WP/20/44

IMF Working Paper

The More the Merrier? A Machine Learning Algorithm for Optimal Pooling of Panel Data

by Marijn A. Bolhuis and Brett Rayner

INTERNATIONAL MONETARY FUND

WP/20/44

IMF Working Paper

The More the Merrier? A Machine Learning Algorithm for Optimal Pooling of Panel Data

by Marijn A. Bolhuis and Brett Rayner

 WP/20/44

IMF Working Paper

European Department

The More the Merrier? A Machine Learning Algorithm for Optimal Pooling of Panel Data[1]

Prepared by Marijn A. Bolhuis and Brett Rayner

Authorized for distribution by Donal McGettigan

February 2020

Abstract

We leverage insights from machine learning to optimize the tradeoff between bias and variance when estimating economic models using pooled datasets. Specifically, we develop a simple algorithm that estimates the similarity of economic structures across countries and selects the optimal pool of countries to maximize out-of-sample prediction accuracy of a model. We apply the new alogrithm by nowcasting output growth with a panel of 102 countries and are able to significantly improve forecast accuracy relative to alternative pools. The algortihm improves nowcast performance for advanced economies, as well as emerging market and developing economies, suggesting that machine learning techniques using pooled data could be an important macro tool for many countries.

JEL Classification Numbers: C53, C45.

Keywords: Machine learning, GDP growth, forecasts, panel data, pooling.

Authors' E-Mail Addresses: marijn.bolhuis@mail.utoronto.ca; brayner@imf.org

[1] The authors would like to thank Donal McGettigan, Alex Culiuc, Vincenzo Guzzo, Romain Lafarguette, Anil Ari, and participants at the IMF EUR seminar for helpful comments and suggestions, and Morgan Maneely for outstanding research assistance. All remaining errors are our own.

CONTENTS

"The more the merrier;
the fewer, the better fare."

–English proverb

I. Introducing Optimal Pooling with Machine Learning

Macro forecasting can be challenging. To improve predictions, economists often pool data from multiple countries to forecast country-specific macroeconomic aggregates. Pooling has been shown to lower forecast errors relative to predictions based on country-specific data, especially in settings with scarce data.[2]

Although pooling is common, there is no consensus on how to select the optimal set of countries for a panel. There is a tradeoff between bias and variance in expanding a dataset to include data from additional countries. Adding countries with a similar economic structure (i.e., data-generating process) may reduce variance and improve forecasts, but adding countries with a dissimilar economic structure may introduce bias to the forecasts. Optimal pooling amounts to solving a version of the bias-variance tradeoff for which machine learning methods have specifically been developed. The aim of this paper is to use insights from machine learning to provide a method for selecting the optimal set of countries for macro panel forecasting.[3]

We develop an algorithm to select the optimal set of countries to be pooled. Our algorithm consists of two steps. First, we estimate the similarity of the economic structure of two countries by measuring the extent to which the data-generating process (DGP) from one country matches the DGP in the other country. Second, we use cross validation techniques to determine the optimal set of countries to include in the panel to minimize out-of-sample forecast errors.

Our method has several advantages over conventional panel pooling methods. Compared to conventional linear panel methods (e.g., fixed effects models with heterogeneous coefficients), our method can be applied to both linear and non-linear models. In addition, our method is designed to maximize out-of-sample, rather than in-sample, prediction accuracy. Our algorithm can also be applied to both small and large panels due to its computational efficiency. Finally, our approach of comparing the similarity of economic structures across countries can be used in a range of contexts. Although we focus on optimal

[2] Baltagi (2008) provides a survey of this literature. Baltagi & Griffin (1997), Baltagi et al. (2000), Hoogstrate et al. (2000), Gavin & Theodorou (2005), and Chen & Ranciere (2019) all find evidence that pooling improves forecasting accuracy.

[3] Machine learning methods have become popular in the recent macroeconomic forecasting literature as they tend to improve accuracy of forecasts relative to expert forecasts and traditional factor models (e.g., Tiffin (2016), Jung et al. (2018), Richardson et al. (2018), Smalter Hall (2018), Medeiros et al. (2018), Bolhuis & Rayner (2019)).

pooling to improve nowcast performance, the same approach could be used to select a group of comparator countries for any panel data analysis.

We are able to reduce forecast errors substantially when using our pooling method to nowcast real output growth. We apply the algorithm to a range of advanced economies and emerging market and developing countries, including Austria, Canada, Costa Rica, El Salvador, Germany, Iceland, Lithuania, Mexico, and Turkey. Our pooling method reduces forecast errors by up to 20 percent relative to the same forecast model using alternate pools, with the largest improvements for Austria, Canada, Costa Rica, Lithuania, and Turkey.

II. A Two-Step Method for Optimal Pooling

There is a tradeoff between bias and variance when estimating an economic model with a pooled dataset of multiple countries. More observations provide more information, making forecasts more stable, thereby reducing the variance of the forecast. However, economic structures and the DGP of the forecast variable may differ across countries. Conditional on the number of observations, this difference can lead to bias in the forecast (Box 1).

Box 1: The Bias-Variance Tradeoff

The bias-variance tradeoff can be demonstrated with a linear example. Suppose we want to forecast a variable $\boldsymbol{y_{n,t}}$ (e.g., real GDP growth) for country *n* using $\boldsymbol{V}$ predictor variables. Let $\boldsymbol{X_{n,t}}$ be the $\boldsymbol{V \times 1}$ data vector that summarizes these predictors at time *t* and denote the h-step ahead forecast of $\boldsymbol{y_{n,t}}$ as $\boldsymbol{y_{n,t+h}}$. Suppose we have training data for two countries indexed *m* and *n*, for M and N time periods. Assume the properties of the predictor data are the same for both countries, and in both cases the DGP is linear, although the processes differ by country:

$$y_{i,t+h} = \beta_i' X_{n,t} + \epsilon_{i,t+h} \; ; \; i \in \{n, m\}$$

Under certain conditions (Stock & Watson, 2006), it can be shown that the difference in expected squared errors from using data from country *m* to predict $y_{n,t+h}$ using OLS would be:[1]

$$(E\left[y_{n,t+h} - \widehat{\beta_m}' X_{n,t}\right])^2 + \sigma^2 V\left[\frac{1}{M} - \frac{1}{N}\right]$$

Here, the first term is the squared bias from forecasting for country *n* based on data from country *m* only. Note that this bias term is weakly positive, and zero if the DGP of the countries are the same. The second term measures the difference in forecast variance from using data from country *m*. Note that this variance term can be negative—contributing to lower forecast errors—if $M > N$ (i.e., if we have more data from country *m* than country *n*). Taken together, we can see that pooling data from countries with additional data may reduce variance and improve forecast results, but adding countries with a dissimilar DGP may introduce bias to the forecasts.

