Can Simple Combination Strategies Improve Forecasts Of Swedish Inflation?

Andreas Leandersson

May 2017

Master Thesis
Lund University, Department Of Economics
Supervisor: Joakim Westerlund

Abstract

In this paper I use a procedure for selecting, combining and forecasting VAR models in order to improve the prediction of Swedish inflation. Three simple average combinations are evaluated with a number of tools. I conclude that a simple combination strategy improve the forecast ability and reduce the root mean square errors. The performance is an improvement over both a simple benchmark and a previous top performing individual model. This result are consistent for shorter horizons, while on longer horizons all models tend to normalize with more similar performance. Furthermore I compare the models forecast performance using the Diebold-Mariano test. I found that the combination models performance is not significantly different from the benchmark models. The conclusion therefore is that even if the results are promising, more studies are required to fully evaluate the used procedure in order to include longer sample and for other inflation environments.

KEYWORDS: Forecasting, VAR models, Combination forecast, Average combination, Swedish inflation

---

Special thanks to Joakim Westerlund for valuable feedback and discussion during the work of this paper.

1 Introduction

Forecast inflation, the central banks dream of perfectly control and stabilize the inflation over time, has been targeted in the academia for several years now. Most central banks these days use a model-based approach together with gut feeling and experience to determine the policy for achieving a stable inflation level. This has been successful to some extent. However, rising critique against central banks inability to keep the inflation at a desired level has emerged in line with negative interest rates and quantitative easing stimulations.

The struggle to produce reliable forecast models that holds over time has been known for several years, where Stock and Watson (2007), Kilian and Taylor (2003), Gürkaynak et al. (2013) among others found it very hard for models to beat a simple or a naive benchmark model. Rossi and Sekhposyan (2010) argue that a reason for this could be the great moderation were the volatility of macroeconomic variables was significantly reduced in the mid 1980s.

The Swedish Riksbank has since 1993 been working with a specified goal to have a stable annual inflation rate of two percent. From this point on, the monetary analysis have been supported by models in order to forecast key economic variables, much of which is well described by Iversen et al. (2016). Several different types of models are used for this purpose and the combined (judgment) analysis is the basis for the monetary policy decision implemented by the executive board.

Goodfriend and King (2015) reviewed the Swedish Riksbank monetary policy for the period 2010 to 2015 and how the bank has managed the post crisis economy. Their critique was profound.

Their main conclusion was that the executive board had been too reliant on the forecast models that assumed full credibility in the quantified policy path. They also criticized the executive board for interpreting the models to mechanical since they did not considered the models relatively bad performance in detecting the crisis. Of course, the Swedish Riksbank was not the only central bank that consistently overestimated the rate of return to a target inflation level (Iversen et al. (2016)).

Research has also shown that it is a bad idea to choose best the forecast model based on historic track record. Simple combination strategies has been superior to previous well performing models in studies by Gupta and Wilton (1987) and Makridakis and Winkler (1983).

A pragmatic approach to deal with such criticism would be to examine a broader set of variables and the weighting of several models. In such set up the reliability towards specified models becomes smaller as numerous combinations are evaluated. A more dynamic selection of models can be adopted for different phases in the business cycle.

The idea to combine several forecast models is intuitive appealing together with

the portfolio diversification argument (see Bates and Granger (1969)). Clemen (1989) argue the reason for this is that the effect from misspecification of individual models is reduced.

Timmermann (2006) address the risk for unobservable information in individual models. For example, structural breaks in a sample can affect different types of models in different ways. Some models might be affected temporarily, while other take longer time for adaption. This is also the main argument for not pooling several models into a large supermodel. Pesaran and Timmermann (2007) argue that combining several models will on average perform better than individual models when both known and unknown structural breaks is present.

This idea, to weight several models in order to improve their forecasting ability has been tested with positive results in an early paper by Bates and Granger (1969). In fact, the well performance of the procedure have been documented in the literature ever since. Interestingly, equally weighted forecasts seem to perform very well according to Clemen (1989), Clark and McCracken (2010) and Aiolfi et al. (2010). Figlewski and Urich (1983) mean that the main reason simpler combination strategies outperform more sophisticated is because of the optimal weights do no have to be estimated. Palm and Zellner (1992) also points out that the diversification argument is stronger for simple combinations since the estimation of weights could be affected by uncertainty and sample errors.

It is therefore quite surprising that such weighted forecast are not included in the Riksbanks official set of models to this day (see Iversen et al. (2016)). However, in recent years some attempts have been made to combine several forecast into a single one, with promising results (see Adolfson et al. (2007)). Other central banks have used this procedure more frequently. Central bank of Switzerland for example, has rather successfully found this method beating a naive comparison forecast (see Lack et al. (2006)). It could be argued if their policy interpretation has been more successful than the Swedish Riksbank, since both countries struggle with low inflation and an expansive central bank.

The purpose of this paper is to evaluate if combination strategies can improve the forecast of Swedish inflation. I evaluate how the methodology for selecting, combining and forecasting models presented in Lack et al. (2006) works when forecasting Swedish inflation. By extending the method to include more solid comparison tools I will be able to provide an extensive and fair evaluation of this methodology when compared to Lack et al. (2006).

