Journal of Quantitative Economics (2021) 19 (Suppl 1):S401–S417
https://doi.org/10.1007/s40953-021-00277-5

ORIGINAL ARTICLE

Modeling Judgment in Macroeconomic Forecasts

Philip Hans Franses[1]

Accepted: 27 October 2021 / Published online: 9 November 2021

Abstract

Many macroeconomic forecasts are the outcome of a judgmental adjustment to a forecast from an econometric model. The size, direction, and motivation of the adjustment are often unknown as usually only the final forecast is available. This is problematic in case an analyst wishes to learn from forecast errors, which could lead to improving the model, the judgment or both. This paper therefore proposes a formal method to include judgment, which makes the combined forecast reproducible. As an illustration, a forecast from a benchmark simple time series model is only modified when the value of a factor, estimated from a multitude of variables, exceeds a user-specified threshold. Simulations and empirical results for forecasting annual real GDP growth in 52 African countries provide an illustration.

Keywords Macroeconomic forecasting · Judgment · Dynamic factors · GDP growth in Africa

JEL Classification C18 · C32 · E37

This paper has been prepared for the special issue of the Journal of Quantitative Economics in honour of the late Professor A.L. Nagar.

Professor Nagar was affiliated with our Econometric Institute, where he worked on this doctorate thesis under the supervision of Henri Theil and graduated in 1959. It must have been a great time back then, when other visitors to our Institute were leading econometricians like Arnold Zellner, Arthur Goldberger, Thomas Rothenberg and Marc Nerlove, and of course when Jan Tinbergen worked there too. Recently I spoke with Teun Kloek (1934), who was my PhD supervisor in the late eighties, and he remembered Nagar as a very industrious person, and a fine colleague with exceptional technical skills. Nagar and his wife lived not far from Rotterdam in Zevenhuizen, and Teun Kloek still remembers that even an after-dinner dessert contained red peppers. In a sense, Nagar was way ahead of his time, as a young student coming to the Netherlands to work on a PhD thesis, something that is now much more common these days, but back then it was very special.

✉ Philip Hans Franses
franses@ese.eur.nl

[1] Econometric Institute, Erasmus School of Economics, Rotterdam, The Netherlands

Introduction

Macroeconomic forecasts are a key input to macroeconomic policies issued by governments and central banks. These forecasts typically concern important variables like growth in real gross domestic product (GDP), unemployment and inflation. The forecasts usually provide an outlook on the short-run (current year and next year) to medium-term (5 years) developments.

Econometric models can provide the basis for macroeconomic forecasts. In reasonably prosperous times these models tend to do well in terms of forecast accuracy. Unfortunately, when times radically change, most econometric models by their very nature are not qualified to predict for example turning points, and perhaps human intervention may direct the forecasts in the proper direction. It is thus common practice to base macroeconomic forecasts on the outcome of an econometric model combined with expert judgment,[1] or sometimes even to use no econometric model at all. As an example, Franses et al. (2011) show that literally all forecasts created by the Netherlands Bureau of Economic Policy Analysis (CPB) are judgmentally adjusted model-based forecasts. The authors also show that the final forecasts are in general more accurate than the model-based forecasts. So, judgmental adjustment can lead to more accuracy. In that study, and in many studies on evaluating judgmentally adjusted forecasts,[2] the researchers have access to model forecasts and the judgmentally adjusted forecasts, but not to the precise motivation that caused judgment.

In many other cases in practice, only the final forecast is available, and it is unknown how the underlying model looked like, how adjustment took place, and what was the precise motivation for such adjustment. This can disadvantageous if one wants to learn from forecast errors. Consider for example the 2008/2009 recession. In those years for the USA, real GDP growth was − 0.3 and − 2.8, respectively. In the June 2008 survey of the Consensus Forecasters[3] the average quote for 2008 was 1.5 (ranging from 0.8 to 1.9), while the average quote for 2009 was 1.7 (with a highest and lowest score of 3.1 and 0.6, respectively). Even in the November 2008 survey, the average quote for the very same year, 2018, was 1.4 (with high and low 1.5 and 1.3, respectively), whereas the average quote for 2009 now was − 0.6 (with individual quotes ranging from 1.2 to − 2.1). Apparently, either econometric model forecasts were off track or expert adjustment or both. Unfortunately, this is unknown.

In this paper I therefore propose a formal way of documenting the creation of a judgmentally adjusted forecast. This methodology is illustrated for a simple time series model (as "the model") and the outcome of a factor analysis of a range of

[1] Recent interesting literature on combining model forecasts with experts' judgemental forecasts make use of entropy-based methods, see for example Robertson et al. (2005), Tallman and Zaman (2020), and Altavilla et al. (2017), among others.

[2] Interesting studies on the evaluation of judgmentally adjusted forecasts are Wallis (1989), McNees (1990), Turner (1990), Stekler (2007), Fildes and Stekler (2002), Bunn and Salo (1996), Clements (1995), Lawrence et al. (2006) and Davydenko and Fildes (2013), amongst many more.

[3] Consensus Forecasters concern the forecasts created by over 20 banks and institutes, including Merrill Lynch, Swiss RE, Morgan Stanley, Goldman Sachs, Bank America Corp, and JP Morgan, amongst various others.

potential predictor variables, which will be added to the model forecast (as "judgmental adjustment") only when it exceeds thresholds. A multitude of other choices could have been made, but here it serves as illustration, not to show that judgment is always better, or that the model must be improved,[4] but just to illustrate the methodology.

