Establishing Prediction Benchmarks of Fertility Outcomes for 12 European Countries

Keywords: benchmarking, fertility, machine learning, predictability of life outcomes, prediction in social sciences

Extended Abstract

Fine-grained fertility predictions are essential for planning and efficient allocating of resources in many areas. However, current local population projections often suffer from considerably larger forecast errors than those for countries and larger subnational regions [1]. That calls for improvements in the methods of small area population forecasts [2].

One of the possible ways to improve fertility forecasting is to use machine learning methods to predict individual-level fertility outcomes. There is extensive research on the mechanisms behind fertility behavior that taught us a great deal about various factors significantly associated with the number and timing of children [3]. This theoretical knowledge holds a huge potential for accurate predictions of fertility based on machine learning models. Yet, the actual ability of these factors to predict new cases is not clear as social sciences traditionally focus on causal explanation rather than measuring predictive performance [4].

In this paper, we utilize a more data-scientific approach and measure the ability of well-established variables in fertility research to predict fertility outcomes out-of-sample. To the best of our knowledge, this paper is the first systematic attempt to quantify the out-of-sample predictive performance of these variables.

We use two waves for 12 European countries (N = 83 244) of the Generations and Gender Survey (GGS). It is the largest currently available longitudinal multinational dataset that includes information about the number of children and factors associated with fertility decisions. Using several machine learning methods we predict the total number of children for participants older than 45 years and having a new child by Wave 2 for participants aged 18-45 years. As predictors, we use gender, age, education, income, religion, current partnership status, satisfaction with the relationship, the total number of marriages (only for predicting the number of children), and fertility intentions and age of the youngest child (only for predicting having a new child), all measured in Wave 1. We also analyze how prediction models generalize across countries by training a predictive model on each country's data and then testing the model on each of the other countries.

Although almost all the factors are highly significant, the predictions are not very accurate. The predictions of the *number of children* are off by an average of 1.15 children (ranging from 0.84 to 1.44). On average, only 9% of the variation in the total number of children is explained; out-of-sample R-squared measured using 10-fold cross-validation and a linear regression model varies from 0.2% for Germany to 20% for Georgia (Figure 1A). Non-linear models do not perform better, except for the Netherlands, where gradient boosting is able to explain 40% of the variation (RMSE = 1.25 vs 1.32 for a linear regression model).

The accuracy of predictions of the likelihood of *having a new child* varies substantially by country. In Bulgaria, Georgia, and the Czech Republic, where the proportion of people who had a new child is low (Figure 1F) the models are only able to correctly identify around 5% of positive cases; in Germany, France, and Austria – about 33%; in the Netherlands, Lithuania, Russia, Hungary, and Poland (where the proportion of participants who had a new child is around 40-50%) – 66-85% of the positive cases. In several countries, predictions are worse than those from a null model that simply uses the proportion of participants who had a child as an estimate for new observations (Figure 1B). Non-linear models do not improve predictions.

Regarding generalizability or the ability of a model to make accurate predictions across different countries, for the case of predicting the *number of children*, the model's performance is fairly consistent across test countries (Figure 1C). The test R-squared is almost the same when the model is trained and tested on the same country as when it is tested on other countries, with the exception of Russia, Georgia, the Netherlands, and Lithuania. In most countries, the standard deviation of the test R-squared ranges from 0.02 to 0.08, while in the mentioned countries, it ranges from 0.11 to 0.18. This could indicate differences in fertility patterns. For instance, predictions for other countries based on data from Georgia may be less accurate due to differences in the average number of children and regression coefficients. Georgia has one of the highest numbers of children in the sample, and also the highest regression coefficient for the number of marriages. However, Georgia can accurately predict Poland and Sweden due to their relatively high regression coefficients for the number of marriages and the average number of children. In the case of predicting *having a new child*, prediction performance seems to reflect mostly the similarity of the proportion of participants who had a new child (Figure 1D).

Our findings indicate that the well-established variables have limited predictive ability for fertility outcomes. To improve predictions, further exploration of machine learning methods is needed. For example, testing another approach to variable selection, as significant variables may not always be good predictors [5], or using additional variables to fully utilize the potential of non-linear models that can identify more complex patterns of fertility. Addressing the issue of imbalanced classes can also improve results when predicting the likelihood of having a new child. The results provide a baseline for predicting fertility outcomes based on the GGS data, which can be used to evaluate the effectiveness of new approaches and facilitate further improvements in prediction models. With the improvement of predictive models for each country, cross-country forecast errors can potentially serve as an interesting integral indicator of the similarity of fertility patterns between countries.

References

1. Wilson, T., & Rowe, F. (2011). The forecast accuracy of local government area population projections: A case study of Queensland. *Australasian Journal of Regional Studies*, *17*(2), 204–243.
2. Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2022). Methods for Small Area Population Forecasts: State-of-the-Art and Research Needs. *Population Research and Policy Review*, *41*(3), 865–898.
3. Balbo, N., Billari, F. C., & Mills, M. (2012). Fertility in Advanced Societies: A Review of Research. *European Journal of Population*, *29*(1), 1–38.
4. Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.
5. Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, *112*(45), 13892–13897.
6. Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... & McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398-8403.

