# Assessing GDP forecasts from autoregressive models: the impact of model complexity and training dataset

J. Daniel Aromí[1] and Paula Bonel[2]

[1]IIEP-Baires UBA-Conicet and FCE, Universidad Católica Argentina
[2]IIEP-Baires UBA-Conicet

August 31, 2021

**Abstract**

We evaluate different strategies to forecast GDP through autoregressive models. The analysis covers 46 emerging and advanced economies. Two key dimensions are explored. We consider alternative training datasets: single country vs. panel of countries. Simultaneously, we evaluate models that differ in terms of complexity. The results indicate that expanding the training dataset from single country to a panel of countries results in significant improvements in performance. In addition, we verify a strong complementarity between dataset size and model complexity. If individual country data is used for training, more complex models do not improve performance. In contrast, when a panel of countries is used for training, higher model complexity results in more accurate forecasts. The performance of these model forecasts is similar to professional forecasts. The results are verified for different forecast horizons and for emerging and advanced economies. Our results can prove valuable for forecasting exercises in other contexts and in the case of alternative forecast models.

JEL classification: C53, E37

Keywords: macroeconomic forecasting, forecast comparisons, GDP growth.

## 1 Introduction

GDP forecasts are important for numerous significant economic decisions. For example, irreversible investment decisions and consumption plans require an assessment of the

path of the economy over several years ahead. Also, policymakers, portfolio managers and multilateral organizations make decisions based on forecasts over long horizons.

Given the relevance of GDP forecasts and the challenges associated with the task, there exists an important literature that proposes alternative forecasting strategies (Canova and Ciccarelli, 2009; Cuaresma et al., 2016), evaluates the performance of professional forecasters (Timmermann, 2007; De Resende, 2014; Hellwig, 2018) and compares the accuracy of forecasting models versus professional forecasts (Aromí, 2019; Timmermann and Zhu, 2019).

Contributing to this literature, in this study we focus on two interrelated aspects: the selection of the training dataset and the choice of model complexity. The choice of training data is a significant one since there exists a trade off between fitting models using more information provided by larger datasets versus focusing on specific information that is deemed more strongly connected with the forecasting task. We evaluate this trade-off jointly with a second choice associated to model complexity. This choice involves a second trade off. While more complex models with more variables and parameters can learn richer associations between inputs and outputs, parsimonious models are able to reduce variance originated in noisy model estimation (Banerjee and Marcellino, 2006; Giannone et al., 2020). These trade offs imply that, a priori, the most convenient strategy is unknown and empirical analyses are needed to guide practitioners.

The analysis is carried out for the simple but relevant case of autoregressive models. Going beyond the more prevalent short term horizon, we implement multiple years ahead forecasts. We propose traditional distributed lag models and heterogeneous autoregressive models (Corsi, 2009). Training datasets range from the small-scale individual country time series to the sizeable history of GDP growth for all countries in the panel. The sample covers 46 advanced and emerging economies from 1950 through 2019.

Our results lead to solid conclusions regarding how the interaction of training dataset and model complexity determines performance. First, we find that there are consistent gains associated with training models with information from multiple countries. That is, the information provided by the growth trajectory of a large collection of countries allows for trained models that anticipate individual country growth trajectories in a more accurate manner. Second, if models are trained with large datasets, more complex models allow for additional gains in forecast accuracy. These gains are not observed when the training dataset is small (i.e. individual country data). In this way, a strong

complementarity between model complexity and training dataset size is established. Third, we verify that these findings are observed for different country groups (advanced and emerging) and for alternative forecast horizons.

To benchmark the performance of these model forecasts, we implement a comparison with professional forecasts. More specifically, we consider World Economic Outlook forecasts generated by the International Monetary Fund staff. The results indicate that model forecast compare well with professional forecasts. This is particularly the case for multiple years ahead forecasts. This relative performance suggests that, while simple, the analysis proposed in this work corresponds to a competitive approach for the forecasting task.

