University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 34/2020 (340)

WHAT FACTORS DETERMINE UNEQUAL SUBURBANISATION? NEW EVIDENCE FROM WARSAW, POLAND

Honorata Bogusz
Szymon Winnicki
Piotr Wójcik

Warsaw 2020

WORKING PAPERS 34/2020 (340)

What factors determine unequal suburbanisation? New evidence from Warsaw, Poland

Honorata Bogusz[1]*, Szymon Winnicki[2] and Piotr Wójcik[3]

[1] *Faculty of Economic Sciences, University of Warsaw and Labfam*
[2] *Faculty of Economic Sciences, University of Warsaw*
[3] *Faculty of Economic Sciences and Data Science Lab WNE UW, University of Warsaw,*
** Corresponding author: h.bogusz@uw.edu.pl*

Abstract: This article investigates the causes of spatially uneven migration from Warsaw to its suburban boroughs. The method is based on the gravity model of migration extended by additional measures of possible pulling factors. We report a novel approach to modelling suburbanisation: several linear and non-linear predictive models are estimated and explainable AI methods are used to interpret the shape of relationships between the dependent variable and the most important regressors. It is confirmed that migrants choose boroughs of better amenities and of smaller distance to Warsaw city center.

Keywords: suburbanisation, gravity model of migration, machine learning models, explainable artificial intelligence

JEL codes: R23, P25, C14, C51, C52

Acknowledgements: Research partly financed by Polish National Science Center, project no. 2016/21/B/HS4/00670

1. Introduction

Suburbanisation is a shift of population from central urban areas into suburbs, resulting in the formation of (sub)urban sprawl. As a consequence of the movement of households and businesses out of the city centers, low-density, peripheral urban areas grow (Caves, 2004). Most of the residents of metropolitan cities work within the central urban area, but choose to live in satellite communities called suburbs. These processes occur in more economically developed countries. The United States is the generic example: it is believed to be the first country, where the majority of the population lives in the suburbs, rather than in the cities or rural areas (Hobbs & Stoops, 2002). Urban sprawl, a direct consequence of suburbanisation, is an unrestricted growth in many urban areas of housing, commercial development, and roads over large expanses of land, with little concern for urban planning (Fouberg & Murphy, 2020). The negative impacts of urban sprawl include: increase in vehicle mileage, residential energy consumption and land use, degradation of air quality, as well as increased usage of natural resources and greenhouse gas emissions (Kahn, 2000); increased infrastructure costs (Downs, McCann, & Mukherji, 2005); decline of social capital, residential segregation resulting in class and racial divisions (Duany, Plater-Zyberk, & Speck, 2010); growing fiscal deficit (Downs et al., 2005); health deterioration as a result of increased vehicle usage (Sturm & Cohen, 2004).

During Communism, most socialist countries in the Eastern Bloc were characterized by under-urbanization, which meant that industrial growth occurred well in advance of urban growth and was sustained by rural-urban commuting (Murray & Szelenyi, 2009). City growth, residential mobility, land and housing development were under tight political control. Consequently, suburbanization in post-communist Europe is not only a recent but also a specific phenomenon. Warsaw is a particular example of such circumstances – 80-90% of the buildings were destroyed during the 1944 uprising (Tung, 2001) resulting in a renewed, communist city planning. Suburbanisation ”in Western sense” has been a recent phenomenon - it is believed to have begun in the post-socialist countries in the 90., after the political transformation (Lisowski, 2004; Nuissl & Rink, 2005; Tim´ar & Varadi, 2001). As of 2019, about 43% of the Warsaw metropolis inhabitants were living in the suburbs.

The causes of metropolitan suburbanisation have been heavily discussed in the literature and a few theories have been offered (Mieszkowski & Mills, 1993), mostly based on the situation in the Western cities. Some papers offering insight about suburbanisation processes in post-socialist cities are: Kok (1999); Lisowski (2004); Murray and Szelenyi (2009); Nuissl and Rink (2005). The quantitative measures of suburbanization determinants existing in the literature are scarce and include works by Jordan, Ross, and Usowski (1998); Kok (1999); Loibl (2004). Jordan et al. (1998) identified several pulling factors of the target suburban areas in 35 US metropoleis while measuring surburbanisation by the population density gradient (Brueckner, 1987). Loibl (2004) identified the attractiveness measures of Vienna suburban areas but used the number of migrants as the dependent variable. Kok (1999) used micro-level data to investigate the motifs of individuals to move out of the city to the suburbs in Budapest and Warsaw. In all of the three mentioned works, simple regression models were used (logit in case of Kok (1999)) and only a few possible pulling factors were included. Moreover, surprisingly, no application of the gravity model of migration (the most well-known quantitative migration model (Poot, Alimi, Cameron, & Mar´e, 2016)) in context of the suburban migration has been offered in the literature.

These three identified gaps point to further research potential of the topic of suburbanisation determinants. They are filled in this article making our approach innovative for several reasons. First of all, we use the gravity model of migration framework to predict the number of migrants choosing different suburban boroughs of Warsaw. On the other hand, we implement a much wider selection of possible pulling factors (30) than so far offered in the literature. When including that many regressors, it is reasonable to assume the relationships between the dependent variable and some of the regressors might be non-linear. In addition, one can expect interactions between various predictors. A wide selection of pulling factors and a possibility of non-linear relationships, interactions and collinearity call for the use of methods robust to such issues. Hence, we apply various predictive models that take into account potential non-linearities without any prior assumptions of their shape and are capable of dealing efficiently with a large number of potential predictors, also highly correlated. OLS is used as a simple benchmark. In addition, to explain the estimated machine learning models and unhide the identified shape of relationship, the novel approach of eXplainable Artificial Intelligence (XAI) is used.