[1] Specifically, if the regressors are orthogonal s.t. $\frac{1}{T}\sum_{t=1}^{T} X_{n,t} X_{n,t}' = I_n$ (the identity matrix), the regressors and estimate of β_n are independently distributed, then the OLS forecast is distributed $N\left(x'\beta, c\sigma^2 \frac{K}{T}\right)$, where *x* is a typical predictor vector used to construct the forecast, and *c* is an arbitrary constant.

Conventional pooling methods are often impractical for macroeconomic forecasting. One type of pooling method is based on fixed effects (e.g., Baltagi and Griffin, 1997; Baltagi et al., 2000). This method pools data from different countries and estimates DGPs that are

potentially heterogeneous by restricting the set of regression coefficients for each country.[4] This method is typically applied to linear models designed for causal inference, belonging to the class of best linear unbiased predictors (Baltagi, 2008). By focusing on minimizing bias, this approach is at one extreme of the bias-variance tradeoff and does not maximize out-of-sample forecast accuracy. More recent pooling methods, such as those based on Bayesian Model Averaging or forecast combinations (Timmermann, 2006; Wang et al., 2019) weight forecasts from different sets of countries in the panel. These methods achieve relatively high forecast accuracy but are computationally expensive, even for modestly large panels.

Our approach differs from conventional methods and leverages the benefits of machine learning techniques. Unlike conventional linear panel methods, our method can be applied to both linear and non-linear models. In addition, through cross-validation techniques, our method is designed to maximize out-of-sample, rather than in-sample, prediction accuracy (Annex I). Our algorithm can also be applied to both small and large panels due to its computational efficiency, and to non-linear models.[5]

By using machine learning methods, our approach is specifically designed to optimize the bias-variance tradeoff to improve forecasting results. Our method for selecting the optimal set of countries for pooling consists of two steps. The first step is to determine which countries contribute the least bias to the panel, which requires constructing some 'proximity' measure of the DGPs of different countries. The second step is to determine the optimal number of countries to include in the panel, for which the expected out-of-sample forecast error is minimized.

In the first step of our algorithm, we infer the 'proximity' of two countries by measuring the extent to which the DGP from one country is similar to that in the other country. Specifically, we use data from each individual country to predict real output growth in that same country using the machine learning model Random Forest (Annex II), yielding an estimate of the DGP for each country. We then compare how well the estimated DGP for each country predicts real output growth in the country of interest, using data only from the country of interest.[6] We then rank the countries accordingly to the predictive power of their DGPs. The intuition behind this method is simple: countries with similar economic structures should respond similarly to economic shocks. As a result, countries with similar DGPs are less likely to introduce bias to the pool.[7]

---

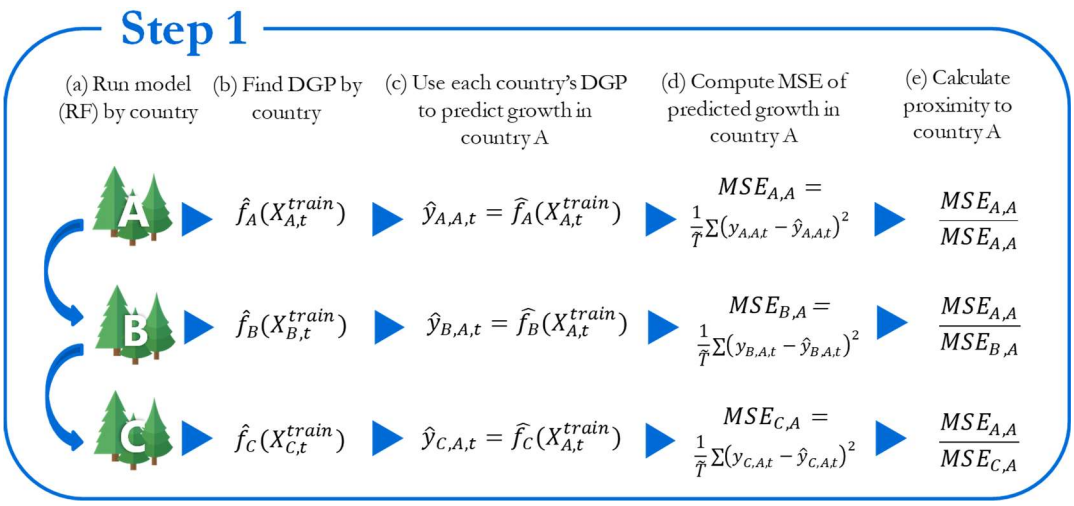
[4] Specifically, this literature estimates a linear model of the form $y_i = Z_i\delta_i{}' + u_i$ where y_i is $(T \times 1)$, $Z_i = [\iota_T, X_i]$, X_i is $(T \times K)$, $\delta_i' = (\alpha_i, \beta_i')$, and u_i is $(T \times 1)$ (Baltagi, 2008). The null hypothesis of homogeneity ($\delta_i = \delta$) is testable with a Chow F-test.

[5] For example, a naïve algorithm that sifts through all potential sets of 100 countries would need to consider more than nonillion (10^{30}) possibilities. This stands in stark contrast with our approach in which the algorithm would only consider 100 different sets in each of the two steps.

[6] In doing so, we follow the principle from the Bayesian literature that "(…) *if the model fits, then replicated data should look similar to the observed data*" (Gelman et al., 2014, p. 143). For a recent theoretical treatment of evaluating model fit by assessing its capability of generating predictions close to the observed data, see Svensson et al. (2018).