My results indicate that combining several forecasts improves the forecast prediction for Swedish inflation over a simple benchmark. This is in line with Lack et al. (2006), and strengthens the methodology approach. However, by extending the evaluation methods, I cannot prove that the difference in performance is significantly better.

The remaining of this thesis is structured as follows: Section 2 Data discussion, Section 3 Forecasting approach and methodology, in section 4 the results are presented, Section 5 conclusion and discussion and lastly, Appendix is provided in section 6.

2 Data

The variable I intend to forecast is the quarterly consumer price index (CPI). When analyzing the historical data for CPI, it is clear that recent years can be characterized as a low inflation environment. As seen in Figure 1, from 1993 and forward the inflation rate has fluctuated in a stable range between -1.4 and 4 percent. Iversen et al. (2016) argue that the low inflation environment could be a consequence of a number of shifts in economic policy. They are mentioning both the change of exchange rate regime in 1992, and the implementation of inflation target in 1993 as possible causes. It was also after the introduction of the inflation target when a more model-based approach was adopted to forecast key variables in the Swedish economy.

The sample contain 96 quarters between 1993Q1 and 2016Q4. All variables are collected from OECD. I use 10 macroeconomic variables and together with CPI, this makes up for a total of 1056 observations.

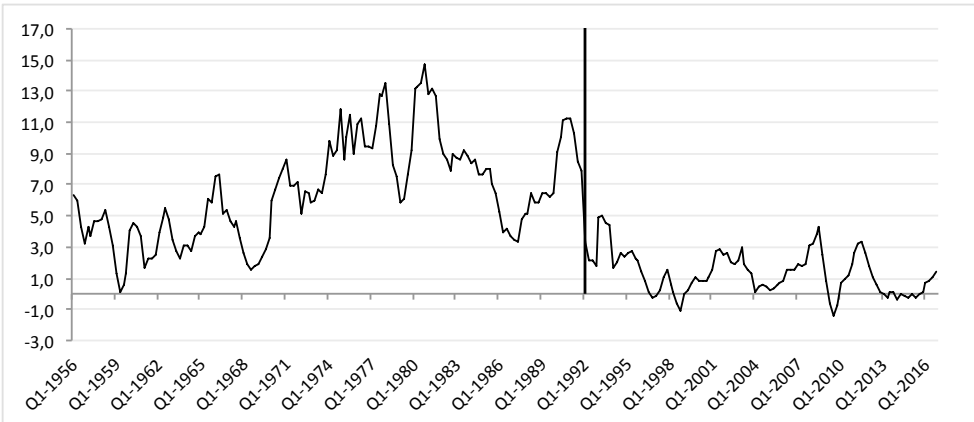

Figure 1: Quarterly inflation, percentage change from previous years same period

I select macroeconomic variables that been proved to be a good predictior of Swedish inflation in the past and what is commonly used at the official Swedish institutions in forecasting inflation. Riksbanken is quite transparent in this matter, and several research papers are available on the subject. Iversen et al. (2016) lay out the foundation of the forecasting made at the central bank. Commonly used macroeconomic variables that I include in this study are CPI, Hourly earnings, GDP and the exchange rate; in my case I use USDSEK. In addition I use the short-term in-

terest rate, which has been used previously at Riksbanken by Adolfson et al. (2007). Same types of variables are also used by Mossfeldt et al. (2016) and Villani (2009). I also want to include some real activity measurement and Rossi and Sekhposyan (2010) has evaluated following variables for US inflation forecasts; Manufacturing index, Import goods, retail sales and industrial production. Finally I also include PPI and M3 to complete the price index and money variables. These 10 variables will be used to forecast Swedish inflation. See appendix A, for full specification and transformation of the data. All variables are verified stationary by rejecting the null hypothesis of a unit root from the augmented Dickey Fuller test.

The selection of time period is particular made by having two main objectives in mind. First, even though CPI data are available since the 1950s other macroeconomic variables are not. The limitation of available macroeconomic variables is narrowing the sample size and is a common constraint. Second, Lucas (1976) argue that previous mentioned policy shifts can lead to poor forecast accuracy. The effects of structural breaks can be minimized with a combination strategy. Pesaran and Timmermann (2007) found that the combination of several models with different speed of adjustment will on average outperform the forecast from individual models.

The implications bounded by these limits are that the chosen sample period only reflects a low-inflation environment. Therefore no comparison with other types of environment is possible. Chortareas et al. (2002) argues that this could be a threat to the robustness of the forecast. However, studies have found that in low inflation environment in general, and after the great moderation in particular, it is more difficult for sophisticated models to perform better than a simple benchmark model (see Gürkaynak et al. (2013), Fisher et al. (2002) and Stock and Watson (2007)). Lack et al. (2006) found that this is also the case in combination forecast. He suggests that the reason for this is that low-inflation environment also implies lower volatility of inflation. It is under these conditions simple models tend to perform better which make them harder to beat. Therefore, one could suspect that the results reported in this paper based on combination forecasts would become even better in a high inflation environment. Although, this would have to be further evaluated.

3 Forecast approach

3.1 The process

The forecast approach is set up as a three-period process. The first part acts as the initial estimation period, while the second part is the evaluation period of all models forecasting performance. From period 2, i will take out the top 10 performing models to use in the final forecast period. The last part is the period of inflation I intend to forecast. An overview of the three-period process is presented in Figure 2. This methodology is also described in Lack et al. (2006).

