointitoimenpiteitä toteutuneen sekä arvioidun myynnin erotukseen. Markkinointidata sisältää usein aikariippuvaista dataa, jonka takia aikasarjamallit soveltuvat erinomaisesti kuvaamaan datan muuttujien riippuvaisuuksia. Aikasarjamalleissa mallin muuttujat riippuvat ajasta [2].

Kaksi suurinta ongelmaa myynnin mallintamisessa ovat vaihtelevat dynamiikat muuttujien välillä ja endogeenisyyden ongelma. Eli markkinoiden reagointikyky ei välttämättä ole vakio ajan suhteen. Mainonnan vaikutus saattaa hiipua tuotteen elinkaaren ajan, mikä puolestaan johtaa muuttujien välisten dynamiikkojen vaihteluun tuotteen tarkasteluvälillä [3]. Endogeenisyyden ongelmalla tarkoitetaan, että eksogeeniset muuttujat kuten hinta, hintapromootiot ja mainonta ovat tosiasiassa endogeenisiä. Tilastotieteen määritelmien mukaisesti eksogeeninen muuttuja on muuttuja, joka vaikuttaa malliin ilman, että malli vaikuttaa siihen. Eksogeenisen muuttujan kvalitatiiviset ominaisuudet ja tuottamismenetelmät eivät ole mallinrakentajan määrittämiä [11]. Endogeeninen muuttuja on taas tuotettu mallin sisällä ja täten sen arvo vaihtuu mallin funktionaalisten suhteiden kautta [12]. Markkinointijohtajat asettavat markkinointimuuttujien arvot markkinainformaation pohjalta, joka saattaa osittain olla piilevää tutkijalle mutta kuitenkin vaikuttaa asiakkaiden käyttäytymiseen. Tämä luo tilanteen, missä markkinointitoimenpidemuuttujat voivat olla korreloituneita virhetermien kanssa [2].

Markkinointitutkimuksessa yleisesti käytetty malli on vektoriautoregressiomalli (VAR), mitä käytetään kuvaamaan markkinointitoimenpiteiden vaikutusta myyntiin. Tässä työssä päädyttiin kuitenkin käyttämään $SARIMAX$-mallia, koska sen jatkokehitysmahdollisuudet ovat huomattavasti laajemmat [4, 5].

$SARIMAX$-malli on aikasarjoihin perustuva tilastollinen malli. Kun $SARIMAX$:ia sovelletaan myynnin perustason mallintamiseen, selitettävänä muuttujana on tuotteen myyntitulot. Myyntituloja

selitetään myyntitulojen omalla historialla. Lisäksi myynnin satunnaisshokkien historiallista dataa käytetään selittäjänä. Mallissa on myös mahdollista käyttää ulkoisia muuttujia, kuten mainospanosdataa [2].

Selitettävän muuttujan datan on oltava stationaarista, jotta $SARIMAX$-mallia voidaan käyttää. Stationaarisuus tarkoittaa, että selitettävän muuttujan aikasarjan keskiarvo sekä varianssi ovat aina vakioita. Toisin sanoen aikasarjan tilastolliset ominaisuudet pysyvät vakiona ajan suhteen. Malli ottaa myös huomioon stokastiset trendit myynnissä sekä myynnin kausivaihtelut. Datan stokastinen trendi voidaan ottaa huomioon differentioimalla data, eli laskemalla erotus peräkkäisten aika-askelien väliltä [2].

$SARIMAX$-mallin soveltaminen dataan ei kuitenkaan ole täysin triviaalia, vaan se suoritetaan ns. Box-Jenkins menetelmän avulla. Box-Jenkins menetelmä voidaan jakaa kolmeen vaiheeseen; mallin identifiointi, estimointi sekä validointi. Identifioinnin tarkoitus on muokata data stationaariseksi sekä määritellä $SARIMAX$-mallin parametrien lukumäärä. Datan stationaarisuutta voidaan tutkia ns. täydennetyllä Dickey-Fullerin testillä. Parametrien lukumäärä taas saadaan määriteltyä tutkimalla muuttujien korrelaatiofunktioita. Korrelaatiofunktioista nähdään millä viiveellä muuttujat korreloivat itsensä ja toisten muuttujien kanssa [4].