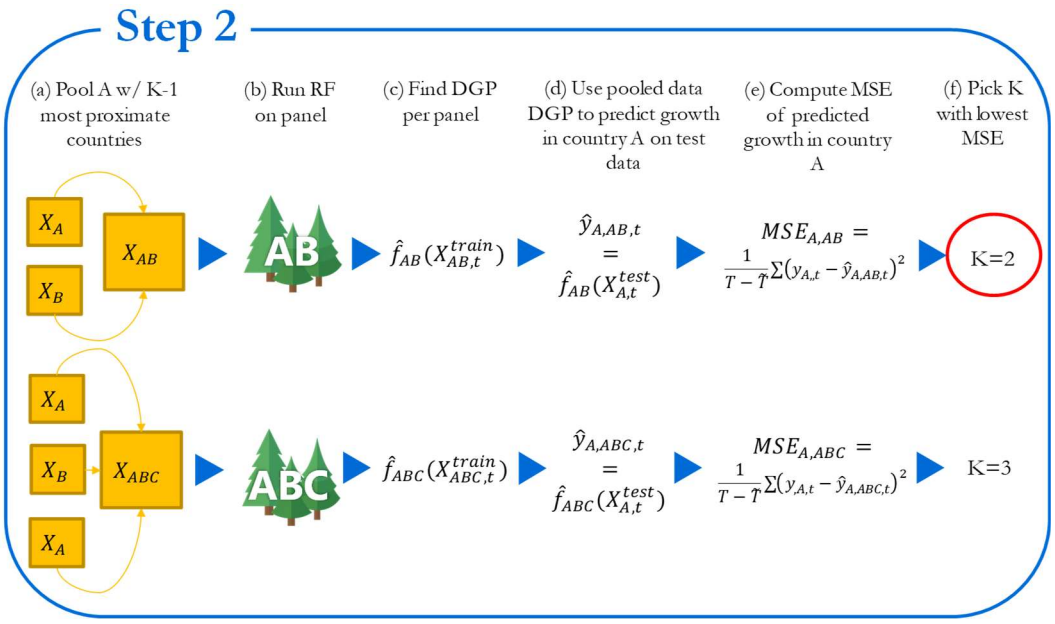
[7] Note that when the underlying DGP is linear, it is relatively straight-forward to express the similarity of countries' economic structures. In this case, two countries with similar estimates of the linear coefficients tend to respond similarly to

Figure 1. Step 1—Proximity



In the second step, we determine the optimal set of countries to pool by selecting the countries that maximize out-of-sample forecast accuracy. To identify the optimal set of countries, we cross validate each possible combination of pooled countries again using Random Forest and select the set of countries that minimizes the out-of-sample mean squared errors of the forecasts. The bias-variance tradeoff implies that, in most cases, neither a model with only country-specific data, nor a model with data from a very large number of countries is likely to deliver the highest forecast accuracy. Rather, the optimal set of countries for pooling is likely to lie somewhere in between these two extremes.

Figure 2. Step 2—Optimal Pool



shocks. This intuition does not generalize to non-linear methods. In contrast, our method can be applied using any underlying type of linear or non-linear forecasting model (e.g., dynamic factor model, Random Forest, neural networks).

III. Applying the Method

We apply our method to a range of countries. Specifically, we use the algorithm to find the optimal pooling set for each of nine countries, chosen to represent different stages of development. For these countries, we nowcast quarterly real output growth with a panel of (up to) 102 countries for the period 1987–2018. The nine example countries are Austria, Canada, Costa Rica, El Salvador, Germany, Lithuania, Mexico, Iceland and Turkey. For details on indicator selection, variable transformations, and missing value imputation, see Annex III.

In determining proximity, the algorithm selects countries that are similar in terms of economic structure. Figures 3 and 4 plot the most and least proximate countries for Turkey. In this case, the most proximate countries include emerging market countries with a substantial manufacturing base and that have experienced relatively volatile growth during the sample period, such as Mexico, Malaysia, Paraguay, Thailand, and Brazil. The algorithm also selects advanced economies with good data coverage, such as Japan and Finland. Tables 3-5 in Annex IV summarize the most and least proximate countries for each of the other eight example countries. In the case of Austria and Germany, the algorithm mainly selects advanced European economies. For Lithuania, it relies mostly on neighboring Central and Eastern European countries.

Figure 3. Turkey: Most Proximate Countries
(Relative inverse RMSE)

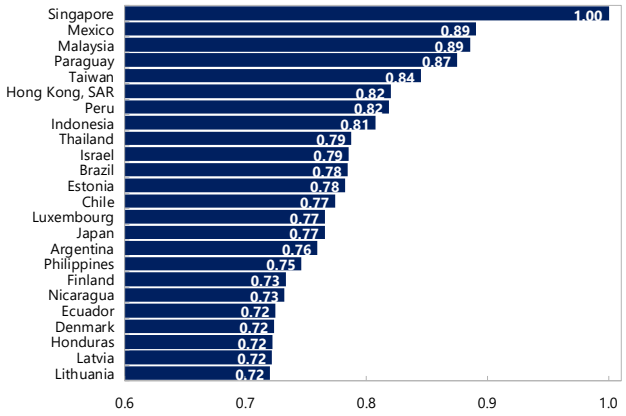
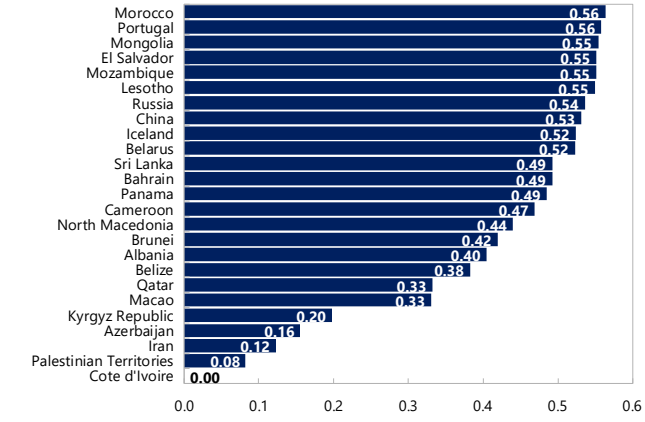


