Modelling Swedish Inflation Using Market Data

Yang Zhou
yangzhou@kth.se

May 12, 2017

Modelling Swedish Inflation Using Market Data

Abstract

This study is an attempt to model Swedish CPI inflation using ARIMA and variations of distributed lag model with market data as explanatory variables. The model will be constructed on the CPI subcomponents level and the results are aggregated to the CPI. Three approaches are tested in this report. In the first approach, only ARIMA model is used to model each of the subcomponents. In the second approach we use a distributed lag model (DLM) on subcomponents with significant correlation to the market data, the residual of the DLM is then modelled using ARIMA. In the third approach we use an restricted finite distributed lag model (RFDLM) instead of DLM. The study found that RFDLM was the best approach to model inflation with 20% RMSE compared to 32% of the naive forecast. However, there is little forecast potential using this approach due to the short lag of market data used as input. The model would be most useful in testing CPI inflation scenarios using predictions or assumptions of market data as input.

Modellering Av Svensk Inflation Med Marknadsdata

Abstrakt

Denna studie är ett försök att modellera svenska inflation genom att använda ARIMA-modell och variationer av distributed lag model med marknadsdata som förklarande variabler. Modellen är konstrukterad på underkomponents nivån av KPI och sedan aggregerad till KPI. Tre metoder prövas i denna studie. I första metoden modelleras underkomponenterna direkt med ARIMA-modeller. I andra metoden används distributed lag model (DLM) på underkomponenter med signifikant korrelation till marknadsdata, residualen från DLM modelleras i sin tur med ARIMA-modeller. I den tredje metoden ersätter vi DLM med restricted finite distributed lag model (RFDLM). Resultaten från studien visar att RFDLM är den bästa metoden att modellera inflationen och hade ett RMSE på 20%. Detta jämfört med den naiva prognosen som hade en RMSE på 32%. Däremot har RFDLM inte särskilt mycket praktiskt nytta i prognostisering av inflationen då man behöver marknadsdatan för prognosperiod i förhand på grund av att modellen använder sig av väldigt korta lagg. Däremot skulle modellen kunna ha nytta i scenario byggande med prognoser and antagande på marknadsdatan som input.

Contents

Chapter 1

Introduction

1.1 Background

This report was written as a Master's thesis for 30 ECTS at the end of a two year Master's Program at KTH. The report documents a statistical modelling project of the Swedish inflation using distributed lag model (DLM) and ARIMA model. The data has mainly been provided by the Macro Research team at Nordea which also has provided with ideas and knowledge regarding inflation measuring. Assistant professor Thomas Önskog at KTH has also provided mathematical insights and ideas to the project.

1.2 What is CPI?

Inflation is a variable that is seen in all investment decisions made by pension funds, risk capital funds, corporations and even households. In some cases and for some asset classes (such as fixed income), inflation is often seen as one of the most important factors to consider and could directly impact the asset's valuation. The most common way to measure inflation is using the Consumer Price Index (CPI), a largely internationlly standardized metrics computed by the country's statistical agency. In Sweden, the CPI statistics is computed by Statistics Sweden (SCB) and is released on a monthly basis [1].

Consumer Price Index tracks the change in goods and services sold to private domestic consumers and should reflect the price consumers actually paid. That means the index should also reflect the cost for households to maintain a constant stan-

dard of living, which sounds simple but is difficult to meassure. The basic idea is to construct a basket of goods that reflects what an average consumer purchase each month and measure the price change of these products on a monthly basis [1].

SCB classifies the CPI basket in 12 main groups of products and services, ranging from food, restaurants to housing. The 12 main groups are in turn divided into 44 subcomponents which can be further split into over 350 product groups. For example the main group "Food and non-alcoholic beverages" include among other, "Coffee, tea and cocoa" which in turn include the product groups "coffee", "tea" and "cocoa" with a price index for each. On the lowest level, product group is aggregated through measurement of actual price of various product in the category from different vendors and locations. The indices for product group is also referred to the elementary indices as it is the building block to aggregating CPI. See Figure 1.1 for an illustration of the decomposition of CPI.

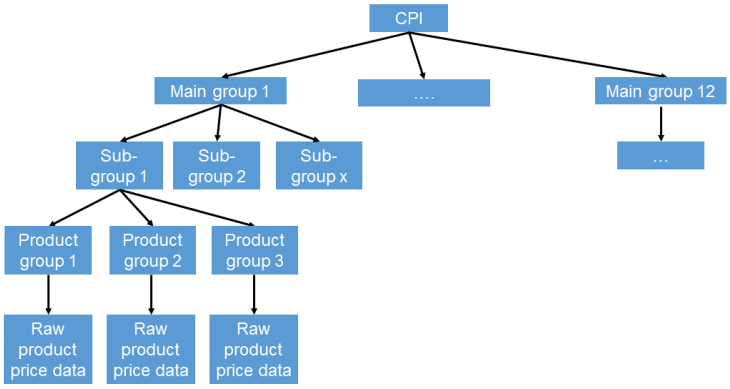


Figure 1.1: Illustrative decomposition of the CPI

As the usage of CPI has changed over the years, the methodology has undergone changes as well and most recently in 2005 and 2012. CPI is used today by many type of organisations such as government institutions, labour unions, central bank, investors and enterprises. It can be referred to in wage discussions, investment evaluations and many more situations so there is no doubt that CPI is one of a key economic metrics of every country[1, 2].

1.3 Purpose

This report will attempt to model CPI by modelling the subcomponents using a combination of time series methods such as ARIMA together with econometric regression methods for lagged correlation with explanatory variables, such as distributed lag models (DLM). The explanatory variables will mainly consist of easy to access market data that is available. The idea is that this model could potentially be used in two ways.

The first is to use the model to predict inflation. However, this will largely be limited by the shortest lag used for the market data as input to the DLM. In this study, we will be selecting lags simply based highest correlation, if the best lag turns out to be zero, we will not be able to predict inflation beyond zero lag, current inflation level. In other words, this would be Nowcasting.

Predicting inflation is known to be difficult and there is a good chance that the model used in this approach would not make any significant improvement to the current methods used by economist. However, a second and maybe more important use of the model would be to, a tool for scenario analysis based on market events. For example, market makers in Sweden typically use simple rule of thumbs that a 10% drop in oil prices corresponds to a 0.1% drop in inflation two years later, which was concluded by a study conducted by the Riksbank in 2008. [14] If the market makers had model that could more accurately indicate how inflation is moving in real time than a rule of thumb or a naive guess, they could maybe price financial products related to the level of inflation more accurately and be able to price as well as hedge their positions better.

Thus, another important goal of this report is to create a model that is modularized and thus easy to maintain. The mode should be able to be updated in live time with automated input data that could be found through Bloomberg and Reuters.

Five explanatory variables have been used in this model, but the concept can easily be expanded to include many more explanatory variables.

1.4 Previous Models

Inflation is one of the main economic indicators for a country and a very complex one to model. There are no universal models that fits every country due to the difference

in economical structure. For example a country with larger amount of imports would be more affected by their currency rate and a country with large domestic market would be more affected by their labour market situation. However, in general there are mainly two categories of approaches, a top-down and a bottom up approach.

The most straightforward and easiest way would be to model inflation using macro variables and thus a top-down approach, Philip's curve being one of the more common ones. Philip's curve states that the is a historical inverse correlation between the rate of unemployment and inflation. Therefore, a decreasing unemployment rate will lead to a higher inflation rate. The model has been modified many times since the original publications of William Philips in 1958, adding the rule of the "money illusion" measured by inflation expectation and also the so called long-run Philip's curve respective short-run Philip's curve. However, many economists have started to questions if the relationship between unemployment rate and inflation has broken up due to the implications of the modern economical structure with supply side policies that allow economies to expand without inflation [3, 4, 5].

The bottom-up approach is easy in theory but is usually not very practical to implement. In reality, the measurement of CPI is in itself is a bottom-up approach to estimate the "real" inflation in the economy from a consumer's point of view by tracking prices of products consumer buy. Statistic Sweden has a large team collecting price data on nearly everything in the economy, assembling them together for the CPI data release each month. If one wish to make a perfect model of the CPI, then the most accurate method would be to copy Statistics Sweden methods and collect various price data to create a "shadow CPI". However, in reality, this would be very resource consuming and not very practical. Given the increasing amount of e-commerce activities, the amount of accessible data have increased tremendously and one could in theory automate the price collection for creating a shadow CPI with web-scrapping algorithms. This has been done to a certain degree by Nordea on a few products such as fuel price, electricity and food. In reality, forecasters will use multiple approaches in order to triangulate the best forecast possible by combining top-down and bottom-up approaches. Any improvements on existing models will be highly demanded.

A study from Yu-chin Chen, Stephen J. Turnovsky, and Eric Zivot tried to predict inflation in five commodity producing countries using commodity prices and least-angle regression and generalized autoregressive distributed lag model and compared it to a AR(1) process. They found that commodity prices did indeed outperform the

AR(1) process, although with only modest improvement in some cases. [16]

Chapter 2

Mathematical methods description

This chapter will provide the reader with a mathematical description of the models used in this project. We will cover the two main models used in this report, DLM, ARIMA models, and their variations. The methodology of aggregating the CPI itself will also be covered in this section.

2.1 Distributed lag model

A distributed lag model (DLM) is a dynamic model that describes the effect on the dependent variable y through an explanatory variable x where the effects occurs over time rather than immediately. The distributed lag model in the finite form can be written as

$$y_t = \alpha + \sum_{l=0}^{q} \beta_l x_{t-l} + \epsilon_t, \tag{2.1}$$

where α is the intercept, β_l is the coifficient for lag l and ϵ is the error term. This is very similar to the MA-model described later in Equation (2.5), with the difference being that the white noise ϵ_{t-l} in the MA-model has been replaced by a explanatory variable x_{t-l}.[6]

In theory one can model a temporary change in x to a permanent change in y, but this would require q to be infinite and β_l to be larger than 0 for all lags, creating an un-stationary relationship between x and y. Therefore one important factor to consider is that x and y both need to be stationary and have the same level of differentiation. In other words, if x measures a change then y should mea-

sure a change as well, if x is a nominal level then y need to be a nominal level as well.

For example if we would model the effect on inflation with the year on year change of oil price at year t denoted as x_t, x_t would be positive for one year and 0 for the subsequent years if oil prices stays stable. If y_t is the inflation index level, β in Equation (2.1) would need to be non-zero and extend into the indefinite future. One approach to fix this would be to redefine x_t to be the price level of oil prices instead of year on year change, the problem is then solved as x_t would be permanent higher with a price increase. A second and maybe easier approach would be to redefine y_t to the change in inflation instead of index level, the problem would be solved as the increase in y_t would only be temporary as well and the number of terms in Equation (2.1) would be less.

There are some disadvantages with the distributed lag model in the finite form. The first problem is multicollinearity, as x can be highly autocorrelated even if it is stationary. This means for example if we would estimate y_t by $y_t = \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-3}$, the lag weights β will be bouncing between positive and negative values and might not be statistically significant, which is inconsistent with economic theories. This is especially common for economic and financial raw data. [6]

A second problem is that the number of observations available for estimation observations would drop quickly with increasing lag length. If we have T observations of data, the data that can be used for estimation is $T-q$ as we need q periods of data in the beginning before starting to estimate the data. This would allow us $T-2q-2$ degree of freedom, assuming that intercept need to be estimated as well. For each lag we add to the model, the degree of freedom would be reduced by 2. [6]

This means that the distributed lag model is a suitable method only when the lag coefficients β decline to zero quickly for each lag, the predicting variables are not highly correlated and the time series of available observation is much longer than the lag length q. In other words, this model can almost never be used as there are nearly no economic or financial data that satisfies all these conditions. However, a possible solution to the problem is to use Restricted Finite DLM. [6]

2.1.1 Restricted Finite DLM

Restricted finite distributed lag models (RFDLM) would solve the two disadvantages of multicollinearity and decreasing number of observations described earlier.

The model is based on the idea that β_l should be a smooth function of the lag l. Even if the lag weights β does not follow the smoothness naturally if regression is made on the variables independently, one could impose it by restricting β_l, thus solving the multicollinearity problem. By predefining the smoothness of the β_l as a function, we can reduce the number of lag weights that need to be estimated to one, as shown in the example of Equation (2.2), thus solving the second problem of the unrestricted model as well [6].

