Erasmus University Rotterdam

Erasmus School of Economics

Master's Thesis Quantitative Finance (FEM21031-19)

The output gap: in search of the hidden truth

An out-of-sample forecasting study

Author
Ewout van Raad (448569)

Supervisor
dr. A. Pick (Erasmus School of Economics)
Second Assessor
prof. dr. C. Zhou (Erasmus School of Economics)

Date: August 25, 2020

The views stated in this paper are those of the authors and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This thesis assesses several real-time vintage-based models that incorporate data revisions in their ability to forecast the output gap and inflation for multiple horizons. The output gap itself is an unobserved variable and requires estimation. For this purpose I consider four detrending methods that generate measures of the output gap. I plug these measures into the models and perform a Bayesian out-of-sample forecasting exercise, where the resulting forecasts are densities and are evaluated by logarithmic score and MSFE. I find that the choice of model is not noticeably important for forecasting both variables. Furthermore, using combinations of forecasted densities belonging to a particular filtering method based on its inflation forecasting performance generally works well in tranquil times, but poorly in times of financial turmoil. Next to that, the results show indecisiveness in which filtering method is the most favourable, although the Beveridge-Nelson decomposition seems to provide some evidence of relative superiority.

Contents

1. Introduction

Having an adequate estimate of the output gap, defined as the difference between an economy's potential and actual output, is of paramount importance for macroeconomic decision making. Finding a reliable estimate is not trivial, since the variable itself is unobserved and therefore not straightforward to measure. Furthermore, the estimation of the Gross Domestic Product (GDP) on which the output gap is based may change over time due to data revisions. This possibly leads to poor initial estimates compared to later measurement of GDP.

Next to the incertitude of the GDP variable, there is vast disagreement on which particular method should be used to extract cyclical components from a time series. The output gap amounts to such a component. For instance, Garratt, Mitchell, and Vahey (2014) consider no less than seven methods to obtain an output gap measure. Much research focuses on this topic. Hamilton (2018) states that one should never opt for the quite famous Hodrick-Prescott (HP) filter proposed by Hodrick and Prescott (1997) and should use a regression-based filter instead. Morley and Wong (2020) opt for using a multivariate Beveridge-Nelson (BN) decomposition which uses multivariate models to obtain a univariate time series of cyclical components. Also, the univariate BN decomposition is popular (Beveridge & Nelson, 1981; Garratt et al., 2014; Morley, Nelson, & Zivot, 2003). Kamber, Morley, and Wong (2018) find that these BN decompositions provide reliable estimates and outperform the HP filter. However, there is no unambiguous conclusion as to which method consistently yields the most favourable results.

In this thesis, I consider models that explicitly model data revisions. Also, I utilise models that incorporate seasonality adjustments for annual and benchmark revisions. Finally, I include models that restrict the predictability of such data revisions. I show that a relatively simple real-time VAR based on vintages is not typically outperformed by seasonality adjusted or restricted models, implying that data revisions are predictable to a certain degree. The choice of model is then not as important.

Furthermore, I assess the quality of the filtering methods measured by out-of-sample forecasting performance of both the output gap and inflation. Particularly, this thesis focuses on density forecasts to measure the uncertainty in both variables. I consider the aforementioned four methods in my research and compare them using various model types and forecast combinations. I show that there is scant support for a persistently auspicious detrending method, but there are some indications that BN decompositions are superior. When combining forecasts based on their ability to forecast inflation in the past, we see that we often obtain propitious results in relatively calm financial regimes but poor outcomes in times of economic distress.

Extensive research has been done with regard to the output gap. Orphanides and Norden (2002) show that many methods of output gap estimation are quite unreliable due to data revision. They show that the magnitude of some revisions of the output gap are of similar size as the estimated gap itself. Moreover, they find that the effects of estimation error in measurement of trend output level and data revision are both significant for US data.

Garratt, Lee, Mise, and Shields (2008) alleviate the data revision issue by estimating a cointegrating VAR model which incorporates the revision process, generating forecasts of current and future output levels. By doing so, the model moderates the effect of revisions in the data. Also, it circumvents end-of-sample measurement problems by forecasting the future output. In this setting it is also possible to describe the uncertainty of the forecasts of the gap more precisely, compared to the approach by Orphanides and Norden (2002). They find that this approach outperforms the univariate methods in Orphanides and Norden (2002). However, this VAR approach uses all information that is available at the respective point in time and not only the vintage of data that is most recent.

Clements and Galvão (2012) follow the idea by Garratt et al. (2008) and propose various VAR specifications that model the dynamics of data releases in general. For instance, they account for seasonality of revisions. Additionally, they assume that after a certain amount of revisions, further revisions cannot be predicted. They show that real-time output gap estimates can be improved if both revisions to data in the past and future values after revision are predicted.

In this study, I incorporate the dynamics of data revision into the forecasting models following Clements and Galvão (2012). In contrast to their work, the work of Garratt et al. (2008), and that of Orphanides and Norden (2002), these models are not analysed in a frequentist setting but in a Bayesian way to shed light on the uncertain nature of the variable of interest.

Often, it is the case that using an ensemble of forecasts from different models leads to more favourable results compared to using forecasts from single models. The idea of combining forecasts is well established in empirical finance and this idea is explored by Garratt et al. (2014) in the output gap and inflation setting. They focus on one step ahead density forecasting based on seven different specifications and combining them by a logarithmic scoring rule. This elaborate approach however still indicates that substantial uncertainty is present.

To accentuate the unobserved state of the output gap, it is interesting to turn to density forecasting following Garratt et al. (2014). Density forecasting is an increasingly popular methodology, which gives an estimate of the probability distribution of the future values of a certain stochastic variable (Timmermann, 2000). In macroeconomics and risk management, it has become commonplace (Diebold, Gunther, & Tay, 1998). It is particularly interesting for our variables of interest since we know from the work by Garratt et al. (2014) that there is considerable uncertainty in these variables. Instead of just giving a point estimate or slightly more elaborate, a confidence interval, we are interested in the entire distribution to have as much information as possible about the possible outcome of the variable of interest.

Given that we have forecasting models (and combinations), we of course need to evaluate their performance. Many researchers use metrics such as average logarithmic (log) score and Mean Squared Forecast Error (MSFE) to evaluate their forecasts of output gap, assuming that there is a 'true' output gap. This truth is typically based on the most recent vintage of data available. Moreover, it assumes that some filtering method, such as the HP filter, decomposes the trend from its cyclical components perfectly. However, there is no such thing as the truth ultimately. It may be more interesting not only to measure the forecasts themselves for the sake of it, but to see whether output gap measures and models have predictive power for forecasting inflation as well. The idea that the output gap is important for inflation stems from economic theory. If the output gap is positive, it means the economy is above its potential output, which corresponds to a period with high inflation. If the output gap is negative, inflation should be low analogously.

Orphanides and Van Norden (2005) perform an exercise in which they research whether output gap forecasts can be useful for predicting inflation. They find that the predictive power of different measures of output gap is negligible. Simple linear models that include information on output growth provide more reliable forecasts than using output gap predictions. However, the models to forecast the output gap were not as elaborate as the methods present in the literature as of now, which may have influenced the results of Orphanides and Van Norden (2005) negatively.

In this thesis, I incorporate many of the previously mentioned forecasting ideas and methodologies, which I explore further than previous researchers. Firstly, the study focuses on density forecasts following Garratt et al. (2014). In contrast to their research I produce forecasts for multiple steps ahead instead of just forecasting one step ahead, for which the predictive density is known. This way we can examine the one-quarter-ahead forecasts and also, for instance, the two-years-ahead forecasts. This horizon is important to decision makers, since their target is often inflation in two years. Secondly, I produce ensemble forecasts following the same weighting rule used by Garratt et al. (2014). However, in their work they only provided a study of one-step-ahead forecast combinations. I construct ensemble forecasts for multiple forecasting horizons to examine their usefulness more thoroughly. Thirdly, I follow the literature in terms of the average log score and MSFE evaluation criteria. However, the evaluation is not only based on the output gap measure forecasts, or solely on inflation forecasts following the idea of Orphanides and Van Norden (2005), but rather both.

It remains to choose methods that gives us output gap measures to put in the forecasting models. In the paper by Garratt et al. (2008), the estimation of the 'true' output gap is performed by usage of the HP filter. However, we now know that there are multiple methods to give an estimate of the output trend. For example, Clements and Galvão (2012) use a band-pass filter proposed by Watson (2007). Orphanides and Norden (2002) consider three more of these methods, on top of the methods discussed earlier in the Introduction. In short, there are a plethora of methods that a researcher can use to separate a cyclical component from a trend component.

In this study, the filtering methods used are relatively well-known in the literature. However, the exact comparison between the four considered methods has not been made to the best of my knowledge, adding novelty to the study.

This brings us back to the research of this thesis. I investigate the suitability of various models and cyclical component extractors and observe that there is still indecisiveness in which output gap measure is optimal. Next to that, model choice seems to be unimportant.

The remainder of the thesis is composed as follows: Section 2 discusses the data set that I employ in this research, Section 3 describes the methods utilised in the thesis, why they are used, and how they are employed, Section 4 displays and interprets the results following from the methodology, Section 5 provides the conclusions of the research, and finally Section 6 discusses the limitations of this paper and suggests possible directions of further research.

2. Data

The data that I consider for this thesis is the same real-time data set that Garratt et al. (2008) and many others use in their research. It is the Philadelphia Federal Reserve Real-Time Macroeconomic Data Set of which I extract the real GDP and the Price Index for Personal Consumer Expenditures (PCE) quarterly data. The reason that I choose for quarterly timesteps and not monthly or yearly is as follows: one highly important forecasting horizon is the two-years-ahead forecast. For monthly data this would mean forecasting 24 steps ahead which makes the forecasts quite noisy and indistinguishable from different models. With annual data however, much information from within the year is (more or less) discarded which could make the forecasts relatively uninformative. Therefore, I deal with this trade-off by selecting quarterly data. The datasets consist of 217 vintages ranging from 1965Q4 to 2019Q4. From the PCE data set, I construct inflation as follows:

$$\pi_t^{t+1} = 400\ln\left(\frac{p_t^{t+1}}{p_{t-1}^{t}}\right), \tag{1}$$

in which p_t^{t+1} denotes the PCE index at time t measured in data vintage $t+1$.

3. Methodology

In this paper, I research the performance of certain predictive models and the eligibility of different filtering methods. First, we focus on the forecasting procedure. Afterwards, I discuss the various filtering methods. Finally, I describe in what way I use the forecasts to compare the usefulness of models and methods.

3.1. Forecasting procedure

Following the spirit of the work by Garratt et al. (2014), the dependent variable in our models is defined as the cyclical component of log output and inflation $y_t^{t+1} = [\text{filter}(Y_t^{t+1})', (\pi_t^{t+1})']'$, where Y_t^{t+1} is the natural logarithm of observation t of the output level in the vintage corresponding to $t+1$, π_t^{t+1} represents inflation of observation t in vintage $t+1$ and filter$(\cdot)$ represents a particular filtering method. The general forecasting procedure is to estimate multiple models with an expanding window to obtain forecasts for every vintage, creating a series of draws of the predictive density of the output gap and inflation for each point in time within the forecast sample. The first forecasts are made using information up to and including 1992Q4 in order to include an ample amount of observations for the models that produce the forecasts.

3.2. Bayesian forecasting

Rather than simply estimating points, I produce h-step-ahead with $h = 1, 2, 4, 8$ density forecasts of the output gap in order to incorporate substantially more information. This comes down to estimating the predictive density of y_{T+h}^{T+h+1}, the future observation, conditional on all observed data $\boldsymbol{Y}_T$, the information set. Note that I drop the superscript indicating vintage in the remainder of this Section for notational ease. To formalise this problem, we need to specify a loss function $\mathcal{L}(\boldsymbol{a}, y_{T+h})$ where y_{T+h} is the unobserved future data point vector and $\boldsymbol{a}$ represents a forecast vector of y_{T+h}. Our (Bayesian) objective is to choose $\boldsymbol{a}$ that minimises the expectation of the loss function conditional on $\boldsymbol{Y}_T$,:

$$\mathbb{E}[\mathcal{L}(\boldsymbol{a}, y_{T+h})|\boldsymbol{Y}_T] = \int \mathcal{L}(\boldsymbol{a}, y_{T+h}) p(y_{T+h}|\boldsymbol{Y}_T) dy_{T+h}, \tag{2}$$

where $p(y_{T+h}|\boldsymbol{Y}_T)$ corresponds to the predictive density of y_{T+h}. I specify the loss function as

$$\mathcal{L}(\boldsymbol{a}, y_{T+h}|\boldsymbol{Y}_T) = (\boldsymbol{a} - y_{T+h})'(\boldsymbol{a} - y_{T+h}), \tag{3}$$

with solution $\boldsymbol{a} = \mathbb{E}(y_{T+h}|\mathbf{Y}_T) = \boldsymbol{a}(\mathbf{Y}_T)$, where $a(\mathbf{Y}_T)$ indicates that the solution depends on the information set (Karlsson, 2013). Now, we are left to determine the form of the predictive distribution. This depends on three other distributions. First, we require the distribution of the future observation conditional on an unknown parameter vector $\boldsymbol{\theta}$ and the information set.

We denote this distribution as $p(y_{T+h}|\mathbf{Y}_T,\boldsymbol{\theta})$. Secondly we need the likelihood of the observed data conditional on the parameter vector, denoted as $L(\mathbf{Y}_T|\boldsymbol{\theta})$. Finally, we need to represent our beliefs on plausible values of the unknown parameters, the so called prior distribution $\pi(\boldsymbol{\theta})$. The likelihood is typically of the following form in time series forecasting (Karlsson, 2013)

$$L(\boldsymbol{Y}_T|\boldsymbol{\theta}) = \prod_{t=1}^{T} f(y_t|\mathbf{Y}_{t-1},\boldsymbol{\theta}). \tag{4}$$

The distribution on the future observation is then straightforward:

$$L(\boldsymbol{Y}_{T+1}|\boldsymbol{\theta}) = f(y_{T+1}|\mathbf{Y}_T,\boldsymbol{\theta}), \tag{5}$$

Using Bayes' rule we obtain the predictive density as

$$p(y_{T+1}|\mathbf{Y}_T) = \frac{p(y_{T+1},\mathbf{Y}_T)}{p(\mathbf{Y}_T)}, \tag{6}$$

$$= \frac{\int f(y_{T+1}|\mathbf{Y}_T,\boldsymbol{\theta})L(\mathbf{Y}_T|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int L(\mathbf{Y}_T|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{7}$$

We can use the following result

$$p(\boldsymbol{\theta}|\mathbf{Y}_T) = \frac{L(\mathbf{Y}_T|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\mathbf{Y}_T|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto L(\mathbf{Y}_T|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{8}$$

to obtain the following expression for the predictive density

$$p(y_{T+1}|\mathbf{Y}_T) = \int f(y_{T+1}|\mathbf{Y}_T,\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y}_T)d\boldsymbol{\theta}. \tag{9}$$

This specification neatly separates the underlying uncertainties. $f(y_{T+1}|\mathbf{Y}_T,\boldsymbol{\theta})$ describes the uncertainty of the future and $p(\boldsymbol{\theta}|\mathbf{Y}_T)$ accounts for the parameter uncertainty. Equation 9 is particularly interesting in the case when a closed form is not available for the predictive density. Often, it is impossible or extremely difficult to integrate out the parameters out of the joint distribution of y_{T+1} and $\boldsymbol{\theta}$. Equation 9 implies a simulation process that can be used to marginalise. If we can sample random values from the posterior $p(\boldsymbol{\theta}|Y_T)$, we can also draw y_{T+1} for each sample of $\boldsymbol{\theta}$ by drawing from $f(y_t|\mathbf{Y}_{t-1},\boldsymbol{\theta})$. This corresponds to a draw of the joint distribution of $(\boldsymbol{\theta}, y_{T+1})$ conditional on the information set. We can simply exclude the draw of the parameter vector and obtain a sample of the predictive distribution if repeated a fixed sufficiently large number R times. The distribution can then be utilised to estimate the desired features and metrics such as conditional mean or quantile values (Karlsson, 2013).

3.3. Forecasting models

3.3.1. Vintage-based VAR (V-VAR) model

The idea of Clements and Galvão (2012) is to not only model output, but also the revisions of data. If we assume that there are $q-1$ revisions in subsequent quarters for any y_t^{t+1} and none after, we can implement the following V-VAR model inspired by Clements and Galvão (2012):

$$y^{t+1} = c + \sum_{i=1}^{p} \Gamma_i y^{t+1-i} + \epsilon^{t+1}, \tag{10}$$

where we define $y^{t+1} = \left[y_t^{t+1}, y_{t-1}^{t+1}, \ldots, y_{t-q+1}^{t+1}, \pi_t^{t+1}\right]'$, $y^{t+1-i} = \left[y_{t-i}^{t+1-i}, y_{t-1-i}^{t+1-i}, \ldots, y_{t-q+1-i}^{t+1-i}, \pi_{t-i}^{t+1-i}\right]'$, and c and ϵ^{t+1} are both $(q+1)\times 1$ vectors. The former represents a vector of intercepts whereas the latter corresponds to a Gaussian error term with zero mean. Finally, Γ_i represents a $(q+1)\times(q+1)$ parameter matrix. This way, the model includes both the contemporary observation y_t^{t+1} and values of past observations that are revised $y_{t-1}^{t+1}, \ldots, y_{t-q+1}^{t+1}$. The first equation models the first-release value y_t^{t+1}. The second one models the second-release data y_{t-1}^{t+1} (data after one revision) and equations go on until $q-1$ revisions are modelled.