We believe that the results here reported provide insights that might prove useful in forecast exercises that go beyond the current application. For example, similar findings might characterize forecasts of other macroeconomic variables such as inflation and exchange rates. In these contexts forecasters need to select model complexity, e.g. number of lags, and the collection of data in training datasets, e.g. number of countries. Our findings suggest that it is plausible that forecast accuracy might improve when large datasets and complex models are implemented. For an example in this direction, in an exercise of exchange rate forecasts, moving from individual country data to pooled data by country group (i.e. advanced and emerging countries) Bonel (2021) is able to generate gains in accuracy for more than half of the cases.

The quantitative findings here reported contribute to an important literature that focuses on GDP forecasts. Chauvet and Potter (2013) present a detailed survey on forecasting methods for GDP, ranging from traditional univariate autoregressive models, VARs, SVARs to DSGE models and, more recently, the combination of these techniques with a Bayesian framework and factor models. They point that simple autoregressive models are hard to beat by large-scale macroeconomic models. They also highlight that several papers find that some financial and macroeconomic series have significant predictive power for future economic activity across a number of countries when they are included in simple models. Some of the methods found to be useful to forecast economic activity are factor models, mixed frequency models, non-linear models, and forecast combinations. For example, recently, Cuaresma et al. (2016) found positive results by using a Bayesian variant of global vector autoregressive (B-GVAR). The element that distinguish our contribution from previous literature is the insights that emerge from the

joint choice of model complexity and training dataset. An exploration of our findings in the context of these data rich environments is an interesting extension of the current work.

It is worth noting that there is an alternative interpretation of some aspects of our exercise that follows previous work by Hoogstrate et al. (2000). Training models using data from a panel of countries can be viewed as imposing the restriction of training a single or few models. This constraint can generate losses associated to disregarding heterogeneities. This loss might be small if true parameters are different but exhibit some similarity. In fact, Hoogstrate et al. (2000) find that pooling leads to a reduction of the errors of the estimates. Other studies that have explored these questions are: Chen and Ranciere (2019) and Garcia-Ferrer et al. (1987). These studies do not explore the issue of model complexity as we do in the current work.

The rest of the paper is organized as follows. The next two sections present, respectively, data and methodology. Section 4 shows the performance of alternative forecasting strategies. Section 5 compares these results to those observed for professional forecasts. The last section presents concluding remarks.

## 2 Data

The dataset consists of annual GDP growth data. The full sample covers the period 1950–2019 for 46 countries. [1] Yearly growth rates are expressed as the difference of the natural logarithm of yearly GDP. Data are from the IMF's World Economic Outlook database and extended back to 1950 using information from Penn World Table 10.0 [2].

Expert forecasts are evaluated and compared to forecast models. In this case, we work with the IMF's staff forecasts that are published in multiple releases of the World Economic Outlook. The analysis focuses on one year ahead to five years ahead forecasts for 46 advanced and emerging countries covering the period from 2006 to 2019.

[1] Sampled countries were given by the largest 50 economies according to 2000 GDP in US dollars as reported by WEO´s database. We excluded oil exporting countries. Sampled countries were: Argentina, Australia, Austria, Bangladesh, Belgium, Bolivia, Brazil, Canada, Chile, China, Colombia, Czech Republic, Denmark, Egypt, Finland, France, Germany, Greece, Hong Kong, Hungary, India, Ireland, Israel, Italy, Japan, Malaysia, Mexico, Morocco, Netherlands, New Zealand, Norway, Pakistan, Peru, Philippines, Poland, Portugal, Puerto Rico, Russia, Singapore, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Thailand, Turkey, Ukraine, United Kingdom and United States. Due to data limitations four countries were excluded: Czech Republic, Puerto Rico, Russia and Ukraine.

[2] https://www.rug.nl/ggdc/productivity/pwt/?lang=en

Table 1: Annual Growth Rate % (Period 1960-2019)

| | Mean | Median | Min | Max | Q1 | Q3 |
|---|---|---|---|---|---|---|
| All countries | 3.5 | 3.4 | -22 | 19.9 | 1.4 | 5.6 |
| Advanced | 2.7 | 2.6 | -13.4 | 19.9 | 1.1 | 4.3 |
| Emerging | 4.3 | 4.6 | -22 | 18.3 | 2.2 | 6.8 |

Note: Descriptive statistics for annual GDP growth rates over the period 1950-2019 for 46 countries. Advanced countries are those classified as high income countries by the World Bank in year 2000. The panel covers 24 advanced countries and 22 emerging countries).

