ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics [1]

Bachelor Thesis Econometrics & Operations Research

Forecasting over a different dataset

july 6, 2019

Author	Diederick Clements (447863)
Supervisor	Professor D. Fok
Second assessor	Assistant professor M. D. Zaharieva

Abstract

The parameters of a model are estimated over a certain dataset. For instance, in the case a rolling window are the last n datapoints used to estimate the parameters of a model. In this paper is investigated, whether the dataset to estimate the parameters could be chosen in such a way that the parameter estimates improve and subsequently provide better forecasts. It is researched in the context of timeseries, such that two choices for the dataset could be made. The parameters in a model could be estimated over the same dataset, or each parameter could be estimated over a different dataset. Secondly, every parameter in a model could for every time t be estimated over, for instance, the last n datapoints, or it could differ per time t how many datapoints are used to estimate the parameters. Choosing the right option could significantly reduce the forecast error, because there could be a different impact over time of some variables on the dependent variable or there could be for some variables a long-term and for others a short-term relation with the dependent variable. In this paper is a chain of formulas developed, to get the best dataset and subsequently better parameter estimates. A real-life example was used, namely forecasting the Dutch GDP, to determine whether this approach improves the forecasts. Only estimating the parameters for every time t over a different dataset has some implications to be better than a simple rolling window.

[1] The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam

Contents

1 Introduction

Forecasting is done in a lot of contexts. For instance, the total amount of calls for next week, the Gross Domestic Product (GDP) of the next quarter or the price for tomorrow of a certain stock are all examples of forecasting. All these examples have in common that they depend on time, such that the models need to be estimated over the past. Which past is good for estimation of the model? The last 10 weeks could provides good estimates of the model, but the last 50 weeks could provide even better estimates. This timeframe, over which the parameters are estimated, is called the estimation timeframe.

Consider a model with a linear relation between the dependent variable and the independent variables. All the parameters in such a model are normally estimated over the same estimation timeframe. However, the impact of a variable may not be constant over time. Moreover, it could be that some independent variables show a clear long-term linear relation with the dependent variable, while others do not. As a result, non-linearity could occur.

Therefore, it could be that estimating some parameters over a different estimation timeframe than others provides a better forecast. Intuitively this may be very logical. Some independent variables may have a clear long-term linear relation with the dependent variable. An estimation timeframe of 100 weeks would then provide for these parameters approximately the same estimates as an estimation timeframe of 10 weeks, which is illustrated in figure 1. Choosing an estimation timeframe of 100 weeks would provide better estimates of the parameters as more data is used. In case of a short-term linear relation between the dependent variable and the independent variables the estimated parameter really depends on which estimation timeframe is chosen, as also illustrated in figure 1. Here, an estimation timeframe of 10 weeks would be much better than an estimation timeframe of 100 weeks.

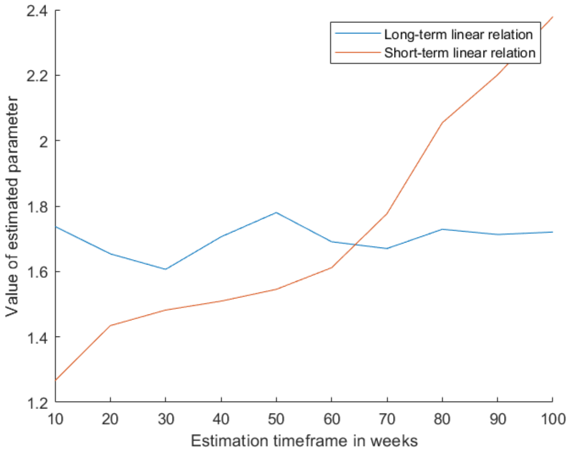

Figure 1: Value of an estimated parameter where the independent variable has a long-term or a short-term relation with the dependent variable

Improving the estimation timeframe could have a significant impact on the forecast accuracy. Therefore, the main research question is: Do the forecasts in a linear model become better if the parameters are calculated over different estimation timeframes? First is investigated how to compute the best estimation timeframe. In this section are also some models for the estimation timeframe proposed. Secondly, a real-life example is used, namely a model forecasting the Dutch GDP. Finally, the forecasts of the Dutch GDP are evaluated and the different ways to calculate and forecast the estimation timeframes are compared.

2 Literature

Clark and McCracken (2009) try to forecast the estimation timeframe. If a structural break happens, their model is made in such a way that it still produces reasonably good forecasts. Incorporating the possibility of a structural break nearly always reduces the mean squared error (*MSE*) of the forecast errors. They also combine rolling and recursive forecasts. A rolling window uses the last n observations to predict the next observation. A recursive window uses all the observations available until a certain point to predict the next observation. Combining different models and estimation windows reduces also the *MSE*.
However, they use only the possibility of one structural change over all variables. It may be very well possible that some independent variables have a long-term linear relation with the dependent variable and others have a short-term linear relation with the dependent variable.
In the latter case the parameter estimates need much more structural breaks then in the first case.

Zellner et al. (1991) assumes that the estimators interact with time. Let β_t be the estimate for a parameter in the model. They propose that v_t is normally distributed in $\beta_t = \beta_{t-1} + v_t$. Here, v_t is normally distributed with mean zero and a certain standard deviation. They also use the idea of regime switching.
In this paper we are not interested in what the value of v_t is or when the regime switching happens. The parameter β_t shall only depend on time in this way that it depends on the estimation timeframe.