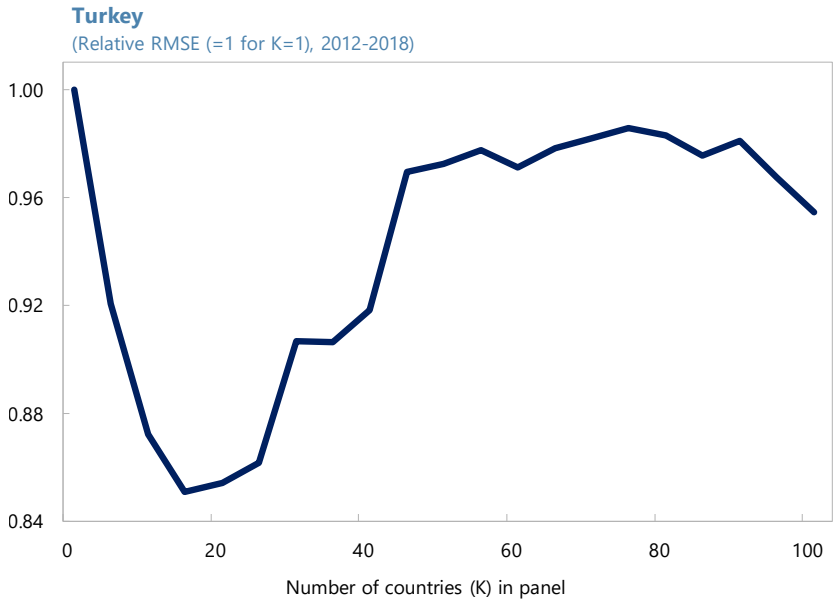
Figure 4. Turkey: Least Proximate Countries
(Relative inverse RMSE)



Our method significantly reduces forecast errors relative to alternative pools. The algorithm produces more accurate forecasts for eight of the nine example countries. In the remaining case (Iceland), the algorithm indicates that the best forecasts are achieved using country-specific data only. Figure 5 plots the out-of-sample root-mean-squared error (RMSE) of the algorithm's predictions across all possible number of countries (K) included in the pool for Turkey. In this case, the forecast errors initially fall substantially as more countries are added to the pool, with a minimum RMSE found using a pool of around 20 countries, where the algorithm lowers forecast errors by about 15 percent. After this point, however, adding more countries increases the forecast error. In line with the

bias-variance tradeoff, this U- shaped pattern suggests that the lower variance from more countries outweighs the additional bias only up to a certain point.

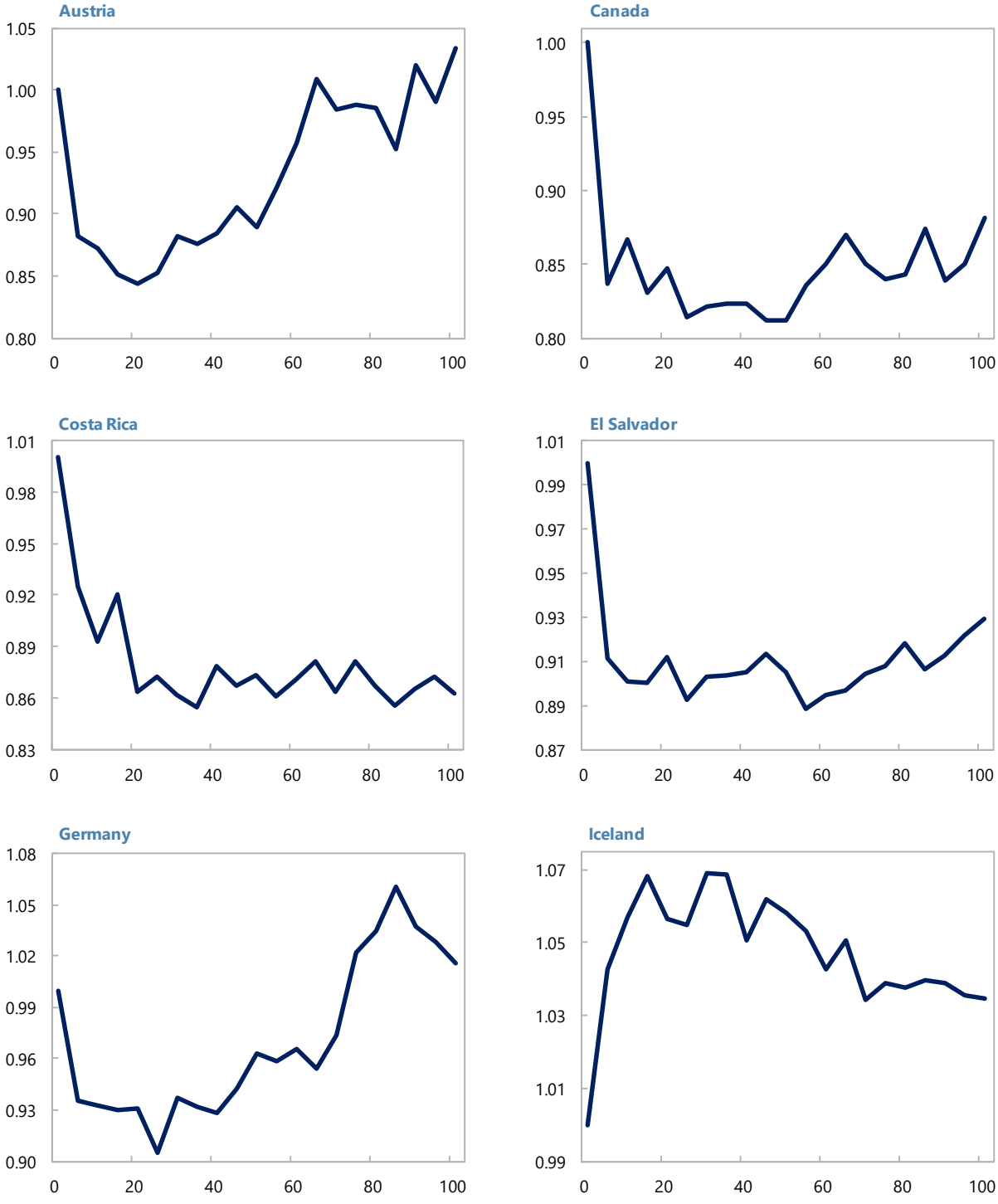
Figure 5. Turkey—Relative Forecast Errors for Different Pools