## 3 Methodology

Model forecasts are generated combining different specifications of autoregressive models with alternative training datasets. Two types of autoregressive models are considered. First, we generate iterated forecasts using distributed lag models with number of lags that range from 1 through 15. In addition, we consider heterogeneous autoregressive models (Corsi, 2009) to generate direct forecasts. Small training datasets are given by the yearly time series of GDP growth for the corresponding country. On the other hand, two version of large training datasets are considered. In the first version, the dataset is given by the history of GDP growth for the panel of 46 countries. For the second version of the large training dataset, we split the sample between advanced and emerging countries. In this case, for each country, the relevant dataset is given by the GDP growth history of countries in the same group to which the relevant country belongs. Advanced countries are those classified as high income countries by the World Bank in year 2000. Countries not classified as high income are labeled "emerging".

In each case, forecasts are estimated recursively as a pseudo-out-of-sample exercise. Forecasts are based only on values of the series up to the date, $t$, on which the forecast is made. That is, parameters are estimated in each period using data from the beginning of the sample through the current forecasting date (i.e. expanding windows). We consider forecasts horizons from one up to five years ahead.

More formally, the iterated forecasts are generated by distributed lag models that are given by:

$$gr_{c,t+1} = \alpha + \sum_{l=0}^{K} \beta_k gr_{c,t-k} + u_{c,t+1}$$

Where $gr_{c,t}$ is the growth rate of country $c$ on year $t$ and $K$ refers to the number of lags in the model. We evaluate model specifications with $K$ equal to 1, 5, 10 and 15. In this case, multiple years ahead forecasts are generated iterating the estimated model.

Heterogeneous autoregressive models (HAR) are proposed considering that they provide a parsimonious way to capture the relationship between rates of GDP growth corresponding to distant years. In this case, direct multiple years ahead forecast are generated through two specifications that we label Model 1:

$$gr_{c,t+h} = \alpha + \beta_1 gr_{c,t} + \beta_5 \tilde{gr}_{c,t,t-5} + u_{c,t+h}$$

and Model 2:

$$gr_{c,t+h} = \alpha + \beta_1 gr_t + \beta_5 \tilde{gr}_{c,t,t-5} + \beta_{15} \tilde{gr}_{c,t,t-15} + u_{c,t+h}$$

where $\tilde{gr}_{c,t,t-k}$ is the average annual GDP growth rate for country $c$ observed between year $t-k$ and year $t$.

The out of sample forecasts simulations start on year 2005, $t = 2005$. The last year on which forecast are generated and evaluated is year 2018. The last realization evaluated corresponds to year 2019.

## 4 Performance under alternative modeling and training strategies

Table 2 shows the root mean square errors (RMSEs) under different forecast horizons and different forecasting strategies. The performance of each strategy is compared to the benchmark in which an autoregressive model with one lag is trained with data from individual countries.

Analyzing the performance of alternative forecasting strategies some clear patterns emerge. The first row shows that in the case of short term forecasts, when models are trained with individual country data, increments in model complexity do not help and sometimes hurt forecast accuracy. This observation is for the most part valid when longer forecast horizons are considered. That is, with models trained using small datasets, com-

plexity does not lead to consistent gains in performance. The only noticeable exemption to this rule is observed in the case of the simple heterogeneous autoregressive model (Model 1). This particular finding suggests that this class models constitutes a convenient way to learn relationships between relatively distant realizations of GDP growth.

Importantly, when the performance of models trained with single country data is compared to the performance of models trained with the country panel it becomes clear that there are important gains associated with using a large dataset. This finding is verified for different forecast horizons and different levels of model complexity.

