Forecasting inflation: An art as well as a science![§]

Ard den Reijer and Peter Vlaar[*]

January 2004

ABSTRACT

In this study we build two forecasting models to predict inflation for the Netherlands and for the euro area. Inflation is the yearly change of the Harmonised Index of Consumer Prices (HICP). The models provide point forecasts and prediction intervals for both the components of the HICP and the aggregated HICP-index itself. Both models are small-scale linear time series models allowing for long run equilibrium relationships between HICP components and other variables, notably the hourly wage rate and the import or producer prices. The model for the Netherlands is used to generate the Dutch inflation projections over a horizon of 11-15 months ahead for the eurosystem's Narrow Inflation Projection Exercise (NIPE). The recursive forecast errors for several forecast horizons are evaluated for all models, and are found to outperform a naive forecast and optimal AR models. Moreover, the same result holds for the Dutch NIPE projections, which have been provided quarterly since 1999. The direct and aggregation methods to predict total HICP inflation perform about equally good.

Key words: inflation, model selection, time series models
JEL codes: C32, C43, C52, C53

[§] We thank participants of the second Conference of the Euro Area Business Cycle Network, in particular Filippo Altissimo, Denise Osborn and Adrian Pagan, for useful comments.

[*] De Nederlandsche Bank, Research Department, P.O. Box 98, 1000 AB Amsterdam, The Netherlands, e-mail: a.h.j.den.reijer@ dnb.nl, p.j.g.vlaar@dnb.nl.

1 INTRODUCTION

The mandate of the European Central Bank (ECB) is to maintain price stability in the euro area. This goal is given a quantitative content by requiring that the year on year growth of the Harmonised Index of Consumer Prices (HICP) for the euro area as a whole should not exceed 2% in the medium term. The ECB is monitoring and forecasting price developments under the second pillar of its monetary policy strategy [1]. Therefore, forecasting inflation rates has become important for both monetary authorities and private agents who try to understand and react to the central bank´s behaviour. The aim of this paper is to describe the procedures used at De Nederlandsche Bank to predict Dutch HICP inflation and a new model to directly predict overall euro area inflation. The models forecast both the components of the HICP, as requested for the euro system's narrow inflation projection exercise (NIPE), and, for comparison reasons, the total HICP itself. The forecast horizon is eleven to eighteen months ahead.

One of the issues that is addressed in this paper is that of aggregation. In the European context, aggregation can be considered in two dimensions. First, there is the aggregation of the forecasts of individual countries to a euro area level, see for instance Marcellino *et al.* (2003). Second, there is the aggregation of HICP component forecasts. In this study we will only address the latter issue [2]. Obviously, there is a clear interest in finding out whether aggregating component forecasts performs better than forecasting the aggregate directly. Aggregating forecasts of component models is potentially beneficial as forecast errors might cancel between components. Moreover, the disaggregate components can be better modelled by choosing a more suitable model for each component separately and by possibly incorporating additional explanatory variables. This argument is indeed apparent in theoretical models, see Lütkepohl (1987) and Hendry & Mizon (1999). The latter authors assume for

[1] In practice, each country provides four times a year its own inflation forecast for an horizon of 11-15 months and these forecasts are used to construct an area wide forecast. This periodic procedure is called the Narrow Inflation Projection Exercise (NIPE).

[2] In contrast, the eurosystem's NIPE creates a euro area forecast by aggregating the individual countries' inflation forecasts. So, the model for the euro area built in this study generates forecasts for the area wide aggregates, while the NIPE aggregates the individual country´s forecasts to the euro area level.

instance a known and constant data generation process. However, Hubrich (2001, 2003) finds empirical evidence for euro area data and across various specifications that directly forecasting the aggregate HICP performs better than aggregating the forecasted components, especially for a forecast horizon up to 12 months ahead. Generally, the relative performance depends on many different aspects: a univariate or multivariate forecasting method, the number and choice of variables included, the model selection procedure, the sample size, the degree of dependence between the disaggregated series and the forecast horizon. Apart from the possible efficiency gain, a policy maker also has an interest in forecasting the HICP components to construct a measure for core inflation, defined as HICP excluding the components unprocessed food and energy. These two components are generally considered more volatile and less susceptive to monetary policy.

Another central issue in this paper is the model selection procedure, for which a new heuristic method is developed. The method involves three steps. The first step involves the visual inspection of the data, primarily to detect changing seasonal patterns. This step determines the model structure, either a VARX model in first differences or a VECMX model in first and twelve month differences. The second step involves calculating all possible models given this structure, allowing for a small set of exogenous and endogenous variables and variable lag lengths. Optimal statistical models are subsequently selected according to goodness-of-fit, parsimony and/or out-of-sample forecasting accuracy. The final selection is based on the economic evaluation of the statistically selected models. Especially, the long run properties are important in this respect.

Although we are aware of the growing popularity of factor models, developed by Forni *et al*. (2001), for forecasting, we decided to investigate only standard linear time series models[3]. Theoretically, factor models can efficiently summarise characteristics of large data sets. The possibility to include more information for forecasting purposes should improve the forecasting quality. However, empirical

[3] Applied research on forecasting with factor models is currently done at, for instance, the European Commission, Banca d'Italia, Oesterreichische Nationalbank, Moser *et al*. (2002) and the Banque de France, Bruneau *et al*. (2002).

research indicates ambiguity in this respect. Banerjee & Marcellino (2002), for instance, find that standard time series models fitted on a large data set by an automated procedure turn out to outperform factor models. The same even holds for single leading indicator models, although their performance seems less stable. The main result from the applications of new techniques seems to be that going beyond standard models can be beneficial, but not necessarily so. As most non-linear models are likely to be more sensitive to structural breaks and less applicable for short time periods, we will restrict ourselves in this paper to multivariate linear time series models.

The rest of this paper is organized as follows. In section 2 we describe our procedure to select an optimal forecasting model. The selected models for the Netherlands and the euro area are described in section 3. Section 4 elaborates on the uncertainty surrounding the forecasts and how to measure and interpret this uncertainty. In section 5 the forecast results of the models are evaluated. First, the recursive root mean squared forecast errors are computed, both for the component models and the direct HICP models, given the realized values for the exogenous variables. Then, the Dutch NIPE results are evaluated. Finally, section 6 concludes.

2 MODEL SELECTION PROCEDURE

The model selection procedure consists of three steps. The first step involves preliminary data analyses. This step is used to select the optimal model structure. That is to say, either a VAR model in first differences or a vector error correction model in both first and twelve month differences. The second step involves the computation of all possible models given the possible set of exogenous and endogenous variables and the allowable lag length. The optimal models according to several statistical criteria are subsequently shown. The last step implies the economic evaluation of the statistically selected models in order to select the final optimal model. This involves both an economic interpretation of the coefficients, primarily with respect to the error correction term, and an analysis of the stability of the model forecasts.

2.1 **Preliminary data analysis**

The sample period runs from 1987(10) and 1990(1) onwards for the Netherlands and the euro area respectively. The notation of the HICP components reads as: HICP unprocessed food P^{uf}, HICP processed food P^{pf}, HICP industrial goods excluding energy P^{i}, HICP energy P^{e}, HICP services P^{s} and the HICP index P^{total}. This last variable can also be constructed by contemporaneously weighting the sum of the disaggregated HICP components:

$$P_t^{agg} = \omega_t^{uf} P_t^{uf} + \omega_t^{pf} P_t^{pf} + \omega_t^{i} P_t^{i} + \omega_t^{e} P_t^{e} + \omega_t^{s} P_t^{s} \qquad (1)$$

where t=1,…,T, T+1,...,T+n for the sample with size T and forecast horizon n. The ω's are the aggregation weights. The 2002 weights for P^{uf}, P^{pf}, P^{i}, P^{e} and P^{s} are 6,8%; 13,7%; 30,6%; 9,2%; 39,7% and 8,1%; 12,3%; 32,0%; 8,7%; 38,9% for the Netherlands respectively the euro area. In order to mitigate the aggregation error in this chain index, these weights refer to price indices that are rebased to 100 in December the year before[4]. These numbers are updated yearly in January, but are assumed fixed over the forecast horizon. Two different notations are in use, namely P^{total} and P^{agg}, for the HICP index in order to distinguish between the forecasts generated by applying (1), P^{agg}, for t=T+1,…,T+n and the forecasts generated by a model for the index, P^{total}. These latter forecast results act as a robustness check of the former ones. Although applying (1) to the component forecasts results in consistent HICP index forecasts it is not obvious that P^{agg} contains better forecasts than P^{total}.

The forecasts of natural gas prices (part of P^{e}) and housing rents (part of P^{s}) for the Netherlands are generated outside the model as they are only adjusted twice respectively once a year, according to some strict rules. Consequently, these elements can better be predicted using institutional knowledge

[4] Instead of rebasing the price indices we divide the original weights by the December year before value of the sub-price index and multiply it by the December year before value of the aggregate price index. These rebased weights do not necessarily sum to exactly one.

than with a statistical model. Also, radio and TV licences for the Netherlands are excluded, as their abolishment in January 2000 had a huge impact on services inflation.

The HICP data are not seasonally adjusted. They are plotted in the appendix, figure A1 and figure A2 for respectively the Netherlands and the euro area. All series in both cases are plotted in raw format, in annual inflation rates, that is 12 month differences of the log-transformed HICP and in the monthly change of the log price index, that is monthly inflation. All sub-indices except P^{uf} and P^{e} show for both areas a relatively steep upward tendency. The sub-indices unprocessed food and energy prices show a rather erratic development.

A visual inspection of the data shows that especially the (log) sub-indices for non-energy industrial goods and services as well as the one for total HICP are subject to a changing seasonal pattern. Particularly for the euro area but also for the Netherlands, this change is not concentrated in just one month. Otherwise, a second set of seasonal dummies should be sufficient to remove the seasonality. The changing seasonal pattern is confirmed by performing a regression analysis of the HICP sub-indices on seasonal dummies. The pattern is filtered out of the data by taking 12 month differences. Moreover first difference are taken to eliminate the (near) unit root in inflation. However, lagged annual inflation is also included to allow for either stationarity of annual inflation, or cointegration with other variables. The same model is used for processed food for the Netherlands, for which the seasonal pattern is less clear. Concerning the series P^{uf}, P^{e}, and P^{pf} for the euro zone, seasonal dummies are included in the model to capture the seasonal effects. As these variables are clearly non-stationary, they are modelled in first differences. Cointegration is not allowed for these models in first differences as it appears that the economic rationale for a long run relationship among price levels is less obvious than among inflation rates. Moreover, even if such a relationship would exist, it is prone to structural breaks due to for instance indirect tax adjustments.

