

A Work Project, presented as part of the requirements for the Award of a Master's degree in Economics from the Nova School of Business and Economics.

EVALUATION OF INFLATION FORECASTING METHODS

ROMAN LENGELING

Work project carried out under the supervision of:

Paulo M. M. Rodrigues

17-12-2024

Abstract

This paper evaluates inflation forecast accuracy for Germany using disaggregated Harmonized Index of Consumer Prices (HICP) data, comparing a bottom-up approach to direct aggregate forecasting. By examining more granular data disaggregation, Autoregressive, Vector Autoregressive, and factor-augmented models are used to forecast HICP. The findings show that direct aggregate forecasting consistently outperforms the bottom-up approach. Additionally, we enhance aggregate forecast accuracy through forecast combination techniques, which deliver better results than individual models.

Keywords: Aggregate forecasts; Disaggregate information; Forecast combination; Inflation; Linear time series models

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1 Introduction

During recent years of high inflation in the European Union, inflation forecasting has become increasingly relevant. In 2021, the European Central Bank (ECB) readjusted its 2% inflation target, measured by the year-on-year increase in the HICP, to a symmetrical target, aiming to keep inflation close to 2% over the medium term. Achieving this target requires constant monitoring and accurate forecasting of price trends. Since monetary policy measures affect the economy with a lag and external shocks can disrupt price stability, inflation forecasting is crucial for the effective management of monetary policy. While forecasting inflation for the entire euro area is essential, country-specific forecasts also play a critical role in understanding how different economies respond to policy measures. Marcellino et al. (2003) showed that aggregating forecasts from individual countries can improve accuracy over directly forecasting the aggregate, highlighting the importance of disaggregated data.

In addition to monetary policy, accurate inflation forecasts are important for fiscal planning, wage negotiations, and financial markets. Improving forecast accuracy can often be achieved by incorporating more detailed data, such as disaggregated inflation components. By breaking down inflation into subcomponents, forecasters can better capture the unique dynamics of each component, which may lead to more precise predictions overall. Additionally, forecast errors from different components may cancel each other out when aggregated, resulting in a more accurate overall forecast.

There are two main approaches to leveraging disaggregated data: i) The bottom-up approach, which involves forecasting individual components and then aggregating them to obtain the overall forecast. This method can improve accuracy by addressing the specific characteristics of each subcomponent, although it risks introducing errors if the component-level models are misspecified. ii) Dimensionality reduction, which condenses a large set of disaggregated information into a few key variables which then can be used to forecast the aggregate directly. This simplifies the forecasting process without losing essential data, though it may not capture all individual component characteristics.

Some argue that forecasting the aggregate directly is a more efficient strategy than aggregating subcomponent forecasts, since disaggregate models might be imperfect, and misspecifications can undermine the potential gains from using detailed data. Furthermore, if unexpected shocks affect multiple components simultaneously, the

advantage of error cancellation may be lost, negating the benefits of disaggregation.

This paper will focus on the evaluation of both approaches for the case of Germany. The key question is whether aggregating subcomponent forecasts outperforms forecasting the aggregate directly. To investigate this, the paper is organized as follows. In section 2, a summary of the existing literature is given. In section 3, the methodology for both methods is discussed. In section 4, the data that we use will be presented and in section 5, inflation forecast results are evaluated. Section 6 includes robustness tests and finally, section 7 concludes.

2 Literature review

For the comparison of aggregating HICP subcomponent forecasts or forecasting the HICP aggregate directly, there is extensive literature. One of the most famous papers in this field is Hubrich (2005), who analyzes whether the accuracy of forecasts of aggregate euro area inflation can be improved by aggregating forecasts of subindices of the HICP. His results indicate that aggregating forecasts by component may outperform aggregate forecasts for a 1-month ahead horizon but not in the case of forecasting year-on-year inflation 12 months ahead, which is more relevant for monetary policy. He finds that subcomponents are affected by certain shocks in a similar way, leading to forecast failures in the same direction. Fritzer et al. (2002) (for Austria) and Reijer and Vlaar (2003) (for the Netherlands) find that aggregating the forecasts of the subcomponents is also important up to 6 and 7 months ahead, respectively.

Hubrich and Hendry (2011) expanded on Hubrich's 2005 work by examining whether forecasts for the aggregate could be enhanced by incorporating disaggregate variables into the aggregate model. They proposed first combining the information from the disaggregate variables before incorporating this combined information into the aggregate model to potentially improve forecasts for U.S. inflation. Their findings show that directly forecasting the aggregate, whether using only past aggregate values or disaggregate information, is more accurate than aggregating individual forecasts from disaggregate components. This suggests that using combined disaggregate information outperforms the approach of combining individual disaggregate forecasts. Additionally, including a select set of disaggregate variables or factors summarizing the disaggregate information

tends to improve forecast accuracy over using only past aggregate data, especially in samples with sufficient variability in the aggregate.

This work is closely related to Stock and Watson (2002) who try to forecast macroeconomic variables using diffusion indices or dynamic common factors, meaning they apply dimensionality reduction schemes without a significant loss of information in cases of scarce time series observations.

Duarte and Rua (2007) try to merge both strands of aggregating subcomponent forecasts and forecasting the aggregate directly for the case of Portugal. In contrast to other existing literature, they introduce three different levels of disaggregation, with the highest level containing almost sixty subcomponents. Also, they extend the Seasonal Autoregressive Integrated Moving Average (SARIMA) models used by including exogenous variables that affect the dynamic behaviour of the dependent variable, thereby converting them to Factor-Augmented *SARIMA* models. Their findings indicate that for very short-term inflation forecasting, a bottom-up approach with a high level of disaggregation and an emphasis on exploiting variables' comovement is most effective. However, for longer horizons, simpler models tend to perform better, and the necessary level of disaggregation decreases as the forecast horizon extends.

Regarding the forecasting techniques, Cobb (2019) investigates an implementation of a bottom-up approach to generate density forecasts for CPI and GDP instead of point forecasts which can be useful for predicting intervals of future inflation. Marcellino et al. (2005) investigate whether iterated forecasts outperform direct multiperiod-ahead forecasts, as researchers can choose between these two methods when generating a multiperiod time series forecast. They find that iterated forecasts do indeed have lower mean squared forecast errors (MSFE) than direct forecasts.

3 Methodology

In this work, different univariate and multivariate models will be used to compare the forecasts with each other. Autoregressive (AR), Moving Average (MA), Autoregressive Integrated Moving Average (ARIMA), and Vector Autoregressive (VAR) models have emerged as popular methodologies according to Zarnowitz and Braun (1993).

The *AR* model utilizes past observations of a single time series to forecast future values,

relying on the principle that past behaviour is indicative of future trends. A process y_t is defined as:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \varepsilon_t, \quad t = 1, \dots, T, \quad (1)$$

which is an Autoregressive process of order p , also known as $AR(p)$, where ε_t is a zero-mean white noise process with variance σ^2 .

Conversely, the Moving Average model predicts future values based on past forecast errors (or shocks), emphasizing the role of unforeseen events in shaping price dynamics. Assuming ε_t is a white noise process, y_t is defined as:

$$y_t = \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots + \psi_q \varepsilon_{t-q}, \quad t = 1, \dots, T. \quad (2)$$

Given that the ε_t are not observed, equation (2) can be adjusted so that $\psi_0 = 1$. Thus,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (3)$$

is defined as a *MA* process of order q , also known as $MA(q)$.

The *ARIMA* model integrates the strengths of both *AR* and *MA* components, making it particularly effective for non-stationary data that exhibits trends or seasonality. It represents the most comprehensive class of forecasting techniques for time series data, which can be rendered stationary through differencing, potentially accompanied by nonlinear transformations such as logarithmic adjustments or deflation, if required. A time series is deemed stationary if its statistical properties remain unchanged over time. Characteristics of a stationary series include the absence of a trend, consistent amplitude in its fluctuations around the mean, and a uniform pattern of short-term variability, meaning its autocorrelations remain stable over time. This stability implies that the power spectrum of the series does not change over time. Such a random variable can be conceptualized as a mixture of signal and noise, where the signal, if present, may manifest as patterns of rapid or slow mean reversion, sinusoidal oscillations, or frequent sign changes, potentially including seasonal components. The *ARIMA* model acts as a filter that aims to differentiate the signal from the noise, allowing for future extrapolation of the signal to generate forecasts.

The *ARIMA* forecasting equation for a stationary time series is formulated as a mixture of lagged values of the dependent variable and/or lagged forecast errors. This can be

expressed as:

$$y_t = \mu + \alpha_1 y_{t-1} + \dots + \alpha_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (4)$$

which is also known as an $ARIMA(p, d, q)$ process where p is the number of autoregressive terms, d is the number of nonseasonal differences needed to achieve stationarity, and q is the number of lagged forecast errors included in the prediction equation.

An extension of the $ARIMA$ model is the factor-augmented, seasonal $ARIMA$ ($FASARIMA$) model, which incorporates additional exogenous factors to capture more complex relationships between the endogenous variable and exogenous regressors. The factors in the $FASARIMA$ model are derived through dimensionality reduction techniques like Principal Component Analysis (PCA). A simple $FASARIMA$ model of lag order 1 can be represented as this:

$$y_t = \mu + \phi y_{t-1} + \theta \varepsilon_{t-1} + \sum_{i=1}^K \gamma_i f_{i,t} + \Phi y_{t-s} + \Theta \varepsilon_{t-s} + \varepsilon_t \quad (5)$$

where μ is the intercept term, ϕ is the autoregressive parameter, θ is the moving average parameter, ε_{t-1} is the white noise error term at time $t - 1$, $f_{i,t}$ is the $i - th$ factor at time t , representing a principal component extracted from the exogenous data, γ_i is the parameter for the $i - th$ factor and K represents the number of principal components included in the model. Φ represents the seasonal autoregressive coefficient, y_{t-s} is the value of the series s -periods ago, Θ represents the seasonal moving average coefficient and ε_{t-s} is the error term s -periods ago.

In contrast to univariate models like $ARIMA$, the VAR model captures the interdependencies between multiple time series, making it especially valuable when analyzing systems where several economic indicators interact. It allows to determine the dynamic behaviour between multiple endogenous variables. This characteristic is critical in HICP forecasting, as inflationary pressures can stem from a variety of sources such as growth in money supply, supply chain disruptions, and fluctuations in asset prices. A simple bivariate $VAR(1)$ model can be expressed as:

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} \quad (6)$$

where we assume that the two time series $x_{i,t}$, $i = 1, 2$ are covariance stationary. Also, we

defined ε_t to be a vector of unobservable zero mean white noise processes which means that the elements of ε_t are serially uncorrelated and independent of each other.

3.1 Forecasting the aggregate

In this work, we employ two different approaches: Forecasting the aggregate versus aggregating the forecasts of the HICP's subcomponents. In this subsection, the methodology for forecasting the aggregate directly will be presented.

For both model types, a rolling framework will be pursued which will deliver predictions for different forecast horizons continuously. This means that for each rolling window, univariate and multivariate models will be fitted which deliver h -step-ahead predictions. The models will be tested for different forecast horizons, ranging from 1-month ahead predictions up to 12-month ahead predictions. Also, a window size of 60 periods for each rolling window is chosen as it offers the best trade-off between stability and sensitivity, meaning a larger window size will result in more stable estimates, as it uses a larger set of data to fit the model, however it may also smooth over important short-term changes in the data.

For the univariate models, the procedure in each rolling window starts with the Augmented-Dickey-Fuller (ADF) test (Dickey and Fuller 1979) which tests for the presence of a unit root in a time series and if a unit root is present, the time series is considered non-stationary as its statistical properties (mean, variance and autocovariance) are not constant over time. In this case, the time series will be differenced and the models end up predicting first differences, rather than level values for HICP.

Next, the appropriate orders for each model type among the univariate models will be determined. This is done by fitting each model and extracting the model selection criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and the Hannan-Quinn Information Criterion (HQIC) (Kuha 2004). The order for each model will be chosen based on which order the majority of information criteria proposes. In case there is a tie, meaning each information criterion proposes a different order, the lowest proposed order among the three will be chosen to ensure model simplicity. Then, each univariate model will be fitted using the corresponding best order based on the results from the information criteria.

The procedure for the *FASARIMA* model is similar to the procedure for the univariate

models, except that the factors will be estimated by PCA and the model will be fitted using the first two principal components as they will capture most of the variance in the exogenous data. Note that *FASARIMA* models require forecasting the exogenous variables in order to use them for forecast generation. Hence, for simplicity reasons, *AR*(1) models were used to forecast the principal components, consistent with the framework of Duarte and Rua (2007).

For the multivariate model such as the presented *VAR* model, the procedure in each rolling window starts with the ADF test as well to find the presence of a unit root in the given time series. The test will be applied on all endogenous variables and if non-stationarity is observed, first differences are applied. Then, the same procedure as for the univariate models is applied to find the best possible order for the *VAR* model. After having found the best order, the Granger Causality test will be applied which is used to determine whether one time series can predict the other time series (Shojaie and Fox 2022). This test is used to determine the two best potential predictors in the *VAR* model with the highest potential explanatory power to predict HICP. These selected predictors are then used, among the lagged endogenous variable itself, to fit the *VAR* model.

To generate the best possible predictions based on the estimated models, two different forecasting methods were tested for both univariate and multivariate models: Iterative and direct forecasting (Marcellino et al. 2005). Iterative forecasting or recursive forecasting generates a forecast for a time period one step at a time. This forecast is then used as an input for the forecast in the next period. For an *AR*(p) model, this can be represented as:

$$\hat{y}_{t+h} = \hat{\alpha}_1 \hat{y}_{t+h-1} + \hat{\alpha}_2 \hat{y}_{t+h-2} + \cdots + \hat{\alpha}_p y_{t+h-p} \quad (7)$$

where the forecast $\hat{y}_{t+h}$ depends primarily on the forecasted values from previous steps $\hat{y}_{t+h-1}, \hat{y}_{t+h-2}, \dots$ and possibly on the observed values from the past up to $t + h - p$ if $p > h$, depending on how many lags are considered in the chosen model.

In contrast to this, direct forecasting generates forecasts for all desired horizons simultaneously, without relying on the predictions from the previous steps. This can be achieved by lagging the independent variables in the model by the desired forecast horizon. Direct forecasting is computationally more expensive while iterative forecasting bears the risk of compounding forecast errors, especially for larger forecast horizons. For

a simple $AR(p)$ model, direct forecasting can be formally represented in the following way:

$$y_{t+h} = \alpha_{1,h}y_t + \alpha_{2,h}y_{t-1} + \dots + \alpha_{p,h}y_{t-p+1} + \varepsilon_{t+h}, \quad t = 1, \dots, T, \quad (8)$$

where h represents the forecast horizon.

If the training data is non-stationary for a rolling window and differencing is applied, the forecasts will be in first differences, meaning the rate of change will be predicted, not level values. Hence, level forecast values can be generated by adding the predicted first differences to the last observed level values:

$$\hat{y}_{t+h} = \Delta \hat{y}_{t+h} + y_t \quad (9)$$

The obtained level forecasts will then be compared to the observed level HICP values.

In order to improve the forecasts for the aggregate even further, forecast combination will be applied. The simplest forecast combination technique describes taking the mean of individual forecasts to potentially improve forecasting accuracy and to get more robust predictions than with any single model alone (Timmermann 2006).

To evaluate the forecasts' performance, error metrics such as the Mean-Squared-Error (MSE), Root-Mean-Squared-Error (RMSE) and Mean-Absolute-Error (MAE) will be later used. While the Diebold-Mariano (DM) test is commonly used to statistically compare forecast accuracy, it was designed to compare model-free forecasts, not models. In the context of models with estimated parameters, the DM test can yield misleading results due to nonstationarities in forecast error differences caused by parameter estimation, especially when models are nested (Diebold, 2015). However, with nested models, test properties are likely to differ because, under the null hypothesis, the forecast errors are asymptotically identical and therefore perfectly correlated (Clark and McCracken 2001). Therefore, McCracken (2007) introduces test statistics to evaluate the predictive accuracy of forecasts generated by two nested parametric models. Under the null hypothesis, the expected loss for each pair of forecasts is identical. Conversely, under the alternative hypothesis, the more complex model is expected to have a lower loss. The null and alternative hypotheses are defined as follows:

$$H_0 : \mathbb{E} [L_{t+1} (\beta_1)] = \mathbb{E} [L_{t+1} (\beta_2)] \quad \text{vs} \quad H_a : \mathbb{E} [L_{t+1} (\beta_1)] > \mathbb{E} [L_{t+1} (\beta_2)] \quad (10)$$

where $L_{t+1}(\cdot)$ denotes a squared error loss, and β_1 and β_2 are the parameter vectors of e.g. the $AR(1)$ and $VAR(1)$ models. McCracken proposes two out-of-sample test statistics: An out-of-sample F- and a t-test, i.e.:

$$\text{OOS-F} = P \left[\frac{\text{MSFE}_1(\hat{\beta}_1) - \text{MSFE}_2(\hat{\beta}_2)}{\text{MSFE}_2(\hat{\beta}_2)} \right]; \quad (11)$$

$$\text{OOS-T} = \hat{\omega}^{-0.5} \left[\text{MSFE}_1(\hat{\beta}_1) - \text{MSFE}_2(\hat{\beta}_2) \right], \quad (12)$$

where P is the number of out-of-sample observations, and $\text{MSFE}_1(\cdot)$ and $\text{MSFE}_2(\cdot)$ are the mean squared forecast errors of the $AR(1)$ and $VAR(1)$ models. $\hat{\omega}$ is a consistent estimate of the asymptotic variance of the mean squared-error loss differential. The $OOS - T$ statistic is based on Diebold and Mariano (2002). Both test statistics have nonstandard limiting distributions. The corresponding 95% critical values can be found in McCracken (2007). If both the $OOS - F$ and $OOS - T$ tests fail to reject the null hypothesis, it indicates that the forecasts from both models have similar predictive accuracy.

Additionally, the $ENC - New$ test, introduced by Clark and McCracken (2001), is a forecast encompassing test. Under the null hypothesis, the forecasts from model 1 encompass those from model 2, implying that model 2's forecasts provide no additional information. Under the alternative hypothesis, model 2's forecasts contain more information than those from model 1. The test statistic is:

$$\text{ENC} - \text{New} = P \left[\frac{\text{MSFE}_1(\hat{\beta}_1) - \frac{\sum_{t=1}^P \hat{e}_{1,t+1} \hat{e}_{2,t+1}}{P}}{\text{MSFE}_2(\hat{\beta}_2)} \right], \quad (13)$$

where $\hat{e}_{1,t+1}$ and $\hat{e}_{2,t+1}$ are one-step-ahead forecast errors. The 95% critical values can be found in Clark and McCracken (2001). If the $ENC - New$ test rejects the null hypothesis, it suggests that the forecasts from the $AR(1)$ model do not encompass those from the $VAR(1)$ model, indicating that the $VAR(1)$ model's forecasts provide additional information.

3.2 Aggregating the forecasts

In this subsection, the second of the two employed approaches will be presented. This approach consists of aggregating the forecasts, meaning we forecast the subcomponents of the HICP individually and later weight the subcomponents' forecasts to end up with a forecast for the HICP.

To forecast the subcomponents, the same models will be used as for forecasting the aggregate directly, namely univariate and multivariate models. The procedure to fit the models and generate the forecasts will be the same as described in section 3.1, however, the new part to this approach is the weighting scheme. Within this forecasting approach, two different levels of disaggregation will be used: A lower and a higher disaggregation level, meaning we chose 5 different subcomponents of the HICP for the lower disaggregation level and 59 subcomponents for the higher disaggregation level. Depending on the disaggregation level, the corresponding weights need to be retrieved and applied. The weighting procedure can formally be represented as:

$$\hat{y}_{t+h} = \sum_{i=1}^S w_{s,t} \hat{y}_{s,t+h}, \quad (14)$$

where t is the period in which the forecast is being generated, S represents the total number of subcomponents, $w_{s,t}$ represents the weight of the subcomponent s at time t , $\hat{y}_{s,t+h}$ represents the h -step-ahead forecast of the subcomponent s and $\hat{y}_{t+h}$ represents the forecasted aggregate of the HICP, equal to the sum of the weighted subcomponents' forecasts. Important to highlight during this procedure is the usage of the weights at time t and not $t+h$. The problem is that the weights only change annually and when generating multiple-step-ahead forecasts where the forecast horizon lies in the next year, the weight in that period will be unknown at the forecast initiation date. So the weight at the forecast initiation period needs to be used in order to prevent a look-ahead bias.

4 Data

The monthly dataset refers to the German HICP, spanning from January 2016¹ to November 2023, with a total of 95 observations. It includes data for the year-on-year

¹January 2016 marks the start of the observation period, as it is the earliest point from which all required data is consistently available at a monthly frequency.

change in overall HICP, its division into five higher level subcomponents (energy, processed food, unprocessed food, non-energy industrial goods, and services) as well as 59 more granular subcomponents (see Table 21) and the corresponding weights for each subcomponent². Since the weights of the HICP components are only updated annually, the weights were resampled to a monthly frequency to further process them when generating the aggregate forecasts by forecasting the subcomponents. The selection of the five high-level and 59 lower-level subcomponents is based on the work of Duarte and Rua (2007).

For the variables in the *VAR* models, the 3-month Euribor, German long term rates, the 5-, 10- and 30-year Bund yields, M1 and M3 in the Euro Area, German GDP, German unemployment, Brent Oil and Natural Gas prices, the Global Supply Chain Pressure index, German Economic Sentiment, 5-year-5-year inflation swap rates for the Euro Area, and the Baltic Dry Index were chosen for the same frequency and observation period as the remaining data as these variables are expected to exhibit leading behaviour for inflation³. Bund yields were initially observed on a daily basis and had to be resampled by averaging them to convert them to monthly frequency.

In each of the models that are built, year-on-year changes in German HICP are the values that are being predicted, hence, this time series is used as the endogenous variable y_t in every model. In the univariate models, lagged values of HICP y_{t-p} are used as regressors, while in the multivariate models, the external potential predictors from above are used as regressors $x_{j,t-p}$ as well as the lagged values of HICP, y_{t-p} . The FASARIMA model contains principal components as exogenous regressors, extracted from the described variables above.

5 Results

In the following section, the results for forecasting the aggregate versus aggregating the forecasts will be presented.

²The data for the HICP and the subcomponents was retrieved from the ECB Data Portal. The weights were retrieved from EuroStat.

³The 3-month Euribor was retrieved from the ECB Data Portal. M1 and M3, Oil and Gas Prices as well as German GDP were retrieved from the Federal Reserve Bank of St. Louis. The Global Supply Chain Pressure Index is provided by the Federal Reserve Bank of New York. German unemployment rates were retrieved from the German Federal Statistical Office. German long term rates as well as the Bund yields were retrieved from the Central Bank of Germany. German Economic Sentiment, swap rates and the Baltic Dry Index were retrieved from Bloomberg.

5.1 Description of the results

The key question of this work is whether aggregating the forecasts of the German HICP subcomponents improves German HICP forecasts in comparison to forecasting the aggregate directly during the given observation period. Two different disaggregation levels were chosen, one consisted of 5 different subcomponents and the other disaggregation level consisted of 59 more granular subcomponents. Furthermore, this work focuses on improving the HICP forecasts as much as possible by comparing and combining different univariate and multivariate models with each other.

5.1.1 Forecasting the aggregate

First of all, the forecasting results for forecasting the aggregate directly will be described. In Tables 1 and 2, the evaluation metrics can be seen for the iterative forecasts of the univariate and the *VAR* models over different forecasting horizons.