The choice of lag order p should theoretically be as high as possible to include all revisions, but this may lead to high-dimensionality problems and egregious computation times. For this reason, the number of lags should be limited and in this paper I set it to one following Clements and Galvão (2012). We can rewrite Equation 10 as

$$y_t' = z_t'\mathbf{\Gamma} + u_t', \tag{11}$$

where $y_t' = (y^t)'$ to avoid double superscript confusion, $z_t' = (y_{t-1}', \mathbf{1})$ is a $k = mp+1$ dimensional vector, $\mathbf{\Gamma} = (\Gamma_i', c')'$ is a k x m matrix and Gaussian noise $u_t' \sim N(0, \boldsymbol{\Psi})$. Then we have $f(y_t|\boldsymbol{Y}_{t-1}, \boldsymbol{\theta}) = N(y_t; z_t'\mathbf{\Gamma}, \boldsymbol{\Psi})$. We take a diffuse prior, in the form of a uniform prior for $\mathbf{\Gamma}$ and a Jeffreys' prior for Ψ,

$$\pi(\mathbf{\Gamma}, \boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}|^{-(m+1)/2}. \tag{12}$$

We can write this model in matrix form as

$$\boldsymbol{Y} = \boldsymbol{Z}\mathbf{\Gamma} + \boldsymbol{U}, \tag{13}$$

where $\boldsymbol{Y} = [y_1', y_2', \cdots, y_T']$, $\boldsymbol{Z} = [z_1', z_2', \cdots, z_T']$ and $\boldsymbol{U} = [u_1', u_2', \cdots, u_T']$ with likelihood

$$L(\boldsymbol{Y}|\mathbf{\Gamma}, \boldsymbol{\Psi}) = (2\pi)^{-mT/2}|\boldsymbol{\Psi}|^{-T/2}\exp\left\{-\frac{1}{2}\sum(y_t' - z_t'\mathbf{\Gamma})\boldsymbol{\Psi}^{-1}(y_t' - z_t'\mathbf{\Gamma})'\right\}, \tag{14}$$

$$= (2\pi)^{-mT/2}|\boldsymbol{\Psi}|^{-T/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[(\boldsymbol{Y}-\boldsymbol{Z\Gamma})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{Z\Gamma})']\right\}, \tag{15}$$

$$= (2\pi)^{-mT/2}|\boldsymbol{\Psi}|^{-T/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{Z\Gamma})'(\boldsymbol{Y}-\boldsymbol{Z\Gamma})]\right\}, \tag{16}$$

We can add and subtract $\boldsymbol{Z}\hat{\boldsymbol{\Gamma}}$, with OLS estimate $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ and multiply with the prior which yields the joint posterior:

$$p(\boldsymbol{\Gamma},\boldsymbol{\Psi}|Y_T) \propto |\boldsymbol{\Psi}|^{-T/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Psi}^{-1}(\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\Gamma}})'(\boldsymbol{Y}-\boldsymbol{Z\Gamma})]\right\}\times \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Psi}^{-1}(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})'\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})\right]\right\}|\boldsymbol{\Psi}|^{-(m+1)/2}. \tag{17}$$

If we consider the part of the equation that includes $\boldsymbol{\Gamma}$ we state that

$$\mathrm{tr}\left[\boldsymbol{\Psi}^{-1}(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})'\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}})\right] = (\gamma-\hat{\gamma})'(\boldsymbol{\Psi}^{-1}\otimes\boldsymbol{Z}'\boldsymbol{Z})(\gamma-\hat{\gamma}), \tag{18}$$

with $\gamma = \mathrm{vec}(\boldsymbol{\Gamma})$ and $\hat{\gamma} = \mathrm{vec}(\hat{\boldsymbol{\Gamma}}) = [\boldsymbol{I}_m\otimes(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\mathrm{vec}(\boldsymbol{Y})$. This corresponds to a kernel of a multivariate normal distribution conditional on $\boldsymbol{\Psi}$,

$$\gamma|\mathbf{Y}_T,\boldsymbol{\Psi} \sim N(\hat{\gamma},\boldsymbol{\Psi}\otimes(\boldsymbol{Z}'\boldsymbol{Z})^{-1}). \tag{19}$$

We can also see it as a matricvariate normal distribution for $\boldsymbol{\Gamma}$ and hence write $\boldsymbol{\Gamma}|\mathbf{Y}_T,\boldsymbol{\Psi} \sim MN_{km}(\hat{\boldsymbol{\Gamma}},\boldsymbol{\Psi},(\boldsymbol{Z}'\boldsymbol{Z})^{-1})$. If we integrate out γ from the joint posterior distribution gives us the marginal posterior distribution for $\boldsymbol{\Psi}$ as follows (Karlsson, 2013):

$$p(\boldsymbol{\Psi}|\mathbf{Y}_T) \propto |\boldsymbol{\Psi}|^{-(T+m+1-k)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}[\boldsymbol{\Psi}^{-1}\boldsymbol{S}]\right\}, \tag{20}$$

with $\boldsymbol{S} = (\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\Gamma}})'(\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\Gamma}})$. We observe that this is a kernel of an inverse Wishart distribution with scale matrix $\boldsymbol{S}$ and $T-k$ degrees of freedom

$$\boldsymbol{\Psi}|\mathbf{Y}_T \sim iW_m(\boldsymbol{S},T-k). \tag{21}$$

Karlsson (2013) states that the one-step-ahead predictive density for y_{T+1} is matricvariate t, $Mt_{1m}(z'_{T+1}\hat{\boldsymbol{\Gamma}},(1+z'_{T+1}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}z_{T+1})^{-1},\boldsymbol{S},T-k)$. For longer horizons, we do not have closed form expressions for those predictive densities. I implement the following algorithm used by Karlsson (2013) to simulate them:

Algorithm 1: Direct sampler for simulating the predictive densities for *h*-step VAR forecasts

```
for $j = 1, \dots, R$ do
    Sample $\boldsymbol{\Psi}^{(j)}$ from marginal posterior $\boldsymbol{\Psi}|\mathbf{Y}_T \sim iW_m(\boldsymbol{S}, T-k)$;
    Generate $\boldsymbol{\Gamma}^{(j)}$ from the conditional posterior $\boldsymbol{\Gamma}|\mathbf{Y}_T, \boldsymbol{\Psi}^{(j)} \sim MN_{km}(\hat{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}^{(j)}, (\boldsymbol{Z}'\boldsymbol{Z})^{-1})$;
    Generate $u_{T+1}^{(j)}, \dots, u_{T+h}^{(j)}$ from $u_t \sim N(0, \boldsymbol{\Gamma}^{(j)})$;
    Compute recursively;
        $$\tilde{y}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \tilde{y}_{T+h-i}^{(j)\prime} \Gamma_i^{(j)} + \sum_{i=h}^{p} y_{T+h-i}^{\prime} \Gamma_i^{(j)} + c^{(j)\prime} + u_{T+h}^{(j)}. \qquad (22)$$
end
Obtain independent sample of joint predictive distribution $\{\tilde{y}_{T+1}, \cdots, \tilde{y}_{T+h}\}_{j=1}^{R}$
```

3.3.2. Accounting for seasonality of data revisions

Data revisions are generally not random. For instance, annual revisions are always performed in the third quarter of the year, although it is not always done. The Bureau of Economic Analysis (BEA) implements benchmark revisions as well, typically once every five years (Clements & Galvão, 2012). The V-VAR model does not account for this seasonality in revisions. Typically, the elements beyond the first and second release data are unrevised unless there is such a benchmark or an annual revision. (Clements & Galvão, 2012). Therefore, I define $D_1^{t+1} = 1$ if there was an annual revision in the third quarter of a year and $D_2^{t+1} = 1$ if either an annual or benchmark revision has been implemented. For $p = 1$ we get the following model (Clements & Galvão, 2012):

$$y^{t+1} = \left[\tilde{c} + \tilde{\boldsymbol{\Gamma}}_1 y^t\right](1 - D_s^{t+1}) + [c + \boldsymbol{\Gamma}_1 y^t] D_s^{t+1} + v^{t+1}, \qquad (23)$$

with zero-mean Gaussian error term v^{t+1} and

$$\tilde{\boldsymbol{\Gamma}}_1 = \begin{bmatrix} & \gamma_{2\times q} & \\ \boldsymbol{0}_{(q-2)\times 1} & \boldsymbol{I}_{(q-2)\times(q-2)} & \boldsymbol{0}_{(q-2)\times 1} \\ & \gamma_{1\times q} & \end{bmatrix}, \qquad (24)$$

$s \in \{1, 2\}$ and $\tilde{c} = (c_1, c_2, 0, \dots, 0, c_{15})'$. When the dummy variable equals zero, the model only incorporates nonzero parameters on the first two rows and the last inflation row. For later than second-release revisions we have unity, corresponding to the notion that later revisions are often equal to zero (first difference is then zero). If the dummy equals one, the model is reduced to the V-VAR. Intercepts for later revisions are only estimated when the dummy variable is equal to one as we can see from the construction of $\tilde{c}$. This model now captures the influence of the annual or benchmark revisions. If the forecasting quarter is not Q4, we have that later revisions are zero. When there is a revision, the revisions are determined by the parameter matrix $\boldsymbol{\Gamma}_1$. The model corresponding to $s = 1$ is defined as the seasonal vintage-based VAR (SV-VAR) and the model that adheres to $s = 2$ is defined as the seasonal and benchmark vintage-based VAR (SBV-VAR).

3.3.3. Restricted VAR model

The questions also arises to which degree revisions are predictable. We expect that the predictability declines over the number of revisions. Therefore, as in the paper by Clements and Galvão (2012), I specify a V-VAR model that imposes the restriction that after $n-1$ revisions, the next revision $y_t^{t+n+1} - y_t^{t+n}$ is independent of previous value y_t^{t+n}. We can apply this restriction to the coefficient matrix $\boldsymbol{\Gamma}_1$ from Equation 10 as follows:

$$\tilde{\boldsymbol{\Gamma}}_1 = \begin{bmatrix} & \gamma_{n\times q} & \\ \mathbf{0}_{(q-n)\times(n-1)} & \boldsymbol{I}_{(q-n)\times(q-n)} & \mathbf{0}_{(q-n)\times 1} \\ & \gamma_{1\times q} & \end{bmatrix}, \tag{25}$$

Following Clements and Galvão (2012), n is set to two which implies that the values after two revisions (third-release values) are efficient. We obtain for $p = 1$ and $n = 2$

$$y^{t+1} = c + \tilde{\boldsymbol{\Gamma}}_1 y^t + v^{t+1}, \tag{26}$$

with $\tilde{\boldsymbol{\Gamma}}_1$ from Equation 24. We allow the revisions to have a non-zero mean by incorporating unrestricted intercepts. This model is defined as the news-restricted vintage-based VAR (RV-VAR).

These models, including the S(B)V-VAR models, need to be estimated as Seemingly Unrelated Regressions (SUR) since their errors are likely correlated (Clements & Galvão, 2012). For this purpose I define the following matrix:

$$x_S^t = \begin{bmatrix} (1, y_t') & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \cdots & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} \\ \mathbf{0}_{1\times(q+2)} & (1, y_t') & \mathbf{0}_{1\times(q+2)} & \cdots & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} \\ \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & (D_s^{t+1}, D_s^{t+1} \times y_t') & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \cdots & (D_s^{t+1}, D_s^{t+1} \times y_t') & \mathbf{0}_{1\times(q+2)} \\ \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \cdots & \mathbf{0}_{1\times(q+2)} & (1, y_t') \end{bmatrix},$$

where again $y_t' = (y^t)'$ to avoid double superscript confusion. This is a $(q+1) \times (q+1)(q+2)$ matrix of independent variables, where subscript S denotes the seasonality VAR models. It is straightforward to see that if we reach a third quarter or benchmark revision, the twelve equations above the bottom inflation equation become nonzero. For the RV-VAR we obtain $(q+1)\times(3(q+1))$ matrix

$$x_R^t = \begin{bmatrix} (1,y_t') & \mathbf{0}_{1\times(q+2)} & 0 & \mathbf{0}_{1\times 11} & \mathbf{0}_{1\times(q+2)} \\ \mathbf{0}_{1\times(q+2)} & (1,y_t') & 0 & \mathbf{0}_{1\times 11} & \mathbf{0}_{1\times(q+2)} \\ \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & 1 & \mathbf{0}_{1\times 11} & \mathbf{0}_{1\times(q+2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times 11} & 1 & \mathbf{0}_{1\times(q+2)} \\ \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times(q+2)} & \mathbf{0}_{1\times 11} & 0 & (1,y_t') \end{bmatrix},$$

where subscript R denotes the restricted VAR model. This way, we restrict the predictability of the revisions to the first two revisions and allow for nonzero higher-order revisions which I assume to be unpredictable in this model. Following Clements and Galvão (2012) we then obtain our model for the S(B)V-VAR

$$y^{t+1} = \boldsymbol{A}y^t(1-D_s^{t+1}) + x_S^t\beta + \nu^{t+1}, \tag{27}$$

with auxiliary $(q+1)\times(q+1)$ matrix

$$\boldsymbol{A} = \begin{bmatrix} & \mathbf{0}_{2\times(q+1)} & \\ \mathbf{0}_{(q-2)\times 1} & \boldsymbol{I}_{(q-2)\times(q-2)} & \mathbf{0}_{(q-2)\times 2} \\ & \mathbf{0}_{1\times(q+1)} & \end{bmatrix}.$$

For the RV-VAR variant, the equation is simply

$$y^{t+1} = x_R^t\beta + \nu^{t+1}, \tag{28}$$

which can be estimated directly with as SUR. Lastly, I define $z_t = (y^{t+1} - \boldsymbol{A}y^t(1-D_s^{t+1}))$ for the seasonality models. Then we can estimate the following model as SUR:

$$z_t = x_S^t\beta + \nu^{t+1}. \tag{29}$$

In contrast to Clements and Galvão (2012) I produce density forecasts instead of (only) point estimates. We do not have a closed form expression for h-step-ahead forecasts for any horizon h (Percy, 1992). For the SUR models, we consider the same prior as for the VAR model. However, we now consider precision matrix $\boldsymbol{\Phi} = \boldsymbol{\Psi}^{-1}$ instead of the regular covariance matrix, which is more common in the SUR and GLS literature. The elements of this matrix follow a Wishart (W) distribution instead of an inverted Wishart distribution. Combining the work of Percy (1992) and Karlsson (2013) I propose the following sampling algorithm for the SUR models:

Algorithm 2: Direct sampler for simulating predictive densities for h-step SUR forecasts

```
for $j = 1, \ldots, R$ do
    Sample $\mathbf{\Phi}^{(j)}$ from marginal posterior $\mathbf{\Phi}|\mathbf{Y}_T \sim W_m((\sum_{t=1}^{T}(z_t - x^t\beta)(z_t - x^t\beta)')^{-1}, T-k)$;
    Generate $\beta^{(j)}$ from the conditional posterior
     $\beta|\mathbf{Y}_T, \mathbf{\Phi}^{(j)} \sim N_{(q+1)(q+2)}\Big((\sum_{i=1}^{T} x'^t\mathbf{\Phi}x^t)^{-1}(\sum_{i=1}^{T} x'^t\mathbf{\Phi}z_t), (\sum_{i=1}^{T} x'^t\mathbf{\Phi}x^t)^{-1}\Big)$;
    Generate $u_{T+1}^{(j)}, \ldots, u_{T+h}^{(j)}$ from $u_t \sim N(0, \mathbf{\Phi}^{-1(j)})$;
    for $i = 1, \ldots, h$ do
        Compute
            $$\tilde{z}_{T+i}^{(j)'} = x_S^{T+i-1}\beta^{(j)} + u_{T+i}^{(j)}. \qquad (30)$$
         Add $Ay^{T+i-1}(1 - D_s^{T+i})$ to $\tilde{z}_{T+i}^{(j)'}$ for transformation to $\tilde{y}_{T+i}^{(j)'}$;
        Construct updated explanatory variable matrix $x_S^{T+i}$ by plugging in $\tilde{y}_{T+i}^{(j)'}$;
    end
end
Obtain independent sample of joint predictive distribution $\{\tilde{y}_{T+1}, \cdots, \tilde{y}_{T+h}\}_{j=1}^{R}$;
```

For the RV-VAR model, the algorithm is similar. We do not need the transformation incorporated within Algorithm 2 since we do not have to transform our y-variable to get an appropriate model. Note that the transformation of variable here is justified without any extra terms since the Jacobian of the transformation is equal to one in this case.

3.4. Extracting cyclical components

The output gap should represent the distance between the (theoretical and unobserved) potential output of an economy and its actual output level. Hence, we know that it is not a variable that is observed such as inflation or GDP, of which we simply have the observations. To combat this issue, I utilise four approaches that aim to separate the potential or trend output from the observed series, creating a cyclical component.

3.4.1. Hodrick-Prescott filter

The first method that I consider for extracting the deviations from the trend is the filter proposed by Hodrick and Prescott (1997). It is one of the most commonly used filters in the output gap framework. It states that there is a additive relation between a trend component denoted as g_t and a cyclical component denoted as c_t for a series y_t, i.e.

$$y_t = g_t + c_t. \qquad (31)$$

We are interested in finding $c_t = y_t - g_t$. Since g_t is unobserved, we instead compute $\hat{c}_t = y_t - \hat{g}_t$, where $\hat{g}_t$ is the trend estimate from the HP filter. The trend of y_t is computed as

$$\min_{\{g_t\}_{t=-1}^{T}} \left\{ \sum_{t=1}^{T} (y_t - g_t)^2 + \lambda \sum_{t=1}^{T} [(g_t - g_{t-1}) - (g_{t-1} - g_{t-2})]^2 \right\}, \tag{32}$$

where T represents the number of observations and λ is a smoothing parameter. If $\lambda \rightarrow 0$, the trend estimate would simply equal the series itself. If $\lambda \rightarrow \infty$, then the result would adhere to a regression on a simple linear time trend, since the generated series $\hat{g}_t$ would have precisely zero second differences. As a rule of thumb for quarterly data, Hodrick and Prescott (1997) suggest a value of $\lambda = 1600$ which I follow. To elucidate the composition of any implied series $\hat{g}_t$, we can write a closed-form expression for it in vector notation. For this purpose, define $\underset{(Tx1)}{y} = (y_T, y_{T-1}, \ldots, y_1)'$ and $\underset{((T+2)x1)}{g} = (g_T, g_{T-1}, \ldots, g_{-1})'$. Further define

$$\underset{(Tx(T+2))}{\boldsymbol{H}} = \begin{bmatrix} \underset{(TxT)}{\boldsymbol{I}_T} & \underset{(Tx2)}{\boldsymbol{0}} \end{bmatrix},$$

in which $\boldsymbol{I}$ is the identity matrix. Finally, denote

$$\underset{(Tx(T+2))}{\boldsymbol{Q}} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix}. \tag{33}$$

We can then write the solution to the minimisation problem as

$$g^* = (\boldsymbol{H}'\boldsymbol{H} + \lambda \boldsymbol{Q}'\boldsymbol{Q})^{-1}\boldsymbol{H}'y, \tag{34}$$

which shows that the solution at any point in time can be written as a linear function of the whole series y including all observations, also from the future. This often leads to end-of-sample measurement problems (Hamilton, 2018).

