What factors determine uneven suburbanisation?
Explaining urban sprawl with machine learning

Honorata Bogusz*, Szymon Winnicki and Piotr Wójcik
University of Warsaw, Faculty of Economic Sciences
Data Science Lab

Background

- suburbanisation is a shift of population from central urban areas into suburbs, resulting in the formation of (sub)urban sprawl
- unrestricted city growth leads to a variety of **formidable consequences**
- in order to execute **adequate social planning**, it is useful to identify factors which push migrants out of the city and pull them to the suburban boroughs

Objectives

1. identify pulling features of boroughs which are key factors in predicting the number of migrants
2. find a machine learning algorithm that exhibits the most accurate predictive performance
3. uncover non-linear relationships between the number of migrants and the most important regressors

Data and Methods

- **Data**: 30 features of boroughs obtained from Polish Statistical Office, Open Street Map, Google Maps, National Electoral Commission, e-podroznik.pl and gratka.pl (most recent observations for year 2018 or 2019)
- **Model benchmark:** we base our 7 predictive models on the *gravity model of migration* framework
- **Models used:** Ordinary Least Squares, Lasso, Ridge Regression, Elastic Net, Support Vector Regression, Random Forest, Extreme Gradient Boosting
- **EDA**: PCA with varimax rotation as a method of choosing variables out of correlated groups for algorithms incapable of variable selection
- **Errors**: assesing models' performance by common benchmarks of RMSE, MAE and R^2 on validation and training samples (LOOCV used)
- **XAI**: *Permutation-based feature importance* and *Accumulated Local Effects* plots

Conclusions

1a. identified important pulling factors (see ALE plots on the right): good amenities, mean relative income, progressiveness of a community, spiritual needs

1b. identified important pushing factors (see ALE plots on the right): distance and population density (the latter in contrast with the existing literature)

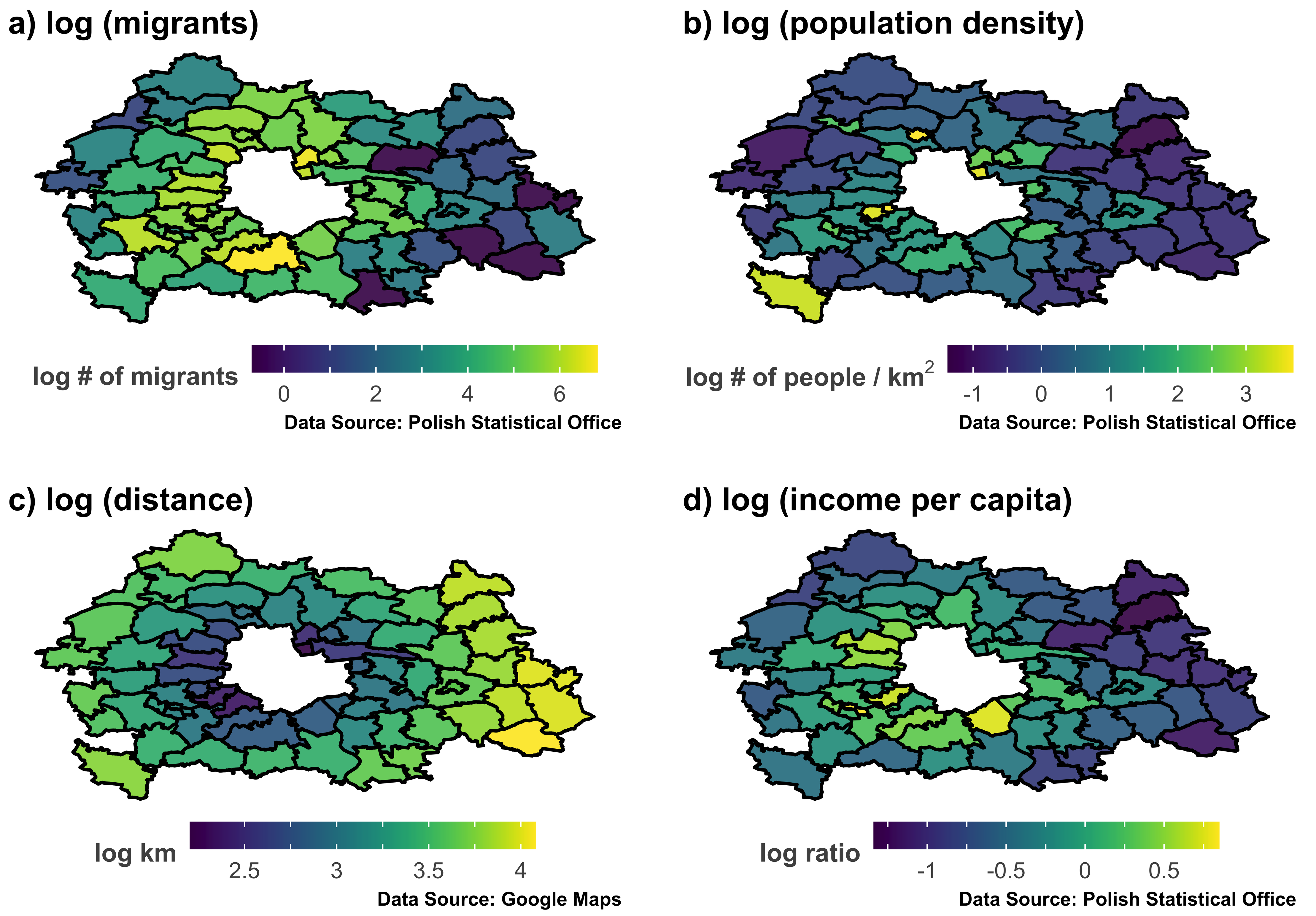
2. XGB outdoes OLS by only a slight margin
3. assuming linear relationships is infeasible with respect to some regressors - non-linear relationships can be interpreted with ALE plots

Contact information

***email** h.bogusz@uw.edu.pl
twitter HonorataBogusz

Empirical Verification

Maps of the key variables of the gravity model of migration, and income per capita (important pulling factor according to the literature) - logarithmed for clearance



Model errors (validation and training sample, calculated by LOOCV)

MODEL	VALIDATION RMSE	VALIDATION MAE	VALIDATION R2	TRAIN RMSE	TRAIN MAE	TRAIN R2
OLS	111.12	79.86	0.62	97.55	71.93	0.71
RIDGE	136.15	87.50	0.43	113.59	73.25	0.61
LASSO	124.63	83.92	0.53	97.99	66.81	0.71
SVR	119.15	78.33	0.57	65.63	34.45	0.87
RF	123.05	75.63	0.54	80.29	53.00	0.80
XGB	109.56	70.07	0.63	78.25	45.76	0.81
ELASTIC	124.63	83.92	0.53	97.99	66.81	0.71

ALE plots for 6 most important features as indicated by mean PFI (of all models)