Table 1: Evaluation metrics for iterative forecasts of univariate models to forecast HICP directly

	AR 1	MA 1	ARIMA 1	AR 3	MA 3	ARIMA 3	AR 6	MA 6	ARIMA 6	AR 9	MA 9	ARIMA 9	AR 12	MA 12	ARIMA 12
MSE	1.017	1.057	1.028	4.438	4.326	4.155	10.135	9.902	9.609	21.985	20.881	21.097	31.800	30.123	30.538
RMSE	1.008	1.028	1.014	2.107	2.080	2.038	3.184	3.147	3.100	4.689	4.570	4.593	5.639	5.488	5.526
MAE	0.770	0.792	0.784	1.758	1.782	1.742	2.825	2.846	2.822	4.133	4.081	4.113	4.979	4.926	4.975

Table 2: Evaluation metrics for iterative forecasts of VAR models to forecast HICP directly

	VAR 1	VAR 3	VAR 6	VAR 9	VAR 12
MSE	1.924	3.717	8.892	23.820	89.012
RMSE	1.387	1.928	2.982	4.881	9.435
MAE	1.043	1.582	2.467	4.045	5.524

It can be inferred that the larger the forecast horizon gets, the worse do the forecasts perform as all of the performance metrics constantly increase. Worth noticing is that when looking at Figures 19 and 20 in the appendix, spikes in the forecasts of the *VAR* model can be seen which may be a result of forecast error compounding when performing iterative forecasts. Hence, the MSE numbers for *VAR* models rise sharply for longer forecast horizons. For forecast horizons of 3 and 6 months, all metrics are lower for the

VAR models than for the univariate models, suggesting that *VAR* models outperform univariate models in the short to medium term.

In Tables 3 and 4, the evaluation metrics can be seen for the direct forecasts of the univariate and the *VAR* models over different forecasting horizons.

Table 3: Evaluation metrics for direct forecasts of univariate models to forecast HICP directly

	AR 1	MA 1	ARIMA 1	AR 3	MA 3	ARIMA 3	AR 6	MA 6	ARIMA 6	AR 9	MA 9	ARIMA 9	AR 12	MA 12	ARIMA 12
MSE	1.244	1.576	1.606	4.127	4.583	5.463	9.367	9.678	10.698	18.861	18.396	18.566	28.302	26.883	25.646
RMSE	1.115	1.256	1.267	2.032	2.141	2.337	3.061	3.111	3.271	4.343	4.289	4.309	5.320	5.185	5.064
MAE	0.813	0.996	1.034	1.732	1.824	1.944	2.856	2.908	2.996	3.936	3.906	3.885	4.749	4.631	4.534

Table 4: Evaluation metrics for direct forecasts of VAR models to forecast HICP directly

	VAR 1	VAR 3	VAR 6	VAR 9	VAR 12
MSE	1.023	4.896	9.443	21.250	24.801
RMSE	1.011	2.213	3.073	4.610	4.980
MAE	0.755	1.820	2.769	4.119	4.435

Again and as expected, it can be inferred that the larger the forecast horizon gets, the worse the forecasts perform as all of the performance metrics constantly increase. Figures 49 and 50 in the appendix show that the spikes for VAR models for longer horizons have decreased in comparison to the iterative forecasts in Figures 19 and 20. For forecast horizons of 9 and 12 months, MSE, RMSE and MAE values are lower than for the iterative forecasts and for the remaining, shorter forecast horizons, MSE, RMSE and MAE values are lower for the iterative forecasts than for the direct forecasts.

5.1.2 Aggregating the forecasts

In this part, the results for aggregating the subcomponents' forecasts will be presented. In Tables 5 to 8, the forecasting metrics for iterative and direct forecasts of both model types of the low disaggregation level, i.e. using 5 different subcomponents, can be seen.

Table 5: Evaluation metrics for iterative forecasts of univariate models with low disaggregation level

	AR 1	MA 1	ARIMA 1	AR 3	MA 3	ARIMA 3	AR 6	MA 6	ARIMA 6	AR 9	MA 9	ARIMA 9	AR 12	MA 12	ARIMA 12
MSE	1.371	1.505	1.398	4.960	5.387	5.125	11.590	12.447	12.320	24.153	23.902	27.333	35.285	32.731	38.096
RMSE	1.171	1.227	1.182	2.227	2.321	2.264	3.404	3.528	3.510	4.915	4.889	5.228	5.940	5.721	6.172
MAE	0.874	0.914	0.896	1.863	1.980	1.879	3.093	3.206	3.200	4.420	4.417	4.743	5.483	5.262	5.598

Table 6: Evaluation metrics for iterative forecasts of VAR models with low disaggregation level

	VAR 1	VAR 3	VAR 6	VAR 9	VAR 12
MSE	3.279	4.345	10.279	40.017	229.546
RMSE	1.811	2.085	3.206	6.326	15.151
MAE	1.585	1.585	2.428	4.788	8.194

Table 7: Evaluation metrics for direct forecasts of univariate models with low disaggregation level

	AR 1	MA 1	ARIMA 1	AR 3	MA 3	ARIMA 3	AR 6	MA 6	ARIMA 6	AR 9	MA 9	ARIMA 9	AR 12	MA 12	ARIMA 12
MSE	2.165	3.049	2.086	6.779	19.560	5.997	16.229	39.331	18.854	22.199	21.547	23.359	29.495	29.211	30.948
RMSE	1.471	1.746	1.444	2.604	4.423	2.449	4.029	6.271	4.342	4.712	4.642	4.833	5.431	5.405	5.563
MAE	1.204	1.301	1.112	2.110	2.740	1.981	3.492	4.086	3.753	4.303	4.222	4.456	4.935	4.881	5.060

Table 8: Evaluation metrics for direct forecasts of VAR models with low disaggregation level

	VAR 1	VAR 3	VAR 6	VAR 9	VAR 12
MSE	20.387	8.106	12.335	19.156	23.967
RMSE	4.515	2.847	3.512	4.377	4.896
MAE	3.125	2.236	3.198	3.871	4.323

Similar to the results for forecasting the aggregate directly, it can be observed that the evaluation metrics constantly increase for both model types the higher the forecast horizon gets. Iterative forecasts get even worse when aggregating the subcomponents' forecasts (as forecast errors compound within each model and accumulate during aggregation), especially for the *VAR* models. Again, the direct forecasts outperform the iterative forecasts for forecast horizons of 9 and 12 months. When comparing the forecasts from the low disaggregation level with the forecasts of the aggregate directly, the forecasts

of the low disaggregation approach constantly show higher performance metrics than its counterparts from forecasting the aggregate directly, except for the direct forecasts of the *VAR* model with low disaggregation. For this occurrence, the *VAR* model shows lower performance metrics for the 9 and 12 month forecast horizons than its counterpart from forecasting the aggregate directly (Table 8 vs. Table 4).

In Tables 9 to 12, the forecasting metrics for iterative and direct forecasts of both model types of the high disaggregation level, i.e. 59 different subcomponents, can be seen.

Table 9: Evaluation metrics for iterative forecasts of univariate models with high disaggregation level

	AR 1	MA 1	ARIMA 1	AR 3	MA 3	ARIMA 3	AR 6	MA 6	ARIMA 6	AR 9	MA 9	ARIMA 9	AR 12	MA 12	ARIMA 12
MSE	4.393	4.060	3.943	10.246	7.264	7.108	18.748	15.408	15.145	194.612	24.319	26.941	156.061	34.460	43.175
RMSE	2.096	2.015	1.986	3.201	2.695	2.666	4.330	3.925	3.892	13.950	4.931	5.190	12.492	5.870	6.571
MAE	1.867	1.804	1.757	2.663	2.426	2.437	3.839	3.636	3.549	7.312	4.583	4.835	8.969	5.279	5.849

Table 10: Evaluation metrics for iterative forecasts of VAR models with high disaggregation level

	VAR 1	VAR 3	VAR 6	VAR 9	VAR 12
MSE	4.266	9.964	20.107	35.393	68.609
RMSE	2.066	3.157	4.484	5.949	8.283
MAE	1.799	2.749	4.059	5.190	6.442

Table 11: Evaluation metrics for direct forecasts of univariate models with high disaggregation level

	AR 1	MA 1	ARIMA 1	AR 3	MA 3	ARIMA 3	AR 6	MA 6	ARIMA 6	AR 9	MA 9	ARIMA 9	AR 12	MA 12	ARIMA 12
MSE	3.402	8.732	4.857	8.778	36.106	15.056	15.880	20.499	16.477	25.394	24.195	25.699	32.414	29.508	31.567
RMSE	1.845	2.955	2.204	2.963	6.009	3.880	3.985	4.528	4.059	5.039	4.919	5.069	5.693	5.432	5.618
MAE	1.586	2.386	1.936	2.614	3.968	3.041	3.717	3.777	3.671	4.671	4.537	4.692	5.111	4.859	5.038

Table 12: Evaluation metrics for direct forecasts of VAR models with high disaggregation level

	VAR 1	VAR 3	VAR 6	VAR 9	VAR 12
MSE	31.695	305.572	48.246	51.660	31.253
RMSE	5.630	17.481	6.946	7.187	5.590
MAE	3.951	10.948	5.144	5.471	4.945

Apparently, the higher the disaggregation level gets, the worse get the performance metrics of the forecasts. When looking at Tables 9 to 12, the metrics become significantly higher even for shorter forecast horizons (Table 12) and no improvement can be seen in comparison to the low disaggregation level or forecasting the aggregate directly. This is most likely caused by the even higher forecasting error accumulation when using high disaggregation. Again, the direct forecasts deliver lower evaluation metrics for longer forecast horizons.

Looking at the extreme discrepancies between the different disaggregation levels, it can be concluded that aggregating subcomponents' forecasts does not improve HICP forecasts.

5.1.3 Further improvements for forecasting the aggregate directly

In this part, results will be presented for the measures that were taken to improve the forecasts of the aggregate directly. For this purpose, FASARIMA models and forecast combinations were used.

Table 13: Evaluation metrics for iterative forecasts of the FASARIMA model to forecast HICP directly

	FASARIMA 1	FASARIMA 3	FASARIMA 6	FASARIMA 9	FASARIMA 12
MSE	1.040	4.719	16.458	39.306	89.808
RMSE	1.020	2.172	4.057	6.269	9.477
MAE	0.780	1.678	2.950	4.907	6.830

Table 14: Evaluation metrics for iterative forecasts of univariate models combined with FASARIMA model to forecast HICP directly

	Combined Model 1	Combined Model 3	Combined Model 6	Combined Model 9	Combined Model 12
MSE	0.946	3.744	8.320	20.052	39.929
RMSE	0.973	1.935	2.884	4.478	6.319
MAE	0.728	1.622	2.436	3.933	5.397

Table 15: Evaluation metrics for iterative forecasts of combined univariate models to forecast HICP directly

	Combined Model 1	Combined Model 3	Combined Model 6	Combined Model 9	Combined Model 12
MSE	1.013	4.273	9.850	21.287	30.780
RMSE	1.007	2.067	3.139	4.614	5.548
MAE	0.770	1.761	2.829	4.109	4.960

Table 16: Evaluation metrics for iterative forecasts of univariate models combined with VAR model to forecast HICP directly

	Combined Model 1	Combined Model 3	Combined Model 6	Combined Model 9	Combined Model 12
MSE	1.005	3.940	8.915	20.412	33.766
RMSE	1.003	1.985	2.986	4.518	5.811
MAE	0.754	1.670	2.697	3.966	4.829

Table 17: Evaluation metrics for direct forecasts of combined univariate models to forecast HICP directly

	Combined Model 1	Combined Model 3	Combined Model 6	Combined Model 9	Combined Model 12
MSE	1.253	4.606	9.807	18.499	26.793
RMSE	1.119	2.146	3.132	4.301	5.176
MAE	0.876	1.831	2.920	3.906	4.632

Table 18: Evaluation metrics for direct forecasts of univariate models combined with VAR model to forecast HICP directly

	Combined Model 1	Combined Model 3	Combined Model 6	Combined Model 9	Combined Model 12
MSE	1.094	4.567	9.434	19.126	26.238
RMSE	1.046	2.137	3.071	4.373	5.122
MAE	0.824	1.809	2.881	3.959	4.577

From Table 13 it can be inferred that the *FASARIMA* model itself exhibits higher MSE, RMSE and MAE values than the other univariate models for iterative forecasts. Figures 6 and 7 in the appendix show that the *FASARIMA* models capture the downward trend in HICP, from Fall 2022 onwards, significantly faster which is why considering *FASARIMA* models in the forecast combination might significantly improve the combination of univariate forecasts, which can be seen in Table 14.

Combining univariate models with each other performs somewhat similar to using them individually as can be inferred from the performance metrics in Tables 15 and 17 for the iterative and direct forecasts. Using *VAR* models in combination with univariate models improves the forecasting results remarkably in comparison to using the univariate models individually as can be inferred from Tables 16 and 18, however, the combined forecasts do not perform necessarily better than using the *VAR* models alone. Due to the exogenous data that *FASARIMA* and *VAR* models allow to consider, these models tend

to anticipate switches in trends of inflation significantly faster than the univariate models. However, in episodes of constant up- or downward trend in inflation, the autoregressive models outperform the multivariate models due to their autoregressive nature. Combining both model types appears to offer a superior approach, as it effectively integrates the strengths of each model, particularly under conditions of heightened inflation volatility observed during the study period.

5.2 Classification of the results in the literature context

The findings of this work are mostly in line with the results of existing literature in this field which mostly follow the conclusion that aggregating the subcomponents' forecasts does not improve forecasts for the aggregate. While Hubrich (2005) found that forecasting the subcomponents may improve the aggregate forecasts only for 1-month forecast horizons, and Fritzer et al. (2002) as well as Reijer and Vlaar (2003) found that aggregating the forecasts of the subcomponents may improve results even up to 6 or 7 months ahead, the results from this work do not support any of these conclusions. The main difference between this work and the preceding literature in this field is the observation period and the country for which this framework was tested.

5.3 Economic interpretation of the results

The chosen observation period from 2021 until 2024 to perform the forecasting exercise is the most volatile inflation regime since the mid-1970s, early 1980s or early 1990s. In the 2000s and 2010s, annual HICP was mostly ranging from 0% to 3% while in the past 3 years, the sharp rise and decline in inflation of up to 10% was unmatched in Germany for the past decades. The observation period is characterized by two major exogenous shocks hitting the German economy, namely the Covid-19 pandemic and the Russo-Ukrainian War. While the post-pandemic rise in demand and consumption in combination with expansive fiscal stimuli already put inflationary pressure on the German economy, the Russo-Ukrainian War caused the supply side of energy-related commodities to collapse as Russia was one of the main providers for Germany for Natural Gas, Crude Oil and Coal (Wiertz et al. 2023), leading to a rapid increase in energy prices and ultimately causing German price levels to jump. Also important to mention are the interest rate hikes by the ECB since July 2022 by a total of 450 basis points to bring inflation back to its target

of 2% which was finally achieved again in Q3 of 2024.

Given these two exogenous shocks and the rate hikes, the subcomponents of HICP appear to be affected in a similar way, leading to forecast failures in the same direction and ultimately causing forecast errors to accumulate rather than canceling out when aggregating the forecasts of the subcomponents. This will likely be the main reason why the proposed framework does not perform as desired in comparison to forecasting HICP directly.

When trying to evaluate which model types perform best for forecasting the aggregate directly in the context of exogenous shocks, models containing exogenous information such as the presented *FASARIMA* or *VAR* models anticipate the effects of such shocks clearly better as they incorporate the information of leading inflation indicators, such as commodity prices, inflation swaps, or economic sentiment. In contrast, univariate autoregressive models - due to their nature - do not account for such exogenous data, making them perform well at extrapolating trends but worse at foreseeing switches in trends.

Inflation forecasts are important in capital markets for trading and hedging purposes against future inflation as well as for monetary policy making. Since maintaining price stability is among the main goals of central banks, monitoring future inflation is of extreme importance. Retrospectively looking, for short term predictions, the built forecasts appear to work well, although they rarely manage to predict inflation on the spot. For long-term predictions, the gap between actual inflation and forecasted values tends to grow significantly. While these forecasts may not outperform non-model-based methods such as surveys or implicit inflation expectations derived from financial assets, they can serve as a valuable complement to these tools. In conclusion, the forecasts developed in this work project may not be suitable for standalone use in trading or policymaking, however, when integrated into a broader framework, they provide a model-driven perspective that might improve the evaluation of inflation forecasts.

6 Robustness

In this section, it will be evaluated whether forecasts from certain model types are statistically significant from other model types or combined forecasts.

Regarding robustness in general, it can be concluded that robustness is ensured through the out of sample comparison of forecasts with the actual values. The rolling framework and continuous recalibration of the models allows parameters to change over time. As could be seen in the results in section 5, the MSE, RMSE and MAE values for forecasts from the aggregation of subcomponents' forecasts were significantly higher than for forecasts of the aggregate directly. This is why the focus of this section is on forecasting the aggregate directly, which single models actually deliver more robust forecasting results in comparison to each other and whether forecast combinations significantly outperform individual models. For this purpose, the forecast accuracy and encompassing tests presented in section 3.1 will be applied.

In Tables 19 and 20, the results of the Clark and McCracken (2001, 2007) tests for forecast accuracy and encompassing for both iterative and direct forecasts can be seen, including the results for individual models versus a combination of univariate and multivariate models. The critical values were retrieved from Clark and McCracken (2001).

As can be concluded from Table 19, for every forecast horizon, when comparing the VAR model with any of the univariate models, the test results indicate that not only the VAR forecasts are statistically different from the univariate forecasts but also, they most often do encompass the forecasts from univariate models, suggesting that they provide more predictive information. This is proven by the results of the *ENC – New* test. When comparing the iterative forecasts of individual models with those of the combined model (Table 19), it is interesting to see that according to the *ENC – New* test, the combined model encompasses the VAR model for every forecast horizon. Even for the direct forecasts (Table 20), this appears to be the case for every forecast horizon except the 12 month ahead horizon. In combination with the statistical non-equivalence between most of the individual models and the combined model for both forecasting methods (direct and iterative), it can be concluded that forecast combinations do improve forecasting performance and according to the forecast accuracy and encompassing tests, this finding is also robust.

Table 19: OOS-F, OOS-T and ENC-New Test results for iterative forecasts. The combined model consists of all univariate models and the VAR model. 0 = False, 1 = True.

Forecast Horizon: 1											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	-1.325	-0.394	-16.507	0.383	0.968	-15.779	1.776	-16.296	0.786	31.966	
Significant	0	0	1	0	0	1	0	1	0	1	
OOS-t	-0.449	-0.182	-2.04	0.115	0.528	-1.903	0.424	-1.917	0.191	2.372	
Significant	0	0	1	0	0	1	0	1	0	1	
ENC-New	0.588	0.888	2.088	1.38	1.243	3.863	2.668	4.578	2.588	28.275	
Significant	0	0	1	0	0	1	1	1	1	1	
Forecast Horizon: 3											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	0.856	2.252	6.399	4.169	1.361	5.403	3.229	3.882	1.795	-1.868	
Significant	0	1	1	1	0	1	1	1	0	1	
OOS-t	0.438	1.392	1.312	2.052	1.126	1.215	2.183	0.859	1.202	-0.616	
Significant	0	1	1	1	1	1	1	0	1	0	
ENC-New	0.914	1.537	7.843	2.506	0.931	7.519	2.035	6.705	1.28	1.505	
Significant	0	1	1	1	0	1	1	1	0	1	
Forecast Horizon: 6											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	0.708	1.644	4.194	4.108	0.914	3.406	3.322	2.418	2.336	-0.076	
Significant	0	0	1	1	0	1	1	1	1	0	
OOS-t	0.486	1.26	0.449	1.484	0.944	0.357	1.25	0.255	0.89	-0.011	
Significant	0	1	0	1	1	0	1	0	1	0	
ENC-New	0.535	0.973	8.336	2.49	0.565	8.015	2.103	7.604	1.621	3.475	
Significant	0	0	1	1	0	1	1	1	1	1	
Forecast Horizon: 9											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	1.427	1.136	-2.08	2.08	-0.276	-3.331	0.621	-3.086	0.906	4.508	
Significant	0	0	1	1	0	1	0	1	0	1	
OOS-t	1.446	1.189	-0.251	1.034	-0.654	-0.399	0.339	-0.363	0.442	0.561	
Significant	1	1	0	1	0	0	0	0	0	0	
ENC-New	0.794	0.657	3.521	1.4	-0.107	2.77	0.619	3.192	0.851	5.246	
Significant	0	0	1	0	0	1	0	1	0	1	
Forecast Horizon: 12											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	1.337	0.992	-15.426	-1.397	-0.326	-15.878	-2.589	-15.766	-2.295	39.268	
Significant	0	0	1	0	0	1	1	1	1	1	
OOS-t	1.21	0.934	-0.885	-0.225	-0.924	-0.904	-0.403	-0.897	-0.354	0.979	
Significant	1	1	1	0	1	1	0	0	0	1	
ENC-New	0.731	0.559	0.479	0.605	-0.144	0.402	0.097	0.562	0.313	31.972	
Significant	0	0	0	0	0	0	0	0	0	1	

Table 20: OOS-F, OOS-T and ENC-New Test results for direct forecasts. The combined model consists of all univariate models and the VAR model. 0 = False, 1 = True.

Forecast Horizon: 1											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	-7.381	-7.896	7.573	4.781	-0.653	18.95	15.411	19.976	16.37	-2.296	
Significant	1	1	1	1	0	1	1	1	1	1	
OOS-t	-1.4	-1.172	1.062	0.746	-0.109	2.108	2.221	2.502	2.538	-0.912	
Significant	1	1	1	0	0	1	1	1	1	1	
ENC-New	0.493	2.718	14.94	4.665	10.755	25.844	12.819	21.483	13.099	3.683	
Significant	0	1	1	1	1	1	1	1	1	1	
Forecast Horizon: 3											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	-3.279	-8.071	-5.184	-3.179	-5.32	-2.115	0.112	3.821	6.476	2.379	
Significant	1	1	1	1	1	1	0	1	1	1	
OOS-t	-1.502	-2.33	-1.418	-1.784	-1.387	-0.647	0.062	0.781	1.894	0.867	
Significant	1	1	1	1	1	0	0	0	1	1	
ENC-New	-1.184	-2.73	-0.505	-1.249	-1.123	0.604	0.352	5.36	4.289	2.395	
Significant	0	1	0	0	0	0	0	1	1	1	
Forecast Horizon: 6											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	-0.963	-3.733	-0.241	-0.212	-2.863	0.746	0.776	3.989	4.022	0.029	
Significant	0	1	0	0	1	0	0	1	1	0	
OOS-t	-1.34	-1.674	-0.065	-0.154	-1.341	0.196	0.547	0.831	1.62	0.011	
Significant	1	1	0	0	1	0	0	0	1	0	
ENC-New	-0.394	-1.165	2.191	0.117	-0.852	2.672	0.573	5.065	2.566	1.359	
Significant	0	0	1	0	0	1	0	1	1	0	
Forecast Horizon: 9											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	0.682	0.429	-3.035	-0.374	-0.247	-3.625	-1.03	-3.41	-0.791	2.998	
Significant	0	0	1	0	0	1	0	1	0	1	
OOS-t	0.709	0.279	-1.734	-0.555	-0.16	-2.122	-1.514	-1.499	-0.656	1.922	
Significant	0	0	1	0	0	1	1	1	0	1	
ENC-New	0.424	0.569	-1.298	-0.121	0.156	-1.637	-0.479	-1.28	-0.224	1.626	
Significant	0	0	0	0	0	1	0	1	0	1	
Forecast Horizon: 12											
Test Statistic	AR/MA	AR/ARIMA	AR/VAR	AR/Comb. Mod.	MA/ARIMA	MA/VAR	MA/Comb. Mod.	ARIMA/VAR	ARIMA/Comb. Mod.	VAR/Comb. Mod.	
OOS-F	1.267	2.486	3.389	1.888	1.157	2.015	0.59	0.818	-0.541	-1.315	
Significant	0	1	1	1	0	1	0	0	0	0	
OOS-t	1.018	1.184	2.232	1.944	0.788	1.804	1.425	0.528	-0.457	-1.82	
Significant	1	1	1	1	0	1	1	0	0	1	
ENC-New	0.703	1.63	1.956	1.041	0.756	1.128	0.306	0.683	-0.146	-0.58	
Significant	0	1	1	0	0	0	0	0	0	0	

7 Conclusion

The key question of this paper was to find out whether aggregating the forecasts of the German HICP subcomponents does outperform forecasting the HICP aggregate directly. Earlier research on this topic such as Hubrich (2005) for the Euro Area or Duarte and Rua (2007) for Portugal concluded that in the very short term, aggregating the subcomponents'

forecast does improve overall forecasting performance but for longer forecast horizons, forecasting the aggregate directly performs best. Regarding the forecasting technique, Marcellino et. al. (2005) found that iterated forecasts deliver better forecast results than direct forecasts.