A set of additional variables is potentially included in the models to improve the forecasting performance. This set of explanatory variables is both motivated by economic theory and data availability and captures key variables that influence the inflation rate. The set of explanatory variables consists of endogenous and exogenous variables. The former ones are endogenously forecasted by the model while the latter enter the model exogenously. In the NIPE-exercise practice, future paths for policy variables and the variables regarding the external environment of the EMU are provided exogenously by the ECB. With respect to interest rates and exchange rates a no-change path is implemented, whereas futures are used to project oil and other commodity prices. Besides these variables nominal wages are also included exogenously. Although prices and wages are clearly related, we prefer to include wages exogenously as overlapping contracts and other institutional features make them relatively easy to predict in the short run. Potential endogenous variables include producer prices, import prices, the nominal money stock, industrial production, credit data, retail turnover and business cycle indicators. All variables except interest rates and business cycle indicators are in logarithms and enter the respective model in the same way as the concerning subindex; mostly by taking the change of the 12 month differences and sometimes by simply taking only the first difference. The interest rates enter in accordance with the annual inflation rate. The exogenous and endogenous variables that are currently selected are displayed in raw format in figures A3 respectively A4 of the appendix [5]. These are the exogenous variables Euro/Dollar exchange rate ($e^{€\$}$), the oil price measured in euros (P^{oil}), the hourly nominal wage rate ($wages^{EU}$) and ($wages^{NL}$), and the world market prices of commodities exclusive energy, also in euros (wmp^{exe}), and the endogenous variables producer prices (P^{prod}) for the euro zone and the import price index of Germany (Pim^{GE}) for the Dutch models[6].

2.2 **Statistical criteria**

This research considers two types of linear time series models, namely a Vector AutoRegressive model (VAR) in first differences for the categories unprocessed food, energy and for the euro zone processed food, and a Vector Error Correction Model (VECM) in both first and twelve month

[5] See table A1 for detailed information on the sources.

[6] German import prices are used as no Dutch import price index is available at a monthly frequency and with a reasonable publication lag.

differences for the other indices. The differencing is dictated by the unstable seasonal patterns of the data, as described in the previous subsection, and consequently does not depend on the outcome of (seasonal) unit root tests. Given this structure, all possible VAR models given the set of explanatory variables are computed in an automated selection procedure. That is to say, a model is computed for every possible combination of explanatory variables and every lag length up to a maximum (usually 12). The lag length is the same for all the variables included. Moreover, for the VECM models every possible combination of lagged twelve month differences is added to check for long run relationships (and to correct for possibly incorrectly imposed unit roots in inflation). For each model, a number of selection criteria is computed and the best models according to the different criteria are subsequently shown.

Certain limitations are imposed in the search procedure, however. In order to limit the number of variables in the models, a maximum number of exogenous, endogenous or total number of variables can be imposed. Moreover, the number of parameters as a fraction of the total degrees of freedom in the model can be restricted. Besides these quantitative restrictions, certain limitations are imposed on the error correction term (the lagged variables in twelve month differences). First of all, if a long term relationship is selected, the lagged inflation level itself should always enter[7]. Second, all the variables in the error correction term should have the expected correct sign in the inflation equation.

With respect to the selection criteria, the VAR literature usually looks at three different criteria: Schwarz (*SC*), Hannan-Quinn (*HQ*) and Akaike (*AIC*). These criteria aim at a parsimonious model by trading off the fit of the model against losing degrees of freedom. The goodness-of-fit will always improve as one includes more lags, which results in a reduction of the sum of squares of the residuals. Incorporating more lags implies including more parameters, which results in loosing degrees of

[7] The selection of the long run relationship is based on the same criteria as the variable selection and the lag length. No formal cointegration rank tests are performed. As the left hand side variable (the monthly change in inflation) is clearly stationary, the twelve month differences will only enter if they are indeed cointegrated or stationary themselves.

freedom. The difference between the criteria lies in the severity of the parameter penalty, with the Schwarz criterion giving the biggest punishment for extra parameters and the Akaike the lowest:

$$AIC = T\ln(|\hat{\Omega}|) + 2k \tag{2}$$

$$HQ = T\ln(|\hat{\Omega}|) + 2k\ln(\ln(T)) \tag{3}$$

$$SC = T\ln(|\hat{\Omega}|) + k\ln(T) \tag{4}$$

where

T	=	size of the sample,
k	=	number of parameters estimated in all equations,
$\lvert\hat{\Omega}\rvert$	=	determinant of the covariance matrix of the residuals of the estimated systems.

Unfortunately, the fact that the three criteria use a different penalty usually also implies that they prefer different models. Besides this practical problem, there are at least three reasons why the criteria can hardly be used directly to select the best model for our purposes. First, these criteria are primarily used to determine the optimal lag length of a given VAR model. In our case, not only the lag length, but also the optimal selection of variables to include has to be chosen. Different models can hardly be evaluated on the basis of covariance matrices, especially if models with different endogenous variables are compared, as the residuals of different equations are computed. As we are primarily interested in the inflation equation in each model, we replace the covariance matrix and the number of parameters of the system by the ones of the inflation equation alone.

A second problem with the information criteria is that these criteria are essentially in-sample. However, the best fitted model in-sample is no guarantee of being the best forecasting model out-of-sample. Within the forecasting literature, an alternative measure of fit has become popular, namely the root mean squared forecast error (RMSFE) defined as:

$$RMSFE = \sqrt{\tfrac{1}{n}\Sigma_{i=1}^{n}(x_{T+i} - \tilde{x}_{T+i})^2}, \tag{5}$$

where

n = number of forecasts,

$\tilde{x}_{T+i}$ = forecast of the variable at time $T+i$,

x_{T+i} = realisation of the variable at time $T+i$.

In order to be able to apply this criterion, a certain fraction of the data is not used in the estimation of the model. These data are used for the out-of-sample evaluation. We will also use this criterion as one of our selection criteria. A problem with this criterion is however, that given our relatively short sample, only a short out-of-sample period can be used to evaluate the model. It might very well be the case that during this short sample exceptional circumstances occurred that are less representative for future developments than experiences in the sample period. So, it seems hardly efficient to ignore the in-sample fit.

A third reason why our specific needs require adjustment of the information criteria is that we want to compare both models with endogenous and with exogenous variables. Exogenous variables have the advantage that they provide a solid anchor for future inflation, especially if they are included in the long run relationship. Christoffersen and Diebold (1998) show that error correction terms among endogenous variables alone do not help to produce better long run forecasts as they have expectation zero in the long run. Exogenous variables in an error correction term on the other hand do positively affect long run forecasting as they steer the long run outcome for the endogenous variables. In order to distinguish between the two, the forecasting performance of all endogenous variables has to be taken into account, which requires the inclusion of an out-of-sample period. Of course, the anchor function of exogenous variables is particularly useful if the future path of the exogenous variables can indeed be predicted with reasonable accuracy. However, even if an exogenous variable is very volatile (for instance the oil price), a forecast based on this variable might be preferred to one based on a hard to predict endogenous variable as the inflation forecast based on the exogenous variable still has a clear interpretation as a *conditional* forecast.

In order to bring the forecast performance into the selection criteria, we have withheld about 20 months of observations from the estimation sample during the selection process. The forecast errors for inflation are incorporated into the information criteria by replacing $\left|\hat{\Omega}\right|$ in (2) to (4) by a weighted average of the in-sample variance of the inflation equation and the out-of-sample forecast error variance. Apart from these mixed in-sample/out-of-sample criteria, the purely in-sample criteria are also computed. Moreover, the root mean squared (forecast) errors are also computed for both the in-sample and the out-of-sample period and a weighted average [8].

This procedure gives us 9 possibly different models, based on in-sample information criteria; SC^{in}, HQ^{in}, AIC^{in}, $RMSFE^{in}$, mixed in-sample and out-of-sample information criteria; SC^{mixed}, HQ^{mixed}, AIC^{mixed}, $RMSFE^{mixed}$, and a purely out-of-sample criterion $RMSFE^{out}$. As to be expected, the nine different criteria often give nine different optimal models. In principal, the model selected by AIC^{mixed} is chosen. The relatively low penalty for extra parameters for this criterion is justified as the risk of overfitting is mitigated by the inclusion of the out-of-sample forecast variance. The other models might give important indications with respect to the preferred specification as well, however. The fact that the model selection choice is not robust with respect to the selection criterion puts some doubts on the existence of *the* optimal model. This is further confirmed by a periodic evaluation of the results. One more year of data often leads to different selected models.

2.3 **Economic evaluation**

Given that different criteria prefer different models, and the fact that these choices are not very robust with respect to the addition of more data, it is obvious that statistical criteria alone are hardly sufficient to select the optimal model. Judgmental issues, based on economic criteria are important as well. Here,

[8] The in-sample variance is hereby computed with a correction for the number of estimated parameters in order to get an unbiased estimate. Otherwise extra variables can only improve the result.

three issues come to mind. First, does the choice of variables make sense? Second, are the parameter values in the model of the right order of magnitude? Third, does the model provide stable forecasts?

With respect to the variable selection, we do not have strong priors, but for instance in the energy model we prefer the oil price above a model with commodity prices, even if the latter includes oil. For services inflation on the other hand, the oil price does not seem to be an obvious candidate.

Regarding the order of magnitude of the coefficients, most attention is concentrated on the error correction term. In the automatic model selection procedure, all variables in the error correction term are already checked for their sign. Apart from that, implausible long run elasticities might be remedied by slightly adjusting the model (for instance changing the lag length). Apart from the error correction term other coefficients might draw attention as well. For the Netherlands for instance, both the within sample and the mixed criteria selected the oil price as an important variable for services inflation. In the estimated models this was reflected in a very significant negative contemporaneous coefficient. As there is no economic rationale for such a negative impact, the oil price was not included in our preferred model. The significance was probably due to an incidental correlation of outliers in the past. Indeed, according to model selection criteria based on the current data set, the oil price would no longer be selected by the mixed criteria, but the within sample criteria would still select it.