One example of a restriction that could be implemented is simply the linear one. Redefine β_l as the following

$$\beta_l = \frac{q+1-l}{q+1}\beta_0, \quad l = 1, 2, \ldots, q. \tag{2.2}$$

Combining Equation (2.1) and Equation (2.2) gives us the RFDLM on the form

$$y_t = \alpha + \beta_0 \sum_{l=0}^{q} \frac{q+1-l}{q+1} x_{t-l} + \epsilon_t. \tag{2.3}$$

We see that in Equation (2.3) only α and β_0 need to be estimated given that q has been selected, thus increasing the number of observations for estimation and simplifying the computation.

The same approach can be used for other types of restricting functions on B_l. Some other common functions are quadratic and "tent" shaped model.

2.2 Auto Regressive Moving Average models

This section will cover the different variations of Auto-Regressive-Moving-Average (ARMA) models used. We will start by giving a very brief description of ARMA, as more details can be found in many textbooks or articles on time-series analysis. ARIMA model and Seasonal ARIMA (SARIMA) will be covered later on in further detail as they are used to a large extent in this report.

2.2.1 ARMA model

The ARMA model itself can be defined as the combination of the autogressive (AR) model and moving average (MA) model. Assuming no constants, the AR model can

be written as: [8]

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + z_t, \tag{2.4}$$

where p is the order of the AR-model, ϕ_i are parameters and z_t is the white noise. In this case, we can model the white noise z_t with a Moving-Average model assuming no no constants

$$z_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \tag{2.5}$$

where q is the order of the MA-model, θ_i are the parameters and ε_i is the white noise error term. Combining Equations (2.4) and (2.5) gives us the $ARMA(p, q)$ model on the form

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \tag{2.6}$$

where ε_t is white noise, denoted $\varepsilon_t \sim WN(0, \sigma^2)$, that satisfy the following conditions

1 Every variable in $\{\varepsilon_t\}$ has zero mean, $E(\varepsilon_t) = E(\varepsilon_{t-1}) = \ldots = 0$

2 Every variable in $\{\varepsilon_t\}$ has a constant variance, σ^2, $E(\varepsilon_t^2) = E(\varepsilon_{t-1}^2) = \ldots = \sigma^2$

2 $\{\varepsilon_t\}$ is serially uncorrelated, $E(\varepsilon_t \varepsilon_{t-s}) = E(\varepsilon_{t-1} \varepsilon_{t-s-1}) = \ldots = 0$ for all s

This also means that the white noise process does not have to be i.i.d as it only requires the time series to be uncorrelated but it does not have to be independent. However, by definition a i.i.d noise is also a white noise.

The estimation of the parameters in a ARMA-model is usually done through Maximum-Likelihood (ML) methods, but there are also analytical solutions for certain model orders. Also, when q is zero, ARMA$(p, 0)$ is simply the AR(p) model which can be estimated through Ordinary Least Square (OLS). The problem becomes simply selecting the parameters $\{\varepsilon\}$ which minimize the sum of squared white noise. In the case of AR(1) with intercept, we have $X_t = c + \phi X_{t-1} + \varepsilon_t$, which can be estimated by $\phi = 1 - \frac{c}{\mu}$, where $\mu = E(X_t)$ due to stationarity. When MA-terms are included, OLS can no longer be used as the ε white noise is not directly observed. Therefore Maximum-likelihood is a common method used to estimate the parameters

in ARMA models.

If we assume that the white noise $\{\varepsilon\}$ follows a normal distribution with mean zero and variance σ^2, the likelihood function L_t for ε_t is defined as

$$L_t = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\varepsilon_t^2/2\sigma^2},$$

since the observations of $\{\varepsilon\}$ occurs independently, the joint likelihood function is

$$L = \prod_{i=1}^{T} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\varepsilon_t^2/2\sigma^2},$$

Taking the natural logarithms on both sides, we get the log likelihood function

$$\ln(L) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}\varepsilon_t^2, \tag{2.7}$$

this function is then maximized by selecting the optimal parameters. [9]

The simplest example of non linear estimation is the MA(1) process given by

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1},$$

which can also be written with the lag operator $Ly_t = y_{t-1}$ as

$$\varepsilon_t = X_t - \theta_1\varepsilon_{t-1} = X_t - \theta_1 L\varepsilon_t.$$

Solving for ε and using the formula of a geometric series, we get

$$\varepsilon_t = \frac{X_t}{1+\theta_1 L} = \sum_{i=0}^{\infty}(-\theta)^i X_{t-i}. \tag{2.8}$$

In reality, the sum in Equation (2.8) needs to be truncated at $t-1$ and using maximum likelihood estimation from Equation (2.7) we get the loglikelihood function

$$\ln(L) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}\Big(\sum_{i=0}^{t-1}(-\theta)^i X_{t-i}\Big)^2, \tag{2.9}$$

where we maximize by choosing appropriate θ and σ. Even though $MA(1)$ is the simplest example, it is difficult to solve this problem analytically. Numerical methods are usually used to maximize the likelihood function [7]. For this study, all estimation was done through R using ML method.

2.2.2 ARIMA Model

The Autoregressive Integrated Moving-Average model (ARIMA) is an ARMA model that has been differentiated to the point where the serie has become stationary. To extend the ARMA model into an ARIMA model, we simply need to differentiate X_t, d times. We can define this using the difference operator $\Delta X_t = X_t - X_{t-1}$ and rewrite Equation (2.6) to

$$\Delta^d X_t - \sum_{i=1}^{p} \phi_i \Delta^d X_{t-i} = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

This can also be represented using the backward operator B where $B^n X_t = X_{t-n}$ as

$$\phi(B)\Delta^d X_t = \theta(B)\varepsilon_t, \tag{2.10}$$

where

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p,$$
$$\theta(B) = 1 + \theta_1 B + \ldots - \theta_q B^q.$$

Also, X_t and ε_t need to be stationary for the model to become an ARMA-process. The stationary requirement is fulfilled for X_t if $1 - \sum_{i=1}^{p} \phi_i r^p$ is outside the unit circle and for ε_t if $\sum_{i=1}^{q} \theta_i^2$ and $(\theta_1 + \theta_1\theta_{s+1} + \theta_2\theta_{s+2} + \ldots)$ are both finite for all s. [8]

In practice, the number of differentiations d is determined with the Kwiatkowski Phillips Schmidt Shin (KPSS) test, which is essentially what we described above. The KPSS test describes $\{X_t\}$ as:

$$X_t = r_t + \beta t + \varepsilon_t,$$

where r_t is a random walk, βt is a deterministic trend and $\varepsilon_t \sim WN(0, \sigma^2)$ is the error term, the error will be stationary if $\sigma^2 = 0$. The test statistics is then given by the Lagrange Multiplier for testing $\sigma^2 = 0$ against the alternative that $\sigma^2 > 0$ defined as

$$\text{KPSS} = (\sum_{t=1}^{T} S_t)/\hat{\sigma}^2 \tag{2.11}$$

Where $S_t = \sum_{i=1}^{t} u_i$ where u_t is the residual of the regression on X_t and $\hat{\sigma}$ is the estimated variance of the error. The time-series is differentiated until this test is passed and by replacing it with a new variable $X_t^* = \Delta^d X_t$, a standard ARMA parameter estimation can be performed on X_t^*. [10]

2.2.3 Seasonal ARIMA

A seasonal component could be added to the ARIMA model and defining the new model as ARIMA$(p,d,q)(P,D,Q)_m$ where P, D and Q are the order of the AR-model, the differentiation and the MA-model respectively for the seasonal part. m is the number of periods per season. With Seasonal ARIMA, we can then simply include any potential seasonal components to the model directly through ARIMA modelling instead of adjusting the raw data for seasonal components beforehand.

The seasonal part of the model is very similar to the non-seasonal components of the model, the only difference is that the backshift operator moves in step of m instead of one. For example, seasonal AR(2) model would become $X_t = \phi_1 X_{t-m} + \phi_2 X_{t-2m}$ where m is the period of a season.[10] The seasonal ARIMA$(p,d,q)(P,D,Q)_m$ can be built on Equation (2.10) and is then written as

$$\Phi(B^m)\phi(B)\Delta_m^D\Delta^d X_t = \Theta(B^m)\theta(B)Z_t, \tag{2.12}$$

where

$$\begin{aligned}\Phi(B^m) &= 1 - \Phi_1 B^m - \ldots - \Phi_P B^{Pm}\\ \phi(B) &= 1 - \phi_1 B - \ldots - \phi_p B^p\\ \Theta(B^m) &= 1 + \Theta_1 B^m + \ldots + \theta_Q B^{Qm}\\ \theta(B) &= 1 + \theta_1 B + \ldots + \theta_q B^q.\end{aligned}$$

The procedure and method for order selection of Seasonal ARIMA is almost the same as ARIMA except from the fact that the order and paramters of the seasonal AR, differentiation and MA parts need to be selected as well.

2.2.4 Model evaluation: AICc

When using ARMA(p,q) model, where we simply estimate $p, q, (\phi,\theta)$, and σ^2 through ML estimation, we will most likely get very high values of p and q which would give us a model that fit the data very well but has low degree of freedom and is less useful as a predictor. Therefore a penalty factor could be introduced for high numbers of p and q to reduce the risk of over-fitting models. One known such criterion is the "Akaike's Information Criterion biased-Corrected" (AICc) where p, q and (ϕ_p,θ_q) are chosen to minimize

$$\text{AICc} = -2ln(L(\phi_p,\theta_q)) + \frac{2(p+1+1)n}{n-p-q-2}, \tag{2.13}$$

where L is the ML estimate. The AICc is rather easy to compute and simplifies ARMA analysis when having a large amount of data [8].

However, AICc is not the perfect indicator for model selection in all situations as the best ARMA(p,q) according to AICc might not always be the best model in reality. There are other criterion as well such as the Bayesian information criterion (BIC) and the Minimum description length (MDL) criterion. In this report, AICc will be used as the main selection criterion.

2.2.5 Model evaluation: Ljung-Box

The residual from the ARIMA model can be checked to make sure that it behaves like white noise with the Ljung-Box statistical test. The test checks if the autocorrelation of the time series is different from zero. The test statistic is defined as

$$Q = n(n+2)\sum_{l=1}^{h}\frac{\rho_l^2}{n-l}, \tag{2.14}$$

where n is the sample size, ρ_l is the sample autocorrelation at lag l and h is the number of lags being tested. The calculated Q value is then compared to the chi-squared distribution and the hypothesis that the series is white noise is rejected if

$$Q > \chi^2_{1-a,d},$$

where a is the significance level and d is the degree of freedom, which is set to be $h - p - q - P - Q$ since the seasonal ARMA components of P and Q need to be included as well [8].

In practice, the calculation is done in R where the function $Box.text()$ is used to calculate the significance level a given the residual from the SARIMA model, p, q, P, Q and the number of lag tested is set to 24. By setting the required level of significance of atleast $a = 0.05$, t our calculated significance level from R a^* need to be $a^* > 0.05$ in order for the series to be white noise.

2.2.6 Model evaluation: RMSE

Root-mean-squared-error is a common way to evaluate the model created for the estimator. It is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2},$$

where $\hat{y}$ is the estimated value, y is the actual measured value and n is the number of observations. In practise, residuals of the ARIMA parameter estimation is equal to $y - \hat{y}$ and can therefore be used directly in the calculation of RMSE.

Normalized RMSE (NRMSE) is also used in many situations, since it makes it easier to compare the results between different series with different scales, as in this study. As we will be modelling the changes in inflation, we will be normalizing RMSE based on the difference between the maximum and minimum observed value instead of using the mean.

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}}. \tag{2.15}$$

In this report, we will only be using Normalized RMSE to track the improvement of the model using various approaches for both forecast and fitted values. From now on, we will refer to it simply as RMSE.

2.3 CPI subcomponents and aggregation

The aggregation of the CPI itself has some complexity to it and is therefore explained in this section.