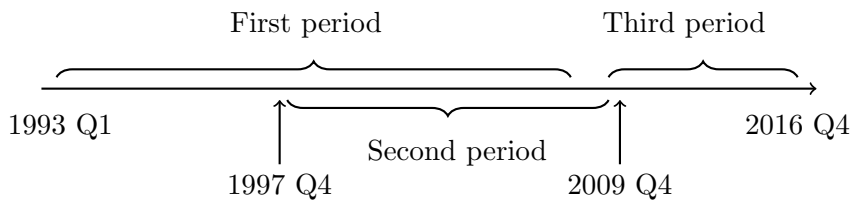

Figure 2: The process of selection, estimation and forecasting models.

The whole sample stretches between period 1993Q1 to 2016Q4. This means that the total period is 96 quarters. Starting off with the first period, this is the sample the estimation is based on for the second periods evaluation forecast. The first forecast in the second period is estimated over 20 quarters between the 1993Q1 and 1997Q4.

Using a recursive estimation window, in contrast to Lack et al. (2006), the sample is increased by one observation for each new quarter that is forecasted in the second period. This means that the last forecast in the second period is based on the estimation window between 1993Q1 and 2009Q2. In contrast to the first forecast for the period, the last one is estimated over 66 quarters.

The forecasted quarter is stored and the next forecast uses the actual value from the estimation sample. The main reason for using a recursive forecast method is due to the limitations that is further addressed in section 2. The risk associated with a recursive forecast method is that the estimation could include structural breaks that distort the robustness and the prediction of the parameter estimates. Such breaks could for example be a shift in monetary policy or a change of exchange rate regime. As this concern is already addressed in the choice of sample period, there is an advantage to include more observations in the estimation sample for every quarter that is forecasted. This will not only improve the fit of the models, but also generate a better forecast.

From the pool of macroeconomic variables I construct several vector autoregressive (VAR) models to test all possible combination. By having CPI fixed and

included in every single model and by restricting the range of each model to include at least 2 and as much as 5 variables, this sums up to a total of 385 models evaluated.

The restrictions are primarily made under consideration of the sample size and to set a limit on how many combinations of variables that is possible.

In the evaluation period all models have the same structure of 1 lag and the performance is evaluated on the out-of-sample forecast for the horizon of 1 quarter ahead. In the third period both optimal lag-structure and horizon will be addressed, but the selection of the top 10 performing models is completed under these premises. All models are estimated with OLS and the forecast procedure is performed in Stata.

The purpose of the second period in the process is to evaluate the forecast performance of each VAR model. A forecast is stored for the 47 quarters between 1998Q1 and 2009Q3. In total, after looping through all models over the whole evaluation period, 18 095 forecast are stored. The top 10 performing models to forecast Swedish inflation, based on their root mean square error (RMSE) are taken out and used in the last part of the process.

The final and third part is the main forecast period. The top 10 performing models selected in previous period is now estimated for the whole sample, period 1 and 2 in Figure 2, to forecast the third period. This generates 29 quarters of forecasts for the period between 2009Q4 and 2016Q4.

In the evaluation of the final performance, each individual forecast will be run through a series of combinations strategies based on averaging techniques. This will test if the combination of several models forecasts can reduce the RMSE and generate a better prediction for Swedish inflation.

3.2 Combination averaging techniques

Clemen (1989), Clark and McCracken (2010), Aiolfi et al. (2010), and many more, have provided evidence that there is an advantage of combining several individual forecast into one. To combine models over a simple averaging technique improve the forecast ability the most. Gupta and Wilton (1987) argue that combination with equal weights perform well when the relative variance of several models forecast error is small. Timmermann (2006) argue that in theory, this is the most critical point when considering averaging combinations. In my case the relative performance between the individual models is small, thus the conditions for adopting a simple equal weighting combination appears optimal.

Therefore, under the justification that using simple averaging techniques beats more sophisticated weighting schemes, the focus on the combination strategies will be limited to simple equal weighted strategies. For this purpose I consider three averaging techniques. The first one is a simple equal weighted average. The second one is the median, and the third is the trimmed mean where the 20 percent of the outline performance is removed, meaning that the top and bottom 2 models is

removed before taking the average. These averaging combinations will be adopted and evaluated by RMSE in the final period. These simple averaging strategies has also been evaluated with promising results by Clark and McCracken (2010).

3.3 Performance evaluation

The performance of the models forecasting ability is evaluated by their RMSE, which is both the standard academic measurement, and frequently used in the Swedish central banks official research paper, for example see Adolfson et al. (2007) and Iversen et al. (2016) among others. Equation 1 represent the calculation of RMSE.

$$RMSE = \sqrt{\frac{1}{T_0}\sum_{t=t_0}^{T_0}(\hat{\pi}_{t+h} - \pi_{t+h})^2} = \sqrt{\frac{1}{T_0}\sum_{t=t_0}^{T_0}(\varepsilon_{t+h})^2} \tag{1}$$

where T_0 is the total number of forecasts made, $\hat{\pi}_{t+h}$ is the forecast at time t_h which is minus the actual inflation π_{t+h} at time t_h. The squared brackets represents the forecasting error, ε_{t+h}. The lower the RMSE the better prediction is made.