In this work, monthly forecasts were being created by using a rolling framework with a fixed window size of 60 months and for an observation period from January 2021 until November 2023. This observation period represents a regime of high and volatile inflation, particularly difficult to forecast, mainly characterized by two exogenous shocks: The Covid-19 pandemic and the Russo-Ukrainian war, which caused post-pandemic consumption to increase and supply chains to disrupt, ultimately leading to high levels of inflation.

The applied framework in this paper is closely related to the work from Duarte and Rua (2007) as two different disaggregation levels were chosen: A lower disaggregation level considering 5 different subcomponents of the German HICP and a higher disaggregation level considering 59 different subcomponents. Two forecasting techniques, namely iterated and direct forecasts, were applied in this work to find the best possible among both. In addition to the key question of this paper, the forecasts for the aggregate directly were tried to be improved by including factors in the autoregressive models and applying forecast combinations to possibly merge the best of both univariate and multivariate models. Forecast accuracy and encompassing tests proposed by Clark and McCracken (2001, 2007) were applied to determine the statistical significance in accuracy and possible encompassing between different forecasting models.

The results of the research conducted are mostly in line with what the preceding literature provides. Aggregating the subcomponents' forecasts for the case of Germany does not deliver better forecasting results for the HICP aggregate than forecasting the aggregate alone as the forecasting errors per subcomponent accumulate because they are being impacted by exogenous shocks in a similar way. Also, iterative forecasts outperform direct forecasts in the short- and medium term, however for forecasting horizons of 9 months and above, direct forecasts perform better as for such horizons, the forecasting errors do not compound over the forecasting periods. Using factor augmented models improves the forecasting performance when combining them with univariate models as the exogenous data allows to anticipate switches in inflationary trends better than simple

autoregressive models. Combining forecasts of univariate and *VAR* models does improve forecasting performance as well in comparison to using univariate models alone as this method also combines the advantages of both autoregressive and multivariate models. In conclusion, it can be inferred that for forecasting the HICP of Germany, forecasting the time series directly works best and considering exogenous information in the models allows to better anticipate switches in inflationary trends. Direct forecasting is useful for forecasting inflation further in the future, otherwise iterative forecasting is recommended.

7.1 Limitations, recommendation for further research

This work mainly focused on forecasting the HICP aggregate directly versus aggregating the forecasts of the HICP subcomponents and determining which approach delivers the better overall forecast for the HICP aggregate. After it was found that the disaggregation of the HICP does not help in improving HICP forecasts, further techniques were applied to improve the forecasts for the aggregate directly, such as incorporating factors into univariate models or combining the forecasts of univariate and multivariate models.

For further research on this topic, it would be interesting to see whether the framework might work better in a more stable inflation regime, such as in the years from 2010 until 2020 where less severe exogenous shocks occurred. This might give a good indication of whether this approach is doomed to fail overall or if the failure of the approach in this work can be attributed to the regime for which it was applied. If one wants to further observe which models are generally the best for forecasting HICP directly, it might be worth considering variations of the used models in this work or other macroeconomic models such as random walk models, dynamic stochastic general equilibrium (DSGE) models or even machine learning algorithms.

References

- Clark, T.E. and McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1), pp.85-110.
- Cobb, M.P., 2020. Aggregate density forecasting from disaggregate components using Bayesian VARs. *Empirical Economics*, 58(1), pp.287-312.
- Dickey, D.A. and Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a), pp.427-431.
- Diebold, F.X., 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, 33(1), pp.1-1.
- Diebold, F.X. and Mariano, R.S., 2002. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1), pp.134-144.
- den Reijer, A.H.J. and Vlaar, P.J.G., 2003. Forecasting Inflation in the Netherlands and the Euro Area (No. 723). *Netherlands Central Bank, Research Department*.
- Duarte, C. and Rua, A., 2007. Forecasting inflation through a bottom-up approach: How bottom is bottom?. *Economic Modelling*, 24(6), pp.941-953.
- Fritzer, F., Moser, G. and Scharler, J., 2002. Forecasting Austrian HICP and its components using VAR and ARIMA models (No. 73). *Working paper*.
- Hendry, D.F. and Hubrich, K., 2011. Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of business & economic statistics*, 29(2), pp.216-227.
- Hubrich, K., 2005. Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy?. *International Journal of Forecasting*, 21(1), pp.119-136.
- Kuha, J., 2004. AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2), pp.188-229.
- Marcellino, M., Stock, J.H., Watson, M.W., 2003. Macroeconomic forecasting in the Euro area: country specific versus area-wide information. *European Economic Review* 47, 1–18.
- Marcellino, M., Stock, J.H. and Watson, M.W., 2006. A comparison of direct and

- iterated multistep AR methods for forecasting macroeconomic time series. *Journal of econometrics*, 135(1-2), pp.499-526.
- McCracken, M.W., 2007. Asymptotics for out of sample tests of Granger causality. *Journal of econometrics*, 140(2), pp.719-752.
- Shojaie, A. and Fox, E.B., 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1), pp.289-319.
- Stock, J.H. and Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2), pp.147-162.
- Timmermann, A., 2006. Forecast combinations. *Handbook of economic forecasting*, 1, pp.135-196.
- Wiertz, T., Kuhn, L. and Mattissek, A., 2023. A turn to geopolitics: Shifts in the German energy transition discourse in light of Russia's war against Ukraine. *Energy Research & Social Science*, 98, p.103036.
- Zarnowitz, V. and Braun, P., 1993. Twenty-two years of the NBER-ASA quarterly economic outlook surveys: aspects and comparisons of forecasting performance. In Business cycles, indicators, and forecasting (pp. 11-94). *University of Chicago Press*.

Appendix

Table 21: Subcomponents for high level disaggregation

Subcomponent	Subcomponent
1. Accommodation services	31. Maintenance and repair of personal transport equipment
2. Actual rentals for housing	32. Major household appliances whether electric or not
3. Administered prices	33. Margarine and other vegetable fats
4. Alcoholic beverages	34. Milk, cheese and eggs
5. Bread	35. Mineral waters, soft drinks, fruit and vegetable juices
6. Bread and cereals	36. Motor cars
7. Butter	37. Motor cycles, bicycles and animal drawn vehicles
8. Catering services	38. Newspapers, books and stationery
9. Cleaning, repair and hire of clothing	39. Non-durable household goods
10. Clothes washing machines, clothes drying machines and dish washing machines	40. Other bakery products
11. Clothing materials	41. Other cereal products
12. Cocoa and powdered chocolate	42. Other edible oils
13. Coffee, tea and cocoa	43. Other meats
14. Dried, smoked or salted fish and seafood	44. Other milk products
15. Education	45. Package holidays
16. Electric appliances for personal care	46. Pasta products and couscous
17. Equipment for the reception, recording and reproduction of sound and picture	47. Potatoes
18. Footwear	48. Poultry
19. Fresh or chilled fish	49. Recreational and cultural services
20. Fresh or chilled seafood	50. Refrigerators, freezers and fridge-freezers
21. Fresh or chilled vegetables other than potatoes and other tubers	51. Repair of furniture, furnishings and floor coverings
22. Frozen vegetables other than potatoes and other tubers	52. Repair of household appliances
23. Fruit	53. Rice
24. Furniture and furnishings, carpets and other floor coverings	54. Sauces, condiments
25. Gas	55. Sugar, jam, honey, chocolate and confectionery
26. Glassware, tableware and household utensils	56. Tea
27. Hairdressing salons and personal grooming establishments	57. Telephone and telefax equipment
28. Household textiles	58. Therapeutic appliances and equipment
29. Insurance	59. Yoghurt
30. Lamb and goat	

Figure 1: 1-month ahead iterative forecasts using univariate models for forecasting HICP directly

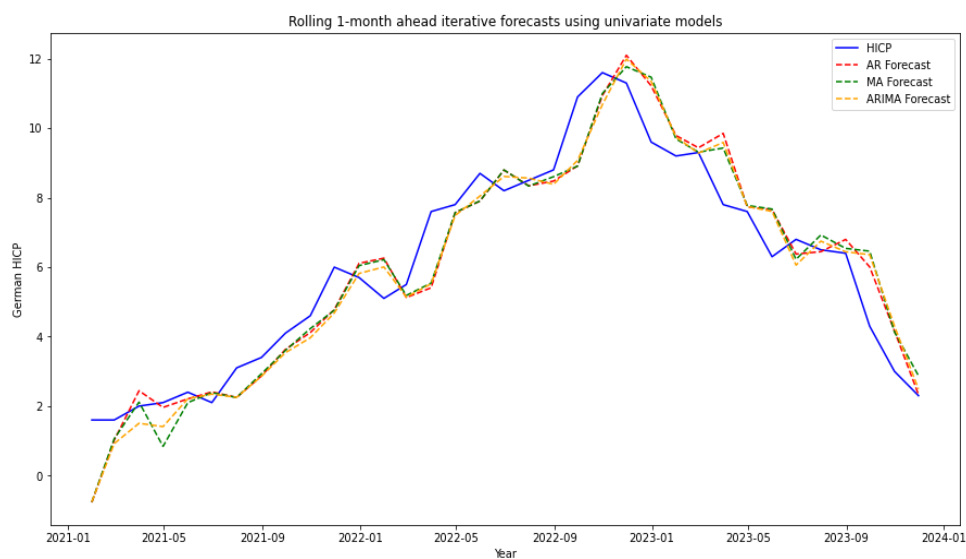


Figure 2: 3-month ahead iterative forecasts using univariate models for forecasting HICP directly

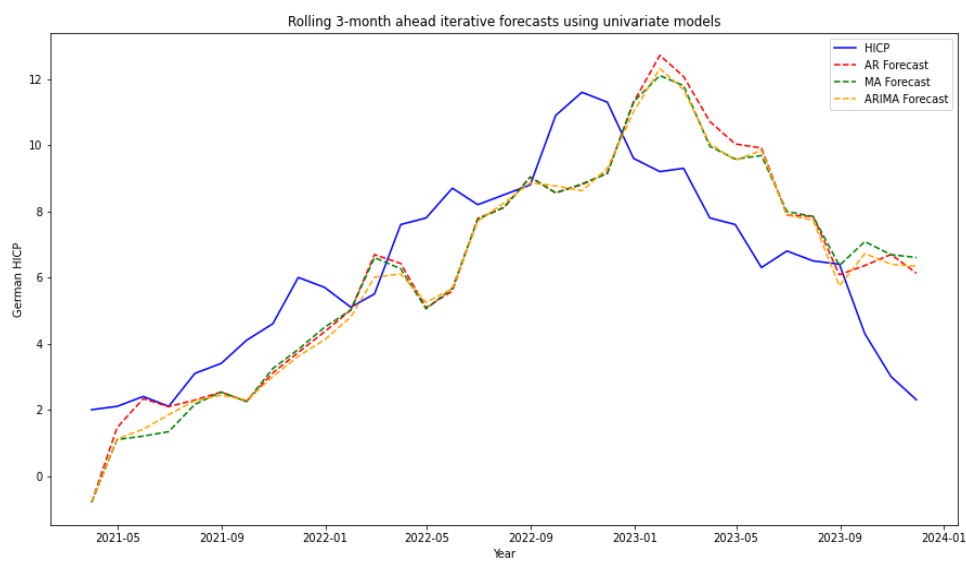


Figure 3: 6-month ahead iterative forecasts using univariate models for forecasting HICP directly

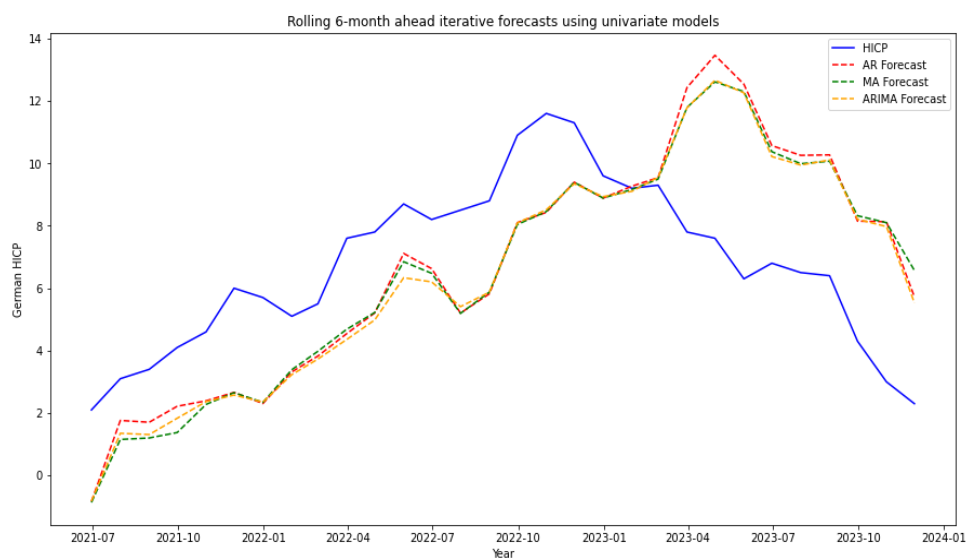


Figure 4: 9-month ahead iterative forecasts using univariate models for forecasting HICP directly

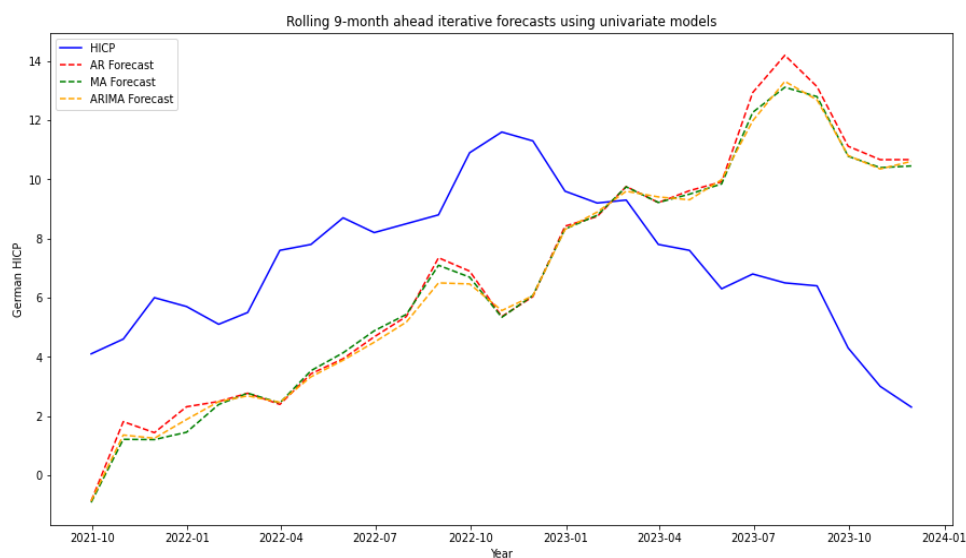


Figure 5: 12-month ahead iterative forecasts using univariate models for forecasting HICP directly

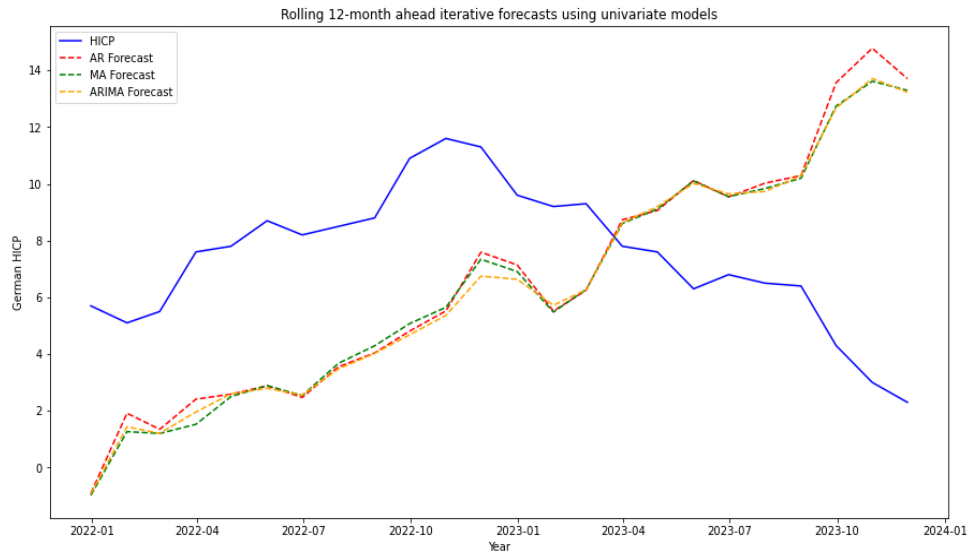


Figure 6: 1-month ahead iterative forecasts using FASARIMA models for forecasting HICP directly

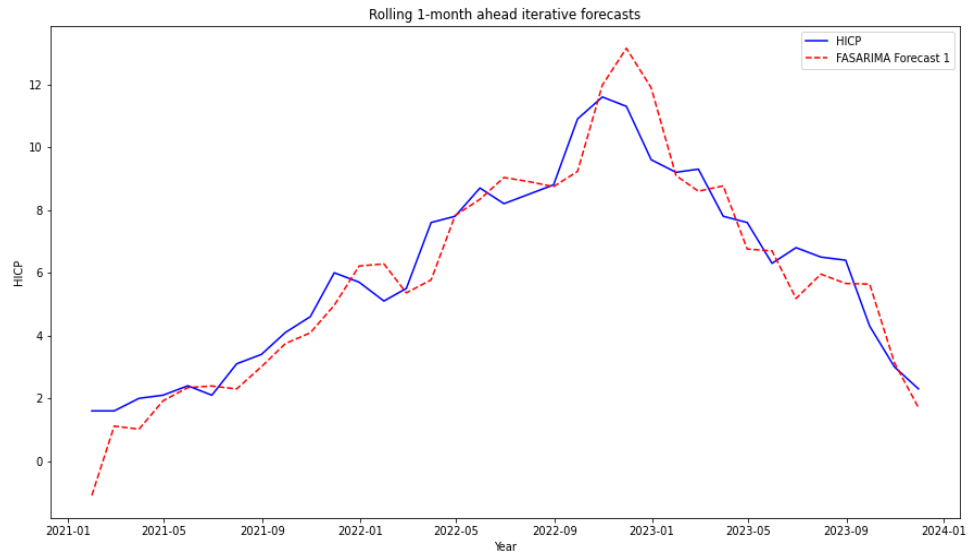


Figure 7: 3-month ahead iterative forecasts using FASARIMA models for forecasting HICP directly

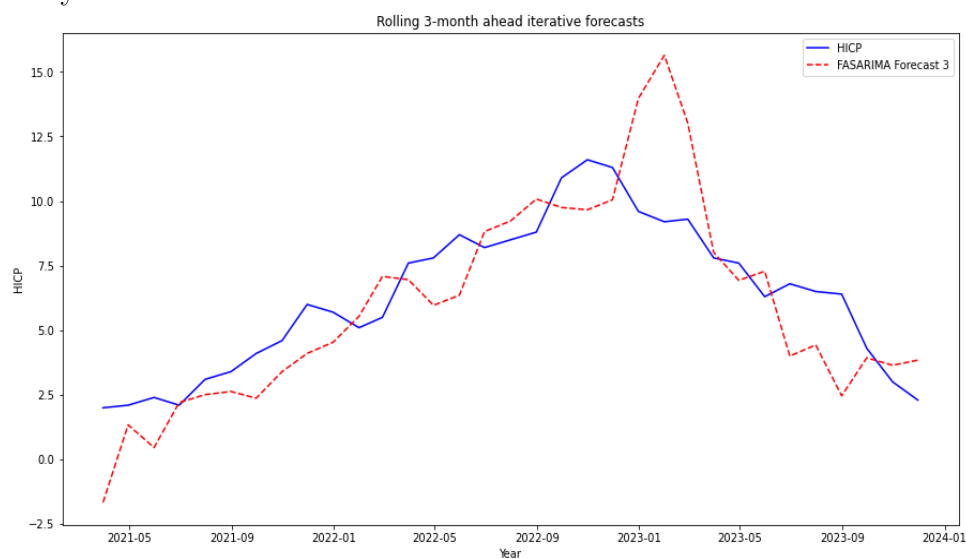


Figure 8: 6-month ahead iterative forecasts using FASARIMA models for forecasting HICP directly

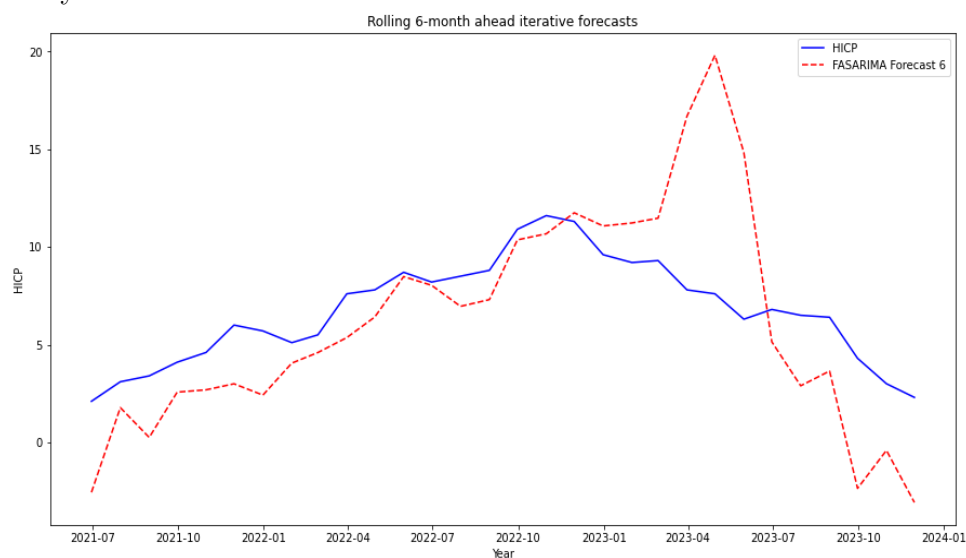


Figure 9: 9-month ahead iterative forecasts using FASARIMA models for forecasting HICP directly

