Legerstee, Rianne; Franses, Philip Hans

**Working Paper**

# Does Disagreement amongst Forecasters have Predictive Value?

Tinbergen Institute Discussion Paper, No. 10-088/4

**Provided in Cooperation with:**
Tinbergen Institute, Amsterdam and Rotterdam

*Suggested Citation:* Legerstee, Rianne; Franses, Philip Hans (2010) : Does Disagreement amongst Forecasters have Predictive Value?, Tinbergen Institute Discussion Paper, No. 10-088/4, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/87065

TI 2010-088/4
Tinbergen Institute Discussion Paper

# Does Disagreement amongst Forecasters have Predictive Value?

*Rianne Legerstee*
*Philip Hans Franses*

*Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.*

**Tinbergen Institute**

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

**Tinbergen Institute Amsterdam**
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

**Tinbergen Institute Rotterdam**
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

# Does disagreement amongst forecasters have predictive value?

Rianne Legerstee[1]
Philip Hans Franses

*Econometric Institute*
*Erasmus University Rotterdam*

September 2, 2010

**Abstract**

Forecasts from various experts are often used in macroeconomic forecasting models. Usually the focus is on the mean or median of the survey data. In the present study we adopt a different perspective on the survey data as we examine the predictive power of disagreement amongst forecasters. The premise is that this variable could signal upcoming structural or temporal changes in an economic process or in the predictive power of the survey forecasts. In our empirical work, we examine a variety of macroeconomic variables, and we use different measurements for the degree of disagreement, together with measures for location of the survey data and autoregressive components. Forecasts from simple linear models and forecasts from Markov regime-switching models with constant and with time-varying transition probabilities are constructed in real-time and compared on forecast accuracy. We find that disagreement has predictive power indeed and that this variable can be used to improve forecasts when used in Markov regime-switching models.

**Keywords:** model forecasts; expert forecasts; survey forecasts; Markov regime-switching models; disagreement; time series
**JEL:** C53

Authors notes:
We are very grateful to participants of the 30th Annual International Symposium on Forecasting in San Diego on June 20-23, 2010 and to participants of the Tinbergen lunch seminar in Rotterdam on June 8, 2010 for their helpful comments.

[1]Corresponding author. Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, Netherlands, e-mail: legerstee@ese.eur.nl

# 1 Introduction

Forecast combinations are often found to outperform their individual component forecasts. An extensive body of research exists on finding the optimal forecast combination across individual forecasts. Most of the time, the conclusion is that simple combinations perform best, see Webby and O'Connor (1996) and Diebold and Lopez (1996) for overviews. However, recently, Elliott and Timmermann (2005) argued that an optimal forecast combination for macroeconomic variables was one that combines forecasts from expert-based survey data and time series models where the weights are driven by a Markov regime-switching model.

As in Elliott and Timmermann (2005), there are many situations in macroeconomics in which there is not just one single expert forecast available. Indeed, quite often, forecasts from surveys are available, consisting of forecasts from various experts who make predictions for the same variable. A well-known example is the Survey of Professional Forecasters (SPF) conducted by the Federal Reserve Bank of Philadelphia. Most studies that analyze SPF-type data focus on the predictive value of the mean or median of these SPF forecasts. Theoretical and empirical research has shown the relevance of these two statistics (see e.g. Einhorn and Hogharth, 1975; Clemen, 1989; Armstrong, 2001).

Recently, a growing literature discusses other features of experts-based survey forecasts. For example, disagreement amongst experts is described (Dovern et al., 2009), an explanation is sought for this disagreement (Capistrán and Timmermann, 2009; Mankiw et al., 2003), and its effects on decision makers are investigated (Baillon and Cabantous, 2009). Besides disagreement amongst experts, various other features are described and evaluated. Survey forecasts typically involve much dependence between forecasts from the same expert (Cooke, 1991). There is also positive serial correlation in the forecast errors (Mankiw et al., 2003; Capistrán and Timmermann, 2009). Expert opinions are biased (Laster et al., 1999) and characteristics of the forecasters as age and experience are related to forecast performance (Lamont, 2002).

Interestingly, although much information on features of survey forecasts is available, there is no research on the predictive power of these features. In this paper, we therefore look at this predictive power where we focus on the disagreement amongst the forecasters. Our conjecture, which we outline in more detail below, is that the degree of disagreement could signal upcoming structural or temporal changes in an economic process or in the predictive power of the survey forecasts.

In our empirical work, we examine a variety of US-specific macroeconomic variables, and we consider different ways to measure the degree of disagreement. The models include, besides a measure for the degree of disagreement, measures for location of the survey data and autoregressive components. Forecasts from simple linear models and forecasts from Markov regime-switching models with constant and with time-varying transition probabilities are constructed in real-time and compared on forecast accuracy. The survey forecasts are from the Survey of Professional Forecasters. Our main finding is that disagreement can indeed have predictive value.

The remainder of the paper is structured as follows. Section 2 discusses in more detail the literature and it explains which variables and models are used. Section 3 describes the data, the models and the methods to evaluate the forecasts. Section 4 gives a summary of the results and section 5 concludes.

# 2 Background

In this section, we first discuss the features of expert forecasts that could be relevant for forecasting, and next, we explain how such features can appear in forecasting models.

## 2.1 Why disagreement may be useful

Without doubting the usefulness of the mean and median of survey forecasts, there are many reasons to also look at other statistics of survey forecasts. Researchers have been puzzled by some (seemingly) irrational characteristics of expert forecasts since

long and explanations have been sought in multiple directions. For example, most theoretical macroeconomic models do not endogenously generate disagreement, while disagreement is prevalent in every survey of (professional) forecasters (Dovern et al., 2009; Laster et al., 1999; Capistrán and Timmermann, 2009). Other characteristics unexplained by full rationality are autocorrelated forecast errors and insufficient sensitivity to recent macroeconomic news (Capistrán and Timmermann, 2009; Mankiw et al., 2003). Note that explanations given in these studies for these features might also explain why these features could have predictive power, and that is what we will consider next.

The new statistics that we consider here are the standard deviation, the 5th percentile and the 95th percentile of the survey forecasts and the number of forecasts collected. We begin by explaining why we address the first three of these statistics, which can be seen as direct measures of disagreement amongst forecasters [2].

Laster et al. (1999) describe that it is questionable that professional forecasts are rational in the sense of being efficient and unbiased. They construct a model in which a forecaster is driven by two conflicting incentives, and these are (1) to forecast as accurate as possible and (2) to generate publicity for their firms. Most ideally, a prediction is accurate and all other predictions are very inaccurate. Being accurate while all others are too, does not generate much publicity. At the same time, being wrong once in a while and being the only one close to the true value at other times, might be better for a firm than always following consensus. Therefore, professional forecasters may behave strategically rational.

Related to this principal-agent model is the work of Lamont (2002). He finds that the age and experience of forecasters are related to forecast accuracy. The older and more established forecasters they are, the more radical their forecasts and the more inaccurate. This can again be explained by reputational factors.

[2] The 5th percentile and the 95th percentile of survey forecasts can be seen as measures of disagreement amongst forecasters, especially when used in combination with the mean or median of the survey forecasts. This is what we do in this study, see the following sections.

Following this line of thought, it might be beneficial for a firm or individual expert to give extreme forecasts at times in which change may come up, even if they are not sure what kind of change it will be or how it will look like. Some forecasters, who may be more dependent on publicity, might react more extreme to certain information than other forecasters would do. In these periods some forecasters might take an additional risk, because correctly forecasting extreme future observations generates positive publicity, while being wrong about it is not that bad.

In our empirical work below, we rely on survey forecasts from anonymous sources. It is possible to follow one and the same forecaster by way of a code and it is known what kind of firm provides the forecast, but the names of the forecasters and the firms are unknown. The question is now, to what extent the forecasters behave strategically as described above. Without having any information about this, it is very well possible that the forecasts provided in the anonymous survey are used in other contexts too, where reputational factors do play a role. Furthermore, within the firms the personal reputation of a forecaster might also be important and factors described by Lamont (2002) might still influence the practice of forecasting.

Forecasters who react in different ways to specific information also fits the arguments in Capistrán and Timmermann (2009). The forecasters are presumed to have asymmetric loss functions, there is heterogeneity in agents' loss functions, and a constant loss component can explain how dispersion in inflation beliefs evolves over time and why it is correlated with the level and volatility of inflation. Without discussing their model in detail, it is intuitively clear that if forecasters have asymmetric and differing loss functions (possibly also convex), they update their forecasts in very different ways and particular information might cause dispersion in forecasts.

Also Mankiw et al. (2003) propose a model that can reproduce the distribution of inflation expectations. They use a sticky-information model in which agents only update their forecasts periodically because of costs involved in gathering information and adjusting projections [3]. Each period, only a fraction of the forecasters gets new

[3]Carroll (2003) uses a very similar model to explain the evolution of variation in inflation expecta-

information and processes it. Mankiw et al. (2003) find that this model is capable of matching the disagreement and its evolution in inflation expectations.

If we were to consider this model to represent how expert opinions are adjusted, then disagreement in expert expectations might very well signal upcoming changes in an economic process. If only some of the survey experts receive information about this and processes it, then the standard deviation, the 95th percentile or the 5th percentile might contain this information, while the mean and the median might do so only partially or perhaps not at all.

A final argument why we focus on disagreement in forecasts originates from the work of Zarnowitz and Lambros (1987) and Lahiri and Sheng (2010), amongst others. Zarnowitz and Lambros (1987) show, also by using the SPF data, that measures of consensus, the degree of agreement amongst point predictions, and uncertainty, lack of confidence, are positively correlated. Lahiri and Sheng (2010) also argue that disagreement amongst experts, measured as the standard deviation of expectations, is a good proxy for forecast uncertainty, assuming that the variability of future aggregate shocks is stable.

When disagreement in expectations increases while the variability of future aggregate shocks does so too or stays the same, the forecast uncertainty of expert opinions also increases and the predictive power of the survey forecasts gets reduced. It might be better in such a situation to rely more on statistical model forecasts and less on expert opinion. However, if the level of disagreement and the variability of future aggregate shocks move in opposite directions, it is unclear what to do.

Finally, the fourth explanatory variable that we consider, which is the number of forecasts collected, is a bit different from the previous three. This variable is related to disagreement, because with more forecasts there is more opportunity for disagreement and more possibility for differing statements. But because more forecasts does not necessarily result in a larger variation in forecasts, we discuss this variable separately.

tions, namely an epidemiological model in which information goes from one person to another in the same manner as diseases go from one person to another.

Firms might go bankrupt and therefore no longer provide forecasts. It is also known that participation in the survey might depend on efforts of the organization collecting the data [4]. But firms might also decide strategically to stop or to start forecasting. If they are unsure about the future they may decide not to provide a forecast, so they cannot be wrong. Or, they provide forecasts if they have information they believe is exclusive. So, the number of forecasts might itself be informative about what might happen next.

Another reason why this might be the case, probably more convincing in case of anonymous data, is that the economic situation can influence the necessity for and the ability of firms to forecast. If firms are in trouble because of the economic situation, research departments may be closed. On the other hand, economic volatile times can increase the need to get insight in what the future will bring and thus the need to create forecasts. So, although it is not clear how the number of forecasts and the predictability of macroeconomic variables are linked exactly, it is clear that there is a good reason why they might be linked.