3.4.2. Regression-based filter (Hamilton, 2018)

Hamilton (2018) states that there are major issues with using the HP filter as a detrending method. First of all, he claims that the generated series have no basis in the data generating process and are hence spurious. Secondly, filtered values at the midpoint of the sample differ substantially from the end-of-sample estimates. Moreover, he finds that setting $\lambda = 1600$ is often false when a statistical approach is utilised to obtain an estimate for this parameter. He proposes a quite simple and robust methodology to obtain the desired cyclical components without the issues inherited by the HP filter. He shows that the residuals of a linear regression of the following form can be used as a reasonable estimate of the transient component for various processes:

$$y_{t+h} = \phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \phi_3 y_{t-2} + \phi_4 y_{t-3} + \epsilon_{t+h}. \tag{35}$$

After fitting the regression, our estimate of the output gap is $\hat{\epsilon}_{t+h}$. Hamilton (2018) suggests to set $h = 8$ for quarterly data, since a business cycle typically lasts for 2 years.

3.4.3. Beveridge-Nelson decomposition

Another approach is the decomposition devised by Beveridge and Nelson (1981). This alternative method defines a permanent component of a non-stationary time series $\{s_t\}_{t=1}^{T}$ with a drift μ as a forecast for an infinite horizon adjusted for the drift rate:

$$g_t = \lim_{h\to\infty} s_{t+h|t} - \mu h, \tag{36}$$

$$g_t = TD_t + BN_t, \tag{37}$$

in which TD_t represents the deterministic part of the trend and BN_t the stochastic part, which is called the BN trend. This way, the trend can be allowed to have a stochastic component as well. The idea that the BN decomposition reflects is that the conditional expectation of the long-term cyclical component of a time series is equal to zero. As a result, this implies that the long-term conditional expectation of a time series is represented by its trend. Then we simply need to fit a model to predict the required trend. Typically ARIMA models are often used for univariate time series (Beveridge & Nelson, 1981; Morley et al., 2003), whereas VAR models are suitable for multivariate time series (Morley & Wong, 2020). In a univariate setting, Beveridge and Nelson (1981) show that if Δs_t has a Wold representation

$$\Delta s_t = \mu + \psi^*(L)\epsilon_t, \tag{38}$$

then

$$BN_t = BN_{t-1} + \psi^*(1)\epsilon_t = BN_0 + \psi^*(1)\sum_{i=1}^{t}\epsilon_i, \tag{39}$$

which corresponds to a random walk model without drift. For the derivation of the BN trend, I refer to the original paper by Beveridge and Nelson (1981). Then, we are interested in $c_t = s_t - TD_t - BN_t$ as our candidate output gap measure. I consider an ARIMA(1,1,0) model for s_t

for this purpose (so AR(1) for Δs_t), for which it is straightforward to compute the trend:

$$
\begin{aligned}
TD_t + BN_t &= \lim_{h\to\infty} (s_{t+h|t} - h\mu), && (40)\\
&= s_t + (\Delta s_t - \mu) \lim_{h\to\infty} \sum_{s=1}^{h} \phi^h, && (41)\\
&= s_t + \frac{\phi}{1-\phi}(\Delta s_t - \mu). && (42)
\end{aligned}
$$

We can then easily compute the cyclical component as

$$
\begin{aligned}
c_t &= s_t - TD_t - BN_t, && (43)\\
&= \frac{\phi}{1-\phi}(\Delta s_t - \mu). && (44)
\end{aligned}
$$

Since this decomposition only relies on historical data, it is highly suitable for generating real-time estimates.

In a multivariate setting, let us denote Δx_t as a vector of exogenous variables including our (differenced) target variable y_t. I assume that Δx_t follows a VAR(p) process which we can rewrite to VAR(1) in the following way:

$$
(\Delta \boldsymbol{X}_t - \mu) = \boldsymbol{F}(\Delta \boldsymbol{X}_{t-1} - \mu) + \boldsymbol{H} e_t, \tag{45}
$$

where $\Delta \boldsymbol{X_t} = \{\Delta x_t', \Delta x_{t-1}', \ldots, \Delta x_{t-p+1}'\}'$, $\boldsymbol{F}$ denotes the companion matrix, μ a vector of means, $\boldsymbol{H}$ an auxiliary matrix to map the forecast errors to companion form and finally e_t a series of forecast errors. We define Σ as the covariance matrix of these forecast errors. If the variables in Δx_t are stationary, then $(\boldsymbol{I} - \boldsymbol{F})^{-1}$ exists which allows us to write the sum of expected deviations of the multivariate series from its mean vector (Morley & Wong, 2020):

$$
\mathbb{E}_t \sum_{j=1}^{\infty} (\Delta \boldsymbol{X}_{t+j} - \mu) = \boldsymbol{F}(\boldsymbol{I} - \boldsymbol{F})^{-1}(\Delta \boldsymbol{X}_t - \mu). \tag{46}
$$

Morley (2002) shows that we can solve for the vectors of BN trends and cycles, denoted as τ_t and c_t respectively, as

$$
\begin{aligned}
\tau_t &= \boldsymbol{X}_t + \boldsymbol{F}(\boldsymbol{I} - \boldsymbol{F})^{-1}(\Delta \boldsymbol{X}_t - \mu), && (47)\\
c_t &= -\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{F})^{-1}(\Delta \boldsymbol{X}_t - \mu). && (48)
\end{aligned}
$$

The question arises which variables should be included in Δx_t. Morley and Wong (2020) propose a set of 23 variables, based on the findings by Bańbura, Giannone, and Reichlin (2010), which are included in the appendix. However, five of these variables are (based on) inflation. Since I forecast inflation itself it would give this method an unfair advantage if I include inflation information in the output gap estimate. Hence, I remove these five variables.

3.5. Assessing the informative value of forecasted measures

Typically, researchers evaluate their forecasts directly based on a loss function such as (Root) Mean Squared Forecast Error. However, only considering such a loss function gives relatively low information about which output gap measure is the most informative, since evaluating the forecasts directly only says something about the models' performance. Hence, it is interesting to develop a methodology that tests the performance of the models and also researches the added value of certain filtering methods.

As mentioned before, there is a causality between inflation and the output gap. They should be positively correlated, since an increase in output gap generally corresponds to increases in inflation. Particularly, the causality is likely such that output gap (Granger) causes inflation. I make predictions using the described VAR and SUR models models including inflation and output gap measures both as endogenous variables, following Garratt et al. (2014).

Using these models, we can conclude whether there are models that are consistently superior in terms of out-of-sample forecasting based on MSFE and average logarithmic score for inflation, which is defined as $\sum_{t=1}^{T} \frac{1}{T} \ln g(\pi_t^{t+h}|Y_t)$, where $g(\cdot|Y_t)$ represents the forecasted probability density function (pdf) conditional on the information up to and including time t. The scoring rule comes down to evaluating the forecasted pdf value at the outcome of inflation. The motivation for logarithmic scoring is relatively intuitive since a good estimate of the pdf should assign a high density value to the actual outcome, meaning that a high log score corresponds to accurate forecasting. Note that this score is not bounded between 0 and 1, since evaluating a pdf at a certain point does not yield the probability of its occurrence but the density at that point. Hence, it is unbounded. I also consider the average log score for the univariate output gaps. This can tell us about how predictable particular output gap measures are and compare how well the various models are able to forecast them relative to each other.

However, computing the log score is not straightforward in this case since we do not have a closed form expression for the density forecasts and only have the draws from the sampler. Hence, we need to estimate the density values. We can do this non-parametrically with a Kernel Density Estimator (KDE). In general, we have that the density estimate of a single point y with explanatory variables $x_i, i = 1, \ldots, T$ is as follows:

$$\hat{p}(y) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{h}K(y - x_i; h), \tag{49}$$

where K(·) denotes a positive function representing a kernel and h is a bandwidth parameter (Sheather & Jones, 1991). The choice of this bandwidth parameter vehemently influences the KDE, more so than the choice of kernel (P. Hall, Sheather, Jones, & Marron, 1991). Following P. Hall et al. (1991) among others, I use Gaussian kernels of which we know that

$$K(x; h) \propto \exp\left(-\frac{x^2}{2h^2}\right). \tag{50}$$

For bandwidth selection we use a rule of thumb proposed by Scott (1992). h is then given by $1.059\hat{\sigma}T^{-\frac{1}{5}}$, where $\hat{\sigma}$ is the (sample) standard deviation of the data. With this setup, we are able to find the pdf values necessary. Next to the direct approach of evaluating log scores of models individually, it is also possible to create a weighting scheme for the output gap densities based on their performance in forecasting inflation following Garratt et al. (2014). To do this, I denote the ensemble inflation density for a model specification $i = 1, \cdots, N$, where $N = 4$ as

$$p^{\text{filter}}(\pi_t^{t+h}) = \sum_{i=1}^{N} w_{i,t}\, g^{(i)}(\pi_t^{t+h}|\mathbf{Y}_t), \tag{51}$$

where $g^{(i)}(\cdot|\mathbf{Y}_T)$ represents a density forecast from model i. I determine the weights $w_{i,t}$ in the following way:

$$w_{i,t} = \frac{\exp[\frac{1}{r}\sum_{k=t-r+1}^{t}\ln g^{(i)}(\pi_k^{k+h}|\mathbf{Y}_t)]}{\sum_{i=1}^{N}\exp[\frac{1}{r}\sum_{k=t-r+1}^{t}\ln g^{(i)}(\pi_k^{k+h}|\mathbf{Y}_t)]}. \tag{52}$$

Here, r denotes the length of a rolling window of included log scores, which I set to eight. This way, the weights from Equation 52 are based on the average log score over the last eight quarters. Using those weights, we can obtain an ensemble output gap forecast:

$$O^{\text{filter}}(y_t^{t+h}) = \sum_{i=1}^{N} w_{i,t}\, g^{(i)}(y_t^{t+h}|\mathbf{Y}_t), \tag{53}$$

for all filtering methods and forecasting horizons. By also evaluating these ensemble forecasts, we know if combining forecasts from different models is useful and we can see if predictive power of inflation is important for accurately forecasting the output gap.

Finally, I examine whether the log scores and MSFEs of the SUR models are significantly better than those of the (simplest) V-VAR model. For the MSFE, I consider the test developed by Diebold and Mariano (1995). In order to compute the Diebold-Mariano (DM) statistic, I construct loss differentials based on the squared forecast errors of a specific SUR model i and the V-VAR model (superscript again dropped for notational ease),

$$D_t = e^2_{\text{V-VAR},t} - e^2_{i,t}. \tag{54}$$

Note that our loss function specified in Subsection 3.2 implies that our point estimate is the mean of the predictive density. The test involves computing the sample mean of the loss differential

$$\bar{D} = \frac{1}{T}\sum_{t=1}^{T} D_t. \tag{55}$$

The series of the loss differentials are known to exhibit significant autocorrelations (Harvey, Leybourne, & Newbold, 1997). Harvey et al. (1997) state that one can compute the asymptotic variance of $\bar{D}$ as follows:

$$V(\bar{D}) \approx \frac{\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k}{T}, \tag{56}$$

where γ_k denotes the k-th order autocovariance of D_t and h is the forecasting horizon. We estimate the autocovariance in the standard way:

$$\hat{\gamma}_k = \sum_{t=k+1}^{T} (D_t - \bar{D})(D_{t-k} - \bar{D}). \tag{57}$$

This leads us to the computation of the DM statistic:

$$DM = \frac{\bar{D}}{\sqrt{\hat{V}(\bar{D})}}, \tag{58}$$

which is standard normally distributed under the null hypothesis of a zero mean.

In the research of this thesis, the number of observations is relatively limited. Harvey et al. (1997) state that the DM test is typically too large when computed over a forecast sample with a small or intermediate size. They show that one can account for this issue by multiplying the original DM statistic with a constant which varies over the forecasting horizon h:

$$DM^* = \left(\frac{T + 1 + 2h + T^{-1}h(h-1)}{T}\right) DM. \tag{59}$$

Moreover, Harvey et al. (1997) argue that the comparison to the standard normal distribution may be inappropriate. Instead, they opt for comparison to the Student's t-distribution with $(n-1)$ degrees of freedom. Hence, I compare the DM statistic with the $t(n-1)$ extreme value at a five percent significance level.

I also use this DM statistic for the log scores. We just consider the other tail of the distribution since for this metric a higher value is better. For log scores, a positive DM statistic implies that the V-VAR is better whereas for MSFE it is then negative. This leads us to the results in the following Section.

4. Results

I compute the results for three different subsamples. The first subsample ranges from 1995Q1 to 2006Q4 which corresponds to the period before the financial crisis around 2007-2009. This naturally leads us to the second subsample which corresponds to this crisis, spanning 2007Q1 to 2009Q4. The final period of interest is the post-crisis period covering 2010Q1 up to and including 2014Q3. This period is interesting considering the preceding period could cause a structural change or break in the time series. Evaluating this period shows us whether the models suffer from a possible break or shift. The Section is split up in three main parts.

First, I present some preliminary results to get a slightly better understanding of the differences between the methods, before considering the out-of-sample forecasting results.

Secondly, we direct our attention to the output gap forecasting results to examine whether there are relatively best performing models given a certain filtering method with respect to the output gap. I find that individual model choice is unimportant. Additionally, I find that combining forecasts usually works well but not in all cases.

Thirdly, we consider the results of the inflation density forecasting where we focus on which cyclical component extractor is most informative for predicting inflation. I show that the choice of model there is likewise not of the utmost importance, but choice of filtering method does influence the results. The BN decompositions seem to perform slightly better in general. However, there is no clear superior method when it comes to directly forecasting inflation through all subsamples and all forecast horizons.

Finally, I link the results between the output gap and inflation results, where similarities and differences are highlighted.

4.1. Preliminary results

Table 1 displays summary statistics on our four selected filtering methods. We observe considerable differences in most metrics. The averages of the HP and Hamilton filter are both positive, while both BN decomposition averages are negative. Also, the standard deviation of the Hamilton filter is roughly ten times larger than the standard deviation that describes the univariate BN decomposition, approximately twice the size of that of the HP filter and similar to the multivariate BN decomposition in terms of variance. More or less the same principle holds for the range (maximum minus minimum) of the variables.

Most extractors produce series with noticeably high first-order autocorrelations above 0.90. The univariate BN decomposition is the odd one out, where the first-order autocorrelation is around 0.44. This is clearly high as well, but relatively low in comparison to the other methods. The eighth-order autocorrelation is negative for the HP filter, implying that for this method the output gap measure often takes an opposite direction compared to two years before. For the other methods this does not hold, where the multivariate BN decomposition even has an positive eighth-order autocorrelation of 0.3445. These statistics unequivocally show that the variables are not identical.

Table 1: Summary statistics of output gap measures

Metrics	HP filter	Hamilton filter	Univariate BN decomp.	Multivariate BN decomp.
Mean	0.0015	0.0026	-0.0010	-0.0029
Std. dev.	0.0112	0.0262	0.0027	0.0238
Min.	-0.0373	-0.0735	-0.0131	-0.0705
Max.	0.0157	0.0405	0.0046	0.0263
$\hat{\gamma}_1$	0.9031	0.9407	0.4419	0.9079
$\hat{\gamma}_8$	-0.0859	0.1429	0.0561	0.3445

Table 1: This table displays the summary statistics of the various output gap measures: Hodrick-Prescott filter, Hamilton filter, univariate and multivariate Beveridge-Nelson decomposition. $\hat{\gamma}_i$ denotes the i-th-order autocorrelation.

Figure 4.1 shows the probabilities of a negative output gap at any quarter in the forecast sample for the four filtering methods, based on one-step-ahead forecasts of the RV-VAR model. The differences between various forecasting models with respect to these probabilities are slim, hence one figure is sufficient to show the dynamics.

We can already see that the methods show various patterns. The Hamilton filter produces the most extreme probabilities in a sense, a probability can seemingly ’swing’ from zero to one in one or two quarters. When the Hamilton filter is applied to more data, we see that this effect lessens at the end of the sample and it starts to follow the BN decompositions to a larger extent.

After that we have the HP filter that is capable of the same ’swings’ but is slightly less extreme than the Hamilton filter, at least at the start of the sample. Visually, the Hamilton filter seemingly follows the HP filter in a lagged way, which is to be expected looking at the direct forecast specification that the regression in the Hamilton filter has. At the end of the sample, the HP filter is relatively certain of positive output gaps, whereas the other models show no such certainty.

Both the univariate and multivariate BN decomposition are more leveled out and seem to be centered around the 50% probability mark. In the beginning of the sample, they seem to agree more than later on in the sample, starting from the crisis around 2008. At the end they seem to converge more though. In times of financial turmoil their differences seem to be more pronounced. In the appendix Figure D.1 is included which depicts the probabilities given by the ensemble models. The probabilities follow the same patterns but seem to be spread out more. Ensemble models are more certain in forecasting negative or positive output gaps. There seem to be little differences between the various methods in ensembles compared to the single model graph. This result is not surprising since the different models produce similar probability predictions, hence the ensembles behave more or less likewise. In any case, these graphs once more show that the methods have distinct differences.