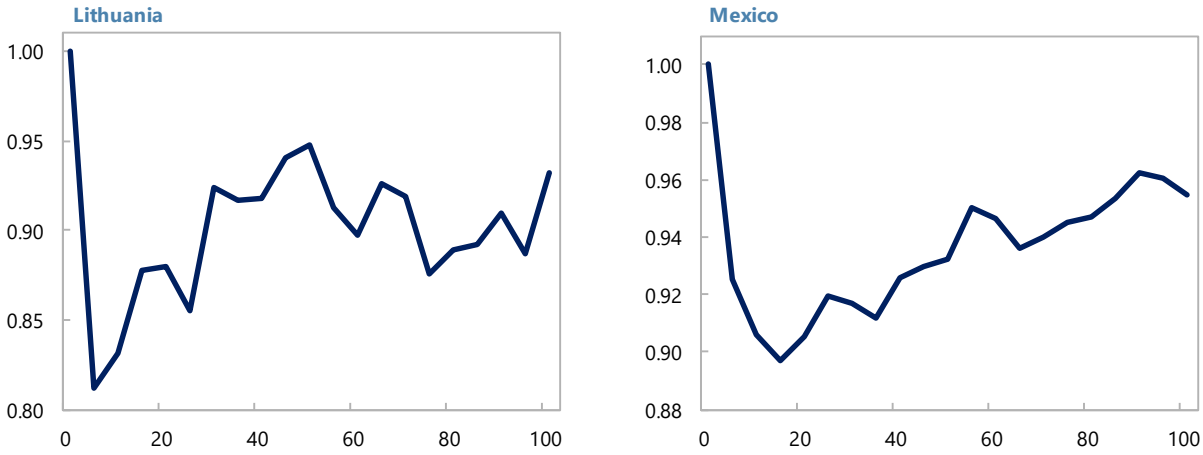


We observe a similar pattern for other countries. In particular, for Austria, Canada, Costa Rica, El Salvador, Germany, and Mexico (Figure 6), out-of-sample accuracy is maximized with a panel of about 20 to 60 countries. For Austria and Germany, adding too many countries to the pool increases forecast errors even beyond the RMSE from using only country-specific data. In the case of Lithuania, the algorithm selects only 5 to 10 similar countries. Overall, the algorithm reduces forecast errors by 10 to 20 percent relative to the benchmarks of one country or the full panel.

One exception to this pattern is Iceland. In this case, any panel of more than one country (i.e., any additional country beyond Iceland itself) has higher forecast errors than simply using data from Iceland alone. This difference is perhaps unsurprising given the unique nature of the Icelandic economy, (e.g., its reliance on fish and aluminum exports, as well as its vulnerable exposure to a few large domestic companies) relative to its two most proximate countries, Lithuania and Croatia.

Figure 6. Other Countries—Relative Forecast Error of Different Pools
(Relative RMSE (=1 for K=1), 2012–2018; Number of countries (K) in panel on x-axis)





IV. Conclusions

Pooling data from multiple countries can improve the performance of economic models. But there is a tradeoff between bias and variance in expanding a dataset to include data from additional countries. Adding countries with a similar economic structure may reduce variance and improve forecast results but adding countries with a dissimilar economic structure may introduce bias to the forecasts.

We use insights from machine learning to provide a method for selecting the optimal set of countries for macro panel forecasting. Our algorithm first infers the similarity of economic structures between countries to gauge bias and then uses cross validation techniques to optimize the bias-variance tradeoff and determine the optimal set of countries to include in the pool. We are able to substantially reduce forecast errors when using our pooling method to nowcast real output growth.

V. REFERENCES

Bai, J., & Ng, S., 2008. "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, *146*(2), 304–317.

Baltagi, B. H., 2008. "Forecasting with Panel Data," *Journal of forecasting*, *27*(2), 153–173.

Baltagi, B. H., & Griffin, J. M., 1997. "Pooled Estimators vs. Their Heterogeneous Counterparts in the Context of Dynamic Demand for Gasoline," *Journal of Econometrics*, *77*(2), 303–327.

Baltagi, B. H., Griffin, J. M., & Xiong, W., 2000. "To Pool or Not to Pool: Homogeneous Versus Heterogeneous Estimators Applied to Cigarette Demand," *Review of Economics and Statistics*, *82*(1), 117–126.

Bolhuis, M. A., & Rayner, B., forthcoming. "*Deus ex Machina? A Framework for Macro Forecasting with Machine Learning*," Mimeo.

Chen, S., & Ranciere, R., 2019. "Financial Information and Macroeconomic Forecasts," *International Journal of Forecasting*, *35*(3), 1160–1174.

Gavin, W. T., & Theodorou, A. T., 2005. "A Common Model Approach to Macroeconomics: Using Panel Data to Reduce Sampling Error," *Journal of Forecasting*, *24*(3), 203–219.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B., 2013. "*Bayesian Data Analysis*," Chapman and Hall/CRC.

Hoogstrate, A. J., Palm, F. C., & 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.

Jung, J. K., Patnam, M., & Ter-Martirosyan, A., 2018. "An Algorithmic Crystal Ball: Forecasts Based on Machine Learning," IMF Working Paper 18/230.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E., 2019. "Forecasting Inflation in a Data-rich Environment: the Benefits of Machine Learning Methods," *Journal of Business & Economic Statistics*, 1–45.

Richardson, A., & Mulder, T., 2018. "Nowcasting New Zealand GDP Using Machine Learning Algorithms," Mimeo.

Smalter Hall, A., 2018. "Machine Learning Approaches to Macroeconomic Forecasting," *Economic Review-Federal Reserve Bank of Kansas City*, *103*(4), 63.

Stock, J. H., & Watson, M. W., 2006. Forecasting with many predictors. *Handbook of Economic Forecasting*, *1*, 515–554.

Svensson, A., Zachariah, D., & Schön, T. B., 2018. "How Consistent Is My Model with the Data? Information-Theoretic Model Check," In *18th IFAC Symposium on System Identification SYSID 2018: Stockholm, Sweden, 9–11 July 2018* pp. 407–412. IFAC Papers Online.

Tiffin, A., 2016. "Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon," IMF Working Paper 16/56.

Timmermann, A., 2006. "Forecast Combinations," *Handbook of Economic Forecasting*, *1*, 135–196.

Wang, W., Zhang, X., & Paap, R., 2006. "To Pool or Not to Pool: What Is a Good Strategy for Parameter Estimation and Forecasting in Panel Regressions?," *Journal of Applied Econometrics, 34*, 724–745.