Finally, the number of forecasts might indicate how reliable the mean or median of those forecasts is. One can imagine that if the number of forecasts is low, that extreme erroneous forecasts are not cancelled out and that the weight put on it in models and forecast combinations should better be small in that case.

## 2.2 How to include disagreement in forecasting models

We will use the above mentioned explanatory variables in three different forecasting models. The first is a simple ARX model, in which autoregressive terms and each time one of the new statistics of the survey data is included. The second is a Markov regime-switching model with constant transition probabilities (MRScons) where one

[4] In case of the Survey of Professional Forecasters: initially the survey was conducted by the American Statistical Association (ASA), together with the National Bureau of Economic Research (NBER), but was taken over by the Federal Reserve Bank of Philadelphia in 1990. They revived the survey by inviting new forecasters to the survey, as participation rates had dropped in the years before.

of the explanatory variables each time is one of the new survey statistics. The third is a Markov regime-switching model with time-varying transition probabilities (MRSvar). In this model, one of the four new statistics is used to predict if a regime switch will happen. The models are explained in more detail in the next section, but here we will give some arguments why we use these MRS models.

First of all, Markov regime-switching models are an often used and popular tool to describe and forecast macroeconomic variables, ever since the publication of the influential paper of Hamilton (1989). In Granger (2001) some of the many applications are discussed, showing its suitability in a variety of forms for macroeconomic variables such as output growth, inflation and interest rates. Timmermann (2000) showed the flexibility of these models by deriving the moments for a range of Markov switching models. MRS models showed to be capable of accounting for specific features of macroeconomic time series such as volatility clustering, asymmetry, and fat-tail behavior.

Second, Elliott and Timmermann (2005) showed that Markov regime-switching models might be useful in combining forecasts. They indicate that the weights of the combination of AR model forecasts and expert forecasts could be driven by a Markov regime-switching process. For three of their six macroeconomic series, that is, unemployment rate, inflation and nominal GDP growth, this combination method performs better than a range of alternative combination methods.

A third and final reason to use MRS models follows from our arguments above why we focus on statistics of disagreement as explanatory variable. This variable might signal upcoming structural or temporal changes in an economic process or in the predictive power of the survey forecasts. A good way to model this is by incorporating these variables as explanatory variables for the transition probabilities in a MRSvar model.

The models include lags as explanatory variables, because many studies have shown that combinations of expert forecasts with statistical model forecasts outperform both individual forecasts. AR models and AR-MRS models have a proven track record.

Furthermore, we include the mean or median of the survey forecasts in some of the models. Most studies analyzing survey type forecasting data restrictively focus on the predictive value of the mean or median. Clemen (1989) found in a broad study on forecast combinations that simple arithmetic averages of forecasts are accurate for many types of forecasts. Einhorn and Hogharth (1975) argue that equal weights produce precise forecasts because there is no estimation error, no degrees of freedom are lost, and no mistakes can be made with the 'true' relative weights, giving the wrong forecast the largest weight. According to Armstrong (2001), who refers to multiple other studies, there is evidence that the median is even more accurate. As the mean is an often used and praised statistic (Zarnowitz and Braun, 1993; Laster et al., 1999; Elliott and Timmermann, 2005), we will incorporate both variables in our empirical study.

# 3 Methodology

In this section we will discuss the data, the models and the evaluation methods in more detail.

## 3.1 Data

The survey forecasts we use are from The Survey of Professional Forecasters (SPF), which is a quarterly survey of macroeconomic forecasts in the United States of America. The survey began in the fourth quarter of 1968 and was conducted by the American Statistical Association and the National Bureau of Economic Research, so it is often called the NBER-ASA survey. The Federal Reserve Bank of Philadelphia took over the survey in the second quarter of 1990.

The respondents to the survey are forecasting professionals. Participants include, amongst others, financial firms, banks, consultancy firms and university research centers. For each variable, 'forecasts' are given for the previous quarter (for which the first release is already available), for the current quarter (for which no realized data

is available yet) and for the four following quarters. We consider one-quarter-ahead predictions, where one quarter ahead is the first quarter for which no data is available yet at the moment of estimating the model parameters and creating forecasts. So, we use the survey forecasts given for the current quarter.

At present, the survey encompasses 31 macroeconomic variables. We use five of these variables in our analysis [5]: the index of industrial production (INPROD), Nominal Gross Domestic Product (GNP before 1992) (NGDP), inflation as measured through the GDP chain-weighted price index (PGDP), the unemployment rate (UNEMP) and private housing starts (HOUSING). Our main focus is on INPROD, for which we will give the most detailed results and we compare the results for INPROD with the results for the other variables. For INPROD and HOUSING, realized monthly figures are available, so averages are taken to obtain quarterly figures. Except for the unemployment rate, we look at growth rates, measured as the first differences in natural logs of the current value and the previous quarter's value.

The SPF forecasts, used for the explanatory variables in the forecasting models, are also transformed into growth rates. We include the rate of change in the forecast for the current quarter over the 'forecast' of that same forecaster for the previous quarter. So, growth is measured as $100 * (\ln(ySPF_{i,t}) - \ln(ySPF_{i,t-1}))$, where $ySPF_{i,t}$ is the forecast of forecaster $i$ for variable $y$ for the current quarter and $ySPF_{t-1}$ is the 'forecast' of that same forecaster for the previous quarter. We use for this the forecasters' own stated value for previous quarter instead of the first release, because we think it gives a better representation of the forecasted growth. Most of the time, these two figures are the same and sometimes they differ because of mistakes in for example the base year used to construct some of the variables. These mistakes cancel out by using forecasts instead of release values.

At the start of our analysis, SPF forecasts were available for the last quarter of 1968 to and including the third quarter of 2009. Realized data were also available up until

[5] Except for corporate profits, we use the same variables as Elliott and Timmermann (2005) do in their study.

the third quarter of 2009. So we work with $n = 164$ data points. Each quarter there are $f_t$ forecasts available from the SPF for the current quarter and the previous quarter.

## 3.2 The general model

The Markov regime-switching model with time-varying transition probabilities nests all models we consider. This MRSvar for the variable to be explained $y_t$, with $m$ regimes and $p$ lags, is

$$y_t = \alpha_{s_t} + \phi'_{s_t}(L)y_{t-1} + \beta'_{s_t}X_t + \varepsilon_{t,s_t}, \tag{1}$$

or stated differently:

$$y_t = \begin{cases} \alpha_1 + \phi_1(1)y_{t-1} + ... + \phi_1(p)y_{t-p} + \beta'_1 X_t + \varepsilon_{t,1} & \text{if in state 1} \\ \alpha_2 + \phi_2(1)y_{t-1} + ... + \phi_2(p)y_{t-p} + \beta'_2 X_t + \varepsilon_{t,2} & \text{if in state 2} \\ \quad\vdots \qquad\qquad \vdots \qquad\qquad \vdots & \qquad\qquad \vdots \\ \alpha_m + \phi_m(1)y_{t-1} + ... + \phi_m(p)y_{t-p} + \beta'_m X_t + \varepsilon_{t,m} & \text{if in state m,} \end{cases} \tag{2}$$

with $\phi_{s_t}(L)$ a polynomial lag of order $p$, $s_t \in [1, m]$ the regime state in period $t$, $X_t$ is a vector with $k$ explanatory variables and $\varepsilon_{t,s_t} \sim N(0, \omega_{s_t})$. The model can assume the $\omega_{s_t}$ to vary per regime, or $\omega_{s_t}$ to be constant over the different regimes. The variable $s_t$ is unobserved and it is assumed to develop according to a first-order Markov chain with transition probabilities

$$p_{ijt} = Pr[s_{t+1} = j | s_t = i, Z_t] = \frac{\exp(\delta_{ij} + \gamma'_{ij}Z_t)}{\sum_{j=1}^{m} \exp(\delta_{ij} + \gamma'_{ij}Z_t)}, \tag{3}$$

with $Z_t$ a vector of $r$ explanatory variables for the regime switching and $\delta_{i1}$ and $\gamma_{i1}$ are set to zero for identification purposes.

If $Z_t$ in (3) is empty (or if all $\gamma_{ij}$ are set to zero), the model reduces to a Markov regime-switching model with constant transition probabilities (MRScons). With all $\delta_{ij}$ and $\gamma_{ij}$ set to zero the regimes have an equal probability of occurring.

If $m = 1$ the resulting model is a linear model with $p$ lags and additional explanatory variables $X$ (ARX). If $m = 1$ and $X$ is empty (or if $\beta$ is set to zero) we get a simple AR model.

## 3.3 Models considered

$X_t$ and $Z_t$ are vectors with one or more variables related to the SPF forecasts (transformed into growth rates). It is also possible that they are empty. The variables used for $Z_t$ are standardized to facilitate the estimation process. The variables that we consider for $X_t$ and $Z_t$ are divided into two sets. The first set consists of two variables, that is the mean and median of the forecasts, and are therefore called the location variables. The second set consists of the standard deviation (std), the 0.05 quantile (5p) and the 0.95 quantile (95p) of the forecasts and the number of forecasts $f_t$ (nr), and each of these four is a different measure for the degree of disagreement amongst forecasters.

For each of the five macroeconomic variables four groups of models are put forward. These are AR models, ARX models, MRScons models and MRSvar models. The first group consists of linear models with lags of the dependent variable as the explanatory variables. Models with zero, one, two, three and four lags are considered, so in total this group consists of five different models.

The second group contains the same linear models as the first group, only now with one or two additional explanatory variables. So, this group consists of models with zero, one or more lags and one of the six explanatory variables described above and it consists of models with zero, one or more lags and two explanatory variables, one being a location variable and one being a variable measuring disagreement. In total, this group encompasses 70 models.

The third group is the group with MRScons models, consisting of 150 models. For each of the models in the first two groups two models are estimated in this group, each with two regimes ($m = 2$) and one with common variance and one with varying variance per regime.

The final and fourth group contains 24 MRSvar models. These models are estimated with $m$ set to two, with one lag, $X_t$ containing zero or one of the variables of the first group of explanatory variables (that is, the mean or the median of the SPF forecasts), $Z_t$ containing one of the variables of the second group of explanatory variables and with a common or varying variance per regime.

All model parameters are estimated 40 times. The first time with 124 data points, leaving out the last 40 observations, the second time again with 124 data points, leaving out the first observation and the last 39 observations and so on. Stated differently, parameter estimation adopts a rolling window of data. Each model estimated with data up till date $t$, will be estimated with the vintage of date $t$ (that is, the last data release available at date $t$). Only for INPROD, which is the focal variable, the model parameters are also estimated 60 times, using a rolling estimation window of 104 data points and leaving out 60 observations.

Estimation of the parameters proceeds by optimizing the likelihood function associated with the Markov regime-switching model or with the linear model. As the underlying state variable $s_t$ is assumed to be unobserved, it is treated as a latent variable and the EM algorithm described in Hamilton (1994) is used for the estimation of the MRScons models. The EM algorithm developed by Diebold et al. (1994) is used for estimation of the MRSvar models.