A more detailed analysis shows that the best performance is observed for relatively complex models that are trained with data from the panel of countries. While in the case of the simplest model, autoregressive model with one lag, there are no noticieably differences in the performance under different training datasets, when more complex models are considered, the benefits associated with larger datasets become evident. For one year ahead forecast the best performing model is the autoregressive model with 15 lags. For longer forecast horizons, the best performance corresponds to the more complex heterogeneous autoregressive model (Model 2).

In this way, the evidence points to a complementarity between complex models and large datasets. This result is not unexepected since it can be explained observing that under larger datasets the variance of estimated parameters drops and, as a result, the disadvantages of complex models become weaker. On the other hand, it is worth emphasizing that these gains in performance could have not materialized. That is, there are plausible alternative scenarios, in which pooling data from other countries or increasing model complexity would hurt performance in a significant manner.

The message provided by these results can be transmitted in a practical manner considering a thought experiment. Consider a forecaster that uses single country data to select a model to forecast one year ahead GDP. Additionally, assume that this model is also used to produce multiple years ahead forecasts. We find that, under traditional selection criteria (AIC or BIC), the selected model would be the parsimonious one lag model. This choice is consistent with the performance observed in out of sample forecasts (first row of table 2). Now, it becomes evident that the outcome of this approach would fail to learn how to use available information. In other words, this standard practice would clearly underperform when compared to the cases in which a larger dataset is used jointly with more complex models.

Table 2: RMSEs and relative performance versus benchmark

| | | Iterative Forecasts | | | | Direct forecasts | |
|---|---|---|---|---|---|---|---|
| | | 1 lag | 5 lags | 10 lags | 15 lags | Model 1 | Model 2 |
| **h=1** (n=644) | | | | | | | |
| Country | RMSE | 0.0253 | 0.0262 | 0.0283 | 0.0341 | 0.0255 | 0.0256 |
| | ratio | 1 | 1.0356 | 1.1186*** | 1.3478** | 1.0079 | 1.0119 |
| Pool | RMSE | 0.0253 | 0.0242 | 0.0238 | 0.0237 | 0.0242 | 0.0238 |
| | ratio | 1 | 0.9565* | 0.9407** | 0.9368** | 0.9565** | 0.9407** |
| **h=2** (n=598) | | | | | | | |
| Country | RMSE | 0.0284 | 0.0289 | 0.0298 | 0.036 | 0.0281 | 0.0270 |
| | ratio | 1 | 1.0176 | 1.0493 | 1.2676* | 0.9894 | 0.9507 |
| Pool | RMSE | 0.0288 | 0.0267 | 0.026 | 0.0259 | 0.0266 | 0.0258 |
| | ratio | 1.0141 | 0.9401* | 0.9155** | 0.912** | 0.9366** | 0.9085** |
| **h=3** (n=552) | | | | | | | |
| Country | RMSE | 0.0296 | 0.0289 | 0.0296 | 0.0421 | 0.028 | 0.0279 |
| | ratio | 1 | 0.9764 | 1 | 1.4223 | 0.9459* | 0.9426 |
| Pool | RMSE | 0.0302 | 0.0276 | 0.0268 | 0.0267 | 0.0268 | 0.0259 |
| | ratio | 1.0203 | 0.9324 | 0.9054* | 0.9020** | 0.9054** | 0.8750*** |
| **h=4** (n=506) | | | | | | | |
| Country | RMSE | 0.0303 | 0.0294 | 0.0298 | 0.0356 | 0.0284 | 0.0293 |
| | ratio | 1 | 0.9703 | 0.9835 | 1.1749 | 0.9373* | 0.9670 |
| Pool | RMSE | 0.0311 | 0.0288 | 0.0275 | 0.0273 | 0.0277 | 0.0265 |
| | ratio | 1.0264 | 0.9505 | 0.9076* | 0.9010* | 0.9142* | 0.8746** |
| **h=5** (n=460) | | | | | | | |
| Country | RMSE | 0.0249 | 0.0242 | 0.0249 | 0.0312 | 0.0235 | 0.0346 |
| | ratio | 1 | 0.9719 | 1 | 1.253** | 0.9438 | 1.3896 |
| Pool | RMSE | 0.0257 | 0.0235 | 0.0224 | 0.0222 | 0.023 | 0.0217 |
| | ratio | 1.0321 | 0.9438 | 0.8996 | 0.8916* | 0.9237 | 0.8715** |