ANNEX I. MACHINE LEARNING AND CROSS VALIDATION

At its core, machine learning is about finding the optimal degree of complexity of a model that maximizes out-of-sample forecast accuracy. More complex forecasting models tend to exhibit lower bias as they are better at capturing nuances in how predictors affect the forecast variable. However, complex models are also more likely to capture perturbations (or 'noise') in the historical data that are uninformative for future predictions. This tendency, known as 'overfitting', increases the variance of forecasts, potentially resulting in higher forecast errors.

Machine learning models typically use cross validation (CV) to find the optimal model parameters. To avoid overfitting, machine learning algorithms assess performance of a particular model configuration by predicting on a new (test) data set. With CV, the entire data set is split into multiple subgroups, which are all used as separate test sets. The type of CV used in this paper is *holdout validation*, which uses only one subgroup. This type is often used in forecast settings to assess how a model 'would have done in the future'.

Any machine learning algorithm—including the one proposed in this paper—can thus be cast as a series of steps (Bolhuis & Rayner, forthcoming):

(a) Given a degree of model complexity, find the model configuration that maximizes forecast accuracy on the training data.

(b) Forecast on the test data using this model configuration.

(c) Across all, pick the degree of model complexity that maximizes forecast accuracy on the test data.

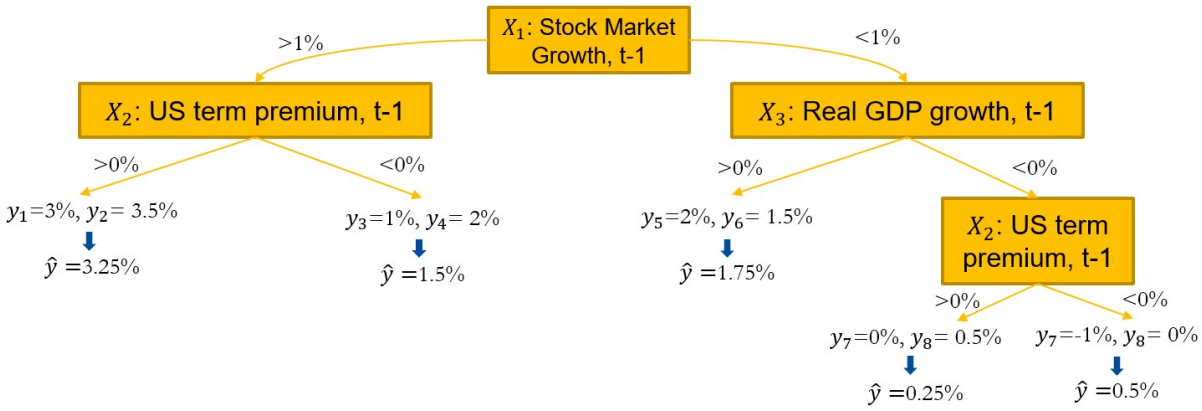
Note that in this paper, the degree of model complexity is the number of countries to pool (K) and the model configuration that minimizes forecast errors for the training data is a Random Forest run on the panel of K countries most proximate to the one under consideration.

ANNEX II. RANDOM FOREST

Random Forest (RF) is a machine learning algorithm that uses forecast combinations of multiple decision trees to construct an aggregate forecast. RF is one of the most popular algorithms available, because it is computationally efficient and requires almost no tuning of model parameters. This second advantage makes it an ideal algorithm for forecasting on time-series data with relatively few observations.

A decision tree is an algorithm that partitions the set of predictor combinations into regions, making a point forecast for each of the regions. A tree creates the partition by splitting the training data using recurrent yes/no questions (Figure A.1). This feature makes decision trees an attractive prediction tool because they translate nonlinear prediction problems into easy-to-understand steps.

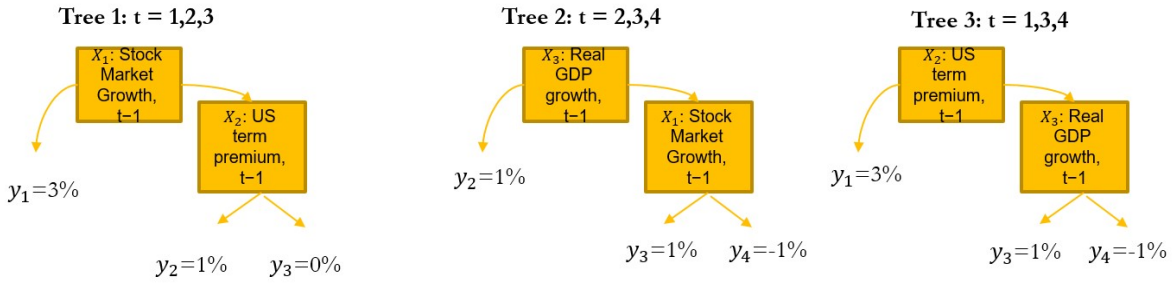
Figure A.1. Decision Tree Example



Notes: Figure plots a hypothetical decision tree nowcasting real GDP growth at time *t* using lags of real GDP growth, stock market growth, and the US term premium. Each leaf contains two training observations, and the trained decision tree predicts the average observed GDP growth of these two observations.

RF maximizes the information content of the training data by using subsamples of observations and predictors. While each individual decision tree in a RF tends to have low forecast accuracy, the aggregate predictions of RFs tend to be surprisingly good. The reason for discrepancy is that each decision tree is only a partial reflection of the information in the training data. RF modifies the decision tree approach in two ways. First, it uses *bootstrap aggregation* ('bagging') by building each individual tree on only a random sample of the observations in the training data. Second, at each split in the tree, they use only a random subsample of the predictors.

Figure A.2. Random Forest Example



Notes: Figure plots a hypothetical decision Random Forest nowcasting real GDP growth at time *t* using lags of real GDP growth, stock market growth, and the US term premium. Each tree uses different observations and considers different variables at each split. In this example, each leaf contains only one training observation. The trained RF predicts the average of the GDP growth rates of the leaves that the new observation belongs to.