For the MRS models, especially the MRSvar models, the estimation results might depend on the starting values of the parameters used in the estimation procedure, as the likelihood function has multiple local optima. Therefore, we use a grid of different starting values for $\delta_{ij}$ and $\gamma_{ij}$ the first time the model parameters are estimated and we select the model with the maximum log-likelihood. The resulting estimated parameters are used as starting values in the next step of the real-time forecasting procedure, and the resulting estimated parameters of that step are used as starting values in the step after that, and so on. Every five steps the model parameters are estimated using the grid of different starting values again. Most ideally, the grid of different starting values for $\delta_{ij}$ and $\gamma_{ij}$ would be used at each step, in order to get the most optimal results.

However, as this takes too much computation time, the model parameters are estimated using the grid every five steps.

## 3.4 Forecast evaluation

For each model and each macroeconomic variable a set of one-step-ahead forecasts is created. The forecasts are created making use of the parameters estimated with the most recent available data if the forecasts are created in real-time and making use of a rolling estimation window. We use 124 data points to estimate the model parameters, so we obtain $P = 40$ forecasts to evaluate.

For INPROD we also look at two other sets of forecasts per model. The first set is created with the models estimated in the first round with the first 124 data points, so for these forecasts we use a fixed estimation window. We also obtain $P = 40$ forecasts in this case. The second set of predictions is created making use of a rolling estimation window again, but now the models are estimated over 104 data points, so here we obtain $P = 60$ forecasts.

Forecasts for the MRS models are constructed as in Hamilton (1994). This means that with two regimes the one-step-ahead forecast is

$$\hat{y}_{t+1|t} = E[y_{t+1}|s_{t+1} = 1, \Omega_t] \cdot P(s_{t+1} = 1|\Omega_t; \theta) + E[y_{t+1}|s_{t+1} = 2, \Omega_t] \cdot P(s_{t+1} = 2|\Omega_t; \theta), \quad (4)$$

where $\theta$ denote the estimated parameters, $\Omega_t$ is all the data available up to date $t$ and $P(s_{t+1} = j|\Omega_t; \theta)$ are the one-step-ahead state probabilities computed from the filtered state probabilities $P(s_t = j|\Omega_t; \theta)$ (which are obtained from the estimation procedure) and multiplied by the transition probabilities in (3). For the fixed estimation window these one-step-ahead state probabilities obtained from the first estimation round are multiplied by the transition probabilities in (3) to obtain forecasts in the second round and so on.

One way to evaluate the forecasts is to select in each step the number of lags for

the models in the first three groups using an information criterion. In the same way, a selection between constant and varying variance in the MRS models could be made. However, we decide to only analyze and compare the forecasts of the models estimated with one lag and to look at the models with varying and constant variance without selecting one of these with an information criterion. This decision is based on our finding that results do not necessarily improve by working with information criteria to select the number of lags and to make a selection between constant and varying variance. Furthermore, it is not that clear which information criterion to use [6]. This way, we focus completely on the predictive value of the disagreement variables without the results being flawed because of a possible inappropriate use of information criteria.

The forecasts are analyzed using two kinds of data realizations. These are the first release and last release data, as they are known in the third quarter of 2009. Root mean squared prediction errors (RMSPE) are constructed and the RMSPE's of our models (including one of the four SPF variables introduced in this study) are compared to 9 benchmarks. As benchmark forecasts we use the mean and median of SPF forecasts and forecasts from 7 different benchmark models, which are the AR model with one lag, the ARX model with one lag and with the mean or median of SPF forecasts and the MRS model with one lag and with mean or median of SPF forecasts and with varying or constant variance of the error terms (as inspired by Elliott and Timmermann (2005)).

We use two different tests to see if the differences in RMSPE's are significant. The first is the well-known test of Diebold and Mariano (1995) (DM). McCracken (2000) and Clark and McCracken (2001) have shown that this test is not valid if the models are nested models, because the asymptotic distribution of the test statistic is not standard in this case. However, Giacomini and White (2006) showed that the DM-test remains valid for nested models when the estimation sample size remains finite, or

[6] See for example Psaradakis and Spagnolo (2006), Smith et al. (2006) and Awirothananon and Cheung (2009). They investigate which information criterion to use to choose between different MRS models, but also focus on the decision on the number of regimes and for example not on the choice to use common or varying variance. Furthermore, their results are conflicting.

stated differently, when a rolling or fixed estimation sample forecasting scheme is used. Thus, although we work with nested models, by using rolling and fixed estimation sample forecasting schemes it is possible to use the DM-test in a standard way.

It is found that the DM-statistic tends to be over-sized in small samples. As we have a rather small sample of forecasts we adjust the DM-test in a way proposed by Harvey et al. (1997) to overcome this problem. To that end we adjust the DM-statistic by multiplying it with the square root of $(P-1)/P$. We also compare this adjusted statistic with critical values obtained from a Student's t-distribution with $P-1$ degrees of freedom, instead of using the standard normal distribution.

The second test we use is proposed by Van Dijk and Franses (2003). This test is put forward to specifically compare the forecasting performance of linear and nonlinear time series models. Van Dijk and Franses (2003) argue that nonlinear time series models often do not outperform linear models in out-of-sample forecasting, despite their superior in-sample fit. They suggest that this might be due to the use of inappropriate evaluation criteria and suggest using a criterium that weights the forecasted observations. Therefore, they use the DM-statistic and adjust it by using different weight functions in such a way that more weight is placed on the relevant observations which are most associated with non-linearity (for example, turning points). Weight functions that they propose focus on one or two tails ($LT$ and $RT$) of the distribution of the dependent variable. We look at the same three weight functions, being: $w_T(y_t) = 1 - \phi(y_t)/\max(\phi(y_t))$, $w_{LT}(y_t) = 1 - \Phi(y_t)$ and $w_{RT}(y_t) = \Phi(y_t)$. Here, $\phi(\cdot)$ is the density function of $y_t$ and $\Phi(\cdot)$ is the cumulative distribution function of $y_t$. The density function of $y_t$ is estimated using the relevant in-sample observations and using a normal kernel function with automatic bandwidth selection. The empirical cumulative density function is used as an estimate of $\Phi(y_t)$.

# 4 Results

As announced in the previous section, to save space our main focus is on the variable INPROD. In the first part of this section we analyze the forecasts obtained for this variable using a rolling estimation window of 124 observations. After that, we check the robustness of these results in three different ways. First of all, we look at fixed estimation window forecasts for INPROD. Second, we look at forecasts for INPROD obtained using a rolling estimation window of 104 observations. Finally, we analyze forecasts for the other four macroeconomic variables obtained using a rolling estimation window of 124 observations.

The RMSPE's of a large part of the estimated models can be found in tables 1 and 2. Although models without the mean or median of SPF forecasts are estimated too, results from these models are omitted from the tables, because these models appeared to have a very poor forecasting performance in all cases.

## 4.1 Industrial Production, rolling window, 40 forecasts

Columns 1 and 4 of table 1 show the RMSPE's of the sets of 40 INPROD forecasts created by using a rolling estimation window. The ten smallest RMSPE's are displayed in boldface.

What is remarkable is that the simple mean and median of the SPF forecasts perform very well over this period (RMSPE's of $0.916$ and $0.909$). It is obviously more precise than any of the 7 benchmark models. Clearly, it is also difficult for the alternative models in this case to outperform these SPF forecasts.

For every new SPF variable (std, 5p, 95p and nr) to use in forecasting, there is at least one model that seems to outperform the benchmark models in forecasting to some extent. We discuss these variables therefore one by one.

**Table 1:** RMSPE's for models estimated for growth of INPROD. Forecasts are created using a rolling estimation window ('roll') or a fixed estimation window ('fix') and RMSPE's are calculated over 40 or 60 forecasts. All models include one lag as explanatory variable and the SPF variables as indicated below. The 'c' or 'v' indicates if a constant or varying variance per regime is used for the error terms in the MRS models. The bold RMSPE's are the ten smallest RMSPE's in that column.

| **INPROD** | **Last release** | | | **First release** | | |
|---|---|---|---|---|---|---|
| | **Roll 40** | **Fix 40** | **Roll 60** | **Roll 40** | **Fix 40** | **Roll 60** |
| **AR** | 1.247 | 1.226 | 1.126 | 1.198 | 1.186 | 1.064 |
| **ARX-mean** | 0.967 | 0.918 | 0.893 | 0.864 | 0.817 | 0.773 |
| **ARX-med** | 0.966 | **0.911** | 0.896 | 0.861 | 0.804 | 0.773 |
| **MRScons-mean-c** | 0.959 | 0.959 | 0.902 | 0.844 | 0.848 | 0.776 |
| **MRScons-mean-v** | 0.994 | 0.969 | 0.899 | 0.880 | 0.877 | 0.784 |
| **MRScons-med-c** | 0.997 | 0.954 | 0.894 | 0.896 | 0.845 | 0.769 |
| **MRScons-med-v** | 0.999 | 0.938 | 0.911 | 0.887 | 0.835 | 0.790 |
| **SPF-mean** | **0.916** | 0.916 | **0.857** | **0.837** | 0.837 | **0.740** |
| **SPF-med** | **0.909** | **0.909** | **0.853** | **0.823** | **0.823** | **0.729** |
| **ARX-mean+std** | 0.977 | 0.939 | 0.918 | 0.876 | 0.832 | 0.793 |
| **ARX-mean+5p** | 0.960 | 0.917 | 0.912 | 0.859 | 0.816 | 0.784 |
| **ARX-mean+95p** | 0.978 | 0.969 | 0.907 | 0.876 | 0.857 | 0.787 |
| **ARX-mean+nr** | 0.965 | 0.915 | 0.901 | 0.858 | 0.814 | 0.776 |
| **ARX-med+std** | 0.978 | 0.936 | 0.920 | 0.874 | 0.823 | 0.793 |
| **ARX-med+5p** | 0.982 | 0.941 | 0.900 | 0.875 | 0.828 | 0.783 |
| **ARX-med+95p** | 0.952 | 0.917 | 0.910 | 0.852 | 0.809 | 0.780 |
| **ARX-med+nr** | 0.972 | **0.912** | 0.908 | 0.861 | 0.805 | 0.780 |
| **MRScons-mean+std-c** | 0.983 | 0.928 | 0.956 | 0.861 | 0.811 | 0.828 |
| **MRScons-mean+std-v** | 0.965 | 0.931 | 0.931 | 0.845 | 0.822 | 0.803 |
| **MRScons-mean+5p-c** | **0.917** | 0.943 | 0.910 | **0.810** | 0.837 | 0.771 |
| **MRScons-mean+5p-v** | 0.981 | 0.918 | 0.903 | 0.870 | 0.827 | 0.761 |
| **MRScons-mean+95p-c** | 0.966 | 0.938 | 0.927 | 0.844 | 0.816 | 0.805 |
| **MRScons-mean+95p-v** | 0.993 | 0.968 | 0.932 | 0.879 | 0.850 | 0.814 |
| **MRScons-mean+nr-c** | **0.917** | 0.921 | 0.907 | **0.799** | **0.795** | 0.786 |
| **MRScons-mean+nr-v** | **0.935** | 0.953 | 0.914 | **0.811** | 0.831 | 0.788 |
| **MRScons-med+std-c** | 0.990 | 0.959 | 0.942 | 0.873 | 0.847 | 0.814 |
| **MRScons-med+std-v** | 1.008 | 0.956 | 0.913 | 0.891 | 0.842 | 0.776 |
| **MRScons-med+5p-c** | 0.978 | 0.974 | 0.892 | 0.865 | 0.861 | 0.774 |
| **MRScons-med+5p-v** | 0.985 | 1.006 | 0.897 | 0.870 | 0.894 | 0.781 |
| **MRScons-med+95p-c** | 0.976 | 0.914 | 0.911 | 0.861 | **0.798** | 0.782 |
| **MRScons-med+95p-v** | 0.960 | 0.924 | 0.896 | 0.844 | 0.816 | 0.754 |
| **MRScons-med+nr-c** | 0.947 | 0.965 | 0.909 | **0.832** | 0.842 | 0.782 |
| **MRScons-med+nr-v** | **0.920** | **0.907** | 0.901 | **0.806** | **0.794** | 0.777 |
| **MRSvar-mean+std-c** | 1.029 | **0.908** | 0.884 | 0.934 | **0.803** | **0.741** |
| **MRSvar-mean+std-v** | **0.945** | 0.923 | **0.844** | 0.863 | 0.821 | **0.712** |
| **MRSvar-mean+5p-c** | 1.035 | 0.932 | 0.978 | 0.943 | 0.832 | 0.844 |
| **MRSvar-mean+5p-v** | 0.976 | 0.931 | **0.823** | 0.889 | 0.827 | **0.694** |
| **MRSvar-mean+95p-c** | 0.978 | 0.982 | 1.027 | 0.872 | 0.904 | 0.928 |
| **MRSvar-mean+95p-v** | 0.975 | 0.913 | 0.946 | 0.844 | **0.793** | 0.828 |
| **MRSvar-mean+nr-c** | 0.952 | **0.880** | **0.873** | 0.840 | **0.792** | 0.750 |
| **MRSvar-mean+nr-v** | **0.882** | **0.891** | **0.870** | **0.773** | **0.799** | 0.752 |
| **MRSvar-med+std-c** | 1.007 | 0.949 | **0.873** | 0.926 | 0.848 | **0.736** |
| **MRSvar-med+std-v** | 1.028 | 0.919 | **0.827** | 0.961 | 0.814 | **0.697** |
| **MRSvar-med+5p-c** | 1.038 | 0.923 | 0.885 | 0.942 | 0.819 | 0.765 |
| **MRSvar-med+5p-v** | 1.030 | 0.961 | **0.854** | 0.946 | 0.855 | **0.728** |
| **MRSvar-med+95p-c** | **0.933** | 1.147 | 0.934 | **0.810** | 1.071 | 0.810 |
| **MRSvar-med+95p-v** | **0.936** | **0.892** | 0.986 | **0.810** | **0.773** | 0.873 |
| **MRSvar-med+nr-c** | 1.071 | **0.885** | 0.874 | 0.976 | **0.783** | **0.745** |
| **MRSvar-med+nr-v** | 1.089 | **0.875** | **0.865** | 0.976 | **0.779** | **0.746** |