I want to evaluate the performance of the models in context of a relatively simple benchmark. This is standard as to see if the implementation of more sophisticated models are justified over the use of simple ones. In fact, Gürkaynak et al. (2013) and others argues that there is particularly hard to beat a simple model in forecasting inflation. For this purpose, I examine a variant of Henri Theil (1971) U-statistic, which is also used by Lack et al. (2006).

$$U_h = \frac{\sqrt{\sum_{t=t_0}^{T_0}(\hat{\pi}_{t+h} - \pi_{t+h})^2}}{\sqrt{\sum_{t=t_0}^{T_0}(\pi^*{}_{t+h} - \pi_{t+h})^2}} \tag{2}$$

The U-statistic is the RMSE of the evaluated model divided by the RMSE of the benchmark model. It implies that if U_h is above one, the simple benchmark forecast inflation better than the evaluated model. The opposite holds if U_h is below one. For a value of 1, both models have the same RMSE.

I will use two simple models as benchmark. The first is a naive model that simply forecast the next quarter with the previous actual inflation rate. This no-change forecast is widely used in evaluating the use of having more sophisticated models or not. This is also used as comparison model in Lack et al. (2006).

As this model generally performs worse at longer horizons I also include another simple model, the first order autoregressive model. This AR(1) is commonly used

and is considered hard to beat (see Gürkaynak et al. (2013), Kilian and Taylor (2003), Jore et al. (2010), Atkeson and Ohanian (2001) and many more).

As final test, I will evaluate if the performance difference between models is statistically significant or not. This test was developed by Diebold and Mariano (2002), with the null hypothesis that the models have the same predictive ability. Later on Diebold (2015) criticized the use of DM as an absolute model comparison. He argues that it instead should be used as a forecast comparison to determine if the performance difference can be proved as statistically significant or rather to be seen as a consequence of pure luck.

Diebold and Mariano (2002) is a two-tailed test where the test statistics is

$$S = \frac{\bar{d}}{(\widehat{LRV}_{\bar{d}}/T_0)^2} \tag{3}$$

and where $\bar{d}$ is equal to:

$$\bar{d} = T_0^{-1} \sum_{t=t_0}^{T} d_t \tag{4}$$

and d_t is the loss differential $d_t = L(\varepsilon^1_{t+h|t}) - L(\varepsilon^2_{t+h|t})$ of the squared forecast error for model 1 and 2. The denominator is the asymptotic long run variance $\widehat{LRV}$, which according to Newey and West (1986) is consistently estimated with Bartlett Kernel method.

The test shows that $S \sim N(0,1)$ under the null hypothesis of equal predictive accuracy.

4 Results

I start with the characteristics of the selected models for the final period. Table 1 presents the variable outcome for the top 10 models. Average model has 3,7 variables which makes up to 37 variables for all 10 models. CPI is by restriction included in every model.

Besides that, retail sales is clearly the most important variable to forecast Swedish inflation. It appears in all top performing models. This is followed by Producer price index (PPI) that appears in 60 percent of the models. Three variables appear each in 3 out of 10 models. These are GDP, Import goods and the exchange rate of USDSEK. Hourly earnings and the short-term interest rate appear only in 10 percent of the models. Three variables did not make it into any one of the optimal models; these are the manufacturing index (MEI), M3 and industrial production.

Variable	Number of apperance	Percent of total apperance	Percent of model apperance
CPI	10	27,03	100
PPI	6	16,22	60
Hourly Earnings	1	2,70	10
Manufacturing index (MEI)	0	0	0
M3	0	0	0
USDSEK	3	8,11	30
Short-term rate	1	2,70	10
Import Goods	3	8,11	30
GDP	3	8,11	30
Retail Sales	10	27,03	100
Industrial production	0	0	0
Average number of variables in model			3,7
Total number of variables in all models			37
Total number of models evaluated			385

Table 1: Variable appearance in top performing models

In the following presentation of the results I will, beside reporting the three average combinations and the two simple benchmark models, also report two more models. The first one is the evaluation periods best performing model. From now on this will be referred to as the "original top model". This comparison will determine if the combination forecast is an improvement over a model that has been the best predictor of inflation in the past. Previous research from Gupta and Wilton (1987) and Makridakis and Winkler (1983) suggests that so is the case. The second model is the optimal top model for each specific test that is undertaken. As it shows, for different test specifics on horizon and lags, different models generate the best forecast. From now on this will be presented as the "optimal top model". The optimal top model is unachievable and should rather be seen as a benchmark that

assumes perfect insights of model prediction.

In the selection period, all models were evaluated by the standard specifics with 1 lag. This might not be optimal, as it could potentially leave out important variable dynamics and bias the estimation of the models. Lack et al. (2006) argue that this results in a trade off between having to many or to few lags. In order to optimize the lag-structure I evaluate two information criterion, Akaike information criterion (AIC) and Bayesian information criterion (BIC). AIC suggests 5 lags while BIC suggests 2 lags. As the general information criterion propose different number of lags I precede by testing it in practice. Figure 3 show the results by number of lags and their RMSE. It is clear that all models achieve the best result for 2 lags, which is in line on what BIC suggested. Therefore, 2 lags will be used in the VAR models in the rest of the paper.