Our aim is to identify the pulling factors of Warsaw suburban boroughs which contribute to the choice of one borough instead of another to the widest extend and constitute the unequal spread of migrants from Warsaw inner city to the suburban ring.

Since these processes are fairly recent in case of Warsaw, adequate spacial planning can be executed in order to hinder the above mentioned negative consequences of urban sprawl. As these factors, such as the total area of greenery in a borough or transport network, can be directly influenced by the local offices planning in most cases, we believe that this work will be useful for both, Warsaw and municipal authorities. On the other hand, the example of Warsaw can contribute to understanding the bigger picture of suburbanisation patterns in post communist cities. The remaining part of this paper is structured as follows.The second section includes a review of suburbanisation theories and existing empirical evidence is offered in order to identify the possible pulling factors. In the third section we introduce methodological issues concerning different predictive models and XAI tools. The fourth section discusses empirical results including variable importance based on each model and the analysis of relationship between the dependent variable and the most important regressors. Discussion and conclusions complete the study.

2. Literature Review

According to Mieszkowski and Mills (1993), two classes of theories of suburbanization have been offered. The first one is called ”natural evolution theory” and bases on a simple chain of events. When the employment is concentrated in the center of a city, residential development takes place from the inside out. To minimize commuting costs to the Central Business District (CBD), central areas are developed first and, as they become filled in, development moves to open lands in the suburbs. The older, smaller, centrally located units, built when average real incomes were lower, filter down to lower income groups. As higher affluent households prefer to reside in outlying suburban areas, this natural working of a housing market leads to social stratification.

Transportation costs further reinforced the tendency of the middle class to live in the suburbs. Historically, when the cost of moving goods and people within cities was high, and

urban areas were dense and spatially small, high income groups located at the center. Today, due to the relatively low costs of public and private communication, this tendency has been reversed. Moreover, decentralisation of residential activity was followed by employment decentralization. Firms followed the population to the suburbs, both to provide services to suburban residents and to take advantage of lower suburban wages and land costs.

A second class of explanations of suburbanization is a generalisation of the Tiebout model (Tiebout, 1956) and stresses fiscal and social problems of central cities: high taxes, low quality public schools and other government services, racial tensions, crime, congestion and low environmental quality. These problems lead high income central city residents to migrate to the suburbs, which leads to a further deterioration of the quality of life and the fiscal situation of central areas, which induces further out-migration. The social affiliation preferences also play a role in that vicious circle: people generally prefer to live in a group of similar income, education level or social background. Hence, the suburbs are often homogeneous entities. Mieszkowski and Mills (1993) argue that the two above-mentioned theories have a number of interactions and thus, it is difficult to distinguish them empirically. Nonetheless, both theories identify factors, which can contribute to the outflow of people from the city center (costs of commuting to the CBD, average income and institutional amenities in the suburbs) and are, hence, relevant to our study.

Several researchers have investigated the outflow to the suburbs in different metropolies quantitatively. Jordan et al. (1998) analysed the outflow in 35 American metropolies in 1980-1990, taking Brueckner's population density gradient (Brueckner, 1987) as the dependent variable. They aimed at comparing the same process in various cities and found that suburbanisation proceeds quicker in areas previously unaffected by it, those of greater population growth rate and those where the employment is concentrated in the city center. It slows down with the decentralisation of the job market. Moreover, the more local authorities in the area (the number of offices), the more intensive suburbanisation occurs. On the other hand, more expensive rents in the suburbs contribute to a decrease of influx. Finally, a developed public transport system in the inner city and closer suburbs hinders motivation to move out to the farther outskirts.

Loibl (2004) offers yet another quantitative study of suburbanisation patterns. He adapted a multi-agent system approach to simulate different urban sprawl trends in Vienna with either restricted or unlimited residential area zoning and higher versus lower target residential density regulations. The simulation runs for a 30 year span were compared with the real observations. The author found out that a remarkable decrease of urban sprawl can be achieved by applying the right planning measures, even when the number of migrating households stays the same. Loibl (2004) hypothesised that the pulling factors have to be examined in detail "as polycentric growth dynamics seem to be dependent on regional attractiveness patterns within the suburban areas neighbouring the core city". Hence, part of the simulation design was to identify these "attractiveness patterns" and 4 were distinguished: landscape attractiveness (measured by the forest area quota in the neighbourhood), local services supply (access and number of attorneys), core-city availability (calculated by applying the shortest-path model to find the minimum travel time to Vienna), residential lot prices and availability of lots information (zoned, but still vacant residential areas). Loibl (2004) tested the influence of these factors on the net number of migrants by linear regression in two groups: high and low educated migrants (proxy for income groups).

He identified core-city accessibility, landscape attractiveness, services supply and the population total in the previous period (additional regressor) as significant pulling factors. Residential lot prices were not included in the final regression, as, according to the author, they are directly dependent on the demand of migrants.