**Table 2:** RMSPE's for models estimated for UNEMP and growth of NGDP, PGDP and HOUSING. The RMSPE's are calculated over 40 forecasts, which are created using a rolling estimation window. For further information, see the caption of table 1.

| **Roll 40** | **Last release** | | | | **First release** | | | |
|---|---|---|---|---|---|---|---|---|
| | **NGDP** | **PGDP** | **UNEM** | **HOUS** | **NGDP** | **PGDP** | **UNEM** | **HOUS** |
| **AR** | 0.776 | 0.331 | 0.392 | 7.390 | 0.702 | 0.349 | 0.391 | 7.683 |
| **ARX-mean** | 0.443 | 0.279 | 0.121 | 5.338 | 0.358 | 0.264 | 0.113 | 5.425 |
| **ARX-med** | 0.433 | 0.278 | 0.120 | **5.089** | 0.349 | 0.266 | 0.115 | **5.176** |
| **MRScons-mean-c** | 0.443 | 0.271 | 0.120 | 5.348 | 0.354 | 0.256 | 0.116 | 5.392 |
| **MRScons-mean-v** | 0.441 | 0.278 | 0.123 | 5.369 | 0.348 | 0.270 | 0.114 | 5.415 |
| **MRScons-med-c** | **0.429** | 0.277 | 0.120 | **5.096** | **0.341** | 0.264 | 0.120 | **5.139** |
| **MRScons-med-v** | 0.437 | 0.278 | 0.122 | **5.091** | 0.346 | 0.267 | 0.118 | **5.142** |
| **SPF-mean** | 0.458 | **0.237** | 0.135 | 6.010 | 0.363 | **0.230** | 0.129 | 6.126 |
| **SPF-med** | 0.450 | **0.241** | 0.127 | 5.624 | 0.354 | **0.239** | 0.123 | 5.735 |
| **ARX-mean+std** | 0.441 | 0.278 | 0.120 | 5.676 | 0.355 | 0.263 | **0.111** | 5.754 |
| **ARX-mean+5p** | 0.440 | 0.281 | 0.122 | 5.593 | 0.353 | 0.265 | 0.114 | 5.679 |
| **ARX-mean+95p** | 0.443 | 0.277 | **0.114** | 5.594 | 0.357 | 0.263 | **0.104** | 5.666 |
| **ARX-mean+nr** | 0.441 | 0.282 | 0.119 | 5.626 | 0.359 | 0.266 | 0.111 | 5.709 |
| **ARX-med+std** | 0.433 | 0.277 | 0.119 | 5.362 | **0.346** | 0.265 | 0.114 | 5.442 |
| **ARX-med+5p** | **0.433** | 0.279 | 0.120 | **5.137** | 0.349 | 0.266 | 0.115 | **5.227** |
| **ARX-med+95p** | **0.432** | 0.280 | 0.117 | 5.401 | **0.345** | 0.268 | **0.113** | 5.465 |
| **ARX-med+nr** | **0.425** | 0.280 | **0.117** | 5.430 | **0.342** | 0.267 | 0.113 | 5.513 |
| **MRScons-mean+std-c** | 0.451 | **0.264** | 0.124 | 5.301 | 0.355 | **0.253** | 0.119 | 5.311 |
| **MRScons-mean+std-v** | 0.447 | **0.267** | 0.123 | 5.550 | 0.352 | 0.261 | 0.117 | 5.567 |
| **MRScons-mean+5p-c** | 0.447 | **0.257** | 0.130 | 5.438 | 0.358 | **0.250** | 0.124 | 5.481 |
| **MRScons-mean+5p-v** | 0.453 | 0.276 | 0.123 | 5.453 | 0.352 | 0.274 | 0.116 | 5.483 |
| **MRScons-mean+95p-c** | 0.450 | **0.263** | **0.112** | 5.312 | 0.364 | **0.251** | **0.107** | 5.315 |
| **MRScons-mean+95p-v** | 0.446 | 0.269 | 0.119 | 5.289 | 0.355 | 0.260 | **0.110** | 5.348 |
| **MRScons-mean+nr-c** | 0.442 | 0.271 | 0.129 | 5.690 | 0.356 | **0.251** | 0.124 | 5.752 |
| **MRScons-mean+nr-v** | 0.433 | 0.275 | **0.115** | 5.623 | 0.354 | 0.261 | **0.108** | 5.695 |
| **MRScons-med+std-c** | 0.443 | 0.270 | 0.117 | 5.512 | 0.351 | 0.265 | 0.114 | 5.505 |
| **MRScons-med+std-v** | 0.449 | 0.270 | 0.120 | 5.394 | **0.345** | 0.269 | 0.118 | 5.414 |
| **MRScons-med+5p-c** | **0.427** | **0.268** | **0.116** | **5.031** | **0.343** | 0.259 | 0.114 | **5.082** |
| **MRScons-med+5p-v** | 0.473 | 0.272 | 0.121 | **5.049** | 0.379 | 0.265 | 0.117 | **5.094** |
| **MRScons-med+95p-c** | 0.445 | 0.268 | **0.110** | 5.389 | **0.344** | 0.260 | **0.111** | 5.381 |
| **MRScons-med+95p-v** | **0.430** | 0.276 | **0.117** | 5.274 | **0.334** | 0.269 | 0.114 | 5.294 |
| **MRScons-med+nr-c** | **0.423** | 0.280 | 0.120 | 5.406 | **0.339** | 0.267 | 0.120 | 5.485 |
| **MRScons-med+nr-v** | **0.420** | 0.272 | 0.120 | 5.303 | **0.338** | 0.264 | 0.116 | 5.367 |
| **MRSvar-mean+std-c** | 0.494 | 0.269 | 0.129 | 5.826 | 0.473 | 0.265 | 0.121 | 5.922 |
| **MRSvar-mean+std-v** | 0.491 | 0.277 | 0.131 | 5.302 | 0.477 | 0.266 | 0.123 | 5.390 |
| **MRSvar-mean+5p-c** | 0.517 | 0.273 | 0.124 | **4.945** | 0.389 | 0.264 | 0.119 | **5.061** |
| **MRSvar-mean+5p-v** | 0.506 | 0.277 | 0.124 | 5.559 | 0.397 | 0.270 | 0.114 | 5.553 |
| **MRSvar-mean+95p-c** | 0.535 | 0.281 | 0.120 | 5.798 | 0.434 | 0.261 | 0.115 | 5.886 |
| **MRSvar-mean+95p-v** | 0.514 | 0.275 | 0.124 | 5.741 | 0.399 | 0.262 | 0.119 | 5.708 |
| **MRSvar-mean+nr-c** | 0.491 | 0.275 | **0.114** | 5.353 | 0.385 | **0.255** | **0.105** | 5.478 |
| **MRSvar-mean+nr-v** | 0.460 | 0.276 | **0.108** | 5.358 | 0.384 | **0.254** | **0.100** | 5.487 |
| **MRSvar-med+std-c** | **0.424** | **0.262** | 0.127 | **5.153** | 0.353 | 0.257 | 0.123 | **5.214** |
| **MRSvar-med+std-v** | **0.423** | 0.275 | 0.127 | 5.391 | 0.355 | 0.267 | 0.120 | 5.510 |
| **MRSvar-med+5p-c** | 0.502 | **0.266** | 0.125 | 5.356 | 0.373 | **0.255** | 0.123 | 5.334 |
| **MRSvar-med+5p-v** | 0.473 | **0.266** | 0.123 | 5.453 | 0.369 | 0.257 | 0.121 | 5.462 |
| **MRSvar-med+95p-c** | 0.504 | 0.280 | 0.123 | 5.644 | 0.407 | 0.263 | 0.121 | 5.742 |
| **MRSvar-med+95p-v** | 0.499 | 0.273 | 0.122 | 5.585 | 0.402 | 0.259 | 0.120 | 5.690 |
| **MRSvar-med+nr-c** | 0.475 | 0.271 | 0.119 | **5.107** | 0.374 | **0.254** | 0.116 | **5.238** |
| **MRSvar-med+nr-v** | 0.451 | 0.276 | **0.115** | **5.136** | 0.365 | 0.262 | **0.111** | **5.249** |

## Std

The first is the standard deviation of SPF forecasts. The MRS model with one lag and the mean of the SPF forecasts as explanatory variables, the standard deviation of SPF

forecasts used to model regime switches and a varying variance of the error terms per regime (MRSvar-mean+std-v) is amongst the ten models with the lowest RMSPE's. Before we discuss the forecasting performance in detail it might be interesting to see how the estimated model parameters look like. To that extent we estimate the model parameters over the complete data set and compared the results with the models estimated over parts of the data set. The results look quite the same, so we discuss the estimates for the complete data set here.