2.3.1 Aggregation of CPI

The CPI is aggregated using in many layers. Jevons and Dutot index are used on the most elementary layer where the actual price is aggregated to the elementary indices, i.e the indices on product group level. For example, for the product groups "coffee" includes prices for several different coffee types from different shops that is aggregated using Jevons index [2]. The Jevons price index P_j is a geometrical mean

of the measured price p at time t for item i compared to the price for base period 0. N is the number of items

$$P_j = \prod_i^n (\frac{p_{i,t}}{p_{i,0}})^{1/n},$$

while the Dutot index is defined as mean price p for a selected product i at time t divided by mean price in time 0

$$P_d = \frac{\frac{1}{N}\sum_i^N p_{i,t}}{\frac{1}{N}\sum_i^N p_{i,0}} = \frac{\sum_i^N p_{i,t}}{\sum_i^N p_{i,0}},$$

The Dutot formula should not be used in heterogeneous item groups, such as electronics, where the products change frequently. In Sweden, the Dutot price index is only used on municipal services for home owners, such as water and sewerage. The Jevons index is used for all other indices. This is the first level of aggregation of inflation and the result from Jevons index and Dutot index are often referred to as the elementary aggregate indices, which is the elementary building block for aggregating CPI [2][11].

For the macro-level aggregation the Walsh and Laspeyres index formulas are used to build the yearly index chains that is used for aggregating the entire CPI index. The yearly index chains I_t^{t+1} is defined as an index where the value is 100 at the beginning of year t and I_t^{t+1} at the beginning of year $t+1$. To construct the yearly index chains, index of each product group $I_{t,g}^{t+1}$ for Walsh index or $I_{t,g}^{t+2,m}$ for Laspeyres index is also needed. These indices are calculated using so called elementary aggregate which is compiled using the Jevon price index and will be denoted as $I_{t,dec,g}^{t+1,m}$ (i.e index of price change for product g from December year t to year $t+1$ at month m) using the following formula:

$$I_{t,g}^{t+1} = I_{t-1,dec,g}^{t,dec} \frac{\sum_{m=1}^{12} I_{t,dec,g}^{t+1,m}/12}{\sum_{m=1}^{12} I_{t-1,dec,g}^{t,m}/12},$$

$$I_{t,g}^{t+2,m} = \frac{I_{t-1,dec,g}^{t,dec}}{\sum_{m=1}^{12} I_{t-1,dec,g}^{t,m}/12} I_{t,dec,g}^{t+1,dec} I_{t+1,dec,g}^{t+2,dec},$$

The Walsh index $\hat{I}_t^{t+1}$ is defined as the following

$$\hat{I}_t^{t+1} = \sum_g W_{w,g}^t \times I_{t,g}^{t+1},$$

Where $I_{t,g}^{t+1}$ is the index for the product group g and $W_{w,g}^{t}$ is the Walsh weight for each product group g and should sum up to one. The Walsh weight for the product group g is defined as

$$W_{w,g} = \frac{\sqrt{U_g^t \times U_g^{t+1}/I_{t,g}^{t+1}}}{\sum_{g'} \sqrt{U_{g'}^t \times U_{g'}^{t+1}/I_{t,g'}^{t+1}}},$$

where U_g^i is the consumption value for product group g during the year t and g' denotes all product groups. The idea for the weight adjustment is that, as the price of an product increase/decrease, the real weight of the product during the year will increase/decrease as well. This effect is adjusted in the Walsh weight by using the consumption value at the year t and $t+1$ and the price change for the product given by $I_{t,g}^{t+1}$.

The Laspeyres index $\tilde{I}_{t-2}^{t,m}$ is used to calculate the last part of the index due to lack of consumption data and calculate index up to a monthly basis. Laspeyres index is defined as the following

$$\tilde{I}_t^{t+2,m} = \sum_g W_{L,g} \times I_{t,g}^{t+2,m},$$

where $I_{t,g}^{t+1,m}$ is the index chain index and $W_{L,g}$ is Laspeyres weight for each product group g and should sum up to one. Laspeyres weight only uses the consumption weight at year t is defined as

$$W_{L,g} = \frac{U_g^t/I_{t,g}^{y,m}}{\sum_{g'} U_{g'}^t/I_{t,g'}^{y,m}},$$

Walsh and Laspeyres index is used to build the actual aggregated CPI index $I_{1980}^{y,m}$ going from year 1980 to year y and month m. Laspeyres index is used for the chain for the last 2 years as the consumption value U_g is calculated based on the GDP which is released later than inflation while the remaining index is calculated using Walsh index [2]. The CPI $I_{1980}^{y,m}$ is then defined as

$$I_{1980}^{y,m} = \prod_{t=1980}^{y-2} \hat{I}_t^{i+1} \times \tilde{I}_{y-2}^{y,m},$$

Chapter 3

Data

This chapter will describe the data used in this study and also discuss the data quality and its applicability in the context of this study.

3.1 CPI data

The most important data in this study is of course the underlying CPI data used. All the CPI data are originally from Statistics Sweden. Inflation index data for the sub groups were downloaded directly from Statistics Sweden's data bank. Data on inflation weights were, however, retrieved from Macrobond, a database software that have compiled macro data from almost every country. Long term historical data on sub group weights were only partly available from Statistics Sweden's data bank.

For this study, the monthly CPI index of the 44 subcomponents between March 1996 and February 2016, will be used resulting in 240 data points for each subcomponent. The reason for not using older data is that the CPI index has gone through structural changes with some subcomponents being removed or added.

Out of the 240 data points, (equivalent to 20 years of data) 18 years of CPI data will be used in constructing the models and the last 2 years will be used to benchmark the accuracy of the forecast.

Some of the subcomponents' historical weight data prior to 2001 were missing and have been assumed to be constant in this study. Since we will be modelling each subcomponent individually, this would only affects the accuracy of the aggregated CPI historically and does affect the forecast. The accuracy of the aggregated CPI

using subcomponents and the approach described in Chapter 2.3.1 is shown in Figure 3.1. As shown, the results from the aggregated CPI assuming constant historical weights and the actual CPI are very similar. From here on when referring the the CPI index, we will be referring to the aggregated index.

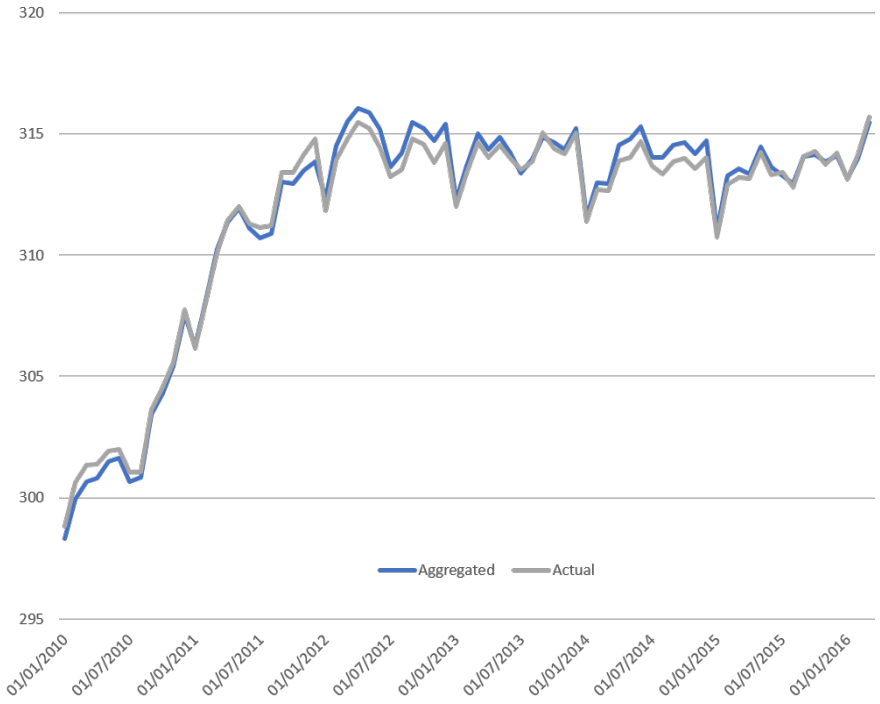


Figure 3.1: Aggregated CPI index vs actual CPI index

The methods of measuring and aggregating price data have changed multiple times historically. The latest change was made in 2005 where the price change during a year chain was changed to reflects the average price change instead of change from December to December. It is rather unlikely that this change would have a considerable impact on forecasts.

3.2 Market data

The market data used in this study are retrieved from some of the most frequently used financial data providers namely Reuters, Bloomberg and Macrobond. For this

study, time series for five different types of market data was selected but this could easily be expanded to include additional data.

3.2.1 KIX currency index

Currency is known to affect the CPI due its direct effects on the price of imported goods. As the Swedish Krona weakens it will become more expensive to import. Therefore CPI will increase as a consequence of more expensive imported goods. A large portion of the product measured in CPI is imported, such as food, electronics and other products. It would therefore be wise to include a currency indicator to a CPI model. However, not all product would be directly affected by changes in currency. Some companies might have hedged their currency exposure and therefore effects of large currency changes might be delayed and small currency fluctuations might have no affect at all on the CPI.

KIX is The Riksbank's weighted currency index for the Swedish Krona. The index has a base of 100, and an increase in value means Swedish Krona has been weakened while a drop in value means a stronger Krona. 32 currencies valued against the Krona is included in the index with different weights. The weights are calculated based on the value of trade between Sweden and the country, which in theory should reflect the importance of the counterpart's currency on Swedish Krona. EUR, NOK and USD are therefore the currencies with the highest weight in KIX. In the best case, one would want to use a currency basket that is specified for the CPI component. For example, most food are imported from countries within Europe therefore weight on EUR/SEK should be higher for subcomponents for food and electronics from Asia, therefore CNY/SEK or JPY/SEK should be higher for the subcomponent for electronics. However, this is not feasible as the trade data is not on this detailed level and it is more practical to use the KIX instead. This could however mean that we do not capture all the correlations between the subcomponent and currencies.

3.2.2 Electricity price

Price of electricity futures play a role in inflation as the subcomponent "Electricity", which measures the consumer price of electricity is directly related to the price of the futures contract. The data we used is based on future contracts for floating electricity price per region from Stockholm Nasdaq's database. The regional data is then weighted and aggregated to a combined price based on SCB's methodology of

measuring the price of electricity. This methodology is a combination of weighting of actual electricity prices with different floating or fixed time contracts for each region. In theory, one has to correct for the energy tax that could differ as well from year to year and from region to region. For simplicity we have not done that in this study.

This weighting and analysis was done by the macro research team at Nordea and the time series used in this study was directly provided by Nordea.

3.2.3 Oil price

Oil price has a clear impact on many subcomponents of the CPI, the most obvious being the impact on energy prices, car fuels and prices of flight tickets. Apart from that, there should be an impact on all imported goods as well since oil prices affects the cost of logistics. With the correlation analysis, this effect is however seen to be very small and hardly notable.

Brent Crude and WTI are the two most common oil prices indices that essentially measure the price of the same commodity, oil. Brent Crude is the price of North Sea oil while WTI is the price of American West Texas oil index. The two indices have a very similar underlying product and are highly correlated. In this study, we have chosen to use Brent Crude and the time series was retrieved using Reuters Eikon's database.

3.3 Interest rates

The relationship between interest rate and the CPI inflation is complex. On the one hand, inflation affects the interest rate directly since, if inflation is higher, investors would be less willing to lend money at a lower interest rate than inflation. On the other hand, with higher interest rate, mortgage rates would increase, thereby increasing housing expenses, which would increase inflation rate as well. Creating a loop of causal connections.

To try to simplify some of the relationships, the fixed mortgage rate of five year maturity was used. Maturity of five year was used as it is the longest maturity and is often least affected by short-term events such as declines in oil prices, reducing the cross correlation with other terms. The data was retrieved directly from the Riksbank's database of mortgage rates.

3.3.1 Cotton price

Cotton is a raw material used in many consumer products, such as clothing. In theory, the price of cotton could have a direct impact on the price of the clothes. However, it could also be the case that the material cost is such a small portion of the total cost and that clothing may not be sensitive to cotton prices at all. But cotton price correlates well with various other agriculture products such as soy beans and coffee and could therefore provide additional information even on the food components. The cotton price data was retrieved from Reuter Eikon's database.

Chapter 4

Forecasting model structure

This section will describe the construction of the model, describing overall approach and methodology based on mathematical models from Chapter 2 are used in the model.