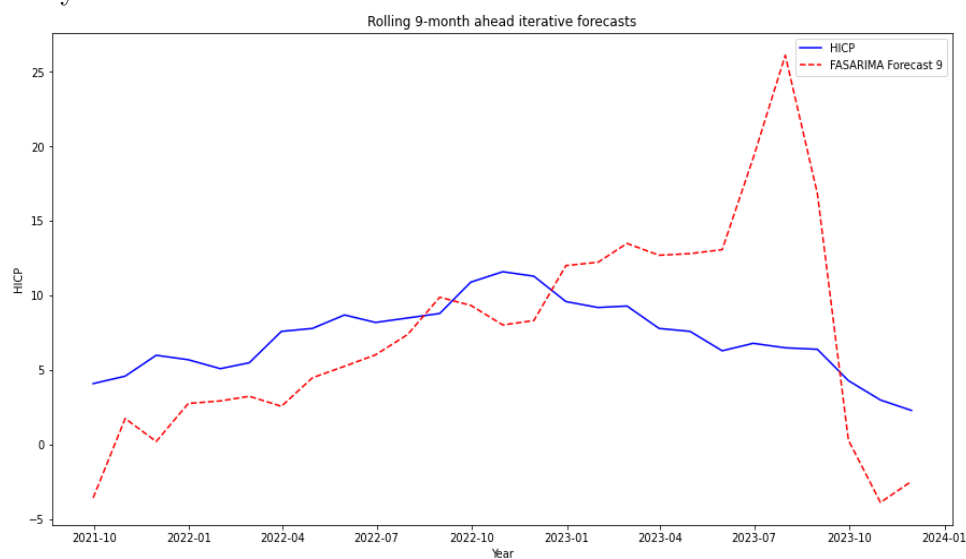


Figure 10: 12-month ahead iterative forecasts using FASARIMA models for forecasting HICP directly

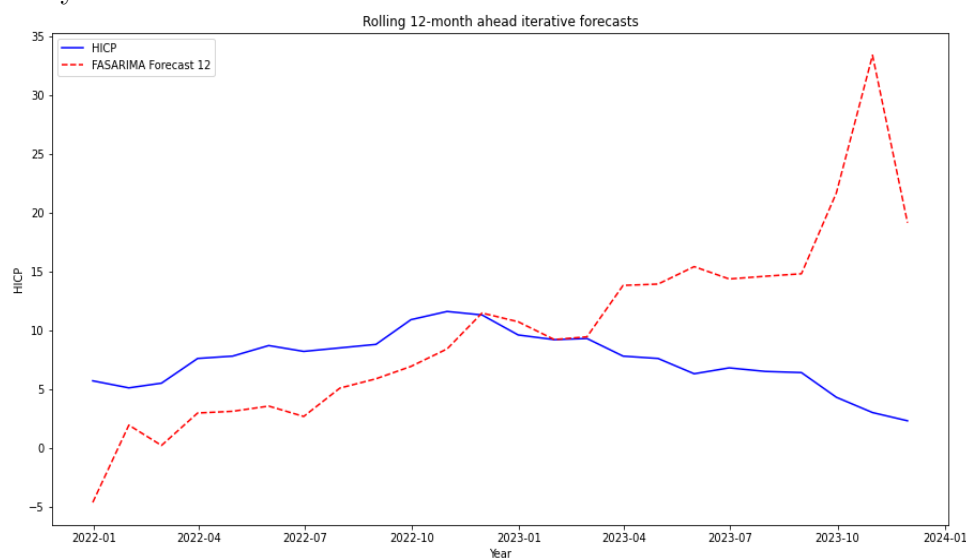


Figure 11: 1-month ahead iterative forecasts using a combination of univariate and VAR models for forecasting HICP directly

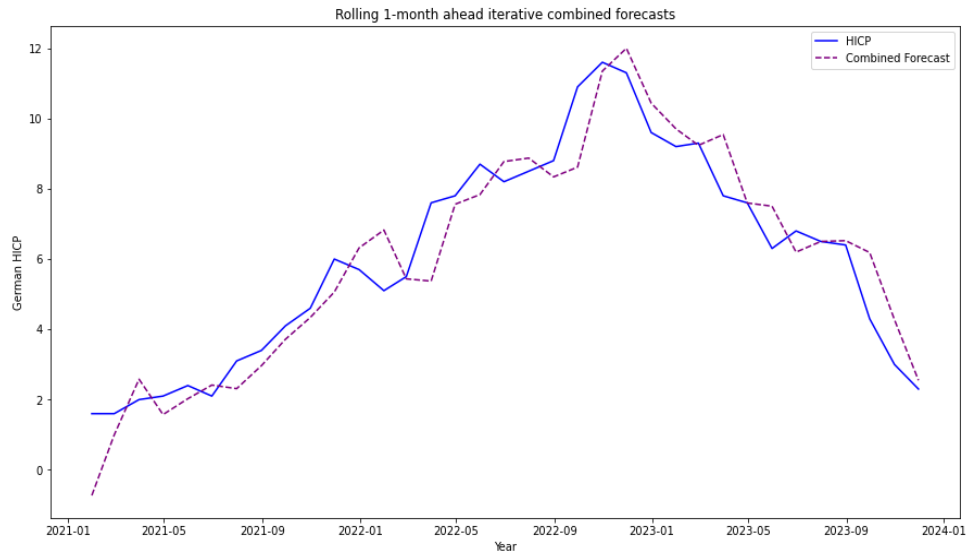


Figure 12: 3-month ahead iterative forecasts using a combination of univariate and VAR models for forecasting HICP directly

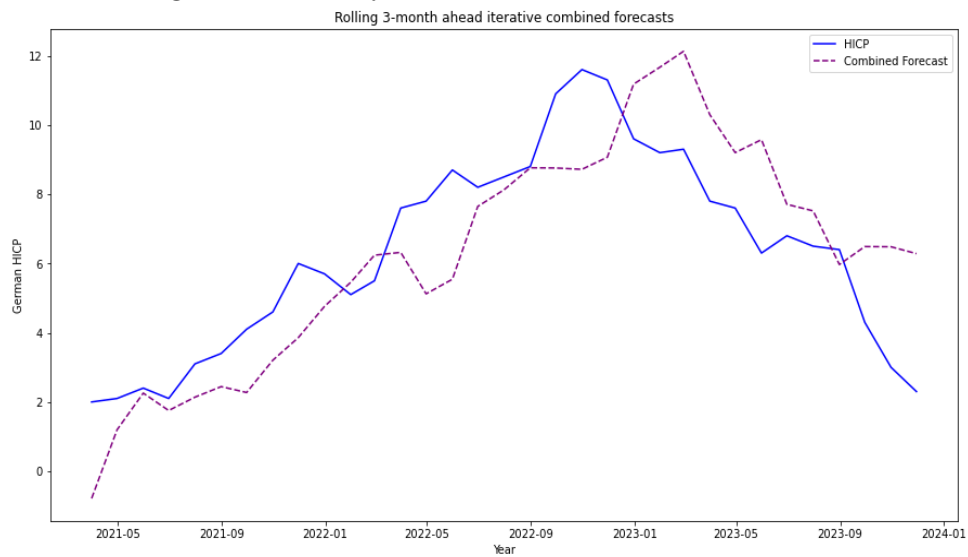


Figure 13: 6-month ahead iterative forecasts using a combination of univariate and VAR models forecasting HICP directly

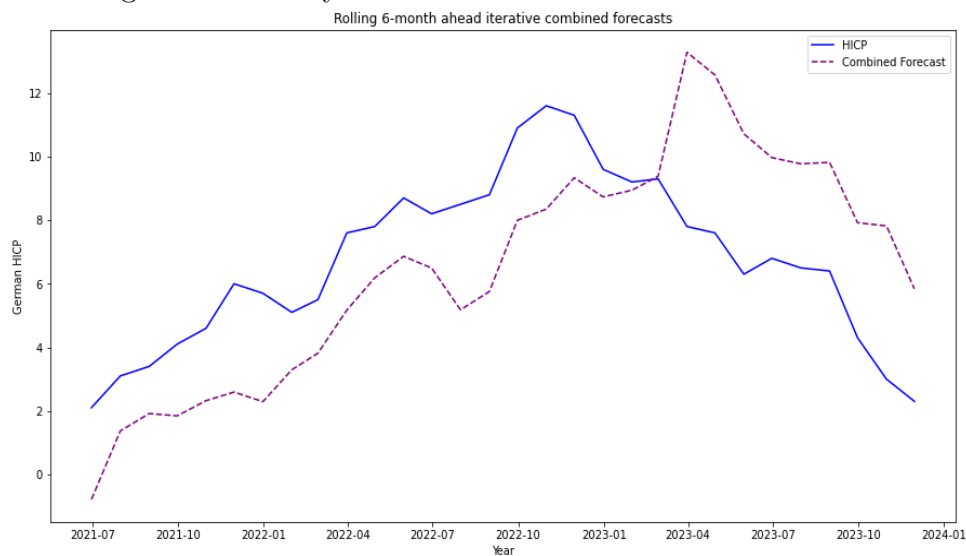


Figure 14: 9-month ahead iterative forecasts using a combination of univariate and VAR models for forecasting HICP directly

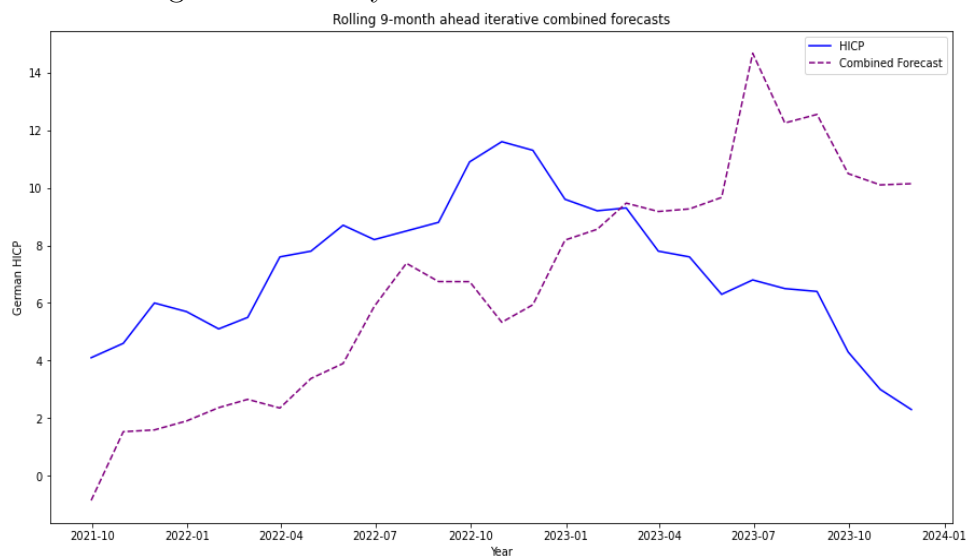


Figure 15: 12-month ahead iterative forecasts using a combination of univariate and VAR models for forecasting HICP directly

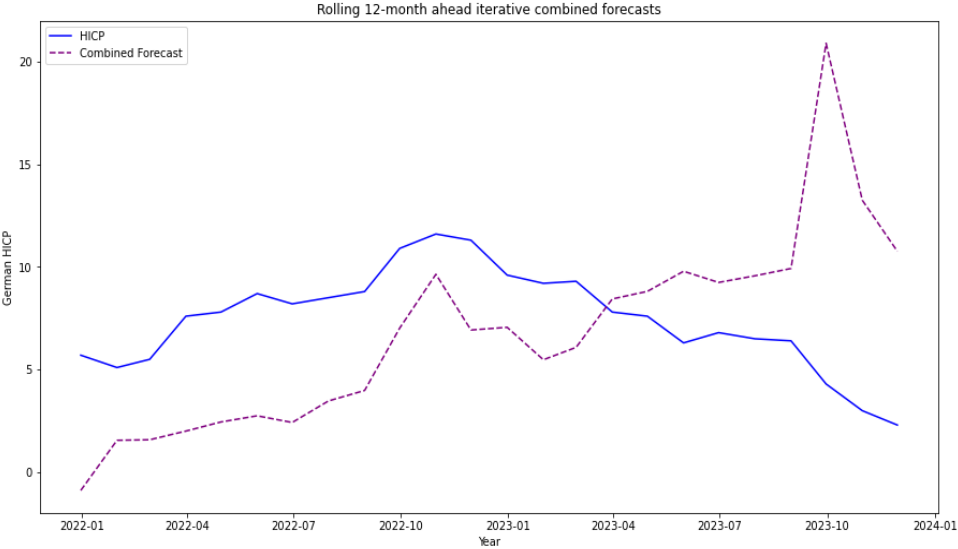


Figure 16: 1-month ahead iterative forecasts using VAR models for forecasting HICP directly

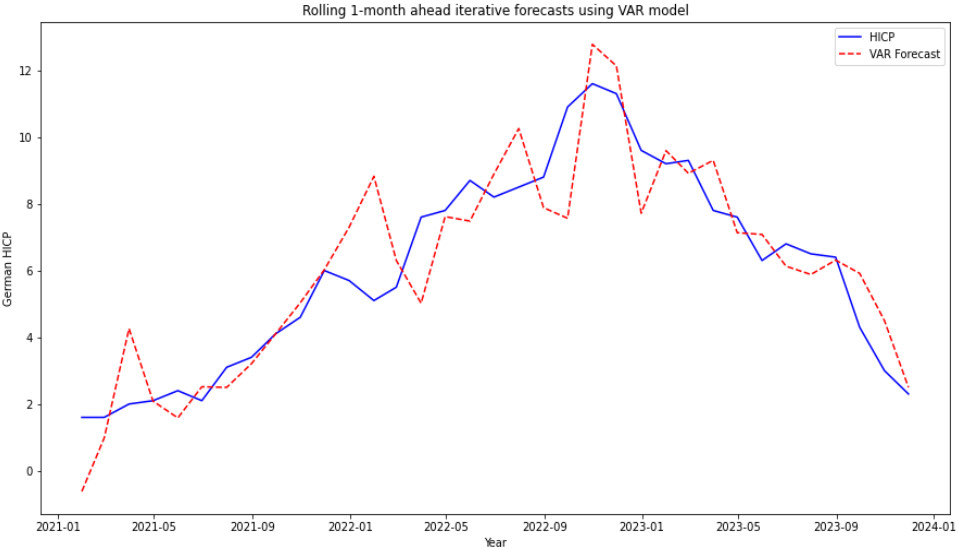


Figure 17: 3-month ahead iterative forecasts using VAR models for forecasting HICP directly

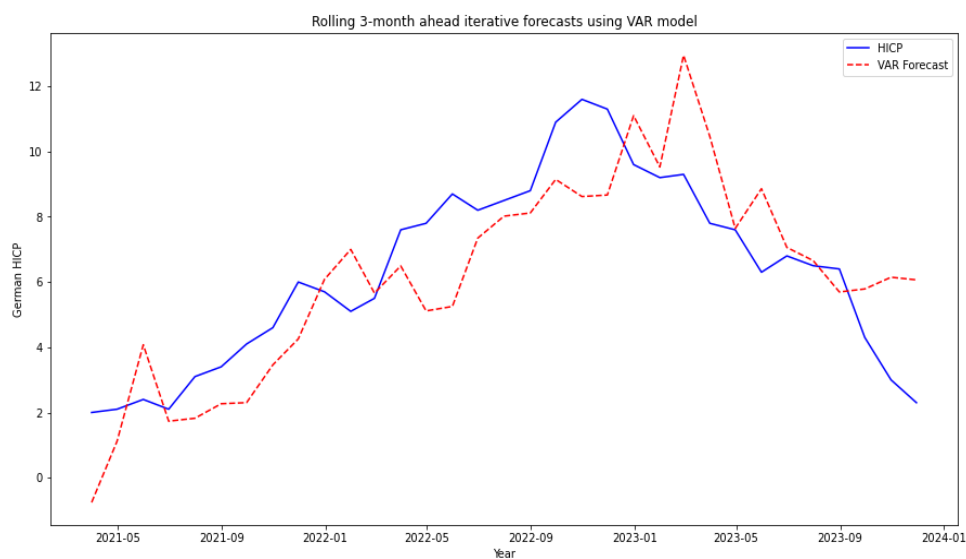


Figure 18: 6-month ahead iterative forecasts using VAR models for forecasting HICP directly

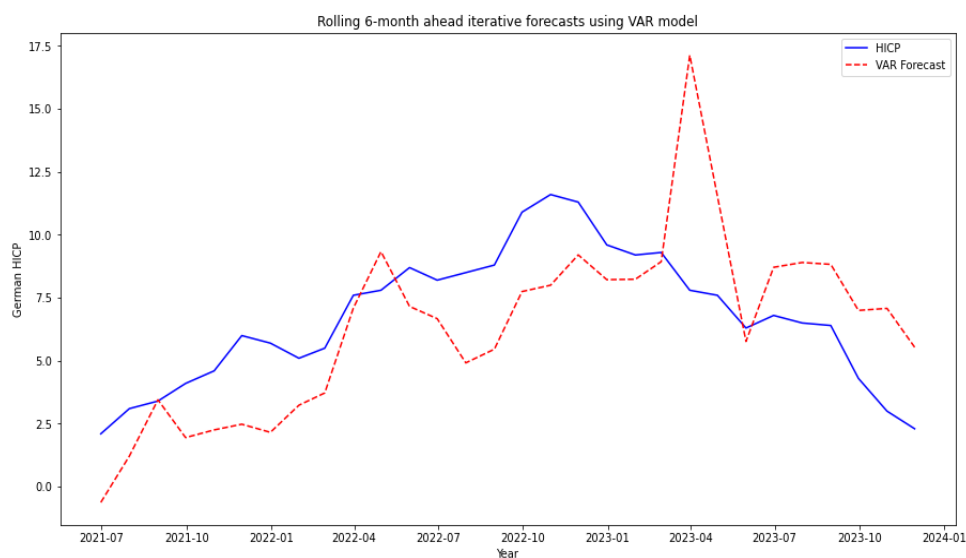


Figure 19: 9-month ahead iterative forecasts using VAR models for forecasting HICP directly

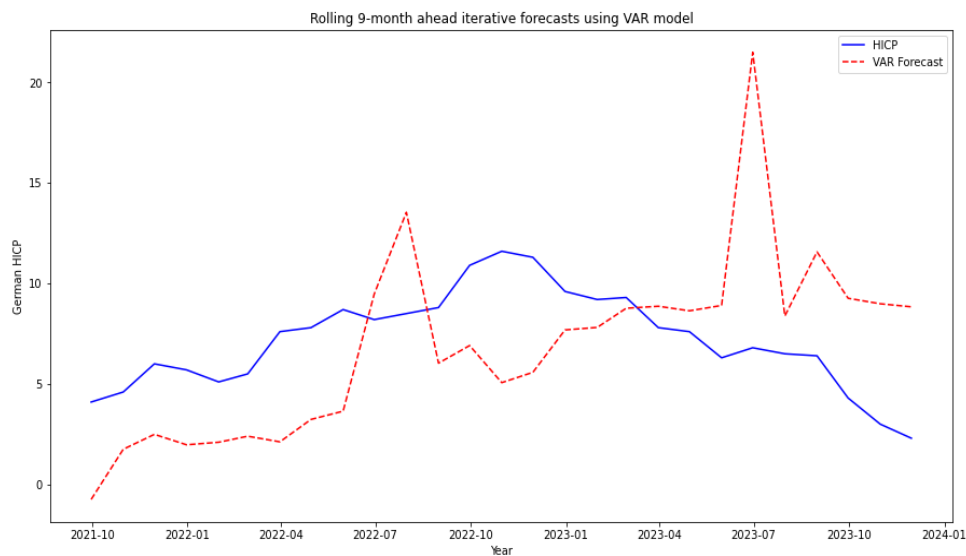


Figure 20: 12-month ahead iterative forecasts using VAR models for forecasting HICP directly

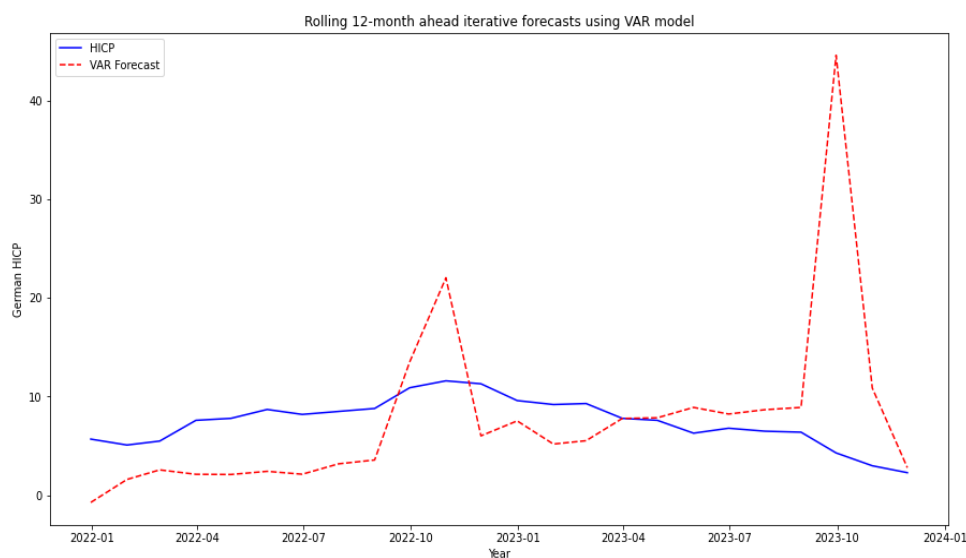


Figure 21: 1-month ahead iterative forecasts using univariate models with low disaggregation level

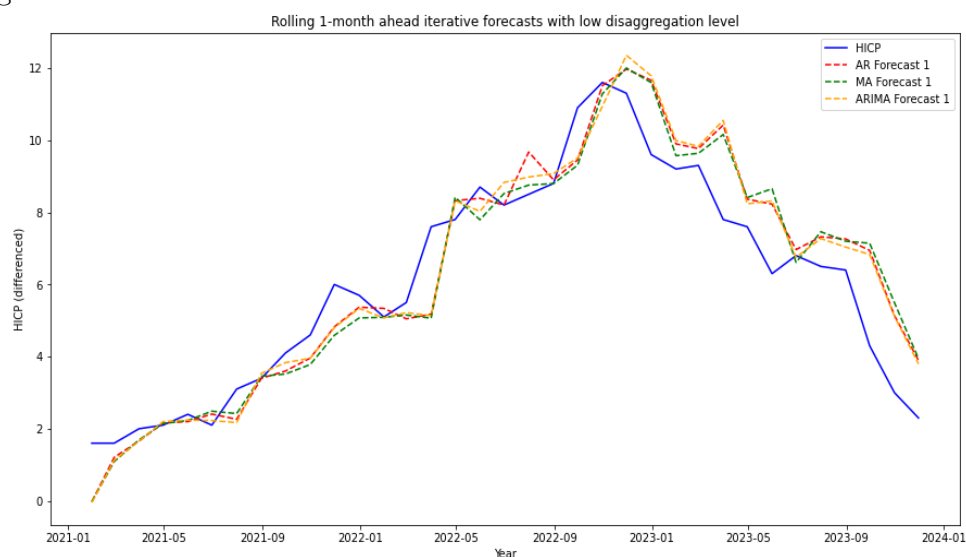


Figure 22: 3-month ahead iterative forecasts using univariate models with low disaggregation level