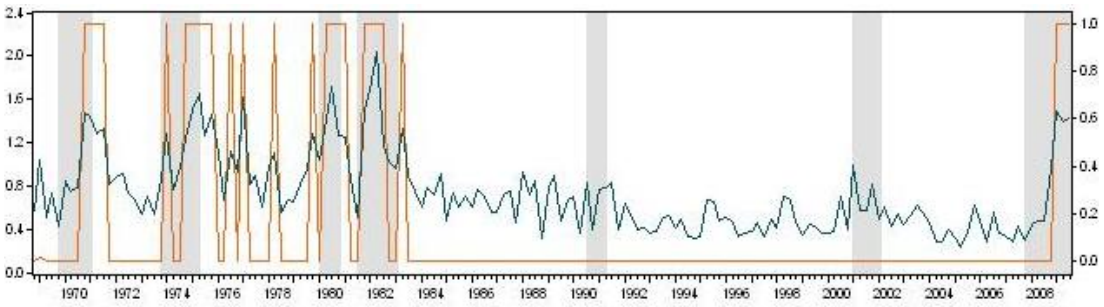


**Figure 1:** This figure shows the probability at regime two as estimated by the MRSvar-mean+std-v model for growth of INPROD in combination with the standard deviation of SPF forecasts. The green line, with its scale on the left, is the standard deviation of SPF forecasts. The orange line, fluctuating between 0 and 1 and with its scale on the right, is the probability at regime two. The shaded area's are recessions as indicated by the NBER.

Figure 1 shows the standard deviation of SPF forecasts, the estimated smoothed probabilities at regime two and recessions as officially declared by the NBER. It can be seen that this model estimates one regime that occurs most of the time, when the standard deviation of SPF forecasts is not too high. When the standard deviation of SPF forecasts rises above approximately 1.1 the process switches to regime two. In panel 1 of table 3, the estimated parameters are given with their significance. In regime one, we see a negative constant, a significant [7] positive coefficient for the lag of INPROD growth, a significant positive coefficient for the mean of SPF forecasts and a variance of the error terms of around 0.6. In regime two, the estimated intercept is much more negative, the coefficient for the lag is not significantly different from zero

[7] A significance level of 5% is used.

**Table 3:** Coefficients of models estimated for INRPOD data from the fourth quarter of 1968 to the third quarter of 2009. If the coefficients are significantly different from 0 at the 5%-level is indicated by '*'.

| | **Regime 1** | **Regime2** |
|---|---|---|
| | **MRSvar-mean+std-v** | |
| **c** | -0.064 | -0.240 |
| **lag** | 0.274* | -0.000 |
| **mean** | 0.808* | 1.038* |
| **var** | 0.563* | 2.170* |
| | **MRScons-mean+5p-c** | |
| **c** | -0.576 | -0.217 |
| **lag** | 0.180* | 0.110 |
| **mean** | 0.902* | 1.744* |
| **5p** | 0.183* | -0.619* |
| **var** | 0.378* | 0.378* |
| | **MRSvar-med+95p-c** | |
| **c** | -0.100 | 1.011 |
| **lag** | 0.244* | -0.127 |
| **median** | 0.905* | 0.514* |
| **var** | 0.787* | 0.787* |
| | **MRSvar-mean+nr-v** | |
| **c** | -0.165 | 0.367 |
| **lag** | 0.100* | 0.626* |
| **mean** | 1.106* | -0.180 |
| **var** | 0.812* | 0.236* |

anymore, the coefficient for the mean of SPF forecasts is a little bit higher and still significant and the variance of the error terms is with a value of around 2 much higher than in regime one.

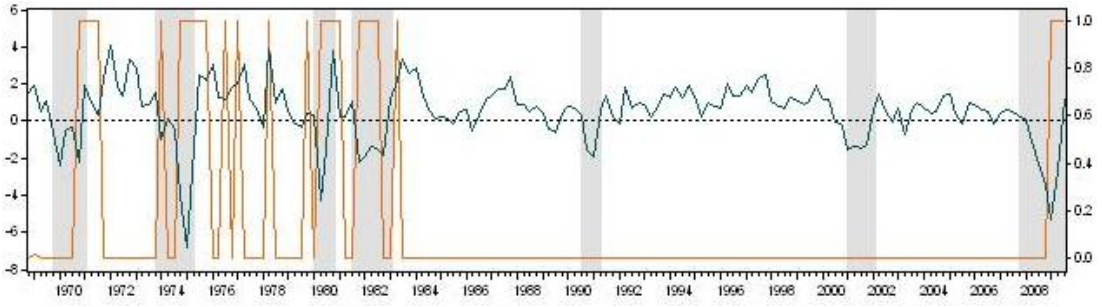


**Figure 2:** This figure shows the probability at regime two as estimated by the MRSvar-mean+std-v model for growth of INPROD in combination with growth of INPROD. The green line, with its scale on the left, is growth of INPROD. The orange line, fluctuating between 0 and 1 and with its scale on the right, is the probability at regime two. The shaded area's are recessions as indicated by the NBER.

In figure 2, growth of INPROD, the estimated smoothed probabilities at regime two and recessions as officially declared by the NBER are shown. We see that regime two often occurs around recessionary periods, but not always. During regime two as indicated by the model, growth of INPROD is on average negative and during regime one it is positive. Furthermore, in regime two growth of INPROD is much more volatile than in regime one and regime two covers more extreme values than regime one.

We can conclude from these results that the standard deviation of SPF forecasts might predict volatile periods in which a relatively higher weight should be put on the mean of SPF forecasts and a lower weight on the lag of INPROD growth in forecasting models/ combinations.

If we take a closer look at the forecasting performance of this model, we see that the RMSPE of this model is lower than all the benchmark models, but it is not lower than the mean or median of the SPF forecasts. In the last column of table 4, we see that this model does produce significantly[8] more accurate forecasts than a few of the benchmark models, but not all.

[8] A significance level of 10% is used here.

**Table 4:** This table shows if the models in the header of the column forecast significantly more accurate than the models in the header of the row, according to the tests as described in section 3.4. The models are estimated for INPROD data, a rolling estimation window is used and 40 forecasts are created. '++' indicates that the model in the column header produces more accurate forecasts than the model mentioned in the row header according to the unweighted test. '+' indicates that not the unweighted test, but at least one of the weighted tests shows a significant difference. A significance level of 10% is used.

| **INPROD Roll 40** | **MRSvar-mean+nr-v** | **MRScons-med+nr-v** | **MRScons-mean+nr-c** | **MRScons mean+5p-c** | **MRSvar med+95p-c** | **MRSvar mean+std-v** |
|---|---|---|---|---|---|---|
| | **Last release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | ++ | + | | + | | |
| **ARX-med** | ++ | + | + | ++ | | |
| **MRScons-mean-c** | ++ | ++ | ++ | ++ | | |
| **MRScons-mean-v** | ++ | ++ | ++ | ++ | ++ | ++ |
| **MRScons-med-c** | ++ | ++ | ++ | ++ | ++ | ++ |
| **MRScons-med-v** | ++ | ++ | ++ | ++ | ++ | |
| **SPF-mean** | + | | | | | |
| **SPF-med** | + | | | | | |
| | **First release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | ++ | | | + | | |
| **ARX-med** | ++ | ++ | ++ | ++ | ++ | |
| **MRScons-mean-c** | ++ | ++ | ++ | ++ | | |
| **MRScons-mean-v** | ++ | ++ | ++ | ++ | ++ | |
| **MRScons-med-c** | ++ | ++ | ++ | ++ | ++ | |
| **MRScons-med-v** | ++ | ++ | ++ | ++ | ++ | |
| **SPF-mean** | + | | | + | | |
| **SPF-med** | + | + | + | + | + | |

## 5p

The MRS model with constant transition probabilities, constant variance of the error terms and as independent variables one lag and SPF mean and 5p (MRScons-mean+5p-c) has even lower RMSPE's. Although there are differences between the models estimated to create forecasts and the model estimated for the complete data set, some general features of the full data model are prevalent in the smaller models. Therefore we will again look at this full data model.

This model distinguishes two regimes that occur about equally often, see figure 3. The smoothed probability at regime two follows the fluctuations in growth of INPROD

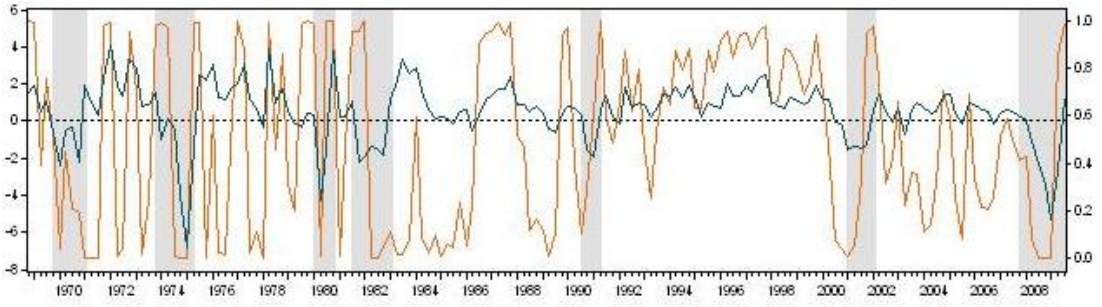


**Figure 3:** This figure shows the probability at regime two as estimated by the MRScons-mean+5p-c model for growth of INPROD in combination with growth of INPROD. For more information, see the caption of figure 2.

very closely. On average this growth is higher in regime two, with also a higher 5th and 95th percentile. In regime one, growth is slightly negative on average. In panel 2 of table 3, we present the estimated coefficients in both regimes. The most obvious difference is that the coefficient of the mean rises from 0.9 in regime one to 1.7 in regime two and that the coefficient of 5p declines from around 0.2 to around -0.6. So, in regime one, where growth of INPROD is most of the time declining, the coefficients of the mean and 5th percentile of survey forecasts have the same sign. The mean of survey forecasts has a coefficient close to 1 and when the most 'negative' forecasters predict a negative growth in this regime the final growth forecast should be lowered, ceteris paribus. In regime two, where growth of INPROD is most of the time increasing, the coefficients of the mean and 5th percentile of survey forecasts have opposite signs. The coefficient of the mean of survey forecasts indicates that the average forecast is too modest and should be inflated, ceteris paribus, and when the most 'negative' forecasters predict a negative growth in this regime the final forecast should be increased and visa versa.

The forecasting performance of this model (MRScons-mean+5p-c) is much better than that of the previous model we discussed. Its RMSPE's are lower than the RMSPE's of all benchmark models and even lower than the RMSPE's of SPF mean and median in case of first release data, see columns 1 and 4 of table 1. It produces significantly more accurate forecasts according to the unweighted test than all benchmark

models except one, both compared to first release and last release data. The remaining benchmark model is beaten according to at least one of the weighted tests, again both compared to first and last release data, and SPF mean and median are beaten significantly according to at least one weighted test if we look at first release data, see column 4 of table 4.