After developing the right method to estimate and forecast the estimation timeframe it is tested on a real-life problem: forecasting the Dutch GDP. For doing so, the Econometric Institute Current Indicator Economy (EICIE) model is used, which is made by de Groot and Franses (2005). It provides two weeks after each quarter a forecast of the Dutch GDP, solely using as explanatory variable the number of people working at the Randstad company (denoted by S_t). As independent variables are taken some quarterly dummy variables, as well as some lags and transformations of S_t and GDP_t. They do some tests, such as a Johansen cointegration test or a unit root test, to determine which model is appropriate (Johansen, 1991; Hylleberg et al., 1990). In their conclusion they propose that modelling non-linearity could provide better forecasts.

Skogholt et al. (2017) added to EICIE some sentiment variables. The above-mentioned sentiment variables partially represent the consumer confidence. These variables where used in a linear model and in a neural network to forecast the Dutch GDP. Using extra variables and modelling in different ways improved significantly the EICIE model.
Although a neural network is able to model non-linearity, it is not easily and intuitively interpretable. In this paper is researched whether this non-linearity could be modeled in an understandable linear way. Moreover, it could be that short-term linearity is present, but for the entire dataset and for all variables together non-linearity is existent. If the length of this linearity could be estimated and forecasted per variable, than a linear model with a different estimation timeframes would be adequate.

3 Estimation timeframe

3.1 General calculation

Consider a linear model consisting of M independent variables:

$$y_t = \beta_{1,t}X_{1,t} + \beta_{2,t}X_{2,t} + ... + \beta_{m,t}X_{m,t} + \epsilon_t \tag{3.1}$$

Here, $X_{j,t}$ is the independent variable, y_t is the dependent variable and ϵ_t is the regression error on time t. The parameters $\beta_{j,t}$ need to be estimated over a certain estimation timeframe. For instance, a timeframe of length γ_0 uses the last γ_0 datapoints until time t-1 to estimate the parameters $\beta_{j,t}$. Subsequently, with these estimated parameters y_t could be forecasted. However, to know which estimation timeframe provides optimal forecasts, this estimation timeframe itself has to be forecasted as well. The central question in this case is: what is the optimal timeframe to get the best estimates of the parameters, such that the forecast of y_t is as good as possible. Otherwise stated, could the estimation timeframe be chosen in such a way that it minimizes the absolute value of the forecast error $\hat{\epsilon}_t$. Minimizing the absolute value of $\hat{\epsilon}_t$ produces more robust estimates of the estimation timeframe than minimizing $\hat{\epsilon}_t^2$. The model of the estimation timeframe, denoted by $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$, could be specified in various ways. It could be a constant γ_0, as mentioned in the previous example. However, it could also be a linear model in y_{t-1} and A_t. Here, A_t consists of all variables which could predict the estimation timeframe and which are available at time t, such as X_t. These and other ways to specify $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ are further explored in section 3.2. To estimate γ, an mixed integer problem (MIP) formulation is considered.

$$\min \sum_{t=1}^{T} c_t \tag{3.2}$$

$$\hat{\epsilon}_t = y_t - \sum_{j=1}^{M}\sum_{i=1}^{N} \hat{\beta}_{j,i,t}X_{j,t}(I_{j,i,t} - I_{j,i+1,t}) \qquad \forall t \tag{3.3}$$

$$I_{j,i,t} \leq I_{j,i-1,t} \qquad \forall j,t,i \in \text{[2,N+1]} \tag{3.4}$$

$$I_{j,1,t} = 1 \qquad \forall j,t \tag{3.5}$$

$$I_{j,N+1,t} = 0 \qquad \forall j,t \tag{3.6}$$

$$c_t \geq \hat{\epsilon}_t \qquad \forall t \tag{3.7}$$

$$c_t \geq -\hat{\epsilon}_t \qquad \forall t \tag{3.8}$$

$$I_{j,i,t} \in \mathbb{B} \qquad \forall j,i,t \tag{3.9}$$

$$I_{j,N,t}Z - 0,5 \leq \sum_{i=1}^{N} I_{j,i,t} - g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t) \leq 0.5 - \frac{1}{Z} + (1 - I_{j,2,t})Z \qquad \forall j,t \tag{3.10}$$

Minimizing the absolute value of the forecast error $\hat{\epsilon}_t$ is achieved by introducing a new variable c_t, which is equal to the maximum of $\hat{\epsilon}_t$ and -$\hat{\epsilon}_t$ and by subsequently minimizing c_t as stated in equation 3.2, 3.7 and 3.8.

The forecast error itself is calculated in equation 3.3. Here, the index j corresponds to the j'th independent variable of model 3.1. The index i, which is an integer from 1 to N, corresponds to the length of a certain estimation timeframe.[2] Time t, is an index from 1 until T. The parameter $\beta_{j,i,t}$ is estimated with ordinary least squares (OLS) and uses for the estimation the data from t minus the length of the estimation timeframe (belonging to i) until t-1. Note that the goal of this MIP formulation is not to estimate $\beta_{j,i,t}$, but γ_j. Therefore, $\beta_{j,i,t}$ is already estimated with OLS and subsequently put as a three-dimensional matrix in this MIP formulation.