ANNEX III. DATA

- **Our framework uses mixed frequency (monthly and quarterly) standardized (across countries) leading and coincident indicator data from Haver Analytics.** As the machine learning method (Random Forest) sifts through a broad range of potential predictors and tends to select the most predictive ones, we do not need to specify the ultimate set of predictors in advance. We thus collect as many country-specific and global indicators as possible. Tables 1 and 2 contain the set of indicators in the panel.[8]

- **We transform each indicator twice, deflate where necessary and include 1- through 12-month lags.** We use two types of transformations. In the case of stationary variables (e.g., capacity utilization, consumer confidence), we use the level and quarter-on-quarter difference. For non-stationary variables (e.g., production, money) we take first- and second-order log differences.

- **To use all available indicators and observations, we impute missing values using K-Nearest Neighbor proximity measures.** This algorithm finds the 10 ('*K*') other observations ('neighbors') in the dataset that are most similar ('nearest') using a Euclidian metric. It then imputes the missing values using the median value from these neighbors.

- **We pre-select indicators by using 'hard thresholding' (Bai & Ng, 2008).** For each indicator, we run regress the forecast variable on its lags and the indicator. We then select all indicators with an absolute t-statistic above 2.5.

Table 1. Stationary Variables—Level and First Difference

Official/policy interest rate	Long term sovereign bond yield
Gross External Debt, percent of GDP	Capacity Utilization
LFS Unemployment Rate	Current Account, percent of GDP
BoP, percent of GDP	Now-Casting Index (NCI)
World Uncertainty Index	Composite PMI
Manufacturing PMI	Composite PMI, flash
Composite PMI, manufacturing, flash	Sovereign CDS Spread
JPM Global Composite PMI	JPM Global Manufacturing PMI
US Corporate High Yield	US Federal Funds Effective Rate
US 10-year Treasury Yield	World Uncertainty Index
Sentix Economic Expectations	Sentix Current Economic Situation
CBOE VIX	CBOE 10-year Treasury VIX
Real short rate	Real long rate
Term spread (long rate—short rate)	Long dollar spread (long rate—US 10 year yield)
Real M2 growth	Real US Federal Funds Rate
Real US 10 year	Real US term spread (10 year yield—FFR)
US credit spread (high yield—10 year yield)	

[8] We also construct the sovereign term spread, sovereign yield spread, the US sovereign term spread, and the US high yield spread.

Table 2. Non-Stationary Variables—First and Second Log Difference
(Y: in real terms)

Nominal ER against USD	Nominal ER against EUR
Nominal Effective ER (38 partners)	Real Effective ER (38 partners)
M2	Private sector debt (Y)
Domestic debt (Y)	MSCI Stock Market, total return, USD
Gross ED, nominal, USD	CPI
Core CPI	PPI
EPI	IPI
ToT	Housing Prices (Y)
Housing Permits	Industrial Production, total (Y)
Industrial Production, manufacturing (Y)	Employment
Earnings (Y)	Manufacturing Shipments (Y)
Retail Sales, value	Retail Sales, volume
Vehicle Registrations	Exports, value (Y)
Imports, value (Y)	Consumer Confidence
Consumer Expectations	Business Confidence
Tourist Arrivals	Gross Operating Surplus or Corporate Profits (Y)
Nominal Final Domestic Demand	Composite CPI for Advanced Economies
HWWI Commodity Price Index	Dallas Fed House Price Index World
World Industrial Production ex Construction	CPB World Trade Volume
Dow Jones Global Index World	WTI Weekly Average Price
BIS Narrow NEER Dollar	AUD/JPY ER

ANNEX IV. PROXIMATE COUNTRIES

Table 3. Most/Least Similar Countries—Austria, Canada, Costa Rica

Austria	Canada	Costa Rica
[1] Austria	[1] Canada	[1] Costa Rica
[2] Belgium	[2] United States	[2] Japan
[3] France	[3] France	[3] Nicaragua
[4] Switzerland	[4] Denmark	[4] Guatemala
[5] Germany	[5] United Kingdom	[5] Malaysia
[6] Denmark	[6] Sweden	[6] Colombia
[7] Italy	[7] South Africa	[7] Honduras
[8] United States	[8] Austria	[8] Poland
[9] Netherlands	[9] Finland	[9] Hong Kong
[10] Canada	[10] Switzerland	[10] Australia
[11] Sweden	[11] Belgium	[11] Taiwan
[12] United Kingdom	[12] Hungary	[12] Chile
[13] Portugal	[13] Netherlands	[13] Tunisia
[14] South Africa	[14] Spain	[14] Thailand
[15] Japan	[15] Mexico	[15] Bolivia
[16] Finland	[16] Germany	[16] Latvia
[17] Spain	[17] Australia	[17] Philippines
[18] Hungary	[18] Italy	[18] Indonesia
[19] Greece	[19] Greece	[19] Zambia
[20] Croatia	[20] Portugal	[20] Cyprus
[21] Czech Republic	[21] Czech Republic	[21] United States
[22] Brazil	[22] New Zealand	[22] Malta
[23] Norway	[23] Japan	[23] New Zealand
[24] Russia	[24] Cyprus	[24] Israel
[25] Mexico	[25] Brazil	[25] Jordan
[78] Ghana	[78] North Macedonia	[78] Russia
[79] Botswana	[79] Vietnam	[79] Kazakhstan
[80] Kenya	[80] Botswana	[80] Belarus
[81] Cameroon	[81] Kenya	[81] Ireland
[82] Belize	[82] Tanzania	[82] Iceland
[83] Georgia	[83] Senegal	[83] Ukraine
[84] India	[84] Georgia	[84] Albania
[85] Ireland	[85] Cameroon	[85] Namibia
[86] North Macedonia	[86] India	[86] Panama
[87] Kazakhstan	[87] Namibia	[87] Serbia
[88] Sri Lanka	[88] Sri Lanka	[88] Cameroon
[89] Dominican Republic	[89] Kazakhstan	[89] Botswana
[90] Namibia	[90] Belize	[90] China
[91] Mozambique	[91] Dominican Republic	[91] Mongolia
[92] Panama	[92] Mozambique	[92] El Salvador
[93] Mongolia	[93] Mongolia	[93] North Macedonia
[94] China	[94] Panama	[94] Brunei
[95] Brunei	[95] China	[95] Belize
[96] Qatar	[96] Iran	[96] Qatar
[97] Iran	[97] Cote d'Ivoire	[97] Kyrgyz Republic
[98] Macao	[98] Qatar	[98] Iran
[99] Cote d'Ivoire	[99] Macao	[99] Azerbaijan
[100] Azerbaijan	[100] Kyrgyz Republic	[100] Macao
[101] Kyrgyz Republic	[101] Palestinian Territories	[101] Palestinian Territories
[102] Palestinian Territories	[102] Azerbaijan	[102] Cote d'Ivoire

Notes: Table presents top and bottom 25 countries that are most predictive of growth in Austria, Canada and Costa Rica. Data up to and including 2017, based on Random Forests with 500 trees.

