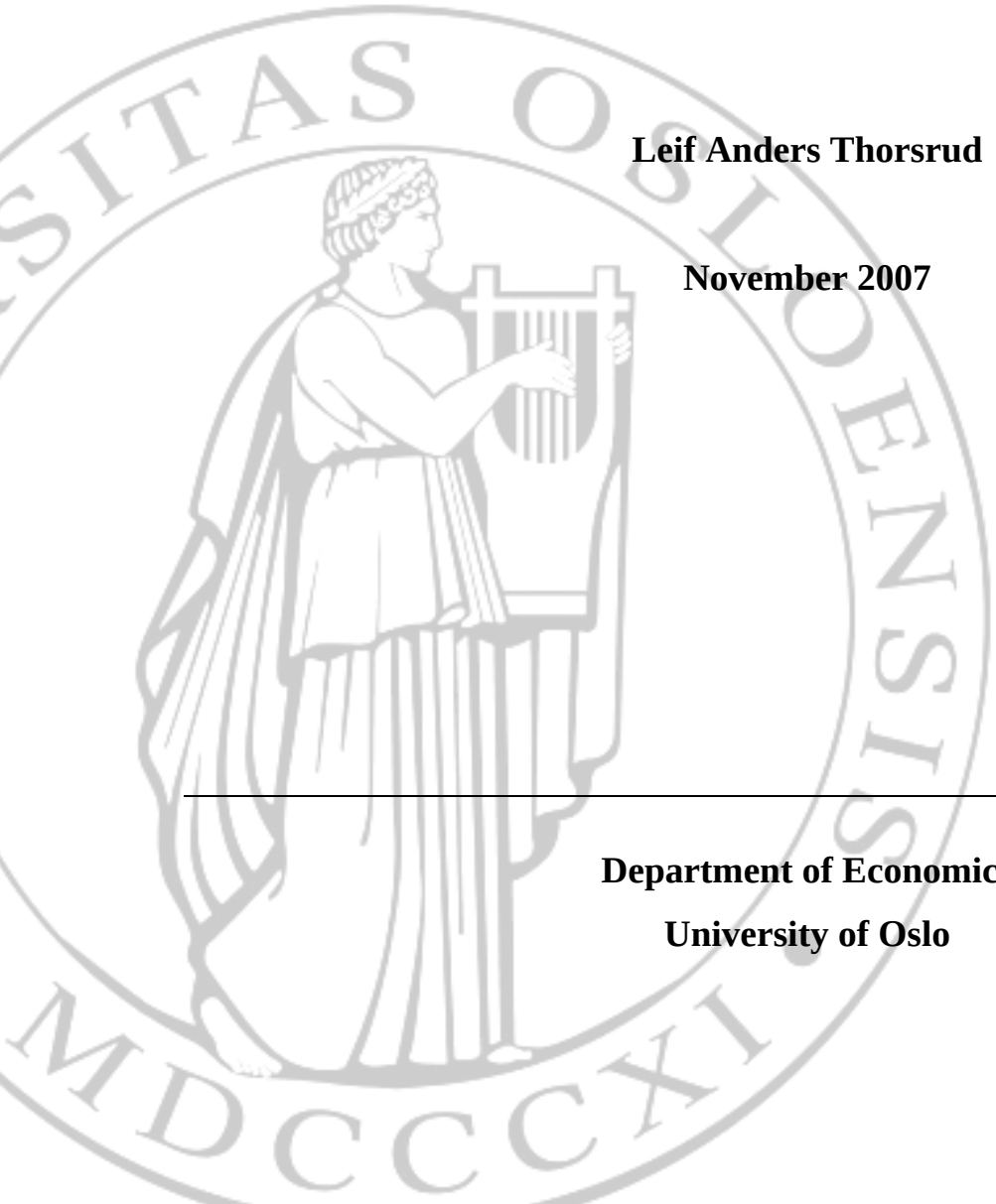
Master thesis for the Master of Philosophy in Economics degree

Forecasting Inflation in Real-time

Leif Anders Thorsrud

November 2007



Department of Economics

University of Oslo

Preface

I would like to thank Norges Bank (the central bank of Norway) for giving me the time, opportunity and economic funding to write this thesis. Christian Kascha and my other colleagues at the Economics Department have come with inspiring help throughout the working process. My supervisor, Qaisar Farooq Akram, has always been available for questions and valuable discussions. I would especially like to thank Anne Sofie Jore at the Economics Department who inspired me to write this thesis, and who has also done a great effort making the real-time database available. The views, conclusions and any remaining errors represented in this thesis are my own only.

Leif Anders Thorsrud

Oslo

9th November 2007

Contents

List of Tables

List of Figures

1 Introduction

Flexible inflation targeting has become the preferred policy among a growing number of central banks over the last decades. Due to the lag between interest rates and inflation, optimal monetary policy in this framework is essentially about forecasting inflation (Svensson and Woodford, 2003). The output gap, measuring the deviation of output from potential, has a key role in this regard. Through different transition mechanisms a positive output gap leads to inflation. For central banks aiming at a flexible inflation target, an appropriate policy response to the observed pressure in the economy will not only help stabilize inflation at a desired level, but also stabilize output (Svensson, 1997 and 2000).

If the policy reactions are going to be proper, the measure of the output gap has to be adequate. As demonstrated in this and other analysis it seldom is.[1] There are basically two factors making the derivation of the output gap difficult. The first concerns the estimation procedure. Since one fails to reject the hypothesis of a unit root in macroeconomic time series, the long run trend of output can no longer be treated as deterministic; see e.g. Nelson and Plosser (1982). Accordingly, the computation of potential output has to take into consideration the estimation of a stochastic trend, which greatly complicates the measuring of potential output and the output gap.

The second factor concerns the real-time nature at which central banks have to conduct monetary policy: Decisions are based on highly uncertain data, which are subjected to substantial revisions. This is especially true of the output. There are three main reasons for changes to official statistics.

1. The earliest estimates are based on preliminary and incomplete information.

2. Changes to the base year.

3. The national accounts are occasionally subject to major revisions.

[1] See for example Orphanides and van Norden (2002), and Bernhardsen, Eitrheim, Jore and Røisland (2004).

Real-time data is data as it was observed at each point in time, and typically categorized into different vintages describing their time of release, thus taking into account these data revision processes.[2]

In the spirit of Orphanides and van Norden (2005), this paper examines two different methods for extracting the output gap in real-time, and evaluates their performance in forecasting Norwegian inflation. Especially, I question whether the inclusion of the output gap gives any value added in forecasting Norwegian domestic inflation compared to simple autoregressive benchmark models. The answer clearly depends on factors as model specifications, evaluation criteria, the forecasting periods and the quality of the data: The output gap models evaluated are the Hodrick-Prescott filter and the Production function method. As a benchmark forecasting model I employ a linear AR(p) model of inflation. My main forecasting model is a Phillips curve relation including the output gap. These specifications makes it possible to relate inflation to real activity.[3] I have used root mean square forecast errors (RMSFE) to assess the forecasting performance, and the forecasting period has ranged from 94q1 to 06q2. By using real-time data this paper highlights the problems and the uncertainties brought forward by the data revision processes.

To my knowledge real-time forecasting exercises of this kind has not been conducted on Norwegian data before. Bjørnland, Brubakk and Jore (2007) found that models including the output gap gave a better predictive power of inflation than models based on alternative indicators, and that they forecasted significantly better than simple benchmark models, but they did not use real-time data.

Based on real-time data estimations my findings suggests that the inclusion of the output gap makes the out-of-sample forecasts less accurate than what would have been attained if the simpler benchmark models had been used, a finding that is consistent with results reported in Orphanides and van Norden (2005). Some output gap models computed in real-time do however forecast better than the benchmark

[2]Orphanides and van Norden(2002), Bernhardsen , Eitrheim, Jore and Røisland (2005) and Mckenzie (2007) provide evidence that the real-time measure of the output and the output gap are exposed to substantial revisions. Mckenzie (2006) give a more thourough list of the different revisions, and notes a total of eight reasons for revisions of official statistics.

[3]The output gap is assumed to be related to the unemployment gap through the so called Okun's law.

models, but the results seems to be very sensitive to the chosen forecasting period. Further I find that there are considerable differences in forecasting performance between using real-time data, and final vintage data.[4]

The reminder of this paper is organized as follows: Section 2 describes the output gap concept, the output gap models and the real-time data sets that I have used. Section 2–2.2 follow Bjørnland, Brubakk and Jore (2004), and Frøyland and Nymoen (2000) closely. For a more thorough exposition of the output gap, and the different methods to extract it, I refer to the cited papers. Section 2.4 illustrates clearly how the real-time issues affect the output gap estimates. Section 3 presents the forecasting methodology. Sections 4 and 5 present the results and conclusions.[5]

2 Measuring the output gap

The output gap is often understood as the difference between observed production and an underlying unobserved trend which output would revert to in the absence of business cycle fluctuations.[6]

While the observed component is easy to grasp in practice, the unobserved trend or potential production, can be a little more complicated. On the one hand, the economy will have a nearly constant increase in labour, capital and technological progress. This will contribute to a smooth annual growth in potential production, and can be considered as a deterministic trend being a function of time only. On the other hand there are clear signs that the economic potential does not grow in a regular manner. Technological breakthroughs, the access to natural resources, different labour market circumstances and the amount of capital in the economy, factors typically considered as representing the supply side of the economy, may all contribute to alteration in the potential production. If the observed production followed the potential production at all times the output gap would have been zero.

[4]The final vintage in the sample has been 06q2.

[5]I have used Matlab computer software and the Econometrics Toolbox provided by James P. LeSage for my computations. Programming codes can be made available on request.

[6]In many recent macroeconomical models the output gap is understood as real wages divided by the marginal product of labor. Only under very strong conditions is this measure of the output gap comparable to the one used in this paper.

This is hardly ever the case. The economy is not only hit by different supply shocks affecting the potential, but also by a variety of demand shocks. These shocks may be of different magnitude and durability, but they all contribute to business cycle fluctuations.

The observed production can in light of this be divided into three parts; a deterministic trend, changes at the supply side of the economy, and changes at the demand side represented by the output gap.

The output gap and potential production are unobserved components. There are however a variety of methods to apply for extracting the output gap and the potential production. Although they all give similar results, there are important differences. These differences may become more pronounced when dealing with real-time data. In this paper I have considered one univariate method (the Hodrick-Prescott filter(HP)), and one multivariate method (the Production function method (PF))for extracting the output gap. The univariate method uses only information from one time-series, while the multivariate method takes into account a variety of variables.

Practical application and earlier research have been important criteria for my selections. By practical application I mean that the methods chosen should be implemented and widely used in the central bank community as a means of computing the output gap and potential production. I find this an important attribute because dealing with real time-data is very much about practicality, and real life simulations. In this respect the HP filter fulfils the first selection criteria. Further, Bjørnland, Brubakk and Jore (2007) provide evidence that the PF method has desirable properties as an input in a forecasting experiment similar to the one conducted here.