An empirical study of sububanisation in post-communist countries was made by Kok (1999), who examined its determinants in Poland (49 province cities) and Hungary (19 province cities) during the political transition time. He investigated individual decision to move out of the city of 4977 people (in the case of Poland), adapting Mulder's (Mulder, 1993) life-course approach. This approach assumes that a decision to change one's place of residence is an outcome of striving for satisfying individual preferences, given constrained resources and that a "trajectory of migration" is closely connected to some spheres of life, such as work or education. Through a logit model, Kok (1999) found out that the variables having significant impact on the probability of moving out are: taking new employment vacancy, obtaining an own real estate, being married, being in age groups 18-24 and 35-39 and making that decision between 1989-1993. These findings confirm the presumption that suburban communities are rather homogeneous and that suburbanisation process started in Poland right after the political change.

Following the distinction of attractiveness measures identified in the existing literature, we include several groups of features. We use different measures of distance from the borough to Warsaw, such as a straight line in km, distance via roads, travel time with an own car and mass transport. We incorporate several measures of urban density such as population density, metropolitan density or mean of minimum distance between houses, as well as different indicators of available amenities, f.e. the number of infant places in nurseries, number of kindergartens, leisure sites and so on. We also use information about residential lot prices and supply as measures of the property market. Unemployment rate is included as a measure of local job market condition.

Percentages of votes obtained by the two biggest political parties in the 2019 parliamentary election (conservative PiS and liberal KO) are used as proxies for progressivity and social affiliations. All of the included variables are described in detail in Table 1.

The reason for taking into consideration as many as 30 features of boroughs, many of which are similar to each other (such as different measures of distance) is identifying the ones that contribute to predicting the number of migrants most accurately.

Apart from the papers of Jordan et al. (1998), Loibl (2004) and Kok (1999), we are not aware of any study, in which statistical techniques were used to explain the influx of migrants to suburban municipalities or in which the gravity model of migration framework was used in such context. Even though the gravity model of migration has recently gained popularity (Poot et al., 2016), it has been mainly used with respect to international (Beine, Bertoli, & Fernandez-Huertas Moraga, 2016; Beine, Docquier, & Ozden, 2011; Belot & Ederveen, 2012; Fan, 2005; Greenwood, 1993; Grogger & Hanson, 2011; Millock, 2015) or, in regional science, intrastate, between-province (Boyle, Flowerdew, & Shen, 1998; Pietrzak, Wilk, & Matusik, 2013; Poot et al., 2016) migration. Additionally, only a few possible pulling factors were identified in the context of suburban boroughs in the previous studies. Finally, the possible non-linearity of the relationships between the number of migrants and the pulling factors has not been yet addressed. Our study aims at filling these gaps. Therefore, in the context of explaining

suburbanisation, the use of machine learning algorithms and XAI presented here is a completely novel approach.

3. Methods

The gravity model of migration is one of the oldest and most popular analytical models of migration flows. According to that model, spatial flows of people depend positively on the size of target areas and negatively on the distance between them. In that sense, it resembles Newton's law of universal gravitation (Newton, 1687) what was first noticed by Ravenstein (1885, 1889). The idea of applying selected laws of physics to sociology and other social disciplines was introduced by Stewart (1950) as "social physics". Poot et al. (2016) deliver a thorough description about that model and following their article, the commonly applied form is:

$$M_{ij} = G\frac{P_i^{\alpha} P_j^{\beta}}{D_{ij}^{\gamma}} \tag{1}$$

where M_{ij} is the migration number of people who previously lived in area j (or i) and move to area i (or j). $P_i(P_j)$ is the population of that area at the beginning of the migration and D_{ij} is the distance between the two areas. G is the constant measuring the proportion and α, β, γ are parameters to be estimated. It is useful to take logarithm of the above equation, in order to express it in common, econometric framework:

$$\log \mathrm{M}_{ij} = \delta + \alpha \log \mathrm{P}_i + \beta \log \mathrm{P}_j - \gamma \log \mathrm{D}_{ij} + \varepsilon_{ij} \tag{2}$$

in which a zero-mean error term ε_{ij} has been added to the equation and the constant term has been replaced by the parameter δ.

Several extended forms of the gravity model of migration have been proposed (Beine et al., 2016; Fan, 2005; Greenwood, 1993; Lowry, 1966; Millock, 2015). The choice of additional variables is context-specific. The variables used in our model are described in detail in the next chapter. In our setting, there are 30 regressors for 70 observations which results is an increased dimensionality. Even though in such a setting the OLS estimator remains unbiased, high variance typically makes it perform very poorly (unless the matrix of observations is orthogonal) resulting in increased Mean Squared Error. Hence, we intend to try Elastic Net penalized regression technique, including Lasso and Ridge Regression as its special cases, which can be thought as an extension of OLS with an additional constraint on model parameters (Hastie, Tibshirani, & Friedman, 2009). As a result of this, constraint model parameters (especially at less important features) are shrunken towards zero, some are even reduced to zero. That is why these methods (except for Ridge Regression) are known for capability of performing variable selection. However, the above mentioned approach still assumes a linear relationship between the number of migrants and the features of the target place.