Kun mallin parametrit on identifioitu, ne estimoidaan suurimman uskottavuuden menetelmällä. Suurimman uskottavuuden menetelmässä käytetään tilastollista otantaa ja arviota siitä, mistä parametrisoidusta jakaumasta kyseinen otanta on otettu. Suurimman uskottavuuden menetelmä estimoi jakaumalle sellaiset parametrit, että otanta on otettu suurimmalla todennäköisyydellä kyseisestä jakaumasta [4].

Mallin validointi koostuu lähinnä residuaalien tutkimisesta. Residuaalien kuuluisi olla riippumattomia, homoskedastisia sekä normaalijakautuneita. Homoskedastisuus tarkoittaa, että residuaalien varianssi ja keskiarvo ovat vakioarvoisia. Lisäksi tarkastellaan parantaako lisäparametrien lisääminen malliin huomattavasti mallin tarkkuutta [4].

Työssä käytettiin $SARIMAX$-mallia tehojuomatuotteen myynnin perustason mallintamiseen. Myyntitulodatana käytettiin yhtiön myyntituloja jälleenmyyjille. Huomattavaa on, että myyntitulodatana ei siis käytetä määrää, joka loppukäyttäjälle on myyty. Koska tarkastelukohteena on tässä tapauksessa maitotuote, jonka hyllyikä on hyvin lyhyt, voidaan olettaa, että jälleenmyyjät osaavat mitoittaa tilauksensa niin, että se vastaa suhteellisen hyvin loppukäyttäjän kulutusta. Tällöin jälleenmyyjiltä saadut myyntitulot toimivat hyvänä approksimaationa loppukäyttäjiltä saadusta myynnistä.

Tuotteen myynnissä esiintyi selkeää kausivaihtelua, mutta sitä ei pystytty mallintamaan datan lyhyen aikajakson takia. Ulkoisiksi muuttujiksi oli mahdollista valita tuotteen mainospanosdata eri medioissa, tuotteen hintatiedot, tuotteen hintapromootiotiedot ja substituuttituotteiden markkinointitoimenpidedata. Korrelaatiomatriisin perusteella kuitenkin huomattiin, että tuotteen hintapromootiodata on yksinään tarpeeksi merkittävä selittämään myyntituloja.

Myyntitulodata muutettiin stationaariseksi ja sen korrelaatiofunktiosta pääteltiin, että sopivin $SARIMAX$-malli olisi sellainen, jossa myyntituloja selitettiin edellisen viikon myyntituloilla,

satunnaisshokilla ja tuotteen hintapromootiodatalla. Mallin validointivaiheessa todettiin, että tämä oli oikea päättely.

Parhaaksi $SARIMAX$-malliksi saatiin $\Delta z_t = -0.5471\varepsilon_{t-1} + 0.177x_t + \varepsilon_t$, missä Δz_t on myyntitulojen erotus ajanhetkillä t ja $t-1$, ε_t on myynnin satunnaishokki ajanhetkellä t ja x_t on mallin ulkoinen muuttuja, joka tässä tapauksessa on tuotteen hintapromootiodata. Mallin tulkinta on se, että myynnin satunnaisshokki ajanhetkellä t vaikuttaa myynnin erotukseen aikavälillä t ja $t-1$ negatiivisesti. Tämä tarkoittaa sitä, että satunnaisshokki aiheuttaa laskevan stokastisen trendin myyntituloissa. On vaikea sanoa, mistä satunnaishokki johtuu. Se voi esimerkiksi johtua tuotteen elinkaaresta tai siitä, että kilpaileva tuote syö tuotteen markkinaosuutta. Ilmeisesti hintapromootio vaikuttaa tuotteen myyntiin positiivisesti. Tässä tapauksessa se kuitenkin vaikuttaa myös myynnin erotukseen aikavälillä t ja $t-1$ positiivisesti. Tämä tarkoittaa sitä, että hintapromootiot luovat nousevan stokastisen trendin myyntituloihin. Selitys tähän voisi olla, että hintapromootiot saavat asiakkaat ostamaan enemmän tuotetta, joka taas lisää bränditietoisuutta, mikä saa asiakkaat ostamaan lisää tuotetta tulevaisuudessa.