Note: The benchmark model is the autoregressive model with one lag trained with individual country data. Significance levels correspond to two way clustered tests of differences in mean square errors (Thompson, 2011). *p<0.1; **p<0.05; ***p<0.01

### 4.1 Results for country groups

The performance of forecast models is evaluated by country groups. In this case, we evaluate two versions of large datasets. First, we consider a training dataset equal to the panel of 46 sampled countries. Second, we use large datasets corresponding to two country groups: advanced and emerging. In this way we can explore in more detail the impact of variations in the training dataset on forecast performance.

Table 3 presents the results for country groups when models are trained with the panel of 46 sampled countries. In the case of emerging countries, when models are trained with country panels, accuracy increases with model complexity. In the case of advanced economies a similar pattern is observed but differences in accuracy with respect to the benchmark model are not statistically significant. We conjecture that this failure to reject the null hypothesis is probably due to the small testing sample size and the associated low power.

Table 4 presents the complementary exercise in which two large training datasets are used for different country groups. In this case, we observe statistically significant gains in forecast accuracy. This result suggest that, in the case of advanced economies, there is a point in which larger datasets with more heterogeneous countries have a negative effect on forecast accuracy.

In the case of emerging economies, a different conclusion is reached. While we observe statistically gains in accuracy versus the benchmark model, these gains are smaller than those observed when the training dataset is given by the panel of 46 countries. That is, for emerging economies, the gains associated to more observations outweight the potential losses resulting from the inclusion of information of countries with different growth trajectories. These contrasting comparisons for the case of emerging and advanced indicate that the selection of an optimal training dataset is case-sensitive and it is worth addressing this question in more detail in future research.

## 5 Model forecast versus professional forecasts

To benchmark the performance of the previous section, we compare model forecast with professional forecasts. In particular, we work with World Economic Outlook forecasts. These forecasts are released in April of each year $t$ for up to five years ahead. We take into account lags related the publication of GDP data in order to simulate the available

Table 3: RMSEs for different country groups

| | | Iterative Forecasts | | | | Direct forecasts | |
|---|---|---|---|---|---|---|---|
| | | 1 lag | 5 lags | 10 lags | 15 lags | Model 1 | Model 2 |
| **h=1** | | | | | | | |
| Advanced | RMSE | 0.0235 | 0.0222 | 0.0216 | 0.0216 | 0.0221 | 0.0215 |
| | ratio | 1.0731 | 1.0137 | 0.9863 | 0.9863 | 1.0091 | 0.9817 |
| Emerging | RMSE | 0.027 | 0.0263 | 0.026 | 0.0259 | 0.0264 | 0.0261 |
| | ratio | 0.9441 | 0.9196** | 0.9091* | 0.9056* | 0.9231** | 0.9126* |
| **h=2** | | | | | | | |
| Advanced | RMSE | 0.029 | 0.026 | 0.025 | 0.025 | 0.0257 | 0.0247 |
| | ratio | 1.1197** | 1.0039 | 0.9653 | 0.9653 | 0.9923 | 0.9537 |
| Emerging | RMSE | 0.0286 | 0.0274 | 0.027 | 0.0268 | 0.0275 | 0.0269 |
| | ratio | 0.9286 | 0.8896** | 0.8766** | 0.8701** | 0.8929** | 0.8734** |
| **h=3** | | | | | | | |
| Advanced | RMSE | 0.0316 | 0.0278 | 0.0265 | 0.0264 | 0.0268 | 0.0255 |
| | ratio | 1.1408** | 1.0036 | 0.9567 | 0.9531 | 0.9675 | 0.9206 |
| Emerging | RMSE | 0.0288 | 0.0274 | 0.0271 | 0.0269 | 0.0269 | 0.0262 |
| | ratio | 0.9143 | 0.8698** | 0.8603* | 0.854* | 0.854** | 0.8317** |
| **h=4** | | | | | | | |
| Advanced | RMSE | 0.0326 | 0.0293 | 0.0274 | 0.0273 | 0.0277 | 0.0264 |
| | ratio | 1.1399** | 1.0245 | 0.958 | 0.9545 | 0.9685 | 0.9231 |
| Emerging | RMSE | 0.0294 | 0.0283 | 0.0276 | 0.0274 | 0.0277 | 0.0266 |
| | ratio | 0.9216 | 0.8871* | 0.8652* | 0.8589* | 0.8683* | 0.8339** |
| **h=5** | | | | | | | |
| Advanced | RMSE | 0.0262 | 0.0231 | 0.0208 | 0.0206 | 0.0212 | 0.0198 |
| | ratio | 1.1644* | 1.0267 | 0.9244 | 0.9156 | 0.9422 | 0.8800 |
| Emerging | RMSE | 0.0251 | 0.0239 | 0.0241 | 0.0238 | 0.0248 | 0.0236 |
| | ratio | 0.9161 | 0.8723* | 0.8796 | 0.8686 | 0.9051 | 0.8613 |