Below I have described the derivation of the different methods more thoroughly.

2.1 The Hodrick-Prescott filter (HP)

The Hodrick-Prescott filter is a fairly simple and technical procedure for extracting the output gap and the potential production. The main idea behind the method is to minimize the distance between the potential production and real production,

while at the same time taking into consideration restrictions on the growth rate of potential production. The expression to minimize is as follows:

$$Min\{y_t^*\}_{t=1}^{T}\{\sum_{t=1}^{T}(y_t - y_t^*)^2 + \lambda\sum_{t=2}^{T-1}[(y_{t+1}^* - y_t^*) - (y_t^* - y_{t-1}^*)]^2\}, \quad (1)$$

where y_t is GDP and y_t^* is potential GDP. λ is a parameter whose value determines how much potential production is allowed to vary. λ is determined outside the model, and in this paper I have considered three values of λ; 1600, 20000 and 40000. λ 1600 is the international standard (for quarterly data). Further, in Inflation Report 2/2004 Norges Bank found that the HP model with a λ value of 20000 described the Norwegian business cycle better than the alternative λ values evaluated, and finally λ 40000 is used by the Statistics Norway as their preferred λ value.

From equation (1) we see that if we let $\lambda = 0$, the minimization problem would imply setting observed production and potential production equal, and consequently the output gap to zero. On the other hand, by setting λ infinitely big, we would get a very large output gap because the trend, or potential production, hardly would be growing.

The HP filter is easy to implement, but at the same time it has its weaknesses. The filter uses information from both $t-1$ and $t+1$. Thus, at the endpoints the estimations of the output gap become less accurate. By manually prolonging the time-series some quarters ahead with the researcher's best guess of the future value of the series, this problem can be managed. This is also often done when computing the output gap with the HP filter in practice, although I have not used such elongation in this paper. Another weakness is that the value of λ has to be decided beforehand. I have applied three different λ values to overcome this objection.

2.2 The Production function method (PF)[7]

The production function describes the supply side of the economy. Typically the production consists of the production factors capital and labour, and the accessible level of technology. The aggregated production function can therefore be represented by a Cobb Douglas production function (in logarithmic form):

$$y_t = \alpha_0 + a_1 l_t + (1 - a_1) k_t + e_t, \tag{2}$$

where y_t is GDP, l_t the number of working hours, k_t represents the capital stock, e is total factor productivity (TFP) and α_0 is a constant. a_1 and $1 - a_1$ is the wage share and the capital share respectively. TFP is computed as the residual from estimating equation (2).

The potential levels of hours worked, capital and TFP can after estimating equation (2) be used to compute the potential production level (y_t^*):

$$y_t^* = \alpha_0 + \frac{2}{3} l_t^* + \frac{1}{3} k_t^* + e_t^*. \tag{3}$$

In equation (3) I have used the factor shares that are applied by Norges Bank in their daily calculations, and also recommended by the Ministry of Finance in Norway.[8]

The potential level of hours worked depend on the potential levels of the working force, working hours per employee and of the equilibrium level of unemployment. The last measure can be understood as the level of unemployment that is consistent with stabile wage- and price development. All of these potential levels are computed with the HP filter.[9] The output gap is computed as the difference between GDP and the potential GDP estimate.

The PF model has a strong theoretical foundation. However, the functional form applied here is just one of many, and the results are typically a result of the

[7]The following description resembles how the PF computation method is used at Norges Bank, and also at the OECD (see OECD Working Paper no.152). The exposition is taken from Nymoen and Frøyland (2000).

[8]See Finansdepartementet (1997): "Fakta og analyser".

[9]See the Appendix for a closer description of the derivations.

functional form applied. In addition the data foundation can be troublesome. This is especially true when it comes to the stock of capital, which is very hard to measure. As mentioned, many of the potential levels in the method are calculated with the use of the HP filter. This makes also this method exposed to the endpoint problems described earlier when discussing the HP filter method.

2.3 The real-time data sets

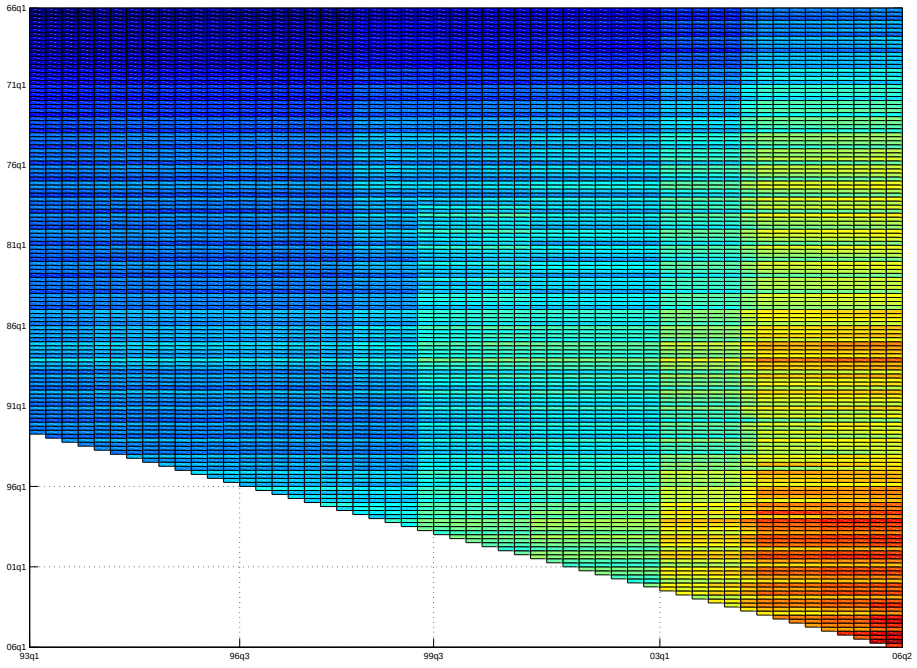
In both output gap models I have used value added at factor costs in manufacturing and construction, and value added at factor costs in private service production as a measure of production. This means that I have not taken into consideration value added in the public sector. Public sector spending can of course also contribute to the cyclical behavior of the business cycle, and in that regard it should perhaps have been modeled. To facilitate model evaluation and make it possible to compare forecasting performance of the different models I have however tried to use the same real-time data sets across the different output gap models as much as possible. Since the production function method is computed without the public sector, I have omitted it from the computation of the other output gap computations as well.[10]

To compute the output gap using the PF method, eleven other data sets have been used, in addition to the production data. These data sets, and their aggregations are described in the Appendix. With one exception, all the data sets starts in 66q1. The different vintages ranges from 93q1 to 06q2.

Figure 1 displays the variable production as value added in manufacturing and construction, and gives an illustration of how the revisions of official statistics influence real-time data sets. Each column represents the time series that a researcher would observe at the different releases, i.e. the different vintages. Each row on the other hand, displays the value for a specific observation in time. The colors give an indication on the magnitude of the value, and warmer colors indicate higher values. Typically the value, and thus the colors, for a specific observation changes across the different vintages due to data revisions. In the figure the base year shifts are

[10]The GDP measure used in this paper covers approximately 3/4 of GDP mainland.

Figure 1: *A real-time data set*



Notes: Production as value added at factor costs in manufacturing and construction. The horizontal axis displays the vintages, the vertical axis displays the different observations, and the color shading indicates the value of the observations. If none of the observations had been revised, colors would have been the same across the different vintages.

clearly visible, and causes the observations to increase in value as we move along the different vintages. The fact that the earliest estimates are based on preliminary and incomplete information can be spotted as more unsystematic shifts in the color shadings.

Figure 2 displays the growth rate of each observation across the different vintages and the standard deviation of this growth rate for two of the data sets in the sample. If one of the observed growth rates deviated more than one standard deviation from the mean, it is showed as a ridge or a dump in the plots. The magnitude of the ridges or dumps are just the observed growth rate at that vintage. Typically the base year effects affects all the observations within a vintage, while the unsystematic revisions are scattered more around the plot.[11]

[11]For estimation and forecasting purposes it would of course have been nice to be ably to detect a pattern in the revisions described above. In the literature this have been tried accomplished by either modeling revisions as noise, news, spillovers within a given data vintage, or as a mixture of all three (Jacobs and van Norden, 2006). Any consensus about the best method have however not

The challenges of evaluating model and forecasting performance in light of real-time data, have only recently been put under intensive study by economists (Bernhardsen et al., 2004). Early contributions to the field was made by Zellner (1958), Morgenstern (1963), and Cole (1969), but only when Dean Croushore and Tom Stark at the Federal Reserve Bank of Philadelphia made available a real-time database including a wide range of US data did research comprised by real-time challenges get easily accessible.[12]

For the Norwegian economy the construction of a real-time database is work in progress, but part of the database have been made available to me by Norges Bank for this project.[13] Bernhardsen et al. (2004) give a profound description of the construction of the database.

None of the real-time data sets in the real-time database that I have used are seasonally adjusted, and therefore I have had to do this manually. For this procedure I have applied the standard X12-ARIMA method, without specifying any special effects (as for example working day adjustments).[14] The seasonal adjustment procedure has been applied in real-time in accordance with the different experiments conducted in this paper.

I have had to adjust and correct some of the data sets and vintages for obvious shortcomings. For all the real-time data sets four vintages have been missing; 93q3, 95q3, 04q2, and 04q4. To fill in these "holes", I simply copied the preceding vintages and extended these series with the growth rates from the subsequent vintages. There have also been data missing for the first observations at some of the vintages for some of the variables in the data sets. I have used growth rates in a reversed order to fill in these gaps. As pointed out in Bernhardsen et al. (2004) these error corrections makes some of the vintages less accurate, but should not constitute major problems for the overall results.