While it might be true for the standard independent variables of the gravity model of migration (population size and distance), there is no reason to expect it in terms of the additional regressors, such as measures of average income, amenities or transport system. If relationships between the dependent variable and the regressors fail to be linear, a linear specification is

inappropriate and may lead to incorrect inference. As the shape of the relationship is not known in advance, we use machine learning tools that can flexibly adjust to data and reveal the true relationships. According to the literature, no machine learning algorithms have been yet applied to predict the number of people migrating from the city to the suburbs. Therefore, it is not known, which group of methods yields the best predictions in this framework. As a consequence, we try a variety representing various estimation approaches. Support Vector Regression is included as another enhanced variant of OLS, in the sense that, just as OLS, it finds a hyperplane that best fits to the data. In addition, with the use of a selected kernel function, SVR applies an implicit non-linear mapping of the matrix with explanatory variables into a higher dimensional feature space, where it is more probable to find an appropriate hyperplane (Vapnik, 1995).

The other approaches are based on tree models which allow for non-linearities and take into account interactions easily. As single trees are not very useful in predicting a continuous outcome, we use two distinct approaches that base on multiple models in different ways – bootstrap averaging (bagging) and boosting. Random forest (Breiman, 2001) is an example of the bagging approach. It consists of estimating multiple independent tree models, each trained on a different bootstrap sample of the original dataset. In addition, at each split of each tree, only a random subset of all predictors is considered. In turn, Extreme Gradient Boosting (XGB) is an example of the boosting approach, which is also usually based on trees models. It builds the model in an iterative fashion at each step trying to improve the previous model by giving higher weights to observations that were not fitting well to the previous step. In addition to capturing non-linear relationships, XGB is also capable of performing regularization, for example by shrinkage like in the Elastic Net. An excellent, thorough description of these algorithms can be found in Hastie et al. (2009).

Each of these models has various hyper-parameters, the values of which have to be assumed before the optimization starts (e.g. number of trees in the random forest). Their optimal values can be found with the use of cross-validation (Hastie et al., 2009). To eliminate randomness from the model validation process, we use the leave-one-out cross-validation. Since machine learning models can flexibly adjust to the data, no functional transformations of predictors are applied.

The high variety of included regressors can be beneficial in terms of finding the pulling factors which contribute to the outflow to boroughs in the widest extent, but it can also result in collinearity problems. However, Lasso and Elastic Net are believed to be successful in selecting the most important out of correlated variables and leaving the redundant features out of the model. In our dataset, many variables are highly correlated in groups, for example we have different measures of distance at disposal. The models based on trees (e.g. random forests, XGB) are not robust to correlated features, which may disturb their results. The same problem might occur for OLS, SVR and Ridge. Hence, a pre-selection of variables out of correlated groups for these models is needed. To address this issue we performed Principal Component Analysis with varimax rotation and choose one feature with the highest loading out of each identified group of highly-correlated variables. We decided to keep 15 rotated components that cover 95% of the variance of the original variables. The dropped variables are mentioned in the Empirical Analysis part of the paper. Even though Lasso and Elastic Net are capable of

perfoming variable selection, even under great collinearity, we decide to use the restricted dataset for all algorithms, due to reasons of reliable comparison.[1]

We intend to compare the performance of all the algorithms by the common benchmarks of Root Mean Squared Error, Mean Absolute Error and R2. We report them both with regard to the training sample, as well as the validation sample. Many machine learning models are automatically optimizing fit on the training sample almost perfectly but it doesn't necessarily mean that they perform well in general. If a model performs radically better on the training sample than on the new data (validation sample), the issue of overfitting is very likely present. Selecting the best model based solely on the training sample could therefore lead to incorrect inference. Hence, introducing a validation sample is highly needed and relevant. We choose the model with the most accurate predictions based on the validation sample and interpret its results. While the outcome of linear algorithms is fairly easy to explain, interpretability of non-linear models poses a challenge. These structures are usually called "black box models" as the shape of the relationship between variables cannot be easily derived from them in the functional form that allows for interpretation. Statistical models allow to fit a specific probability model of a defined form and usually require a set of assumptions, such as normal distribution of the variables. On the other hand, machine learning methods find patterns in rich and ponderous data with minimal set of assumptions.

Explainable Artificial Intelligence (XAI) is a group of methods which allow to understand the complex structure of the black-box inner working. Multiple methods on a global (whole sample) and local (a single observation) level have been proposed and a review of them can be found in Molnar (2019). Here, we focus on a brief description of the two methods we use: Permutation Feature Importance (PFI) and Accumulated Local Effects (ALE). In our case, we are only interested in interpretability on the global level, not in the performance of the model with respect to individual boroughs.

Permutation based variable importance was first introduced by Breiman (2001) in a Random Forest algorithm. Further research was done by Fisher, Rudin, and Dominici (2019), who proposed a model agnostic tool for calculating contribution of individual features into prediction accuracy. Variable importance is calculated by randomly permuting a variable and computing an increase in prediction error with the newly created learning sample. The loss function, which quantifies the goodness-of-fit of a model for each variable is plotted for visual inspection of variable importance. Permutation Feature Importance allows for ranking the used regressors, in terms of their contribution to prediction accuracy and is model agnostic. This ranking is applied to the results of each of our models.

Furthermore, according to Zhao and Hastie (2019), the most commonly used black-box visualisation tool is the Partial Dependence Plot (PDP) introduced by Friedman (2001). It depicts the marginal effect of an input variable and the model outcome (ceteris paribus) and is a graphical representation of predictions. For a given variable, PDP averages model predictions keeping Xi feature values constant. However, this method assumes no correlation between predictors, as averaging incorporates the dependence between two features. As we show in the next chapter, that assumption is unrealistic in our case. Recently Apley and Zhu (2019) proposed an extension of PDP, which takes the correlation bias into account, called

[1] We don't report the results of PCA in this work. However, they are available on request.