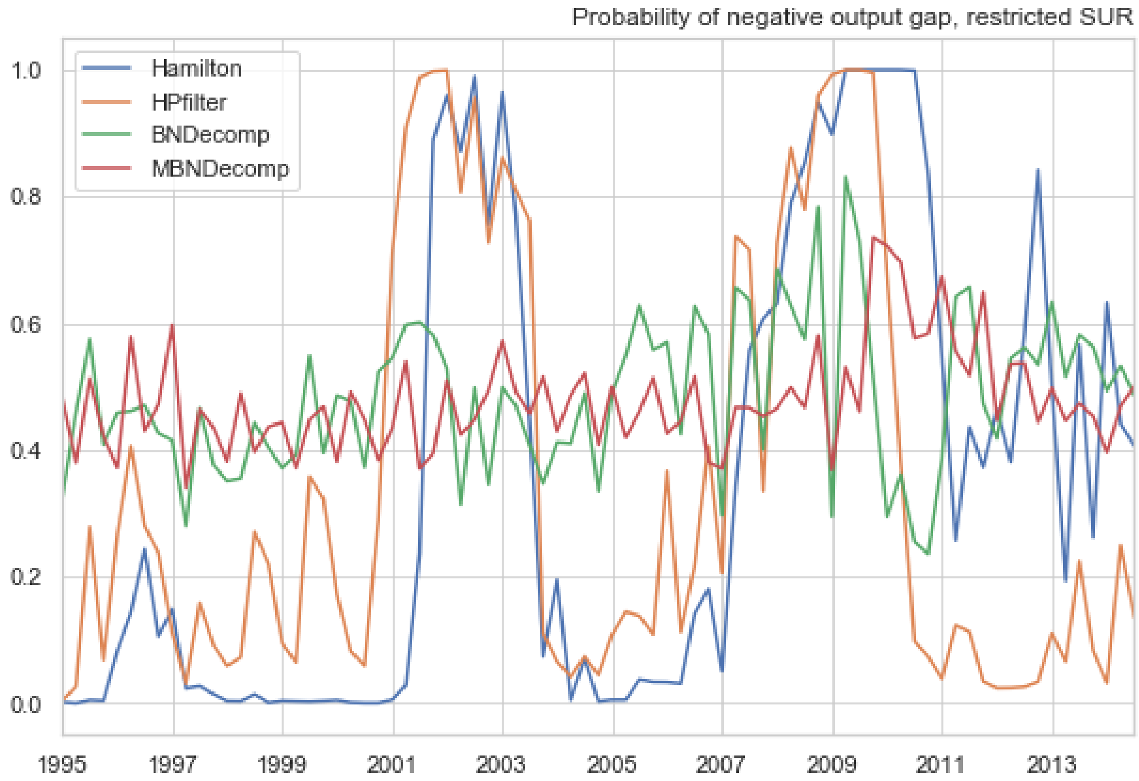

Figure 4.1: Probabilities of negative output gap over forecast sample by individual RV-VAR model

4.2. Output gap forecasting performance

Table 2 displays the results of the HP filter output gap forecasts for the three mentioned subsamples. The tables in general in this Subsection display the results over different forecast horizons rather than comparing various filtering methods in the same table. It is not possible to compare the values of the same models between filtering methods since the underlying 'truth' is unequal. To see this, consider the results of the multivariate BN Decomposition in Table 5 with for example the univariate results in Table 4. Specifically, the MSFE outcomes are immensely larger for the former and the log scores are substantially lower. However, (the scaling of) the output gap measure produced by the multivariate BN is substantially different (see for example the scaling in Figure E.4, Figure E.1, and the summary statistics). In short, now we compare the results within and over different forecasting horizons for a given filtering method. We commence with the HP filter measure.

In the first subsample, we see that the ensemble forecasts yield the best average log scores for all forecast horizons except for eight-steps-ahead. In terms of MSFE, they are not typically the best performing for any horizon but for shorter horizons it holds. This result is in line with the findings of Hall and Mitchell (2007) that combining density forecasts can boost efficacy. We also observe that the efficiency gradually decreases over the forecast horizon. The finding that performance generally worsens for larger horizons is familiar since the MSFE and variance of the previous forecast are clearly passed on to the next forecast when applying multi-step-ahead forecasting in our manner (Cheng, Tan, Gao, & Scripps, 2006).

Comparing models within their respective forecast horizon, the outcomes are noticeably close to each other for both log score and MSFE. Often the RV-VAR has the highest log score with relatively low MSFE, but it is only slightly better in most cases. The only significantly better performing forecast relative to the V-VAR is that for the two-step-ahead ensemble model in log score.

For the crisis subsample we observe some similar and some deviant results in comparison to the first period. The individual models seem to perform relatively similarly, but just on a generally lower level. Here, we see a clearly well-performing individual model for the HP filter measure in a crisis situation for all forecast horizons: the V-VAR. Its log score is always superior and its MSFE is never outperformed by any other individual model. There is only one significant MSFE improvement over the V-VAR by the ensemble of the two-step-ahead forecasts.

The most noticeable change is the performance of the ensemble models, which are (by far) the worst for all forecast horizons in terms of log score. If we consider fan charts of the period around the crisis of various forecasts, we can see why this phenomenon occurs. First, consider Figure E.3 and Figure E.5. Note that the fan charts are in the Appendix to prevent crowding out the main text too much. We observe that the one-step-ahead forecast at 2008Q4 is outside the 90% bounds for the former and within the same bounds for the latter. We observe that combining forecasts leads to more dense distributions around the mean relative to the best individual model, the V-VAR. This makes the resulting density sensitive to outliers, producing the exorbitantly low log scores that we see in Table 2. This result also complies to the findings of Hall and Mitchell (2007), who state that combining forecasts can help, but not always. In terms of MSFE the ensemble forecasts are not noticeably different from the individual models which we can also see in the mentioned figures. The position of the one-step-ahead forecast is similar for ensembles and the respective individual model. A reason why the average log scores of ensembles of different forecasting horizons behave erratically is partly outlier-based. Note that the sample size of this period is small and for some quarters, there can be a log score of around -80 where for another forecasting horizon it can be ’only’ -20. This makes an enormous difference for the average outcome. It is an important driving force behind the differences among horizons. Those differences can happen under certain circumstances. One example would be the following: the quarter just before a crisis quarter could be relatively good (positive output gap) which could imply a new positive output gap in the next quarter, only to find out that it would become remarkably negative. However, a forecast from further steps back can be better suited to forecast a crisis period if the period leading up to the production of the forecast was, for instance, a (small) recession. This can explain the substantial differences to some extent.

Although the ‘real’ output gap measure is further away from the bounds for larger h (higher MSFE), on average we regularly have an improved log score. The reason why log scores seem to improve over longer forecast horizons during this period, I dub the ’fanning’ effect. If we consider Figure E.3, we see that the bounds become more wide or as the name of the plot suggests, it generally fans out over longer horizons because the uncertainty increases. This means that the density becomes less susceptible for outliers since the density is more flat and hence it is not punished as harshly in the log score metric.

The third subsample seems similar to the first period. Now, the ensembles produce significantly high quality forecasts in terms of log score for shorter horizons, with relatively low MSFE, although not usually the lowest. For longer forecast horizons the performance declines. The crisis does not seem to affect the predictions too much, implying that the model has not become misspecificied because of the previous aberrant period. In this period, there does not seem to be a clear winner when it comes to individual model choice. Although the V-VAR model has relatively low MSFE over larger forecast horizons, the log score is never the highest. We see that most models are not far apart when it comes to their metrics in any subsample or for any-step-ahead forecast horizons. This strengthens the idea that individual model choice is not the driving force for accurate output gap predictions. This complies to an extent with the results of Clements and Galvão (2012), who also find (in a frequentist setting) that the V-VAR model is not considerably outperformed by the SUR models. This implies that the revisions in the data are predictable using vintages of past data.

Table 2: HP filter output gap forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE*	Average log score	MSFE*	Average log score	MSFE*
One-step-ahead						
V-VAR	3.914	0.017	**3.512**	**0.053**	4.050	0.010
SV-VAR	3.921	0.017	3.324	0.065	4.054	0.010
SBV-VAR	3.918	0.017	3.370	0.064	4.055	0.010
RV-VAR	3.922	0.017	3.414	0.062	4.051	0.009
Ensemble	**4.010**	**0.017**	-14.734	0.060	**4.358**†	**0.009**
Two-step-ahead						
V-VAR	3.475	0.034	**2.872**	0.173	3.559	0.022
SV-VAR	3.504	0.037	2.653	0.203	3.597	0.023
SBV-VAR	3.501	0.038	2.628	0.202	3.599	0.025
RV-VAR	3.497	0.035	2.707	0.196	3.643†	**0.016**
Ensemble	**3.840**†	**0.027**	-4.294	**0.115**†	**3.824**†	0.028
Four-step-ahead						
V-VAR	3.118	**0.062**	**2.693**	**0.254**	3.161	**0.049**
SV-VAR	3.150	0.086	2.450	0.303	3.226	0.067
SBV-VAR	3.155	0.085	2.415	0.302	**3.263**	0.060
RV-VAR	3.160	0.065	2.561	0.286	3.248	0.054
Ensemble	**3.196**	0.082	-8.028	0.302	3.137	0.085
Eight-step-ahead						
V-VAR	2.867	0.113	**2.647**	0.289	3.004	**0.061**
SV-VAR	2.855	0.169	2.585	0.312	2.832	0.193
SBV-VAR	2.862	0.165	2.607	**0.302**	2.960	0.136
RV-VAR	**2.958**	**0.090**	2.543	0.340	**3.020**	0.106
Ensemble	2.709	0.182	-14.786	0.620	2.659	0.154

Table 2: This table displays the results of the output gap density forecasting with the HP filter output gap measure. Note: the bold numbers represent the most favourable result per metric for a given filter. *MSFE is multiplied by 1000. A dagger † implies significant (95% level) improvement over V-VAR model.

Table 3 shows the results of the Hamilton filter forecasts of the output gap. Considering the first subsample, we observe many similarities compared to the HP filter results, typically on a slightly lower level in an absolute sense.

Similar to the results of Table 2, the ensemble models outperform the individual models for shorter horizons. For larger horizons, we once more find that individual models do better. The V-VAR and RV-VAR seem to better suited to forecast the output gap than the seasonality models in this period. Again, there are few significant improvements over the V-VAR model in this period.

The crisis results are also similar since they again show that the ensembles in most cases have exorbitantly poor performance in looking at the log score. Considering Figures E.2 & E.6, we see that the ensemble density is more centered around the mean whereas the SBV-VAR model is more spread out, which means its log score is not penalised as heavy as the ensemble. We recognise the 'fanning' effect in the Figure of the ensemble, which is an explanation for the improvement in log scores over a longer forecast horizon. Also, we have the erratic behaviour which is partly caused by outlier sensitivity. The best individual model per horizon has also changed. For one-step-ahead forecasts the RV-VAR is optimal and for higher horizons it is typically the SBV-VAR in terms of log score. Although the ensembles can have relatively low MSFE in this subsample, the egregiously poor log scores suggests that their efficiency is remarkably low. Most models are again noticeably similar in terms of their criteria, with little significant improvements.

The final period results also resemble the results from Table 2. However, ensembles generally do not dominate the individual models. For two-step-ahead forecasting, it is the worst. Also for $h = 4$, the log score is even lower than in the crisis sample, which upon investigation is caused by a heavily penalised outlier which influences a mean of a series substantially. Other than that the results are as expected: slightly lower log scores and slightly higher MSFE over longer forecast horizons. The RV-VAR does remarkably well over longer forecasting periods, hence the RV-VAR is a decent forecasting model for both the Hamilton as the HP filter. It is not clearly the best, but it is the individual model that most frequently outperforms others. Still, the results in most subsamples imply that the models are similar and significance is not present for shorter horizons.

Table 3: Hamilton filter output gap forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE*	Average log score	MSFE*	Average log score	MSFE*
One-step-ahead						
V-VAR	3.393	0.049	2.567	0.228	3.459	0.036
SV-VAR	3.394	**0.048**	2.538	0.255	3.443	0.040
SBV-VAR	3.398	0.048	2.571	0.244	3.470	**0.036**
RV-VAR	3.398	0.048	**2.642**	**0.224**	3.455	0.038
Ensemble	**3.488**	0.048	-19.460	0.238	**3.675**	0.037
Two-step-ahead						
V-VAR	3.026	0.089	1.253	0.760	2.983	0.106
SV-VAR	2.950	0.138	1.029	0.887	2.812	0.205
SBV-VAR	2.923	0.150	**1.549**	0.865	2.808	0.199
RV-VAR	2.927	0.123	1.101	0.865	**3.006**	**0.079**
Ensemble	**3.329**†	**0.074**	-16.184	**0.431**†	2.438	0.225
Four-step-ahead						
V-VAR	**2.652**	**0.179**	1.262	1.693	2.467	0.356
SV-VAR	2.518	0.334	1.311	1.800	1.752	1.485
SBV-VAR	2.525	0.314	**1.372**	1.593	1.907	1.263
RV-VAR	2.538	0.272	1.197	1.891	**2.634**	**0.139**†
Ensemble	2.598	0.262	-5.168	**0.880**†	-8.797	1.343
Eight-step-ahead						
V-VAR	2.253	**0.272**	1.465	2.297	1.724	1.615
SV-VAR	1.829	1.516	1.501	2.105	0.714	4.629
SBV-VAR	1.799	1.609	**1.592**	**1.979**†	1.323	3.223
RV-VAR	**2.306**	0.413	1.340	2.170	**2.225**†	**0.587**†
Ensemble	1.649	1.057	-16.359	4.504	1.696	1.429

Table 3: This table displays the results of the output gap density forecasting with the Hamilton filter output gap measure. Note: the bold numbers represent the most favourable result per metric for a given filter. *MSFE is multiplied by 1000. A dagger † implies significant (95% level) improvement over V-VAR model.

Table 4 displays the univariate BN decomposition output gap forecasting results. The results in the first subsample are slightly different to what we have seen so far. The ensemble models are significantly superior in terms of log score for any horizon and have close-to-optimal MSFEs. Hence, for longer horizons, combining density forecasts is a good idea for the univariate BN decomposition but not as much for the other two discussed methods. Relative to these other methods, the BN decomposition output gap measure yields the highest log scores and lowest MSFEs but as discussed this is likely scale related and we cannot conclude that this is an indication that the univariate BN decomposition is relatively superior.

The best individual model in the first subsample is rather ambiguous. The V-VAR performs well in terms of log score but never has the lowest MSFE. Other than that, there are no standout results to observe other than the fact that the models once again produce similar results for all subsamples.

In the second subsample we do not observe surprising results. We observe the fanning effect again in Figure E.1 and see the same differences in density width between the ensemble and the best individual model in the second sample looking at Figures E.1 & E.7, explaining why log scores seem to improve for ensembles over longer forecast horizons and why the individual models do substantially better in a crisis regime.

In the second period, the RV-VAR is the best choice for longer forecast horizons, whereas the SBV-VAR is the best for one-step-ahead forecasting in terms of log score. The V-VAR model often has the lowest MSFE, but differences are slim.

In the final period, we also see that ensembles significantly dominate in terms of log score with decent MSFE performance relatively for any h, which differs from the other two methods we have seen. In Tables 2 & 3 combining forecasts in this period is only helpful in the short-horizon forecasts. It confirms the notion from Hall and Mitchell (2007) that for different variables and different forecast horizons, combining forecasts can boost forecasting performance, but not at all times. The RV-VAR has the best MSFE results in this final period, often significantly or close to being significant.

Table 4: Univariate BN decomposition output gap forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE*	Average log score	MSFE*	Average log score	MSFE*
One-step-ahead						
V-VAR	4.269	0.005	3.830	0.028	4.307	0.007
SV-VAR	4.282†	**0.005**	3.815	**0.028**	4.342†	**0.005**†
SBV-VAR	4.277	0.005	**3.843**	0.030	4.320	0.006†
RV-VAR	4.275	0.005	3.773	0.031	4.300	0.006†
Ensemble	**4.662**†	0.005	-7.787	0.029	**4.600**†	0.006
Two-step-ahead						
V-VAR	4.195	0.005	3.800	**0.030**	4.155	0.010
SV-VAR	4.188	0.005	3.757	0.031	4.144	0.010
SBV-VAR	4.194	0.005	3.757	0.031	4.122	0.012
RV-VAR	4.180	0.005	**3.828**	0.030	4.253	**0.005**
Ensemble	**4.656**†	**0.005**	-9.657	0.040	**4.454**†	0.008
Four-step-ahead						
V-VAR	4.122	0.005	3.871	0.025	4.121	0.010
SV-VAR	4.086	0.007	3.838	0.027	3.981	0.018
SBV-VAR	4.093	0.007	3.852	0.026	4.012	0.016
RV-VAR	4.111	**0.005**	**3.900**	**0.023**†	4.238	**0.003**†
Ensemble	**4.581**†	0.006	-2.064	0.036	**4.484**†	0.008
Eight-step-ahead						
V-VAR	4.022	0.007	3.862	**0.024**	4.151	0.005
SV-VAR	3.888	0.011	3.790	0.030	3.803	0.029
SBV-VAR	3.893	0.011	3.790	0.029	3.958	0.014
RV-VAR	4.051†	**0.005**†	**3.863**	0.025	4.209	**0.003**
Ensemble	**4.440**†	0.008	2.963	0.042	**4.650**†	0.004

Table 4: This table displays the results of the output gap density forecasting with the univariate BN decomposition. Note: the bold numbers represent the most favourable result per metric for a given filter. *MSFE is multiplied by 1000. A dagger † implies significant (95% level) improvement over V-VAR model.

Table 5 shows the multivariate BN decomposition output gap forecast results. In the first period, we observe that only for the eight-step-ahead forecasts the ensemble is unfavourable. The results seem to deteriorate for the SUR models over longer horizons, whereas the V-VAR model behaves expectedly: a gradual decline of performance over the length of the forecast horizon. In an absolute sense, the log scores and MSFEs seem slightly high relative to the other filtering methods. Although this variable has the largest variance (see summary statistics), the differences are substantial. This result might be explained by the complexity of the variable. The multivariate BN decomposition uses 18 different variables in its filtering process. The information of these variables is not included in the forecasting models, which can lead to poor estimates. The V-VAR model seems to be quite robust, perhaps because of its relative simplicity. The V-VAR is outperformed significantly for one-step-ahead forecasts by all models, which is somewhat deviant to the earlier results.

In the second period we see some surprising results compared to earlier methods. The ensembles are relatively good for this method in a crisis situation. To investigate this anomaly, consider Figures E.4 and E.8. We observe that the ensemble fan chart depicts something different than the other three methods. Looking at the scale, we see that this ensemble has relatively wide density regions, although still less wide than the SBV-SUR model forecasts. The Figures look closer to each other than for any other method I consider, although the scale of the individual model is about two times that of the ensemble. The uncertainty of this output gap measure causes the densities to become more flat which is to be expected. It may also explain why MSFEs are relatively high for this method, but log scores are quite low in an absolute sense. It is robust against outliers, yet good predictions are not exceptionally rewarded in the log score metric. The best individual model is the V-VAR for longer horizons. For one-step-ahead forecasting it is the SBV-VAR and for two steps it is the RV-VAR. Hence, there is no clear optimal model for all horizons in this period, although for longer horizons the V-VAR is clearly superior.

In the final period, we see the same patterns as in the first period. Ensembles perform efficiently for shorter horizons and for longer horizons only the V-VAR seems adequate. The ensemble is not extremely poor as well, however this is caused by the fact that the weights of the ensemble focus particularly on the V-VAR. The best individual model for the short-term is the RV-VAR.