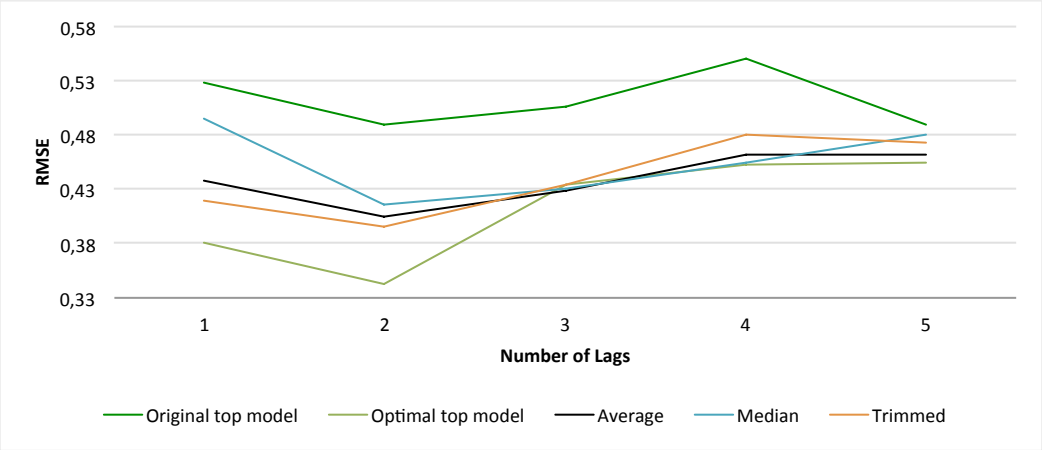

Figure 3: Model performance for different number of lags

It is interesting to evaluate how well the combination models perform for horizons longer than a quarter. Especially since Iversen et al. (2016) has shown that the models at the Swedish central banks generally performs well on shorter horizons, but has harder to predict inflation in the long run. Goodfriend and King (2015) also criticized the Riksbank for holding on to models after the crisis that didnt even predicted it. By successively increasing the horizon I can evaluate the performance and determine the optimal horizon forecast. Figure 4 present the results for all combination models as well as for the optimal- and original top model.

The best performing model is the optimal top model that has the lowest RMSE in horizon 2. For horizons longer than that it gradually increase its RMSE to perform as well as the other ones. The combination strategies tend to perform similarly for all horizons. While it peak at horizon 4 it gradually perform worse after that. The most stable model is the original top model, however it also has the worst

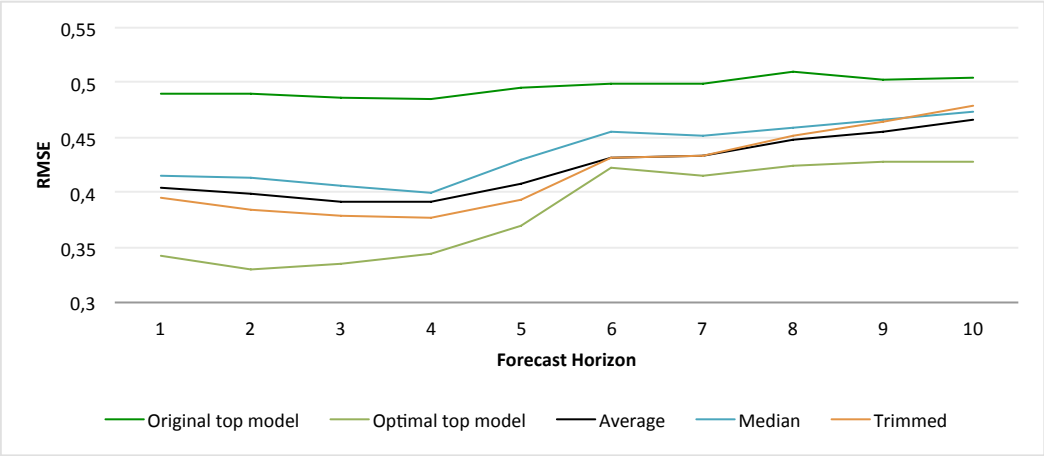

Figure 4: Performance of combination strategies for different horizons

performance. This indicates that the models perform better in the short run and that the combination strategies improve the forecast ability when comparing it to a top performing model from a previous period. This is in line with previous research by Gupta and Wilton (1987) and Makridakis and Winkler (1983).There seems to be small differences between the types of averaging techniques. For longer horizons, all models tend to move together as their RMSE becomes more alike.