Accumulated Local Effects (ALE). ALE calculates the average predicted outcome with respect to the predictor value with slight modifications. For a given predictor, ALE calculates average changes in prediction for observations in close neighbourhood to the original one. The graphical representation is then analogous to the PDP. In that way, we plot the relationships of the most important predictors (as indicated by Permutation Feature Importance) on the outcome variable for each of the models using ALE plots.

All calculations and visualizations presented in this paper were prepared in R software.

4. Dataset Description

Our dataset consists of 70 observations corresponding to the boroughs defined as a part of the Warsaw Metropolitan Area according to the EU NUTS2 norm. The observations are for one year only and we use the latest data available, which means years 2018–2020 depending on the variable. We intend to measure differences between boroughs, rather than in time and hence we assume that no serious changes in the variables' levels happened between these years[2]. The map of boroughs can be seen on Figure 1.

Our dependent variable is the number of migrants who previously lived in Warsaw and moved in to one of the suburban boroughs. The source of the information regarding migrants is registration of residence data provided by Polish Statistical Office on a borough level. Hence, we have the best possible measurement of migration for permanent residence at disposal. We want to predict the number of migrants using 30 regressors. Their description and sources can be found in Table 1. Population density and distance are the standard explanatory variables of the gravity model of migration (population density can be used instead of population total, in order to make the population measure robust to the spatial size of the borough). Following the literature and accounting for data availability, we have chosen 28 supplementary measures. We included relative income as a measure of relative wealthiness in a borough and unemployment rate as a measure of the job market condition. The percentage of votes for a ruling conservative political party (Prawo i Sprawiedliwo´s´c – PiS) and a liberal opponent in 2019 parliamentary election (Koalicja Obywatelska – KO) should depict the conservatism/liberalness of the community. The three levels of borough type classification (according to Polish administrative law, they can be classified as rural, urban-rural or urban) are an alternative measure of urbanisation. Area is an alternative (to population) borough size characteristic. The total forest area captures the extend to which forest areas hinder habitable spaces. Total greenery area, number of kindergartens, nurseries, shops, tourist sites, leisure sites, sport sites, restaurants and places of worship are measures of institutional amenities. The mean of minimum distance between two houses and number of dwelling units per km2 capture metropolitan density. The mean of a minimum distance from a house to a large road, presence of suburban train station, driving distance from borough to the city center (separately in meters and minutes), as well as, travel time from the borough to the city center with public transport depict the transport system condition. The price of m2 of housing, number of parcels available for sell and their mean price measure the real estate market. All data were scraped from open source databases (see Table 1: Source).

[2] The mean number of migrants in the boroughs has been relatively constant in 2008-2019. The time series for the number of migrants in each borough can be shared on request.

In spite that the features of spatial entities are expressed as absolute values in the existing gravity model of migration literature, we decided to take three characteristics as ratios. Relative income was calculated with respect to the average income weighted by the population total in boroughs. We hypothesise that relative income can depict 'wealthiness' in a borough more accurately than a crude measure. Moreover, in our preliminary exploratory data analysis by OLS, using this measure resulted in higher prediction accuracy than when regular income per capita was used. The second variable is the ratio of a forest area to the total area. While the presence of a forest in suburbs might be an attractive pulling factor, a good way of measuring green spaces available to the common is needed. Total greenery spaces is such a measure in our study: it includes parks, green plazas and forest areas available for the public. Both the total area of greenery and forests are provided by Polish Statistical Office. However, not all forest areas are classified as (sub)urban forests available for the common. The total area of greenery will likely capture the nature attractiveness in a borough. By relating the forest area to the total area, we intend to measure the extend to which the forests hinder habitable spaces. For example, one of the biggest national parks in Poland is located north-east of Warsaw and covers many boroughs in our analysis. By Polish law, it is forbidden to settle in close proximity to the national park. This is why the ratio should depict this effect better than the crude measure. The third variable is metropolitan density – the number of dwelling units per km2. Metropolitan density is traditionally defined as such a ratio and is therefore considered in that way.

Table 1. Dataset description: the features of boroughs.

Variable	Description	Source
check_in	number of migrants[a]	Polish Statistical Office
pop_dens	population density in # of people/ha	Polish Statistical Office
dist	distance between Warsaw and borough centers (straight line) in km	Google Maps
income	measure of average relative income[b]	own calculation based on data from PSO
unempl	unemployment rate (percentage)	Polish Statistical Office
pis	percentage of votes obtained by PiS in 2019 parliamentary elections	National Electoral Commission
ko	percentage of votes obtained by KO in 2019 parliamentary elections	National Electoral Comission
bu	borough type - urban[c] (dummy)	Polish Statistical office
bur	borough type - urban-rural (dummy)	Polish Statistical office
br	borough type - rural (dummy)	Polish Statistical office
area	total area in ha	Polish Statistical Office
forest	ratio of forest area to the total area	Polish Statistical Office
greenery	area of green amenities (parks etc.) in ha	Polish Statistical Office
kinder	number of kindergartens	Polish Statistical Office
nursery	number of infant places available in nurseries	Polish Statistical Office
shops	count of places tagged 'shop'	Open Street Map
tourist	count of places tagged 'tourist'	Open Street Map
leisure	count of places tagged 'leisure'	Open Street Map
sport	count of places tagged 'sport'	Open Street Map
restaurant	count of places tagged 'restaurant'	Open Street Map
worship	count of places tagged 'worship'	Open Street Map
bldgs_dist_mean	mean of minimum distance between two houses in m	Open Street Map
bldgs_hghws_dist_mean	mean of minimum distance from a house to a large road in m	Open Street Map
md	metropolitan density: number of dwelling units per km^2	Open Street Map
train	presence of a suburban train station (dummy)	Google Maps
dist_waw_drive	driving distance from the borough to the city center in m	Google Maps
time_waw_drive	driving distance from the borough to the city center in min	Google Maps
min_dur	travel time from the borough to the city center with public transport in min	e-podroznik.pl[d]
price_m2	price of m^2 of housing in PLN [e]	Polish Statistical Office
parcel_n	number of parcels available for sell on 03.08.2020	gratka.pl[f]
parcel_mean	average price of a parcel on 03.08.2020 per m2	gratka.pl