In this paper the focus is on a setting where there are many predictors. More precise, the aim is to forecast annual real GDP growth in each of 52 African countries, where there are data since 1960. The predictors are the growth rates of the other countries. The variables are summarized in dynamic factors, to be estimated from the data, and they only enter the forecast equation, which is based on a simple autoregressive model, when their values pass some pre-set threshold, based on the standard deviation of the estimated factors. This way, judgmental adjustment is formalized and as such one can learn from past forecast errors.

The outline of this paper is as follows. "Judgmental Adjustment of Model-Based Forecasts" summarizes the current knowledge base on judgmental adjustment of econometric model forecasts. "Modelling Judgement" formalizes the methodology. "Simulation Results" presents some simulation results. "Forecasting Real GDP Growth in Africa" implements the methodology to real GDP forecasting in Africa. "Conclusion" concludes with limitations and looks forward.

Judgmental Adjustment of Model-Based Forecasts

The starting point in this paper is that we basically know little about what professional forecasters in macroeconomics do. We do have monthly quotes for example from the Survey of Professional Forecasters[5] or from Consensus Forecasters,[6] and we can analyze their accuracy and perhaps with some assumptions what they might do to arrive at their quotes, but this is all some kind of a reduced form analysis. What we do know is that if the forecasters use an econometric model, they almost always seem to adjust the forecasts from such a model.

The setting can consist of two actors. The first actor is the forecaster who is a staff member of IMF, OECD, Central Banks or of institutes who advice governments. He or she can also work for a company or is an individual who publishes forecasts, for example under the umbrella of the Survey of Professional Forecasters (SPF) or the Consensus Economics panel. Forecasts often include updates, which means that there are usually more forecasts for the same future event. For example, forecasts for

[4] In recent years, we have seen much research on improving econometric models, sometimes using higher frequency data, and in other cases including more variables or components. Also, modern machine learning tools have been developed, where novel search and variable selection algorithms are implemented see for example Kim and Swanson (2018). Big data in some dimensions (think of retrieving prices data from the internet with availability per minute) are exploited, and text mining using web crawlers also seem to be promising avenues.

[5] https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/.

[6] http://www.consensuseconomics.com/.

year T + 1, the forecast horizon, are created in various months of year T, the forecast origin, and even in the months of year T + 1.

The second actor is an analyst or a policy advisor who needs to evaluate the forecasts for macroeconomic variables and to make policy recommendations. The analyst or advisor may be the same individual as the forecaster but does not have to be. The analyst may have to report to policy makers, management, or to investors on the usefulness, relevance and reliability of the forecasts.

Generally, the situation is that the second actor does not know how the first actor creates the forecasts, as usually no detailed documentation is provided. There is no documentation on the specific econometric model used, nor on the size, direction or motivation for eventual adjustment. Forecasts from the IMF, OECD and many others usually do not come with any information on an econometric model that could have been consulted. In addition, the Survey of Professional Forecasters and the often-considered Consensus Economics Forecasters panel does not include information on how the professional forecasters create their quotes. Therefore, we do not know to what extent we can learn from observed forecast errors to improve the model, the judgment or both. In the present paper, I propose a methodology that models judgment in such a way that learning from errors becomes feasible.[7]

Outside of macroeconomics, Mathews and Diamantopoulos (1986) were the very first to investigate how forecasters perform relative to statistical models in terms of out-of-sample forecast accuracy. Their data concern sales of automobiles and their main findings are that manually adjusted forecasts can be better in terms of out-of-sample root mean squared prediction error (RMSPE). In that same period, an important study is Blattberg and Hoch (1990), who compare adjusted forecasts and the original model forecasts and find that adjusted forecasts can be better than model forecasts, but only just a little bit better. In addition, and interestingly, they show that combining model forecasts and final forecasts, these combined forecasts are even better, which shows that judgment does matter.[8]

Recently, large databases with model forecasts, the final adjusted forecasts and the actual realizations have become available in sales forecasting and in macroeconomic forecasting. This has spurred a revived interest in analysing judgmentally adjusted forecasts.

Franses et al. (2011) document that experts with domain-specific knowledge manually adjust the model-based forecasts from the 1945-founded CPB, which originate from a 1000+ equations econometric model. As the CPB did not store past model-based forecasts, the authors had to re-run earlier econometric models. One outcome of that study is that literally all forecasts from the econometric model are manually adjusted. In sales forecasting, where nowadays managers need large numbers of forecasts at high

[7] There are various studies that analyse past forecast errors from adjusted forecasts, and these give insights in general statistical properties of across multitudes of forecasters, see for example Reifschneider and Tulip (2019), Jo and Sekkel (2019) and Lahiri et al. (2015). Some of these studies also provide insights about the construction prediction intervals around adjusted forecasts.

[8] In Franses and van Dijk (2019), it is shown that combined expert-adjusted model forecasts can improve on the combined model forecasts, even in the case when the individual expert-adjusted forecasts are not more accurate than their associated model-based forecasts.

frequency intervals like weeks or months, there is a longer tradition of an interaction between forecasting tools and individuals. Studies as Fildes et al. (2009), and Franses and Legerstee (2009, 2010) indicate that usually over 95% of all statistical model forecasts for sales are manually modified.

There are three main insights from these and other studies, which are reviewed in Franses (2014). The first insight is that adjustment of model forecasts is recommendable in various situations. Forecasters may have knowledge about future forecast errors due to known or foreseen structural changes, which cannot be included in the model. The forecasters thus may know that the observation to predict contains an outlier component. Other examples can be the introduction of an isolated event, like a new law that becomes effective at the time of the next observation, and which could not have been included in the forecasting model. Forecasters may also know more about a possible measurement error in one of the explanatory variables. Re-estimation of the model parameters usually does not lead to dramatic changes in the longer-term relation between variables, but it may well be that one of the explanatory variables will experience a large shift. Manual adjustment can thus be beneficial because it can reduce the forecast error, if the adjustment is independent from the model forecast. Even though this independence cannot be guaranteed, it is important, as a forecaster should not replicate what is already in the model. This is taken aboard below in my novel methodology.