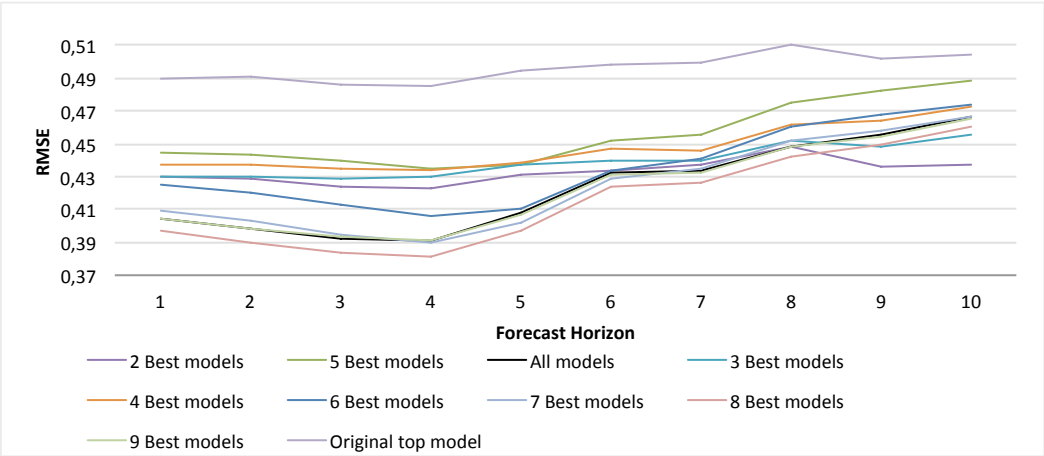

Figure 5: Performance of combination forecast when number of models varies

The small variance between the combination strategies suggests that the choice of averaging combination is negligible. Further test on the average set up is presented in Figure 5, where averaging over different number of models is evaluated for both horizon and RMSE. Only the equal weighted average combination is addressed.

The result from Figure 5 show more variation in shorter horizons, while on longer horizons all models tends to move together and perform equally good (bad). The optimal horizon is as before equal to 4. The gain for combining forecasts is clear in

comparison to the original top model. What is striking is that the RMSE improves when averaging over more models. This despite the fact that the average of smaller number contains better performing models from the evaluation period. Since it is hard to pick the optimal model, combination over several well-performing models improve the results, at least for shorter horizons.

Finally, when lag-length, horizon and averaging combination have been optimized, I can now evaluate them against the simple benchmark models. Using a variant of Theil (1971) U-statistic the relative performance is comparable. Again, if U_h is above 1, the simple benchmark forecast inflation better than the evaluated model. The opposite holds if U_h is below one. For a value equal to 1, both models have the same RMSE.

In Figure 6, I compare the original top model, optimal top model and one average combination against the naive benchmark. Following previous result where choice of average technique had little effect and where taking the average over more models are favorable, I only present the equal weighted average over all top 10 models in Figure 6. The models are estimated with two lags as previous determined being the optimal choice.

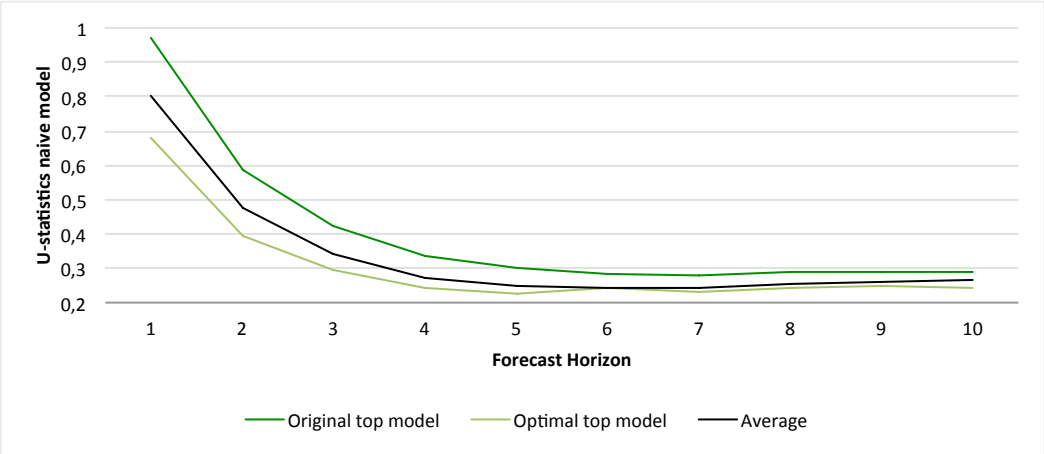

Figure 6: Model performance when comparing with the naive benchmark

As presented in Figure 6 the naive benchmark is outperformed by all models. For 1 horizon, the forecast ability for the compared models is only marginally better. The marginal improvement of forecasting models decreases as the horizon increases. The optimal model is still the best performing one, but as horizon increases, all models tend to normalize and perform equally good. As previous discussed, it is impossible to pick the optimal model before the forecast is made. There is no consistent top model for all horizons in the final period. This makes the argument for combining several models stronger. Especially when the combination strategy

outperform the original top model for all consecutive horizons. Again the selection periods top model has the worst performance in the final period. It has marginally better RMSE for the first horizon compared to the benchmark, while there is an improvement over longer horizons.

It is clear that the naive benchmark fall behind on longer horizon and has, in that sense, to be considered a bad forecast model. This makes intuitively sense, when increasing the horizons, the further away the naive models last observation gets. Since there are some fluctuation in CPI, comparing with a naive benchmark on longer horizons strongly bias the result in favor of the more dynamic models. The naive forecast is more suitable for evaluation of shorter horizons.