**95p**

The third explanatory variable of interest, 95p, seems to have predictive value in the MRS model with one lag and the median of survey forecasts as explanatory variables, with 95p as explanatory variable for regime switches and with a constant variance of the error terms (MRSvar-med+95p-c). If we estimate this model over the complete data set it again partly resembles the estimated models used to create the forecasts. We find one regime that occurs the most and that is when the 95th percentile of growth forecasts is lower than approximately 3%, see figure 4. In this regime (see the third panel of table 3) the constant is slightly negative, the coefficient of the lag is significantly positive and the coefficient of the median is significantly different from zero with a value of around 0.9. When 95p rises above 3 we find a regime with an intercept around 1, a coefficient of the lag that is not significantly different from 0 and a coefficient of the median which is significantly different from 0 with a value of around 0.5. In both regimes the variance of the error terms is approximately 0.8. In figure 5 we see that regime 2 often occurs right after a recession or otherwise right after a dip in INPROD growth. During regime two growth of INPROD is on average much higher than during regime one and always positive.

This all indicates that when the 95th percentile of survey forecasts increases it is likely that a recovery period is coming up with on average high growth of industrial production. This is a plausible result. In this period the lag of INPROD should not receive much weight in the forecast, while the constant should be higher and the coef-

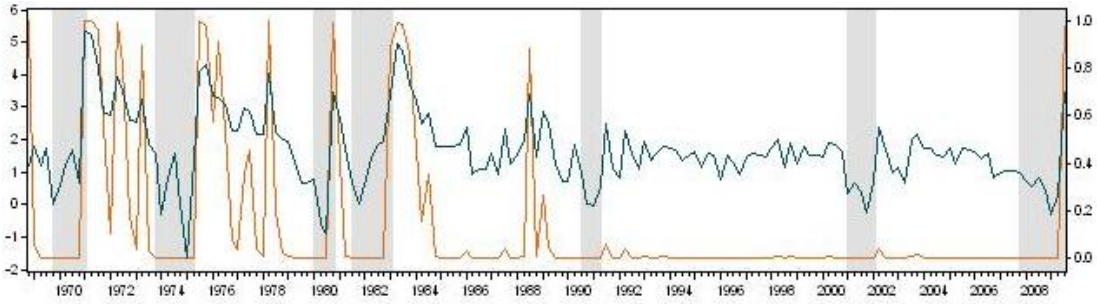

**Figure 4:** This figure shows the probability at regime two as estimated by the MRSvar-med+95p-c model for growth of INPROD in combination with the 95th percentile of SPF forecasts. For more information, see the caption of figure 1.

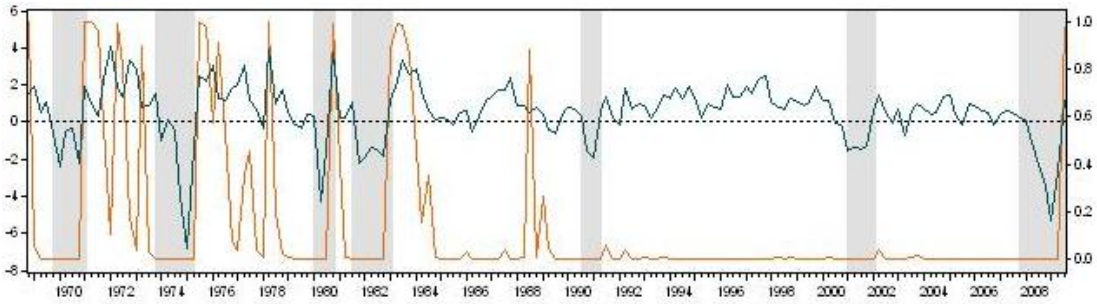

**Figure 5:** This figure shows the probability at regime two as estimated by the MRSvar-med+95p-c model for growth of INPROD in combination with growth of INPROD. For more information, see the caption of figure 2.

ficient of SPF median should be lower.

The forecasting performance of this model is quite good, see again table 1, columns 1 and 4. The RMSPE is clearly lower than the RMSPE's of all benchmark models, both calculated over first release and last release data. However, the difference is not always significant (see column 5 of table 4) and SPF mean and SPF median are more precise predictors.

**Nr**

Finally, for the number of survey forecasts we find multiple models that seem capable of outperforming the benchmark models in forecasting, see columns 1 and 4 of table 1 and the first 3 columns of table 4. The model with the best forecasting performance is a MRS model with as explanatory variables one lag and the mean of survey forecasts,

with as explanatory variable for the regime switches nr and with a varying variance of the error terms. We again estimated this model over the complete data set to see what it looks like. In figure 6 we see the probability at regime 2 in combination with the number of SPF forecasts. The first regime estimated by the model occurs most of the time, notably when the number of forecasts is above approximately 27. When the number of forecasts is lower than 27, the probability that the process switches to regime two increases. See table 3 for the estimated coefficients and its significance. In regime one, when the number of forecasts is high, we find a negative intercept, a significantly positive coefficient for the lag, a significantly positive coefficient for the mean of SPF forecasts and a variance of the error terms of around 0.8. In regime two, when the number of forecasts is low, we find a positive intercept, a coefficient for the lag that is higher than in regime one and still significant, a coefficient for the mean of survey forecasts that is not significant anymore and a lower variance of the error terms. In figure 7, we observe when regime two occurs according to this model in combination with growth of INPROD. Regime two is obviously not related to recessionary periods. During regime two the dependent variable is much more stable and not that volatile as during regime one.

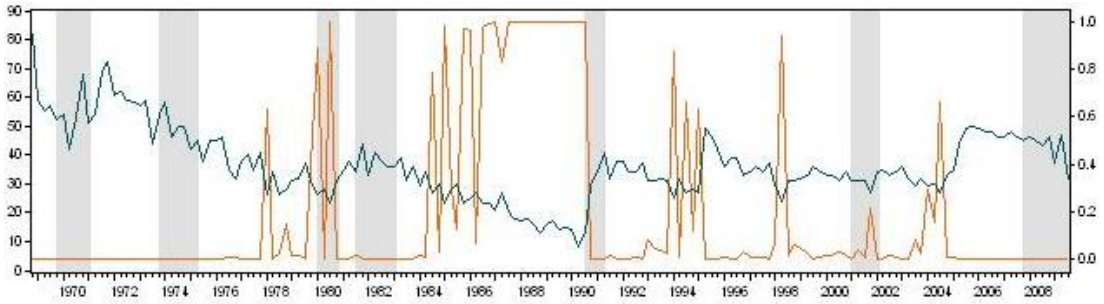


**Figure 6:** This figure shows the probability at regime two as estimated by the MRSvar-mean+nr-v model for growth of INPROD in combination with the number of SPF forecasts. For more information, see the caption of figure 1.

There are two conclusions which could be drawn from the results. The first is that when the dependent variable is more volatile, the expert forecasts should receive more

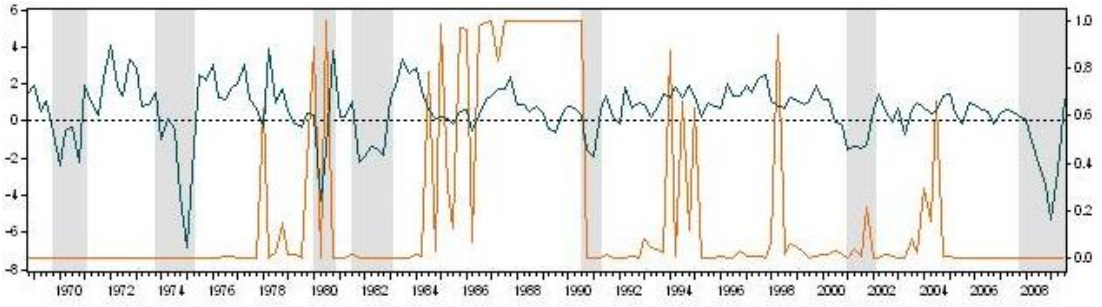


**Figure 7:** This figure shows the probability at regime two as estimated by the MRSvar-mean+nr-v model for growth of INPROD in combination with growth of INPROD. For more information, see the caption of figure 2.

weight and the number of forecasts can predict these different regimes. Thus, when the economic process is in a volatile regime, more experts produce forecasts than in more stable periods, and this coincides with our premises in section 2. This also fits the results of the first model presented in this section, where std predicts how volatile the variable of interest will be.

Another interpretation could be, that when there are not enough forecasts the mean or median of the forecasts is less reliable and informative and should not be used anymore or to a lesser extent. Whether the link between volatility of the dependent variable and the number of forecasts is a coincidence is not clear and should be investigated using other data sets.

The forecasting results of this model are quite convincing. The RMSPE is lower than the RMSPE's of all the benchmark models and even lower than the SPF mean and SPF median, see columns 1 and 4 of table 1. According to the standard test the difference is significant for all the benchmark models and according to the weighted tests the difference is also significant for SPF mean and SPF median, see the first column of table 4.

## 4.2 Industrial Production, fixed window, 40 forecasts

The results in the previous section are encouraging and support the notion that disagreement amongst forecasters might have predictive value. But how do these results

hold in different forecast situations? In this section we look at forecasts created using a fixed estimation window. All the models are estimated once and these estimated models are used to create forecasts for 40 consecutive quarters. RMSPE's can be found in the second and fifth column of table 1. For a few of the estimated models, information on test results can be found in table 5.

**Table 5:** This table shows if the models in the headers of the columns forecast significantly more accurate than the models in the headers of the rows, according to the tests as described in section 3.4. The models are estimated for INPROD data, a fixed estimation window is used and 40 forecasts are created. See for more information the caption of table 4.

| **INPROD Fix 40** | **MRSvar-med+nr-v** | **MRSvar-mean+nr-c** | **MRScons-med+nr-v** | **MRScons-mean+nr-c** | **MRSvar-med+95p-v** | **MRScons-med-95p-c** |
|---|---|---|---|---|---|---|
| | **Last release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | + |
| **ARX-mean** | | + | | + | | |
| **ARX-med** | | | + | | ++ | |
| **MRScons-mean-c** | ++ | ++ | ++ | ++ | ++ | ++ |
| **MRScons-mean-v** | ++ | ++ | ++ | + | ++ | ++ |
| **MRScons-med-c** | ++ | ++ | ++ | ++ | ++ | ++ |
| **MRScons-med-v** | ++ | ++ | ++ | + | ++ | + |
| **SPF-mean** | | + | | | | + |
| **SPF-med** | | + | + | + | + | + |
| | **First release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | | | | + | | |
| **ARX-med** | | | ++ | + | ++ | |
| **MRScons-mean-c** | | | ++ | ++ | ++ | ++ |
| **MRScons-mean-v** | | | ++ | ++ | ++ | ++ |
| **MRScons-med-c** | + | | ++ | ++ | ++ | ++ |
| **MRScons-med-v** | | | ++ | ++ | ++ | ++ |
| **SPF-mean** | | | | + | | + |
| **SPF-med** | | | + | + | ++ | + |

One conclusion that immediately follows from inspection of the RMSPE's, is that the forecasts created using a fixed estimation window are more precise than using a rolling estimation window. For six of the seven benchmark models the RMSPE's are lower and for the seventh model RMSPE's are approximately the same. Also the remaining models often produce more accurate forecasts in this case.