The second insight is that the actual behaviour of forecasters often seems to be far from an ideal situation. Such an ideal situation would be that what an individual adds to the model forecast is unpredictable, as otherwise it could have been included in the model. However, in practice, the differences between adjusted forecasts and model-based forecasts are often found to be predictable. Additionally, for example when it comes to sales and GDP growth, forecasters seem to adjust more often upwards than downwards. Upon actually asking individual forecasters, what it is that they do, it can be learned that forecasters themselves say that they quite often eventually ignore the model forecasts and create their own model. In addition, more experienced forecasters show signs of overconfidence, see Lamont (2002). This notion that the difference between adjusted forecasts and model-based forecasts should not be predictable is also considered in the new methodology.

The third and final insight is that there is substantial room for improvement of adjusted forecasts, to alleviate biases and inappropriate heuristics. There is evidence that infrequent adjustments are more beneficial. In addition, combinations of adjusted forecasts and model-based forecasts can improve on each of the components. This third insight that only infrequent adjustments can be useful is also incorporated in the methodology below.

To introduce some notation, consider the following. Suppose the interest is in the one-step-ahead prediction from origin T of a variable Y for forecast horizon $T + 1$, to be denoted as $Y_{T+1|T}$. Suppose further that there is an econometric model (M) that gives a prediction for $Y_{T+1|T}$ and label this as $M_{T+1|T}$. When a forecaster fully relies on an econometric model, we have

$$Y_{T+1|T} = M_{T+1|T}.$$

We know however that forecasters may not fully on an econometric model. In fact, what they often seem to do is to rely on

$$Y_{T+1|T} = M_{T+1|T} + E_{T+1|T},$$

where $E_{T+1|T}$ is what might be called Expertise. Some would call it intuition or judgmental adjustment. When there is no econometric model at all, or when the forecaster fully ignores an available econometric model, then the final forecast simply is

$$Y_{T+1|T} = E_{T+1|T},$$

meaning that the forecast is fully based on judgement.[9]

In practice we often do not know the values of $E_{T+1|T}$ or $M_{T+1|T}$ for each of the forecasters as we only observe $Y_{T+1|T}$, nor do we know the balance between the two[10]. Basically, this means that many if not almost all available forecasts from SPF, Consensus, IMF, OECD and the like are irreproducible forecasts. And, we cannot learn much from the forecast errors, as these cannot tell us in what way we can modify $E_{T+1|T}$ or $M_{T+1|T}$ or both for future use.

The best one can do with irreproducible forecasts is indeed to combine them and look at their accuracy. Moreover, one can study properties of the empirical distributions of the forecast quotes.[11]

Modelling Judgement

To learn from forecast errors, it seems helpful to formalize how forecasters create their forecasts and how they incorporate any adjustment. This should lead to a trail of decisions made in the past, so that we can learn from past errors. It would also be useful to see to what extent a professional forecaster improves on a model.

A building block of a methodology that can explicitly incorporate judgment is the well-known dynamic factor model, see Stock and Watson (2002), Bai and Ng (2002), and Kim and Swanson (2018), and many others. Other models can also be useful, but the dynamic factor model is well suited to incorporate many predictor variables, as we also have in the illustration section below.

[9] Kahneman (2012) and Tetlock and Gardner (2015) argue that this strategy is not recommend for various reasons.

[10] Franses and Legerstee (2009) provide one of the exceptions. They analyze a unique database with one-step-ahead model-based forecasts adjusted by many forecasters, located in 37 countries, and who make forecasts for pharmaceutical products within seven distinct categories. These authors find that forecasters make frequent adjustments and that these tend to be upward. They also document that judgmental adjustment itself is largely predictable, and that judgmental adjustment is not independent of the model-based forecasts.

[11] To evaluate the quality of the forecasts from professional forecasters, one often takes the average quote (the consensus) or the median quote, and sometimes one also uses measures of dispersion like the standard deviation or the variance. The latter measures give an indication to what extent the forecasters disagree. Recent relevant studies are Capistran and Timmermann (2009), Dovern et al. (2012), Lahiri and Sheng (2010), Laster et al. (1999), and Legerstee and Franses (2015).

The standard model is

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{W}_{t-1}\boldsymbol{\beta}_W + \boldsymbol{F}_{t-1}\boldsymbol{\beta}_F + \boldsymbol{\varepsilon}_t, \tag{1}$$

$t = 1,2,\ldots,T$. Usually, in the relevant literature one takes for $\boldsymbol{W}_{t-1}$ the first p lags of $\boldsymbol{Y}_t$, which makes the model to look like

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{\phi}_1\boldsymbol{Y}_{t-1} + \ldots + \boldsymbol{\phi}_p\boldsymbol{Y}_{t-p} + \boldsymbol{F}_{t-1}\boldsymbol{\beta}_F + \boldsymbol{\varepsilon}_t.$$

The factors F_t are unobservable factors, typically collecting a large number of variables. To create the factors, it is usually assumed that there are N predictors $X_{i,t}$ where $i = 1,2,..,N$, where N can be large. Next, the predictors are associated with the factors as follows:

$$\boldsymbol{X}_{i,t} = \boldsymbol{F}_t\lambda_i + \boldsymbol{\eta}_{i,t},$$

where F_t is an $T \times k$ matrix with k factors, and λ_i is a $k \times 1$ vector with factor loadings. These loadings and the factors are to be estimated from the empirical data and they are not fixed beforehand. There are various ways to estimate the factors, and a prominent method is factor analysis (FA). Methods to select factors are around, and there are also methods to preselect which $X_{i,t}$ variables will be included in FA.[12]