There are N estimation timeframes used to estimate $\beta_{j,t}$. However, only one $\hat{\beta}_{j,i,t}$ should be taken as parameter estimate of $\beta_{j,t}$. Consider the case, where an estimation timeframe belonging to i^* is taken to estimate $\beta_{j,t}$, written as $\hat{\beta}_{j,i^*,t}$.
To take only into account $\hat{\beta}_{j,i^*,t}$ and not $\hat{\beta}_{j,i,t}$ for $i \neq i^*$ a new variable $I_{j,i,t}$ is created. This binary indicator $I_{j,i,t}$ is 1 if i is smaller than or equal to i^* and otherwise zero (restriction 3.4 and 3.9). A typical example of the indicator function $I_{j,i,t}$ for the first variable (j=1):

$$\begin{bmatrix} I_{1,1,2} & I_{1,1,3} & I_{1,1,4} & I_{1,1,5} \\ I_{1,2,2} & I_{1,2,3} & I_{1,2,4} & I_{1,1,5} \\ I_{1,3,2} & I_{1,3,3} & I_{1,3,4} & I_{1,1,5} \\ I_{1,4,2} & I_{1,4,3} & I_{1,4,4} & I_{1,1,5} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

In this example an estimation timeframe with $i^* = 3$ is used to estimate the first parameter of the model on $t = 2$. Note that $i^* = \sum_{i=1}^{N} I_{j,i,t}$. Notice also that $I_{j,i,t} - I_{j,i+1,t}$ equals one if $i = i^*$ and otherwise equals zero. Therefore, multiplying $\hat{\beta}_{j,i,t}$ with $I_{j,i,t} - I_{j,i+1,t}$ takes only the estimated parameters with an estimation timeframe belonging to i^* ($\hat{\beta}_{j,i^*,t}$), as stated in equation 3.3. Restrictions 3.5 and 3.6 are imposed to ensure that the parameter $\beta_{j,t}$ is included and that it is estimated over at least one estimation timeframe.

Finally constraint 3.10 makes it possible to use a certain model for the estimation timeframe by using the function $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$. As stated above this could be a constant or a linear model. The function $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ calculates the length of the estimation timeframe for parameter j on time t. It is continuous and goes from minus infinity to infinity. However, the estimation timeframe and its corresponding index i^*, is an integer between 1 and N. As stated before, $i^* = \sum_{i=1}^{N} I_{j,i,t}$. Therefore, the function $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ and $\sum_{i=1}^{N} I_{j,i,t}$ should have a maximal difference of 0.5, as the first is continuous and the latter is an integer. The constant Z is a large number, for instance, 10000. If the function $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ is smaller than 0.5 it forces i^* to be equal to one. The opposite is also true. If the function $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ is bigger than N+0.5 it forces i^* to be equal to N. Finally, an MIP cannot model smaller than, but only smaller than or equal to, so on the right side $\frac{1}{Z}$ is subtracted.

[2] The length of the estimation timeframe is equal to i+M-1. As there are M independent variables, the estimation timeframe should have at least a length of M.

3.2 Different models for the estimation timeframe

As mentioned in the previous section $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ could be specified in different ways. It could be specified constant or different over time and constant or different between variables. The goal of $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ is to forecast the estimation timeframe, such that the parameters $\beta_{j,t}$ are estimated as good as possible and subsequently the forecast error $\hat{\epsilon}_t$ is minimized. Therefore, to make a model for the estimation timeframe and estimate the parameters in this model the data from t=1 until t=T is used. Subsequently, the estimation timeframe from t=T+1 until t=T+k is forecasted.

The first case is constant over time and constant between variables ($t_{\text{con}}v_{\text{con}}$):

$$g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t) = \gamma_0 \qquad \forall j, t \in [1, T] \tag{3.11}$$

$$g_{j,t}(y_{t-1}, A_t|\hat{\gamma}_0) = \hat{\gamma}_0 \qquad \forall j, t \in [T+1, T+k] \tag{3.12}$$

In equation 3.11 is $g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t)$ specified. γ_0 is estimated with the MIP formulation of section 3.1. This estimated $\hat{\gamma}_0$ is used in equation 3.12 to forecast the estimation timeframe from t=T+1 until t=T+k.

Until now, the same window for every variable was considered. However, some variables may have a long-term linear relation and other a short-term linear relation with the dependent variable. Therefore, the second case is constant over time, but different between variables ($t_{\text{con}}v_{\text{dif}}$):

$$g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t) = \gamma_j \qquad \forall j, t \in [1, T] \tag{3.13}$$

$$g_{j,t}(y_{t-1}, A_t|\hat{\gamma}_j) = \hat{\gamma}_j \qquad \forall j, t \in [T+1, T+k] \tag{3.14}$$

The parameters are estimated over different estimation timeframes within model 3.1. Consider the case where $\hat{\gamma}_1 = 5$ and $\hat{\gamma}_2 = 10$. Model 3.1 is then estimated over the last 5 data-points with OLS and only the estimated parameter $\hat{\beta}_{1,t}$ is used as estimate of $\beta_{1,t}$. Subsequently, model 3.1 is estimated over the last 10 datapoints with OLS and only $\hat{\beta}_{2,t}$ is used as estimate of $\beta_{2,t}$. To conclude, the parameters in one model are estimated over different estimation timeframes. Again γ_j of equation 3.13 is estimated with the above mentioned MIP formulation and this estimated $\hat{\gamma}_j$ is used to forecast the estimation timeframe (equation 3.14).