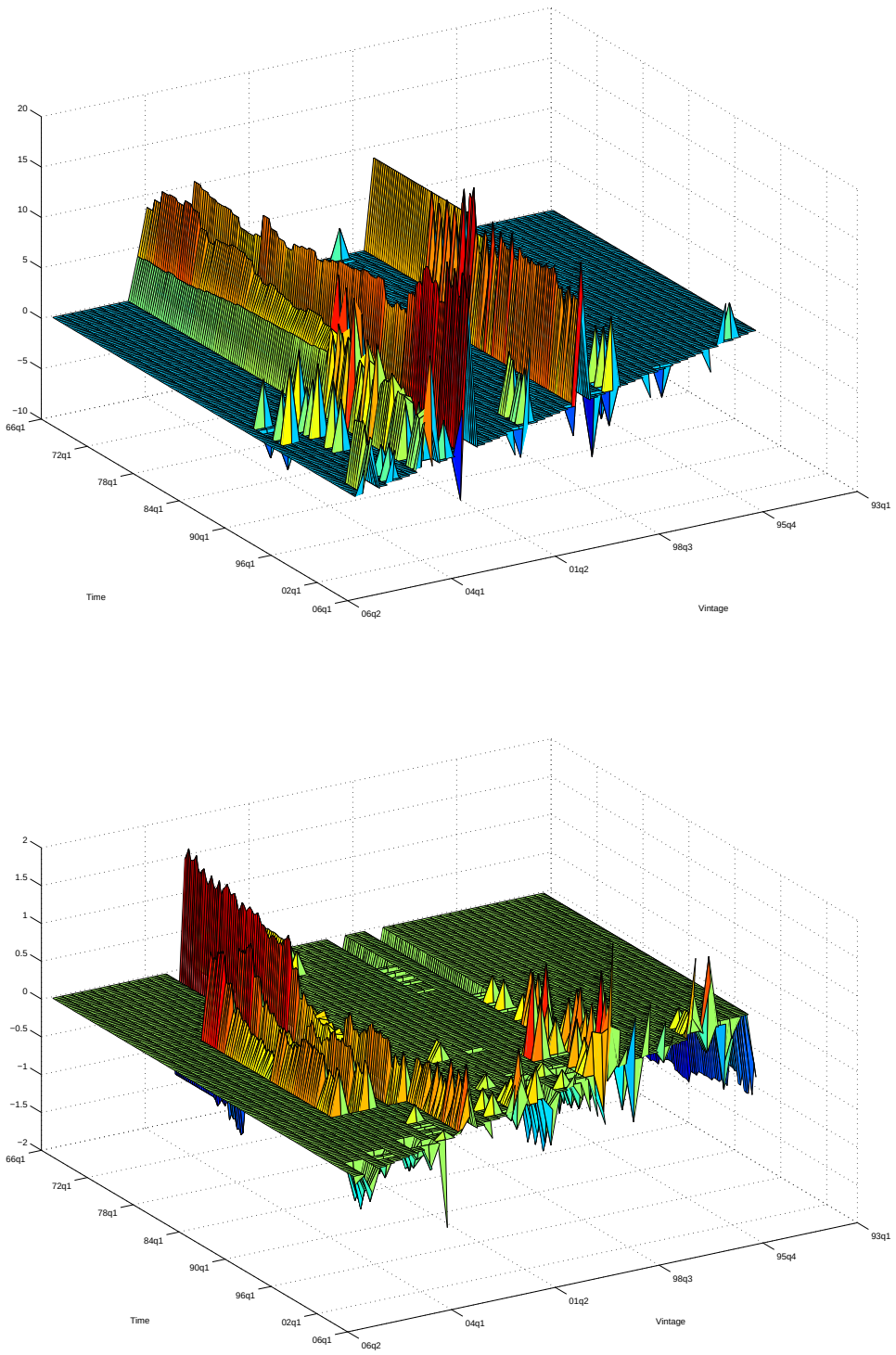
been reached.

[12]See Croushore and Stark (2001). Croushore do also provide a nice overview of the real-time literature, see "http://oncampus.richmond.edu/ dcrousho/docs/realtime_it".

[13]The Datawarehouse Group at Norges Bank and Anne Sofie Jore at the Economics Department at Norges Bank have been to great help in this respect.

[14]X12-ARIMA is the seasonal adjustment software produced and maintained by the U.S. Census Bureau. See "http://www.census.gov/srd/www/x12a".

Figure 2: *Noise in real-time data*



Notes: The upper figure displays production as value added in manufacturing and construction, while the lower figure displays the variable employed wage earners. The ridges and dumps are observations that deviate more than one standard deviation from the mean of the growth rate across vintages. The magnitude of the ridges or dumps are just the observed growth rates at that specific vintage.

2.4 The output gap and real-time estimations

While the preceding section described and illustrated how the revision processes affects the real-time data sets, Orphanides and van Norden (1999 and 2002) have shown how the challenges posed by revisions of real-time data become even more sophisticated when the data are applied in different output gap models. To further enhance the understanding of these issues I have followed Orphanides and van Norden (2002) and Bernhardsen et. al (2005) and decomposed the estimated output gaps into three different gaps; real-time gaps, quasi real-time gaps and final gaps.

I define "*final*" estimates as the estimates produced from the last vintage of data that I have available (06q2). Final is put into quotes here to emphasize the somewhat ephemeral character of the data: These data and estimates are of course also revised. Real-time estimates of the output gap are constructed by first de-trending every vintage, and then taking the last observation of each de-trended vintage as the observation for that point in time. Finally quasi real-time estimates are constructed the same way as real-time estimates, but instead of using real-time data vintages, I use final data truncated at the relevant period.

By constructing three different output gap measures; final, real-time and quasi real-time, I was able to decompose the output gap revisions into three effects; total revisions, data revisions and other revisions. Total revisions, equaling the difference between final and real-time output gap estimates, have two main sources; revisions of national accounts data and effects stemming from new observations as time passes. The difference between quasi real-time and real-time output gaps describes the amount of data revisions, while other revisions, calculated as the difference between final and quasi real-time output gaps, gives a measure of how new observations affects the estimates, and how the results from the different output gap models are affected by new information (Bernhardsen et. al, 2005).

The experiment shows that data revisions do not play a prominent role for the overall results, but that the model specifications do, i.e. different models responds differently to the real-time challenges. The results are summarized in Tables 1, 2 and 3.

Table 1: Output gap statistics

Method	Mean	S.D	Min	Max	Corr	AR
HP1600						
RTgap	0.3148	1.5787	-3.8911	4.1799	0.1230	
QRgap	0.7707	1.9178	-4.1162	5.7789	0.5295	
FLgap	-0.0757	1.6701	-3.8621	4.0540	1.0000	
Total revisions	-0.3905	2.1524	-4.5873	4.1004		0.4690
HP20000						
RTgap	0.7030	2.1290	-4.5949	3.5563	0.2668	
QRgap	1.8222	2.6967	-3.5308	9.0550	0.5730	
FLgap	-0.3485	2.5468	-5.5914	4.3801	1.0000	
Total revisions	-1.0515	2.8505	-6.7959	4.5544		0.7007
HP40000						
RTgap	0.9785	2.1440	-3.9257	4.3347	0.3825	
QRgap	2.2677	2.8411	-2.9037	9.6102	0.6426	
FLgap	-0.4599	2.8161	-6.4228	4.1147	1.0000	
Total revisions	-1.4385	2.8122	-7.3548	4.0234		0.6980
PF						
RTgap	1.2538	1.7377	-3.0007	4.6801	0.6473	
QRgap	1.4364	2.2598	-2.4932	8.1260	0.7686	
FLgap	0.3069	2.5528	-4.8305	5.1425	1.0000	
Total revisions	-0.9469	1.9478	-5.6592	3.9435		0.3408

Notes: Vintages 1993q1–2006q2. RTgap is the real-time output gaps, QRgap is the quasi real-time output gaps and FLgap is final output gaps. Total revisions are calculated as the difference between the FLgap and the RTgap. Mean is the mean value, S.D is the standard deviation. Min and Max is the minimum and maximum values respectively. Corr is the correlation between the final output gaps and the RTgap and the QRgap. AR is the first order autocorrelation coefficient.

Table 1 shows that correlations between the final estimated output gaps and the real-time output gaps are relatively low for all of the HP models, but considerably higher for the PF model.[15] For all the methods the standard deviations of total revisions are large, and typically larger than the standard deviations of the final gap estimates, indicating the relevance of real-time data evaluations. Again the PF gap is the exception. Further, the mean of total revisions are higher in absolute value than the mean for the final gap for all the models. All HP models indicate a high degree of persistence in total revisions. The production function shows the lowest persistence with an autocorrelation coefficient of only 0.3408.

It is also interesting to notice the disparities between the maximum and minimum values of the real-time output gap measures compared to the final gap measures.

[15]It is however worth nothing that the HP models correlation with the final gap increases with the λ value.

Table 2: Output gap credibility

Method	**Corr**	**N/S**	**Opsign**	**Xsize**
HP1600	0.1230	1.2888	0.5370	0.6481
HP20000	0.2668	1.1192	0.4074	0.4630
HP40000	0.3825	0.9986	0.3889	0.5185
PF	0.6473	0.7630	0.3519	0.4074

Notes: Vintages 1993q1–2006q2. Corr is the correlation between the final output gaps and the RTgap. N/S is the noise to signal ratio, computed as the standard deviation of total revisions divided by the standard deviation of the final gaps. Opsign indicates the rate at which the RTgaps and the final gaps have opposite signs. Finally Xsize indicates the rate at which the absolute value of total revisions is larger than the absolute value of the FLgaps.

For the HP method applied with a λ value of 20000 for example, the maximum value is much higher for the final gap measure than for the real-time gap measure. On the other hand, the minimum values for the same gap method displays the opposite characteristics. Accordingly, monetary policy conducted in real time may be prone to react too little to the observed pressure in the economy in a downturn, and react to soft in an upturn. However, the correspondence between the two measures varies a whole lot across the different methods, and the experiment is very fragile towards the properties of what I have labeled the final gap.

Table 2 displays measures that are independent of the size of the estimated output gaps, making it easier to compare models. Note that the statistics do not tell anything about the models ability to say something about the true output gap. Instead the statistics gives a measure of the disparities between final output gaps and real-time output gaps. For the HP method the trend is clear. A higher λ value improves all the measures: The correlation between the real-time output gap and the final output gap increases, the noise to signal ratio improves, and both the Opsign and Xsize measure gets smaller.[16] The PF method performs well compared to the other methods on all the statistics, and have a very low noise to signal ratio compared to the other models.

As can be seen from Table 3, the mean of other revisions are considerably higher

[16]The HP method applied with a λ value of 40000 do however display a higher Xsize than the HP method applied with a λ value of 20000

Table 3: Output gap decomposition

Method	Mean	S.D	Min	Max	AR	N/S
HP1600						
Total revision	-0.3905	2.1524	-4.5873	4.1004	0.4690	1.2888
Data revisions	0.4559	1.3325	-4.0686	4.9624	-0.1715	0.7979
Other revisions	-0.8464	1.7536	-4.2271	2.0631	0.9333	1.0500
HP20000						
Total revision	-1.0515	2.8505	-6.7959	4.5544	0.7007	1.1192
Data revisions	1.1192	1.5457	-3.0302	5.8516	0.0578	0.6069
Other revisions	-2.1708	2.4265	-6.1588	0.9821	0.9632	0.9527
HP40000						
Total revision	-1.4385	2.8122	-7.3548	4.0234	0.6980	0.9986
Data revisions	1.2892	1.6204	-2.5064	6.0928	0.1387	0.5754
Other revisions	-2.7277	2.3916	-6.5327	0.3462	0.9619	0.8493
PF						
Total revision	-0.9469	1.9478	-5.6592	3.9435	0.3408	0.7630
Data revisions	0.1825	1.3699	-4.4914	5.0254	-0.1552	0.5366
Other revisions	-1.1294	1.6601	-3.9268	1.4172	0.9256	0.6503

Notes: Vintages 1993q1–2006q2. Total revisions are calculated as the difference between FLgaps and RTgaps. The difference between QRgaps and RTgaps describes the amount of data revisions, while other revisions are calculated as the difference between FLgaps and QRgaps. See notes in Table 1 for further explanations.

than the mean of data revisions for all the models, and thus contributes more to total revisions. Further, the persistence in data revisions are smaller in magnitude than other revisions. This observation is consistent with the lack of predictability of future revisions of output-growth data reported in Bernhardsen et al.(2004). Consequently the inclusion of new information and model properties play a prominent role for the results of estimating the output gap in real-time.