In practice, one usually compares this dynamic factor model with an autoregression, that is

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{\phi}_1\boldsymbol{Y}_{t-1} + \ldots + \boldsymbol{\phi}_p\boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t$$

The main idea of a new methodology, to be proposed in the present paper, is to make judgment explicit by taking this autoregression as "the model", and to create a forecasting scheme as follows:

$$Y_{T+1|T} = \mu + \phi_1 Y_T + \ldots + \phi_p Y_{T-p+1} + v_T + w_T$$

where v_T and w_T follow from

$$v_t = F_t \text{ if } v_t \geq \tau_U$$

$$v_t = 0 \text{ if } v_t < \tau_U$$

and

$$w_t = 0 \text{ if } w_t \geq \tau_L$$

$$w_t = F_t \text{ if } w_t < \tau_L.$$

[12] There are many variations possible on the proposed methodology. One of these could be to use alternative methods to extract factors. There is a huge literature on this issue, see for example Bai and Ng (2008), and Kim and Swanson (2018) but a detailed discussion or comparison is beyond the scope of the present paper.

Hence, as such v_t and w_t represent "judgment". The key features of this judgment are the upper and lower bounds τ_U and τ_L, respectively. These upper and lower bounds are to be set by the forecaster. Of course, these bounds need not be fixed over time, but in the present study they do. It could make sense to take τ_U as a number of times the standard deviation of F_t and τ_L as minus a number of times the standard deviation of F_t. If $\tau_U = \tau_L = 0$, the full model in (1) appears. In the simulation below I set $\tau_U = -\tau_L = 3$, and in the illustration to forecasting real GDP growth rates I opt for $\tau_U = -\tau_L = 1.5$. Of course, other choices can be made too, and there may also be no need to have symmetric thresholds. Also, the thresholds may vary over time if volatility changes over time. The main issue is that the forecaster reports the thresholds used for each forecast.

This methodology to add judgment to a forecast from a simple time series model obeys the commendable requirement that the model-based forecast and the added judgment are independent. Clearly, v_t and w_t are independent from the lags of Y_t. A second feature of sensible adjustment is that adjustment is not predictable. If it were, then one could just as well modify the model with that predictable term. Given that we only allow adjustment to enter the forecasting scheme once it exceeds some preset threshold value, it is difficult to predict the value of adjustment in advance, also because the threshold value is only known to the forecaster, at the time of the forecast origin. The third feature of our adjustment process is that adjustment occurs infrequently. Depending on the number of times the standard deviation is taken, the frequency of adjustment can be governed.

Simulation Results

The general idea of the approach to formalize the incorporation of judgment is to have an autoregression as the basic model, to which in some cases information is added, based on judgment. In this section I examine the methodology using some simulation experiments.

There are N potential predictors $X_{i,t}$ where $i = 1,2,..,N$. We will set $N = 5, 10, 20$ or 50. The predictors are associated with a single (in this experiment) factor F_t like the data generating process (DGP):

$$\text{DGP}: \boldsymbol{X_{i,t}} = \boldsymbol{F_t}\lambda_i + \boldsymbol{\eta_{i,t}}.$$

The setting in the experiment are

$$\boldsymbol{\eta_{i,t}} \sim N(0, \sigma_\eta^2)$$

$$\lambda_i = 1, \text{ for all } i$$

$$\boldsymbol{F_t} = \rho \boldsymbol{F_{t-1}} + \boldsymbol{\xi_t} \text{ with } \boldsymbol{\xi_t} \sim N\left(0, \sigma_\xi^2\right),$$

with $F_0 = 0, t = 1,2,\ldots,T$, where $T = 100, 500$ or 1000, and $\rho = 0.5, 0.8, 0.9$ or 0.95. Next, the data for Y_t are created as follows:

$$\text{DGP}: \boldsymbol{Y_t} = \boldsymbol{\phi}_1 \boldsymbol{Y_{t-1}} + \boldsymbol{v_{t-1}} + \boldsymbol{w_{t-1}} + \boldsymbol{\varepsilon_t} \, with \, \boldsymbol{\varepsilon_t} \sim N\left(0, \sigma_\varepsilon^2\right),$$

with $Y_0 = 0, t = 1,2,\ldots,T$, where $\phi_1 = 0.5, 0.8, 0.9$ or 0.95, and where v_t and w_t are defined by

$$v_{t-1} = F_{t-1} \text{ if } F_{t-1} > \tau_U$$

$$v_{t-1} = 0 \text{ if } F_{t-1} \leq \tau_U$$

and

$$w_{t-1} = 0 \text{ if } F_{t-1} \geq \tau_L$$

$$w_{t-1} = F_{t-1} \text{ if } F_{t-1} < \tau_L.$$

In our experiments we set the thresholds at 3 and -3 times the standard deviation of F_t, respectively.

The simulation proceeds as follows, which mimics what one can do in practice, that is

Step 0: Split the sample T into T_1 and T_2. In the simulations: $T_1 = T_2 = \frac{T}{2}$.

Step 1: Consider the sample T_1. Use factor analysis to estimate the factors and the loadings, and take only the first factor to create $\boldsymbol{F_t}$.

Step 2: Decide on the values of τ_u and τ_L. In our simulations we will set at 0 and at 3 and -3 (the latter matching the DGP).