Probably, the most important economic criterion in the evaluation of models, is the ability to provide stable forecasts. Here, the difference between endogenous and exogenous variables is essential. Including endogenous variables that are themselves hard to predict might lead to the drifting of inflation forecasts to unlikely regions, especially if this endogenous variable is included in the error correction term. Therefore for instance, the variable M3 is not allowed to appear in the cointegration relationship. This endogenous explanatory variable is very difficult to forecast over a longer horizon in this VAR setting, and a bad forecast would imply a severe bias in the long run forecast of inflation. For wage developments on the other hand, the opposite holds. This variable is imposed in the error correction term of the selected model if validated by the data. Wage development is an exogenous explanatory variable, which is well predictable due to sluggishness in the wage formation process. Due to its imposed prominence, wages act as an anchor of the model. Another anchor is formed by import prices although they are endogenous. However, the import prices themselves are well predictable from (exogenous) oil price and exchange rate developments.

3 THE EMPIRICALLY OPTIMAL MODELS

We applied the selection criteria on a sample running from 1987(10) and 1990(1) until 2002(8), respectively, for the Netherlands and the euro area. The sample is split into a part used for the in-sample information criteria and a part used to evaluate the out-of-sample forecasting performance. The in-sample errors are calculated from the model based on the sample up until 2000(12) and the forecasting errors are obtained from the sample 2001(1) onwards. The model selection is therefore based on the number of 20 forecasts. As stated before, apart from the components unprocessed food and energy as well as processed food for the euro area, all models are specified in changes of 12 month differences. So, for most models this differencing implies a loss of 13 observations and a remaining sample size of T=166 and T=139 for respectively the Netherlands and the euro area. The sample for fitting the model is much larger than the sample for obtaining the out-of-sample forecast residuals. Although these forecast errors are less numerous, they get a weight of 0.4 for the mixed criteria, so as to emphasise the importance of good forecasting performance. We acknowledge that the forecasting track record of specific models and leading indicators are not invariant over time. Checking the robustness of the model specification by evaluating out-of-sample forecasts for different time periods is a difficult job due to small sample availability. Moreover, the different criteria produce different optimal models. Therefore, the model selection procedure is rerun regularly.

The selected models for the 5 components and the HICP-index are presented in table 1 for the Netherlands and in table 2 for the euro area. The optimal model for the component unprocessed food turns out to be a univariate random walk for the Netherlands and an AR(1) process for the euro area (both including seasonal dummies). The component energy price depends mainly on the oil price, and in the euro area also on the producer price. The most dominating explanatory variable for the other sub-indices is the wage rate that is selected in all cointegration relationships for both the Netherlands and the euro area. It forms an important ingredient for the long run properties of the models. Wages tend to be more important for the Netherlands than for the euro area as revealed by the twice as high Dutch long term coefficients. Besides wages, a relatively dominating leading indicator for the Netherlands is the import price index of Germany showing up in the cointegration relationship for all four indices. For the euro area on the other hand, the producer price index is taking this role. These endogenous variables are themselves primarily driven by the oil price, the Euro/Dollar exchange rate and the commodity prices excluding energy. Finally, unprocessed food prices turn out to be important for processed food and services prices in the euro area. The latter impact can be explained by restaurant prices. The endogenous unprocessed food prices in these models are forecasted according to the same specification, even though the optimal model for this variable seems to be an univariate AR(1) model. Using different models for unprocessed food might reduce positive correlation among forecast errors of these three components of HICP.

Table 1 Netherlands: HICP (sub)indices

HICP-index	P^{uf}	P^{pf}	P^{i}	P^{e}	P^{s}	P^{total}
Exogenous	-	$wages^{NL}$, $e^{\text{€}\$}$	$wages^{NL}$, P^{oil}	P^{oil}	$wages^{NL}$, $e^{\text{€}\$}$	$wages^{NL}$, $e^{\text{€}\$}$, P^{oil}
Endogenous	-	Pim^{GE}	Pim^{GE}	-	Pim^{GE}	Pim^{GE}
EC term	-	P^{pf} $wages^{NL}$ Pim^{GE}	P^{i} $wages^{NL}$ Pim^{GE}	-	P^{s} $wages^{NL}$, Pim^{GE}	P^{total} $wages^{NL}$ Pim^{GE}
Number of lags	0	1	0	0	1	0
Specification*	Δ_1	$\Delta_1\Delta_{12}$	$\Delta_1\Delta_{12}$	Δ_1	$\Delta_1\Delta_{12}$	$\Delta_1\Delta_{12}$

*Δ_x is defined as the x month difference of the variable. The error-correction (EC-)term is specified in annual inflation rates. The models in first differences include seasonal dummies.

For both areas, the small number of lags selected for all models is noticeable. In previous specifications lag lengths of up to 12 were included, but it seems that the few significant coefficients with longer lags are not very stable. Consequently, the models with longer lags produced more volatile forecasts, which did not match very well with the subsequent realisations. Over the latest sample, especially the selection criteria with relatively strong penalty for extra parameters suggested at most one lag.

Table 2 Euro area: HICP (sub)indices

HICP-index	P^{uf}	P^{pf}	P^{i}	P^{e}	P^{s}	P^{total}
Exogenous	-	$wages^{EU}$	$wages^{EU}$, $e^{€\$}$	P^{oil}	$Wages^{EU}$	$wages^{EU}$, $e^{€\$}$, P^{oil}, wmp^{exe}
Endogenous	-	P^{uf}	P^{prod}	P^{prod}	P^{uf}	P^{prod},
EC term	-	-	P^{i}, P^{prod}, $wages^{EU}$	-	P^{s}, P^{uf}, $Wages^{EU}$	P^{total}, P^{prod}, $wages^{EU}$
Number of lags	1	1	0	1	0	0
Specification*	Δ_1	Δ_1	$\Delta_1\Delta_{12}$	Δ_1	$\Delta_1\Delta_{12}$	$\Delta_1\Delta_{12}$

*Δ_x is defined as the x month difference of the variable. The error-correction (EC-)term is specified in annual inflation rates. The models in first differences include seasonal dummies.

Interest rates and money are never included in the models. For the euro area, M3 was selected initially in the optimal model for services according to the AIC^{mixed} criterion. However, as forecasting this endogenous variable showed to be rather difficult in this VECM setting, the variable was skipped. The inflation projections including this variable turned out to be too volatile, with sometimes very unlikely outcomes as a result. The short term interest rate is the instrument used by most central banks to fight inflation. This variable is widely acknowledged to exert an influence on inflation with a transmission lag of more than one year. This probably explains the absence of the short term interest rate in the model selection given the restriction not to include more than twelve lags due to data availability. Due to the lag in monetary transmission, the central bank's interest rate actions to fight inflation might create a positive short term relation between high nominal and real interest rates and an inflationary environment and *vice versa*.

4 FORECAST UNCERTAINTY

The constructed models aim at providing forecasts for the inflation rates in the short to medium term. In order to do so, we used all available information to estimate the models. The predicted inflation numbers are called point forecasts and they represent the most likely future outcomes. These numbers are, however, to some extent uncertain. Garatt *et al*. (2003) state that all model-based forecasts are subject to four types of uncertainty: 1) measurement uncertainty, that is data inadequacies and measurement errors, 2) model uncertainty, that is the uncertainty whether the specified model correctly describes the structure of the inflation process including uncertainty surrounding policy reactions, 3) parameter uncertainty given a correctly specified model; this concerns the robustness of forecasts to the choice of the parameter values and 4) future uncertainty, that is the occurrence and the effects of unexpected shocks.

Forecasting models cannot be expected to predict unexpected shocks. In the first half of 2001 we have experienced inflationary pressure on the unprocessed food index caused by the two animal diseases (BSE and Foot-and-Mouth disease) and quite recently a poor harvest in southern Europe due to bad weather conditions. Upward effects on industrial goods prices came from the depreciation of the euro and at a later stage from the almost doubling of oil prices. The introduction of the euro notes and coins in January 2002 triggered a price increase to partly finance this huge operation. Finally, changes in value added taxes induce incidental price changes. While forecasting models cannot foresee unexpected shocks, foreseeable institutional changes or base effects (exceptional shocks during the current year) are incorporated in the predictions by adjusting the model based forecasts ex-post.

Wages, the exchange rate and the oil and commodity prices are exogenous explanatory variables in our models by assumption and have therefore to be predicted outside the model [9]. The uncertainty surrounding these exogenous predictions is not created by the model and will not be reflected in the

[9] In practice, these predictions equal the assumptions of the Eurosystem’s BMPE.

quantification of this uncertainty [10]. We will supplement the point forecasts generated by the models with prediction intervals that provide a quantitative content for the uncertainty surrounding them. So, both the point forecasts and the prediction intervals should be interpreted as conditional on the assumptions. In the same spirit, the Bank of England quantifies uncertainty by publishing [11] density forecasts. A density forecast is an estimate of the complete probability distribution of the possible future values of a variable [12]. In this study we will use non-parametric bootstrapping to construct such a probability distribution and calculate the corresponding prediction intervals. The basic idea underlying non-parametric bootstrapping is that one performs a simulation experiment in which the error terms are not drawn from an assumed distribution, such as the normal one, but rather from the actual residuals of the estimated models.

The bootstrap involves the following steps:

- Estimate the models for the five sub-indices using the same maximum sample period for all five models.
- Compute the residuals of all five models (both for inflation, and the other endogenous variables) and put them in one matrix. The rows represent the different months in the estimation sample, and the columns all endogenous variables (endogenous variables that are included in more than one VAR model are included more than once).
- Compute bootstrap forecasts by repeating the following steps 1000 times:

1 construct recursively new endogenous variables for the estimation sample given the original coefficients, starting values, exogenous variables and randomly selected rows (with replacement) of the residual matrix.

2 compute new coefficients for the five models with this new data.

[10] For this reason the word ‘projection’ is used to indicate that the forecast is actually conditional on exogenous assumptions.

[11] The Bank of England has published a density forecast of inflation in its quarterly *Inflation Report* since February 1996.

[12] Density forecasts are more common practice in the fields of finance and risk-management. For instance, the widely used Value at Risk-measure is defined as the α-th percentile of the predicted probability density.

3 compute forecasts recursively for every subindex as expected values given new coefficients plus random rows of residual matrix (to reflect future uncertainty).

- Compute the aggregated HICP forecast for every bootstrap sample.
- Sort these 1000 bootstrap forecasts in ascending order.