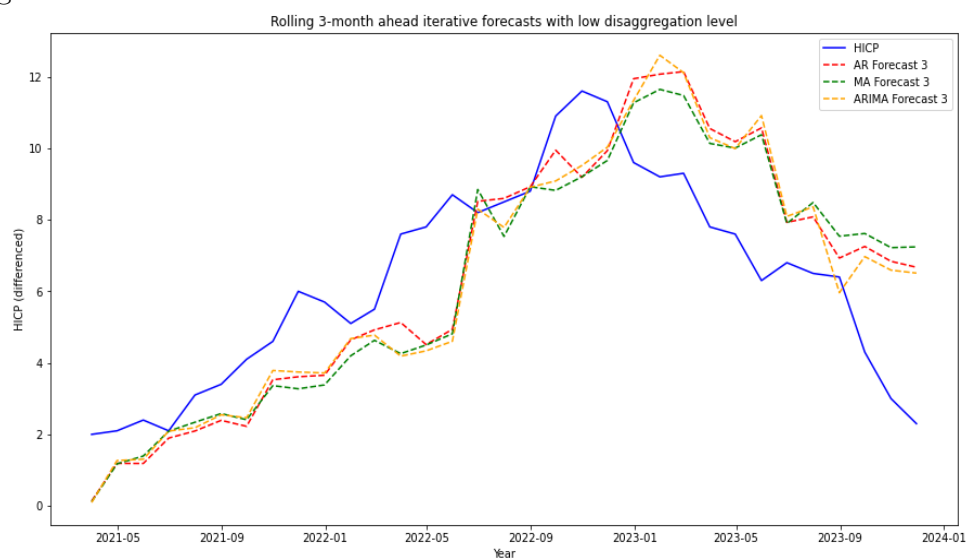


Figure 23: 6-month ahead iterative forecasts using univariate models with low disaggregation level

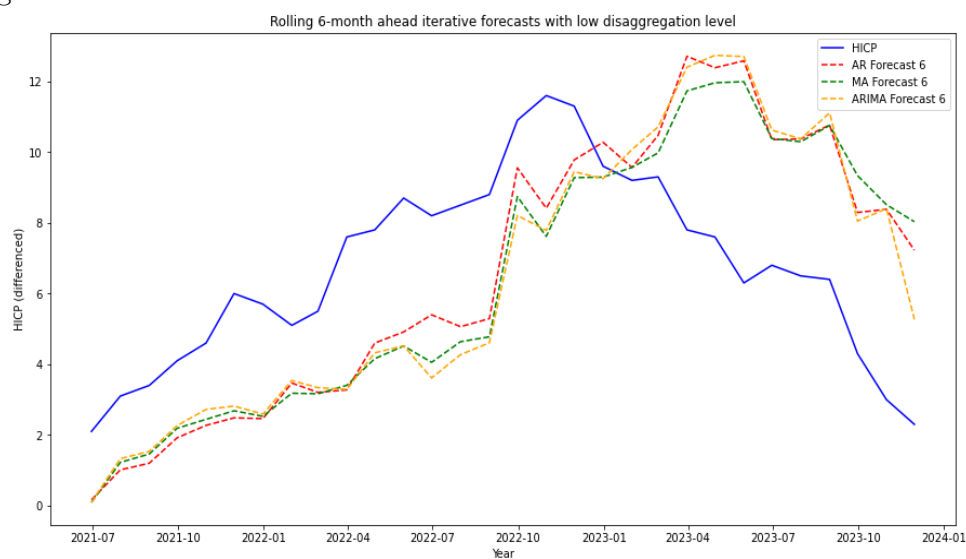


Figure 24: 9-month ahead iterative forecasts using univariate models with low disaggregation level

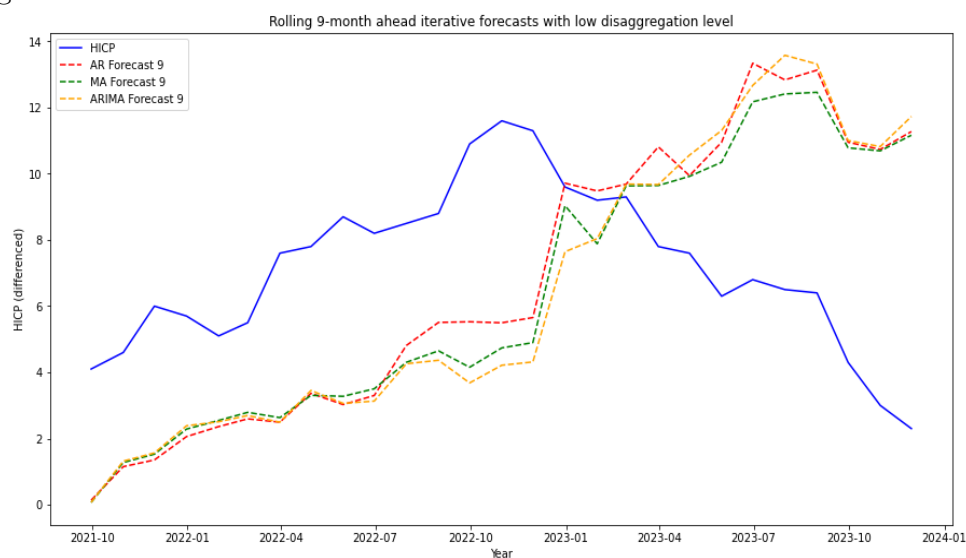


Figure 25: 12-month ahead iterative forecasts using univariate models with low disaggregation level

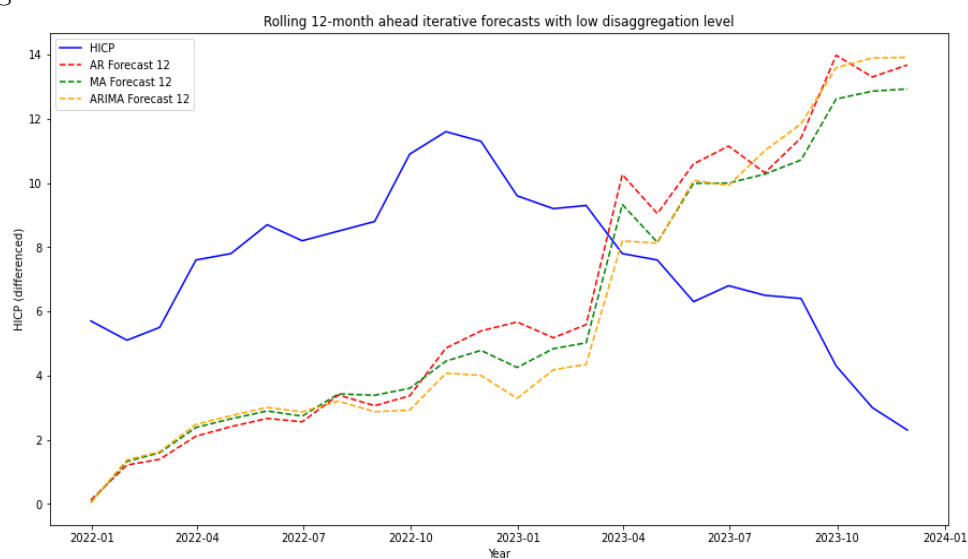


Figure 26: 1-month ahead iterative forecasts using VAR models with low disaggregation level

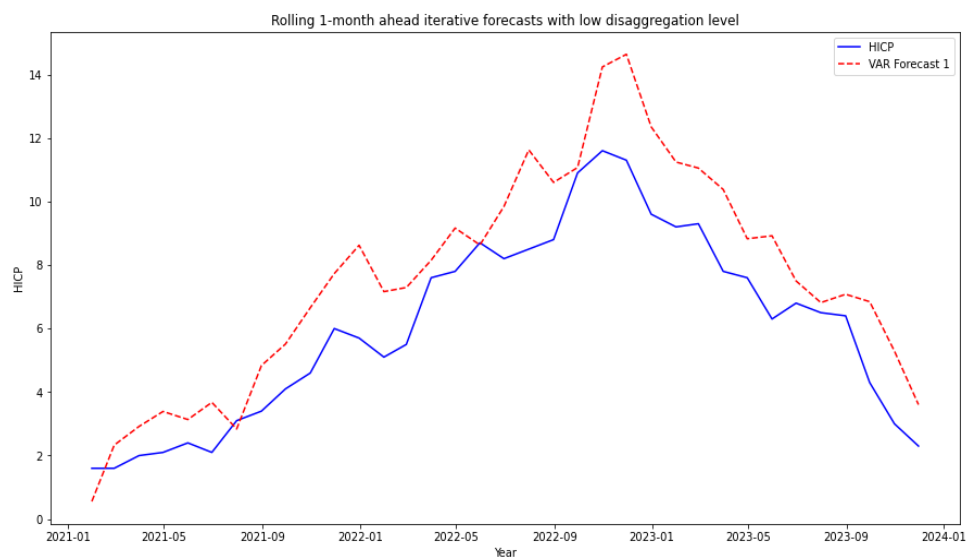


Figure 27: 3-month ahead iterative forecasts using VAR models with low disaggregation level

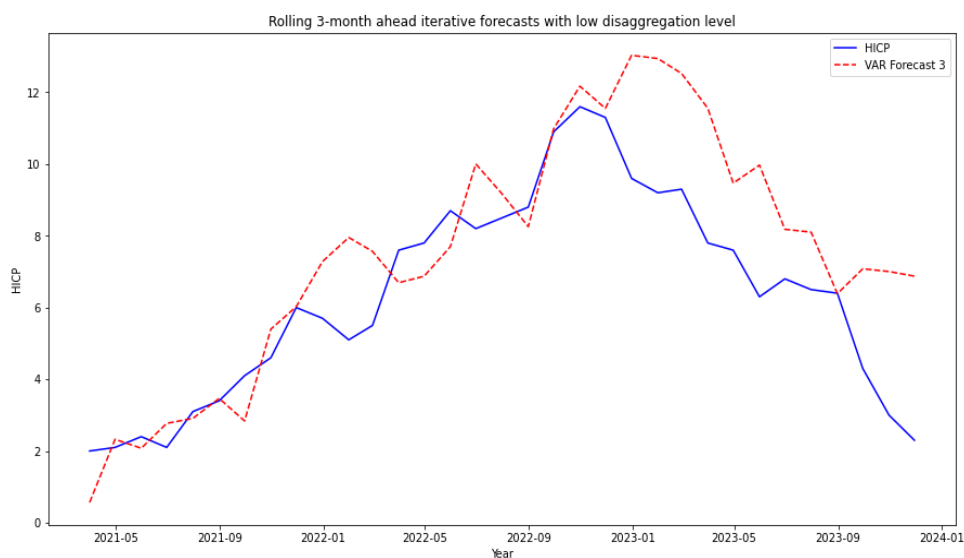


Figure 28: 6-month ahead iterative forecasts using VAR models with low disaggregation level

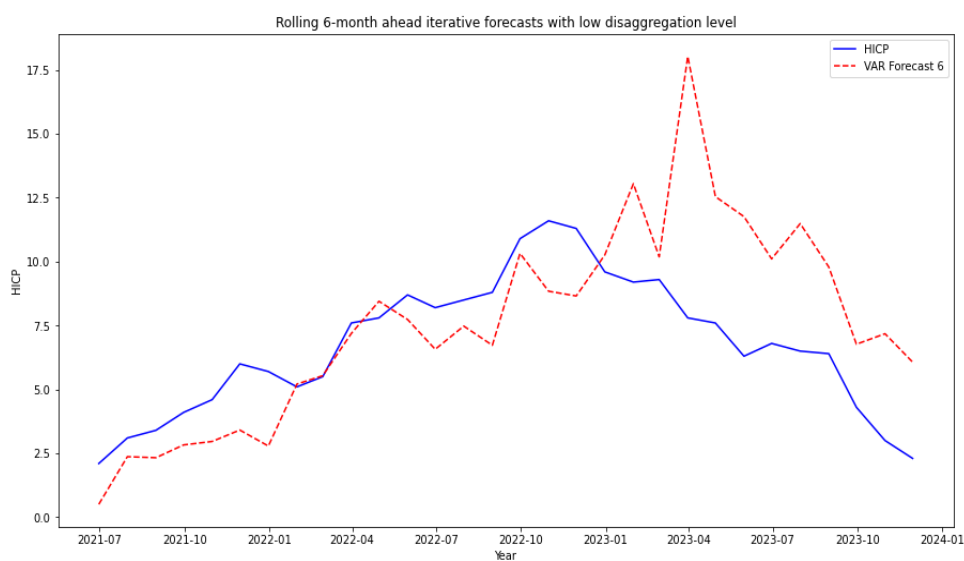


Figure 29: 9-month ahead iterative forecasts using VAR models with low disaggregation level

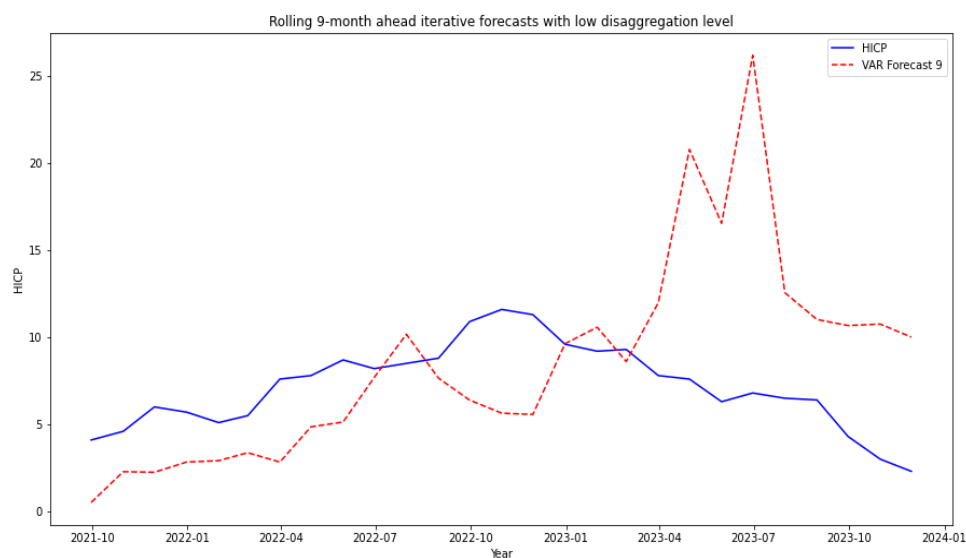


Figure 30: 12-month ahead iterative forecasts using VAR models with low disaggregation level

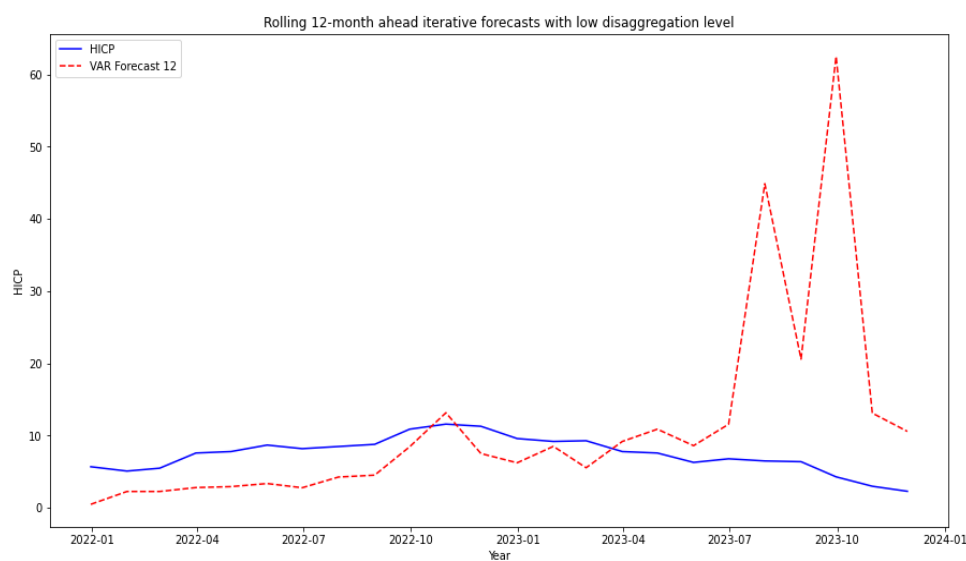


Figure 31: 1-month ahead iterative forecasts using univariate models with high disaggregation level

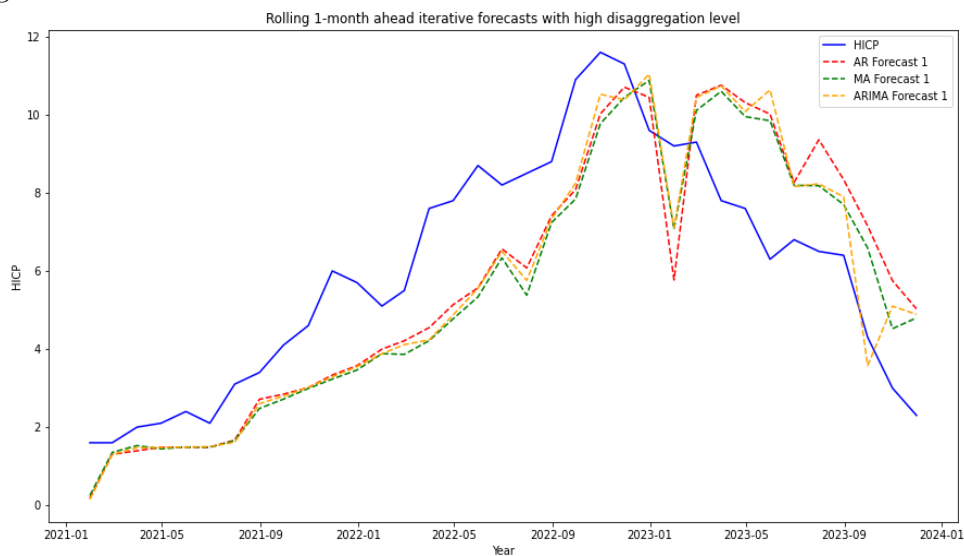


Figure 32: 3-month ahead iterative forecasts using univariate models with high disaggregation level

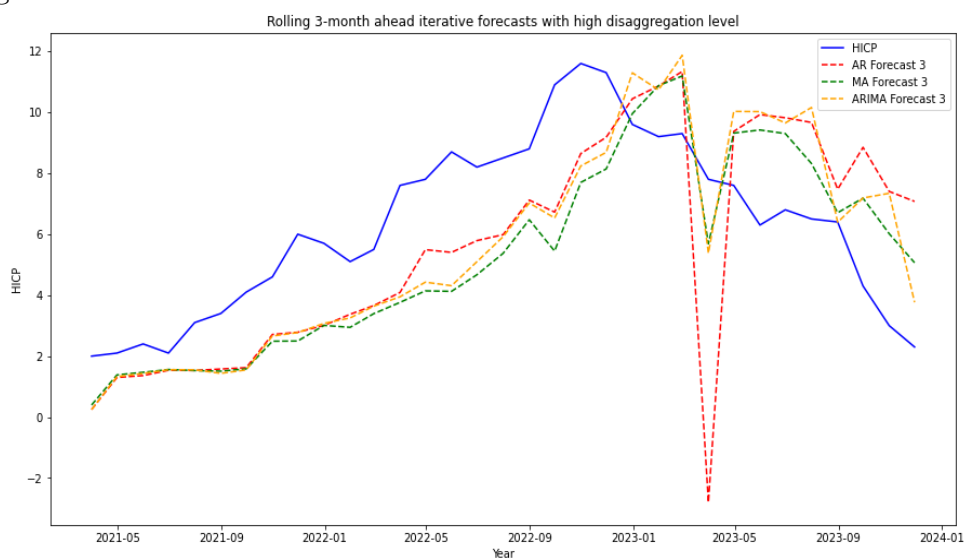


Figure 33: 6-month ahead iterative forecasts using univariate models with high disaggregation level

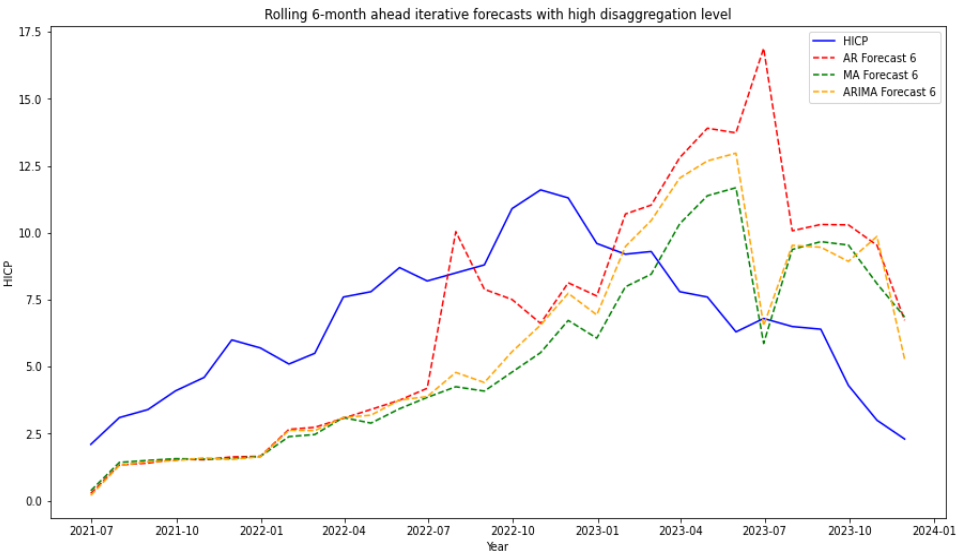


Figure 34: 9-month ahead iterative forecasts using univariate models with high disaggregation level

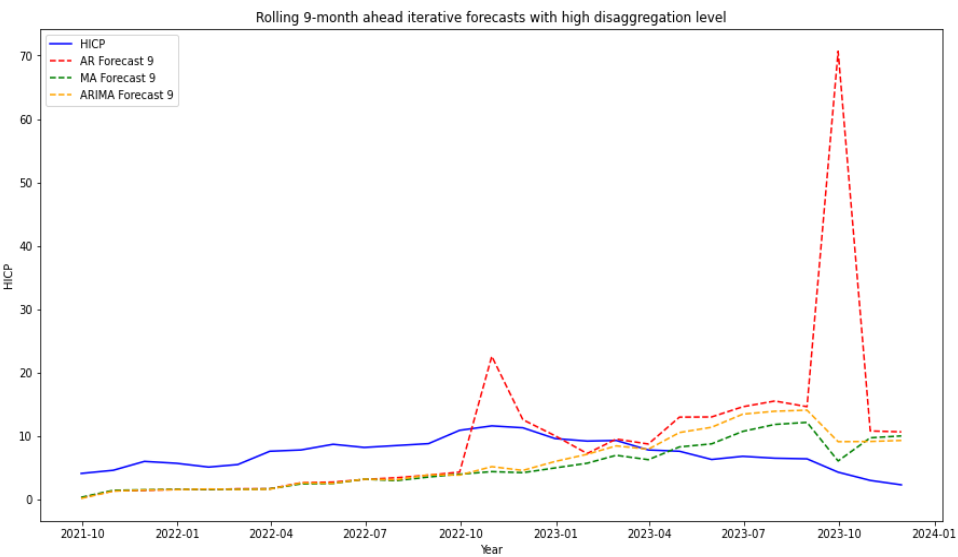


Figure 35: 12-month ahead iterative forecasts using univariate models with high disaggregation level