Step 3: Estimate the parameter ϕ_1 in the first order autoregression

$$\boldsymbol{Y_t} = \boldsymbol{\phi}_1 \boldsymbol{Y_{t-1}} + \boldsymbol{\omega_t}$$

using ordinary least squares (OLS), and create the one-step-ahead forecast

$$\widehat{Y}_{T_1+1|T_1} = \widehat{\phi}_1 Y_{T_1}$$

Step 4: Move the sample to $T_1 + 1$, and repeat Step 3. This gives T_2 recursively created one-step-ahead forecasts. Compute the root mean squared prediction error $RMSPE_M$, where the M refers to "model".

Step 5: Take again the sample $t = 1,2,\ldots,T_1$ and create the one-step-ahead forecast

$$\widehat{Y}_{T_1+1|T_1} = \widehat{\phi}_1 Y_{T_1} + v_{T_1} + w_{T_1}$$

Step 6: Move the sample to $T_1 + 1$, and repeat Step 5. This gives a second set of T_2 recursively created one-step-ahead forecasts. Compute the root mean squared prediction error $RMSPE_J$, where J refers to judgmental adjustment.

Step 7: repeat Steps 1–6 K times and compare the two average root mean squared prediction errors.

Table 1 Simulation results

$\tau_U = 0, \tau_L = 0$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
$\phi = 0.5$	0	0.26	0.74	0.44
$\phi = 0.8$	0	0.26	0.85	0.91
$\phi = 0.9$	0	0.32	0.96	0.99
$\phi = 0.95$	0	0.58	0.96	1.00
$\tau_U = 3SD, \tau_L = -3SD$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
$\phi = 0.5$	0.99	0.99	0.98	0.94
$\phi = 0.8$	1.00	1.00	0.99	0.98
$\phi = 0.9$	0.96	1.00	1.00	1.00
$\phi = 0.95$	0.98	1.00	1.00	1.00

The cells are the fractions (out of 100 replications) that the RMSE of the "Model plus judgment" is smaller than the RMSE of the "Model" (the first order autoregression). *SD* is the estimated standard deviation of the estimated factor from FA, $T = 500$, $N = 5$

In the first simulation exercise the following configuration is employed. There are $K = 100$ replications. The DGP for $X_{i,t}$ involves $T = 500$, $N = 5$, $\eta_{i,t} \sim N(0, 0.5)$, $\lambda_i = 1$, $\xi_t \sim N(0, 1)$, $F_0 = 0$, and $\rho = 0{,}5, 0.8, 0.9$ or 0.95. Set the DGP for Y_t as

$$\boldsymbol{Y_t} = \boldsymbol{\phi Y}_{t-1} + \boldsymbol{v}_{t-1} + \boldsymbol{w}_{t-1} + \boldsymbol{\varepsilon_t} \, with \, \boldsymbol{\varepsilon_t} \sim N(0, 1)$$

with $Y_0 = 0$, with $\phi = 0.5, 0.8, 0.9,$ or 0.95 where v_t and w_t are

$$v_{t-1} = F_{t-1} \text{ if } F_{t-1} > 3 \, standard \, deviations$$

$$v_{t-1} = 0 \text{ if } F_{t-1} \leq 3 \, standard \, deviations$$

and

$$w_{t-1} = 0 \text{ if } F_{t-1} \geq -3 \, standard \, deviations$$

$$w_{t-1} = F_{t-1} \text{ if } F_{t-1} < -3 \, standard \, deviations$$

Table 1 reports on the fractions of times (out of $K = 100$) that $RMSPE_J < RMSPE_M$. From this table it can be learned that when the full model in (1) is used, that is, when the $\tau_U = 0, \tau_L = 0$, this model outperforms the simple time series model when both ρ and ϕ approach 1. The second panel shows that when the modelling and adjustment approach matches the DGP, that then the additional judgment makes the combined forecasting scheme almost always more accurate than just using the time series model only.

Table 2 reports on the same frequencies, but now where $\phi = 0.8$, and where the number of components in the factor ranges from 5, 10, 20 to 50. Comparing the rows in the two panels shows that the methodology works similarly across

Table 2 Simulation results

$\tau_U = 0, \tau_L = 0$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
$N = 5$	0	0.25	0.94	0.79
$N = 10$	0	0.30	0.86	0.88
$N = 20$	0	0.38	0.84	0.84
$N = 50$	0	0.26	0.88	0.80
$\tau_U = 3SD, \tau_L = -3SD$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
$N = 5$	0.98	0.99	0.97	1.00
$N = 10$	1.00	1.00	0.98	0.99
$N = 20$	0.98	1.00	0.99	1.00
$N = 50$	0.99	1.00	1.00	0.99

The cells are the fractions (out of 100 new replications) that the RMSE of the "Model plus judgment" is smaller than the RMSE of the "Model" (the first order autoregression). *SD* is the estimated standard deviation of the estimated factor from FA. $\phi = 0.8$, and $N = 5, 10, 20$ or 50, $T = 500$

Table 3 Simulation results

$\tau_U = 0, \tau_L = 0$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
$T = 100$	0.02	0.31	0.60	0.64
$T = 500$	0.00	0.26	0.89	0.83
$T = 1000$	0	0.26	0.96	0.95
$\tau_U = 3SD, \tau_L = -3SD$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
$T = 100$	0.98	1.00	0.98	0.98
$T = 500$	0.99	1.00	0.99	0.99
$T = 1000$	0.99	1.00	0.99	0.97

The cells are the fractions (out of 100 new replications) that the RMSE of the "Model plus judgment" is smaller than the RMSE of the "Model" (the first order autoregression). *SD* is the estimated standard deviation of the estimated factor from FA. $\phi = 0.8$, and $N = 5, T = 100, 500$ or 1000

the range of N. So, including more (or less) variables to create the factor has no impact on practical performance.