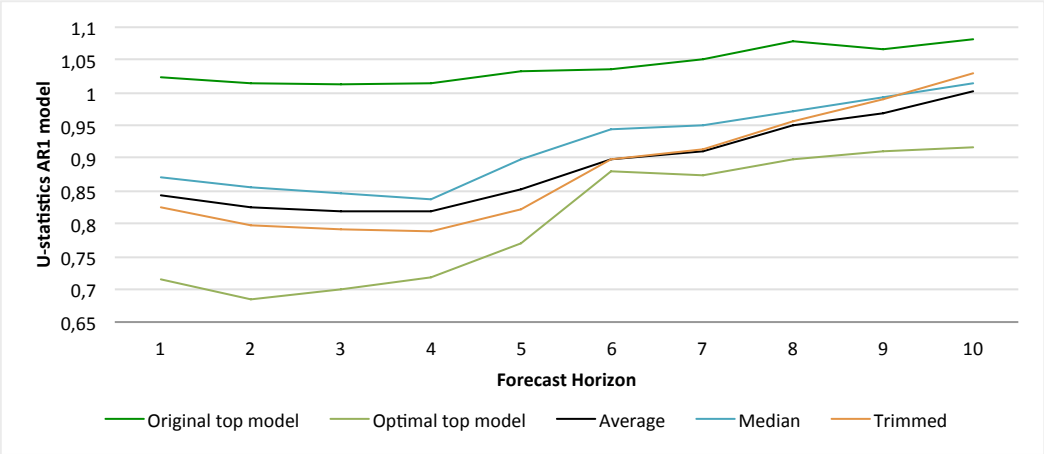

Figure 7: Model performance when comparing with the AR(1) benchmark

The shortcomings of the naive model as benchmark for longer horizon suggests that the simple AR(1) model would be a more realistic and fair comparison. In Figure 7 the results from the U_h-statistic is presented with AR(1) as the simple benchmark model. The results demonstrates that the original top model is performing worse than the AR(1) for all horizons. Earlier empirical research has shown it is hard to outperform a simple benchmark model. As before, the optimal top model is the best performing model for all horizon. There is a gain to combine models on shorter horizon in comparison to the AR(1). However, as the horizon increase all evaluated models tend to move together and forecast inflation with the same accuracy. Optimal horizon for the combination strategies is still at 4. Overall the performance is much worse than when comparing to the naive model.

Figure 7 provides the importance of choosing a realistic comparison measurement. Since it is generally hard to beat a simple model, one has to be careful in order to make the optimal policy decision. This is something that Lack et al. (2006) fails with which undermines the evaluation of the methodology. Even though

	Naive model	AR(1)
Optimal top model	-0.526	-0.222
Original top model	-0.734	0.016
Average	-0.522	-0.172
Median	-0.517	-0.136
Trimmed mean	-0,521	-0,177

P-values: ***=1%,**=5%,*=10%

Table 2: Diebold Mariano test

this paper has presented the same results that a combination strategy improves the forecast ability, the effect is not as strong as previous suggested.

Still, the improved performance when combining several models into one forecast cannot be ignored. However, evaluating the results solely based on the RMSE is not sufficient. To test if the difference in forecast performance is significant I continue with the Diebold-Mariano test. From previous results, the optimal performance is reached with 2 lags and at a 4 step ahead forecast. Table 2 show the result of this test for the all models compared with the two simple benchmarks.

The Diebold Mariano test cannot reject the null-hypothesis for any of the averaging models or for the optimal- and original top model. This means that the forecast performance of the evaluated models and the combination strategies are not significantly different from the two benchmark models for this period of time.

The overall result in this paper strengthens the outcomes and the methodology approach used by Lack et al. (2006). However, I cannot significantly show that either of the combination models or the optimal model performs different from the benchmark models. This means that even though the result show promising indications that averaging combinations improve the prediction of Swedish inflation over simple benchmark models; I cannot prove that so actually is the case.

5 Conclusion & Discussion

I have in this paper tested a method for selecting, forecasting and evaluating VAR models to investigate if simple combination strategies can improve the prediction of Swedish inflation. My results strengthen the use and understanding of this method in some sense, which can provide a more dynamic evaluation of a larger set of variables and model constructions. In the light of the criticism the Swedish Riksbank received in Goodfriend and King (2015) review of the post crisis policy actions, there is a demand for a more dynamic approach towards existing models. I believe this paper will be a valuable contribution in that sense. Interestingly, the top perform-

ing model from the evaluation period consistently underperformed in every test. It highlights the danger of being fixed to previous top performing models. My results indicates that combination strategies are a way around this problem, as it reduces the RMSE. This is in line with previous research by Makridakis and Winkler (1983) and Gupta and Wilton (1987).
Since the optimal top model is outperforming all other models for shorter horizons, one can thus understand that central banks and analysts are trying to achieve this model. However, it seems very difficult to do so. My results show that there was no single model that consistently outperformed the other ones. For different horizon, different model was the best performing one. Therefore the pursuing of finding the optimal model has to be set in relation to the difficulties of doing so, especially over time where models performance are changing. The risk of using the wrong or an outdated set of models as base for policy decision can be devastating. Simple combination strategies could in such cases fill a meaningful part.