Note: The benchmark model is the autoregressive model with one lag trained with individual country data. Significance levels correspond to two way clustered tests of differences in mean square errors (Thompson, 2011). *p<0.1; **p<0.05; ***p<0.01

Table 4: RMSEs: models trained by country group

| | | Iterative Forecasts | | | | Direct forecasts | |
|---|---|---|---|---|---|---|---|
| | | 1 lag | 5 lags | 10 lags | 15 lags | Model 1 | Model 2 |
| **h=1** | | | | | | | |
| Advanced | RMSE | 0.0211 | 0.0207 | 0.0209 | 0.0211 | 0.0207 | 0.0206 |
| | ratio | 0.9635 | 0.9452** | 0.9543 | 0.9635 | 0.9452** | 0.9406* |
| Emerging | RMSE | 0.0276 | 0.0269 | 0.0263 | 0.0261 | 0.0269 | 0.0264 |
| | ratio | 0.965 | 0.9406** | 0.9196** | 0.9126** | 0.9406** | 0.9231*** |
| **h=2** | | | | | | | |
| Advanced | RMSE | 0.0251 | 0.0241 | 0.0244 | 0.0247 | 0.0241 | 0.0236 |
| | ratio | 0.9691 | 0.9305* | 0.9421 | 0.9537 | 0.9305* | 0.9112* |
| Emerging | RMSE | 0.0298 | 0.0285 | 0.0276 | 0.0272 | 0.0283 | 0.0276 |
| | ratio | 0.9675 | 0.9253* | 0.8961** | 0.8831** | 0.9188** | 0.8961*** |
| **h=3** | | | | | | | |
| Advanced | RMSE | 0.0267 | 0.0253 | 0.0255 | 0.0256 | 0.0252 | 0.0242 |
| | ratio | 0.9639 | 0.9134 | 0.9206 | 0.9242 | 0.9097* | 0.8736** |
| Emerging | RMSE | 0.0306 | 0.029 | 0.0281 | 0.0278 | 0.028 | 0.0271 |
| | ratio | 0.9714 | 0.9206 | 0.8921* | 0.8825* | 0.8889** | 0.8603*** |
| **h=4** | | | | | | | |
| Advanced | RMSE | 0.0278 | 0.0263 | 0.0261 | 0.0263 | 0.026 | 0.0246 |
| | ratio | 0.972 | 0.9196 | 0.9126 | 0.9196 | 0.9091 | 0.8601** |
| Emerging | RMSE | 0.0313 | 0.0302 | 0.029 | 0.0286 | 0.029 | 0.0283 |
| | ratio | 0.9812 | 0.9467 | 0.9091 | 0.8966 | 0.9091* | 0.8871** |
| **h=5** | | | | | | | |
| Advanced | RMSE | 0.0213 | 0.0202 | 0.0195 | 0.0197 | 0.02 | 0.0186 |
| | ratio | 0.9467 | 0.8978 | 0.8667 | 0.8756 | 0.8889 | 0.8267* |
| Emerging | RMSE | 0.0263 | 0.0251 | 0.0248 | 0.0245 | 0.0252 | 0.0243 |
| | ratio | 0.9599 | 0.9161 | 0.9051 | 0.8942 | 0.9197 | 0.8869 |