Table 4. Most/Least Similar Countries—El Salvador, Germany, Iceland

El Salvador	Germany	Iceland
[1] El Salvador	[1] Germany	[1] Iceland
[2] Brazil	[2] Japan	[2] Lithuania
[3] Morocco	[3] Netherlands	[3] Croatia
[4] Mexico	[4] France	[4] Finland
[5] Norway	[5] Denmark	[5] Romania
[6] United States	[6] Austria	[6] Serbia
[7] New Zealand	[7] Italy	[7] Netherlands
[8] Australia	[8] Switzerland	[8] Guatemala
[9] Canada	[9] Sweden	[9] Japan
[10] Guatemala	[10] Belgium	[10] Nicaragua
[11] Taiwan	[11] Mexico	[11] Morocco
[12] Belgium	[12] Finland	[12] Hong Kong, SAR
[13] Honduras	[13] Hungary	[13] Costa Rica
[14] France	[14] Canada	[14] Honduras
[15] Denmark	[15] United States	[15] El Salvador
[16] Nicaragua	[16] Portugal	[16] New Zealand
[17] Netherlands	[17] Czech Republic	[17] Sweden
[18] Tunisia	[18] Greece	[18] Philippines
[19] Japan	[19] Spain	[19] Cyprus
[20] Austria	[20] Croatia	[20] Greece
[21] Israel	[21] South Africa	[21] Poland
[22] Colombia	[22] United Kingdom	[22] Latvia
[23] South Africa	[23] Brazil	[23] Czech Republic
[24] Zambia	[24] Cyprus	[24] Hungary
[25] Sweden	[25] Norway	[25] Estonia
[78] Belize	[78] Ghana	[78] Panama
[79] Brunei	[79] Botswana	[79] Senegal
[80] Tanzania	[80] Vietnam	[80] Sri Lanka
[81] Mozambique	[81] Kenya	[81] Lesotho
[82] Ireland	[82] Senegal	[82] Belarus
[83] India	[83] Tanzania	[83] Kazakhstan
[84] Cameroon	[84] Georgia	[84] Cameroon
[85] North Macedonia	[85] Namibia	[85] Tanzania
[86] Belarus	[86] India	[86] Paraguay
[87] Dominican Republic	[87] Cameroon	[87] Albania
[88] Serbia	[88] Kazakhstan	[88] Ireland
[89] Sri Lanka	[89] Sri Lanka	[89] Ghana
[90] Ghana	[90] Belize	[90] Brunei
[91] Kazakhstan	[91] Dominican Republic	[91] Namibia
[92] Botswana	[92] Mozambique	[92] Belize
[93] Mongolia	[93] Mongolia	[93] China
[94] Panama	[94] Panama	[94] Mongolia
[95] China	[95] China	[95] North Macedonia
[96] Iran	[96] Iran	[96] Qatar
[97] Qatar	[97] Cote d'Ivoire	[97] Iran
[98] Kyrgyz Republic	[98] Qatar	[98] Macao
[99] Azerbaijan	[99] Macao	[99] Cote d'Ivoire
[100] Macao	[100] Palestinian Territories	[100] Azerbaijan
[101] Palestinian Territories	[101] Kyrgyz Republic	[101] Kyrgyz Republic
[102] Cote d'Ivoire	[102] Azerbaijan	[102] Palestinian Territories

Notes: Table presents top and bottom 25 countries that are most predictive of growth in El Salvador, Germany and Iceland. Data up to and including 2017, based on Random Forests with 500 trees.

Table 5. Most/Least Similar Countries—Lithuania and Mexico

Lithuania	Mexico
[1] Lithuania	[1] Mexico
[2] Finland	[2] Germany
[3] Estonia	[3] Japan
[4] Bulgaria	[4] Canada
[5] Japan	[5] Hungary
[6] Romania	[6] United States
[7] Mexico	[7] Netherlands
[8] Germany	[8] Sweden
[9] Latvia	[9] Denmark
[10] Costa Rica	[10] Czech Republic
[11] Malaysia	[11] France
[12] Honduras	[12] Finland
[13] Croatia	[13] Hong Kong, SAR
[14] Argentina	[14] Belgium
[15] Greece	[15] Honduras
[16] Iceland	[16] Switzerland
[17] Hungary	[17] Guatemala
[18] Turkey	[18] Norway
[19] Singapore	[19] Austria
[20] Nicaragua	[20] Estonia
[21] Luxembourg	[21] Malaysia
[22] Czech Republic	[22] Tunisia
[23] Sweden	[23] Greece
[24] Netherlands	[24] South Africa
[25] Ecuador	[25] Italy
[78] Mozambique	[78] Bahrain
[79] Russia	[79] Dominican Republic
[80] El Salvador	[80] Senegal
[81] Ukraine	[81] El Salvador
[82] Ghana	[82] Mozambique
[83] Belarus	[83] Tanzania
[84] Ireland	[84] Ghana
[85] Lesotho	[85] Botswana
[86] Bahrain	[86] Brunei
[87] Sri Lanka	[87] India
[88] Mongolia	[88] Sri Lanka
[89] Cameroon	[89] North Macedonia
[90] Namibia	[90] Kazakhstan
[91] China	[91] Belize
[92] Brunei	[92] Cameroon
[93] Albania	[93] Mongolia
[94] North Macedonia	[94] Panama
[95] Belize	[95] China
[96] Qatar	[96] Iran
[97] Kyrgyz Republic	[97] Cote d'Ivoire
[98] Azerbaijan	[98] Macao
[99] Iran	[99] Qatar
[100] Macao	[100] Kyrgyz Republic
[101] Palestinian Territories	[101] Palestinian Territories
[102] Cote d'Ivoire	[102] Azerbaijan

Notes: Table presents top and bottom 25 countries that are most predictive of growth in Lithuania and Mexico. Data up to and including 2017, based on Random Forests with 500 trees.