It could also be the case that the independent variables depend on a common factor. For instance, all the variables increase if a shock happens. As a result could they sometimes all have a long-term linear relation with the dependent variable and at other times all have a short-term linear relation with the dependent variable. Therefore, the third case is different over time and constant between variables ($t_{\text{dif}}v_{\text{con}}$). To know which independent variables should be used to forecast the estimation window a two-step approach is needed. The first step is shown below:

$$\begin{aligned}
&g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t) = \gamma_t && \forall j, t \in [1, T] && (3.15)\\
&V_t \subseteq (y_{t-1} \cup A_t), |V_t| = q : && && \\
&\hat{\gamma}_t = \theta_1 V_{t,1} + ... + \theta_q V_{t,q} + \mu_t && \forall t \in [1, T] && (3.16)
\end{aligned}$$

Here, γ_t is again estimated with the MIP formulation stated in section 3.1. As a result, $\hat{\gamma}_t$ represents the best estimation timeframe for time t. However, the parameter γ_t depends on time, so it cannot be forecasted directly. Therefore, a model is made, where $\hat{\gamma}_t$ is the dependent variable. The independent variables could be y_{t-1} and some variables of A_t, denoted by V_t. They should be chosen in such a way that it optimizes a criterion when model 3.16 is calculated. For instance, minimizing the Akaike information criterion (AIC) could be a good method to chose the best independent variables (Akaike, 1974). Forecasting with the model in 3.16 is possible. However, the estimates of the model are based on minimizing the residual $\hat{\mu}_t$ and not on minimizing the forecast error $\hat{\epsilon}_t$ of model 3.1. Therefore, the second step minimizes the forecast error and uses the independent variables, which are obtained by optimizing a certain criterion in the first step.

$$\begin{aligned}
&g_{j,t}(\theta, V_t) = \theta_1 V_{t,1} + ... + \theta_q V_{t,q} && \forall t \in [1, T] && (3.17)\\
&g_{j,t}(V_t|\hat{\theta}) = \hat{\theta}_1 V_{t,1} + ... + \hat{\theta}_q V_{t,q} && \forall j, t \in [T+1, T+k] && (3.18)
\end{aligned}$$

In the second step are the parameters of equation 3.17 estimated using the MIP formulation of section 3.1. The estimated parameters are used to forecast the estimation timeframe (equation 3.18).

The last case combines the latter two cases: different over time and different between parameters ($t_{\text{dif}}v_{\text{dif}}$). It is modeled in nearly the same way as different over time and constant between parameters. In the first step are the best independent variables obtained by optimizing a certain criterion.

$$\begin{aligned}
&g_{j,t}(\gamma_{j,t}, y_{t-1}, A_t) = \gamma_{j,t} && \forall j, t \in [1, T] && (3.19)\\
&V_{j,t} \subseteq (y_{t-1} \cup A_t), |V_{j,t}| = q_j : && \forall j && \\
&\hat{\gamma}_{j,t} = \theta_{0,j} + \theta_{1,j} V_{1,j,t} + ... + \theta_{q_j,j} V_{q_j,j,t} + \mu_{j,t} && \forall j, t \in [1, T] && (3.20)
\end{aligned}$$

Again, $\gamma_{j,t}$ in equation 3.19 is estimated by using the MIP formulation. It finds for every parameter on every time t, the best estimation timeframe. The estimates $\hat{\gamma}_{j,t}$ could be seen as a time series for every variable j. Therefore, $\hat{\gamma}_{j,t}$ is regressed on some variables using OLS (equation 3.20). For every j should a criterion be optimized, such as the AIC, in order to obtain the best independent variables. The parameters of these independent variables are estimated in step two.

$$g_{j,t}(\theta_j, V_t) = \theta_{0,j} + \theta_{1,j}V_{1,j,t} + ... + \theta_{q_j,j}V_{q_j,j,t} \qquad \forall t \in [1, T] \tag{3.21}$$

$$g_{j,t}(V_t|\hat{\theta}_j) = \hat{\theta}_{0,j} + \theta_{1,j}V_{1,j,t} + ... + \hat{\theta}_{q_j,j}V_{q_j,j,t} \qquad \forall j, t \in [T+1, T+k] \tag{3.22}$$

In the second step is for every variable j a model for its timeframe is estimated. Two approaches could be done to estimate the parameters. The first approach is minimizing the forecast error $\hat{\epsilon}_t$ of model 3.1. The parameters of equation 3.21 could than be estimated by using the MIP formulation stated in section 3.1. Solving this MIP formulation may be computationally expensive. Therefore, coordinate descent (Tseng, 2001) could be applied, where the parameters are iteratively optimized.

The second approach is estimating the parameters of equation 3.20 using OLS. It minimizes the residuals of the best estimation timeframes ($\hat{\mu}_t$). This approach could be useful, if the MIP approach is too computationally expensive and does not give reasonable results. With both approaches could the estimated parameters be used to forecast the estimation timeframe, as presented in equation 3.22.