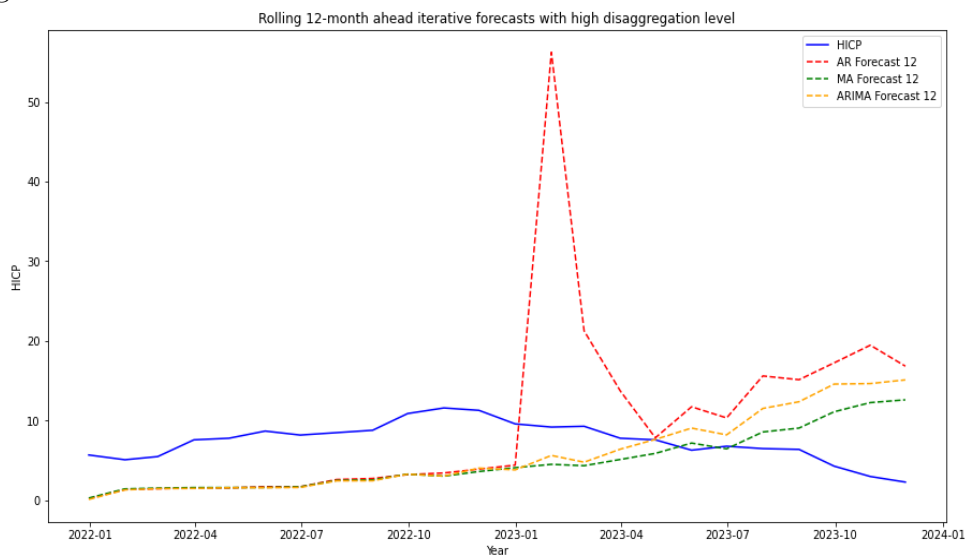


Figure 36: 1-month ahead iterative forecasts using VAR models with high disaggregation level

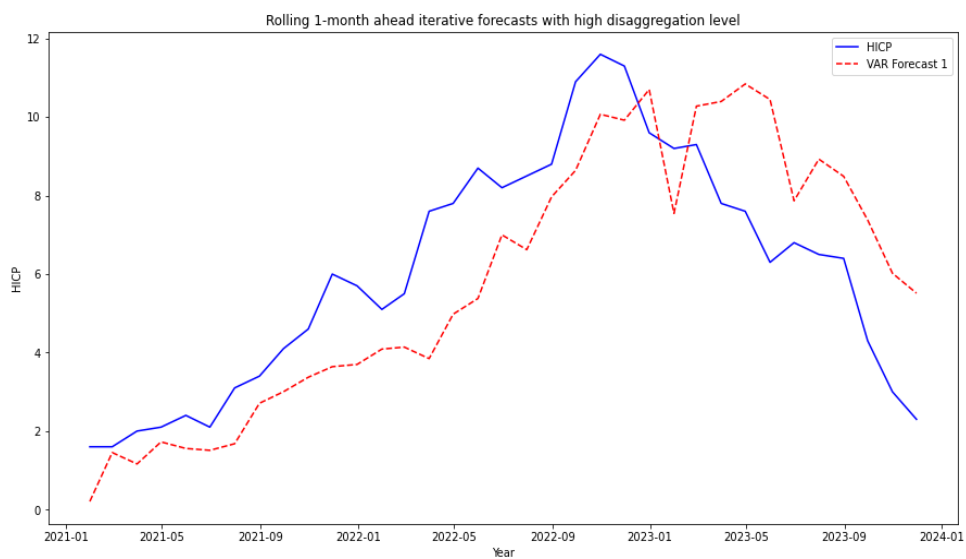


Figure 37: 3-month ahead iterative forecasts using VAR models with high disaggregation level

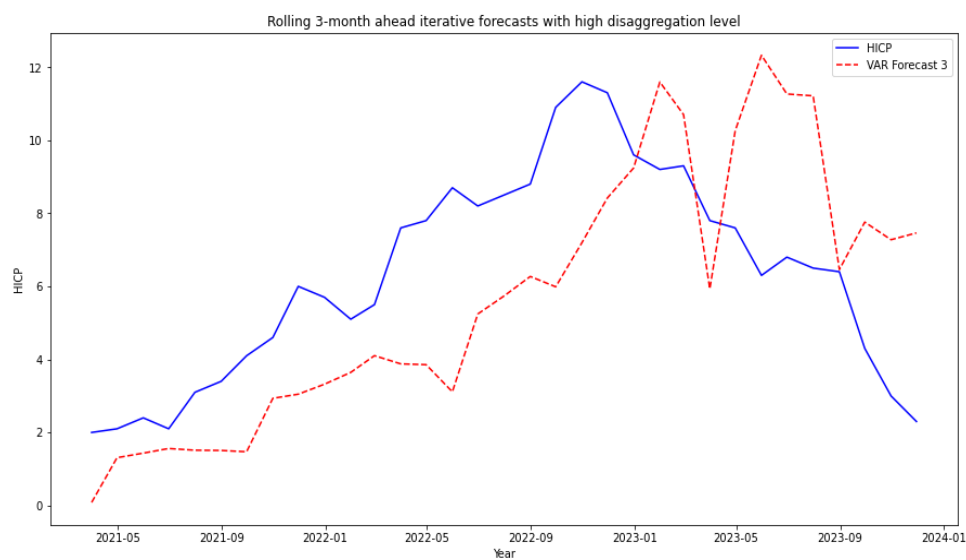


Figure 38: 6-month ahead iterative forecasts using VAR models with high disaggregation level

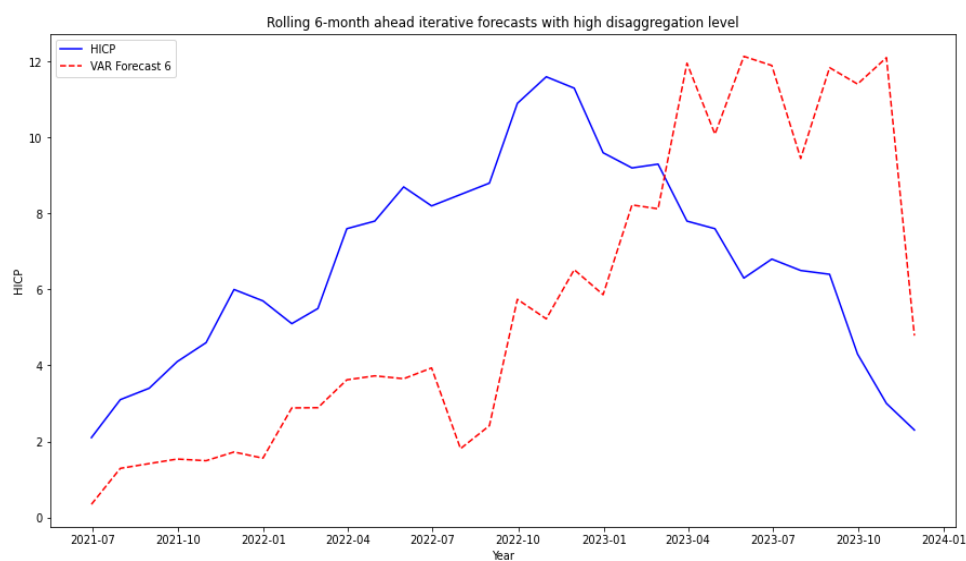


Figure 39: 9-month ahead iterative forecasts using VAR models with high disaggregation level

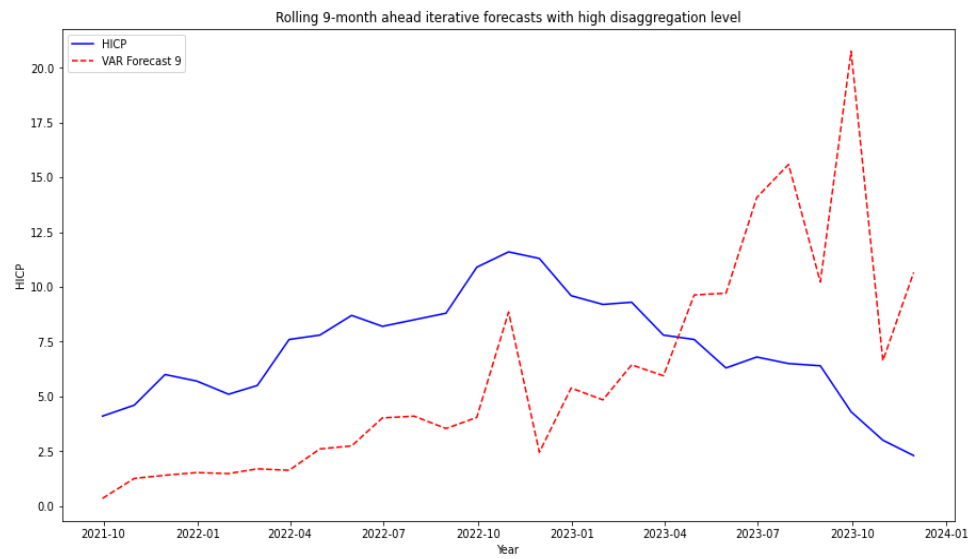


Figure 40: 12-month ahead iterative forecasts using VAR models with high disaggregation level

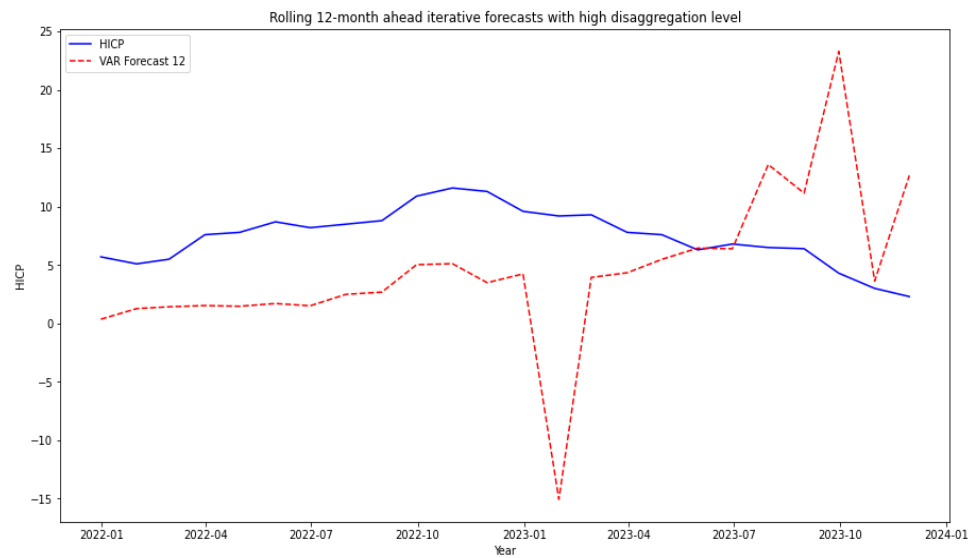


Figure 41: 1-month ahead direct forecasts using univariate models for forecasting HICP directly

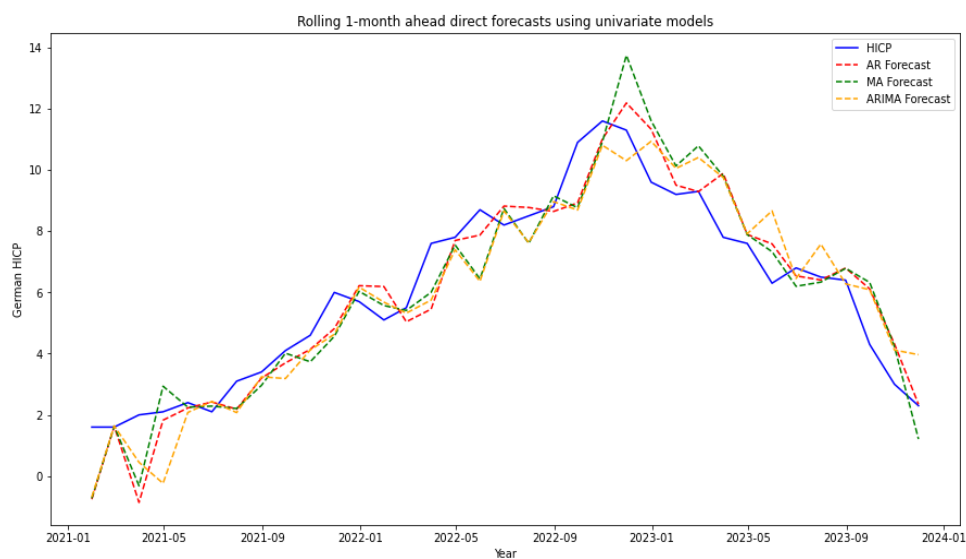


Figure 42: 3-month ahead direct forecasts using univariate models for forecasting HICP directly

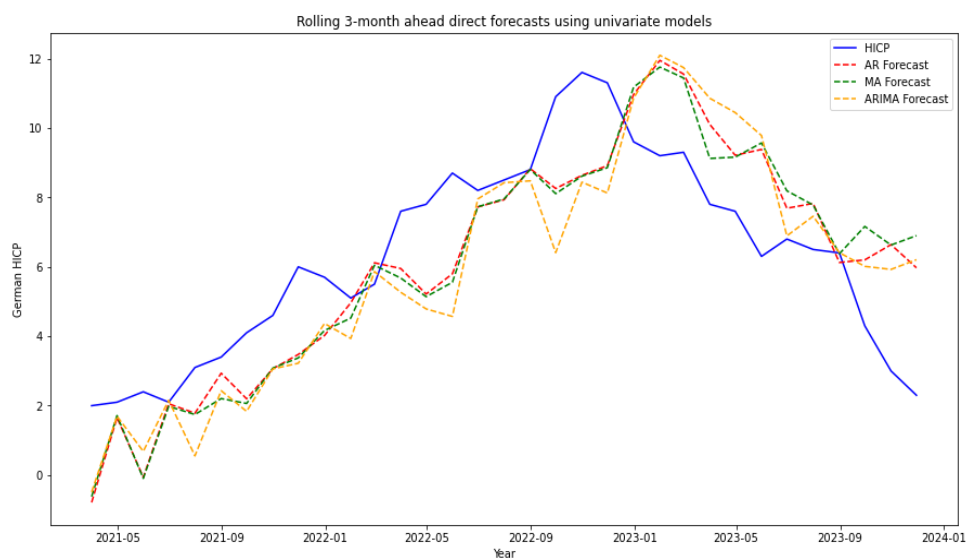


Figure 43: 6-month ahead direct forecasts using univariate models for forecasting HICP directly

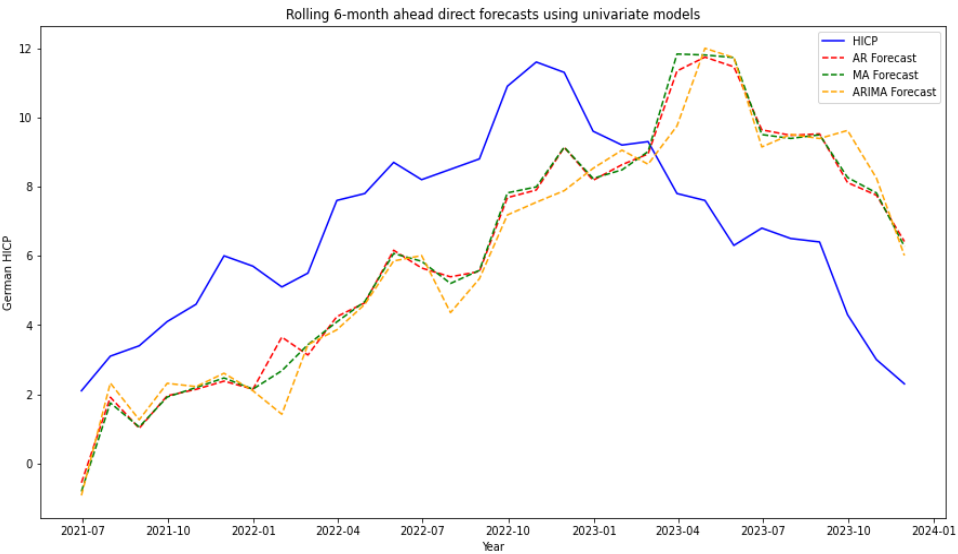


Figure 44: 9-month ahead direct forecasts using univariate models for forecasting HICP directly

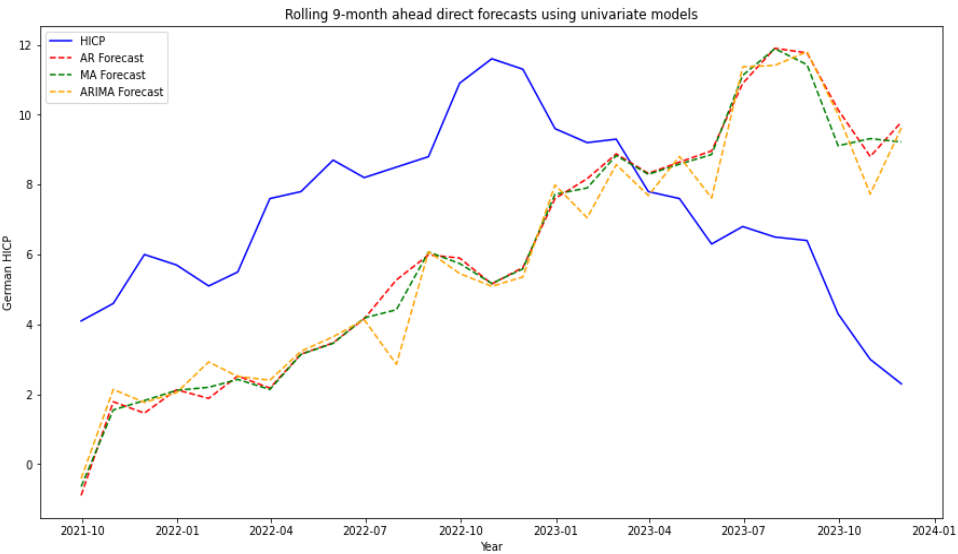


Figure 45: 12-month ahead direct forecasts using univariate models for forecasting HICP directly

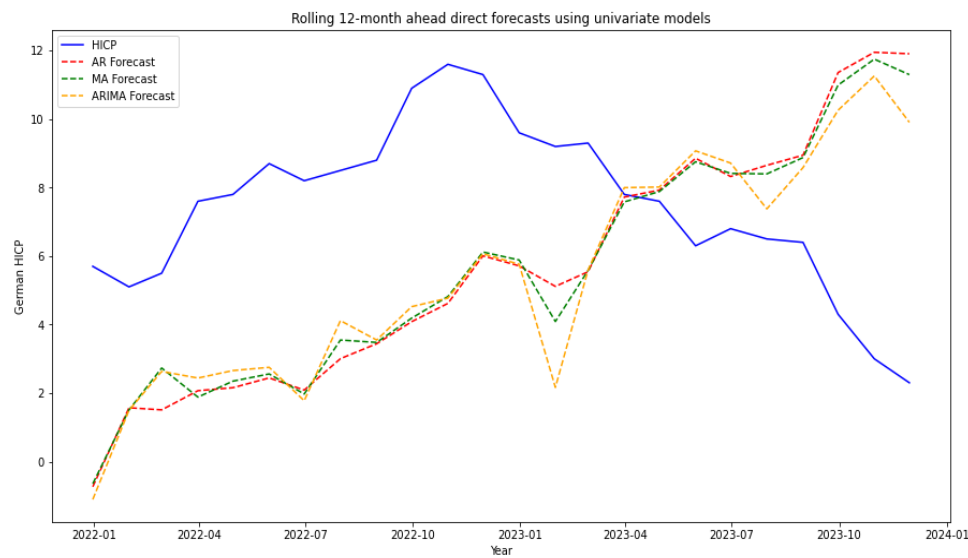


Figure 46: 1-month ahead direct forecasts using VAR models for forecasting HICP directly

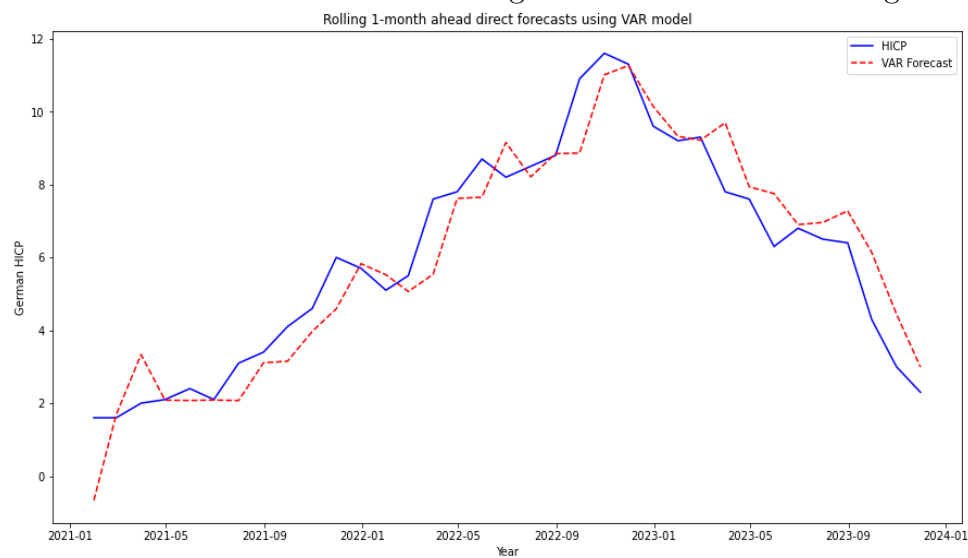


Figure 47: 3-month ahead direct forecasts using VAR models for forecasting HICP directly



Figure 48: 6-month ahead direct forecasts using VAR models for forecasting HICP directly

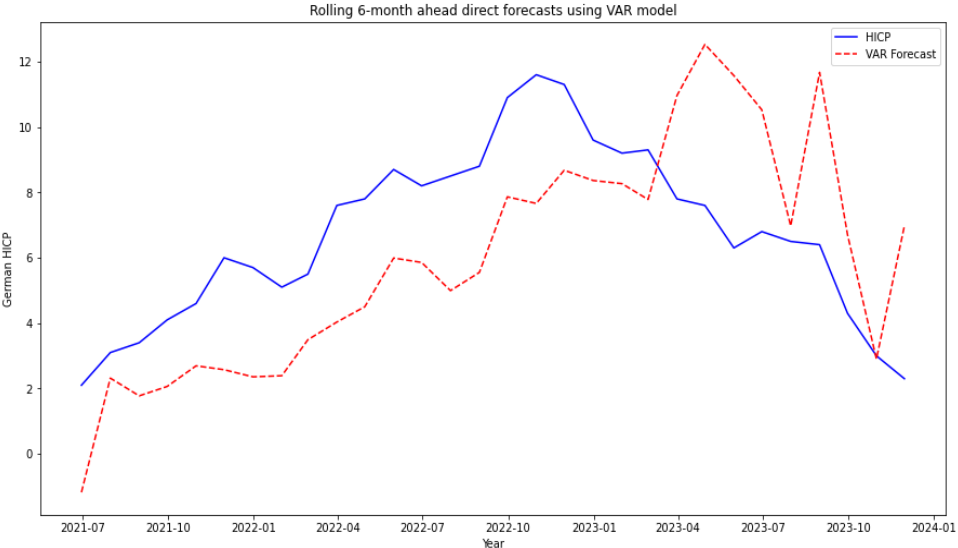


Figure 49: 9-month ahead direct forecasts using VAR models for forecasting HICP directly