To see how these results compare with the results presented in the previous section,

we again discuss the four new variables one by one.

**Std**

For std, the RMSPE's of the two MRSvar-mean+std models are quite low and the RMSPE's of the model with a constant variance for the error terms is amongst the ten models with the lowest RMSPE's and has lower RMSPE's than all the benchmark models and than mean and median of SPF (see table 1). However, according to both the unweighted and weighted tests there is no evidence that these models produce significantly more accurate forecasts than the benchmark models and SPF mean and median. Table 5 does not show any test results for the MRSvar-mean+std models, because all test results were insignificant.

**5p**

For 5p we find even less evidence in this situation that the variable has predictive value. None of the models with 5p is capable of outperforming the benchmarks.

**95p**

This is different for 95p. The same model that performed well using a rolling estimation window also performs well using a fixed estimation model, namely a MRSvar-med+95p model. Only here the model that performs well has varying variance for the error terms and in the previous section we looked at the model with constant variance. However, the differences between these two models were very small in case of the rolling estimation window.

The estimation and interpretation of this model is much the same as shown in the previous section using the estimation of the model over the complete data set. The variance of the error terms is in the first regime around 0.9 and in the second regime around 1, so this is essentially the same as a MRSvar-med+95p-c model. The MRSvar-med+95p-c model did not perform well using a fixed estimation window,

because the estimation of the first of the 40 rolling window estimations (used to create the 40 fixed estimation window forecasts) did not converge to parameters similar to the complete data estimation and similar to most of the other 39 estimations. The first set of MRSvar-med+95p-v parameters estimated, was however similar to the parameters estimated over the complete data set and most of the other 39 estimations.

Another model that performs well, is the MRS model with constant transition probabilities, with explanatory variables the median and 95p and with a constant variance for the error terms. However, the interpretation for this model is hard, as the probabilities at the different regimes are often not close to zero or one and the coefficients in both regimes are often insignificantly different from zero.

**Nr**

For nr we find again multiple models capable of producing more accurate forecasts than at least part of the benchmark models. As for the variable 95p, part of the models we find capable of producing accurate forecasts using a fixed estimation window are similar to the models we find using a rolling estimation window. The MRSvar-mean+nr-c, MRSvar-mean+nr-v, MRSvar-median+nr-c and the MRSvar-median+nr-v all produce forecasts with RMSPE's lower than all benchmark RMSPE's and in most cases the differences are significant in case of last release data, see table 1, columns 2 and 5, and see table 5, columns 1 and 2. The estimated parameters and interpretations of these models are similar to that of the MRSvar-mean+nr-v estimation over the complete data set.

Also two MRS models with constant transition probabilities and nr as one of the explanatory variables perform well in forecasting, especially comparing to first release data, see table 1, columns 2 and 5, and see table 5, columns 3 and 4. In one regime the mean or median has a coefficient of around 0.8 and the coefficient of nr is insignificant. In the other regime, occurring less often, the mean or median has a coefficient of around 1.5 and nr a coefficient of around 0.02. This indicates that in specific situations

more forecasters signify a higher growth of INPROD, ceteris paribus.

## 4.3 Industrial Production, rolling window, 60 forecasts

If we create 60 INPROD forecasts by using a rolling window of 104 observations, we find that the variables std and 5p do have significant forecasting value, while now 95p and nr do not seem to contain significant forecasting information. See columns 3 and 6 of table 1 and see table 6.

**Table 6:** This table shows if the models in the headers of the columns forecast significantly more accurate than the models in the headers of the rows, according to the tests as described in section 3.4. The models are estimated for INPROD data, a rolling estimation window is used and 60 forecasts are created. See for more information the caption of table 4.

| **INPROD Roll 60** | **MRSvar-mean+5p-v** | **MRSvar-med+5p-v** | **MRSvar-med+std-v** | **MRSvar-mean+std-v** | **MRSvar-med+nr-v** | **MRSvar-mean+nr-v** |
|---|---|---|---|---|---|---|
| | **Last release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | + | | ++ | + | | |
| **ARX-med** | + | | ++ | ++ | | |
| **MRScons-mean-c** | | | ++ | ++ | | |
| **MRScons-mean-v** | + | | ++ | + | | |
| **MRScons-med-c** | | | ++ | | | |
| **MRScons-med-v** | ++ | | ++ | ++ | ++ | |
| **SPF-mean** | + | | | | | |
| **SPF-med** | + | | | | | |
| | **First release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | | | ++ | ++ | | |
| **ARX-med** | + | | ++ | ++ | | |
| **MRScons-mean-c** | | | ++ | ++ | | |
| **MRScons-mean-v** | ++ | | ++ | ++ | | |
| **MRScons-med-c** | | | ++ | | | |
| **MRScons-med-v** | ++ | | ++ | ++ | | |
| **SPF-mean** | + | | | | | |
| **SPF-med** | | | | | | |

From table 1 it can be seen that the ARX and MRScons models do not produce very accurate forecasts. In this situation, the MRSvar models are the only models capable of outperforming the benchmark models and SPF forecasts.

**Std**

The SPF mean and SPF median give again lower RMSPE's than all the benchmark models in this period. There are, however, two MRSvar models using std which create RMSPE's lower than these SPF mean and SPF median, namely MRSvar-mean+std-v and MRSvar-median+std-v. The MRSvar-mean+std models also produced relatively low RMSPE's using a rolling window to create 40 forecasts and using a fixed window to create 40 forecasts, but there we were not able to show that the improvement of the forecasts was significant and the mean and median of SPF produced lower RMSPE's. Here we find that MRSvar-mean+std-v significantly improves on almost all the benchmark models and has a lower RMSPE than the mean and median of SPF forecasts, although not significantly. The MRSvar-median+std-v is even more accurate. Interpretation of the models is again similar to the interpretation of the MRSvar model with std estimated over the complete data set and discussed in section 4.1.

**5p**

Using 5p, the models MRSvar-mean+5p-v and MRSvar-median+5p-v are amongst the ten models with the lowest RMSPE's. Especially the first one is interesting. It has the lowest RMSPE based on first and last release data, also lower than mean and median of SPF forecasts and the differences in RMSPE's are often significant. The interpretation is quite similar to the same model with std to predict regime switches. There is one regime that occurs less often than the other regime and that occurs around the same quarters as the second regime occurs in the model with std [9]. In that model the probability at a regime switch increases if std increases, but here this probability increases if 5p declines. The coefficient for the mean of SPF is different in both regimes and the standard deviation of the error terms is much higher in the second regime than in the first. So besides a higher std of survey forecasts, also a lower 5p is a signal that

[9] For example, in this model, the probability at regime 2 increases around the recessions in the 70s and the beginning of the 80s, like in the model with std, but also at the end of the recession in 1990-1991.

a different regime is coming up in which INPROD growth is more volatile and thus different weights should be used in forecasting models/ combinations.

**Nr**

MRSvar models in which nr is used to model the regime switches produce low root mean squared prediction errors too, but not significantly lower than the RMSPE's of the benchmarks.

## 4.4 Other variables, rolling window, 40 forecasts

Finally, we look at four other macroeconomic variables for which SPF forecasts are available. These are NGDP, PGDP, UNEMP and HOUSING. We discuss the four SPF statistics one by one again and see what their predictive value is for these four variables. See table 2 for RMSPE's and see table 7 for the test results on forecast accuracy of some of the models.

Before we start with std, note from the second and sixth column of table 2 that, as with INPROD, the mean and median of SPF forecasts for PGDP are more accurate than any of the benchmark models. For this data set these two simple forecasts seem even harder to beat than for INPROD, because also not one of the alternative models has a lower RMSPE. Therefore, it is maybe needless to say that none of the models improves (significantly) on SPF mean and median.

**Std**

There are four models interesting if we look at std (not making a distinction here between a constant and a varying variance for the error terms): MRSvar-med+std estimated over NGDP data, MRSvar-med+std estimated over HOUSING data and MRSvar-med+std and MRScons-mean+std estimated over PGDP data.

**Table 7:** This table shows if the models in the headers of the columns forecast significantly more accurate than the models in the headers of the rows, according to the tests as described in section 3.4. The models are estimated for NGDP, PGDP, UNEMP and HOUSING data, a rolling estimation window is used and 40 forecasts are created. See for more information the caption of table 4.

| **Roll 40** | **PGDP MRScons-mean+std-c** | **HOUSING MRSvar-mean+5p-c** | **UNEMP ARX-mean+95p** | **UNEMP MRScons-mean+95p-c** | **NGDP MRScons-med+nr-c** | **HOUSING MRSvar-med+nr-c** |
|---|---|---|---|---|---|---|
| | **Last release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | ++ | ++ | ++ | ++ | + | ++ |
| **ARX-med** | ++ | | | | + | + |
| **MRScons-mean-c** | ++ | ++ | | ++ | ++ | + |
| **MRScons-mean-v** | ++ | ++ | ++ | ++ | | + |
| **MRScons-med-c** | ++ | + | | + | ++ | + |
| **MRScons-med-v** | ++ | + | ++ | ++ | ++ | + |
| **SPF-mean** | | ++ | ++ | ++ | ++ | ++ |
| **SPF-med** | | ++ | ++ | ++ | + | ++ |
| | **First release** | | | | | |
| **AR** | ++ | ++ | ++ | ++ | ++ | ++ |
| **ARX-mean** | ++ | ++ | ++ | + | ++ | ++ |
| **ARX-med** | ++ | | ++ | + | | + |
| **MRScons-mean-c** | + | ++ | ++ | ++ | ++ | + |
| **MRScons-mean-v** | ++ | ++ | ++ | + | | + |
| **MRScons-med-c** | ++ | | ++ | ++ | | + |
| **MRScons-med-v** | ++ | | ++ | ++ | | + |
| **SPF-mean** | | ++ | ++ | ++ | + | ++ |
| **SPF-med** | | ++ | ++ | ++ | | ++ |

The first three models are very similar to the estimated MRSvar-mean+std model for INPROD data. In these estimated models one regime occurs when std rises and that is around periods declared as recessions by the NBER. For NGDP the intercept is much lower and negative in this regime, the coefficient of the lag is negative as opposed to around 0 in the first regime and the coefficient of the median of SPF is around 1.7 as opposed to 1 in the first regime. For HOUSING we see a much higher positive intercept in the 'recessionary' regime, a coefficient of the lag that does not differ much between the regimes and a coefficient of median of SPF that is around one in the first regime and around 1.2 in the second. For PGDP the differences between the regimes in coefficient estimation differ greatly over the 40 models estimated.

The fourth model that performs quite well, is MRScons-mean+std estimated for

PGDP. It is however hard to define where this success comes from exactly, as the 40 estimated models differ very much. Some only show one switch between the regimes, whereby regime one occurs the first half of the estimation period and the second regime the second half, where the median of SPF receives relatively more weight. Other models show multiple switches and different parameter estimates.

Although all these models are in the top ten of models with the lowest RMSPE's (see table 2), they in general do not improve significantly on the benchmarks. The model for NGDP is on average over last release data more precise than all benchmarks, but only for the AR model is the difference significant and the model is not more precise over first release data. The models for PGDP do not beat the mean and median of SPF, but are weighted or unweighted significantly more accurate than (almost) all benchmark models (see the first column of table 7). The RMSPE for the model for HOUSING is significantly lower than most benchmark models and than SPF mean and median, but is less accurate on average than three of the benchmark models.