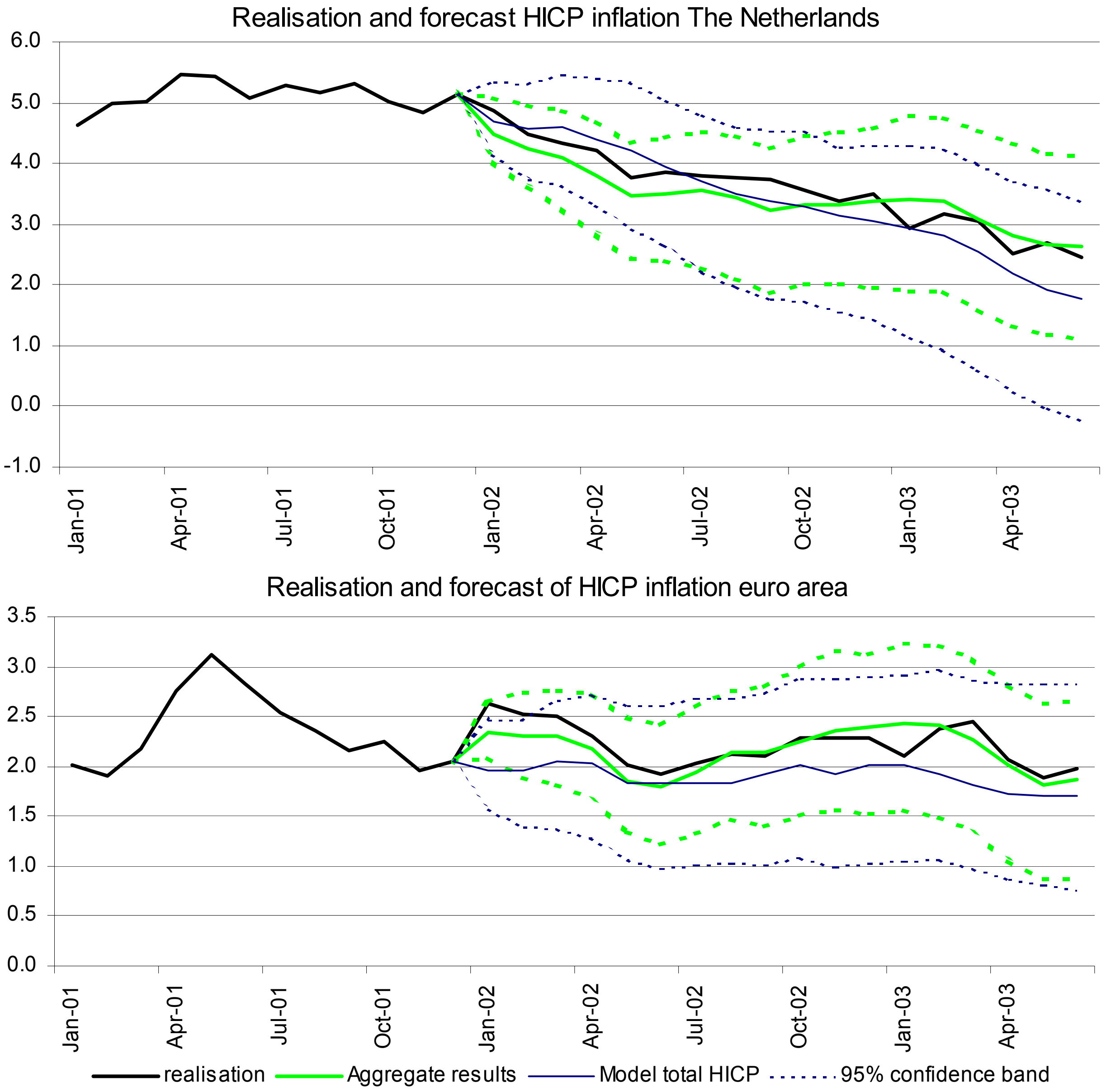
The reason for storing the residuals of all models in the same residual matrix (to use the residuals of the same month in all five models) is that the residuals of the five models are clearly not independent contemporaneously. If the bootstrap procedure would be computed for all five categories separately, the overall HICP confidence band would become narrower due to the neglected positive correlation among components.

The confidence intervals are generated by the percentile method. This means ranking the generated forecasts for each month of the forecasting horizon and taking the 2.5th and the 97.5th percentiles. So, exactly 2.5% of the bootstrapped forecast replications are below the lower prediction bound and 2.5% are above the upper prediction bound.

The bootstrap procedure is merely applied to provide prediction intervals, which quantify the uncertainty surrounding the point forecasts. The Bank of England does a different job by fitting a two-piece normal distribution and they derive point forecasts and prediction intervals analytically, see for instance Wallis (1999). This two-piece normal distribution is an asymmetric one with a fat upper tail indicating that inflation outliers are dominantly on the upside. The asymmetry triggers a debate on the way of presenting the inflation forecasts since the mode, the median and the average of an asymmetric distribution do not coincide.

Moreover, one can choose a prediction interval which covers the stated proportion in the centre of the distribution with equal probability in each of the tails. Alternatively, one can choose a prediction interval to be the shortest possible for the assigned probabilities, which for an asymmetric distribution implies a shift away from the fatter tail. The Bank of England chooses the mode of the density forecast as the point forecast. Moreover, they choose the shortest possible prediction intervals given the assigned probabilities. This implies that the probability that inflation will lie above the interval is higher than the probability that inflation will lie below the interval.

Figure 1 Realisation and forecast of HICP inflation for the Netherlands and the euro area



For our models, these choices are not important as the bootstrap distribution turns out to be fairly symmetric, see Figure 1. For both models, the model forecasts are in the centre of the confidence band. Consequently, the bootstrap median and mode are almost identical to its mean. In the past, for the Netherlands we sometimes found that the point forecast was not in the middle of the confidence band. Also in these cases, however, the bootstrap distribution turned out to be almost symmetrical, but the bootstrap mean did not coincide with the point forecast of the model. The difference between the two was probably due to a bias in the AR-coefficients. The bias correction methodology of Kilian (1998), which implies using the bootstrap twice, might reduce the difference between the two under those circumstances.

Figure 1 shows the forecasts for the Netherlands and the euro area over 2002 and the first half of 2003, based on the models given in Table 1 respectively 2 and the realisations for the exogenous variables. For both areas, two forecasts and corresponding confidence bands are given: one based on aggregation of the models for the components and one based on the model for total HICP inflation. Although the models are estimated with data up until 2001, this exercise is not fully out-of-sample in the sense that the data for 2002 was previously included to select the optimal models. Despite these facts, the realised inflation in January 2002 for the euro area was clearly above the prediction interval for the total HICP model and just on the border for the aggregate model. The most likely reason for this underestimation of inflation seems to be the cash changeover to the euro. This event is not incorporated in the forecasts. Especially services prices have risen more than predicted. Current evidence indeed seems to suggest that the services sector is affected most by the changeover with a price impact about 0.3 to 0.6 percentage point in 2002. Also for the Netherlands, realised inflation in the beginning of 2002 was higher than expected, though not significantly so. Although the size of this effect was of course hard to foresee, some impact could be expected, if only to offset the costs related to the cash changeover. The significance of the effect for the euro area is probably due to the fact that this ‘shock’ is experienced in all countries at the same time, whereas most of the time incidental factors only affect a subset of countries. This example clearly shows the importance of including institutional knowledge into the forecasts. Besides the services prices, food prices contributed to the higher than expected rise in inflation in January 2002. These were due to exceptional bad weather conditions in southern Europe. An attractive feature of both models for both areas is that the forecast accuracy does not deteriorate over the forecast horizon. Probably, the inclusion of the error correction term including exogenous variables guarantees reasonable long-term forecasts. Of course, this is only true in as far as we are able to predict the exogenous variables with reasonable accuracy.

With respect to the difference between the aggregated and the total HICP model results, for both areas the aggregated results seem better. First of all, they follow the realised inflation more closely. Moreover, the confidence bands for the total HICP models are wider, especially for longer forecast horizons. However, this forecast evaluation is only based on one period. In the next section, we will evaluate the forecast performance of all models more systematically.

5 MODEL EVALUATION

In order to systematically evaluate the models, we have computed the root mean squared error of recursive dynamic out-of-sample forecasts (Stock and Watson, 1999), for the model specifications given in Tables 1 and 2. For this purpose, the realisations of the exogenous variables are used. For the Netherlands this includes gas prices, housing rents and, from 2000 on, radio and television (RTV) licences. The latter is taken as exogenous as the abolition of them in January 2000 had a huge negative impact on services inflation in that year (see figure A1). If no account is taken of this event, the fit of the models deteriorate. Our first out-of-sample evaluation is for the period 1998:1 up until 1999:6 based on data up until 1997:12, whereas the last exercise involves 2003:1 up until 2003:6, leading to at most 61 recursive forecast errors for each horizon.

Tables 3 and 4 show the RMSFE for the Netherlands respectively the euro area. The forecast errors are evaluated for the year-on-year percentage change in the respective HICP component. The forecast errors of the estimated models are compared to those of a naive forecast, which sets all the forecasts ahead equal to the latest observed annual inflation rate and optimal univariate AR models. The AR models are in first differences, with seasonal dummies, and the lag length is chosen based on the Schwarz criterion for the full sample. With respect to the total HICP, both the results of the own model for total HICP and the one based on an aggregate of components are reported.

Table 3 Netherlands: Recursive root mean squared forecast error 1998-2002, 1 to 18 months ahead.