a The dependent variable: number of migrants - people who checked out of Warsaw and reported residence in one of the suburban boroughs.

b Average income in a borough (in PLN) divided by the average income in all boroughs weighted by population (in PLN). Calculated based on personal income tax data.

c Three borough types can be distinguished according to Polish Administrative Law classification: urban, urban-rural, rural.

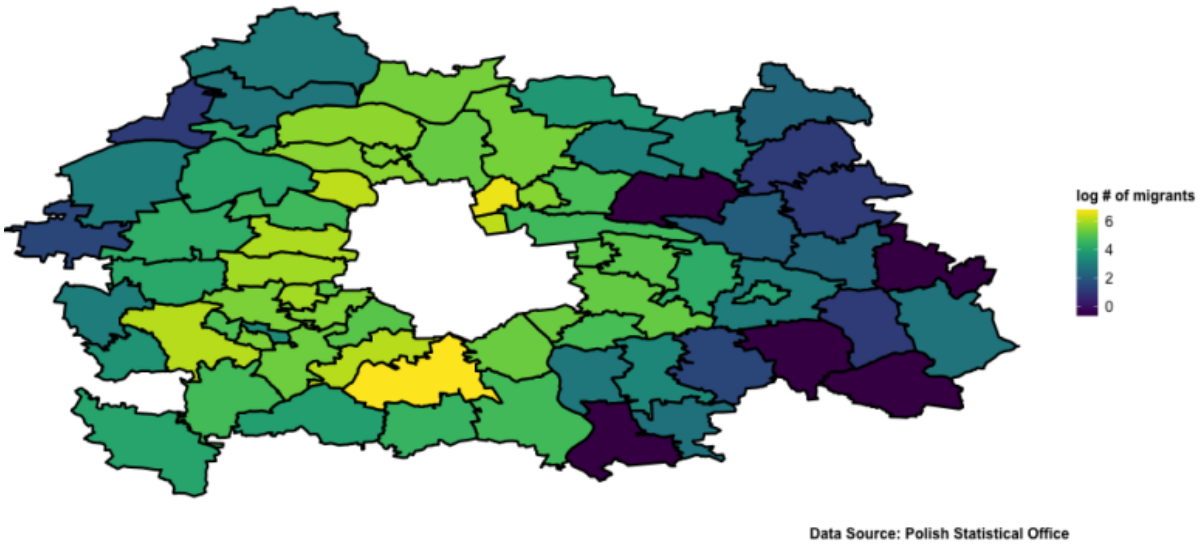
d e-podroznik.pl is the largest website collecting travel possibilities between places in Poland. It has information about even the smallest carriers at disposal and was therefore chosen as the most thorough data base. No information was available for 3/70 boroughs on this platform. However, public means of transport are functioning in that boroughs but carriers are not registered on any platform. In these cases we called the carriers directly and asked about the travel time.

e Only a price on a county level was available in the Polish Statistical Office and is hence included.

f gratka.pl is the largest online real estate platform and was therefore chosen as the most adequate data base.

Figure 1 shows the visual representation of the number of migrants. Supplementary Figures A2, A3, A4 are given in the Appendix and present the key variables of the gravity model of migration (population density and distance), as well as relative income, which, according to the above mentioned literature, is expected to have large influence on the number of migrants. All values correspond to the year 2019 and were logarithmed for clearance. The spatially uneven outflow of migrants is visible at first glance. Less inequalities are visible on the map presenting population density. While most of the boroughs are sparsely populated, a few of them are dense. Average income appears to be higher in the boroughs closer to Warsaw. Visual inspection leads to a conclusion that the number of migrants seems to be higher in the boroughs of greater population density, smaller distance to Warsaw and of higher relative income.

Figure 1. Number of migrants (log) in 70 suburban boroughs of Warsaw, in 2019.



Summary statistics of all variables are attached in Table A1 in the Appendix. The mean number of migrants has been 146.14 in 2019, however the standard deviation exceeds this value (182.96). There were boroughs, where no migrants reported residence, while 909 people moved to the most popular one. The average population density is 5.59 people/km2, but here as well it differs greatly between boroughs (0.25 min, 39.93 max). The average distance measured as a straight line is 29.05 km. The closest borough is only 9.07 km away from Warsaw city center,

while the farthest one is located almost 60 km away. Figure A1 presents the correlation matrix. It can be seen that the variables are correlated up to 75%, which justifies the use of methods robust to correlation, such as Accumulated Local Effects.