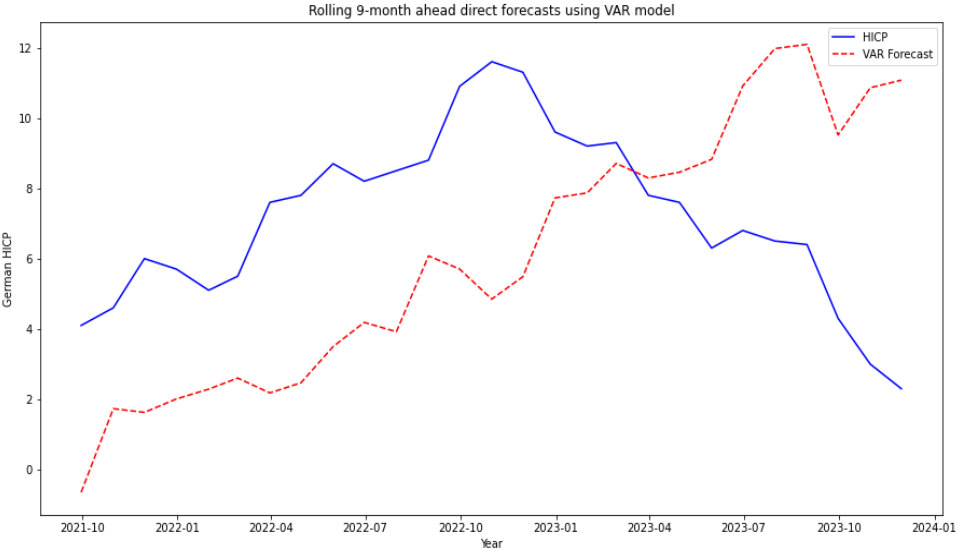


Figure 50: 12-month ahead direct forecasts using VAR models for forecasting HICP directly

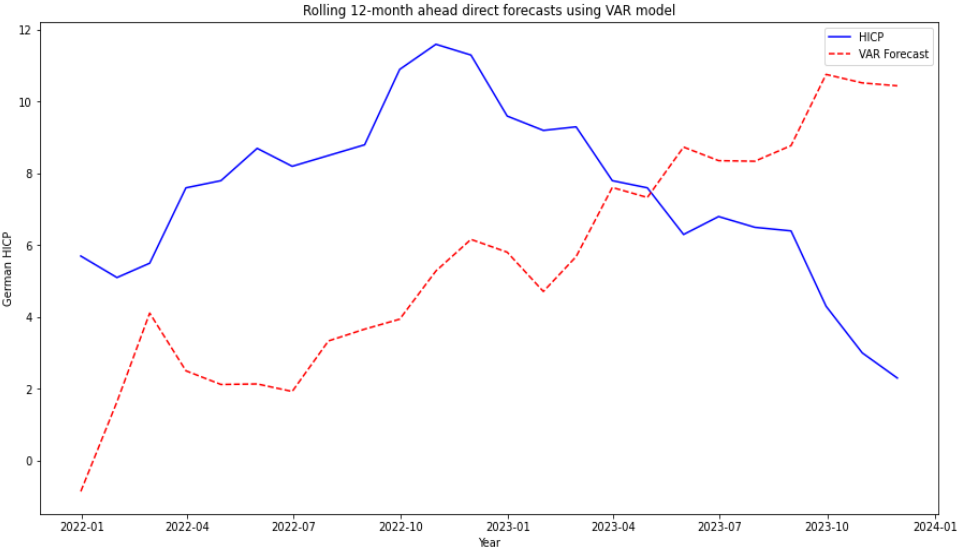


Figure 51: 1-month ahead direct forecasts using a combination of univariate and VAR models for forecasting HICP directly

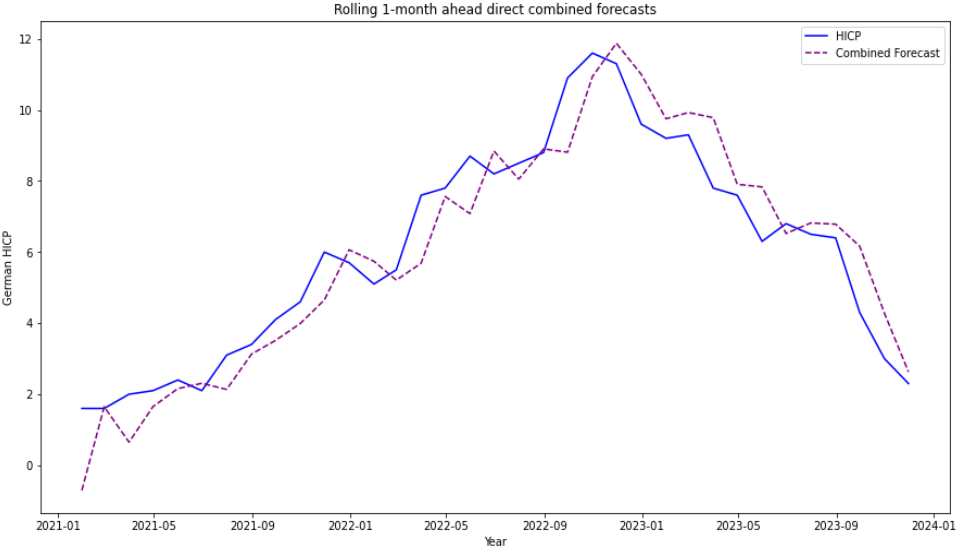


Figure 52: 3-month ahead direct forecasts using a combination of univariate and VAR models for forecasting HICP directly

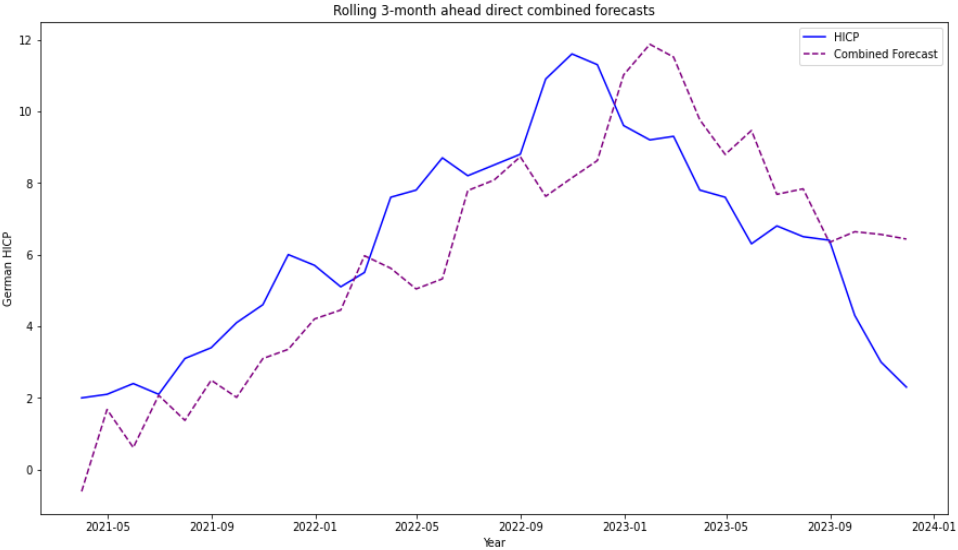


Figure 53: 6-month ahead direct forecasts using a combination of univariate and VAR models forecasting HICP directly

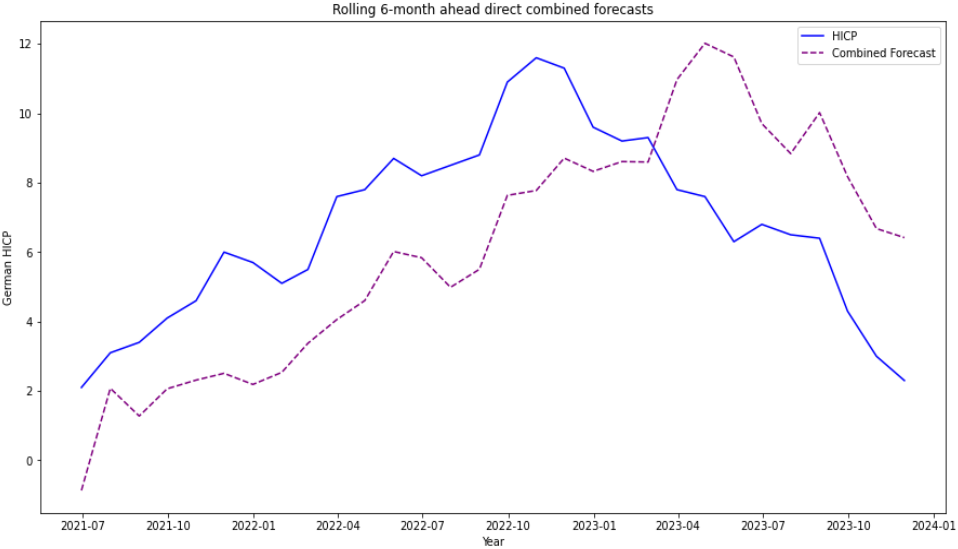


Figure 54: 9-month ahead direct forecasts using a combination of univariate and VAR models for forecasting HICP directly

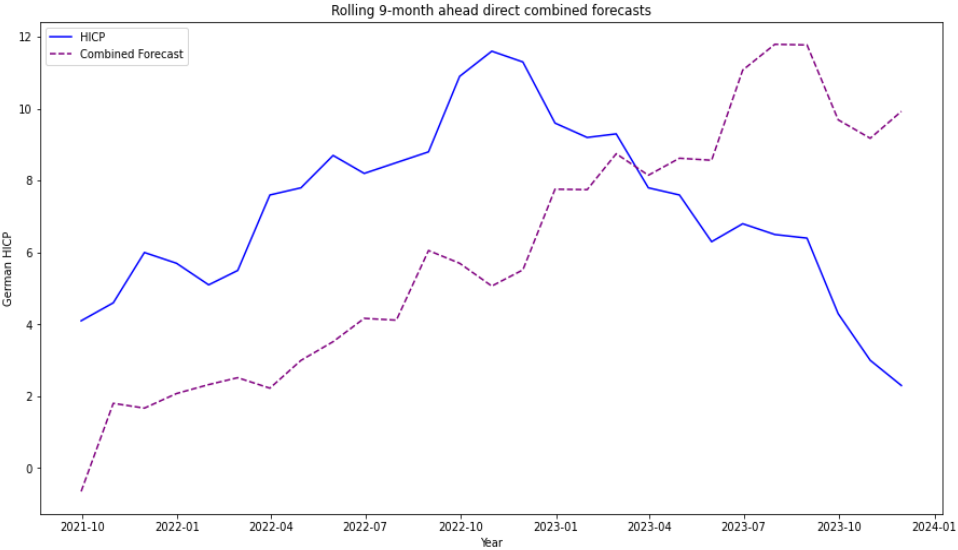


Figure 55: 12-month ahead direct forecasts using a combination of univariate and VAR models for forecasting HICP directly

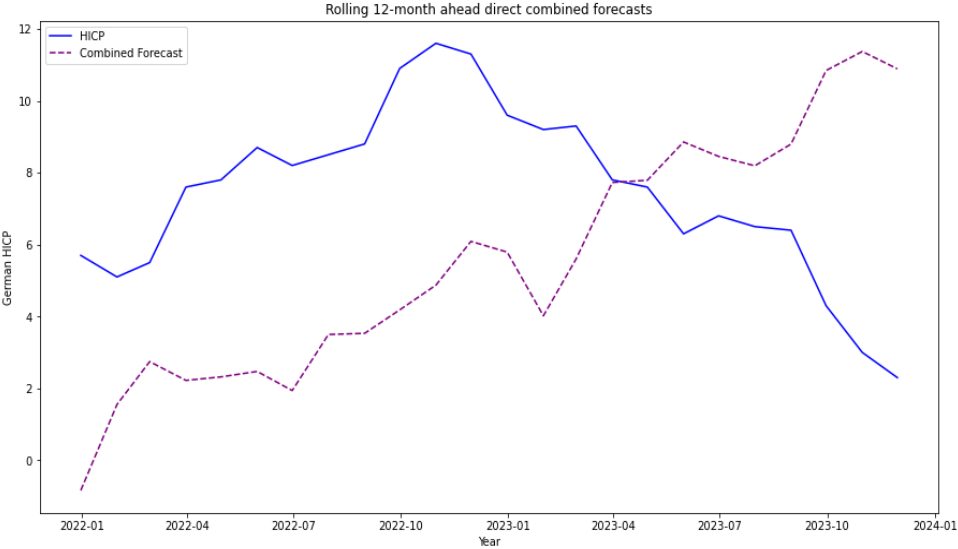


Figure 56: 1-month ahead direct forecasts using univariate models with low disaggregation level

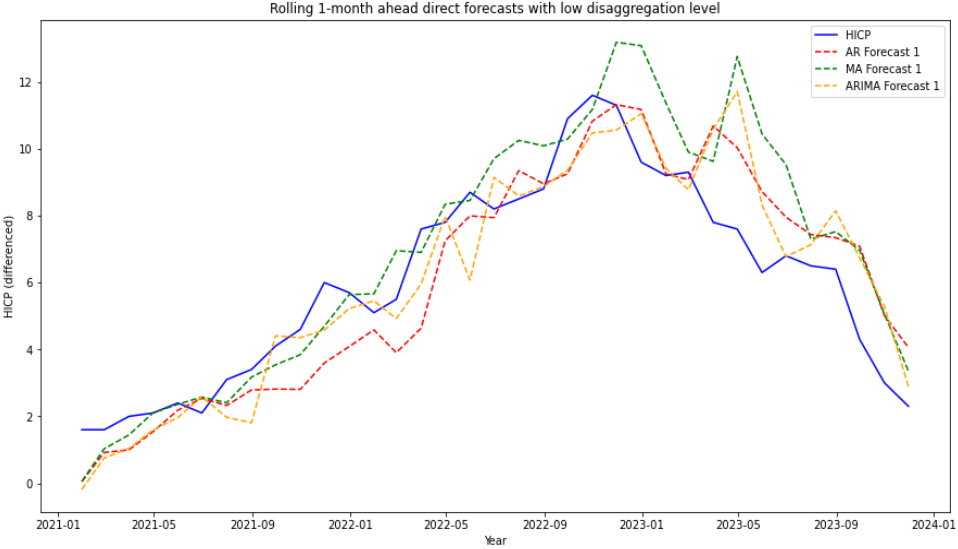


Figure 57: 3-month ahead direct forecasts using univariate models with low disaggregation level

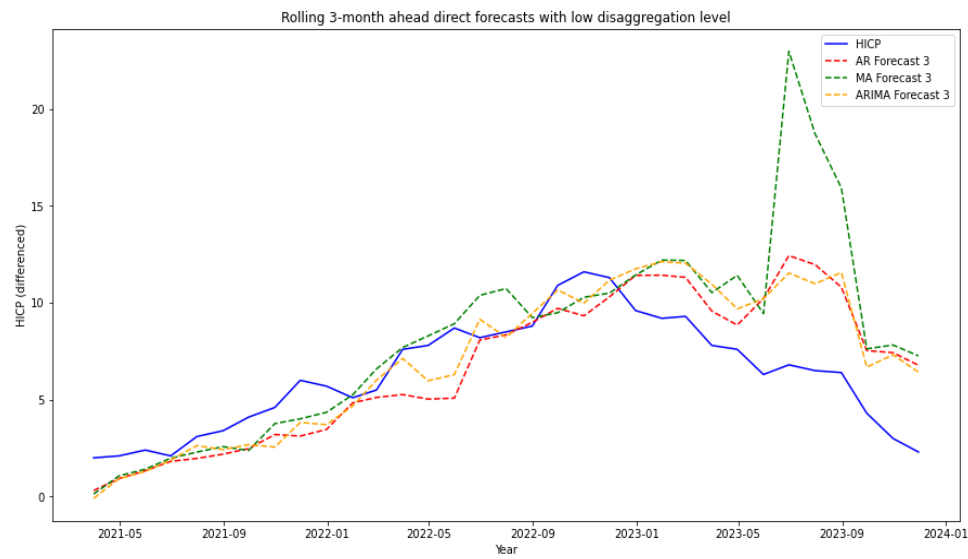


Figure 58: 6-month ahead direct forecasts using univariate models with low disaggregation level

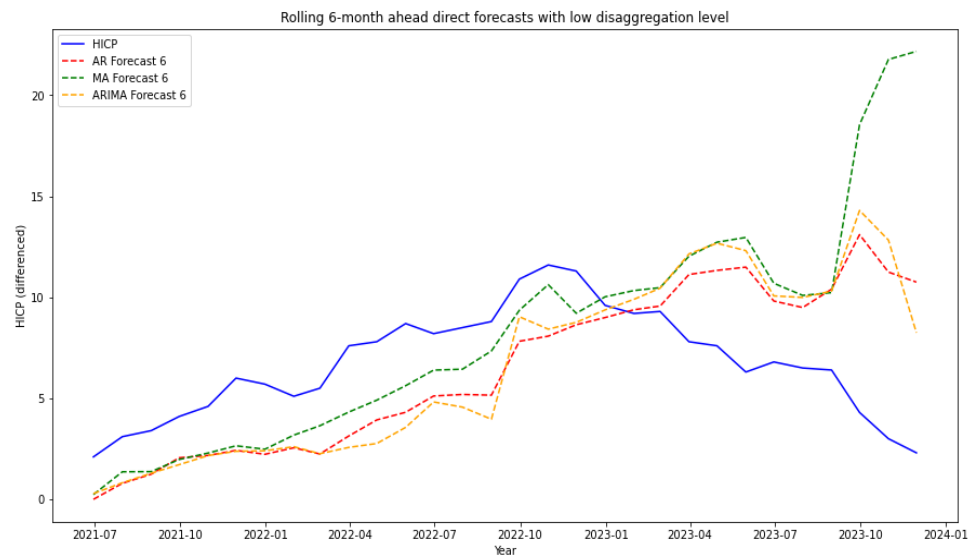


Figure 59: 9-month ahead direct forecasts using univariate models with low disaggregation level

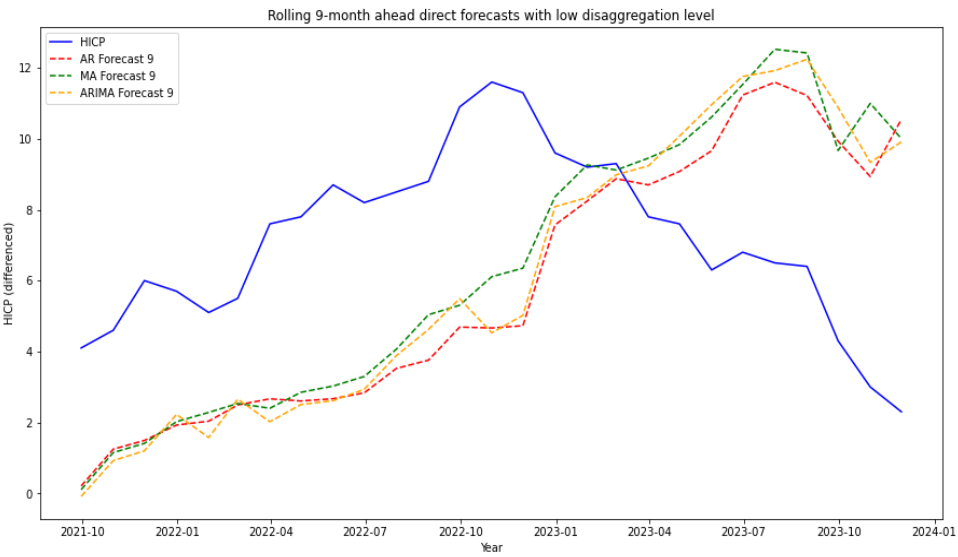


Figure 60: 12-month ahead direct forecasts using univariate models with low disaggregation level

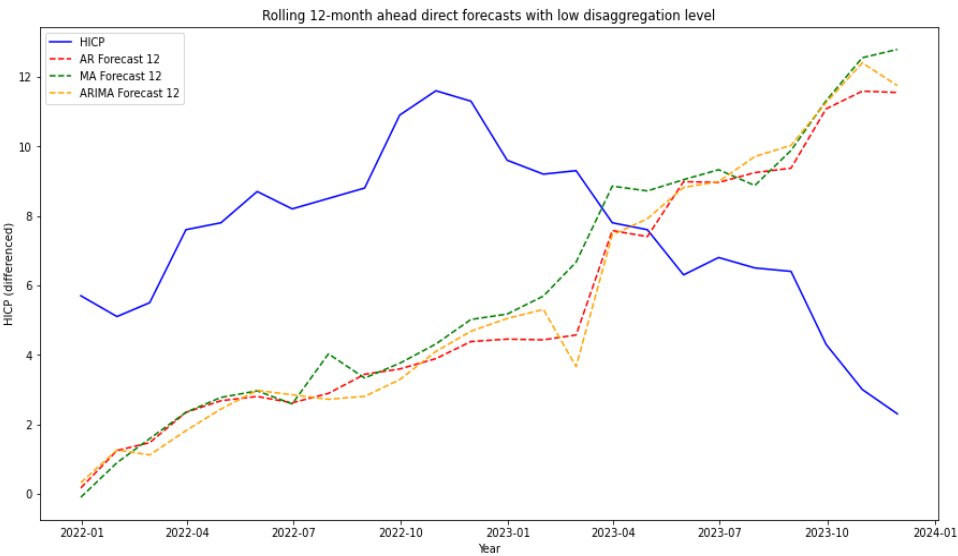


Figure 61: 1-month ahead direct forecasts using VAR models with low disaggregation level

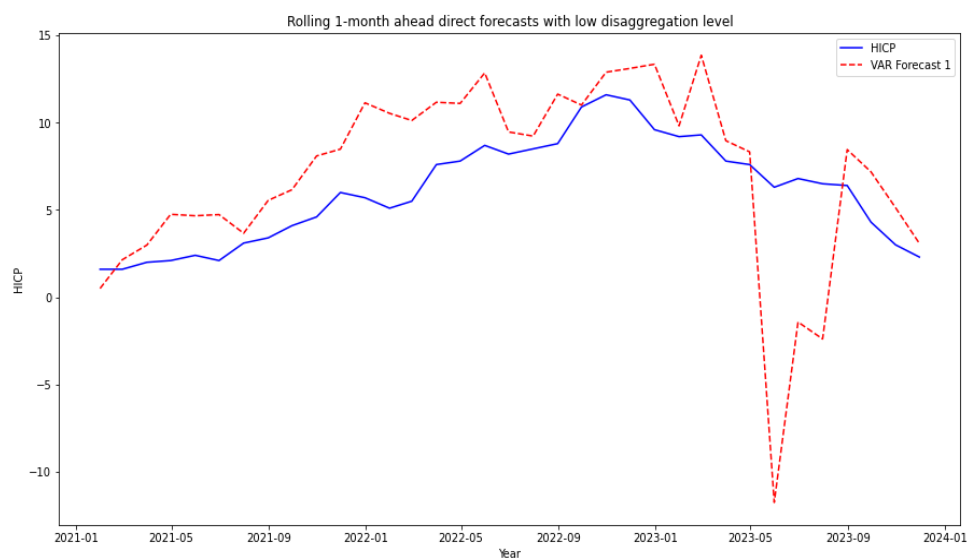


Figure 62: 3-month ahead direct forecasts using VAR models with low disaggregation level