	P^{uf}			P^{pf}			P^{i}			P^{e}			P^{s}			P^{total}			P^{agg}	
horizo	naive	AR(0)	model	naive	AR(1)	model	naive	AR(12)	model	naive	AR(0)	model	Naive	AR(0)	model	naive	AR(0)	model	AR	model
1	*1.89*	**1.49**	**1.49**	*0.42*	**0.30**	0.39	*0.45*	**0.40**	0.44	*1.94*	1.62	**0.88**	*0.38*	0.34	**0.19**	0.36	*0.36*	0.31	0.35	**0.27**
2	*2.75*	**2.20**	**2.20**	*0.67*	**0.49**	0.62	*0.61*	**0.55**	0.59	*2.77*	2.34	**1.25**	*0.56*	0.51	**0.29**	0.50	*0.55*	0.43	0.53	**0.38**
3	*3.38*	**2.70**	**2.70**	*0.87*	**0.67**	0.76	*0.70*	**0.64**	0.69	*3.30*	2.92	**1.45**	*0.69*	0.65	**0.34**	0.61	*0.71*	0.51	0.66	**0.45**
4	*3.87*	**3.12**	**3.12**	*1.05*	**0.85**	0.88	*0.82*	**0.74**	0.78	*3.88*	3.37	**1.69**	*0.80*	0.78	**0.37**	0.74	*0.86*	0.60	0.81	**0.54**
5	*4.25*	**3.47**	**3.47**	*1.20*	1.01	**0.98**	*0.99*	**0.86**	0.89	*4.71*	3.95	**1.99**	*0.90*	0.88	**0.41**	0.88	*0.99*	0.67	0.94	**0.62**
6	*4.67*	**3.76**	**3.76**	*1.35*	1.19	**1.10**	*1.14*	0.97	**0.96**	*5.32*	4.40	**2.11**	*0.99*	0.96	**0.40**	0.97	*1.11*	0.71	1.05	**0.68**
7	*5.05*	**4.02**	**4.02**	*1.48*	1.38	**1.22**	*1.30*	1.12	**1.03**	*5.92*	4.81	**2.19**	*1.10*	1.04	**0.41**	1.07	*1.21*	0.73	1.13	**0.72**
8	*5.42*	**4.19**	**4.19**	*1.62*	1.54	**1.28**	*1.43*	1.25	**1.08**	*6.59*	5.24	**2.40**	*1.21*	1.11	**0.42**	1.18	*1.32*	**0.75**	1.23	0.77
9	*5.88*	**4.45**	**4.45**	*1.75*	1.69	**1.33**	*1.52*	1.34	**1.11**	*7.18*	5.72	**2.59**	*1.32*	1.18	**0.43**	1.29	*1.43*	**0.76**	1.33	0.82
10	*6.39*	**4.76**	**4.76**	*1.88*	1.85	**1.37**	*1.60*	1.43	**1.15**	*7.60*	6.04	**2.65**	*1.44*	1.26	**0.44**	1.39	*1.54*	**0.76**	1.42	0.86
11	*6.96*	**5.03**	**5.03**	2.01	*2.02*	**1.41**	*1.70*	1.53	**1.18**	*8.09*	6.39	**2.74**	*1.55*	1.36	**0.46**	1.48	*1.65*	**0.78**	1.51	0.91
12	*7.48*	**5.35**	**5.35**	2.12	*2.20*	**1.46**	*1.81*	1.63	**1.20**	*8.64*	6.77	**2.87**	*1.65*	1.45	**0.48**	1.60	*1.79*	**0.80**	1.63	0.96
13	*7.79*	**5.36**	**5.36**	2.19	*2.27*	**1.45**	*1.87*	1.67	**1.21**	*9.01*	6.79	**2.88**	*1.73*	1.47	**0.49**	1.67	*1.81*	**0.80**	1.66	0.98
14	*7.97*	**5.35**	**5.35**	2.25	*2.31*	**1.42**	*1.93*	1.70	**1.22**	*9.36*	6.85	**2.89**	*1.81*	1.49	**0.51**	1.74	*1.84*	**0.82**	1.70	1.00
15	*8.11*	**5.36**	**5.36**	2.32	*2.35*	**1.41**	*1.99*	1.73	**1.24**	*9.79*	6.92	**2.91**	*1.88*	1.50	**0.53**	1.82	*1.86*	**0.85**	1.73	1.02
16	*8.18*	**5.36**	**5.36**	2.38	*2.38*	**1.41**	*2.05*	1.75	**1.25**	*10.23*	7.00	**2.93**	*1.94*	1.52	**0.54**	1.88	*1.89*	**0.89**	1.77	1.05
17	*8.13*	**5.40**	**5.40**	*2.43*	2.41	**1.38**	*2.08*	1.77	**1.26**	*10.53*	7.08	**2.95**	*1.98*	1.54	**0.56**	*1.92*	1.91	**0.94**	1.79	1.07
18	*8.07*	**5.44**	**5.44**	*2.48*	2.44	**1.38**	*2.10*	1.78	**1.27**	*10.80*	7.15	**2.98**	*2.03*	1.56	**0.59**	*1.97*	1.94	**0.99**	1.82	1.09

The forecast errors are computed over the annual inflation rates. The models for services and total HICP are corrected for the abolition of RTV licences. The lag length of the AR models is based on the Schwarz criterion using the full sample. The lowest RMSFE for each index is printed in bold face, the highest one in italics.

Table 4 Euro area: Recursive root mean squared forecast error 1998-2002, 1 to 18 months ahead.

	P^{uf}			P^{pf}			P^{i}			P^{e}			P^{s}			P^{total}			P^{agg}	
horizo	naive	AR(1)	model	naive	AR(3)	model	naive	AR(12)	model	naive	AR(0)	model	naive	AR(12)	model	naive	AR(12)	model	AR	model
1	*0.80*	**0.60**	**0.60**	*0.16*	**0.13**	0.14	**0.21**	*0.24*	0.21	*1.87*	1.40	**0.84**	0.19	*0.22*	**0.18**	*0.21*	0.20	0.19	0.20	**0.15**
2	*1.36*	**0.99**	**0.99**	*0.25*	**0.21**	0.23	0.33	*0.36*	**0.32**	*2.79*	2.09	**1.05**	0.21	*0.26*	**0.19**	*0.32*	0.29	0.27	0.30	**0.23**
3	*1.79*	**1.30**	**1.30**	*0.36*	**0.28**	0.31	0.38	*0.40*	**0.37**	*3.56*	2.65	**1.29**	0.26	*0.31*	**0.23**	*0.39*	0.32	0.30	0.36	**0.28**
4	*2.14*	**1.57**	**1.57**	*0.46*	**0.37**	0.39	0.40	*0.42*	**0.39**	*4.39*	3.30	**1.57**	0.29	*0.33*	**0.23**	*0.43*	0.33	0.33	0.41	**0.31**
5	*2.41*	**1.80**	**1.80**	*0.54*	**0.44**	0.48	0.40	*0.42*	**0.37**	*5.05*	3.73	**1.76**	0.34	*0.38*	**0.27**	0.43	0.33	**0.31**	*0.44*	0.34
6	*2.61*	**1.97**	**1.97**	*0.62*	**0.51**	0.57	0.40	*0.42*	**0.36**	*5.68*	4.13	**1.95**	0.39	*0.45*	**0.30**	0.43	0.35	**0.32**	*0.47*	0.37
7	*2.80*	**2.13**	**2.13**	*0.70*	**0.59**	0.65	0.44	*0.51*	**0.39**	*6.31*	4.56	**2.17**	0.45	*0.51*	**0.34**	0.44	0.40	**0.33**	*0.51*	0.40
8	*3.06*	**2.30**	**2.30**	*0.76*	**0.67**	0.72	0.50	*0.62*	**0.43**	*6.94*	4.94	**2.35**	0.48	*0.56*	**0.35**	0.45	0.46	**0.35**	*0.57*	0.44
9	*3.37*	**2.52**	**2.52**	*0.83*	**0.75**	0.79	0.55	*0.66*	**0.44**	*7.61*	5.32	**2.58**	0.53	*0.63*	**0.40**	0.48	0.51	**0.40**	*0.62*	0.49
10	*3.71*	**2.75**	**2.75**	*0.90*	**0.84**	0.87	0.57	*0.68*	**0.44**	*8.27*	5.73	**2.80**	0.57	*0.67*	**0.44**	0.53	0.57	**0.45**	*0.68*	0.54
11	*4.07*	**2.98**	**2.98**	*0.97*	**0.93**	0.95	0.60	*0.68*	**0.43**	*8.89*	6.14	**2.95**	0.61	*0.72*	**0.49**	0.59	0.63	**0.52**	*0.74*	0.59
12	*4.38*	**3.20**	**3.20**	*1.03*	**1.01**	1.02	0.62	*0.67*	**0.42**	*9.58*	6.60	**3.13**	0.66	*0.75*	**0.53**	0.65	0.71	**0.59**	*0.80*	0.63
13	*4.60*	**3.25**	**3.25**	*1.09*	1.05	**1.04**	0.65	*0.70*	**0.41**	*10.08*	6.62	**3.24**	0.69	*0.80*	**0.57**	0.68	0.74	**0.63**	*0.81*	0.65
14	*4.79*	**3.29**	**3.29**	*1.14*	1.09	**1.06**	0.68	*0.75*	**0.41**	*10.46*	6.63	**3.29**	0.73	*0.86*	**0.62**	0.70	0.77	0.68	*0.82*	**0.67**
15	*4.93*	**3.33**	**3.33**	*1.19*	1.12	**1.07**	0.69	*0.76*	**0.39**	*10.87*	6.64	**3.31**	0.78	*0.91*	**0.67**	0.72	0.81	0.75	*0.81*	**0.69**
16	*5.03*	**3.36**	**3.36**	*1.24*	1.14	**1.08**	0.70	*0.76*	**0.37**	*11.20*	6.66	**3.32**	0.82	*0.96*	**0.73**	0.72	*0.86*	0.81	0.82	**0.70**
17	*5.09*	**3.40**	**3.40**	*1.28*	1.15	**1.09**	0.69	*0.77*	**0.35**	*11.47*	6.72	**3.32**	0.85	*1.01*	**0.76**	0.75	*0.91*	0.89	0.83	**0.71**
18	*5.16*	**3.43**	**3.43**	*1.32*	1.15	**1.10**	0.69	*0.78*	**0.33**	*11.74*	6.78	**3.31**	0.89	*1.07*	**0.80**	0.78	*0.97*	0.96	0.85	**0.72**

The forecast errors are computed over the annual inflation rates. The lag length of the AR models is based on the Schwarz criterion using the full sample. The lowest RMSFE for each index is printed in bold face, the highest one in italics.

For both areas, we find that the models outperform the naive forecast almost uniformly. For the Netherlands, this is always the case, whereas for the euro area the one period ahead model forecast for non-energy industrial goods inflation is about equally good as the naive one. Moreover, for the total HICP, the naive forecasts outperforms the own model 15 to 18 months ahead, and the aggregate model 9 and 10 months ahead. Although, outperforming the naive forecast seems hardly demanding, the results for the optimal AR models for P^i and P^s for the euro area and the ones for P^{total} and P^{agg} show it is far from trivial. Moreover, the it is promising that the outperformance holds uniformly, especially since another criterion was used to select the models. Relative to the AR benchmark, the models still perform very good, although the models for processed food, and the one for non-energy industrial goods for the Netherlands, are slightly outperformed by the AR models for short horizons.