The model will be constructed on a subcomponent basis, as basing the model on a main group level would mean that some important information about the subcomponent might be lost in the aggregation. For example it would be difficult to say if the main group "Food and non-alcoholic beverage" have seasonal patterns, but it might be easier to say something about the subcomponents "Fruits" or "Vegetables". Another example is that it is difficult to say how the main group "Housing" correlate with market data, but the subcomponents "Electricity" and "Fuels" will be more likely to correlate with the market data for electricity and oil prices. One can also argue that one should model by product groups instead. However, this would not be practical as the number of variables that need to be modelled increase exponentially and data for product groups is not available to the public and need to be purchased from SCB. Also data of product groups are much more sensitive to the effect of changing items in the product groups [11]. This means for example that, if the Swedish retailer ICA have sales on coffee one month or if they decide to bring in new coffee brands, thus would affect the product group data "Coffee" dramatically. This type of situation would be difficult to model by regression and ARIMA.

The models will be built using both the Distributed Lag Models and ARIMA described earlier on each subcomponent as shown in Figure 4.1. Depending on the approach used and the level of correlation between subcomponent and the market data, DLM might not be on used on all subcomponent while ARIMA model will

always be used. This means that each subcomponent will either be:

- fitted by distributed lag model if the correlation is high enough and the residual modelled using ARIMA, or
- modelled directly through ARIMA

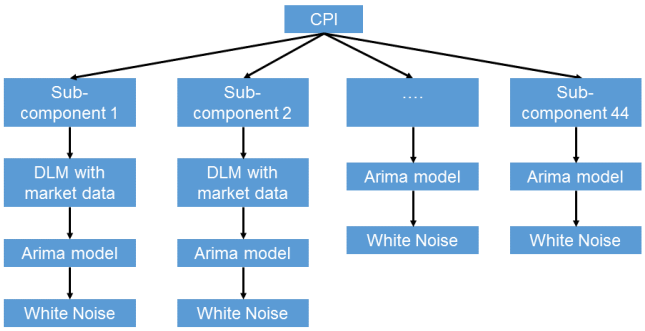


Figure 4.1: Illustrative overview of general process

Three forecasting approaches will be tested and compared to see which yields the best result in terms of RMSE. All modelling calculations will be performed in R and Excel.

For all approaches, the model input will be the 12 month change in CPI instead of the actual index. For each subcomponent i, the change is simply calculated as

$$Y_{i,t} = \frac{I_{i,t+12}}{I_{i,t}} - 1,$$

where $I_{i,t}$ is the index level for subcomponent i at time t measured in month.

In this report, we will be comparing the result from our three approaches with two benchmarks, The Riksbank's historical inflation forecast as well as the naive forecast. The Riksbank's forecasts are retrieved from March 2014 as that is the start of our forecast horizon and the naive forecast is simply defined as the same value as one year ago, i.e. inflation for April 2014 will be equal to April 2013.

4.1 Approach 1: Direct ARIMA

In the first approach, we will use the most straight forward way to model inflation simply by modelling the time series of each subcomponent using seasonal ARIMA and aggregating the results in the end.

Considering that in each ARIMA model there are 6 variables that need to be identified and that the number of models that need to be checked increases exponentially with each additional variable, this is an optimization problem that can not be brute forced. This problem is resolved by using a step-wise algorithm created by Hyndman and Khandakar [7]. The actual calculation is performed in R using the *auto.arima* function that optimizes the order of the model by minimizing AICc.

Step 1: The series are differentiated d and D times until the time-series pass the KPSS stationarity test described earlier in Equation (2.11).

Step 2: Try four different models to start with

- ARIMA$(2, d, 2)(1, D, 1)$
- ARIMA$(0, d, 0)(0, D, 0)$
- ARIMA$(1, d, 0)(1, D, 0)$
- ARIMA$(0, d, 1)(0, D, 1)$

The model with the smallest AICc is selected and denoted the current model

Step 3: Consider up to thirteen variations of the current model

- where one of p, q, P and Q is allowed to vary by ± 1 from the current model
- where p and q both vary by ± 1 from the current model
- where P and Q both vary by ± 1 from the current model
- where the intercept c is included if the current model has $c = 0$ or excluded if the current model has $c \neq 0$.

If any of the variation yields a lower AICc, it is set to be the new current model. Step 2 is repeated until no model with lower AICc can be found. [10]

After the model has been selected, the residual will be checked to see if it is consistent with white-noise using Ljung-Box. Then the forecast for each subcomponent will be made and aggregated using the methods described earlier in Chapter 2.3 regarding CPI aggregation [2]. Then we will plot the forecast compared with naive forecast and the Riksbank's forecast. Finally, we calculate the RMSE as shown in Equation (2.15).

4.2 Approach 2: Simple DLM and ARIMA

By adding regression of market data to the model, we provide the model with additional information and should therefore provide a better result than only using the historical time-series as information.

In the second approach, we will be using the simple Distributed Lag Model with simply one lag. This means that for each subcomponent with significant correlation with any of the market data at a certain lag, we will perform the regression on the one lag that resulted in the highest correlation. Essentially rewriting Equation (2.1) as:

$$y_t = \alpha + x_{t-\hat{l}} + \epsilon_t, \tag{4.1}$$

where $\hat{l}$ is the lag with the highest correlation between x and y.

4.2.1 When to use DLM

Not all subcomponents will correlate with any of the chosen market data, and it makes little sense to use DLM in situations where the data does not correlate. Before we start identifying the correlation of the subcomponent and the market data, we need to define what should be an appropriate level of correlation in order to include it in the model.

The cross correlation function (CCF) for lag j of the time series $\{X_i\}_{1\leq i\leq N}$ and $\{Y_i\}_{1\leq i\leq N}$ assuming that they are two independent $N(0,\sigma)$-distributed random variables, can be calculated as

$$\rho(j) = \frac{\frac{1}{N-j}\sum_{t=1}^{N-j} X_t Y_{t+j}}{\sqrt{\frac{1}{N-j}\sum_{t=1}^{N-j} X_t}\sqrt{\frac{1}{N-j}\sum_{t=1}^{N-j} Y_{t+j}}} \approx \frac{1}{N\sigma_x\sigma_y}\sum_{t=1}^{N} X_t Y_{t+j}$$

Where the last step is valid for $N \gg n$ where n is the number of month include in the CCF. The product $X_t Y_{t+j}$ can then be written as $\frac{\sigma_x \sigma_y}{2}(Q_1 - R_1)$, where Q_1, R_1 are $\chi^2(1)$-distributed random variables. Hence

$$\rho(j) \approx \frac{1}{2N}(Q_N - R_N)$$

We are the interested in finding the largest correlation, for the set of investigated lags, i.e.

$$E\Big[\max_{1\leq j\leq n} \rho(j)\Big] = \frac{1}{2N} E\Big[\max_{1\leq j\leq n} (Q_N - R_N)\Big]$$

Using that the moment generating function for $(Q_N - R_N)$ is $(1 - 4s^2)^{-N/2}$ and n being the number of lags included and N the number of data points[15], we end up with

$$E\Big[\max_{1\leq j\leq n} \rho(j)\Big] \leq \frac{2\log(n)}{N\sqrt{1 - n^{-2/N}}} \tag{4.2}$$

A maximum lag of $n = 24$ month has been set in this study, and with $N = 216$ i.e 216 month (or 18 years) of CPI data, the absolute value of the correlation should be significantly higher than 0.17 to be included in regression.

For simplicity, a threshold of 0.40 has been set for performing the regression analysis. In Table 4.1, significant correlations are marked in green. As seen in the table, a majority of the subcomponents did not have a significant correlation with the market data for this approach. A similar table was also created with the limit that the maximum lag must be larger than six and the result is shown in Table 4.2

Table 4.1: Cross correlation matrix between the CPI subcomponents and market data at optimal lags

Series number	Series name	KIX index		Electricity		Oil		Interest rate		Cotton	
		ρ	At lag	ρ	At lag	ρ	At lag	ρ	At lag	ρ	At lag
1	Bread & Cereals	0.21	0	0.15	0	0.11	3	0.42	4	0.17	5
2	Meat	0.43	0	0.37	3	0.28	21	0.51	22	-0.08	3
3	Fish	0.53	7	0.16	0	0.40	24	0.35	24	-0.37	9
4	Milk, Cheese & Eggs	0.21	24	0.29	2	-0.18	16	0.33	24	0.21	3
5	Oils & Fats	-0.11	9	0.19	1	-0.22	14	0.32	24	0.09	2
6	Fruit	0.49	1	0.22	2	0.39	19	0.51	22	-0.11	2
7	Vegetables	0.30	1	-0.30	18	-0.36	0	0.26	22	-0.15	23
8	Sweets	0.37	11	0.26	14	-0.28	9	0.33	24	0.20	19
9	Food Products N.E.C.	0.37	7	0.19	16	-0.32	8	0.30	24	0.16	20
10	Coffee, Tea & Cocoa	-0.26	15	-0.15	0	0.20	14	0.12	2	0.46	5
11	Cold Non-Alcoholic Beverages	0.12	24	-0.17	22	-0.18	11	0.20	4	0.22	16
12	Spirits	0.11	7	-0.15	9	-0.15	11	-0.22	0	-0.17	17
13	Wine	-0.16	10	-0.41	10	-0.36	24	-0.22	24	-0.15	17
14	Beer	0.28	9	0.29	0	-0.05	18	0.47	20	0.12	8
15	Tobacco	-0.34	18	-0.16	3	0.20	14	0.18	14	-0.08	9
16	Clothing	0.33	6	0.23	5	0.36	22	0.19	0	-0.11	17
17	Footwear	0.23	8	-0.16	24	-0.10	11	0.40	0	0.34	4
18	Owner-Occupied Housing	0.18	24	0.36	14	-0.10	6	0.32	13	-0.21	24
19	Electricity	0.23	24	0.59	3	-0.23	0	0.24	21	-0.52	22
20	Fuel	-0.15	6	0.29	0	0.72	2	0.27	14	-0.18	15
21	Rented & Housing Co-Operative Dwellings	-0.35	14	0.28	17	0.20	16	-0.46	0	0.22	22
22	Imputed Rent for Owner Occupiers	-0.41	7	0.14	2	0.35	5	0.57	3	0.51	9
23	Furniture	0.32	5	0.24	10	0.30	22	0.48	19	-0.43	1
24	Household Textiles	0.20	20	0.21	11	0.32	9	0.39	7	0.25	16
25	Household Utensils	0.28	3	0.17	19	-0.32	3	0.40	15	-0.21	24
26	Household Appliances	0.31	7	0.27	19	-0.29	7	0.27	13	-0.20	2
27	Tools & Garden Equipment	0.43	21	0.31	12	-0.12	1	0.38	10	0.28	4
28	Household Maintenance	0.21	9	0.24	9	-0.36	3	-0.28	0	0.29	15
29	Medicaments, Spectacles, etc.	0.14	3	-0.14	24	0.24	0	-0.22	11	-0.19	9
30	Outpatient Services	-0.22	0	-0.21	5	0.31	0	-0.22	10	-0.19	6
31	Purchase of Vehicles	-0.35	0	0.29	16	-0.35	9	-0.19	8	0.40	0
32	Operation of Vehicles	-0.21	2	-0.28	7	0.83	1	0.23	0	0.25	2
33	Transport Services	-0.32	22	0.26	19	0.17	19	0.33	19	-0.33	0
34	Postal Services	0.17	3	-0.20	0	-0.27	2	-0.29	5	-0.27	11
35	Audio, Photo & Information Equipment	0.17	9	0.17	14	-0.30	6	-0.25	10	-0.25	3
36	Other Durables for Recreation & Culture	0.41	9	0.32	6	0.21	14	0.37	24	-0.18	24
37	Other Recreational Equipment, Garden and Pets	0.27	11	0.17	4	0.15	2	0.29	3	0.30	2
38	Recreational & Cultural Services	0.30	24	0.29	6	0.16	18	0.40	15	0.45	11
39	Newspapers, Books & Stationary	-0.25	11	-0.23	17	0.31	10	-0.17	0	0.25	8
40	Package Holidays	0.55	11	0.18	17	-0.48	6	0.19	24	-0.27	7
41	Education	-0.18	6	-0.19	18	-0.30	24	-0.24	24	-0.26	24
42	Catering Services	0.29	13	0.34	2	0.26	24	0.47	21	0.22	6
43	Accommodation Services	0.20	20	-0.20	14	-0.34	15	-0.14	19	-0.21	24
44	Miscellaneous Goods & Services	0.38	7	0.33	18	0.24	23	0.14	16	-0.23	8