These findings are qualitatively well in line with what Orphanides and van Norden (1999 and 2002) found analyzing US data, and what Bernhardsen et al. (2005) found analyzing Norwegian data: First, the reliability of the various output gap models estimated in real-time are in general poor. Second, the calculations show large and persistent revisions, and low correlation between real-time estimates and final estimates.

3 Forecasting inflation using real-time output gap estimates

My objective in this paper is to assess the value added in forecasting inflation using an uncertain output gap estimated in real-time. The preceding sections have documented the uncertainty of the output gap estimations. I now proceed with the forecasting experiment, beginning by first describing the forecasting model and evaluation criteria that I have applied more accurately.

3.1 The model

I examine forecasts of inflation at two different horizons, 4 and 8 quarters. Given data for quarter $t-1$ and earlier periods, my objective is to forecast π^h_{t+h}, where $h = 4$ or $h = 8$.[17]

I have used quarterly changes in the prices of goods and services produced domestically as a measure of inflation. This is commonly known as domestic inflation. The rational for using this measure, instead of e.g. regular inflation, is that import prices are less likely to be influenced by the domestic output gap (Bjørnland, Brubakk and Jore, 2007).[18] The real-time issues are assumed to be of minor importance for the inflation measure, thus I have used the same time-series across the different vintages. The series starts in 79q3, and I have used information up to 07q2 (vintage 07q3).

The forecasting equation takes the following form:

$$\pi^h_{t+h} = \alpha + \sum_{j=1}^{n} \beta_j \pi^1_{t-j} + \sum_{j=1}^{m} \lambda_j I_{t-j} + \varepsilon_{t+h}, \tag{4}$$

where α is a constant, I_{t-j} represents the output gap, n and m is the number of lags of inflation and the output gap respectively, and ε_{t+h} is the residual. The

[17] Note that because of reporting lags, information for quarter $t-1$ is only available at time t, i.e. my 4 quarter forecast is accordingly 4 quarters ahead of the current t, but 5 quarters ahead of the data that I have information.

[18] Domestic inflation is also used by the monetary authorities in Norway when conducting monetary policy (among many other indicators of inflation of course), and therefore it has an important practical application as well.

lagged inflation measures, π^1_{t-j}, are annualized. I estimate the unknown coefficients (α, β, λ) by ordinary least squares. The values of n and m are evaluated by two different methods, namely the Bayes information criterion (BIC) and the Akaike information criterion (AIC).[19] I have also estimated and evaluated a model with fixed lag structure, using $n = 8$ and $m = 4$: Keeping the model fixed makes it easier to compare the forecast performance across different input arguments (output gap estimates). An obvious disadvantage is that better forecast accuracy could have been attained if an information criterion had been used.

I compare the forecast performance of the forecasting equation above with an autoregressive AR(p) model, referred to as my benchmark model. Here the value of p is determined either by BIC or AIC, or held fixed with $p = 8$.[20] This is very much the same comparison carried out in Orphanides and van Norden (2005), and Bjørnland, Brubakk and Jore (2007).

3.2 Forecast evaluation

To assess forecasting performance I have compared the output gap models root mean square forecast errors (RMSFE) with the benchmark models RMSFE at different horizons. I have also reported whether the output gap models RMSFE are statistically significant different from the benchmarks RMSFE. Many tests of equal forecasting accuracy can be applied. I have used the modified Diebold and Mariano (1995) test statistics. This test is described in Harvey, D. et al. (1997). Failure to reject the null-hypothesis implies that the inclusion of the output gap did not significantly improve or worsen the forecasting accuracy compared to the AR benchmark.

An admonition should be noted: The use of the Diebold and Mariano test statistic is justified only if the two models compared are not nested. As pointed out by

[19] Generally the BIC method removes more lags than the AIC method. The AIC method is however not a consistent estimator. Still I have used both estimators to assess the optimal lag length because using to few lags can decrease forecasting accuracy (Stock and Watson, 2007).

[20] According to Orphanides and van Norden (2005) this must be considered a weak test. In reality a forecaster will have access to a wider information set, and probably use more complex models. On these grounds Orphanides and van Norden argue that a output gap model might forecast better than a simple univariate benchmark, but compared to a more sophisticated benchmark model it will be outperformed.

Orphanides and van Norden (2005) however, the inclusion of information criterions do unfortunately nest the benchmark models and the output gap models, making the test statistics unreliable. On the other hand, Clark and McCracken (2001) find that the limiting distribution of the Diebold and Mariano statistics is non-pivotal for forecast horizons greater than one period, making the problem of minor importance here.[21]

3.3 The experiments

I have run three main forecasting exercises. Firstly, I have followed Orphanides and van Norden (2005) and estimated the model up to the last observation in each vintage, for each vintage, in the sample. For every estimation I have made a 4 quarter forecast and a 8 quarter forecast, and the RMSFE have been computed as the sum of the forecasting mistakes made at each vintage. Equation (4) have been estimated with and without the four output gap measures described in section 2.1 and 2.2. The estimation of the output gaps have been carried out by the same logic as the estimation of the forecasting equation: For every output gap model this procedure has produced 54 vintages of output gap estimates. (Note the difference between this line of action compared to the one taken in section 2.4, where I only used the last estimated observation of each vintage to construct the real-time output gaps.) Figure 5 in the Appendix shows the estimated output gaps across vintages. Because of data revisions, and the properties of the output gap models, the assessment of the business cycle clearly changes as new observations are taken into consideration.

Secondly, to assess the contribution and importance of the real-time data and estimation issues I have compared the real-time results from the experiment described above with the results from a final gap exercise. That is, I took the last estimated output gap vintage for each output gap method, and truncated this into the respective observations in the real-time matrix. Then I run the forecasting experiment as

[21] An additional objection against the p-values reported comes from Ashley (2003), who argued that more than 100 observations are necessary to establish significant difference in predictive accuracy across models. The number of vintages evaluated in this study falls short of this number. See also Bjørnland, Brubakk and Jore (2007).

described above, and computed the new RMSFE values.

Thirdly, to enhance the understanding of how the inclusion of real-time data affects the results, and the importance of the chosen lag length, I have computed the differences in forecasting performance between a set of different models. More specifically I label the results from the experiment where the lag length have been kept fixed and final output gap estimates have been used as FL-FL. Results labeled VL-FL refers to final data results, but now with the inclusion of varying lag length. Finally, the results from running the forecasting experiment with real-time output gap estimates and variable lag length have been labeled VL-RT. As explained in Orphanides and van Norden (2005), differences in outcomes between FL-FL and VL-FL indicates the affect of variations in lag length, while differences between VL-FL and VL-RT isolates the affect of output gap revisions.

Thus, the first experiment evaluates how the inclusion of the output gap affects the forecasting performance relatively to the benchmark models, while the second experiment explores the difference in forecasting performance between using real-time data versus final data. The results from the third experiment are ment to give a description of the difficulties of choosing the optimal lag length, and how the real-time data issues affects the forecasting performance.

4 Results

4.1 Do the output gap give any value added in forecasting inflation?

Table 4 shows the RMSFE results from the 4 and 8 step forecasting experiment, with and without the use of information criterion, applying real-time output gap measures. The RMSFE value for the benchmark models (AR, AR bic and AR aic) are shown as they were computed, the other RMSFE estimates are displayed as the fractional improvement (or deterioration) relatively to the benchmark model $(RMSFE^*_{Benchmark} - RMSFE^*_{Gap}/RMSFE^*_{Gap})$.

Table 4: Forecasting evaluation. Real-time estimates

Model	**4 step forecast**	**p-value**	**8 step forecast**	**p-value**
AR	*0.9097*		*1.1274*	
HP1600	-0.0383	0.6281	-0.1742	0.0012
HP20000	-0.0266	0.8816	-0.1216	0.4435
HP40000	-0.0356	0.8307	-0.1043	0.5220
PF	0.0246	0.8445	-0.0740	0.3369
AR bic	*0.9336*		*1.1261*	
HP1600 bic	-0.0113	0.8875	-0.0759	0.5242
HP20000 bic	0.0319	0.7751	-0.1680	0.4618
HP40000 bic	0.0120	0.9134	-0.1838	0.4358
PF bic	0.0223	0.8199	-0.1575	0.3578
AR aic	*0.9128*		*1.1059*	
HP1600 aic	-0.0212	0.7899	-0.1231	0.0001
HP20000 aic	0.0248	0.8857	-0.1525	0.4014
HP40000 aic	0.0082	0.9608	-0.1493	0.4062
PF aic	0.0217	0.8778	-0.1251	0.1811

Notes: The AR models are univariate autoregressive forecasts of domestic inflation. The other models are Phillips curve relationships with different real-time output gap estimates. bic and aic suffixes signals that the forecasting equations have been estimated with an information criterion. The RMSFE values are shown relatively to the benchmark models, measured as (A-B)/B where A is the RMSFE of the benchmark model and B is the RMSFE of the output gap model. P-values are calculated by the modified Diebold and Mariano test statistics, and are shown as a two-sided test statistic with a null hypothesis of A=B. At the 4 quarter horizon 50 forecasts have been evaluated. At the 8 quarter horizon 46 forecasts have been evaluated. The forecast equation estimations starts in 79q3 for both forecasting horizons.

Considering the 4 quarter forecasts first we see that overall, 7 out of 12 output gap models performs better than the benchmark models. The gain in terms of forecasting accuracy are however very modest. At best only 3.2 percent. More specifically, in the case of no information criterion and a lag structure of 8 and 4 on inflation and the output gaps respectively, the HP models are inferior to the benchmark model while the PF model outperforms the benchmark model. These results change substantially when I include information criterions. Now all the models performs better than the benchmark models. The exceptions are the HP models with a λ value of 1600, which do worse than the benchmark models. On the 4 quarter horizon the best relative improvement from the benchmark model is demonstrated by including the output gap measure computed by the HP method with a λ value of 20000 evaluated by the BIC, while the PF method seems to be the most robust method in this experiment.

The p-values generally show very high numbers, and none of the differences in RMSFE are significant. As pointed out in section 3.2 though, the Diebold and Mariano test statistics can be misleading when the models evaluated are nested, and the p-values reported for the AIC and BIC models (here and below) should therefore be interpreted with caution.

The 4 quarter forecast horizon results do not apply at the 8 quarter forecast horizon. Now none of 12 output gap models performs better than the benchmark models. At the same time the results indicate that the inclusion of an information criterion makes the output gap models perform less favorable. Further, the forecasting accuracy deteriorates quite a lot. The relative RMSFE values for the output gap models are as much as 18.4 percent below the benchmark models. Accordingly the p-values have become smaller than they were at the 4 quarter horizon, but still a significant difference in forecasting performance is hard to prove. Only the differences in forecasting performance between the benchmark models and the HP model with a λ value of 1600 evaluated with and without AIC are significant at the 5 percent level.