Table 5: Multivariate BN decomposition output gap forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE*	Average log score	MSFE*	Average log score	MSFE*
One-step-ahead						
V-VAR	1.551	0.682	1.390	3.481	1.588	1.340
SV-VAR	1.599†	0.245†	1.578	1.413	1.684†	0.599†
SBV-VAR	1.599†	0.241†	**1.594**†	**1.340**	1.670†	0.599†
RV-VAR	1.602†	**0.222**†	1.585	1.421	1.700†	**0.408**†
Ensemble	**2.209**†	0.305†	1.360	1.789†	**2.154**†	0.669†
Two-step-ahead						
V-VAR	1.483	0.353	1.373	3.206	1.556	1.065
SV-VAR	1.434	0.486	1.362	3.159	1.393	2.377
SBV-VAR	1.430	0.428	1.389	2.441	1.388	2.472
RV-VAR	1.454	**0.230**	1.418	2.398	1.584	**0.176**†
Ensemble	**2.090**†	0.322	**1.529**	**2.380**†	**1.829**†	1.512
Four-step-ahead						
V-VAR	1.426	0.199	1.385	2.608	1.525	0.876
SV-VAR	1.053	0.618	1.179	1.407	1.101	1.389
SBV-VAR	1.026	0.591	1.133	2.061	1.100	1.647
RV-VAR	1.204	0.699	1.228	3.661	1.265	1.232
Ensemble	**1.822**†	**0.148**	**1.708**†	**1.400**†	**1.869**†	**0.529**†
Eight-step-ahead						
V-VAR	**1.357**	**0.192**	**1.400**	**1.685**	**1.476**	**1.019**
SV-VAR	-0.018	3.890	0.238	2.338	0.024	21.488
SBV-VAR	-0.126	13.598	0.178	2.337	0.166	15.728
RV-VAR	0.531	10.656	0.897	6.028	0.619	23.595
Ensemble	1.063	0.562	1.136	2.196	1.069	1.816

Table 5: This table displays the results of the output gap density forecasting with the multivariate BN decomposition. Note: the bold numbers represent the most favourable result per metric for a given filter. *MSFE is multiplied by 1000. A dagger † implies significant (95% level) improvement over V-VAR model.

This concludes the output gap forecasting results. In summary, we see that combining forecasts, based on their informative value on inflation forecasting, works quite well for short-horizon forecasting and less so for longer-horizon forecasting in relatively tranquil times. This adheres to the idea that a adequate output gap measure has informative value for inflation. We see that if we select models based on inflation forecasting performance, we get more reliable output gap measures. For longer horizons, we do not observe this idea for all methods. Only the BN decomposition ensemble generally does better than the individual models. However, in times of crisis we see that ensembles perform poorly in terms of log score, because they are relatively centered around the mean making them susceptible for aberrant observations. Although the MSFE is often comparable to individual models, the process predicted is inaccurate. Next to that, the models do not seem to suffer from any breaks caused by the financial crisis and hence the third period results are comparable to the first period. The performance of individual models in general is similar and there is no model which clearly dominates (most of) the results. Hence, choice of model is of less importance. However, if we were to choose a model the RV-VAR outperforms the others on most occasions followed by the V-VAR. The choice of filtering method is more important and hence we continue our research with the inflation forecasting results, where we can directly compare the performance of different filtering methods.

4.3. Inflation forecasting results

In the previous Subsection, a direct comparison between filtering methods is not possible due to the fact that different methods give different output gap measures, making log scores and MSFEs disparate. Therefore, I present the inflation forecasting results where it is in fact possible to directly compare results from different filtering methods, since the inflation outcome is the same for every model.

Table 6 displays the one-step-ahead inflation forecasting results. We see that in terms of log score and MSFE, the multivariate BN decomposition produces superior results compared to any other filtering method. A somewhat deviant result is that ensemble forecasts do not perform well relative to the individual models in terms of log score. Note that the weights of the ensembles are the same for forecasting inflation as for forecasting output gap. The MSFE is comparable in most cases. The best models now seem method-dependent and show little consistency for different filtering methods, without any significant improvements over the V-VAR. A recognisable result is that the HP filter performs relatively poorly and that the Hamilton filter works better. This complies with the results of Hamilton (2018).

The second period shows somewhat deviant results compared to the one-step-ahead output gap forecasts. For instance, the ensembles are usually not the worst except for the Hamilton filter. Looking at the Figures in Appendix Section F, we see that both the ensemble models and the individual models are quite sensitive to outliers, although the ensemble densities seem to be more narrow as we observe in the output gap results as well. We also see that the forecasts from the Hamilton filter are not suitable to forecast inflation in a crisis situation as their log scores and MSFEs are substantially lower than any other model. The best model is again not consistently the same but rather different for every filtering method.

In the third period we see similarities and changes compared to the first subsample. In an absolute sense the results of the first and the third period are seemingly close to each other. The Hamilton filter generally has the worst performance in terms of log score. However, in terms of MSFE the HP filter and multivariate BN decomposition produce inferior results. The univariate BN decomposition yields the best individual outcomes. The role of ensembles has also changed since they are a viable choice now. In fact, the worst individual models together produce the best performing ensemble, being the Hamilton filter ensemble.

Table 6: One-step-ahead model inflation forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE	Average log score	MSFE	Average log score	MSFE
HP filter						
V-VAR	-1.773	**1.828**	-14.634	11.299	-1.767	1.580
SV-VAR	-1.780	1.859	-10.609	**11.057**	**-1.755**	1.569
SBV-VAR	-1.776	1.861	**-8.929**	11.079	-1.759	**1.566**
RV-VAR	**-1.773**	1.843	-14.085	11.454	-1.782	1.697
Ensemble	-2.067	1.843	-12.136	11.149	-1.777	1.599
Hamilton filter						
V-VAR	**-1.758**	**1.662**	-19.020	13.069	-1.775	1.486
SV-VAR	-1.766	1.719	-19.012	**12.958**	-1.778	1.494
SBV-VAR	-1.769	1.727	-16.312	13.059	-1.768	1.478
RV-VAR	-1.763	1.669	**-15.671**	13.167	-1.773	1.485
Ensemble	-1.908	1.686	-41.177	12.998	**-1.671**	**1.473**
Univariate BN						
V-VAR	-1.738	1.707	**-8.810**	11.693	-1.731	1.421
SV-VAR	-1.743	1.721	-16.214	11.509	-1.734	**1.404**
SBV-VAR	-1.743	1.724	-13.131	**11.459**	-1.720	1.411
RV-VAR	**-1.735**	**1.683**	-11.764	12.010	-1.750	1.544
Ensemble	-1.987	1.705	-9.528	11.703	**-1.685**	1.439
Multivariate BN						
V-VAR	-1.708	1.474	-16.691	11.808	-1.787	1.733
SV-VAR	**-1.706**	1.477	-12.842	11.557	**-1.786**	**1.726**
SBV-VAR	-1.706	**1.457**	**-5.677**	11.459	-1.793	1.766
RV-VAR	-1.718	1.511	-14.754	**11.279**†	-1.793	1.778
Ensemble	-1.803	1.475	-5.769	11.408	-1.855	1.742

Table 6: This table displays the results of the one-step-ahead inflation density forecasting from each model for every type of filtering method. Note: the bold numbers represent the most favourable result per metric for a given filter. A dagger † implies significant (95% level) improvement over V-VAR model.

With regard to the two-step-ahead and four-step-ahead forecasts, the results are almost identical to the results of Table 7. The Tables 8 & 9, depicting these results, can be found in the appendix. Three important notes from these results are firstly that the individual models of the univariate BN decomposition improve in the first and last period. Secondly, some of the individual models' performance from the multivariate BN decomposition start to deteriorate, although their ensemble in the first sample is the best by far. Finally, we see that the 'fanning' effect of the ensembles over longer horizons does not lead to an increase in log scores, which is different from the output gap results. Considering the fan charts in Appendix Section F we see that the densities do indeed fan out slightly less than for the output gap forecasts.

This leads us to the final table, being the two-year-ahead inflation forecast results. For policymakers, this table is quite important since their target is typically the inflation in two years. In the first sample, we see that, somewhat surprisingly, the ensemble of the Hamilton filter yields the best result in terms of log score. The ensemble of the multivariate BN decomposition yields the lowest MSFE. In times of crisis, the best model is given by the V-VAR of the multivariate BN decomposition, with the lowest MSFE coming from the RV-VAR of the same method. In the third period, the best model performance comes from the ensemble of the HP filter, with the lowest MSFE coming from the univariate BN decomposition ensemble. There are slightly more significant improvements over the V-VAR in this table, particularly for ensemble metrics, but not as much for individual models.

The results from the eight-step-ahead inflation forecasts are indecisive to a high degree. We typically see that ensembles often yield the best log score values, but typically not the lowest MSFE. Again, this follows the spirit of Hall and Mitchell (2007). In some ways combining forecasts is useful, but not in every aspect and certainly not in crisis situations. From Table 7, it is unclear which of the four methods is best suited to forecast inflation in two years since it varies notably among different periods. Although, there is some credence that the BN decompositions perform relatively well. This concludes the results of the inflation forecasting.

Table 7: Eight-step-ahead model inflation forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE	Average log score	MSFE	Average log score	MSFE
HP filter						
V-VAR	-2.442	7.097	**-2.501**	9.890	-2.376	5.965
SV-VAR	-2.496	7.273	-2.566	10.567	-2.540	7.207
SBV-VAR	-2.488	7.206	-2.531	10.747	-2.461	6.371
RV-VAR	**-2.388**	**5.442**†	-2.487	**9.381**	-2.328	4.393†
Ensemble	-2.413	5.611	-14.528	9.612	**-2.206**†	**4.195**
Hamilton filter						
V-VAR	-2.431	6.893	-2.521	9.873	-2.462	7.379
SV-VAR	-2.511	7.854	-2.509	9.662	-2.794	16.78
SBV-VAR	-2.534	8.382	-2.559	10.113	-2.686	13.778
RV-VAR	-2.351	**3.158**†	**-2.463**	**8.880**	**-2.417**	**6.112**
Ensemble	**-2.241**†	4.571†	-13.283	10.066	-2.774	8.348
Univariate BN						
V-VAR	-2.408	6.315	**-2.507**	9.837	-2.347	5.234
SV-VAR	-2.446	6.380	-2.520	10.451	-2.512	7.535
SBV-VAR	-2.479	7.128	-2.526	10.439	-2.440	6.565
RV-VAR	–2.367†	5.272†	-2.491	9.351	-2.329	5.107
Ensemble	**-2.366**†	**5.231**†	-11.652	**9.287**	**-2.234**†	**4.179**†
Multivariate BN						
V-VAR	**-2.322**	4.343	**-2.434**	8.302	**-2.371**	**5.910**
SV-VAR	-3.536	6.014	-3.306	7.397	-3.538	35.838
SBV-VAR	-3.587	9.192	-3.307	6.521	-3.346	15.312
RV-VAR	-2.937	6.283	-2.732	**6.419**	-2.970	35.06
Ensemble	-2.462	**2.831**†	-2.670	10.887	-2.614	9.381

Table 7: This table displays the results of the eight-step-ahead inflation density forecasting from each model for every type of filtering method. Note: the bold numbers represent the most favourable result per metric for a given filter. A dagger † implies significant (95% level) improvement over V-VAR model.

4.4. Comparison between output gap and inflation results

Considering the results from both variables, we do not directly see a contender for absolute best model to use. The V-VAR model is occasionally outperformed by another model, nevertheless there is no real consistent superior performance by any of the SUR models in both forecasting schemes. The RV-VAR outperforms the V-VAR most often, implying that it can be a slightly better specification. This complies with the results of Clements and Galvão (2012) since they also find that a V-VAR type model is adequate, but there can be improvements at times.

When it comes to choosing the best filtering method, the inflation forecasting results show some support for the superiority of the BN decompositions. These methods seem to have the best performing models most often. For the output gap, this comparison is as explained not suited to answer this research question. However, there does not seem to be ample evidence that these methods are always preferred over the Hamilton and HP filter. This follows the results from many researchers, for instance Orphanides and Norden (2002) and Garratt et al. (2008).

Combining forecasts works better for output gap forecasting than for inflation forecasting. In most cases, the ensemble models outperform the benchmark V-VAR model in terms of log score for the output gap. In terms of MSFE this is not typically the case. For inflation, forecast combinations improve over the forecasting horizon. For both variables, the ensembles perform poorly in crisis situations. These results fall in line with the findings of Hall and Mitchell (2007) that forecast combinations can help but not necessarily do.

5. Conclusion

In this thesis, I compare the forecasting ability of sets of various models where the variable of interest, the true output gap, is measured by a range of different cyclical component extractors. I express the forecasts as densities, since uncertainty is a known problem in the output gap forecasting field (Garratt et al., 2014; Orphanides & Norden, 2002). I forecast the output gap itself as well as inflation, because an adequate output gap measure should hold valuable information for forecasting inflation as well. Following on this idea, I construct ensemble forecasts where the weights are determined by the relative (historical) inflation forecasting performance to see if this improves output gap measures and inflation predictions. I perform all forecasting for multiple horizons, because policymakers are typically interested in periods further away from the present than one quarter ahead, such as two years ahead.

I consider four model types in the analysis: a VAR model which explicitly models data revisions (V-VAR), a SUR model which accounts for seasonality (SV-VAR), which amounts to correcting for annual revisions in (almost always) the third quarter of any year, a SUR model which accounts for seasonality and also incorporates benchmark revisions (SBV-VAR), and finally a restricted SUR model that restrict the predictability of data revisions (RV-VAR). The results imply that for both output gap and inflation forecasting, there is little consistency in which model has optimal prediction performance. Their performance is typically similar for all forecasting horizons, lending credence to the notion of Clements and Galvão (2013) that data revisions are predictable and hence the V-VAR is generally not outperformed by more complex models. If a best individual model has to be chosen, I would opt for the RV-VAR followed by the V-VAR.

I also test the idea that output gap measures that are informative for inflation, can produce density combinations that improve the performance of the prediction. We observe that combining forecasts works effectively for output gap forecasting in tranquil times, particularly for smaller forecast horizons. However, in crisis situations the performance declines drastically in terms of log score, caused by vastly inaccurate process predictions. For longer forecast horizons, the benefits of combining forecast diminish. The efficacy of combining inflation forecasts is subsample-dependent for smaller forecast horizons. We see that in crisis situations the performance also deteriorates for this variable. The first period is dominated by individual models for one-step-ahead forecasts, whereas the best forecasts in the final period come from ensembles. As the forecast horizon increases, ensembles become relatively better options in the first period and remain of similar quality in the final period. However, there is no ensemble belonging to a particular filtering method that consistently outperforms the others, which corresponds to the results of Hall and Mitchell (2007).

With regard to the filtering methods, I employ four variants: the univariate and the multivariate BN decomposition, the HP filter, and the Hamilton filter. If we consider the inflation forecasting results, we see some but not extensive support for the hypothesis that the BN decompositions are optimal for filtering. However, the best method varies among subsamples and among forecast horizons. Even within subsamples, it can be ambiguous what the best method is since the optimal log score does not typically correspond to the lowest MSFE.

6. Discussion

As with most if not all research, this research is not without its limitations, which lead to new directions conducive for further research. First of all, more model types can be considered. All models I use in this thesis are (variations of) the V-VAR model, which assume that the input is an observed variable. To be more complete in my analysis state space models could have been included to emphasise the unobserved nature of the variable of interest.

The same idea applies to the number of filtering methods. Unobserved component (UC) models are quite typical in the literature, which also accentuate the unobserved state of the variable. However, Morley et al. (2003) show that under relaxation of a certain restriction the

BN decomposition and UC models yield identical results, hence it is excluded in this thesis.

To keep computations manageable, I set all lags of VAR and ARIMA models for both forecasting and filtering to one. For further research, lags may be selected based on some out-of-sample forecasting criterion to minimise the risk of model misspecification.

The computation of the log score requires pdf values of the forecasted density. In this thesis, I use Gaussian kernels and a specific rule of thumb to obtain a certain density. Although the choice of Gaussian kernels is quite common in the literature (Sheather & Jones, 1991), the bandwidth selection is of most importance. By using a rule of thumb, we do not use a particularly elaborate method of determining this parameter. In further research, a grid search could be applied to find possibly better bandwidth parameters. However, since this thesis focuses on comparison, the results remain valid.

References

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of applied Econometrics*, *25*(1), 71–92.

Beveridge, S., & Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary economics*, *7*(2), 151–174.

Cheng, H., Tan, P.-N., Gao, J., & Scripps, J. (2006). Multistep-ahead time series prediction. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 765–774).

Clements, M. P., & Galvão, A. B. (2012). Improving real-time estimates of output and inflation gaps with multiple-vintage models. *Journal of Business & Economic Statistics*, *30*(4), 554–562.

Clements, M. P., & Galvão, A. B. (2013). Forecasting with vector autoregressive models of data vintages: Us output growth and inflation. *International Journal of Forecasting*, *29*(4), 698–714.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International economic review*, 863–883.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3).

Garratt, A., Lee, K., Mise, E., & Shields, K. (2008). Real-time representations of the output gap. *The Review of Economics and Statistics*, *90*(4), 792–804.

Garratt, A., Mitchell, J., & Vahey, S. P. (2014). Measuring output gap nowcast uncertainty. *International Journal of Forecasting*, *30*(2), 268–279.

Hall, & Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, *23*(1), 1–13.

Hall, P., Sheather, S. J., Jones, M., & Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, *78*(2), 263–269.

Hamilton, J. D. (2018). Why you should never use the hodrick-prescott filter. *Review of Economics and Statistics*, *100*(5), 831–843.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, *13*(2), 281–291.

Hodrick, R. J., & Prescott, E. C. (1997). Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, 1–16.

Kamber, G., Morley, J., & Wong, B. (2018). Intuitive and reliable estimates of the output gap from a beveridge-nelson filter. *Review of Economics and Statistics*, *100*(3), 550–566.

Karlsson, S. (2013). Forecasting with bayesian vector autoregression. In *Handbook of economic forecasting* (Vol. 2, pp. 791–897). Elsevier.

Morley, J. (2002). A state–space approach to calculating the beveridge–nelson decomposition. *Economics Letters*, *75*(1), 123–127.

Morley, J., Nelson, C., & Zivot, E. (2003). Why are the beveridge-nelson and unobserved-components decompositions of gdp so different? *Review of Economics and Statistics*, *85*(2), 235–243.

Morley, J., & Wong, B. (2020). Estimating and accounting for the output gap with large bayesian vector autoregressions. *Journal of Applied Econometrics*, *35*(1), 1–18.

Orphanides, A., & Norden, S. v. (2002). The unreliability of output-gap estimates in real time. *Review of economics and statistics*, *84*(4), 569–583.

Orphanides, A., & Van Norden, S. (2005). The reliability of inflation forecasts based on output gap estimates in real time. *Journal of Money, Credit and Banking*, 583–601.

Percy, D. F. (1992). Prediction for seemingly unrelated regressions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(1), 243–252.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690.

Timmermann, A. (2000). Density forecasting in economics and finance. *Journal of Forecasting*, *19*(4), 231.

Watson, M. W. (2007). How accurate are real-time estimates of output trends and gaps? *FRB Richmond Economic Quarterly*, *93*(2), 143–161.