Table 4.2: Cross correlation matrix between the CPI subcomponents and market data at optimal lags with restriction of minimum lag of 6

Series number	Series name	KIX index		Electricity		Oil		Interest rate		Cotton	
		ρ	At lag	ρ	At lag	ρ	At lag	ρ	At lag	ρ	At lag
1	Bread & Cereals	0.15	24	0.05	6	-0.10	15	0.40	6	0.17	6
2	Meat	0.35	6	0.29	6	0.28	21	0.51	22	0.04	24
3	Fish	0.53	7	0.14	8	0.40	24	0.35	24	-0.37	9
4	Milk, Cheese & Eggs	0.21	24	0.15	24	-0.18	16	0.33	24	0.18	6
5	Oils & Fats	-0.11	9	-0.12	13	-0.22	14	0.32	24	-0.08	14
6	Fruit	0.42	6	0.13	6	0.39	19	0.51	22	-0.05	6
7	Vegetables	0.24	6	-0.30	18	0.25	24	0.26	22	-0.15	23
8	Sweets	0.37	11	0.26	14	-0.28	9	0.33	24	0.20	19
9	Food Products N.E.C.	0.37	7	0.19	16	-0.32	8	0.30	24	0.16	20
10	Coffee, Tea & Cocoa	-0.26	15	0.11	14	0.20	14	0.09	6	0.46	6
11	Cold Non-Alcoholic Beverages	0.12	24	-0.17	22	-0.18	11	0.18	24	0.22	16
12	Spirits	0.11	7	-0.15	9	-0.15	11	-0.17	6	-0.17	17
13	Wine	-0.16	10	-0.41	10	-0.36	24	-0.22	24	-0.15	17
14	Beer	0.28	9	0.12	6	-0.05	18	0.47	20	0.12	8
15	Tobacco	-0.34	18	0.15	18	0.20	14	0.18	14	-0.08	9
16	Clothing	0.33	6	0.20	6	0.36	22	-0.19	9	-0.11	17
17	Footwear	0.23	8	-0.16	24	-0.10	11	-0.21	14	0.31	6
18	Owner-Occupied Housing	0.18	24	0.36	14	-0.10	6	0.32	13	-0.21	24
19	Electricity	0.23	24	0.52	6	0.22	10	0.24	21	-0.52	22
20	Fuel	-0.15	6	0.25	22	0.57	6	0.27	14	-0.18	15
21	Rented & Housing Co-Operative Dwellings	-0.35	14	0.28	17	0.20	16	0.31	18	0.22	22
22	Imputed Rent for Owner Occupiers	-0.41	7	0.08	6	0.34	6	0.53	6	0.51	9
23	Furniture	0.29	8	0.24	10	0.30	22	0.48	19	-0.30	6
24	Household Textiles	0.20	20	0.21	11	0.32	9	0.39	7	0.25	16
25	Household Utensils	0.24	6	0.17	19	-0.29	6	0.40	15	-0.21	24
26	Household Appliances	0.31	7	0.27	19	-0.29	7	0.27	13	-0.14	24
27	Tools & Garden Equipment	0.43	21	0.31	12	-0.09	21	0.38	10	0.24	6
28	Household Maintenance	0.21	9	0.24	9	-0.27	6	-0.16	19	0.29	15
29	Medicaments, Spectacles, etc.	0.09	24	-0.14	24	-0.22	15	-0.22	11	-0.19	9
30	Outpatient Services	0.22	24	-0.19	6	-0.31	12	-0.22	10	-0.19	6
31	Purchase of Vehicles	0.33	12	0.29	16	-0.35	9	-0.19	8	-0.20	14
32	Operation of Vehicles	0.17	24	-0.28	7	0.40	6	0.17	6	0.19	24
33	Transport Services	-0.32	22	0.26	19	0.17	19	0.33	19	0.27	18
34	Postal Services	-0.17	12	0.17	12	-0.20	6	-0.27	6	-0.27	11
35	Audio, Photo & Information Equipment	0.17	9	0.17	14	-0.30	6	-0.25	10	-0.23	11
36	Other Durables for Recreation & Culture	0.41	9	0.32	6	0.21	14	0.37	24	-0.18	24
37	Other Recreational Equipment, Garden and Pets	0.27	11	0.13	6	-0.07	10	0.27	24	0.28	6
38	Recreational & Cultural Services	0.30	24	0.29	6	0.16	18	0.40	15	0.45	11
39	Newspapers, Books & Stationary	-0.25	11	-0.23	17	0.31	10	0.16	16	0.25	8
40	Package Holidays	0.55	11	0.18	17	-0.48	6	0.19	24	-0.27	7
41	Education	-0.18	6	-0.19	18	-0.30	24	-0.24	24	-0.26	24
42	Catering Services	0.29	13	0.27	8	0.26	24	0.47	21	0.22	6
43	Accommodation Services	0.20	20	-0.20	14	-0.34	15	-0.14	19	-0.21	24
44	Miscellaneous Goods & Services	0.38	7	0.33	18	0.24	23	0.14	16	-0.23	8

For the significant correlations, only one market data variable will be used in the regression. For the variable with multiple market data with correlation above 0.4, we will select the market data that are more likely to have a causal effect on the component. For example in Table 4.1, we can see that components 22 and 23 correlates with cotton and KIX but clearly the component actually measures price for rent, therefore the interest rate will be used for regression.

Another observation from Table 4.1 is that the model will have little predictive power since the maximum lag was zero for many of the significant correlations between subcomponent and market data. Also, restricting the value of the maximum lag to be above six as shown in Table 4.2 did not improve the prospect of the predictive power. In most cases, the components that had the maximum correlation at a lag below six previously now had the maximum correlation at lag six (or larger in some cases) with lower correlation than previously. Considering that such a model would be based on the autocorrelation of the market data, it is difficult to imagine that a DLM model based on setting the minimum lag allowed would perform better than constructing the DLM model with no restrictions and then forecasting the market data using ARIMA as input for forecasts.

After the appropriate market data correlations and lags have been identified, we perform the DLM regression on each of the subcomponent with the significant lag and model the residual with ARIMA. For the subcomponent without any significant correlations, results from ARIMA model in Approach 1 could be used directly. We then make forecasts based on the model and compare it to the other approaches and calculate RMSE of the forecast.

4.3 Approach 3: RFDLM and ARIMA

Building on Approach 2, the idea of the third approach is to capture a larger amount of the information in the market data by including more steps of lag in the regression. Ideally, one would want to fit a function to the cross-correlation function and use it as the restriction for regression. However, such a model would be difficult to maintain and unpractical to implement, therefore we will use the linear RFDLM described in Equation (2.3) using a linear function to restrict the lags.

It is easy to reduce Equation (2.3) to a new single point regression as shown in

Equation (4.1) by recalculating the time series $\{X_t\}_{1\leq t\leq N}$ as

$$X_t^* = \beta_0 \sum_{j=0}^{q} \frac{q+1-j}{q+1} X_{t-j} \tag{4.3}$$

We can then calculate the new cross-correlation between our new market data time series $\{X_t^*\}_{1\leq t\leq N-q}$ and $\{Y_t\}_{1\leq t\leq N-q}$ and then determine which correlations are significant enough to include in the regression, just as in Approach 2. Also note that the number of observations has decrease by q as a number of data points will be used to construct the new time series.

Another factor that needs to be determined is q, the number of lags that is to be included in Equation (4.3) for the best possible fit. This is a crucial variable in the RFDLM model but also difficult to determine. The simplest method would be to calculate the average correlation of all subcomponents and market data and see how it varies with different lags. This is shown the first column in Table 4.3 and we can see that, of course, the correlation improves with every lag we add but with diminishing improvements for each lag added. This is shown in the third column of the table. For the purpose of this study, we have chosen to build the model with lag 6 because the marginal improvement is still high at lag 6 and from a rational perspective, it's hard to argue how for example an increase in oil prices over 6 month ago should have a significant impact on inflation today.

Table 4.3: Improvement in correlation vs lag used

Lag (q)	**Average correlation**	**Δ Average correlation**	**Δ Average corr. per lag**
0	0.271	-	-
3	0.288	0.017	0.006
6	0.299	0.028	0.005
9	0.310	0.039	0.004
12	0.317	0.046	0.004
24	0.342	0.071	0.003
36	0.348	0.077	0.002

From Table 4.4 we see that the correlation using RFDLM with lag 6 improved for most but not all components. In theory, one could create an algorithm that selects the optimal RFDLM for each component and market data instead of using a general

model such as in this study.

After selecting $q = 6$, Equation (4.3) can be calculate and $\{X_t\}_{1 \leq t \leq N}$ replaced with$\{X_t^*\}_{1 \leq t \leq N}$. The modelling process then becomes similar to in Approach 2 where we model the subcomponent with market data where correlation was above 0.4. After the model has been created, forecasts will be calculate and compared to the naive forecast and the Riksbank's forecast same as previously. RMSE will then be calculated after the subcomponents have been aggregated and compared to the benchmarks.

Table 4.4: Improvement in max correlation between market data and CPI subcomponent after using RFDLM with 6 lag

Series number	Series name	KIX index	Electricity	Oil	Interest rate	Cotton
1	Bread & Cereals	0.04	0.00	0.00	0.02	-0.02
2	Meat	-0.04	0.10	0.06	0.09	0.08
3	Fish	0.07	0.02	0.14	0.07	0.07
4	Milk, Cheese & Eggs	0.06	0.07	0.02	0.02	-0.02
5	Oils & Fats	0.05	0.02	0.04	0.02	0.05
6	Fruit	-0.01	0.01	0.05	0.05	0.06
7	Vegetables	-0.06	0.04	0.04	0.02	0.02
8	Sweets	0.05	0.07	0.07	0.06	0.03
9	Food Products N.E.C.	0.03	0.06	0.06	0.06	0.02
10	Coffee, Tea & Cocoa	0.01	0.03	0.03	-0.02	0.02
11	Cold Non-Alcoholic Beverages	0.08	0.03	0.04	-0.01	0.03
12	Spirits	0.14	0.05	0.04	-0.03	-0.01
13	Wine	0.01	0.09	0.05	0.03	0.00
14	Beer	-0.04	0.04	0.03	0.05	0.03
15	Tobacco	-0.13	0.02	0.01	0.04	0.02
16	Clothing	0.04	0.07	0.00	-0.03	-0.01
17	Footwear	-0.04	0.11	0.00	-0.06	-0.01
18	Owner-Occupied Housing	-0.01	0.08	0.00	0.01	0.09
19	Electricity	0.04	0.18	-0.01	0.02	0.02
20	Fuel	-0.01	0.01	0.04	0.02	0.01
21	Rented & Housing Co-Operative Dwellings	-0.01	0.05	0.02	-0.04	0.02
22	Imputed Rent for Owner Occupiers	0.08	0.01	0.03	0.03	0.02
23	Furniture	0.11	0.05	0.04	0.06	0.00
24	Household Textiles	0.01	0.03	-0.01	0.02	0.08
25	Household Utensils	0.04	0.05	0.02	0.03	0.05
26	Household Appliances	0.09	0.08	0.05	0.05	0.04
27	Tools & Garden Equipment	0.05	0.08	0.00	0.03	0.06
28	Household Maintenance	0.00	0.07	0.04	-0.03	0.01
29	Medicaments, Spectacles, etc.	0.00	0.08	0.01	-0.04	-0.01
30	Outpatient Services	0.01	0.03	0.02	0.08	0.00
31	Purchase of Vehicles	-0.03	0.09	0.04	0.05	-0.05
32	Operation of Vehicles	0.02	0.09	-0.02	0.03	0.01
33	Transport Services	0.04	0.03	0.00	0.03	-0.01
34	Postal Services	0.03	0.04	0.02	0.00	-0.05
35	Audio, Photo & Information Equipment	0.08	0.06	0.03	0.01	0.00
36	Other Durables for Recreation & Culture	0.01	0.08	0.02	0.04	0.07
37	Other Recreational Equipment, Garden and Pets	-0.07	0.02	-0.01	0.02	0.00
38	Recreational & Cultural Services	0.10	0.07	0.00	0.02	0.00
39	Newspapers, Books & Stationary	0.02	0.08	0.04	0.00	0.04
40	Package Holidays	0.04	0.04	0.03	0.06	0.04
41	Education	-0.01	0.02	0.04	0.03	0.02
42	Catering Services	-0.03	0.07	0.02	0.01	-0.01
43	Accommodation Services	0.01	-0.03	0.02	0.00	-0.03
44	Miscellaneous Goods & Services	0.02	0.10	0.02	-0.01	0.02

Chapter 5

Results and discussion

In this section we will present the results from the three approaches described in Chapter 4 and discuss the results.