Table 5 shows the same measures as in Table 4, but now the forecast experiments are conducted with final output gap estimates. At the 4 quarter horizon the forecasting results are much better than they were using real-time data. Now all the output gap models, except the HP model with a λ value of 40000, performs better than the different benchmark models. Further, the relative improvements in the RMSFE values are generally of a greater magnitude than they were using real-time data.

On the 8 quarter horizon the results from the real-time experiment stands: 0 out of 12 output gap models performs better than the benchmark models. Still, the p-values are generally poor, and only the HP model with a λ value of 20000, and evaluated with AIC performs better than the benchmark model at the 5 percent significance level.

Figure 6 and Figure 7 in the Appendix shows the different forecasts compared to actual inflation.

Table 5: Forecasting evaluation. Final estimates

Model	**4 step forecast**	**p-value**	**8 step forecast**	**p-value**
AR	*0.9097*		*1.1274*	
HP1600	0.0029	0.9778	-0.1999	0.2730
HP20000	0.0080	0.9580	-0.1847	0.3556
HP40000	-0.0018	0.9893	-0.1761	0.3542
PF	0.0225	0.8899	-0.1200	0.6306
AR bic	*0.9336*		*1.1261*	
HP1600 bic	0.0675	0.1025	-0.0688	0.4148
HP20000 bic	0.0969	0.0390	-0.2283	0.1805
HP40000 bic	0.0608	0.1526	-0.2367	0.1374
PF bic	0.0432	0.5190	-0.1136	0.1532
AR aic	*0.9128*		*1.1059*	
HP1600 aic	0.1139	0.0569	-0.1512	0.1929
HP20000 aic	0.0323	0.7896	-0.2571	0.2303
HP40000 aic	0.0167	0.8809	-0.2340	0.2509
PF aic	0.0391	0.7551	-0.2054	0.3681

Notes: The AR models are univariate autoregressive forecasts of domestic inflation. The other models are Phillips curve relationships with different final output gap estimates. bic and aic suffixes signals that the forecasting equations have been estimated with an information criterion. The RMSFE values are shown relatively to the benchmark models, measured as (A-B)/B where A is the RMSFE of the benchmark model and B is the RMSFE of the output gap model. P-values are calculated by the modified Diebold and Mariano test statistics, and are shown as a two-sided test statistic with a null hypothesis of A=B. All vintages in the sample have been included, giving a total of 50 forecasts for each model on the 4 quarter horizon, and 46 forecasts at the 8 quarter horizon. The forecast equation estimation starts in 79q3 for both forecasting horizons.

The importance of the chosen lag length, and the real-time output gap estimates are clearly seen in Table 6. Results from the 4 quarter horizon forecasts are displayed in the upper box of the table, while the 8 quarter horizon results are displayed in the lower box.

Looking at the final gap RMSFE values and the FL-VL column first, we see that for all the output gap models the inclusion of an information criterion to assess the optimal lag length makes the forecasting accuracy on the 4 quarter horizon better. The mean improvement across the different models is 2.8 percent. If we compare the forecasting performance between final and real-time output gaps we see that the forecasting accuracy worsens (the benchmark model is of course not affected), with a mean drop in accuracy of 3.4 percent. Believing that real-time data causes the output gap estimates to be less precise than final data estimates, as the experiments in section 2.4 indicates, these findings are as expected.

Table 6: Forecasting evaluation. Effects of lag selection and forecasting with real-time output gaps versus final output gaps

	RMSFE			Change in RMSFE(percent)		
Method	**FL-FL**	**VL-FL**	**VL-RT**	**FL to VL**	**FL to RT**	**Total**
AR	0.9097	0.9128	0.9128	-0.3475	0.0000	-0.3475
HP1600	0.9071	0.8195	0.9326	9.6521	-13.7967	-2.8130
HP20000	0.9025	0.8842	0.8908	2.0221	-0.7376	1.2993
HP40000	0.9113	0.8978	0.9054	1.4806	-0.8455	0.6476
PF	0.8896	0.8785	0.8934	1.2477	-1.6939	-0.4252
Mean				2.8110	-3.4148	-0.3277
Std. Dev				3.9247	5.8347	1.5641
AR	1.1274	1.1059	1.1059	1.8997	0.0000	1.8997
HP1600	1.4090	1.3030	1.2611	7.5246	3.2139	10.4966
HP20000	1.3827	1.4887	1.3050	-7.6602	12.3361	5.6208
HP40000	1.3684	1.4437	1.3001	-5.5027	9.9498	4.9946
PF	1.2811	1.3919	1.2641	-8.6514	9.1843	1.3275
Mean				-2.4780	6.9368	4.8678
Std. Dev				6.9522	5.1298	3.6610

Notes: The upper box displays the 4 quarter horizon forecast results, the lower box displays the 8 quarter horizon forecasts results. FL-FL refers to final output gap estimates, and fixed lag lenghts. VL-FL refers to final output gap estimates, and variable lag lenghts. VL-RL refers to real–time output gap estimates and variable lag lenghts. Only the results from the AIC experiment are shown since these generally displayed the best forecasting performance of the two information criterions evaluated. The three last columns shows improvement or decay of the RMSFE values of moving from one estimation procedure to another: FL to VL is the change in RMSFE between FL-VL and VL-FL, FL to RT is the change in RMSFE between VL-FL and VL-RT, and finally Total is the change in RMSFE between FL-FL and VL-RT.

The 8 quarter horizon results are more difficult to explain. Firstly the inclusion of varying lag length show ambiguous results. The forecasting accuracy improves for the benchmark model, and the HP model with a λ value of 1600. For the three other models it deteriorates. Secondly, and perhaps more surprising are the results comparing the final data estimates with the real-time data estimates. In contrast to the results on the 4 quarter horizon, now all the output gap models get more precise. The mean improvement is 6.9 percent. One explanation for this rather odd result can of course be assigned to the quality of what I have labeled the final output gap estimates. These final estimates will of course also be revised in due time, and only in retrospect can we assess their properties.

To sum up the results I find that:

- The forecasting performance of the different output gap models compared with simple benchmark models are strongly affected by the forecasting horizon, and by the inclusion of varying lag length.
- Generally the results indicate that the inclusion of the output gap is redundant or even damaging for the forecasting performance on the longer horizons. On the shorter horizons some output gap models estimated in real-time forecast inflation better than the benchmark models.
- On the 4 quarter horizon forecasts made by final data outperforms forecasts made with real-time data.

4.2 Sensitivity analysis

To check the robustness of my findings and especially how different forecasting periods affect the results, I have performed a fourth forecasting experiment. This experiment has been carried out using a somewhat different method than the experiments described above. More precisely, I have estimated the model in equation (4) up to time $t-1$ across every vintage, and then made a forecast of π^4_{t+4} (π^8_{t+8}) at every vintage. Then I estimated the model up to time t, and made a new forecast of π^4_{t+5} (π^8_{t+9}) at that point in time. I repeated this procedure until a satisfying number of forecasts were reached. The RMSFE was computed as the sum of the forecasting mistakes made at each forecast within each vintage.

The output gap estimations were computed the same way as the forecasting equation. Thus, for every vintage I have computed as many output gap estimates as I have made forecasts.

Two forecasting periods have been considered on each forecasting horizons. First I used all the vintages in the sample (93q1-06q2), and forecasted 4 quarter inflation over the period 89q2 to 94q1, and 8 quarter inflation over the period 90q2 to 95q1. This gave me a total of 20 forecasts on each horizon, while the number of observations available for estimation ranged from 41 up to 60. Then I changed the forecasting

period, and forecasted 4 quarter inflation from 95q2 to 00q1, and 8 quarter inflation over the period 96q2 to 01q1. Accordingly, the number of forecasts are the same as above, but the number of vintages considered had to be reduced. The number of observations available for estimation have accordingly increased, now ranging between 65 and 84.

Running the experiment like this ensured that the amount of information used in the forecasting equations were the same across the different vintages. By keeping all but the vintages the same, I were able to evaluate more directly how the different forecasting periods affected forecasting performance, and at the same time assess which of the output gap estimations that performed best in real-time.[22] An inconvenience with this method is that the number of observations and information used in the forecasting exercise are restricted by the number of observations in the first vintage considered.

Figure 3 and Figure 4 shows the results. The RMSFE for the models including the output gaps are shown relative to the benchmark models.[23]

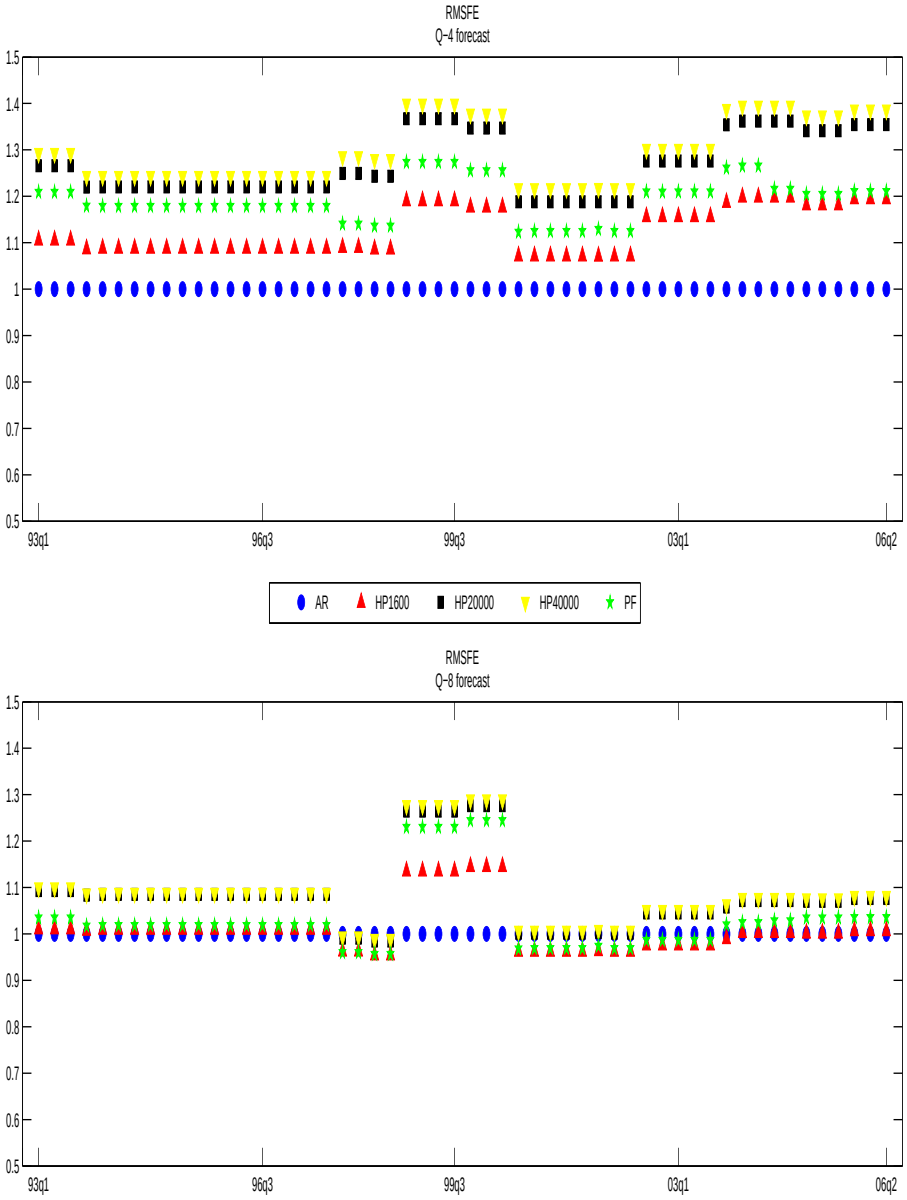
As can be seen from Figure 3, which displays the results from the earlier forecasting period, the benchmark model generally performs better than the output gap models on both horizons. In contrast to my earlier findings though, the output gap models seems to perform relatively better at the longer horizon. In relation to forecasting performance there are also some disparities concerning the ranking of the output gap models.