4 Short-term forecast of Dutch GDP

To explore whether forecasting the estimation timeframe reduces the forecast error, a real-life example is used, namely the EICIE model. The EICIE model provides a short-term forecast of the Dutch GDP and is developed by de Groot and Franses (2005). As independent variables are used a constant, some seasonal dummies and some lags of people working at Randstad (S_t) and Dutch GDP (GDP_t). The data of the GDP_t and S_t are also provided in the paper of de Groot and Franses (2005) and are available until 2003. They built their model with the entire dataset. In this paper, the model is built with the data until 1998. From 1999 until 2003 is the Dutch GDP forecasted and subsequently evaluated. The latter is necessary in order to compare the different methods of section 3.2. The model of de Groot and Franses (2005) cannot be used directly, as in this paper a different part of the dataset is used to built the model. However, all their tests are also done in this paper in order to construct an appropriate model.

The natural logarithm, denoted by $ln(.)$, is taken of GDP_t and S_t. For each of these variables the entire dataset until 1998 is analyzed to research deterministic trends and seasonality. To research whether there are unit roots the HEGY test is applied (Hylleberg et al., 1990). Let the log of GDP_t be denoted by y_t. $y_t - y_{t-4}$ is regressed on a intercept, seasonal dummies, a trend, some lags of $y_t - y_{t-4}$, $(1+L)(1+L^2)y_t$, $-(1-L)(1+L^2)y_t$, $-(1-L)(1+L)y_t$ and $-(L)(1-L)(1+L)y_t$. The last four element are tested on their significance, to determine whether there are unit roots. If one of these is insignificant, then a unit root is present. There unit roots are respectively, 1, -1, and the pair i, -i. The number of lags of $y_t - y_{t-4}$ is chosen in such a way that the Aike Information Criterion (AIC) is minimized (Akaike, 1974). For GDP_t and for S_t provided two lags the lowest AIC. All the statistics of the HEGY test with their corresponding significance levels are in table 1.

Table 1: Statistics of the HEGY test

Null hypothesis	GDP	S_t
Nonseasonal unit root (no cycle)	-1.489	-3.42*
Seasonal unit root (2 quarter per cycle)	-2.786*	-3.16**
Seasonal unit root (4 quarter per cycle)	8.02**	6.731**

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Both variables seem to have a nonseasonal unit root (unit root 1). The variable GDP may have unit root -1, but this is not very clear. Therefore, only root 1 is considered and all variables are first-differenced. To determine whether they have a stochastic trend in common, the Johansen cointegration test is considered (Johansen, 1991). The variable $ln(GDP_t)$ has a significant deterministic trend and a significant intercept. The cointegration relation could be with or without trend With these assumptions the Johansen cointegration test is done. The vector error correction model (VECM) with a lag order minimizing the AIC is selected (Johansen and Juselius, 1990). It resulted in a best lag order of 5.

It is clearly visible in table 2 that it does not really matter which assumption is taken. There is according to the Johansen cointegration test one cointegration relation between GDP_t and S_t.

Table 2: number of cointegration relations based on maximum-eigenvalue of Johansen cointegration test

Data trend In cointegration relation	None No intercept No trend	None Intercept No trend	Linear Intercept No trend	Linear Intercept Trend
GDP_t and S_t	1	1	1	1

Therefore, a VECM model is considered. The dependent variable is the first difference between the log of the GDP_t at time t: $ln(GDP_t) - ln(GDP_{t-1})$. Variables which possibly come in the model are seasonal dummies, the log of S_t and the log of GDP_{t-1}. The first difference of the log of the GDP, namely $ln(GDP_{t_1}) - ln(GDP_{t-2})$ and all its lags could also be added. The same holds for the staffing data of Randstad company, such that $ln(S_t) - ln(S_{t-1})$ and all its lags can be added. The model is built from specific to general. It begins with just a constant. The depended variable is regressed on a constant and iteratively over every possible variable. The variable with the lowest AIC is added if it reduces the AIC and if it is significant with a five percent significance. Once there is no variable reducing the AIC or the variable with the lowest AIC is not significant the procedure is stopped. We end up with the following model:

$$\begin{aligned}(ln(GDP_t) - ln(GDP_{t-1})) =& \beta_1 + \beta_2 Q_{1,t} + \beta_3(ln(GDP_{t-1}) - ln(GDP_{t-2})) + \\ & \beta_4(ln(GDP_{t-4}) - ln(GDP_{t-5})) + \beta_5(ln(S_t) - ln(S_{t-1})) + \\ & \beta_6(ln(S_{t-1}) - ln(S_{t-2})) + \beta_7(ln(S_{t-2}) - ln(S_{t-3})) + \epsilon_t.\end{aligned} \tag{4.1}$$

The Bai-Perron test determines the number of significant structural breaks in the data (Bai and Perron, 1998). A significant structural break leads to different parameter estimates before and after the structural break. The results are presented in the appendix section 8.1 table 7. There are according to the Bai-Perron test two significant structural breaks, namely in the third quarter of 1981 and the last quarter of 1987.

The Breusch-Pagan-Godfrey test determines whether heteroskedasticity is present (Breusch and Pagan, 1979). With a 5-percent significance there is no heteroskedasticity, as stated in table 8 in the appendix section 8.1. The Breusch-Godfrey serial correlation LM test investigates the presence of serial correlation (Breusch, 1978). For the entire dataset there is with a 5-percent significance serial correlation, which is presented in the appendix section 8.1 table 9. However, if the dataset after the last structural break is taken (from 1988 until 1998), then there is no serial correlation. To conclude, these statistics are not very good for the whole dataset until 1998. However, the statistics are much better if the parameters are estimated over a better dataset. Therefore, could the approach of using a different estimation timeframes per time t and also between the variables be very effective.