5.1 Approach 1

After fitting an ARIMA model for each of the components, the Ljung-Box test and RMSE in percentage was calculated for every time series as shown in Table 5.1. Out of the 44 series, the ARIMA model for 37 series passed the Ljung-Box test. However, the error of fitted values are still smaller than the naive approach. On the other hand, 4 of the series that passed the Ljung-Box test still performed slightly worse than the naive forecast. Seasonal components of 12 months were found in every ARIMA model, despite the fact that the data was already year-on-year inflation rate.

Comparing at Table 5.2 and 5.1 we notice that RMSE of the forecast and fitted data have rather large difference, this is common as past correlations does not always predict the future. In total, forecasts for 10 of the components performed worse than the naive forecast. The components that did not pass the Ljung-Box test generally had worse RMSE for the forecast as well in comparison to the naive forecast but there are some components that passed the Ljung-box test but still performed worse than the naive forecast. This is in-line with the finding from RMSE of the fitted data points. Some of the components, such as "Electricity" and "Fuel", had particularly bad results using the ARIMA model and had RMSE of over 90% as shown in Table 5.2. As the period forecasted was April 2014 to March 2016, the large RMSE could also be explained by the sharp fall of oil prices during the period. The fall in oil prices would have affected both components directly and many of the other components

indirectly, this fall would of course not have been possible to predict using ARIMA models.

Table 5.1: Resulted ARIMA model for each subcomponents and result of tests

Series number	Series name	p	d	q	P	D	Q	Seasonal time	Ljung-Box test (Q value)	Passed Ljungbox	AIC value	RMSE fitted	RMSE naive fitted	Difference fitted
1	Bread & Cereals	4	1	2	1	0	2	12	0.10	YES	-1487	4%	10%	-6%
2	Meat	1	1	1	0	0	1	12	0.24	YES	-1475	4%	14%	-10%
3	Fish	1	0	2	1	0	1	12	0.50	YES	-1282	6%	23%	-17%
4	Milk, Cheese & Eggs	1	1	0	2	0	2	12	0.00	NO	-1465	5%	16%	-11%
5	Oils & Fats	2	1	2	1	0	1	12	0.27	YES	-1328	5%	15%	-9%
6	Fruit	0	1	1	2	0	2	12	0.07	YES	-974	9%	22%	-13%
7	Vegetables	1	0	1	0	0	1	12	0.10	YES	-766	8%	17%	-8%
8	Sugar, Jam, Honey, Chocolate & Confectionery	3	1	1	2	0	1	12	0.00	NO	-1447	6%	15%	-9%
9	Food Products N.E.C.	5	1	3	2	0	1	12	0.00	NO	-1538	5%	14%	-9%
10	Coffee, Tea & Cocoa	3	1	1	2	0	1	12	0.23	YES	-1069	4%	34%	-31%
11	Cold Non-Alcoholic Beverages	0	1	1	2	0	2	12	0.00	NO	-1592	6%	9%	-4%
12	Spirits	0	1	0	1	0	0	12	1.00	YES	-2019	4%	8%	-4%
13	Wine	0	1	1	0	0	1	12	0.99	YES	-1492	6%	5%	1%
14	Beer	4	1	3	2	0	1	12	0.95	YES	-1378	6%	6%	0%
15	Tobacco	2	0	1	2	0	0	12	0.15	YES	-1034	4%	15%	-12%
16	Clothing	1	0	2	1	0	2	12	0.04	NO	-1197	12%	24%	-13%
17	Footwear	1	1	1	0	0	1	12	0.05	YES	-1070	9%	20%	-11%
18	Owner-Occupied Housing	0	1	0	0	0	1	12	0.98	YES	-1732	8%	11%	-3%
19	Electricity	1	0	1	2	0	1	12	0.88	YES	-1060	5%	20%	-15%
20	Fuel	2	0	0	2	0	0	12	0.00	NO	-937	6%	15%	-9%
21	Rented & Housing Co-Operative Dwellings	0	1	0	0	0	0	12	0.09	YES	-1997	6%	8%	-3%
22	Imputed Rent for Owner Occupiers	2	0	1	1	0	1	12	0.09	YES	-1160	3%	33%	-31%
23	Furniture	1	1	0	2	0	1	12	0.13	YES	-1577	6%	25%	-18%
24	Household Textiles	0	1	1	0	0	1	12	0.90	YES	-1085	7%	31%	-24%
25	Household Utensils	0	1	1	1	0	1	12	0.67	YES	-1335	7%	29%	-22%
26	Household Appliances	0	1	1	0	0	1	12	0.38	YES	-1367	7%	27%	-20%
27	Tools & Garden Equipment	4	0	1	2	0	1	12	0.13	YES	-1398	6%	20%	-14%
28	Household Maintenance	1	0	0	2	0	0	12	0.63	YES	-1481	5%	12%	-7%
29	Medicaments, Spectacles, etc.	0	1	0	2	0	0	12	0.47	YES	-1064	9%	5%	4%
30	Outpatient Services	1	1	1	1	0	0	12	0.00	NO	-1182	6%	11%	-4%
31	Purchase of Vehicles	0	1	0	2	0	0	12	0.41	YES	-1610	5%	15%	-10%
32	Operation of Vehicles	1	0	0	2	0	0	12	0.07	YES	-1097	10%	24%	-14%
33	Transport Services	1	1	1	0	0	1	12	0.80	YES	-1253	11%	43%	-32%
34	Postal Services	0	1	0	2	0	2	12	0.58	YES	-1099	8%	7%	1%
35	Audio-Visual, Photographic & Information Proces	1	1	2	0	0	1	12	0.66	YES	-1355	4%	12%	-8%
36	Other Major Durables for Recreation & Culture	2	0	0	2	0	0	12	0.06	YES	-1496	6%	11%	-6%
37	Other Recreational Items & Equipment, Gardens	1	1	1	0	0	1	12	0.63	YES	-1535	8%	39%	-30%
38	Recreational & Cultural Services	0	1	2	0	0	1	12	0.60	YES	-1642	8%	37%	-29%
39	Newspapers, Books & Stationary	1	0	1	1	0	1	12	0.69	YES	-1352	7%	24%	-18%
40	Package Holidays	2	0	2	0	0	1	12	0.94	YES	-1044	9%	38%	-29%
41	Education	1	1	0	2	0	0	12	0.32	YES	-1263	9%	9%	0%
42	Catering Services	0	1	0	0	0	1	12	0.29	YES	-1822	5%	33%	-28%
43	Accommodation Services	0	1	1	1	0	1	12	0.33	YES	-1104	9%	16%	-7%
44	Miscellaneous Goods & Services	1	1	1	2	0	2	12	0.08	YES	-1762	7%	29%	-23%

Table 5.2: RMSE of the forecasts during the period April 2014 to March 2016 for each component using approach 1

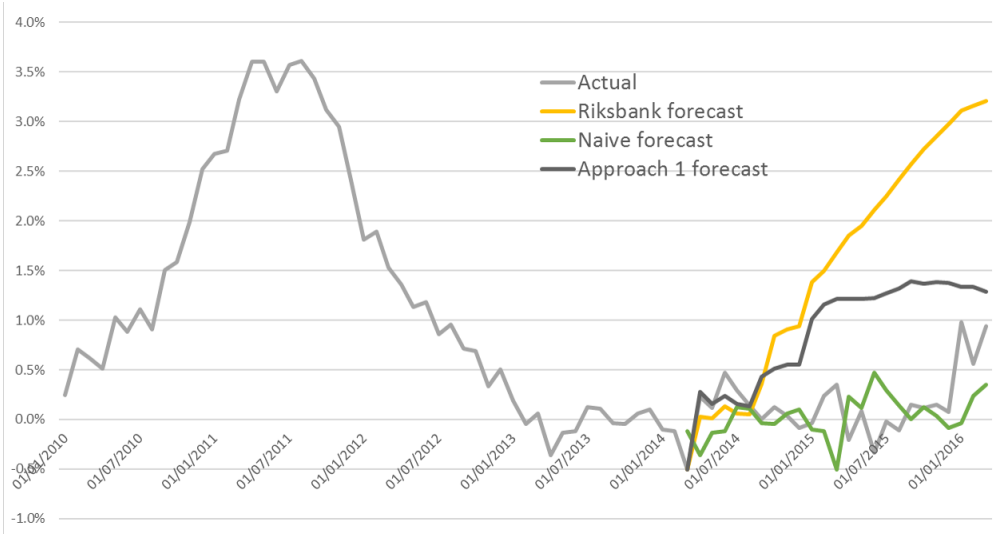
Series number	Series name	RMSE forecast	RMSE naive forecast	Difference forecast
1	Bread & Cereals	27%	41%	-14%
2	Meat	25%	80%	-55%
3	Fish	24%	50%	-26%
4	Milk, Cheese & Eggs	67%	74%	-6%
5	Oils & Fats	28%	56%	-28%
6	Fruit	37%	48%	-11%
7	Vegetables	31%	42%	-11%
8	Sugar, Jam, Honey, Chocolate & Confectionery	25%	40%	-14%
9	Food Products N.E.C.	30%	45%	-15%
10	Coffee, Tea & Cocoa	37%	65%	-27%
11	Cold Non-Alcoholic Beverages	39%	35%	3%
12	Spirits	21%	35%	-14%
13	Wine	33%	33%	1%
14	Beer	25%	38%	-13%
15	Tobacco	31%	57%	-26%
16	Clothing	33%	50%	-17%
17	Footwear	25%	32%	-7%
18	Owner-Occupied Housing	34%	70%	-36%
19	Electricity	92%	44%	48%
20	Fuel	126%	46%	80%
21	Rented & Housing Co-Operative Dwellings	41%	49%	-7%
22	Imputed Rent for Owner Occupiers	58%	59%	-1%
23	Furniture	21%	42%	-21%
24	Household Textiles	31%	39%	-8%
25	Household Utensils	31%	57%	-26%
26	Household Appliances	34%	45%	-11%
27	Tools & Garden Equipment	46%	40%	6%
28	Household Maintenance	24%	47%	-22%
29	Medicaments, Spectacles, etc.	33%	32%	1%
30	Outpatient Services	59%	59%	0%
31	Purchase of Vehicles	56%	43%	13%
32	Operation of Vehicles	68%	39%	29%
33	Transport Services	32%	45%	-13%
34	Postal Services	32%	52%	-20%
35	Audio-Visual, Photographic & Information Processing Equipment	74%	56%	19%
36	Other Major Durables for Recreation & Culture	43%	49%	-6%
37	Other Recreational Items & Equipment, Gardens & Pets	26%	48%	-22%
38	Recreational & Cultural Services	27%	56%	-29%
39	Newspapers, Books & Stationary	22%	33%	-11%
40	Package Holidays	33%	45%	-12%
41	Education	28%	34%	-7%
42	Catering Services	20%	51%	-32%
43	Accommodation Services	19%	53%	-34%
44	Miscellaneous Goods & Services	35%	37%	-2%

Aggregating the data, we see from Figure 5.1 that the ARIMA model appear to have performed worse than the naive forecast in the long term, diverging further away from the actual values than the naive forecast especially during January 2015 to January 2016. However, it still performed better than the Riksbank's that had a forecast which was far off from the actual inflation. The RMSE percentage was 70% for the ARIMA model, 138% for the Riksbank's forecast and only 32% for the naive forecast for the forecast period April 2014 to March 2016.