Figure 4 shows the results from the latter forecasting period. Two factors stands out: The benchmark models are still hard to beat. The RMSFE values have become relatively much poorer.

[22]The experiment conducted here is strictly speaking not a real-time experiment as the one conducted in section 4.1, but it gives an indication of how the different models performs across the different real-time vintages.

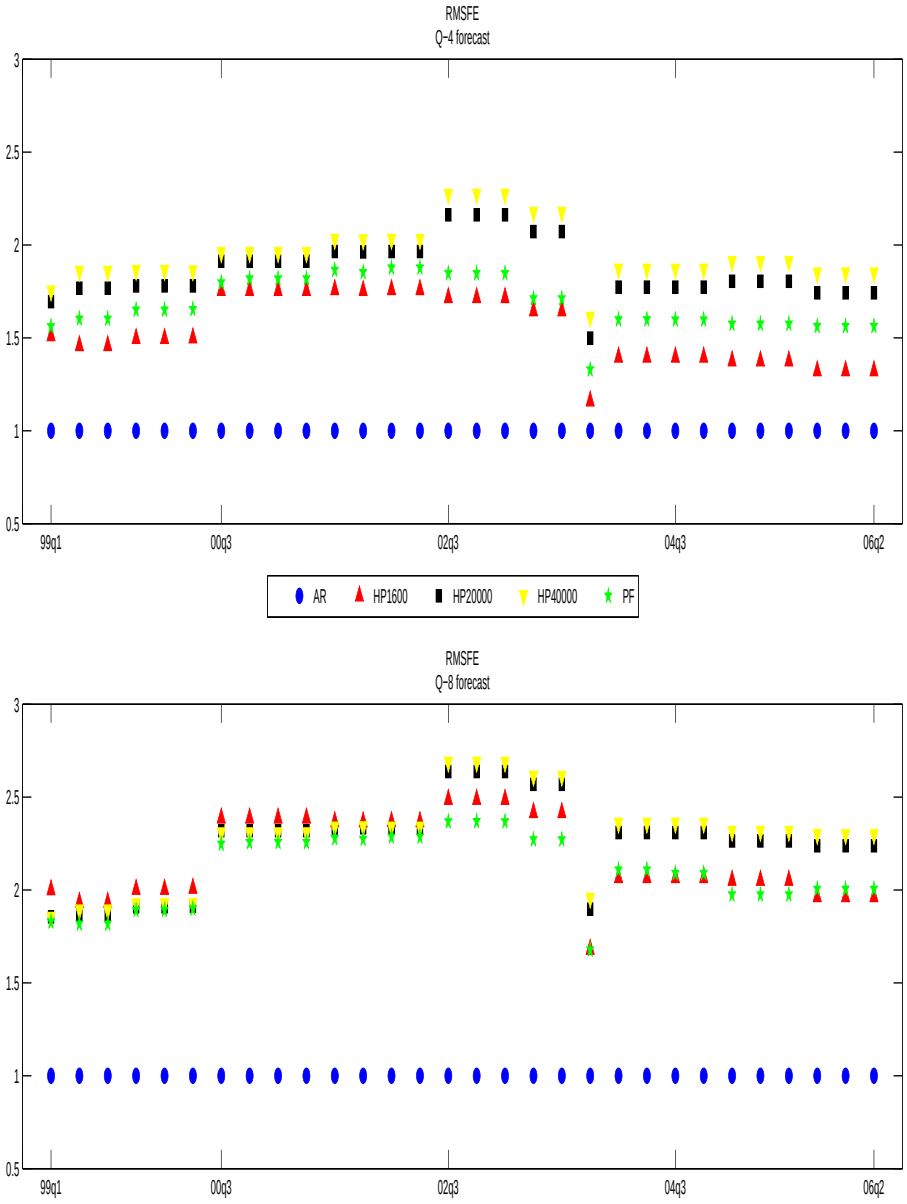
[23]I have not used an information criterion to assess the optimal lag structure on any of the calculations considered in Figure 3 and Figure 4.

Figure 3: *Relative improvement in RMSFE, all vintages*



Notes: The sample starts in 78q1 for every vintage and every forecasting horizon. The forecasts cover the period 89q2 to 94q1 for the 4 quarter forecast horizon, and the period 90q2 to 95q1 for the 8 quarter forecast horizon. For both horizons the number of forecasts are 20. The relative improvement is computed simple as A/B, where A is the RMSFE of the Phillips curve model and B is the RMSFE of the benchmark model.

Figure 4: *Relative improvement in RMSFE, restricted number of vintages*



Notes: The sample starts in 78q1 for every vintage and every forecasting horizon. The forecasts cover the period 95q1 to 00q1 for the 4 quarter forecast horizon, and the period 96q2 to 01q1 for the 8 quarter forecast horizon. For both horizons the number of forecasts are 20. The relative improvement is computed simple as A/B, where A is the RMSFE of the Phillips curve model and B is the RMSFE of the benchmark model.

Both of these additional experiments confirms many of the findings from section 4.1. In addition, the results from this experiment highlights how different vintage samples and forecasting periods may influence the forecasting performance. A general impression is that the inclusion of the output gaps in the forecasting equation gave a better forecasting performance at the earlier forecasting period than the later one. However, it is not easy to tell if this difference is due to the fact that the forecasting period has changed, or that the number of observations have changed.

The results reported here and in section 4.1 are very sensitive to the choice of GDP series. In an alternative experiment I used GDP mainland (seasonally adjusted) for the HP method calculations, and GDP for the private sector for the PF method calculations. This yielded very different results for the HP model calculations. Generally the forecasting performance was still better for the benchmark models compared to the Phillips curve models, but the effects of varying lag length and forecasting horizon showed different properties. To make the comparison between the two output gap models as feasible as possible I have however used the same GDP measure in all output gap calculations reported, namely GDP for the private sector.

I have also tried to estimate the forecasting equation using other explanatory variables than the output gap (unemployment gap and output growth). The forecasting performance of these alternative variables did not outperform the output gap models.[24]

4.3 Discussion

Below I compare some of the findings above with earlier findings in the literature, and I point at some aspects of the analysis that have been conductive for the results.

The 4 quarter horizon results reported in Table 4 and Table 5 are well in line with the results reported in Orphanides and van Norden (2005), while the results on the longer horizon do show some disparities.[25] As demonstrated in section 4.2

[24]The detailed results from these alternative experiments can be attained on request.

[25]Typically the results presented in Orphanides and van Norden (2005) were very sensitive to sample and vintage selection.

the chosen forecasting period probably plays an important role in explaining these differences.

Bjørnland, Brubakk and Jore (2007) argue that the inclusion of the output gap makes the forecasts of inflation better than what would have been the results of simple benchmark models, both on a 4 quarter forecasting horizon and on a 8 quarter forecasting horizon. These findings are very different from what I have reported in this paper. Bjørnland, Brubakk and Jore (2007) did however not conduct their analysis on real-time data, and they did not evaluate their models with any information criterions. I have shown that both of these factors affects the results considerably.[26] Interestingly, one of the best performing models in both this experiment and in Bjørnland, Brubakk and Jore (2007) were the PF output gap model.[27]

The forecasting results on the longer horizon reported above can be advocated by findings in Staiger, Stock and Watson (1997) According to their analysis, forecasting inflation 2 years ahead with a Phillips curve relationship give less favorable results than alternative forecasting models. The Phillips curve relationship applied in Staiger, Stock and Watson (1997) do not resemble the one considered in this paper though, and a direct comparison can and should therefore not be conducted.

In the literature many models for extracting the output gap in real-time have been evaluated against their value added in forecasting inflation. In this paper I have only analyzed two models, and as the results in section 2.4 indicates the models behaved very differently when confronted with real-time data and estimation issues. I can not rebut that some other output gap model than those considered in this paper might turn out to be more robust against these issues, and accordingly will perform better in an forecasting experiment like this. Still, the results from Orphanides and van Norden (2005) suggests that they probably will not.

On the other hand, Stock and Watson (1999) found that Phillips curves specified with alternative measures of real economic activity could forecast inflation better than unemployment-based Phillips curves. As the Norwegian real-time database

[26] At the same time I found that 3 of 4 output gap models estimated with final data performed better than the benchmark model, confirming the results in Bjørnland, Brubakk and Jore (2007).

[27] This finding coincide with the results reported in section 2.4, which confirmed that the PF method had favorable characteristics compared with the HP method.

becomes more comprehensive, the real-time forecasting performance of these alternative indicators can and should be tested.

Further, the forecasting performance of different models depends crucially on how possible structural breaks are managed. Domestic inflation in Norway has from the beginning of the 1980's fallen sharply until the beginning of the 1990's, when it leveled off. Many studies have documented this possible break statistically. Eitrheim and Nordbø (2005) investigated Norwegian CPI and 132 subgroups. They found evidence of a break in the aggregate CPI series in the late 1980's. Levin and Piger (2003) analyzed inflation for 12 OECD countries and found strong evidence of a break in the late 1980's or early 1990's. To enhance the performance of the forecasting models the possible break in the time-series should somehow have been taken into consideration. I have however not done so. In real-time it is highly unlikely that a break would have been detected, and it would not have been proper to lay restrictions on the models or estimations, that seen in retrospect most likely would have enhanced the forecasting performance of the models. I have however done part of the forecasting experiment on different combinations of vintages, and on different forecasting periods, just to emphasize the importance and vulnerability of these facts.

5 Conclusion

In this paper I have questioned whether the inclusion of output gaps give any value added in forecasting Norwegian domestic inflation, compared to simple autoregressive benchmark models using real-time data. My results suggests that the value added is modest, at best.