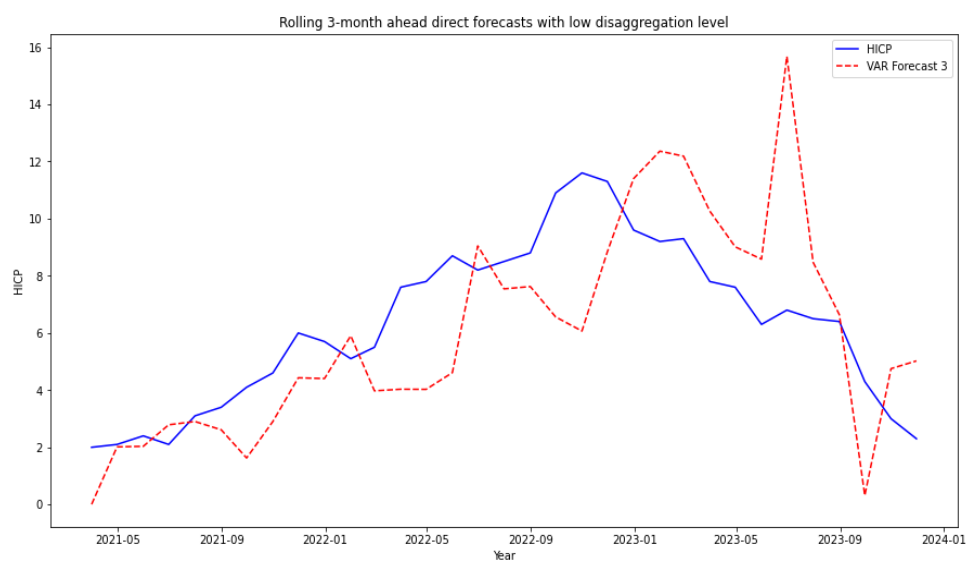


Figure 63: 6-month ahead direct forecasts using VAR models with low disaggregation level

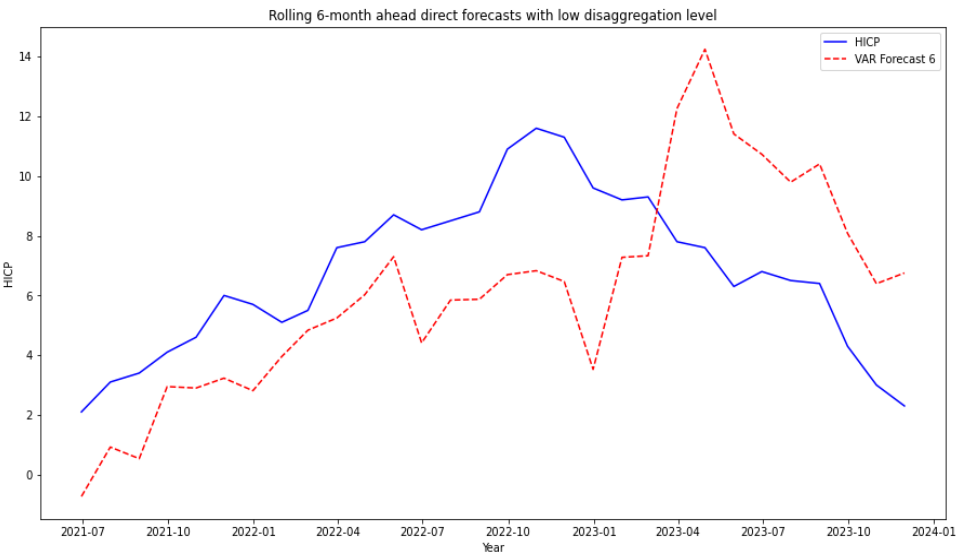


Figure 64: 9-month ahead direct forecasts using VAR models with low disaggregation level

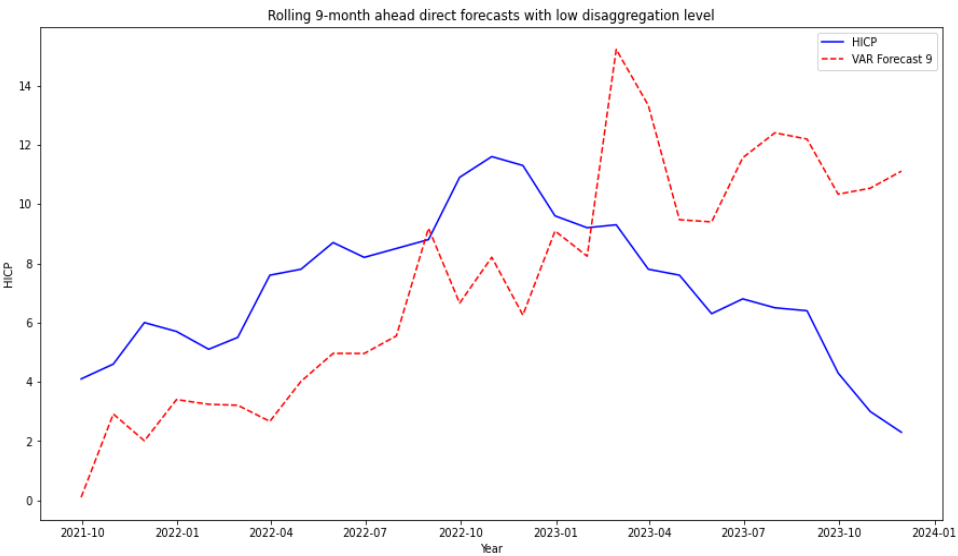


Figure 65: 12-month ahead direct forecasts using VAR models with low disaggregation level

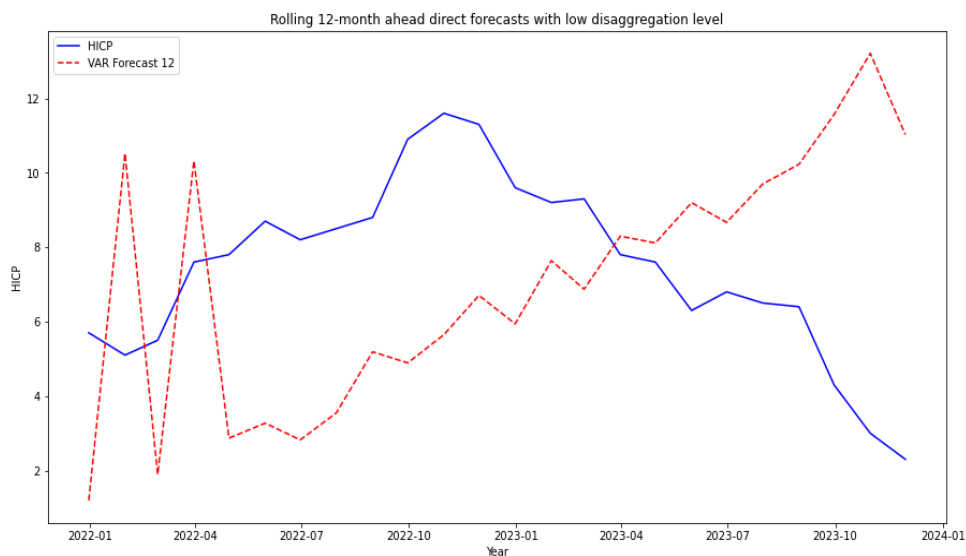


Figure 66: 1-month ahead direct forecasts using univariate models with high disaggregation level

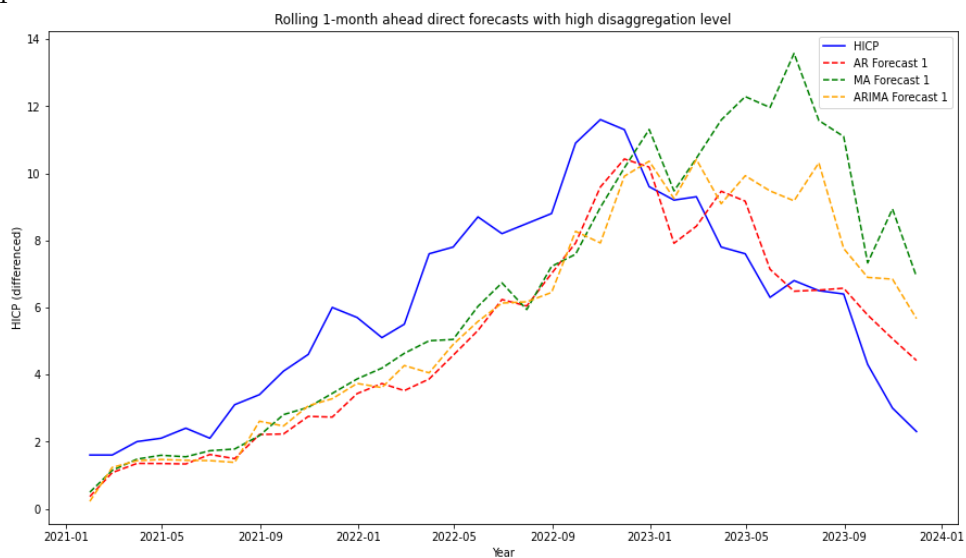


Figure 67: 3-month ahead direct forecasts using univariate models with high disaggregation level

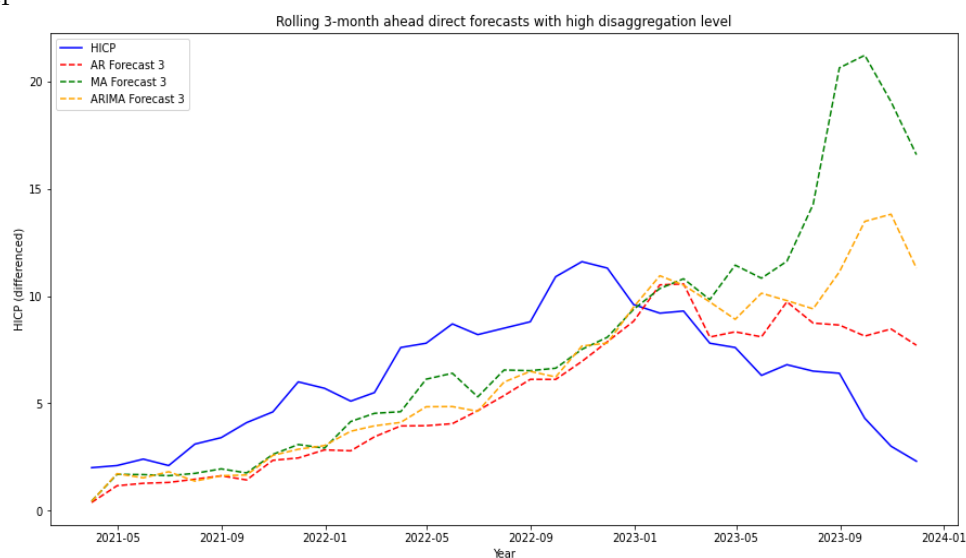


Figure 68: 6-month ahead direct forecasts using univariate models with high disaggregation level

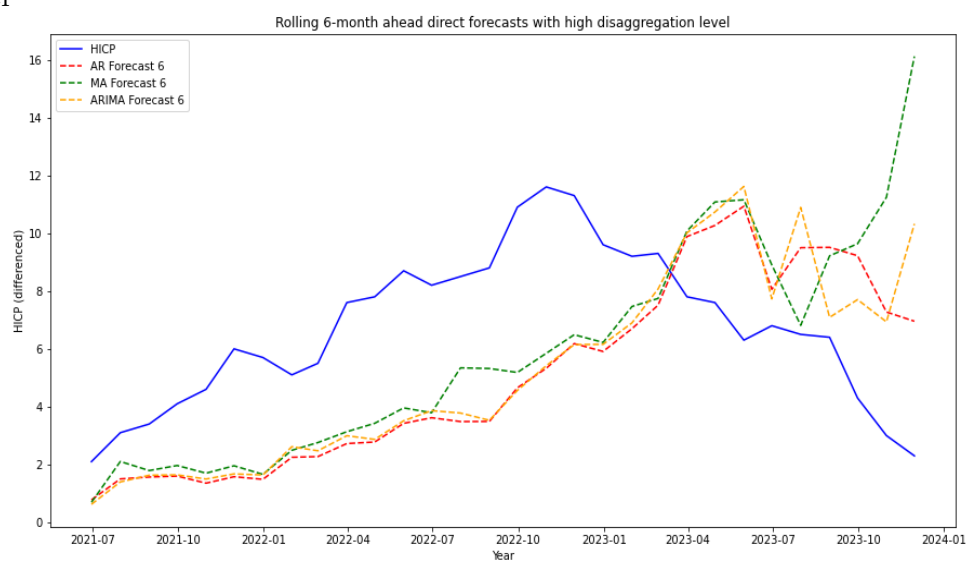


Figure 69: 9-month ahead direct forecasts using univariate models with high disaggregation level

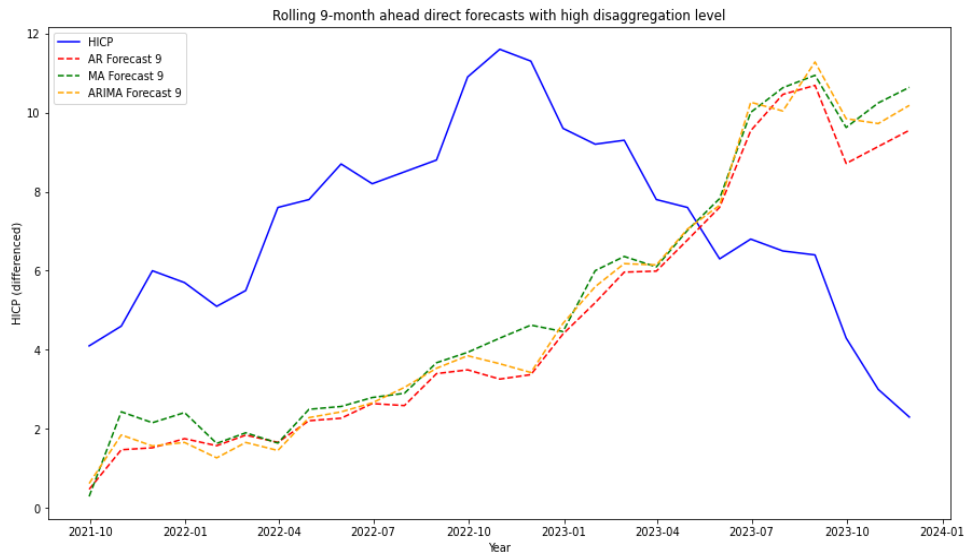


Figure 70: 12-month ahead direct forecasts using univariate models with high disaggregation level

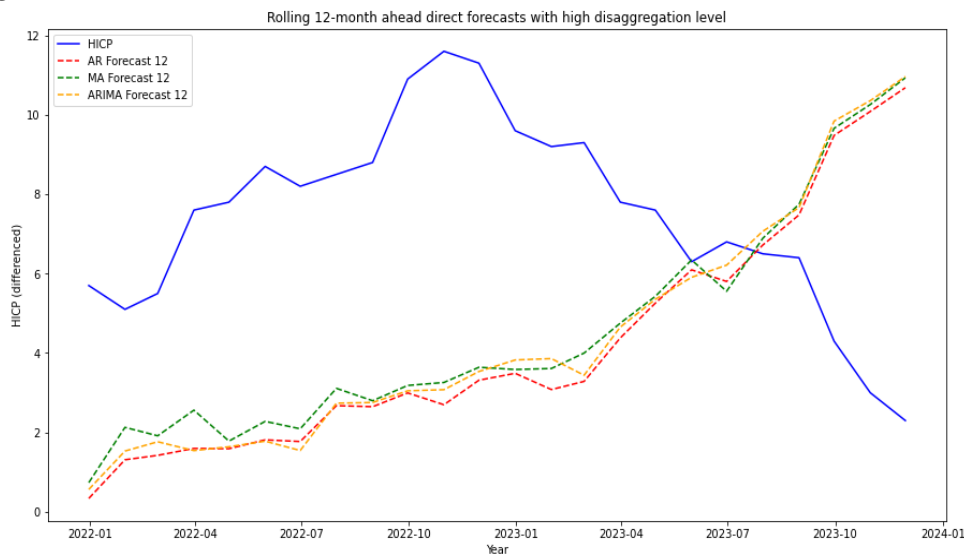


Figure 71: 1-month ahead direct forecasts using VAR models with high disaggregation level

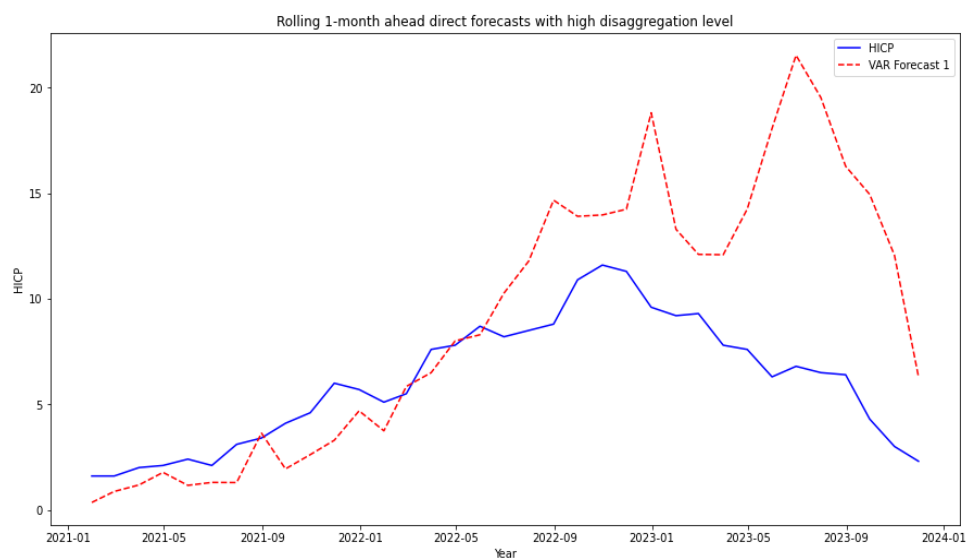


Figure 72: 3-month ahead direct forecasts using VAR models with high disaggregation level

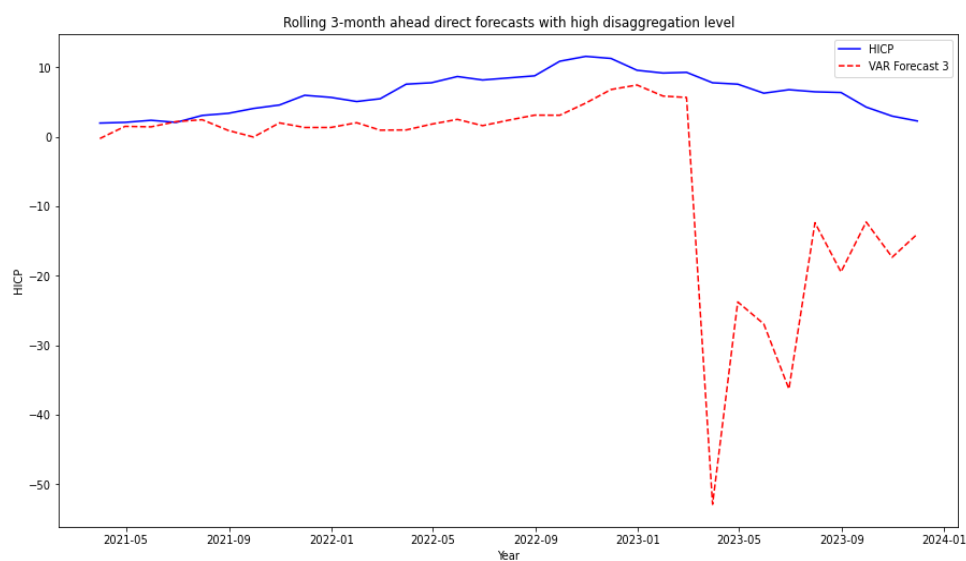


Figure 73: 6-month ahead direct forecasts using VAR models with high disaggregation level

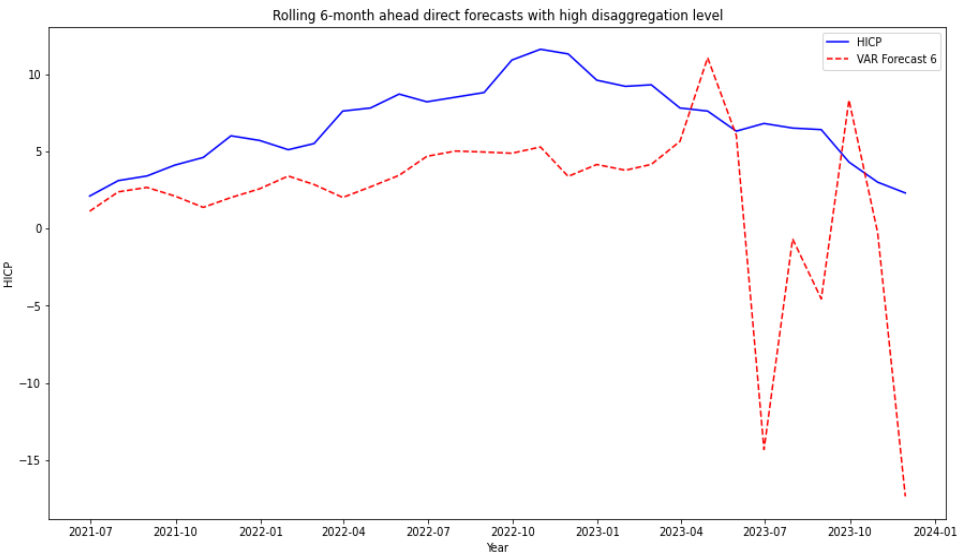


Figure 74: 9-month ahead direct forecasts using VAR models with high disaggregation level

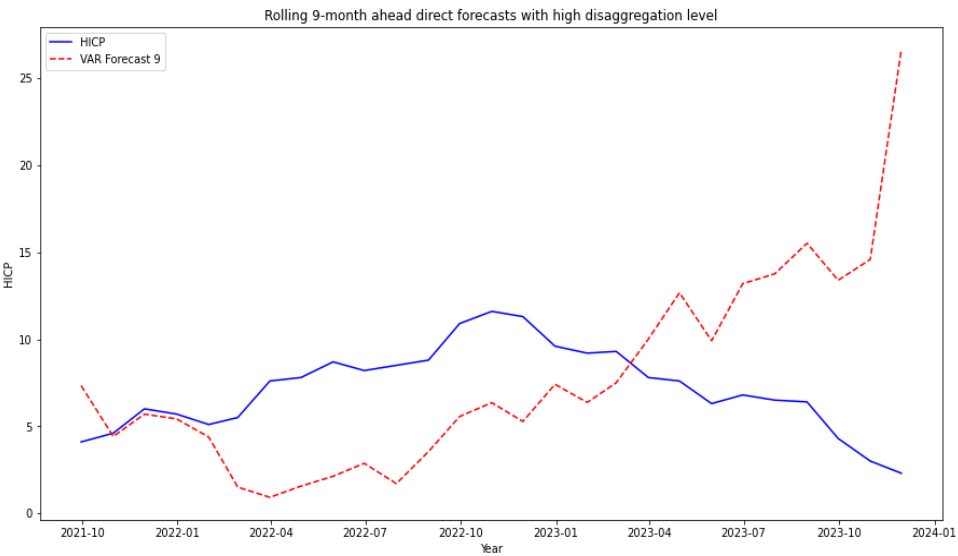


Figure 75: 12-month ahead direct forecasts using VAR models with high disaggregation level