5 Results

The Dutch GDP is forecasted quarterly over the period from 1999 until 2003. Before every forecast of a certain quarter is model 4.1 estimated. The parameters in the aforementioned model are estimated over a certain estimation timeframe, as stated in section 3.2. There are 7 independent variables and the constant N is set equal to 20. Therefore, the timeframe has a minimal length of 7 and a maximal length of 26.

Table 3: Statistics of forecast errors for different window computations

Window computation	average	standard deviation	MAE	MSE
$t_{\text{con}}v_{\text{con}}$	-0.00298	0.00868	0.006732	0.00008
$t_{\text{con}}v_{\text{dif}}$	-0.00236	0.014769	0.011878	0.000213
$t_{\text{dif}}v_{\text{con}}$	-0.00222	0.008174	0.007044	0.000068
$t_{\text{dif}}v_{\text{dif}}$ using MIP	-0.00272	0.014798	0.01132	0.000215
$t_{\text{dif}}v_{\text{dif}}$ using OLS	-0.01001	0.035016	0.024763	0.001265

The first case was constant over time and constant between variables. The estimated parameter $\hat{\gamma}_0$ is equal to 13. All the parameters of model 4.1 are estimated over a timeframe with a length corresponding to $\hat{\gamma}_0 = 13$. With these estimated parameters is the Dutch GDP forecasted for every quarter from 1999 until 2003. In table 3 are some statistics of the forecast error given. The average is a little bit below zero. The standard deviation as well as the mean absolute error (MAE) and mean squared error (MSE) are also given. In figure 2 are the forecast errors plotted.

The second case is constant over time and different between variables. The estimates for the timeframes are given in the appendix, section 8.2 table 10. The estimated γ_j's are used to forecast the estimation timeframe and subsequently also forecast the Dutch GDP. The forecasts seem to be less accurate (table 3). The standard deviation, MAE and MSE are higher than if the case of constant over time and between variables. The average is the only statistic which is a little bit better.

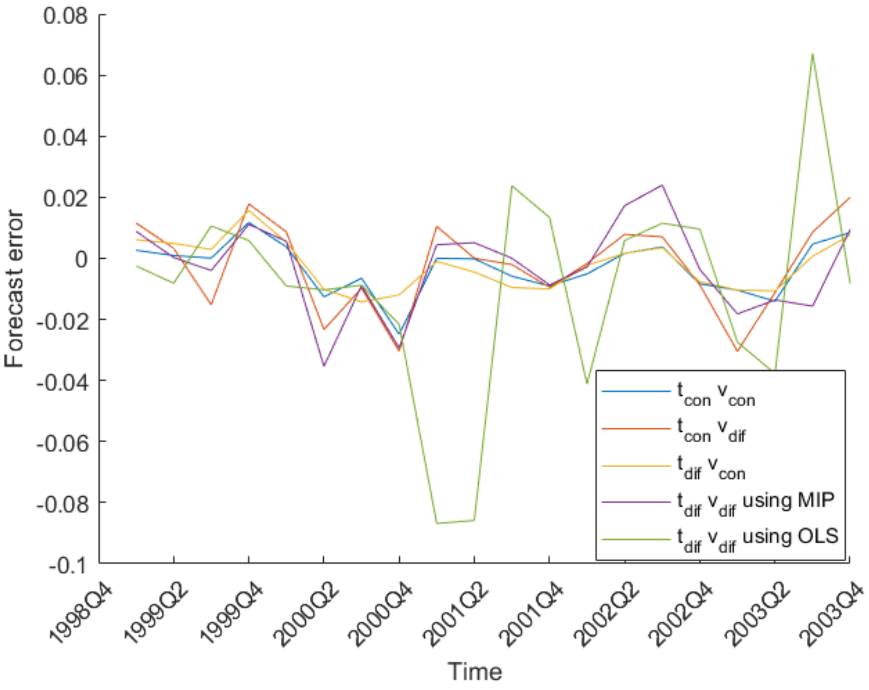

Figure 2: Forecast errors of different methods for modelling estimation timeframes

The third case is different over time and constant between variables. The independent variables selected to forecast the estimation timeframe where a constant, $log(S_{t-3} - S_{t-4})$ and $log(GDP_{t-5} - GDP_{t-6})$. These were selected by minimizing the AIC criterion. Subsequently were the parameters for these variables estimated with the MIP formulation of section 3.1. These estimates are respectively 19.23, 67.997 and 94.664. With these estimates are the estimation timeframes of the parameters of model 4.1 forecasted. Subsequently, are the parameters of model 4.1 repeatedly estimated to forecast the Dutch GDP for the period from 1999 until 2003. The statistics of the forecast errors are given in table 3 and seem to be better than the previous two methods, except for the MAE of constant over time and constant between the parameters.

The last case is different over time and between variables. Again the relevant variables to forecast the estimation timeframe of each parameter of model 4.1 are presented in table 4. With variable 1, 2, 3 and 4 is the estimation timeframe of parameter j of model 4.1 predicted. Subsequently is the parameter β_j estimated over this predicted estimation timeframe. The parameters of variable 1, 2, 3 and 4 are estimated in two ways. The first approach is using the MIP formulation to estimate these variables. The estimates of the estimated parameters are presented in the appendix, section 8.2 table 11.