5. Empirical Analysis

The choice of variables used in all models is based on exploratory data analysis with PCA and we include only one feature out of each identified group of correlated regressors. Hence, 8 following regressors are dropped: driving time and distance from borough to Warsaw, the number of restaurants and sport sites, metropolitan density, the percentage of votes obtained by the ruling conservative political party and three borough type classifiers. As a result, 21 regressors are used in all models. We choose the optimal values of hyper-parameters with the use of leave-one-out cross-validation (LOOCV). The optimal values of hyper-parameters are provided in the Appendix A. It can be seen that the optimal α for Elastic Net is 1, resulting in Lasso. Hence, 6 algorithms are considered and run: OLS, Ridge, Lasso, RF, SVR and XGB. We then compare RMSE, MAE and R2 of the all models and choose to interpret the outcome of the most accurate one. The summary of model accuracy measures is presented in Table 2, for both, the training and validation sample. The validation errors correspond to the average from 70 models run by LOOCV, while the training errors are for one, final model with optimal hyper-parameters. In addition, the errors and R2 were calculated for an OLS model that was obtained by the general-to-specific approach. It includes only four variables significant at 10% level: population density, distance, number of infant places in nurseries and presence of a suburban train station.

Table 2. Model errors comparison.

Sample	Validation			Training		
Model	RMSE	MAE	R^2	RMSE	MAE	R^2
Ordinary Least Squares	111.12	79.86	0.62	97.55	71.93	0.71
Ridge Regression	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
Lasso (full dataset)	126.20	81.99	0.51	106.31	71.06	0.66
Random Forest	123.05	75.63	0.54	80.29	53.00	0.90
Support Vector Regression	119.15	78.33	0.57	65.63	34.45	0.87
Extreme Gradient Boosting	109.56	70.07	0.63	78.25	45.76	0.81

Mind that we report model statistics for Lasso on the restricted and full dataset. It can be seen that the statistics on the training sample are much in favour of Lasso on the restricted dataset. However, model errors and R2 on the validation sample, which reflects real models' performance more accurately, are highly comparable for both models (but slightly better for Lasso on the restricted dataset). In conclusion, Lasso is capable of variable selection, even if the features are highly correlated. Yet, for the purpuse of direct comparability of results, we use the restricted dataset for all models in further analysis.

In terms of the training sample, Extreme Gradient Boosting yields the lowest RMSE, while Support Vector Regression – the lowest MAE, and Random Forest – the highest R2. The

errors and R2 for OLS, Ridge and Lasso are significantly worse. This result confirms that the models which catch non-linear relationships between predictors and the outcome explain substantially more variability of the modelled phenomenon than the linear models. However, due to potential risk of overfitting of machine learning models, their performance is usually assessed based on the validation sample. This can be viewed as checking predictive performance of models on the new data, not used in the estimation process. It is clear to see that Extreme Gradient Boosting yields the most accurate predictions there – its RMSE and MAE are the lowest while R2 is highest. Next comes the OLS with only a slightly worse performance, followed by RF and SVR. In terms of the validation sample, penalized regression algorithms exhibit the poorest accuracy of predictions.

In order to identify the pulling factors, which contribute to the prediction accuracy in the widest extent, we calculate Permutation based Feature Importance (PFI) and report it as a percentage change in RMSE after permuting each variable, for each of the models. We also calculate the average importance of a feature taking into account all models and then rank the features with respect to this value, from the most to the least important one. Table 3 presents the resulting ranking of importance.

Table 3. Permutation Feature Importance for different models

Variable	OLS	Ridge	Lasso	RF	SVR	XGB	Average
nursery	1.18	0.12	0.63	0.25	0.94	0.41	0.59
dist	0.50	0.09	0.33	0.38	0.75	0.33	0.40
income	-	0.01	-	0.24	0.25	0.52	0.26
ko	-	0.04	0.02	0.12	0.38	-	0.14
worship	-	0.01	-	0.18	0.06	0.27	0.13
pop_dens	0.20	0.00	0.05	0.01	0.14	0.15	0.09
area	-	0.00	-	-	0.26	0.00	0.09
greenery	-	0.05	0.01	0.05	0.25	-	0.09
forest	-	0.00	-	-	0.14	-	0.07
tourist	-	0.01	0.01	0.00	0.28	0.02	0.06
shops	-	0.05	0.03	0.08	0.05	0.03	0.05
train	0.06	0.01	0.03	-	0.08	-	0.05
kinder	-	0.05	0.03	0.04	0.05	0.05	0.04
bldgs_hghws_dist_mean	-	0.00	-	0.00	0.11	0.05	0.04
leisure	-	0.00	-	0.03	0.07	0.04	0.04
parcel_mean	-	0.01	-	0.00	0.08	0.04	0.03
bldgs_dist_mean	-	0.00	-	0.02	0.08	0.03	0.03
unempl	-	0.01	-	0.03	0.04	0.04	0.03
parcel_n	-	0.02	0.01	0.02	0.05	-	0.02
price_m2	-	0.01	-	0.00	0.08	0.00	0.02
min_dur	-	0.01	-	0.01	0.05	0.01	0.02

Note: The importance and average importance expressed as a change in RMSE after permuting the variable are given in percents for all models (f.e. 0.59 means 59%).