Note: The benchmark model is the autoregressive model with one lag trained with individual country data. Significance levels correspond to two way clustered tests of differences in mean square errors (Thompson, 2011). *p<0.1; **p<0.05; ***p<0.01

information that forecasters have when making their estimations. More specifically, WEO's forecast released in April of year $t$ are compared with forecast generated with data in which the most recent observation is GDP growth for year $t-1$. Table 5 summarises the results when models are trained by country groups. The last column reports the performance of WEO's forecasts which used as benchmarks to evaluate the performance of model forecasts.

We find that experts forecasts tend to have a better out of sample performance for shorter horizons, where that improvement is usually statistically significant with respect the simple one lag model but not for the more complex models (except Model 1). Even though RMSEs for Model 2 tends to be lower than for WEO forecasts, that difference is not statistically significant. Limited amount of observations may affect the results.

In nontabulated exercises we observe that similar comparisons are obtained when we compare WEO's forecasts against forecast from models trained with the alternative large dataset (46 country panel). Overall, these analysis suggest that the analysis of forecast from autoregressive exercises coorespond to model forecast that are competitive versus professional forecasts.

## 6 Concluding remarks

This work explores the performance of GDP forecasts in the simple but relevant case of autoregressive models. The analysis shows that there are important accuracy gains to be realized from the joint use of datasets from country panels and complex models. These findings are robust to changes in forecast horizon and the subset of countries under consideration.

We believe that these results are relevant for the implementation of forecasting tasks in other settings. For example, going beyond GDP forecast, it is plausible that similar findings could be observed in the case of other macroeconomic indicators such as inflation. Also, our findings could be evaluated in the case in which GDP forecasts are informed by a richer set of variables. It must be noted that in this case, emerges an additional concern regarding the dataset size. That is, the availability of time series to build training datasets with this enlarged set of variables.

Table 5: Experts and Models: RMSE comparison

| | | Iterative Forecasts | | Direct Forecasts | | Experts |
|---|---|---|---|---|---|---|
| | | 1 lag | 15 lags | Model 1 | Model 2 | WEO |
| **h = 2** | | | | | | |
| All | RMSE | 0.0269 | 0.0256 | 0.0258 | 0.0252 | 0.0248 |
| | ratio | 1.0859 | 1.0352 | 1.0400 | 1.0184 | 1 |
| Advanced | RMSE | 0.0244 | 0.0240 | 0.0236 | 0.0230 | 0.0210 |
| | ratio | 1.1628 | 1.1431 | 1.1228 | 1.0979 | 1 |
| Emerging | RMSE | 0.0294 | 0.0273 | 0.0280 | 0.0274 | 0.0283 |
| | ratio | 1.0372 | 0.9649 | 0.9872 | 0.9678 | 1 |
| **h = 3** | | | | | | |
| All | RMSE | 0.0281 | 0.0263 | 0.0262 | 0.0254 | 0.0257 |
| | ratio | 1.0907*** | 1.0203 | 1.0192*** | 0.9890 | 1 |
| Advanced | RMSE | 0.0259 | 0.0248 | 0.0246 | 0.0236 | 0.0235 |
| | ratio | 1.1009** | 1.0545 | 1.0447** | 1.0043 | 1 |
| Emerging | RMSE | 0.0303 | 0.0278 | 0.0279 | 0.0273 | 0.0280 |
| | ratio | 1.0827** | 0.9931 | 0.9991 | 0.9770 | 1 |
| **h = 4** | | | | | | |
| All | RMSE | 0.0292 | 0.0272 | 0.0272 | 0.0261 | 0.0274 |
| | ratio | 1.0659** | 0.9914 | 0.9917 | 0.9538 | 1 |
| Advanced | RMSE | 0.0276 | 0.0260 | 0.0257 | 0.0243 | 0.0253 |
| | ratio | 1.0907** | 1.0301 | 1.0192* | 0.9599 | 1 |
| Emerging | RMSE | 0.0309 | 0.0283 | 0.0286 | 0.0280 | 0.0295 |
| | ratio | 1.0456 | 0.9594 | 0.9693 | 0.9488 | 1 |
| **h = 5** | | | | | | |
| All | RMSE | 0.0298 | 0.0276 | 0.0279 | 0.0266 | 0.0282 |
| | ratio | 1.0558** | 0.9788 | 0.9880 | 0.9434 | 1 |
| Advanced | RMSE | 0.0283 | 0.0267 | 0.0264 | 0.0248 | 0.0259 |
| | ratio | 1.0957** | 1.0329 | 1.0203 | 0.9588 | 1 |
| Emerging | RMSE | 0.0313 | 0.0286 | 0.0294 | 0.0285 | 0.0306 |
| | ratio | 1.0237 | 0.9344 | 0.9620 | 0.9312 | 1 |
| **h = 6** | | | | | | |
| All | RMSE | 0.0240 | 0.0223 | 0.0223 | 0.0217 | 0.0228 |
| | ratio | 1.0541* | 0.9805 | 0.9809 | 0.9529 | 1 |
| Advanced | RMSE | 0.0217 | 0.0198 | 0.0195 | 0.0186 | 0.0193 |
| | ratio | 1.1199*** | 1.0238 | 1.0078 | 0.9614 | 1 |
| Emerging | RMSE | 0.0263 | 0.0248 | 0.0251 | 0.0247 | 0.0260 |
| | ratio | 1.0123 | 0.9535 | 0.9643 | 0.9478 | 1 |