Table 4: Variables selected for prediction of estimation timeframe

Parameters to estimate	variable 1	variable 2	variable 3	variable 4
β_1	c	$ln(S_{t-1}) - ln(S_{t-2})$	$ln(S_{t-2}) - ln(S_{t-3})$	$ln(S_{t-3}) - ln(S_{t-4})$
β_2	c	$ln(S_t)$	$ln(S_{t-2}) - ln(S_{t-3})$	$ln(GDP_{t-2}) - ln(GDP_{t-3})$
β_3	c	$ln(S_{t-1}) - ln(S_{t-2})$	$ln(S_{t-2}) - ln(S_{t-3})$	$ln(GDP_{t-1})$
β_4	c	$ln(S_{t-5}) - ln(S_{t-6})$	$ln(GDP_{t-2}) - ln(GDP_{t-3})$	$ln(GDP_{t-3}) - ln(GDP_{t-4})$
β_5	c	$ln(S_t)$	$ln(GDP_{t-3}) - ln(GDP_{t-4})$	$ln(GDP_{t-6}) - ln(GDP_{t-7})$
β_6	c	$ln(S_t) - ln(S_{t-1})$	$ln(S_{t-1}) - ln(S_{t-2})$	$ln(GDP_{t-5}) - ln(GDP_{t-6})$
β_7	c	$ln(S_t) - ln(S_{t-1})$	$ln(GDP_{t-1})$	$ln(GDP_{t-2}) - ln(GDP_{t-3})$

With these estimated parameters is the estimation timeframe repeatedly forecasted and the parameters of model 4.1 estimated, such that the Dutch GDP could be forecasted. The statistics of the forecast errors are given in table 3. The statistics are all worse than $t_{\text{con}}v_{\text{con}}$ and $t_{\text{dif}}v_{\text{con}}$. Compared to $t_{\text{con}}v_{\text{dif}}$ are the standard deviation and MSE better and the average and MAE worse.

The second approach uses OLS to estimate the parameters of variable 1, 2, 3 and 4 of table 4. The estimates are presented in the appendix, section 8.2 table 12. The statistics of the forecast error are also given in table 3 and are all far worse then the other methods.

This is also quite logical. The parameter estimates for predicting the estimation timeframe are obtained by minimizing the residual $\hat{\mu}_t$. This residual only minimizes the distance to the best estimation timeframe. However, is does not minimize the forecast error ϵ_t. All the statistics are based on minimizing the forecast error, so therefore is this method worse than the other methods.

To know whether these forecast methods provide unbiased forecasts, is a Mincer-Zarnowitz

Table 5: Mincer-Zarnowitz regression to evaluate unbiasedness

Window computation	F-statistic
$t_{\text{con}} v_{\text{con}}$	2.185754
$t_{\text{con}} v_{\text{dif}}$	1.390089
$t_{\text{dif}} v_{\text{con}}$	0.759566
$t_{\text{dif}} v_{\text{dif}}$ using MIP	2.128296
$t_{\text{dif}} v_{\text{dif}}$ using OLS	8.94384***

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

regression done (Mincer and Zarnowitz, 1969). The results are in table 5. All the methods provide unbiased forecasts except for $t_{\text{dif}} v_{\text{dif}}$ using OLS.

Finally, are the forecasts compared with the Diebold-Mariano test (Diebold and Mariano, 2002). The results are provided in table 6. All statistics are significant, so all the different methods have a significantly different forecast accuracy. Therefore, we can order the window computation methods. The best method and the second-best method seems to be different over time and constant over all parameters or constant over time and constant over all parameters. These two differ according to the Diebold-Mariano test significantly. However, it is not clear which method is better. The average and MSE are better of the first and the MAE is better of the latter. The first, which differs over time, may be better, as more statistics are in favour of $t_{\text{dif}} v_{\text{con}}$ then $t_{\text{con}} v_{\text{con}}$. After that, it is again not very clear whether constant over time and different per parameter or different over time and different per parameter using MIP is the best, due to the same reasons. Worst method is different over time and different between parameters using least squares.

Table 6: Diebold-Mariano statistics which compare the forecast accuracy of different methods

estimation method	$t_{\text{con}} v_{\text{con}}$	$t_{\text{con}} v_{\text{dif}}$	$t_{\text{dif}} v_{\text{con}}$	$t_{\text{dif}} v_{\text{dif}}$ with MIP	$t_{\text{dif}} v_{\text{dif}}$ with OLS
$t_{\text{con}} v_{\text{con}}$	-	-5.303***	-4.526***	-4.096***	-3.981***
$t_{\text{con}} v_{\text{dif}}$	-5.303***	-	-5.03***	-4.919***	-4.369***
$t_{\text{dif}} v_{\text{con}}$	-4.526***	-5.03***	-	-5.078***	-4.326***
$t_{\text{dif}} v_{\text{dif}}$ using MIP	-4.096***	-4.919***	-5.078***	-	-4.172***
$t_{\text{dif}} v_{\text{dif}}$ using OLS	-3.981***	-4.369***	-4.326***	-4.172***	-

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

6 Conclusion

This paper investigated the research question whether the forecasts become better if the parameters in a model are estimated over a different estimation timeframe. First, was a method developed to estimate a model for the estimation timeframe. Subsequently, were four models proposed, as the timeframe could be constant or different over time and constant or different between variables.