**5p**

The most interesting models in which 5p is used are the MRScons-median+5p-c and MRSvar-mean+5p-c models for HOUSING. The first of these two shows much resemblance with the MRScons-mean-5p-c model estimated for INPROD, as the parameters have approximately the same estimated values (except the intercepts).

The MRSvar-mean+5p-c model estimates one regime again that occurs the least. The economic/ forecasting process switches to regime 2 when 5p decreases sharply and does not switch back as long as 5p does not increase much. The first regime has an intercept, a coefficient for the lag and a coefficient for the mean of around 1.6, 0.15 and 1.15, respectively and regime two of around -10, -0.1 and 0.55, respectively. So, one can safely say that regime two is a recession regime in which there is a large negative intercept, but in which changes in the mean of SPF forecasts should be used only partially. This model shows much resemblance with the MRSvar-mean+5p-v

estimated over a small horizon for INPROD, discussed in section 4.3 and with the same model estimated for INPROD with std to predict regime switches, discussed in sections 4.1 and 4.3.

Both models have a lower first release and last release RMSPE than all the benchmarks. In most of the cases this difference is significant, see column 2 of table 7.

Also for the other three variables there are models in which 5p is included that perform well in forecasting. These are mainly MRScons models. However, these models are in general not capable of significantly outperforming the benchmarks. Only for PGDP are these models capable of outperforming all benchmark models, but SPF mean and median remain superior here.

**95p**

For 95p, there are again multiple models worth analyzing. These are all ARX or MRScons models estimated for NGDP, PGDP or UNEMP data. The most interesting models are probably the models estimated for UNEMP, because these models are capable of outperforming the benchmarks significantly. We find for example an ARX model where mean has a coefficient of approximately 1.5 and 95p a coefficient of -0.35. So the higher the 'pessimists' predict unemployment, the lower the forecast should be for given values of mean SPF and previous unemployment. Stated differently, if there is a small group giving extremely high forecasts for unemployment, instead of a large group giving moderately high forecasts, there should be an adjustment in the final forecast. Also in the MRScons models 95p has a negative coefficient in both regimes, only in one regime more significantly than in the other. The regime with the more significant negative coefficient for 95p occurs more often and has a coefficient for mean or median SPF of above 1, while the other regime has a coefficient of around 1.

**Nr**

Finally, we pay some attention to the statistic nr. In each of the four data sets this statistic seems to have predictive value. For UNEMP and HOUSING we find, as for INPROD, MRSvar models that outperform all the benchmarks in some way, see the last column of table 7. The model that performs best for HOUSING, MRSvar-median+nr-c, is estimated 40 times almost exactly the same: one regime occurs if there are more than approximately 25 forecasters, has a constant of around 1.8, a lag coefficient of around 0.2 (significant) and a SPF median coefficient of around 1 (significant) and the other regime has a constant of around -2, a lag coefficient that is insignificant and a SPF median coefficient of around 1.7. The same kind of model that performs best for UNEMP, MRSvar-mean+nr-v, has two different sets of estimated parameters. One is similar to those for INPROD and HOUSING, with one minority regime occurring when nr is low. The other has a minority regime occurring when nr is high, above approximately 50.

For UNEMP we find that MRScons-mean+nr-v performs very well. In that model there is one regime with an insignificant coefficient for nr and one regime with a small, significantly positive coefficient for nr. When this coefficient is significant, the coefficient for the mean of SPF is much closer to zero. Also for NGDP and PGDP MRScons models with nr forecast accurately, see column 5, table 7.

# 5 Conclusions

Many studies have shown that the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia contains valuable information for forecasting macroeconomic variables. In our study we have shown that more forecast accuracy can be gained if not just the simple mean or median of SPF forecasts is used, but also if measures of disagreement amongst forecasters is used. All four new SPF statistics proposed, namely standard deviation, 5th percentile, 95th percentile and number of

forecasts, showed to contain useful information for forecasting, especially when used in Markov regime-switching models.

One of the interesting relationships that were found, is for example that the standard deviation of survey forecasts tends to go up around recessionary periods. Therefore, this statistic signals the beginning of an economic/ forecasting regime in which different weights should be used for the mean or median of SPF forecasts and the lag of the dependent variable. Also the 5th percentile and 95th percentile signal regime switches, but are also found to have predictive value in a linear way in specific regimes. Finally, the number of forecasts seems to be very useful, for example to predict regimes in which the mean or median should receive a substantially different weight relative to the lag of the dependent variable than in the other regime.

The results found in our basic analysis, using industrial production data, a rolling estimation window and creating 40 forecasts, could be generalized to other situations, but we found a lack of robustness. This may be due to the fact that the MRS model with varying transition probabilities between the regimes, is not easy to estimate, while it turns out that this model is necessary to include the forecasting information contained in the SPF statistics.

Correct estimation of the MRSvar models is easier if the proper starting values are used in the estimation procedure. As no information was available what these starting values could be, a grid of starting values was used. The results found in our study could be used in further research to choose more specific starting values or to use Bayesian techniques.

Furthermore, using a fixed estimation window does not seem such a bad idea in terms of forecasting accuracy according to the INPROD data. It might be a good idea to explore this method further. As only one model needs to be estimated in this case, more effort and time, for example by using a finer grid of starting values for the parameters, could be devoted to find an adequate MRS model.

Another idea could be to use other nonlinear time series models, such as smooth-transition models.

There are many issues unresolved in this paper and more research is needed on this. It is not clear for example if the number of forecasters predicts that a different economic regime is coming up or that it indicates that the mean or median of survey forecasters is less reliable and should be used in a different way or even not at all in forecasting models or forecast combinations. Furthermore, it might also be interesting to see if changes in survey forecasts have predictive value. If the average of the forecasts does not change much, but every forecaster changes their forecast substantially compared to previous period, does this say anything about the economic situation or about the accuracy of the survey forecasts?

# References

ARMSTRONG, J., "Combining Forecasts," in J. Armstrong, ed., *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Boston: Kluwer Acadamic Publishers, 2001).

AWIROTHANANON, T. AND W. CHEUNG, "On Joint Determination of the Number of States and the Number of Variables in Markov-Switching Models: A Monte Carlo Study," *Communications in Statistics Simulation and Computation* 38 (2009), 1757–1788.

BAILLON, E. AND A. CABANTOUS, "Combining Imprecise or Conflicting Probability Judgments: A Choice Based Study," ICBBR Working Paper Series No 2009_03, 2009.

CAPISTRÁN, C. AND A. TIMMERMANN, "Disagreement and Biases in Inflation Expectations," *Journal of Money, Credit and Banking* 41 (2009), 365–396.

CARROLL, C., "The Epidemiology of Macroeconomic Expectations," in L. Blume and S. Durlauf, eds., *The Economy as an Evolving Complex System, III* (Oxford: Oxford University Press, 2003).

CLARK, T. AND M. MCCRACKEN, "Test of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105 (2001), 85–110.

CLEMEN, R., "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting* 5 (1989), 559–583.

COOKE, R., *Experts in Uncertainty; Opinion and Subjective Probability in Science* (New York: Oxford University Press, 1991).

DIEBOLD, F., J.-H. LEE AND G. WEINBACH, "Regime Switching with Time-Varying Transition Probabilities," in C. Hargreaves, ed., *Nonstationary Time Series Analysis*

*and Cointegration*, Advanced Texts in Econometrics (Oxford: Oxford University Press, 1994).

DIEBOLD, F. AND J. LOPEZ, "Forecast Evaluation and Combination," in G. Maddala and C. Rao, eds., *Handbook of Statistics* (Amsterdam: North-Holland, 1996).

DIEBOLD, F. AND R. MARIANO, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics* 13 (1995), 253–263.

DOVERN, F., U. FRITSCHE AND J. SLACALEK, "Disagreement Among Forecasters in G7 Countries," Europen Central Bank Working Paper Series No 1082, 2009.

EINHORN, H. AND R. HOGHARTH, "Unit Weighting Schemes for Decision Making," *Organizational Behavior and Human Performance* 13 (1975), 171–192.

ELLIOTT, G. AND A. TIMMERMANN, "Optimal Forecast Combination under Regime Switching," *International Economic Review* 46 (2005), 1081–1102.

GIACOMINI, R. AND H. WHITE, "Tests of Conditional Predictive Ability," *Econometrica* 74 (2006), 1545–1578.

GRANGER, C., "Overview of Nonlinear Macroeconometric Empirical Models," *Macroeconomic Dynamics* 5 (2001), 466–481.

HAMILTON, J., "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica* 57 (1989), 357–384.

———, *Time Series Analysis*, chapter Modeling Time Series with Changes in Regime (Princeton, NJ: Princeton University Press, 1994).

HARVEY, D., S. LEYBOURNE AND P. NEWBOLD, "Testing the Equality of Prediciton Mean Squared Errors," *International Journal of Forecasting* 13 (1997), 281–291.

LAHIRI, K. AND X. SHENG, "Measuring Forecast Uncertainty by Disagreement: The Missing Link," *Journal of Applied Econometrics* 25 (2010), 514–538.

LAMONT, O., "Macroeconomic Forecasts and Microeconomic Forecasters," *Journal of Economic Behavior & Organization* 48 (2002), 265–280.

LASTER, D., P. BENNETT AND I. GEOUM, "Rational Bias in Macroeconomic Forecasts," *The Quarterly Journal of Economics* 114 (1999), 293–318.

MANKIW, G., R. REIS AND J. WOLFERS, "Disagreement about Inflation Expectations," in M. Gertler and K. Rogoff, eds., *NBER Macroeconomics Annual* (Cambridge: MIT Press, 2003).

MCCRACKEN, M., "Robust Out-of-Sample Inference," *Journal of Econometrics* 99 (2000), 195–223.

PSARADAKIS, Z. AND N. SPAGNOLO, "Joint Determination of the State Dimension and Autoregressive Order for Models with Markov Regime Switchin," *Journal of Time Series Analysis* 27 (2006), 753–766.

SMITH, A., P. NAIK AND C.-L. TSAI, "Markov-Switching Model Selection Using Kullback-Leibler Divergence," *Journal of Econometrics* 134 (2006), 553–577.

TIMMERMANN, A., "Moments of Markov Switching Models," *Journal of Econometrics* 96 (2000), 75–111.

VAN DIJK, D. AND P. FRANSES, "Selecting a Nonlinear Time Series Model Using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics* 65 (2003), 727–744.

WEBBY, R. AND M. O'CONNOR, "Judgemental and Statistical Time Series Forecasting: A Review of the Literature," *International Journal of Forecasting* 12 (1996), 91–118.

ZARNOWITZ, V. AND P. BRAUN, "Twenty-Two Years of the NBER-ASA Quarterly Economic Oulook Surveys: Aspects and Comparisons of Forecasting Performance,"

in J. Stock and M. Watson, eds., *Business Cycles, Indicators, and Forecasting* (University of Chicago Press, 1993).

ZARNOWITZ, V. AND L. LAMBROS, "Consensus and Uncertainty in Economic Prediction," *The Journal of Political Economy* 95 (1987), 591–621.