Comparing the naive forecast errors for the Netherlands with those for the euro area, it is clear that Dutch inflation is much more volatile than inflation in the total euro zone. Many of the shocks to inflation are country-specific and these shocks partly cancel for the euro area. The exception is energy inflation. For energy, oil prices are most important and these shocks hit all countries at the same time. For the models, the difference in RMSFE between the two areas is less extreme. Consequently, the improvement relative to the naive forecast is bigger for the models for the Netherlands than for those of the euro area. For forecasts 10 or more months ahead, the Dutch model RMSFE for energy and the one for services is even smaller than the corresponding euro area ones. This is probably due to the assumption of exogenous natural gas prices, housing rents and RTV licences for the Netherlands.

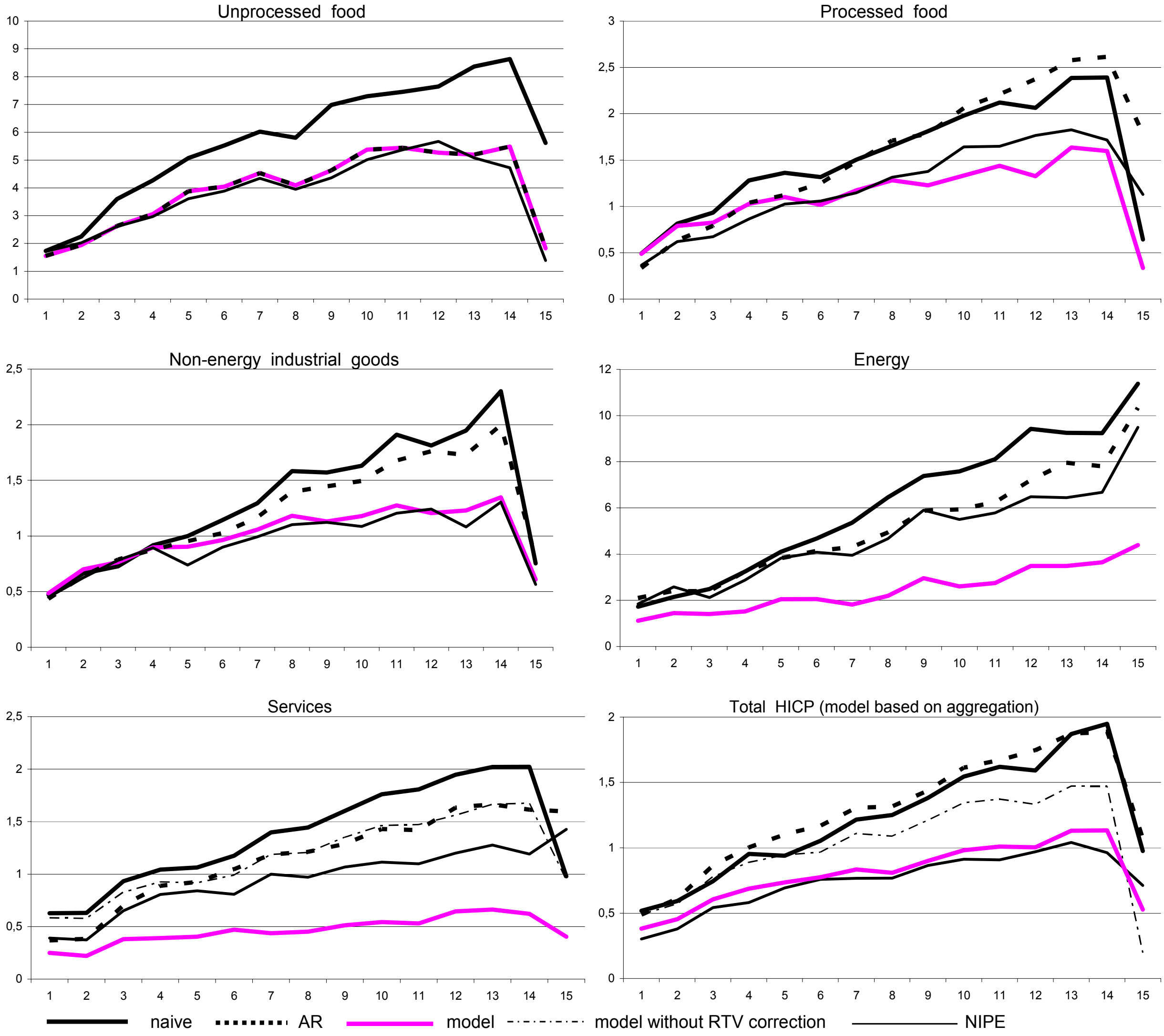
With respect to the advantage of splitting up the HICP index to forecast total HICP inflation, the results are somewhat mixed. For the Netherlands, we find that aggregation of components leads to a lower RMSFE for forecasts up to 7 months ahead, whereas for longer forecast horizons the opposite holds. For the euro area on the other hand, the aggregation method not only performs best for short horizons, but also for very long ones, whereas the direct method shows the lowest RMSFE only for 5 to 13 months ahead. These results imply that the dominance of the indirect approach for 2002 and the first half of 2003, which appeared clear from Figure 1, is certainly not a general feature. The relative good performance of the aggregation method for short forecasts horizons runs counter to the results of Fritzer, Moser and Scharler (2002). For VAR models, they found the direct approach to perform better for horizons up to 9 months ahead, after which the aggregation approach was to be preferred. Hubrich (2001, 2003) on the other hand found that aggregation performed especially worse at long horizons. Also with respect to the AR models, no common feature is found. Whereas for the Netherlands the disaggregated approach produces better results, the opposite holds for the euro area. In general, it seems that forecast errors among HICP sub-indices are too positively correlated to be able to gain a lot by aggregating component models.

The relatively good forecast performance of our models does of course depend on our ability to predict exogenous variables correctly. In Tables 3 and 4, the realisations are used to make forecasts, but obviously these are not available when making really out-of-sample forecast. Moreover, even though we did not use the minimal RMSFE as a criterion to select our models, our model selection criterion is certainly not completely out-of-sample. For the Netherlands, we do have a way to check the relevance of these two objections as the Dutch models have been used for the NIPE since December 1998. Consequently, we have projections for HICP inflation and its five components for 16 forecast rounds.

In Figure 2 the root mean squared forecast errors of the NIPE projections are shown together with the ones for the naive forecast, the optimal AR models and those for the currently selected models. The squared forecast errors are averaged over all projections that were made for a certain forecasting horizon, that is 16 for 1 to 11 months ahead, 12 for 12 months ahead, 8 for 13 and 14 months ahead and 3 for 15 months ahead. This explains the sudden drop in RMSFE at horizon 15.

For every HICP component, the NIPE projections outperformed the naive and AR benchmarks almost uniformly. Compared to the model cum realised exogenous variables forecasts the results are more mixed. In principle, there are three reasons for differences between the model forecasts and the NIPE results. First, the assumptions regarding the exogenous variables differ. Second, different models were used. Third, the NIPE also includes ‘add-factors’ to account for judgmental issues. Unfortunately, we do not have a complete track record of the models and assumptions used. Otherwise, we could identify the exact relevance of each of the three factors. Nevertheless, some interesting conclusions can be drawn. Of the three factors, wrong assumptions probably only lead to worse predictions. For the model specification it can go either way, whereas judgement hopefully only improves the results.

Figure 2: Root mean squared forecast error Dutch HICP inflation 1 to 15 months ahead



The only two sub-indices for which the NIPE performed systematically worse than the model are energy and services. For energy, this is not at all surprising as the development of oil prices is highly unpredictable and very influential on energy prices. Moreover, gas prices, which account to almost 40% of Dutch energy budget, also sometimes moved more than expected. With respect to services, the result is partly due to the abolition of RTV licences in January 2000, which was not foreseen in the NIPE projections of 1999. The figure also shows the impact of ignoring this event for our selected model. It is clear that the model falls apart, as the NIPE now outperforms the model. Besides this effect, the housing rents, which accounted for about 30% of the Dutch services budget, were not always perfectly predictable.

Given the relatively big forecast errors for energy and services, it is surprising to see that for the overall index the NIPE performs even slightly better than the model forecast, even though both are based on the sub-indices. Apparently, the correlation among HICP components was higher for models with correct exogenous variables than for the NIPE. Consequently, the disadvantage of not knowing the future values of exogenous variables was compensated enough by the possibility to add judgement.

6 CONCLUSION

This paper describes the procedures we use to predict monthly Dutch and euro area HICP inflation. The HICP prediction is constructed by aggregating forecasts for the five HICP sub-indices unprocessed food, processed food, non-energy industrial goods, energy and services, whereas total HICP is also modelled directly for comparison reasons. All models are linear vector autoregressive or error correction models, possibly including exogenous variables.

In order to select the appropriate models, the first step is a visual inspection of the data. Those price indices which show a clear changing seasonal pattern are modelled in both first and twelve month differences including an error correction term representing long run equilibrium relationships between inflation and other variables (if they have the correct sign and reasonable order of magnitude). Price indices without clear structural breaks in seasonal pattern (unprocessed food and energy) are modelled in first differences. Here, no error correction term is included. The second step involves the calculation of all possible VARX or VECMX models, using a small set of exogenous and endogenous variables. We select the best models according to nine different statistical selection criteria, using both in-sample goodness-of-fit, parsimony and out-of-sample forecasting accuracy. In the third step, the optimal models are chosen, based both on the statistical criteria and economic evaluation. Especially, the long run properties are important here. Expected wage developments form a very important anchor in this respect.

Once an appropriate model is chosen, all available data are used to generate forecasts. Foreseeable shocks over the forecast horizon (for instance a change in indirect taxes) are incorporated ex-post. Apart from the point forecasts, prediction intervals are generated using a bootstrap methodology. In this context, it is important to use the bootstrap residuals of the same timing for all five sub-indices, so as to preserve the correlation between the five inflation series.

According to a recursive root mean squared forecast error evaluation exercise, all models outperform the naive forecast and optimal AR models on most forecast horizons. The comparison of forecast errors for total HICP between direct HICP models on the one hand and an aggregation of sub-index models on the other does not show a clear preference for either procedure. For short forecast horizons the aggregate is better, but for intermediate horizons the direct approach is to be preferred. These evaluations do depend on perfect knowledge of future values for exogenous variables however.