It is however known that ARIMA models tend to work well in the short-term and worse for the long term. Looking at the 6 month RMSE value from April 2014, the ARIMA model had a value of 44%, the Riksbank had 52% and the naive forecast had 77%. The ARIMA approach forecast tends to converge to a value in the long term. We can see from the beginning of 2015, the forecast start to become less volatile and start converging towards a single value.

It is evident that some form of external data is needed to improve the long term forecast of this model. As mentioned earlier, the oil price dropped by around 50% and this had a large impact on the CPI levels. If the oil price drop had not happened, ARIMA could have been an more accurate forecast.

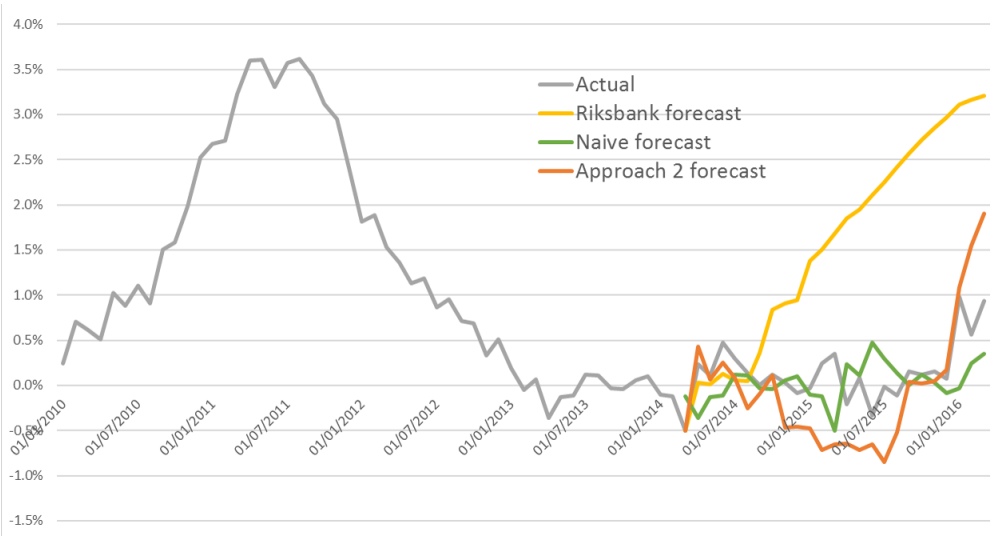
Figure 5.1: Aggregated and comparison for Approach 1



5.2 Approach 2

In the second approach, we have added the market data to the model by using DLM regression and combined it with the ARIMA model. The result of the forecast improved significantly, as shown in Figure 5.2. The RMSE value is now 40% which is a 30% improvement compared to only using the ARIMA model. We can see that the model is tracing the actual values much more than in Approach 1.

Figure 5.2: Aggregated and comparison for Approach 2



There are also some issues with this approach. It can be observed from the chart that the drop in oil price at the end of 2014 and the beginning of 2015 have had an immediate impact on the model, making the modelled CPI undershoot the actual CPI. Similarly, the reverse effect can be seen in the beginning of 2016 when the mortgage rate as well as oil price increased drastically in a short amount of time, making the modelled CPI overshoot the actual CPI. This model is therefore sensitive to changes in market data, overshooting or undershooting the actual CPI with sudden changes in market data.

Looking at the result at a more detailed level, we can investigate the performance of the forecast for each component compared to the naive forecast. In Table 5.3 we note that DLM-model was used on 17 components and out of these, 10 had better results than the naive forecast while 7 had worse results, an improvement over Ap-

proach 1. However, we note that using mortgage rate to predict series 21 and 22 had particularly bad results and the naive forecast outperformed the DLM model by a large amount. On the other side, market data for electricity and oil price appears to have improved the forecasts for a few components, especially "Electricity", "Fuel" and "Operation of Vehicles", by a large margin compared to Table 5.2 in Approach 1. As oil prices had very large movements during the forecast period, the effect on annual change of CPI has been high. Therefore including oil prices appears to have helped to improve the prediction on these affected components drastically and therefore improved the overall results. Cotton price and KIX had smaller effects on the forecast, some minor improvements or worsening of RMSE.

One limitation due to the inclusion of the market data is that the actual forecast ability is limited by the shortest lag used in this model, which was zero for many components as shown in Table 5.3. This means that the model can actually only forecast current results. However, one can always use ARIMA model on these components or on the market data itself to use as input for the forecast.

Table 5.3: RMSE of the forecasts during the period April 2014 to March 2016 for each component using Approach 2

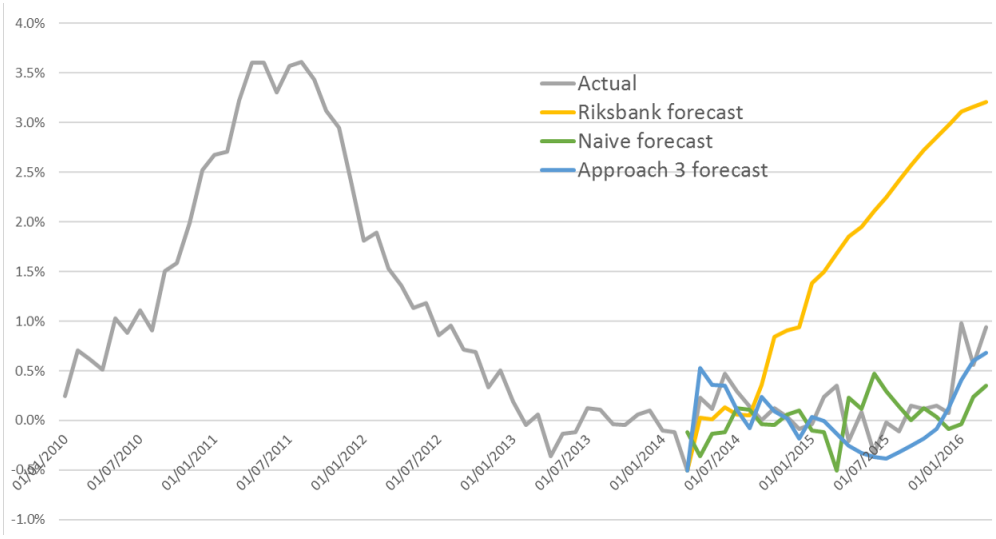
Series number	Series name	Regression with	Lag used	RMSE forecast Approach 2	RMSE naive forecast	Difference
1	Bread & Cereals	None	-	28%	41%	-13%
2	Meat	KIX	0	61%	80%	-18%
3	Fish	KIX	7	24%	50%	-26%
4	Milk, Cheese & Eggs	None	-	53%	74%	-21%
5	Oils & Fats	None	-	29%	56%	-27%
6	Fruit	KIX	1	27%	48%	-21%
7	Vegetables	None	-	30%	42%	-12%
8	Sugar, Jam, Honey, Chocolate & Confectionery	None	-	24%	40%	-15%
9	Food Products N.E.C.	None	-	31%	45%	-14%
10	Coffee, Tea & Cocoa	Cotton price	5	38%	65%	-27%
11	Cold Non-Alcoholic Beverages	None	-	35%	35%	0%
12	Spirits	None	-	21%	35%	-15%
13	Wine	Electricity	10	35%	33%	3%
14	Beer	None	-	25%	38%	-12%
15	Tobacco	None	-	34%	57%	-23%
16	Clothing	None	-	35%	50%	-15%
17	Footwear	None	-	29%	32%	-3%
18	Owner-Occupied Housing	None	-	34%	70%	-36%
19	Electricity	Electricity	3	39%	44%	-5%
20	Fuel	Oil price	2	60%	46%	14%
21	Rented & Housing Co-Operative Dwellings	Mortgage rate	0	119%	49%	70%
22	Imputed Rent for Owner Occupiers	Mortgage rate	3	123%	59%	65%
23	Furniture	Mortgage rate	19	46%	42%	4%
24	Household Textiles	None	-	31%	39%	-8%
25	Household Utensils	None	-	31%	57%	-26%
26	Household Appliances	None	-	35%	45%	-10%
27	Tools & Garden Equipment	KIX	21	44%	40%	5%
28	Household Maintenance	None	-	27%	47%	-20%
29	Medicaments, Spectacles, etc.	None	-	28%	32%	-4%
30	Outpatient Services	None	-	60%	59%	1%
31	Purchase of Vehicles	Cotton price	0	46%	43%	4%
32	Operation of Vehicles	Oil price	1	23%	39%	-16%
33	Transport Services	None	-	27%	45%	-18%
34	Postal Services	None	-	33%	52%	-20%
35	Audio-Visual, Photographic & Information Processing Equipment	None	-	69%	56%	14%
36	Other Major Durables for Recreation & Culture	KIX	9	24%	49%	-25%
37	Other Recreational Items & Equipment, Gardens & Pets	None	-	27%	48%	-21%
38	Recreational & Cultural Services	Cotton price	11	37%	56%	-19%
39	Newspapers, Books & Stationary	None	-	22%	33%	-10%
40	Package Holidays	KIX	11	41%	45%	-3%
41	Education	None	-	25%	34%	-9%
42	Catering Services	Mortgage rate	21	15%	51%	-36%
43	Accommodation Services	None	-	22%	53%	-31%
44	Miscellaneous Goods & Services	None	-	35%	37%	-2%

5.3 Approach 3

In Approach 3, we used RFDLM regression on the market data and combined it with the ARIMA model. From the correlation analysis in Table 4.3, it was shown that correlation increased with larger lags but with diminishing marginal effects and $q = 6$ was selected for modelling. We will also present the result of a test made with $q = 24$ and call it "Approach 3b".

In Figure 5.3 we can see that Approach 3 appears to be the best performing model out of the three approaches. Indeed the RMSE was only 20% compared to 32% of the naive forecast. It's clear that the forecast in Approach 3 displayed in Figure 5.3 is more "smooth" compared to forecast made in Approach 2 displayed in Figure 5.2, which is natural since the restricting function in RFDLM is similar to a moving average.

Figure 5.3: Aggregated and comparison for approach 3



Looking at the results for the subcomponents in Table 5.4 we can see that out of the 22 components we have used RFDLM on, the RFDLM forecast performed better than the naive forecast on 15 and worse on 7 of them just as in Approach 3, but with a larger number of subcomponents getting correlation above 0.40. It appears that mortgage rate became a better predictor in Approach 3 than in Approach 2 while Electricity has become worse. It is hard to identify exactly why the aggregated

CPI from Approach 3 performed better than Approach 2, just by comparing subcomponent by subcomponent between Table 5.4 and 5.3. It appears that Approach 2 actually performed better on many of the component, but on average Approach 3 performed better and more components were modelled using market data. The improvement is difficult to pin-point to a single factor.

5.3.1 Approach 3b

According to Table 4.3, the correlation improved with increasing q and a RFDLM was $q = 24$ was tested and the result is displayed in Figure 5.4. It appears in the figure that the model performed a lot worse with $q = 24$ than $q = 6$ and the forecast had RMSE of 76% which is even worse than the ARIMA model. The forecast appear to have a negative offset compared to Approach 3 and one reason for this could be due to the large drop in oil prices which the model overestimated due to the large number of lags. This had never occurred in the data used for calibration and the coefficient was therefore not adjusted to predict this type of effects.