These models were tested on a real-life problem, namely forecasting the Dutch GDP. As a baseline was taken the EICIE model of de Groot and Franses (2005). Unfortunately, this model was made with the entire dataset. The models for the estimation timeframe could only be compared, if a part of the dataset is forecasted. Therefore, was the model built from 1977 until 1998, following the same procedure as de Groot and Franses (2005) in order to forecast quarterly from 1999 until 2003.

The forecast errors of the different models of the estimation timeframe were compared. The best method to calculate the estimation timeframe is different over time and constant between variables or constant over time and constant between variables. The estimation timeframe which is constant over time but different between variables and the estimation timeframe which is different over time and different between variables provided both a bit worse forecasts for the GDP. The worst forecasts were with the timeframe different over time and different between variables using OLS. In all cases resulted differencing the timeframe between variables in bad statistics about the forecast error.

For further research it could be investigated, whether differencing the timeframe between variables could be done in another way. When differencing the timeframe between the variables in this paper, the estimates mostly provided unbiased forecast errors, but the standard deviation and the mean absolute error increased enormously. It may be that the procedure in this paper is not optimal in reducing the mean absolute error. Therefore, it should be researched whether it is possible to develop a mathematical formula which provides unbiased forecasts and provably minimizes the mean absolute error.

7 Bibliography

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17(31):334–355.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294.

Clark, T. E. and McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, 50(2):363–395.

de Groot, B. and Franses, P. H. (2005). Real time estimates of gdp growth. Technical report.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Hylleberg, S., Engle, R. F., Granger, C. W., and Yoo, B. S. (1990). Seasonal integration and cointegration. *Journal of econometrics*, 44(1-2):215–238.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, pages 1551–1580.

Johansen, S. and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration—with applications to the demand for money. *Oxford Bulletin of Economics and statistics*, 52(2):169–210.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER.

Skogholt, M., Glorie, K., van de Velden, M., and reader Kim Schouten, C. (2017). Erasmus q-intelligence forecasting dutch gdp incorporating sentiment from dutch news.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Zellner, A., Hong, C., and Min, C.-k. (1991). Forecasting turning points in international output growth rates using bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. *Journal of Econometrics*, 49(1-2):275–304.

8 Appendix

8.1 General tests for the model which forecasts the Dutch GDP

Table 7: Bai-Perron test for number of structural breaks

Number of breaks	Scaled F-staticstic	Critical Value**
0 vs. 1	27.16305	21.87
1 vs. 2	26.54531	24.17
2 vs. 3	11.57291	25.13

Note: ** $p < 0.05$.

Table 8: Breusch-Pagan-Godfrey test for heteroskedasticity

F-statistic	2.016683*

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Breusch-Godfrey serial correlation LM test

Data	F-statistic
1977 until 1998	5.2533***
1988 until 1998	1.5177

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

8.2 Estimates for an optimal timeframe

Table 10: Estimates of the estimation timeframe, which is different per variable

$\hat{\gamma}$ per variable	estimation timeframe
$\hat{\gamma}_1$	19
$\hat{\gamma}_2$	16
$\hat{\gamma}_3$	13
$\hat{\gamma}_4$	13
$\hat{\gamma}_5$	13
$\hat{\gamma}_6$	13
$\hat{\gamma}_7$	9

Table 11: The estimates of the variables of table 4 for an optimal estimation window using MIP

Variables of model 4.1 for which the parameters should be estimated over a certain timeframe	variable 1	variable 2	variable 3	variable 4
$Constant$	19,055	-5,123	0,000	-8,962
$Q_{1,t}$	16,000	0,147	0,000	31,240
$ln(GDP_{t-1}) - ln(GDP_{t-2})$	13,000	-5,071	-0,603	0,000
$ln(GDP_{t-4}) - ln(GDP_{t-5})$	13,000	-5,425	5,781	0,000
$ln(S_t) - ln(S_{t-1})$	13,000	-0,047	-13,198	6,084
$ln(S_{t-1}) - ln(S_{t-2})$	12,475	-7,512	18,711	-0,858
$ln(S_{t-2}) - ln(S_{t-3})$	8,984	23,600	-0,001	0,353

Table 12: The estimates of the variables of table 4, obtained by OLS, to predict the estimation timeframe of the parameters of model 4.1

Variables of model 4.1 for which the parameters should be estimated over a certain timeframe	variable 1	variable 2	variable 3	variable 4
$Constant$	8.032***	-21.827*	13.495	-29.782
$Q_{1,t}$	-32.709**	3.253**	14.639*	8.006
$ln(GDP_{t-1}) - ln(GDP_{t-2})$	-120.85	8.658	7.8	11.318
$ln(GDP_{t-4}) - ln(GDP_{t-5})$	4.846***	15.688*	16.155	23.083
$ln(S_t) - ln(S_{t-1})$	-18.212	2.327	47.693	40.556
$ln(S_{t-1}) - ln(S_{t-2})$	6.48***	-7.607	25.98***	31.112
$ln(S_{t-2}) - ln(S_{t-3})$	-129.411*	-1.427	12.11*	7.21

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.