Firstly, the revisions of official statistics makes the real-time data that a professional forecaster relies upon highly uncertain. Secondly, the reliability of the various output gap models estimated in real-time is in general poor, the calculations show large and persistent revisions, and low correlation between real-time estimates and final estimates. At the longer horizon none of the Phillips curve models forecasted

inflation better than the simple benchmark models, while some of the Phillips curve models outperformed the benchmark models at the shorter forecasting horizon. Typically the relative forecasting performance of the models evaluated was very sensitive to the chosen forecasting period. Still the benchmark models were hard to beat.

The models evaluated in this paper must be considered very simplistic though. Different results could perhaps have been obtained if a more sophisticated model structure had been applied. Akram, Eitrheim and Nymoen (2007) argue that well specified econometric models tend to inhabit better forecasting properties than entirely data based time-series models. They argue that: "... within sample properties of an econometric model may be a reliable guide to its out-of-sample forecasting performance even when data is heavily revised."

In the literature there are also different methodological approaches taken to construct and estimate real-time data. Koenig, Dolmas and Piger (2003) give an illustrative description. They suggest that real-time data should be modeled as what they label "real-time-vintage data" and not "end-of-sample-vintage data" as done in this paper. Their out-of-sample forecasting results indicate that the forecasting performance of using the former methodology are substantially better than using the latter methodology. However, their forecasting experiment was conducted on quarterly GDP measures, with a set of monthly explanatory variables.

That said, my results support earlier research conducted on real-time data. Forecasting inflation in real-time is a difficult task, and monetary policy conducted in real-time should therefore be careful of responding to strongly to the output gap as a measure of forecasting inflation.

References

Akram, Q.F, Ø. Eitrheim and R. Nymoen (2007): "Forecasting under data and model uncertainty", forthcoming Working Paper.

Ashley, R. (2003): "Statistically significant forecasting improvements: How much out-of-sample data is likely necessary?", International Journal of Forecasting 13, 229–239.

Bernhardsen, T, Ø. Eitrheim, A.S. Jore and Ø. Røisland (2004): "Real-time data for Norway: Challenges for Monetary Policy", Discussion Paper Series 1: Economic Studies 2004, 26, Deutsche Bundesbank, Research Centre.

Bernhardsen, T, Ø. Eitrheim, A.S. Jore and Ø. Røisland (2005): "Real-time data for Norway: Challenges for Monetary Policy", The North American Journal of Economics and Finance 16(2005), 333–349.

Bjørnland, H.C., L. Brubakk and A.S. Jore (2004): "Produksjonsgapet i Norge - en sammenlikning av beregningsmetoder", Penger og Kreditt 4/04.

Bjørnland, H.C., L. Brubakk and A.S. Jore (2007): "Forecasting inflation with an uncertain output gap", forthcoming in Empirical Economics.

Clark, T.E and M.W. McCracken (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models", Journal of Econometrics 105, 85–110.

Croushore, D. and T. Stark (2001): "A real-time data set for macroeconomists", Journal of Econometrics 105, 111–130.

Cole, R. (1969): "Data errors and forecasting accuracy", in Mincer J.(ed.): Economic Forecasts and Expectations: Analyses of Forecasting Behavior and Performance. National Bureau of Economic Research, New York, chapter 2, 47–82.

Diebold, F.X. and R.S. Mariano (1995): "Comparing Predictive Accuracy", Journal of Business and Economic Statistics, 13 (1995), 253–265.

Eitrheim, Ø. and E.W. Nordbø (2005): "How persistent is disaggregate inflation in Norway?", unpublished paper. Presented at the 28. National Research Convention for Economists 2006, The Norwegian School of Economics and Business Administration.

Finansdepartementet (1997): "Fakta og analyser", Særskilt vedlegg til St. meld nr.4(1996–1997) Langtidsprogrammet 1998–2001, 74.

Frøyland, E. and R. Nymoen (2000): "The Output Gap in the Norwegian economy - different methodologies, same result?", Economics Bulletin 2/00, 46–52.

Giorno, C., P. Richardson, D. Roseveare and P.van den Noord (1995): "Estimating Potential Output, Output Gaps and Structural Budget Balances", OECD Economics Department Working Papers, No. 152. OECD.

Hamilton, J.D. (1994): Time Series Analysis. Princeton University Press, Princeton.

Hanselman, D. and B. Littlefield (2005): Mastering Matlab7. Pearson Prentice Hall, New Yersey.

Harvey, D., S. Leybourne and P. Newbold (1997): "Testing the equality of prediction mean squared errors", International Journal of Forecasting 13 (1997), 281–291.

Jacobs, J.P.A.M. and S.van Norden (2006): "Modeling Data Revisions: Measurement Error and Dynamics of "True"Values", CCSO Working Paper, December 2006.

Koenig, E.F., S. Dolmas, and J. Piger (2003): "The Use and Abuse of "Real-Time"Data in Economic Forecasting", Review of Economics and Statistics 85 (2003), 618–628.

Levin, A.T. and J.M. Piger (2003): "Is inflation persistence intrinsic in industrial economies?", European Central Bank, Working Paper, 334.

McKenzie, R. (2006): "Undertaking revisions and real-time data analysis using the OECD main economic indicators original release data and revisions database", Technical Report STD/DOC(2006)2, OECD.

Mckenzie, R. (2007): "Relative size and predictability of revisions to GDP, Industrial Production and Retail Trade - a comparative analysis across OECD Member countries", forthcoming, OECD.

Morgenstern, O. (1963): On the Accuracy of Economic Observations. Princeton University Press, Princeton NJ.

Nelson, C.R. and C.I. Plosser (1982): "Trends and Random Walks in Macroeconomic Time Series", Journal of Monetary Economics 10, 129–162.

Norges Bank (2004): "Inflation Report 2/2004", Reports from the Central Bank of Norway No 3/2004, 45–47.

Orphanides, A and S.van Norden (1999): "The Reliability of Output Gap Estimates in Real Time", unpublished paper.

Orphanides, A and S.van Norden (2002): "The Unreliability of Output Gap Estimates in Real Time", Review of Economics and Statistics 84, 569–583.

Orphanides, A and S.van Norden (2005): "The reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time", Journal of Money, Credit and Banking 37, 583–601.

Rotenberg, J.J. and M. Woodford (1997): "An Optimization-Based Econometric Model for the Evaluation of Monetary Policy", NBER Macroeconomics Annual 12, 297–346.

Staiger, D., J.H. Stock and M.W. Watson (1997): "The Nairu, Unemployment and Monetary Policy", The Journal of Economic Perspectives, Vol. 11, No. 1 (Winter, 1997), 33–49.

Stock, J.H. and M.W. Watson (1999): "Forecasting Inflation", NBER Working Paper Series, Working Paper 7023.

Stock, J.H. and M.W. Watson (2007): Introduction to Econometrics. Pearson, Boston.

Svensson, L.E.O (1997): "Inflation Forecast Targeting: Implementing and Monitoring Inflation Targets", European Economic Review 41, 1111–1146.

Svensson, L.E.O (2000): "Open-Economy Inflation Targeting", Journal of International Economics 50, 155–183.

Svensson, L.E.O. and M. Woodford (2003): "Indicator Variables for Optimal Policy", Journal of Monetary Economics 50, 691–720.

Zellner, A. (1958): "A statistical analysis of provisional estimates of Gross National Product and its components, of selected National Income components, and of personal saving", Journal of the American Statistical Association 53, 54–65.

A Data definitions

A.1 Notes

1. All the time series, except the inflation measure, have been extracted from the RIMINI databases and organized into a real-time data base maintained by Norges Bank.

A.2 Definitions

$YIBA$ Value added at factor costs in manufacturing and construction, fixed base year prices. Mill. NOK. RIKMOD sectors 01–05.Source: QNA.

YTV Value added at factor costs in private service production, fixed base year prices. Mill. NOK. RIKMOD sectors 06 and 12. Source: QNA.

$KIBA$ Industry stock of fixed capital, fixed base year prices. Mill. NOK. RIKMOD investment sectors 01–05. Source: NA, KVARTS.

KTV Stock of fixed capital in private service production, fixed base year prices. Mill. NOK. RIKMOD investment sector 06. Source: NA, KVARTS.

$FHIBA$ Average quarterly working hours in manufacturing and construction. 1000 hours per employee. Source: KVARTS.

$FHTV$ Average quarterly working hours in private service production. 1000 hours per employee. Source: KVARTS.

$TWIBA$ Man-hours by employees in manufacturing and construction. Including overtime and absence from work due to vacation, sick leave etc. Also influenced by calendar effects. Mill. hours. RIKMOD sectors 01–05. Source: KVARTS.

$TWTV$ Man-hours by employees in private service production. Including overtime and absence from work due to vacation, sick leave etc. Also influenced by calendar effects. Mill. hours. RIKMOD sectors 06 and 12. Source: KVARTS.

NW Employed wage earners. 1000 persons. Includes part time workers, conscripts and persons temporarily absent from work. Sum all RIKMOD sectors. Source: KVARTS (1962Q1–1994Q4), QNA (1995Q1–).

NS The number of self-employed. 1000 persons. Source: KVARTS.

$TILT12$ Number of participants on labour market programmes, 1000 persons. Source: NORMAP.

$REGLED$ Number of registered unemployed. 1000 persons. Source: NORMAP.

$UAKU$ Labour force survey (AKU) unemployment rate. Source: NORMAP.

$PCPIJAEI$ Consumer Price Index Domestic Sources (KPIJAEI). Seasonally adjusted.

A.3 Production function aggregates

All the variables have been seasonally adjusted. An asterix indicates that the variable have been de-trended with the HP filter.