Figure 5.4: Aggregated and comparison for Approach 3b

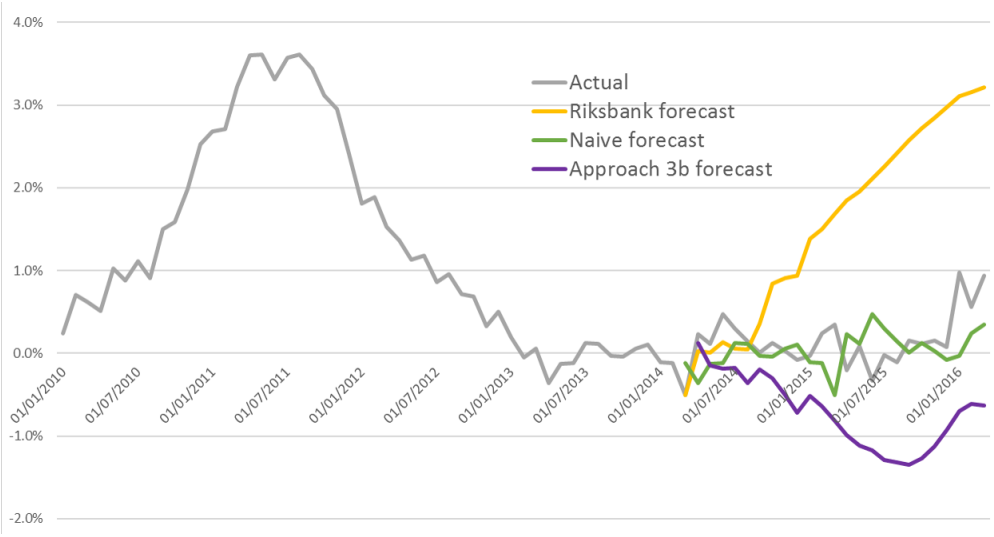


Table 5.4: RMSE of the forecasts during the period April 2014 to March 2016 for each component using approach 3b

Series number	Series name	Regression with	Lag used	RMSE forecast Approach 3	RMSE naive forecast	Difference
1	Bread & Cereals	None	-	28%	41%	-13%
2	Meat	None	-	29%	80%	-50%
3	Fish	KIX	7	23%	50%	-27%
4	Milk, Cheese & Eggs	None	-	53%	74%	-21%
5	Oils & Fats	None	-	29%	56%	-27%
6	Fruit	KIX	1	20%	48%	-28%
7	Vegetables	Oil price	0	19%	42%	-22%
8	Sugar, Jam, Honey, Chocolate & Confectionery	KIX	8	22%	40%	-18%
9	Food Products N.E.C.	KIX	7	23%	45%	-22%
10	Coffee, Tea & Cocoa	Cotton price	3	40%	65%	-25%
11	Cold Non-Alcoholic Beverages	None	-	35%	35%	0%
12	Spirits	None	-	21%	35%	-15%
13	Wine	Oil price	24	30%	33%	-2%
14	Beer	None	-	25%	38%	-12%
15	Tobacco	None	-	34%	57%	-23%
16	Clothing	None	-	35%	50%	-15%
17	Footwear	None	-	29%	32%	-3%
18	Owner-Occupied Housing	Electricity	12	45%	70%	-25%
19	Electricity	Electricity	2	56%	44%	12%
20	Fuel	Oil price	1	68%	46%	22%
21	Rented & Housing Co-Operative Dwellings	Mortgage rate	0	49%	49%	0%
22	Imputed Rent for Owner Occupiers	Mortgage rate	2	71%	59%	13%
23	Furniture	Mortgage rate	16	34%	42%	-8%
24	Household Textiles	Mortgage rate	7	37%	39%	-2%
25	Household Utensils	Mortgage rate	12	52%	57%	-5%
26	Household Appliances	None	-	35%	45%	-10%
27	Tools & Garden Equipment	KIX	20	60%	40%	20%
28	Household Maintenance	None	-	27%	47%	-20%
29	Medicaments, Spectacles, etc.	None	-	28%	32%	-4%
30	Outpatient Services	None	-	60%	59%	1%
31	Purchase of Vehicles	None	-	52%	43%	9%
32	Operation of Vehicles	Oil price	0	32%	39%	-6%
33	Transport Services	None	-	27%	45%	-18%
34	Postal Services	None	-	33%	52%	-20%
35	Audio-Visual, Photographic & Information Processing Equipment	None	-	69%	56%	14%
36	Other Major Durables for Recreation & Culture	KIX	7	55%	49%	6%
37	Other Recreational Items & Equipment, Gardens & Pets	None	-	27%	48%	-21%
38	Recreational & Cultural Services	Cotton price	10	49%	56%	-7%
39	Newspapers, Books & Stationary	None	-	22%	33%	-10%
40	Package Holidays	KIX	9	34%	45%	-11%
41	Education	None	-	25%	34%	-9%
42	Catering Services	Mortgage rate	20	24%	51%	-28%
43	Accommodation Services	None	-	22%	53%	-31%
44	Miscellaneous Goods & Services	KIX	6	53%	37%	16%

5.4 Comparisons

Comparing all the models at once in Figure 5.5 and Table 5.5, we can clearly see that the Riksbank's forecast performed the worst, it is not entirely fair to compare the Riksbank's forecast with Approach 2 or Approach 3 where market data during the forecast period was used, which the Riksbank of course did not have access in March 2014 when the forecast was made. However, it still performed worse than Approach 1 using only ARIMA and also worse than the naive forecast. Approach 2 and 3 that used market data as explanatory variables performed close to naive forecast, this is not surprising considering the information advantage of having future market data as input, thus the predictive value in practice is small using DLM or RFDLM. But if the market data are known in advance, Approach 3 would be the best method to model inflation using this knowledge advantage. If the accuracy of this model could further be improved, it could have some applications for market makers that trades products related to the level of inflation in order to monitor inflation on a live basis.

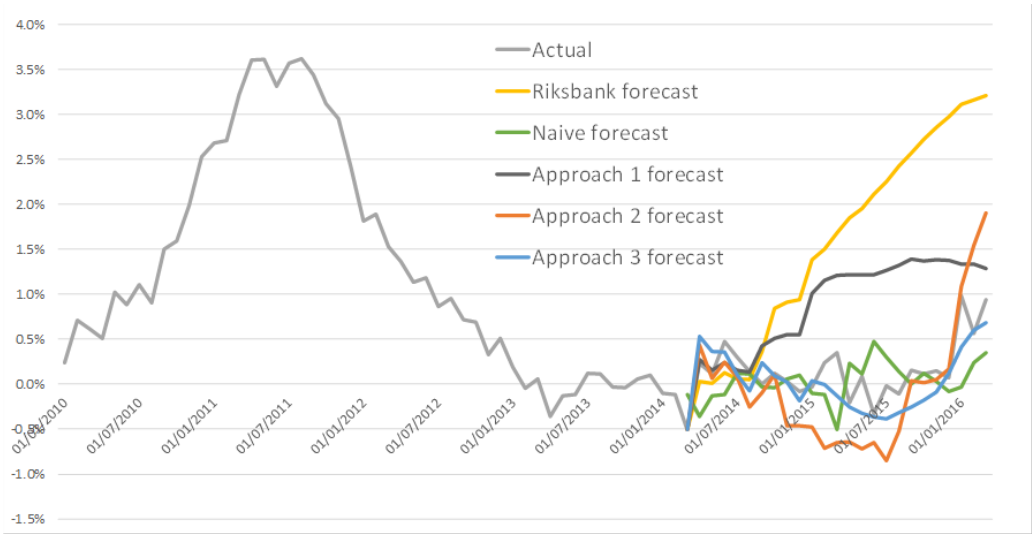


Figure 5.5: Aggregated comparison for all approaches

Table 5.5: RMSE in percentage and maximum forecast error for the different approaches

	Riksbank	**Naive forecast**	**Approach 1**	**Approach 2**	**Approach 3**
RMSE	138%	32%	70%	40%	20%
Max error	2.89%	1.01%	1.55%	1.00%	0.57%

Chapter 6

Conclusions

It is difficult to draw any decisive conclusion on the causal effect between the explanatory variables and the response variable as the underlying relationships based on a statistical study. Even thought we have narrowed it down quite a lot by using the individual CPI subcomponents as the starting point, the causal effects still can not be fully understood. For the two year forecast period of April 2014 to March 2016, it was shown that inflation does indeed correlate with market data and Restricted Finite Distributed lag model (RFDLM) and ARIMA is the model that creates the best model results compared to ARIMA or DLM and ARIMA. It has been shown in previous study that commodity prices offer improvement to the AR(1) forecast [16], although not entirely compare able to this study, we can also conclude that market data does indeed offer more accurate model of the CPI, using RFDLM but cannot forecast as market data need to be known. On the other hand, this modelling approach could still be useful for someone monitoring the market in live time and speculating on fixed income that with priced in inflation.

Even though the study showed the correlation between market data and inflation, the forecast was not always improved by the market data. Oil price and electricity price improved some of the components directly, KIX and cotton price had a mixture of effects and mortgage rate actually decreased the accuracy of our forecast on the component level. Even with just oil price and electricity price as input, it would be difficult to make accurate long term forecast, due to the fact that the lag used was very short. This means that only a forecast of the current or near-term inflation could be made, this is also called nowcast or livecast. Alternatively, the forecaster would need to make forecasts on the market data to use the model. This is of course not very useful, since making forecasts on the market data is probably just as diffi-

cult, if not more difficult, than forecasting the inflation. With this said, there might be a value in using the model to create inflation scenarios of for example, low and high oil prices to estimate the effect on inflation these scenarios.

One mechanism not fully understood by the author while performing the test is that, even if the accuracy decreased component by component the result of the aggregated forecast could still be improved for some reason. For example, a test was made to replace the component 21, 22 and 23 that had worse RMSE using mortgage rate as forecast to the actual values. The RMSE of the aggregated result decreased from 20% in the model to 65%, despite using the actual values in the model.

As always, further research could be done in this area to potentially improve the results. One improvement of high potential would be to test a larger class of restricting functions in RFDLM such as a roof function, exponential functions or setting the restricting function dynamically for each subcomponent and market data to fit the cross correlation curve. Also, instead of using a general model for all subcomponent, one could in theory create an algorithm that determines the best model to use for each market data and subcomponent. More types of explanatory data could also be used, such as additional market data or even economical data that are release on a monthly basis. Unemployment rate and inflation expectation surveys are two variables that are currently used by many economist to forecast CPI inflation. One could also consider to remove some elements of the model, such as the 5 year mortgage rate which decreased the model accuracy on a component level. In addition, adjusting for structural factors such as tax hikes and financial crises could improve the forecast even further.

Bibliography

[1] Oxana Tarassiouk: CPI Statistics description, *http://www.scb.se/contentassets/a1e257bb3a574420b9d3f2ff59851c0a/pr0101-bs-2015-ot-150407-eng.pdf*

[2] P. Bäckström and M. Sammar. *The use of Superlative Index Links in the Swedish CPI*. 2012.

[3] William Philips. *The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom*. 1958.

[4] Economics Online, *www.economicsonline.co.uk/Global_economics/Phillips_curve.html*

[5] Michael Owyang: Has the Phillips Curve Relationship Broken Down? *https://www.stlouisfed.org/on-the-economy/2015/september/phillips-curve-unemployment-down-inflation-low*

[6] Jeffery Parker: Theory and Practice of Econmetrics at Reed College, *http://www.reed.edu/economics/parker/312/tschapters/S13_Ch_3.pdf*

[7] Rob J. Hyndman, and Yeasmin Khandakar. *Automatic Time Series Forecasting: The Forecast Package for R*. Journal of Statistical Software, 2008.

[8] Petre J. Brockwell and Richard A. David. *Introduction to Time Series and Forecasting*. 2010.

[9] Richard Baillie: Maximum Likelihood Estimation of Time Series Models, *https://msu.edu/ baillie/822/MLE.pdf*

[10] Rob J. Hyndman, and George Athanasopoulos. *Forecasting: Principles and practice*. 2016.

[11] Mick Silver and Saeed Heravi: Why Elementary Price Index Number Formulas Differ: Price Dispersion and Product Heterogeneity,
https://www.imf.org/external/pubs/ft/wp/2006/wp06174.pdf

[12] R. W. Hafer and Scott E. Hein. *Forecasting Inflation Using Interest-Rate and Time-Series Models: Some International Evidence*. 2010.

[13] Mehmet Pasaogullari: Do Oil Prices Predict Inflation?,
www.clevelandfed.org/newsroom-and-events/publications/economic-commentary/2014-economic-commentaries/ec-201401-do-oil-prices-predict-inflation.aspx

[14] Mårten Bjellerup and Mårten Löf. *Oljeprisets effekter på svensk inflation*. 2008.

[15] Gautam Dasarathy: A Simple Probability Trick for Bounding the Expected Maximum of n Random Variables,
http://www.cs.cmu.edu/ gautamd/Files/maxGaussians.pdf

[16] Yu-chin Chen, Stephen J. Turnovsky, and Eric Zivot *Forecasting Inflation using Commodity Price Aggregates*. 2012.