Table 3 focuses on the effect of the sample size. Comparing the numbers across the rows in the bottom panel shows little differences across the sample sizes, and hence the method shows consistency. When the full dynamic factor model in (1) is used, the first panel shows that more persistence in the factor makes the time series model to work better.

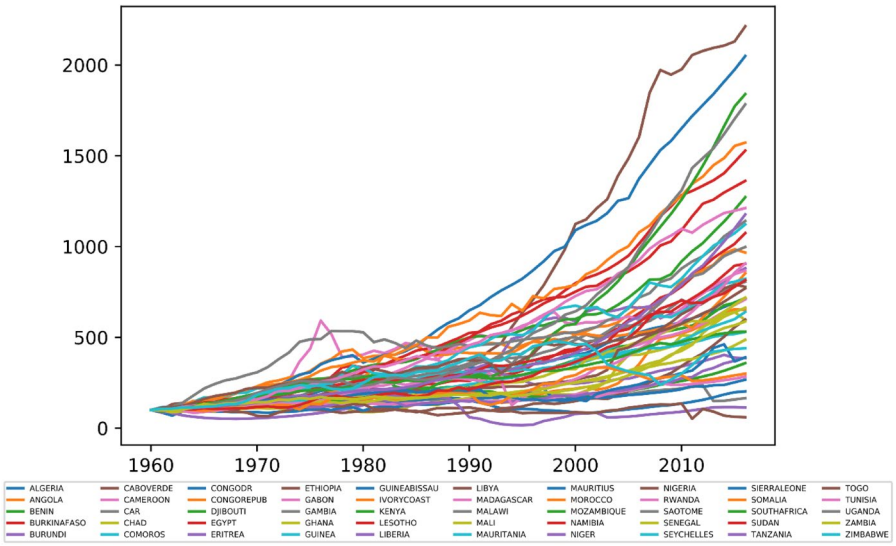

Fig. 1 Indexed gross domestic product in levels in countries in Africa 1961–2016, where Equatorial Guinea and Botswana are excluded from the graph (Source: Franses 2020)

Forecasting Real GDP Growth in Africa

To illustrate the methodology for actual data, I consider annual data for real GDP growth rates for 52 countries in Africa. These data are selected for illustrative purposes, also as it matches with the interest of the journal. Of course, there is no guarantee that judgmental forecasts in this illustration shall be better than model forecasts. The GDP levels for 1960–2016 are presented in Fig. 1, and for all countries the 1960 GDP value is standardized at 100. The sample of growth rates runs from 1961 to 2016, which is 56 time series observations. Write $y_{i,t}$ as the growth rate for country i in year t. The sample T is divided in T_1 and T_2, where $T_1 = 34$. I create a forecast for the first year in sample T_2, then move the estimation sample to $T_1 = 35$, and so on. This gives $T_2 = 22$ recursive one-step-ahead forecast errors. For each model and each country, the RMSPE is computed for various forecasting schemes. To increase the degrees of freedom, I will not apply factor analysis per country to all 51 other countries, but I will apply some pre-selection rules to reduce the number of variables when estimating the factors and the factor loadings.

The following models (forecasting schemes) are considered

No change forecast $y_{i,T_1+1} = y_{i,T_1}$

AR(1) $y_{i,T_1+1} = \alpha_i + \beta_i y_{i,T_1}$

AR(1) with Factor, version 1 $y_{i,T_1+1} = \alpha_i + \beta_i y_{i,T_1} + \gamma_i F_{i,T_1}$

Where the parameters are estimated using OLS, and where $F_{i,t}$ is computed per each country using the growth rates for ten countries with $y_{j,t-1}$ which have the largest R^2 in the lagged regression

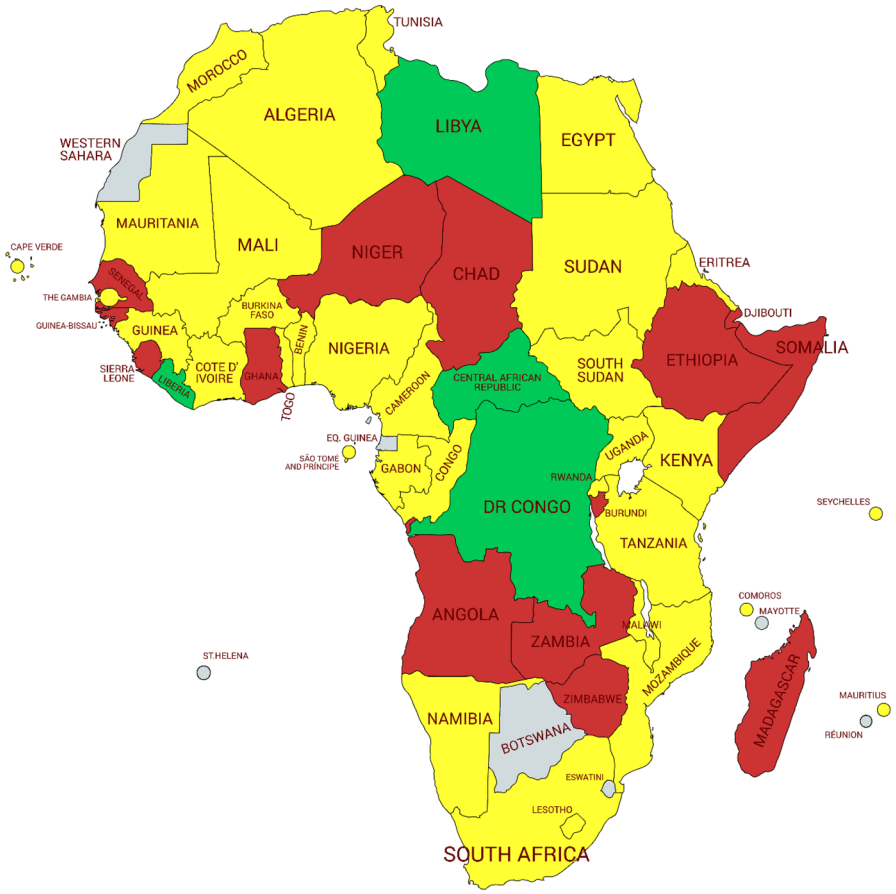

Fig. 2 Three clusters of countries with similar growth according to the dynamic time warping technique combined with k-means clustering (Source: Franses 2020)