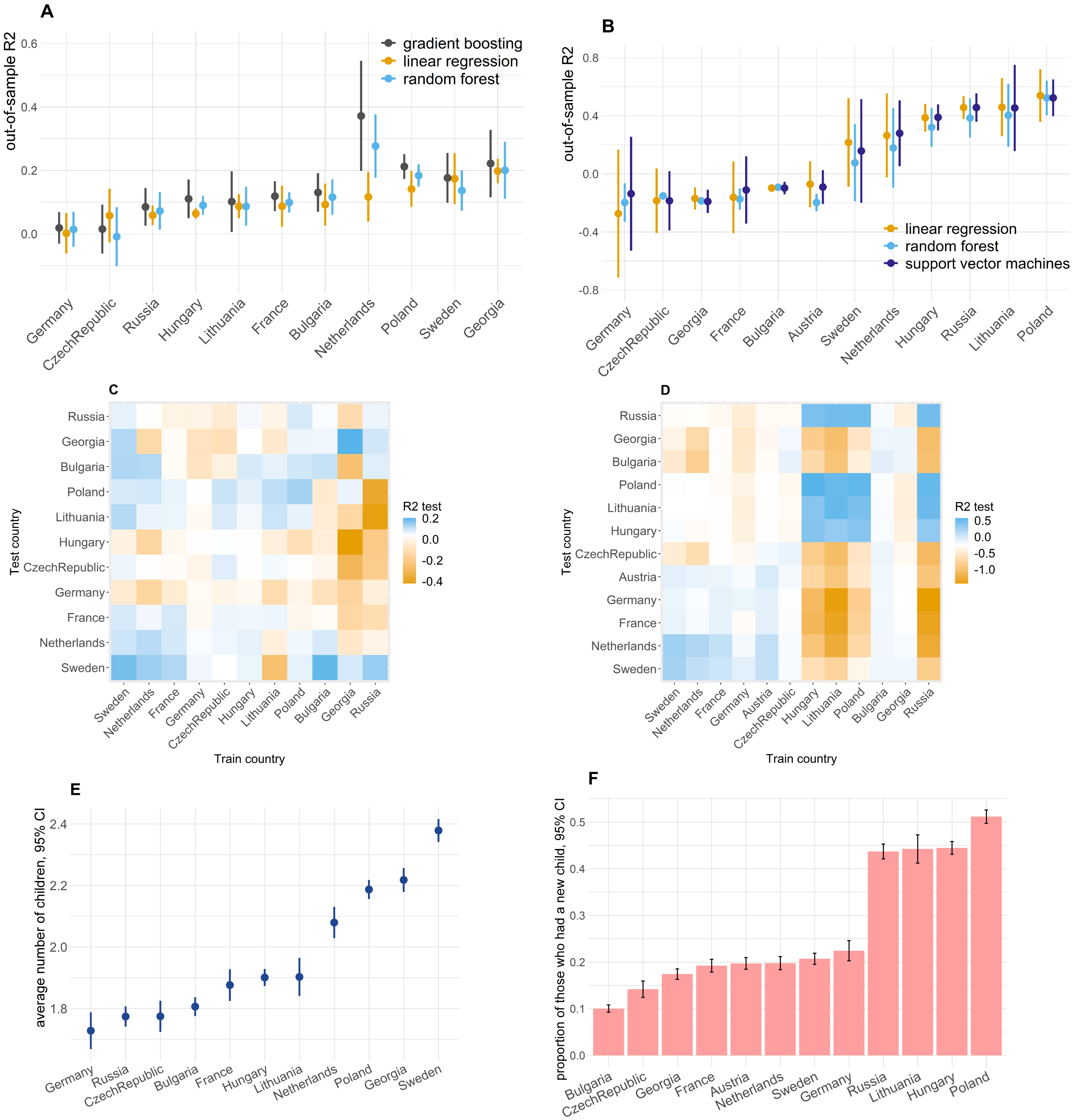

Figure 1. Comparison of the performance of different models for predicting **A)** the total number of children and **B)** having a new child. Austria is missing from (A) because the sample did not include people older than 45. Out-of-sample R^2 is measured using 10-fold cross-validation. We use formula (1) from [6] to measure how much the model adds in comparison with simply predicting the mean of the training sample (in this case, training folds). Error bars represent 95% confidence intervals, calculated based on the R^2 for each of the test folds. **C)** Cross-country predictions of the total number of children, and **D)** having a new child in the next three years. We use the same formula (1) to measure R^2 for predictions based on a train country for each test country. White color represents median R^2; blue – above the median, and orange – below the median. **E)** an average number of children, **F)** the proportion of participants who had a new child in three years (before Wave 2).

$$(1) \quad R^2_{test} = 1 - \frac{\sum_{i\in test}(y_i - \hat{y_i})^2}{\sum_{i\in test}(y_i - \overline{y}_{training})^2}$$