For the Netherlands, a really out-of-sample exercise is performed by evaluating the first 16 NIPE rounds. The forecast performance of the NIPE projections is even slightly better than the one for the selected models with perfect foresight of exogenous variables. Again, the naive forecast is outperformed on every forecast horizon for every (sub-)index. Apparently, judgement more than compensates for the lack of knowledge on the future values of exogenous variables. Indeed, forecasting inflation seems to be an art as well as a science!

Overall, the robustness of the inflation forecasting models, both with respect to the selection criterion used and over time, does not stem overly optimistic though. *The* optimal model is not likely to exist, making regular evaluation of models and the permanent good use of common sense all the more important. The apparent instability of models worsens the problem of overfitting. To compensate for this regularity, our models currently in use at most carry one lag only. Moreover, the instability of linear models is hardly supportive to the introduction of very complicated non-linear models.

REFERENCES

Banerjee, A. and M. Marcellino, 2002, Are there any reliable leading indicators for US inflation and GDP growth?, *unpublished document.*

Bruneau, C., O. de Bandt and A. Flageollet, 2002, Forecasting inflation in the euro area, *unpublished document*, Banque de France.

Canova, F., 2002, G-7 Inflation forecasts, *CEPR discussion paper,* No. 3282.

Christoffersen, P.F. and F.X. Diebold, 1998, 'Cointegration and Long-Horizon Forecasting', *Journal of Business and Economic Statistics*, No. 16, pp. 450-458.

Clements, M.P. and D.F. Hendry, 2001, 'Economic forecasting: some lessons from recent research', *ECB Working Paper,* No. 82.

Forni, M., M. Hallin, M. Lippi and L. Reichlin, 2000, The generalized factor model: identification and estimation, *Review of economics and statistics,* No. 82, pp. 540-554.

Fritzer, F., G. Moser and J. Scharler, 2002, Forecasting Austrian HICP and its components using VAR and ARIMA models, *Working Paper OENB,* No. 73.

Garatt, A., K. Lee, M.H. Pesaran and Y. Shin, 2003, Forecast uncertainties in macroeconometric modelling: an application to the UK economy, *Journal of the American Statistical Association*, forthcoming, available on http://www.econ.cam.ac.uk/faculty/pesaran.

Hendry, D.F. and G.E. Mizon, 1999, On selecting policy analysis models by forecast accuracy, *University of Southampton Discussion Papers in Economics and Econometrics,* No. 9918.

Hubrich, K., 2001, Forecasting euro area inflation: does contemporaneous aggregation improve the forecasting perfomrance?, *Research Memorandum WO&E*, No. 661.

Hubrich, K., 2003, Forecasting euro area inflation: does aggregating forecasts by HICP component improve forecast accuracy?, *ECB Working Paper*, No. 247.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl and T. Lee, 1988, Introduction to the theory and practice of econometrics, John Wiley & Sons.

Kilian, L., 1998, Small-Sample Confidence Intervals for Impulse Response Functions, *Review of Economics and Statistics*, No. 80(2), pp. 218-230.

Lütkepohl, H., 1987, Forecasting aggregated vector ARMA processes, Springer-Verlag.

Marcellino, M., 1999, Some consequences of temporal aggregation in empirical analysis, *Journal of Business and Economic Statistics,* No. 17(1), pp. 129-136.

Marcellino, M., J.H. Stock and M.W. Watson, 2003, Macroeconomic forecasting in the euro area: country specific versus area-wide information, *European Economic Review*, No. 47(1), pp. 1-18

Moser, G., F. Rumler and J. Scharler, 2002, Evaluating factor model forecasts for Austrian inflation, *unpublished document*, Oesterreichische Nationalbank.

Stock, J.H. and M.W. Watson, 1999, Forecasting inflation, *Journal of Monetary Economics,* No. 44(2), pp. 293-335.

Wallis, K.F., 1999, 'Asymmetric density forecasts of inflation and the Bank of England's fan chart', *National Institute Economic Review,* No. 167 ,106-112.

APPENDIX DATA

The sample period of the data set is October 1987 respectively January 1990 for the Netherlands and the euro area until August 2002. Table A1 lists all the variables that are currently included. Apart from these selected variables, other variables have been tested but are not selected in the final models.

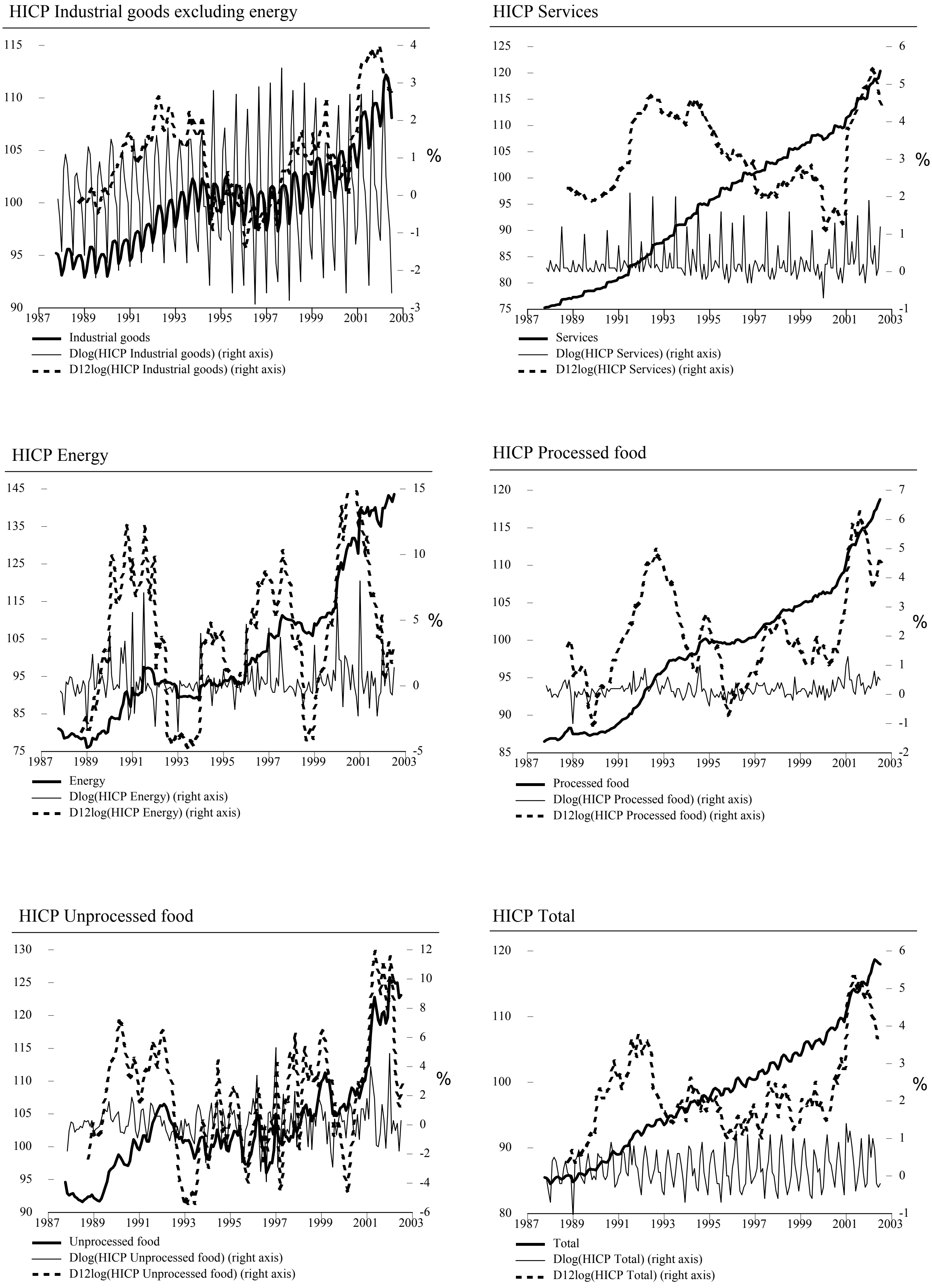
Table A1 Data, notation and source code

Variable	Notation	External source
Harmonised Index of Consumption Prices		
Euro area		
HICP	Pea^{total}	Eurostat
HICP unprocessed food	Pea^{uf}	Eurostat
HICP procecessed food	Pea^{pf}	Eurostat
HICP industrial production excl. Energy	Pea^{i}	Eurostat
HICP energy	Pea^{e}	Eurostat
HICP services	Pea^{s}	Eurostat
Netherlands		
HICP	Pnl^{total}	Eurostat
HICP unprocessed food	Pnl^{uf}	Eurostat
HICP processed food	Pnl^{pf}	Eurostat
HICP industrial goods excl energy	Pnl^{i}	Eurostat
HICP energy	Pnl^{e}	Eurostat
HICP services	Pnl^{s}	Eurostat
Endogenous variables		
Import price index Germany	Pim^{GE}	Federal Statistical Office Germany
Producer prices (euro area)	P^{prod}	BIS
Exogenous variables		
Euro/dollar exchange rate	$e^{€\$}$	ECB
Oil price (Brent crude) in euro	P^{oil}	IFS / Bloomberg [a)]
Hourly wages industry, euro area	$wages^{EU}$	[b)]
Hourly wages private sector, Netherlands	$wages^{NL}$	CBS
Commodity prices (excl.energy) in euro	wmp^{exe}	HWWA [a)]

[a)] Recent data as well as projections for the forecast horizon are obtained from the ECB. The projections are based on futures prices.

[b)] The euro area hourly wage is an average of the individual country´s hourly wage rates, weighted by the GDP-share in 1995. For Belgium, Denmark, Spain, Finland, France, Greece and Ireland only data on quarterly basis is available. This is interpolated to monthly data by the Lisman-procedure. Portugal and Luxembourg are not considered due to lack of data. Moreover, a 12-months centered moving average is applied to smooth the aggregated hourly wage rate in order to get more reliable parameter estimates.