$$\boldsymbol{y_{i,t} = \mu_i + \pi_i y_{j,t-1} + v_i}$$

with $j = 1,2, i-1, i+1, .., N = 52$.

AR(1) with Factor, version 2 $y_{i,T_1+1} = \alpha_i + \beta_i y_{i,T_1} + \gamma_i F_{i,T_1}$

Now, $F_{t,i}$ is computed per each country for using the growth rates of ten countries with $y_{j,t}$ which have the largest R^2 in the contemporaneous regression

$$\boldsymbol{y_{i,t} = \mu_i + \pi_i y_{j,t} + v_i}$$

with $j = 1,2, i-1, i+1, .., N = 52$.

The next three models are based on a cluster analysis using the dynamic time warping technique and k-means clustering proposed in Franses and Wiemann (2020), where the results appear in Franses (2020). Figure 2 gives the three clusters. Cluster 1 includes Angola and the thirteen countries with the same colour, Cluster 2

Table 4 Frequency of highest and lowest accuracy across twelve different models for 52 countries in Africa (lowest RMSPE)

	Highest accuracy	Lowest accuracy
No change	8	24
AR(1)	4	0
Model 1	5	1
Model 2	7	5
Model 3.1	3	2
Model 3.2	6	6
Model 3.3	2	0
Model 1, with judgment	2	1
Model 2, with judgment	5	1
Model 3.1, with judgment	5	8
Model 3.2, with judgment	2	2
Model 3.3, with judgment	3	2
Total	52	52

involves the Central African Republic and three countries with the same colour and Cluster 3 is the largest cluster with Algeria and thirty-one other countries (hence the choice for $T_1 = 34$).

AR(1) with Factor, version 3.1 $y_{i,T_1+1} = \alpha_i + \beta_i y_{i,T_1} + \gamma_i F^1_{i,T_1}$

With $F^1_{i,t}$ is computed using the countries in cluster 1 with $y_{j,t-1}$ (excluding the own country).

AR(1) with Factor, version 3.2 $y_{i,T_1+1} = \alpha_i + \beta_i y_{i,T_1} + \gamma_i F^2_{i,T_1}$

With $F^2_{i,t}$ is computed using the cluster 2 countries with $y_{j,t-1}$ (excluding the own country).

AR(1) with Factor, version 3.3 $y_{i,T_1+1} = \alpha_i + \beta_i y_{i,T_1} + \gamma_i F^3_{i,T_1}$

With $F^3_{i,t}$ is computed using the countries in cluster 3 with $y_{j,t-1}$ (excluding the own country).

This sums to five types of "model forecasts".

Next, all five factor models are again analysed, but now only including the value of the factor if it only takes a value more than 1.5 standard deviations away from the mean or more negative than -1.5 standard deviations from the mean. These are the five types of "adjusted forecasts".

In total, this gives the results for twelve forecasting schemes, where the last five could be seen as corresponding with "simple time series model plus adjustment".

Table 4 reports on the number of times each of these twelve models provides most accurate forecasts. The no change model provides an interesting outcome as it is most often best (for 8 out of 52 countries), yet at the same time it is most often worst (24 out of 52 countries). In general, the factor models outperform in 40[13] of the 52 cases, and hence this shows the merits of factor models in general.

[13] 52 minus 8 (no change) and 4 (AR(1)).

Table 5 How often does a factor model provide more accurate forecasts than an AR(1) model in a pairwise comparison?

Model	Countries (out of 52)
1	23
2	20
3.1	22
3.2	18
3.3	21
1 with thresholds	23
2 with thresholds	23
3.1 with thresholds	16
3.2 with thresholds	20
3.3 with thresholds	23

Table 5 presents the results when we compare ten factor models (five without, five with judgement) relative to a simple AR(1) time series model. On average across countries, there are no stark differences, and it seems that the AR(1) model performs quite well in general.

How often does a model with judgment provide more accurate forecasts than a factor model without judgment model in a pairwise comparison? For models based on the preselection methods 1, 2, 3.3, 3.2, and 3.3, the outcome is that judgment is more accurate for 23, 30, 20, 27 and 32 (out of 52 countries), respectively. We see that for the versions in 2 and 3.3 (with the cluster 3 countries) some improvement can be obtained by incorporating judgment.

Conclusion

This paper proposed a first attempt to formalize judgmental adjustment of model-based forecasts. An important reason to do so is that this allows to track and trace forecast errors and to improve either the model, the adjustment or both. In the application, the first order autoregression was used, and perhaps a second order model could have been better. At the same time, judgment involved the inclusion of the factor value when it exceeded 1.5 times its standard error, and perhaps 2 times or 3 times or even 0.5 times could have been better. Much further analysis can be pursued.