I have concentrated this study to VAR models, and with the previous research by Adolfson et al. (2007) this advances the groundwork of combining models to forecast Swedish inflation. But there is more to do. While the results are promising for combination models, I cannot prove that the performance is significantly different from the benchmarks. Depending on what benchmark one compare with, combination models performs better for longer or shorter horizons. I have argued and showed in this paper that the naive benchmark bias the results in favor of the sophisticated models, at least for longer horizons. Even though the performance advantage could not be proved significant, I would not ignore the importance to further investigate combination models. Because of the limitations in data I had to restrict the research for a period of time that is characterized as low-inflation environment. It is not certain that this environment will go on forever, thus it would be interesting to evaluate if the results changes under different circumstances. Since Lack et al. (2006) found that the performance was much stronger in high-inflation environments, one could suspect that so also is the case for Swedish inflation forecasts. Maybe then it is also possible to prove that the models performance is significantly different from the benchmark. For policy implications, this methodology would also have to be evaluated and compared to the models currently used at the Riksbank.

References

Malin Adolfson, Michael Andersson, Jesper Lindé, Mattias Villani, and Anders Vredin. Modern forecasting models in action: Improving macroeconomic analyses at central banks. 2007.

Marco Aiolfi, Carlos Capistrán, and Allan G Timmermann. Forecast combinations. 2010.

Andrew Atkeson and Lee E Ohanian. Are phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis. Quarterly Review-Federal Reserve Bank of Minneapolis*, 25(1):2, 2001.

John M Bates and Clive WJ Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969.

Georgios Chortareas, David Stasavage, and Gabriel Sterne. Does it pay to be transparent? international evidence from central bank forecasts. *Review-Federal Reserve Bank of Saint Louis*, 84(4):99–118, 2002.

Todd E Clark and Michael W McCracken. Averaging forecasts from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(1):5–29, 2010.

Robert T Clemen. Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583, 1989.

Francis X Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1, 2015.

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

Stephen Figlewski and Thomas Urich. Optimal aggregation of money supply forecasts: Accuracy, profitability and market efficiency. *The Journal of Finance*, 38 (3):695–710, 1983.

Jonas DM Fisher, Chin T Liu, and Ruilin Zhou. When can we forecast inflation? *Economic Perspectives-Federal Reserve Bank of Chicago*, 26(1):32–44, 2002.

Marvin Goodfriend and Mervyn King. Review of the riksbanks monetary policy 2010–2015. *Report prepared for the Finance Committee of the Swedish Parliament (Riksdag)*, 2015.

Sunil Gupta and Peter C Wilton. Combination of forecasts: An extension. *Management Science*, 33(3):356–372, 1987.

Refet S Gürkaynak, Burçin Kısacıkoğlu, and Barbara Rossi. Do dsge models forecast more accurately out-of-sample than var models? the views expressed in this article are those of the authors. In *VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims*, pages 27–79. Emerald Group Publishing Limited, 2013.

Jens Iversen, Stefan Laséen, Henrik Lundvall, and Ulf Söderström. Real-time forecasting for monetary policy analysis: The case of sveriges riksbank. 2016.

Anne Sofie Jore, James Mitchell, and Shaun P Vahey. Combining forecast densities from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(4): 621–634, 2010.

Lutz Kilian and Mark P Taylor. Why is it so difficult to beat the random walk forecast of exchange rates? *Journal of International Economics*, 60(1):85–107, 2003.

Caesar Lack et al. Forecasting swiss inflation using var models. Technical report, Swiss National Bank, 2006.

Robert E Lucas. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pages 19–46. Elsevier, 1976.

Spyros Makridakis and Robert L Winkler. Averages of forecasts: Some empirical results. *Management Science*, 29(9):987–996, 1983.

Marcus Mossfeldt, Pär Stockhammar, et al. Forecasting goods and services inflation in sweden. Technical report, 2016.

Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix, 1986.

Franz C Palm and Arnold Zellner. To combine or not to combine? issues of combining forecasts. *Journal of Forecasting*, 11(8):687–701, 1992.

M Hashem Pesaran and Allan Timmermann. Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161, 2007.

Barbara Rossi and Tatevik Sekhposyan. Have economic models forecasting performance for us output growth and inflation changed over time, and when? *International Journal of Forecasting*, 26(4):808–835, 2010.

James H Stock and Mark W Watson. Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39(s1):3–33, 2007.

Henri Theil. Applied economic forecasting. 1971.

Allan Timmermann. Forecast combinations. *Handbook of economic forecasting*, 1: 135–196, 2006.

Mattias Villani. Steady-state priors for vector autoregressions. *Journal of Applied Econometrics*, 24(4):630–650, 2009.

A Appendix

Variable	Transformation	Source
CPI	Δlv	OECD.Stat
PPI	Δlv	OECD.Stat
Hourly Earnings	Δlv	OECD.Stat
Manufacturing index (MEI)	Δlv	OECD.Stat
M3	Δlv	OECD.Stat
USDSEK	Δln	OECD.Stat
Short-term rate	Δlv	OECD.Stat
Import Goods	Δln	OECD.Stat
GDP	Δln	OECD.Stat
Retail Sales	Δlv	OECD.Stat
Industrial production	Δlv	OECD.Stat

Table 3: Macroeconomic variables, transformation and information, Δ = first difference, ln = natural logarithm and lv = level.