Production:

$$Y = YIBA + YTV \tag{5}$$

Capital

$$K = KTV + KIBA \tag{6}$$

Labour (number of working hours):

$$L = TWTA + TWIBA \tag{7}$$

Not modeled employment:

$$NIM = NS + NW - TWIBA/FHIBA - TWTV/FHTV \tag{8}$$

Average working hours:

$$\begin{aligned} FH &= (TWIBA/(TWTA + TWIBA)) * FHIBA + \\ & \quad (TWTA/(TWTA + TWIBA)) * FHTV \end{aligned} \tag{9}$$

Working force:

$$AS = NS + NW + TILT/1000 + REGLED \tag{10}$$

Potential employment:

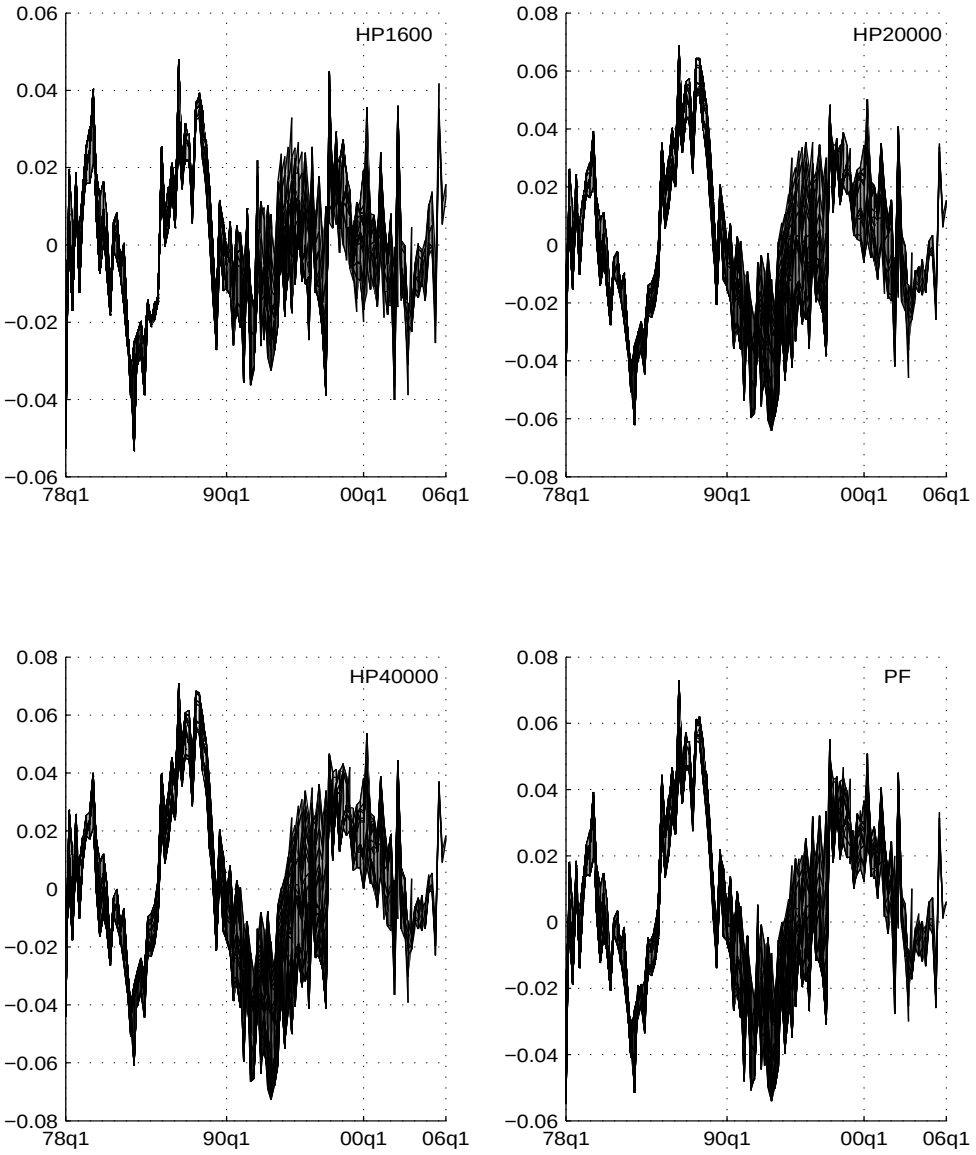
$$n = AS^* * (1 - UAKU^*) - NIM^* \tag{11}$$

Potential hours worked:

$$l = log(n * FH^*) \tag{12}$$

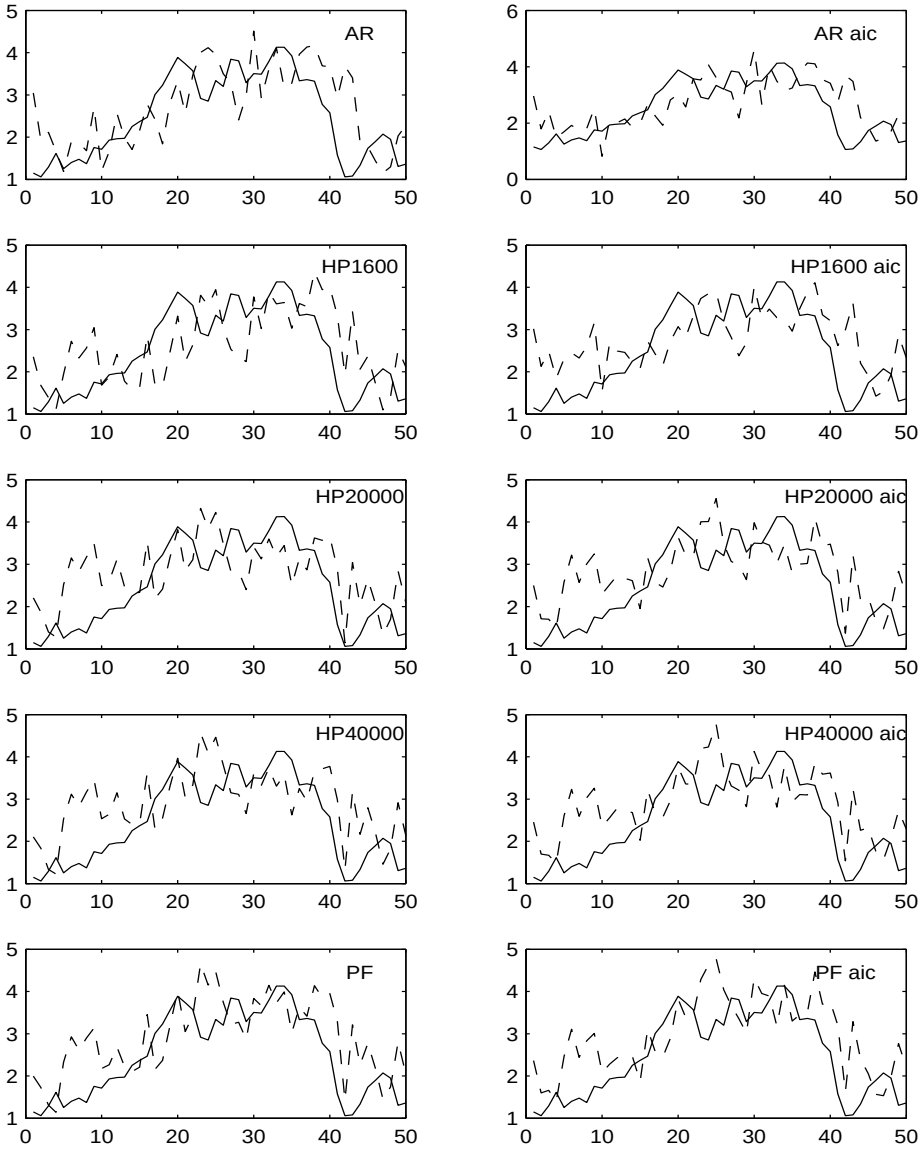
B Figures

Figure 5: *Real-time output gap estimates. "Thick modelling"*



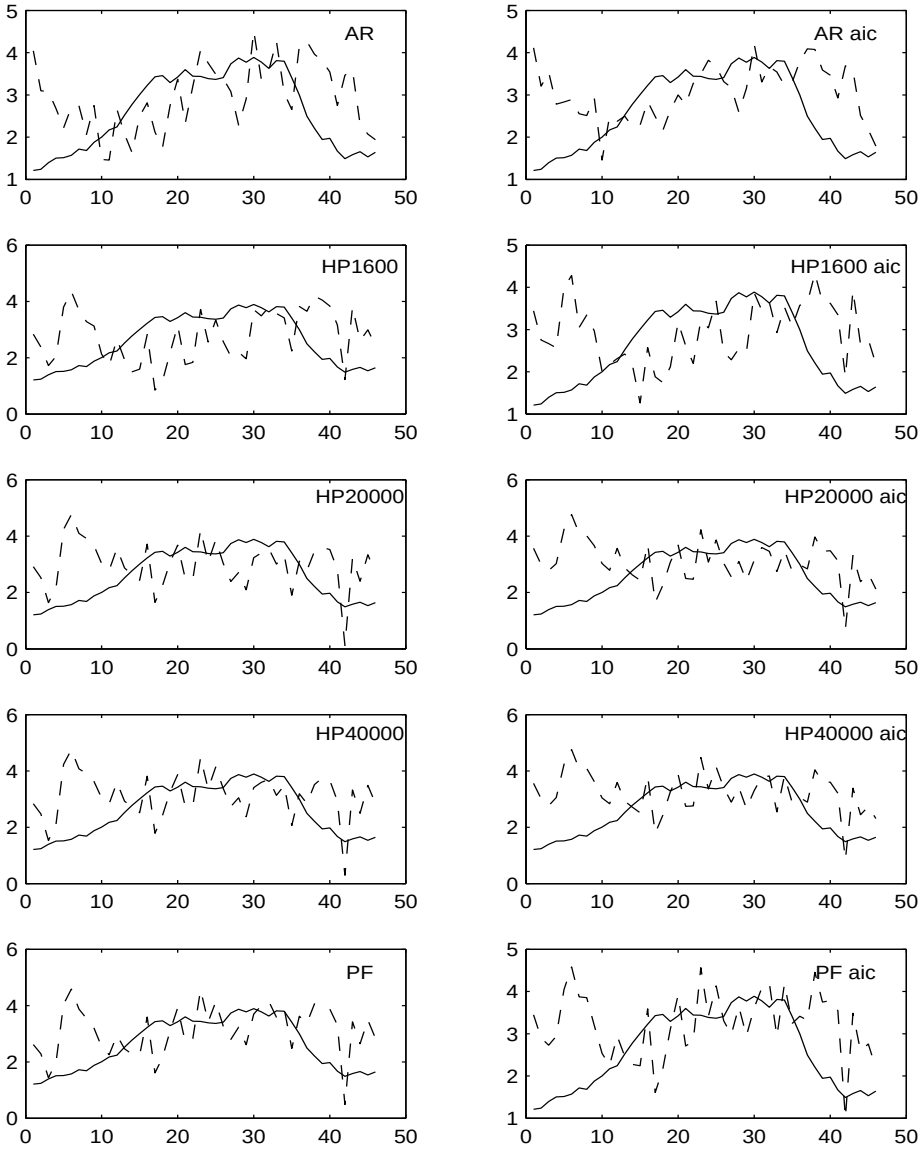
Notes: Output gap estimates are computed for each vintage, ranging from 93q1 to 06q2. In the subfigures all vintage estimates are shown for each output gap model. The horizontal axis displays the observations at each point in time, and the vertical axis displays the output gap. Typically the uncertainty of the output gap estimates becomes bigger at the endpoints of each vintage.

Figure 6: *Inflation forecasts, 4 quarter horizon*



Notes: The dashed lines are forecasts, and solid lines are observed inflation. The models leading to the forecasts displayed in figures in the second column are evaluated by AIC. The output gap is measured at the vertical axis, while time is measured at the horizontal axis. The forecasting period runs from 94q1 to 06q2.

Figure 7: *Inflation forecasts, 8 quarter horizon*



Notes: The dashed lines are forecasts, and solid lines are observed inflation. The models leading to the forecasts displayed in figures in the second column are evaluated by AIC. The output gap is measured at the vertical axis, while time is measured at the horizontal axis. The forecasting period runs from 95q1 to 06q2.