Figure A1 HICP (sub)indices in original, monthly and annual inflation format for the Netherlands



HICP Industrial goods excluding energy

HICP Services

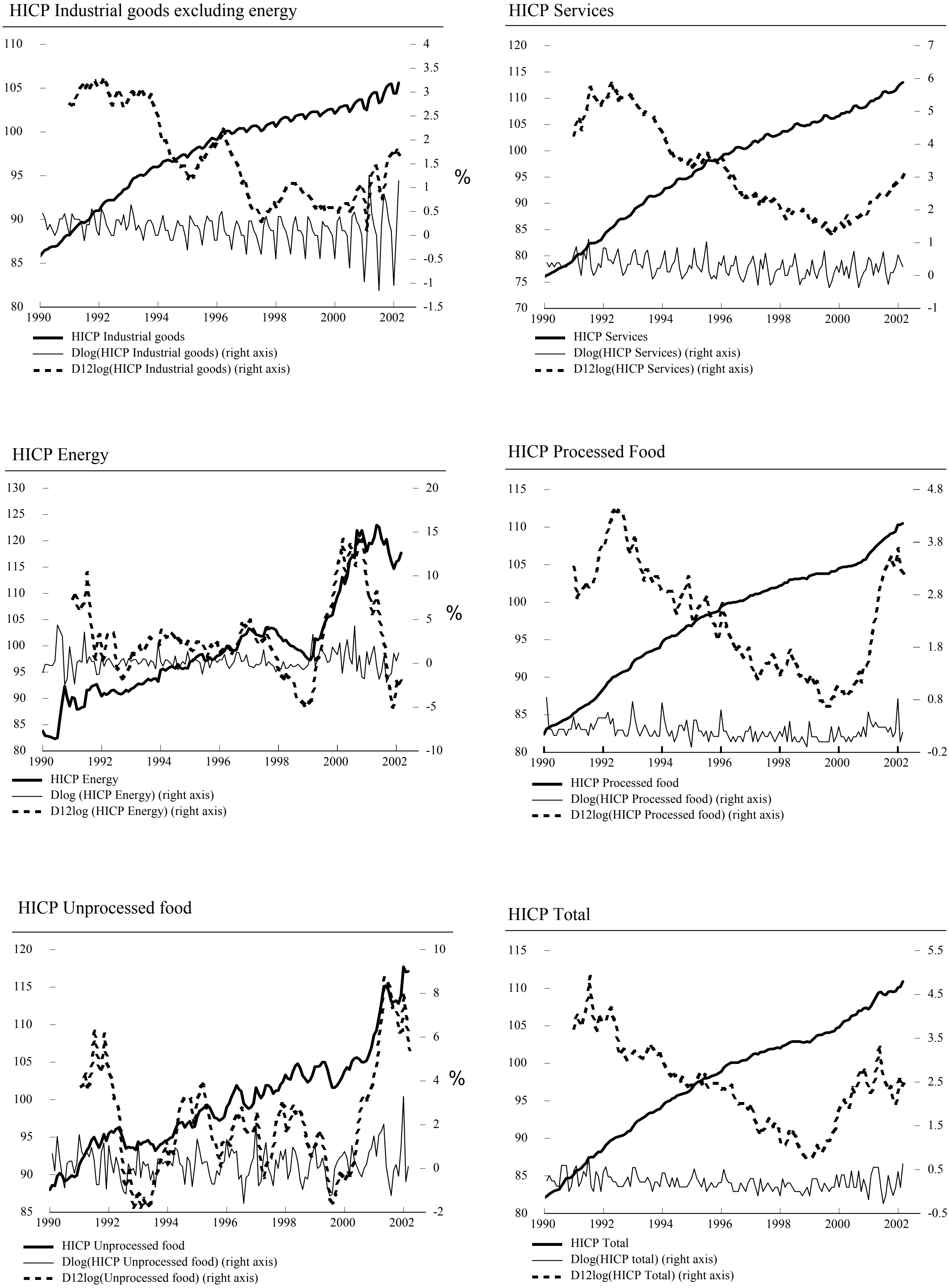
HICP Energy

HICP Processed food

HICP Unprocessed food

HICP Total

Figure A2 HICP (sub)indices in original, monthly and annual inflation format for the Euro area



HICP Industrial goods excluding energy

HICP Services

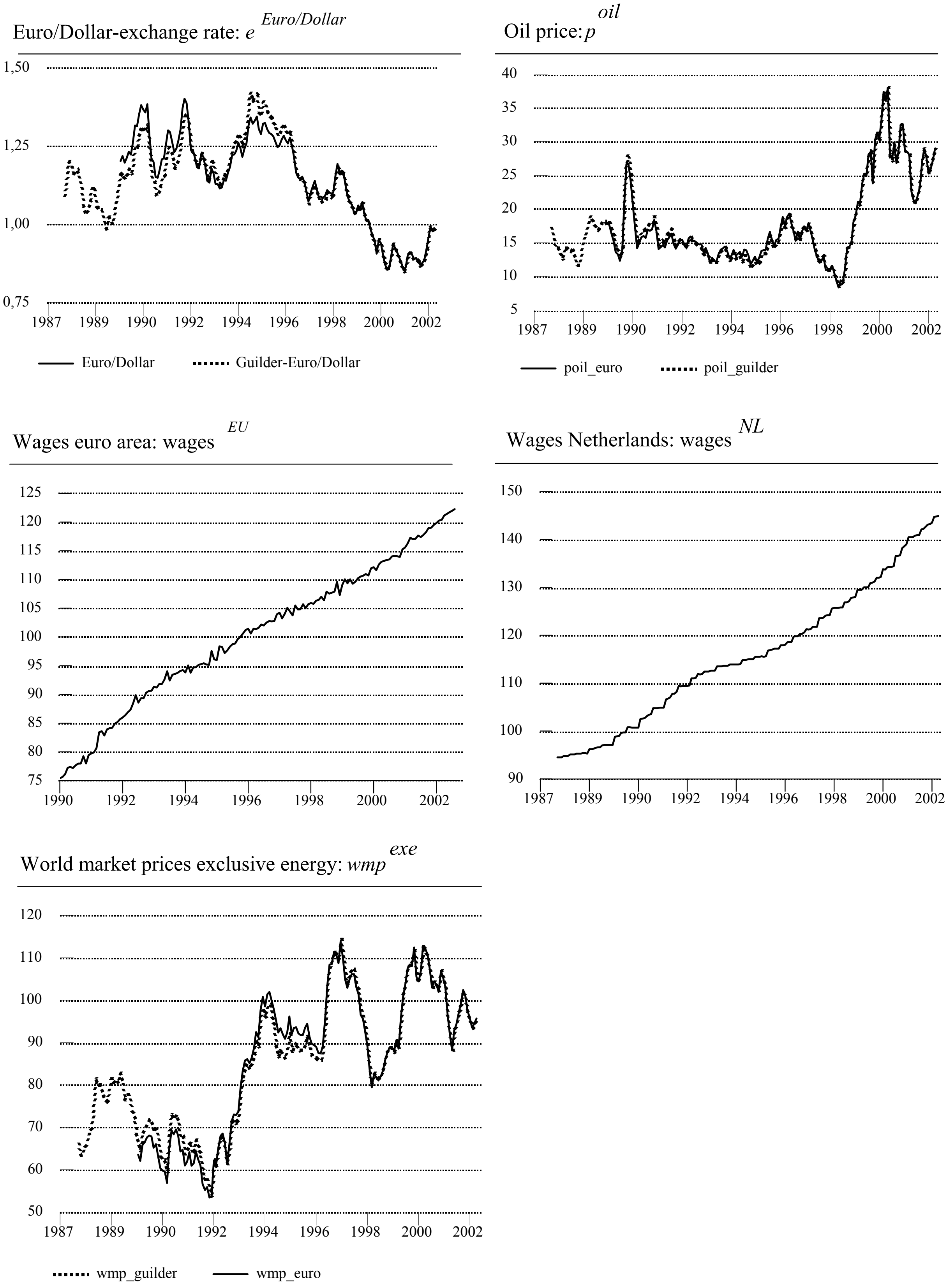
HICP Energy

HICP Processed Food

HICP Unprocessed food

HICP Total

Figure A3 Exogenous variables



Euro/Dollar-exchange rate: $e^{Euro/Dollar}$

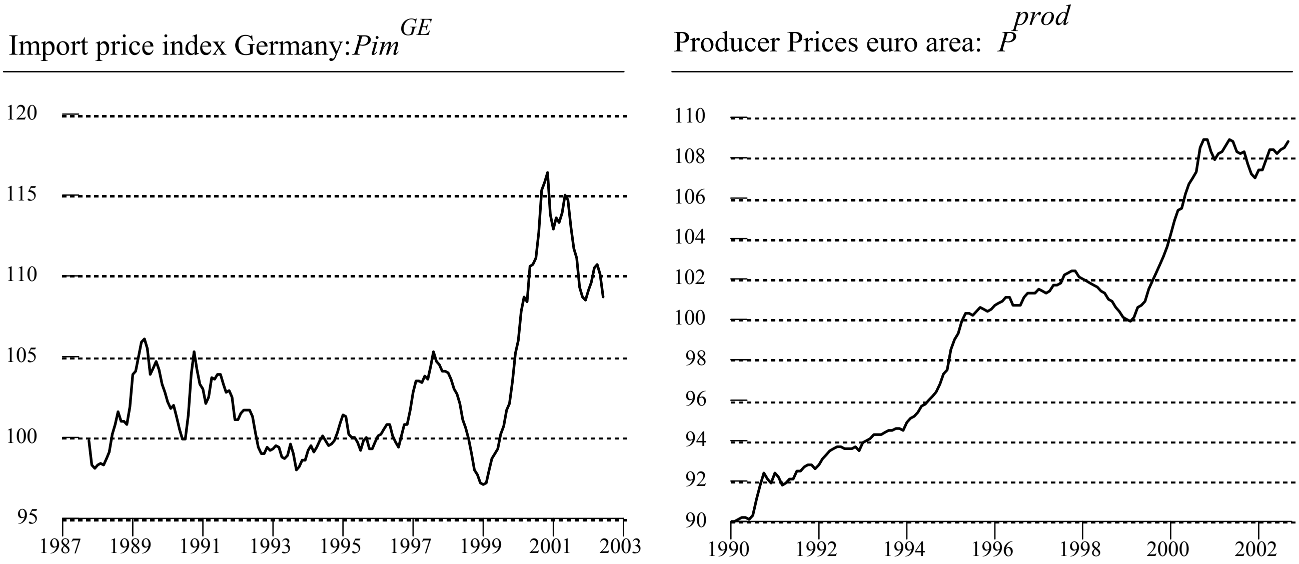
Oil price: p^{oil}

Wages euro area: $wages^{EU}$

Wages Netherlands: $wages^{NL}$

World market prices exclusive energy: wmp^{exe}

Figure A4 Endogenous variables



Import price index Germany: Pim^{GE}

Producer Prices euro area: P^{prod}