A simulation experiment showed that, when the DGP is indeed a model plus adjustment, the methodology provided a replication of the data features of the DGP. Also, the sample size and the number of variables involved did not seem to matter much. The empirical illustration on forecasting real GDP growth using various versions of models and forecasting schemes did not provide overwhelming evidence in favour of judgmental adjustment of model-based forecasts.

Further experience with the proposed methodology thus seems warranted. Given that judgement might be done because experts somehow may expect future outliers,

applications to situations, where more exceptional observations may be expected, might perhaps be more informative.

Acknowledgements The author is grateful to Olivier Mulkin for research assistance and to two anonymous reviewers for their detailed and very helpful comments.

Funding The author has received no funding for this research.

Data Availability The data can be obtained from the author upon request.

Declarations

Conflict of interest The author has no conflict of interest.

References

Altavilla, C., R. Giacomini, and G. Ragusa. 2017. Anchoring the yield curve using survey expectations. *Journal of Applied Econometrics* 32 (6): 1055–1068.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221.

Bai, J., and S. Ng. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146 (2): 304–317.

Blattberg, R.C., and S.J. Hoch. 1990. Database models and managerial intuition: 50% model + 50% manager. *Management Science* 36: 887–899.

Bunn, D.W., and A.A. Salo. 1996. Adjustment of forecasts with model consistent expectations. *International Journal of Forecasting* 12 (1): 163–170.

Capistran, C., and A. Timmermann. 2009. Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking* 41: 365–396.

Clements, M.P. 1995. Rationality and the role of judgement in macroeconomic forecasting. *Economic Journal* 105: 410–420.

Davydenko, A., and R. Fildes. 2013. Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting* 29 (3): 510–522.

Dovern, F., U. Fritsche, and J. Slacalek. 2012. Disagreement among forecasters in G7 countries. *The Review of Economics and Statistics* 94: 1081–1096.

Fildes, R., and H.O. Stekler. 2002. The state of macroeconomic forecasting. *Journal of Macroeconomics* 24 (4): 435–468.

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikopoulos. 2009. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* 25: 3–23.

Franses, P.H. 2014. *Expert adjustments of model forecasts; theory, practice and strategies for improvement*. Cambridge: Cambridge University Press.

Franses, P.H. 2020. Do African economies grow similarly? *Cybernetics and Systems. an International Journal* 51: 746–756.
Franses, P.H., and R. Legerstee. 2009. Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting* 25: 35–47.
Franses, P.H., and R. Legerstee. 2010. Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting* 29: 331–340.
Franses, P.H., and D.J.C. van Dijk. 2019. Combining expert-adjusted forecasts. *Journal of Forecasting* 38: 415–421.
Franses, P.H., and T. Wiemann. 2020. Intertemporal similarity of economic time series: an application of dynamic time warping. *Computational Economics* 56: 59–75.
Franses, P.H., H. Kranendonk, and D. Lanser. 2011. One model and various experts: evaluating Dutch macroeconomics forecasts. *International Journal of Forecasting* 27: 482–495.
Jo, S., and R. Sekkel. 2019. Macroeconomic uncertainty through the lens of professional forecasters. *Journal of Business & Economic Statistics* 37 (3): 436–446.
Kahneman, D. 2012. *Thinking, fast and slow*. London: Penguin.
Kim, H.H., and N.R. Swanson. 2018. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting* 34 (2): 339–354.
Lahiri, K., and X. Sheng. 2010. Measuring forecast uncertainty by disagreement: the missing link. *Journal of Applied Econometrics* 25: 514–538.
Lahiri, K., H. Peng, and X. Sheng. 2015. Measuring uncertainty of a combined forecast and some tests for forecaster heterogeneity, CESifo Working Paper Series 5468.
Lamont, O.A. 2002. Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization* 48: 265–280.
Laster, D., P. Bennett, I.S. Geoum, and I. S. 1999. Rational bias in macroeconomic forecasts. *The Quarterly Journal of Economics* 114: 293–318.
Lawrence, M., P. Goodwin, M. O'Connor, and D. Onkal. 2006. Judgmental forecasting: a review of progress over the last 25 years. *International Journal of Forecasting* 22 (3): 493–518.
Legerstee, R., and P.H. Franses. 2015. Does disagreement amongst forecasters have predictive value? *Journal of Forecasting* 34: 290–302.
Mathews, B.P., and A. Diamantopoulos. 1986. Managerial intervention in forecasting: an empirical investigation of forecast manipulation. *International Journal of Research in Marketing* 3: 3–10.
McNees, S. 1990. Man vs. model? The role of judgment in forecasting. *New England Economic Review* 41–52. (issue July)
Reifschneider, D., and P. Tulip. 2019. Gauging the uncertainty of the economic outlook using historical forecasting errors: the Federal Reserve's approach. *International Journal of Forecasting* 35: 1564–1582.
Robertson, J.C., E. Tallman, and C.H. Whiteman. 2005. Forecasting using relative entropy. *Journal of Money, Credit and Banking* 37 (3): 383–401.
Stekler, H.O. 2007. The future of macroeconomic forecasting: understanding the forecasting process. *International Journal of Forecasting* 23 (2): 237–248.
Stock, J.H., and M.W. Watson. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–1179.
Tallman, E., and S. Zaman. 2020. Combining survey long-run forecasts and nowcasts using BVAR forecasts using relative entropy. *International Journal of Forecasting* 36 (2): 373–398.
Tetlock, Ph., and D. Gardner. 2015. *Superforecasting: the art & science of prediction*. London: Random House Books.
Turner, D.S. 1990. The role of judgment in macroeconomic forecasting. *Journal of Forecasting* 9 (4): 315–345.
Wallis, K.F. 1989. Macroeconomic forecasting: a survey. *Economic Journal* 99 (March): 28–61.