Note: IMF's experts forecasts were used as benchmark models. Model's results were generated using models trained by country group. Significance levels correspond to two way clustered tests of differences in mean square errors (Thompson, 2011). *p<0.1; **p<0.05; ***p<0.01

## Bibliography


Aromí, J. D. (2019). Medium term growth forecasts: Experts vs. simple models. *International Journal of Forecasting*, 35(3):1085–1099.

Banerjee, A. and Marcellino, M. (2006). Are there any reliable leading indicators for us inflation and gdp growth? *International Journal of Forecasting*, 22(1):137–151.

Bonel, M. P. (2021). Combination of theoretical forecasts for exchange rate forecasting. Working paper.

Canova, F. and Ciccarelli, M. (2009). Estimating multicountry var models. *International economic review*, 50(3):929–959.

Chauvet, M. and Potter, S. (2013). Forecasting output. *Handbook of Economic Forecasting*, 2:141–194.

Chen, S. and Ranciere, R. (2019). Financial information and macroeconomic forecasts. *International Journal of Forecasting*, 35(3):1160–1174.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196.

Cuaresma, J. C., Feldkircher, M., and Huber, F. (2016). Forecasting with global vector autoregressive models: A bayesian approach. *Journal of Applied Econometrics*, 31(7):1371–1391.

De Resende, C. (2014). An assessment of imf medium-term forecasts of gdp growth. *IEO Background Paper No. BP/14/01 (Washington: Independent Evaluation Office of the IMF)*.

Garcia-Ferrer, A., Highfield, R. A., Palm, F., and Zellner, A. (1987). Macroeconomic forecasting using pooled international data. *Journal of Business & Economic Statistics*, 5(1):53–67.

Giannone, D., Lenza, M., and Primiceri, G. E. (2020). Economic predictions with big data: The illusion of sparsity. *SSRN Electronic Journal*.

Hellwig, K.-P. (2018). *Overfitting in judgment-based economic forecasts: The case of IMF growth projections*. International Monetary Fund.

Hoogstrate, A. J., Palm, F. C., and Pfann, G. A. (2000). Pooling in dynamic panel-data models: An application to forecasting gdp growth rates. *Journal of Business & Economic Statistics*, 18(3):274–283.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of financial Economics*, 99(1):1–10.

Timmermann, A. (2007). An evaluation of the world economic outlook forecasts. *IMF Staff Papers*, 54(1):1–33.

Timmermann, A. and Zhu, Y. (2019). Comparing forecasting performance with panel data.