A. Codes

resultscode.py
Code which imports files with all predictive density draws from every model and produces the results in the tables, figures, and fan charts.
gibbs.py
Code that produces all output gap measures and density forecasts.
dm_test.py
Code originally by John Tsang that I modified to compute DM statistics for my metrics.
Only_Benchmark_Gap.m
Code to obtain multivariate BN decomposition measure, code originally from Morley and Wong (2020) and modified by me to give proper results.
MorleyWongPackage.zip
Package by Morley and Wong (2020) for computing multivariate BN decomposition. I modified some things to get the regular multivariate BN decomposition without Bayesian shrinkage.

B. List of variables in multivariate BN

1. Real GDP
2. Oil price
3. Personal income
4. Unemployment rate
5. Hourly earnings
6. Fed funds rate

7. Stock prices
8. Slope of the yield curve
9. Employment
10. Industrial production
11. Capacity utilisation
12. Housing starts
13. Hours worked
14. Nonfarm Real Output per hour
15. Total reserves
16. Non-borrowed reserved
17. Real M1
18. Real M2

C. Inflation forecasting results

Table 8: Two-step-ahead model inflation forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE	Average log score	MSFE	Average log score	MSFE
HP filter						
V-VAR	-2.007	2.803	-4.096	13.450	-1.959	**1.981**
SV-VAR	-2.017	2.896	**-3.160**	11.751	-1.999	2.257
SBV-VAR	-2.030	3.005	-3.455	11.757	-1.990	2.167
RV-VAR	**-1.976**	**2.392**†	-4.370	13.494	-1.977	2.101
Ensemble	-2.021	2.483	-47.670	**9.745**†	**-1.855**	2.133
Hamilton filter						
V-VAR	-1.995	2.565	-4.723	14.553	-2.008	2.152
SV-VAR	-2.011	2.671	-3.702	14.878	-2.044	2.488
SBV-VAR	-2.017	2.672	-5.310	14.976	-2.032	2.300
RV-VAR	-1.977	**2.089**†	**-3.415**	14.866	-2.003	2.117
Ensemble	**-1.918**†	2.239	-5.133	**11.403**†	**-1.824**	**2.062**
Univariate BN						
V-VAR	-1.971	2.491	-3.129	13.735	-1.926	**1.592**
SV-VAR	-1.968	2.430	-3.662	12.434	-1.952	1.735
SBV-VAR	-1.984	2.581	**-3.000**	12.326	-1.942	1.679
RV-VAR	-1.950	2.217	-3.963	14.044	-1.953	1.919
Ensemble	**-1.913**†	**2.172**	-13.539	**9.846**†	**-1.722**†	1.765
Multivariate BN						
V-VAR	-1.911	1.847	-3.534	13.457	-2.041	2.803
SV-VAR	-1.974	2.145	-3.127	14.409	-2.103	3.308
SBV-VAR	-1.969	2.071	**-3.019**	13.698	-2.097	3.302
RV-VAR	-1.985	2.086	-3.264	12.567	-2.110	3.568
Ensemble	**-1.716**†	**1.724**	-24.159	**10.338**	**-1.966**	**2.522**†

Table 8: This table displays the results of the two-step-ahead inflation density forecasting from each model for every type of filtering method. Note: the bold numbers represent the most favourable result per metric for a given filter. A dagger † implies significant (95% level) improvement over V-VAR model.

Table 9: Four-step-ahead model inflation forecasting performance

	1995Q1-2006Q4		2007Q1-2009Q4		2010Q1-2014Q3	
Models	Average log score	MSFE	Average log score	MSFE	Average log score	MSFE
HP filter						
V-VAR	-2.237	4.344	-2.755	12.529	-2.182	3.361
SV-VAR	-2.265	4.742	-2.750	13.054	-2.324	5.621
SBV-VAR	-2.277	5.084	-2.674	11.053	-2.248	4.477
RV-VAR	-2.195	3.530	**-2.667**	11.512	-2.161	**2.918**
Ensemble	**-2.128**†	**3.420**†	-26.840	**8.874**†	**-2.142**	3.531
Hamilton filter						
V-VAR	-2.213	4.088	-2.699	12.174	-2.222	3.617
SV-VAR	-2.302	5.232	-2.774	12.008	-2.313	4.843
SBV-VAR	-2.305	5.345	**-2.689**	11.870	-2.289	4.409
RV-VAR	-2.175	**2.505**	-2.784	12.198	-2.243	3.925
Ensemble	**-2.101**†	3.341†	-26.089	**9.535**†	**-2.072**†	**3.339**†
Univariate BN						
V-VAR	-2.185	3.738	-2.799	12.343	-2.104	**2.357**
SV-VAR	-2.186	3.705	-2.707	11.861	-2.185	3.081
SBV-VAR	-2.219	4.154	**-2.646**†	10.867	-2.154	2.824
RV-VAR	-2.166	3.232	-2.695	12.004	-2.135	2.903
Ensemble	**-2.040**†	**3.024**†	-18.172	**8.917**†	**-1.862**†	2.361
Multivariate BN						
V-VAR	-2.127	2.645	-2.615	11.470	**-2.197**	**3.698**
SV-VAR	-2.498	3.132	-2.616	10.341	-2.618	5.611
SBV-VAR	-2.469	2.976	-2.606	10.306	-2.565	5.569
RV-VAR	-2.315	2.694	**-2.533**	10.402	-2.424	6.976
Ensemble	**-1.900**†	**2.281**†	-7.951	**9.403**†	-2.239	4.771

Table 9: This table displays the results of the four-step-ahead inflation density forecasting from each model for every type of filtering method. Note: the bold numbers represent the most favourable result per metric for a given filter. A dagger † implies significant (95% level) improvement over V-VAR model.

D. Probabilities of negative output gap ensembles

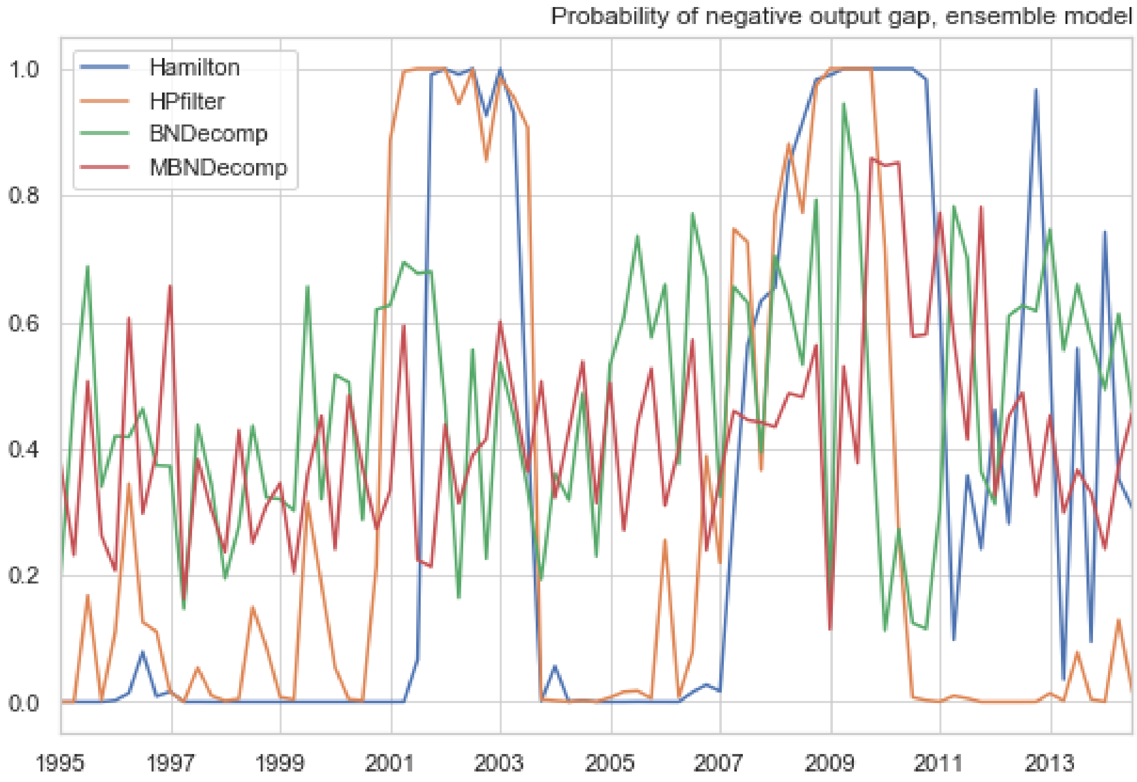

Figure D.1: Probabilities of negative output gap over forecast sample by ensemble model

E. Output gap fan charts

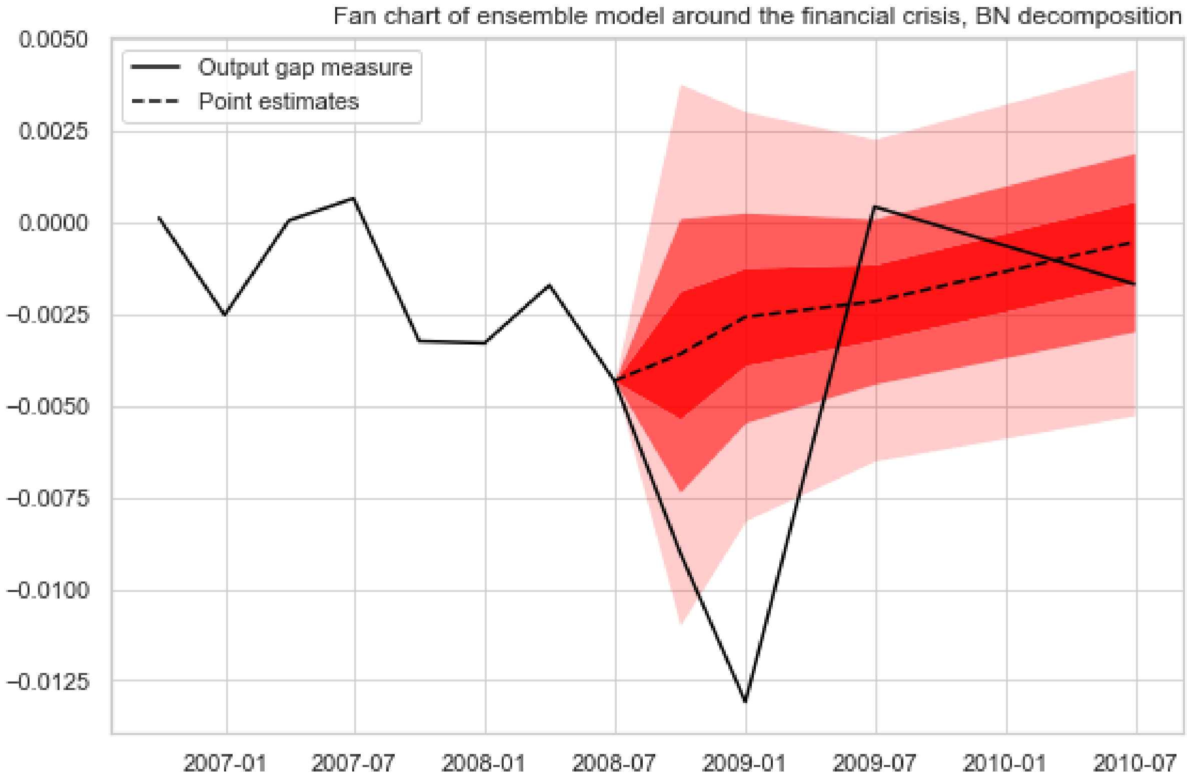

Figure E.1: Fan chart of ensemble model around the financial crisis, BN Decomposition

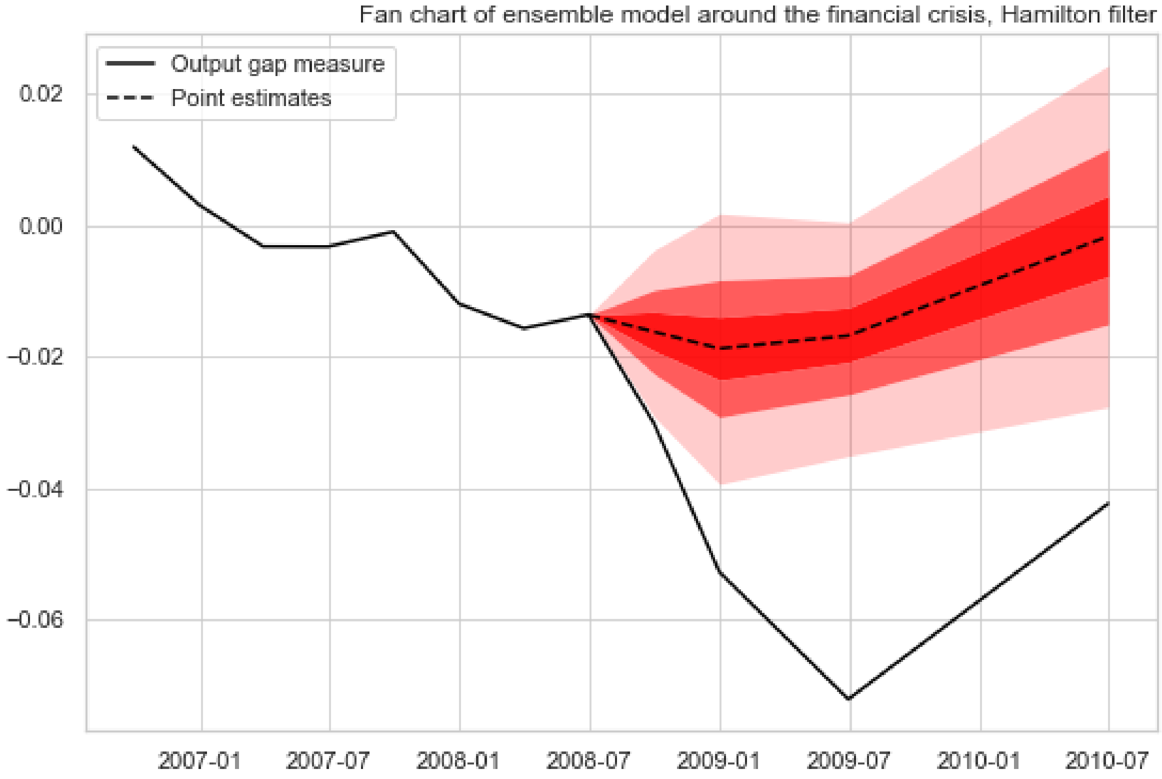

Figure E.2: Fan chart of ensemble model around the financial crisis, Hamilton filter

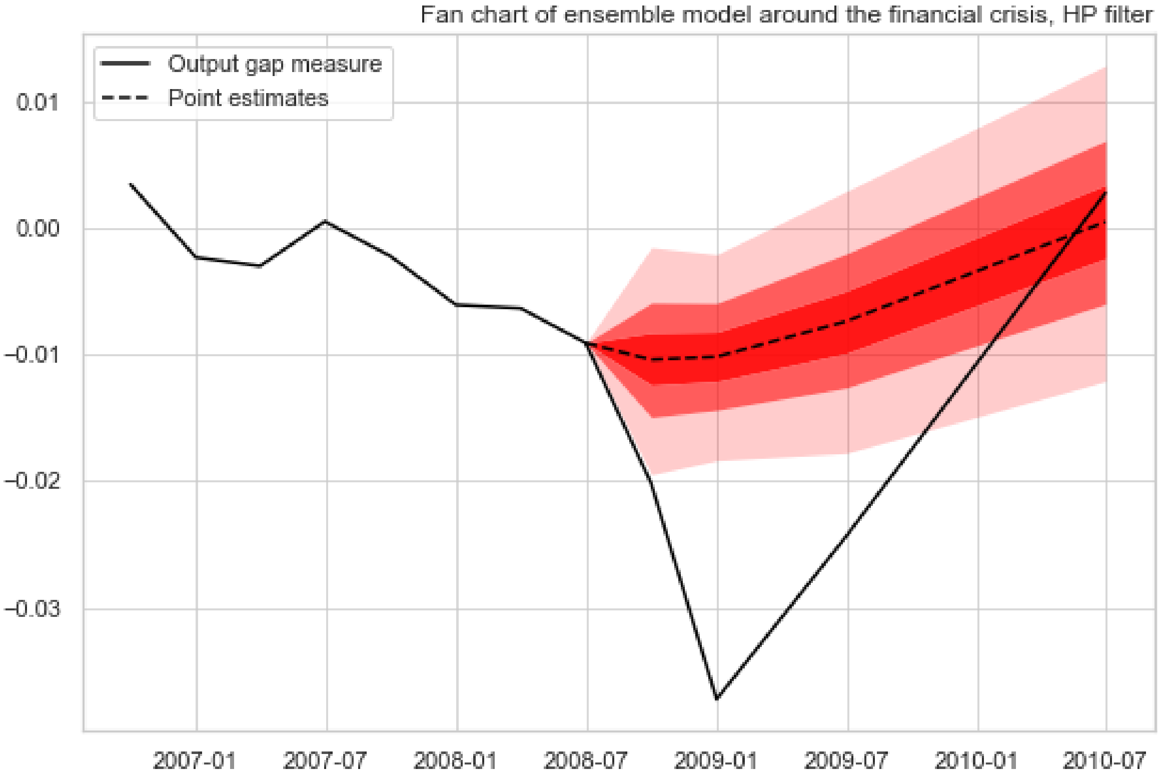

Figure E.3: Fan chart of ensemble model around the financial crisis, HP filter

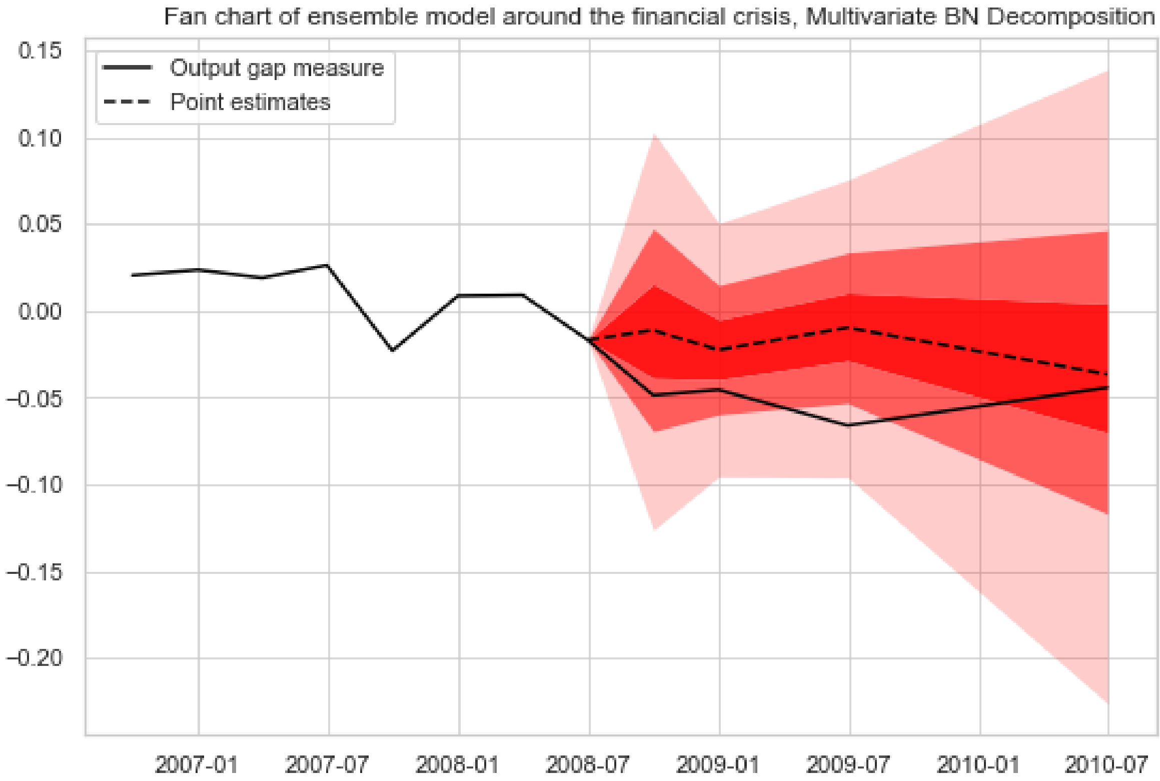

Figure E.4: Fan chart of ensemble model around the financial crisis, multivariate BN Decomposition

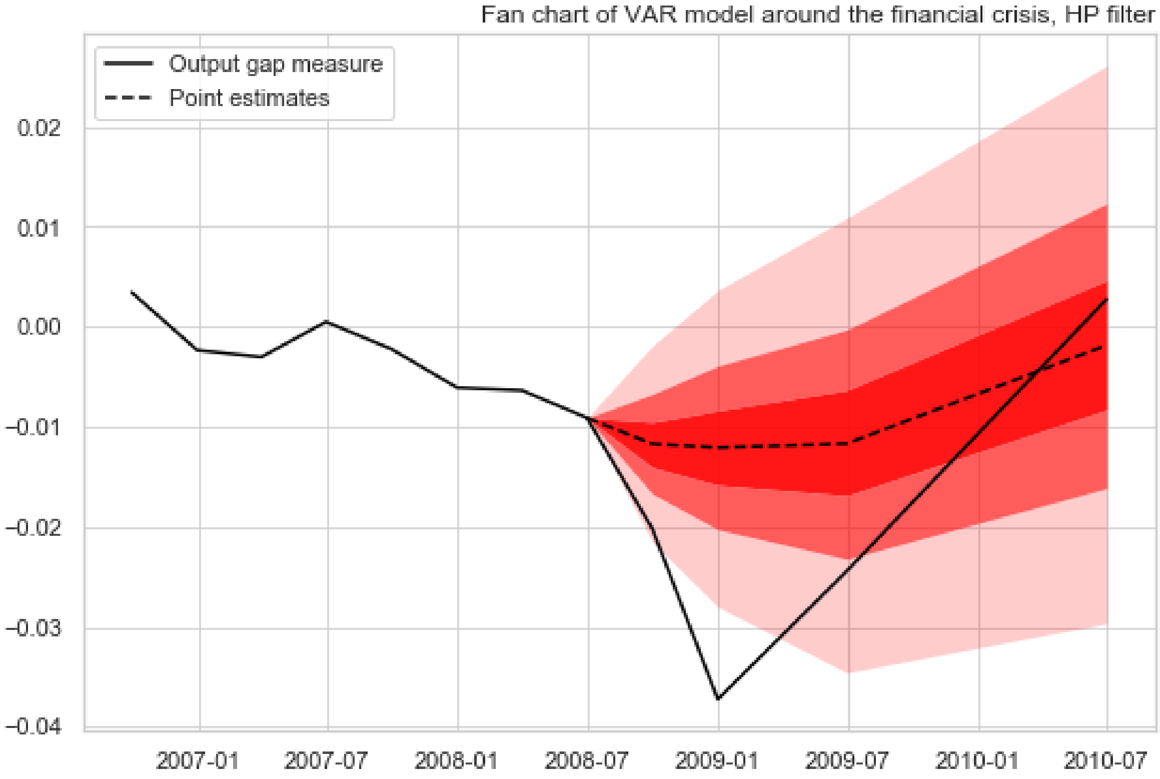

Figure E.5: Fan chart of VAR model around the financial crisis, HP filter

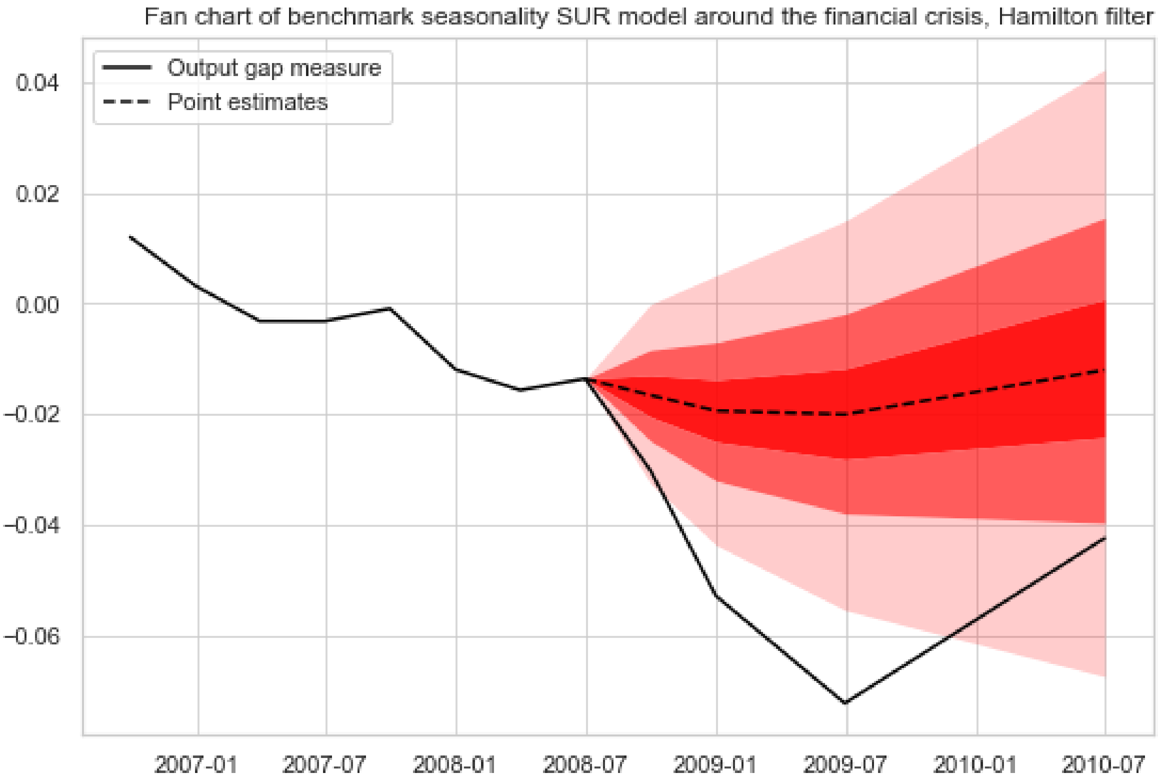

Figure E.6: Fan chart of the benchmark seasonality model around the financial crisis, Hamilton filter

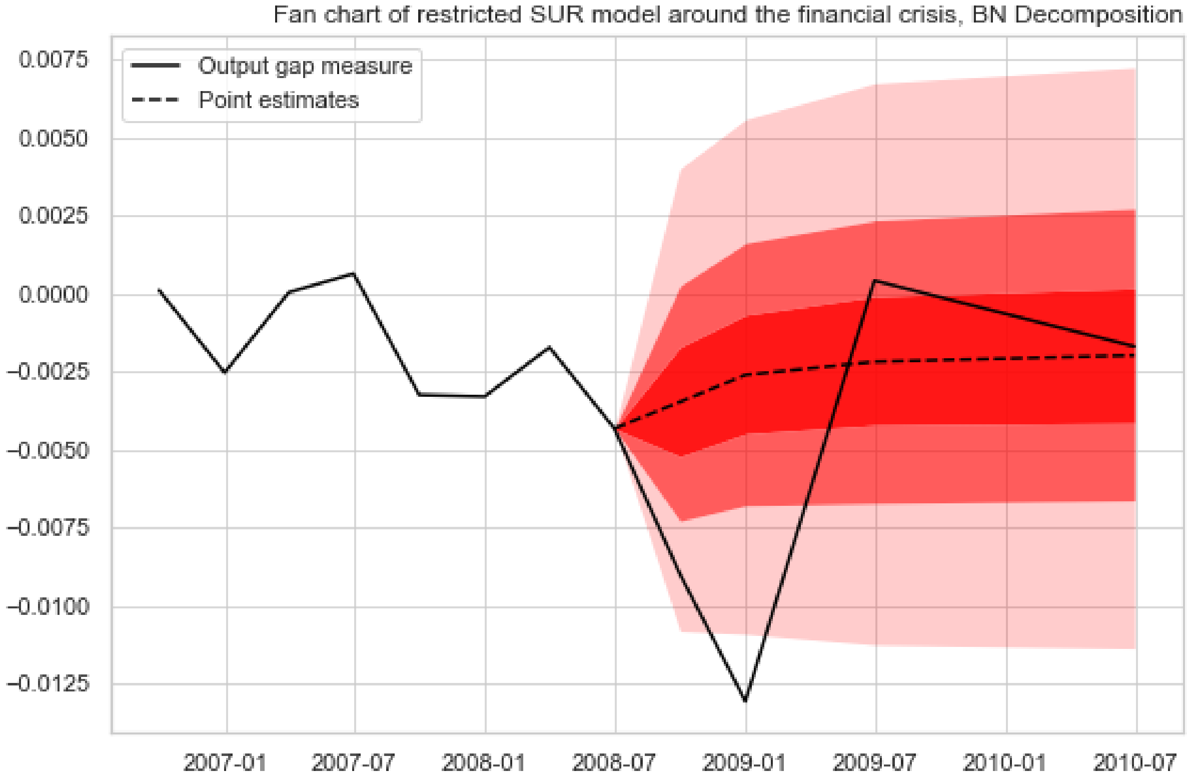

Figure E.7: Fan chart of the benchmark seasonality model around the financial crisis, BN Decomposition

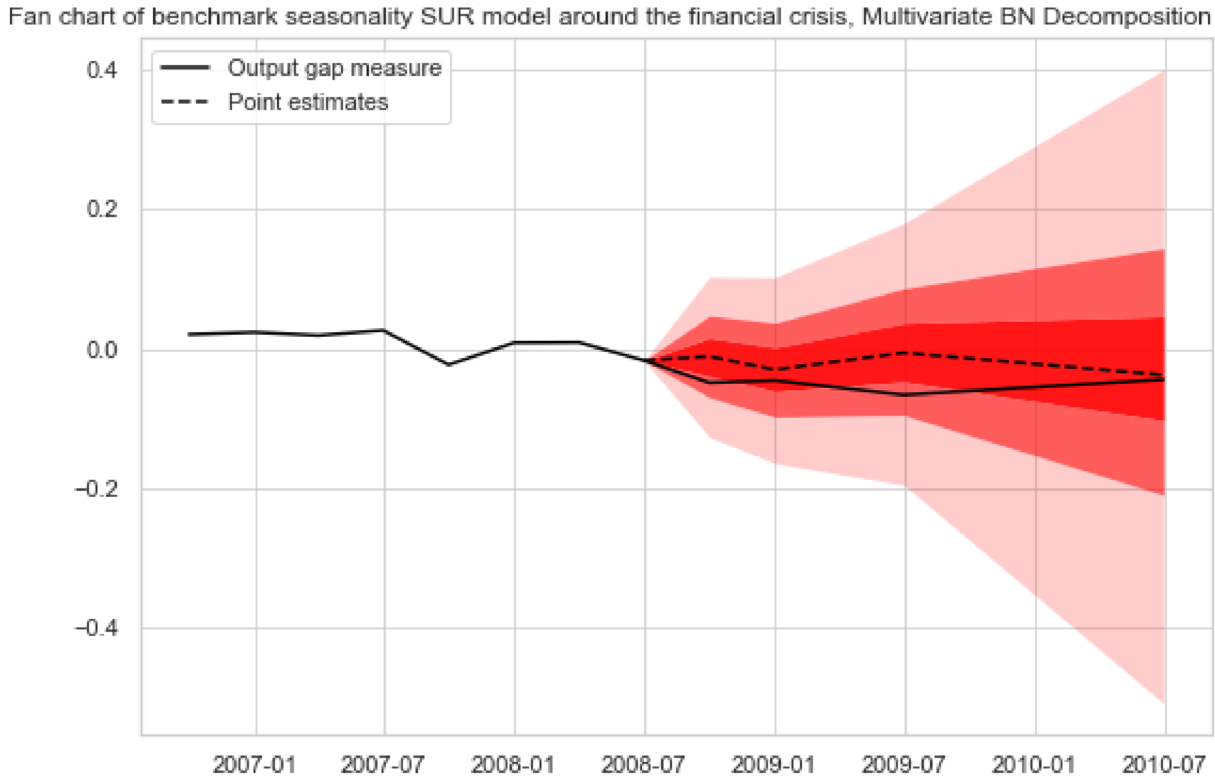

Figure E.8: Fan chart of the benchmark seasonality model around the financial crisis, multivariate BN Decomposition

F. Inflation fan charts

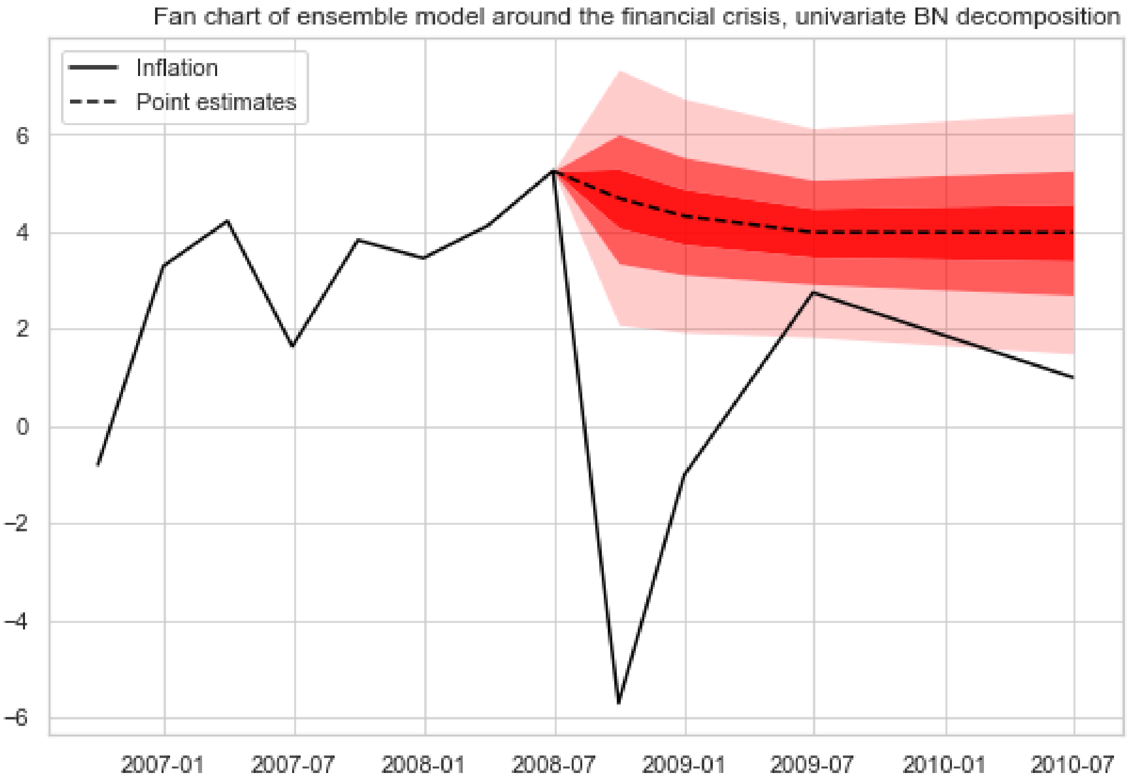

Figure F.1: Fan chart of ensemble model around the financial crisis, univariate BN decomposition

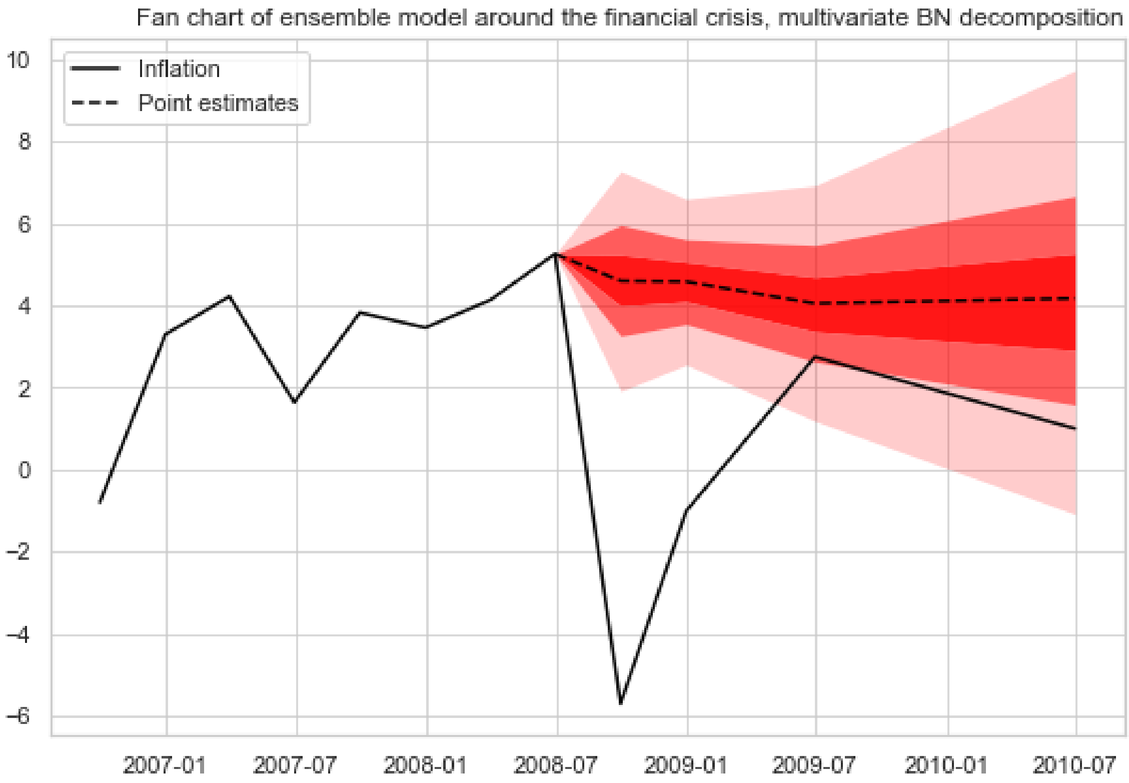

Figure F.2: Fan chart of ensemble model around the financial crisis, multivariate BN decomposition

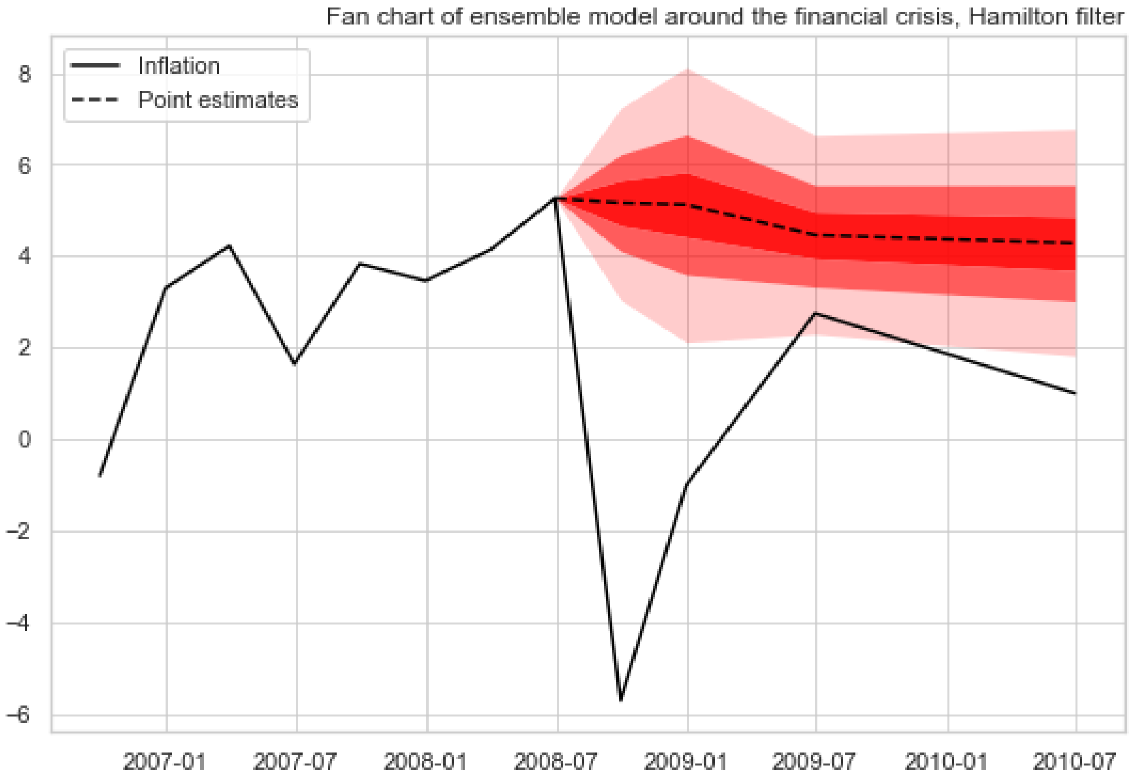

Figure F.3: Fan chart of ensemble model around the financial crisis, Hamilton filter

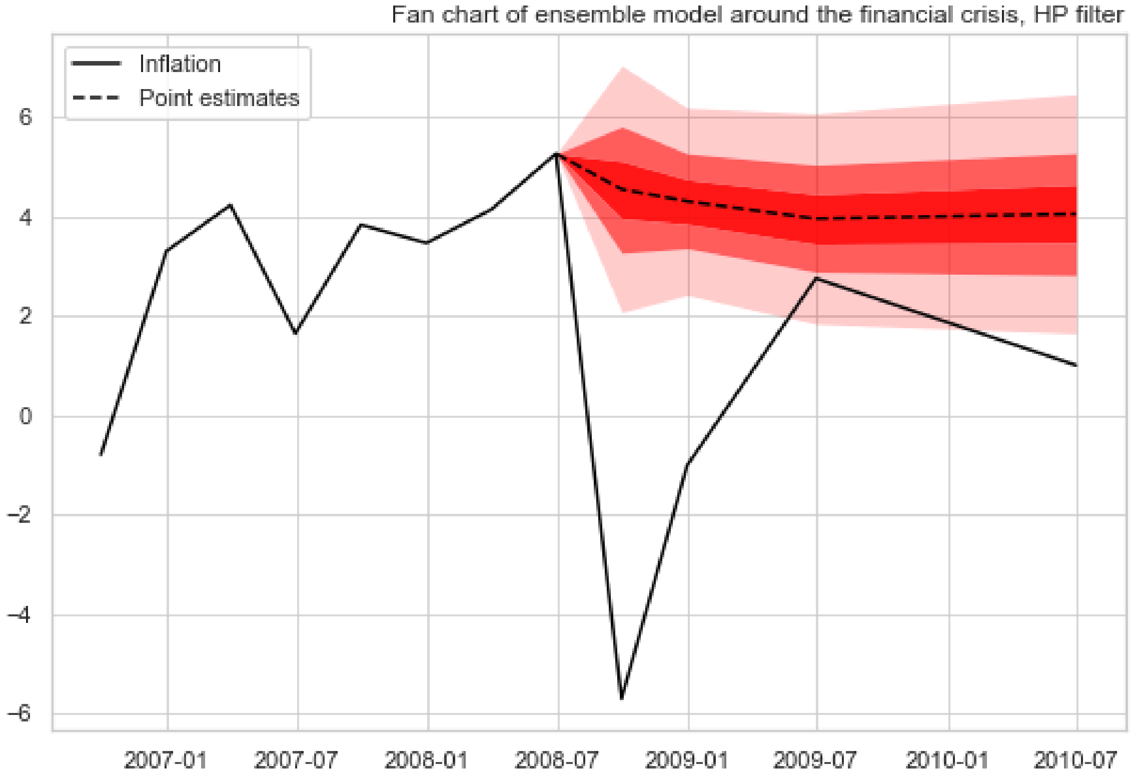

Figure F.4: Fan chart of ensemble model around the financial crisis, HP filter

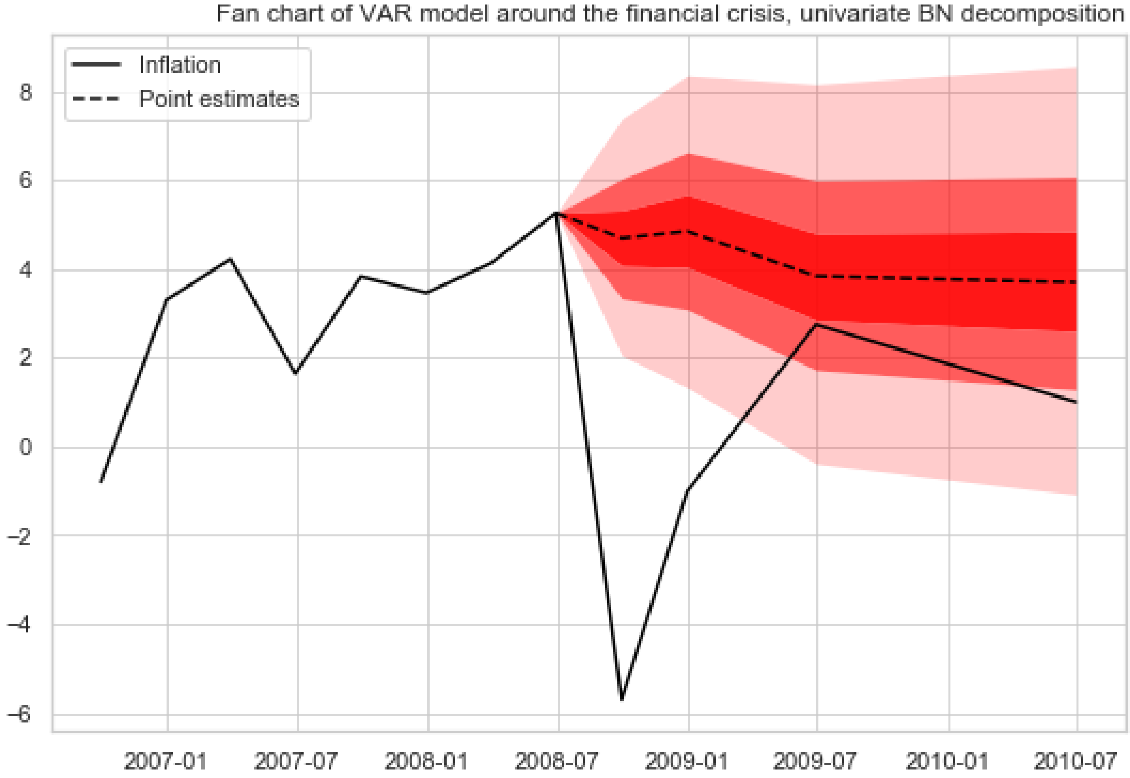

Figure F.5: Fan chart of VAR model around the financial crisis, univariate BN decomposition

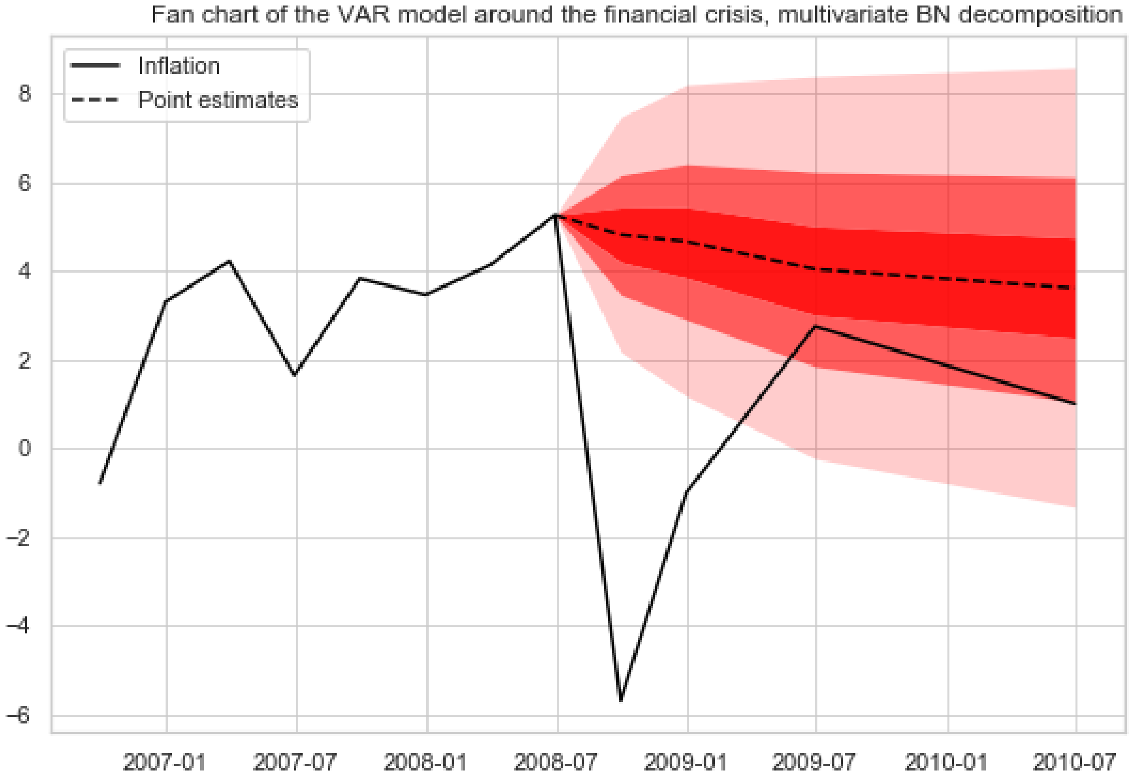

Figure F.6: Fan chart of VAR model around the financial crisis, multivariate BN decomposition

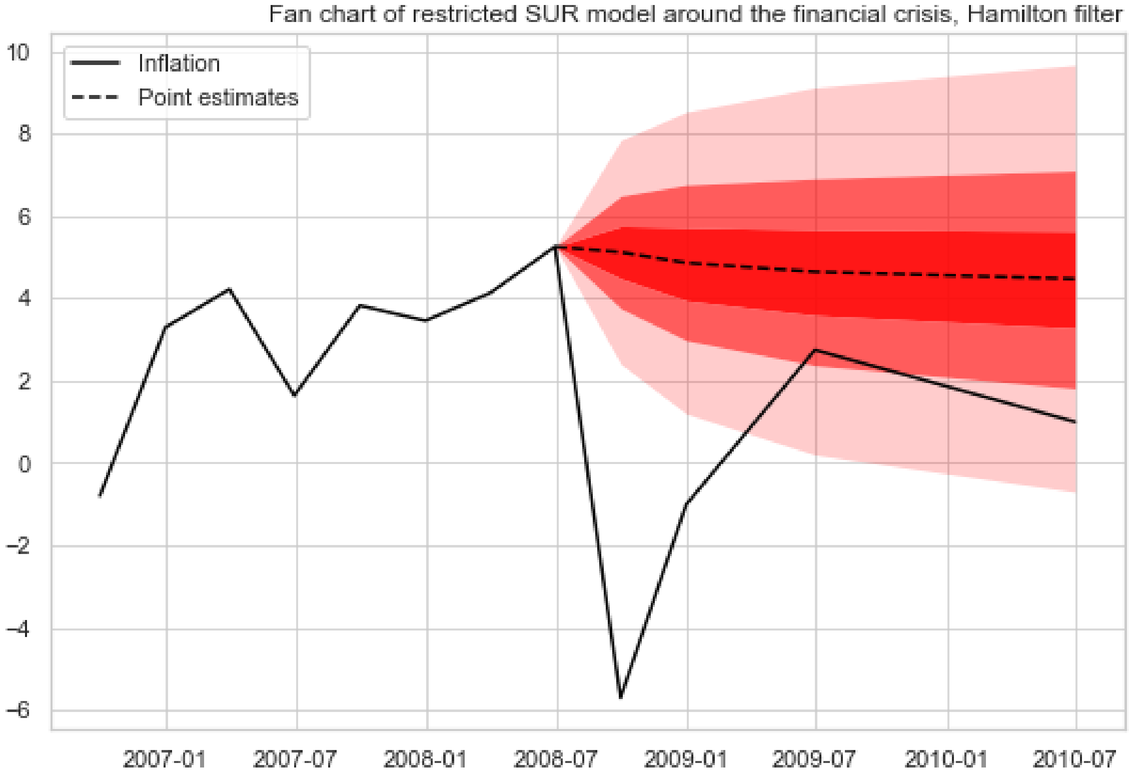

Figure F.7: Fan chart of restricted SUR model around the financial crisis, Hamilton filter

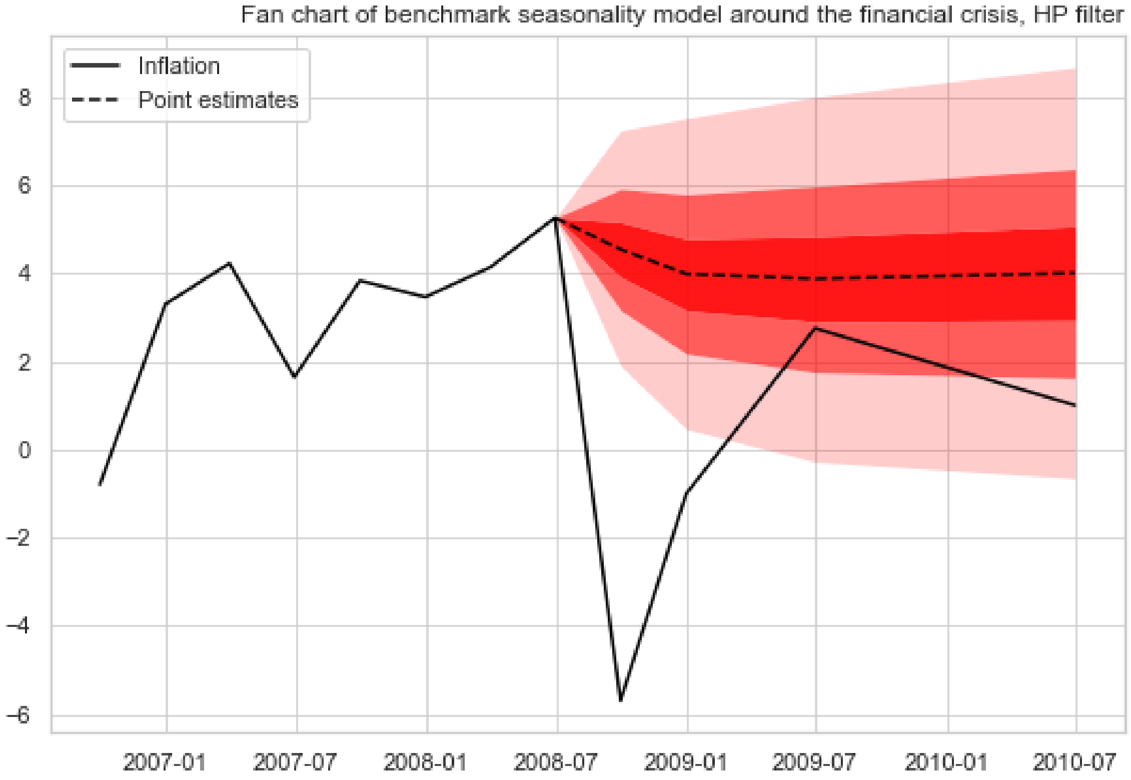

Figure F.8: Fan chart of benchmark seasonality SUR model around the financial crisis, HP filter