Random Forest and Extreme Gradient Boosting are methods based on trees. When setting the hyper-parameters, one has to take into consideration the trade-off between predictive power of

a model and a risk of over-fitting. The larger the number of trees or boosting iterations, the better the model is fitted to the data. However, this comes at the expense of prediction accuracy on new data. In the models based on trees, some variables might be never used for the splits if their contribution to the prediction accuracy is negligible. Hence, the hyphens in Table 3 – they mean that a particular feature was never used within a particular model. This is acceptable for the reason that our purpose is to identify the most important pulling factors, out of a great variety. Lasso and Elastic Net are also capable of variable selection by shrinking some coefficients to 0. Ridge and SVR do not perform variable selection.

The top 6 on average rated features are number of infant places in nurseries, distance to Warsaw as a straight line, relative income, percentage of votes obtained by the liberal party in 2019 election, number of worship sites and population density. The number of places in nurseries is a strong pulling factor since migrants are mostly young couples and families with children moving out of flat to a single family house.

Moreover, suburban migrants usually still work in the core city and commute to work by an own vehicle or public transport. Distance measured as a straight line is another feature ranked highly – a pushing factor. As can be seen on Figure A2 in the Appendix, the boroughs populated most densely are also the ones located closer to Warsaw core.

Relative income is ranked the third and population density – the sixth most important measure, which is strongly in line with the existing literature on determinants of suburbanisation. The percentage of votes obtained by the liberal party is possibly a proxy for age in boroughs, progressivity of its inhabitants or both. The number of worship sites, which can be summarized as an institutional amenity, was ranked fifth.

In Poland, religious beliefs play a significant role in peoples lives – approximately 90% of Poles are baptized Catholics. Hence, it is not surprising that availability of a church in close proximity plays a role in settlement processes. The other features ranked high on average are total area, total greenery spaces, ratio of forests to the total area, number of shops and presence of a suburban train station. For brevity reasons, we plot Accumulated Local Effects for all models to interpret the (non-linear) relationship between the top 6 ranked features and the number of migrants.

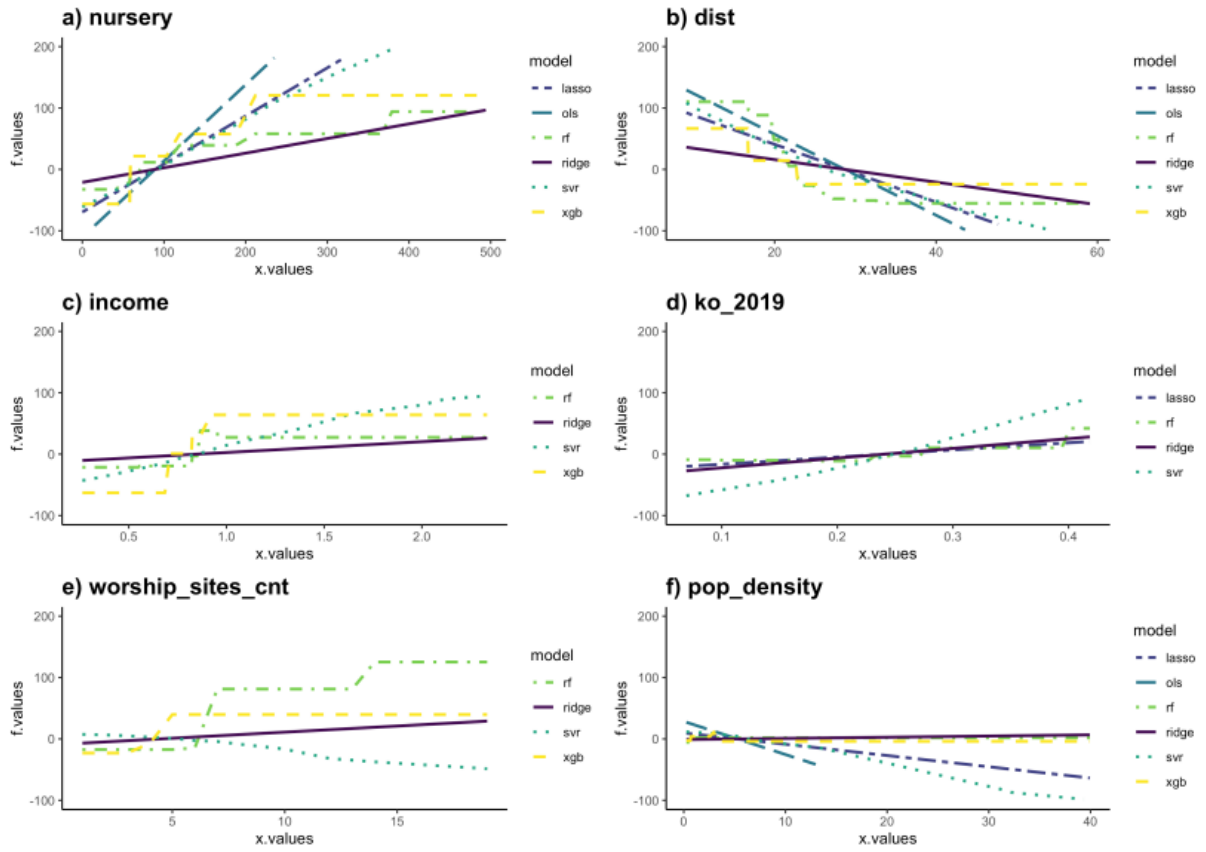


Figure 2. ALE plots for 6 most important determinants of migrations identified by the average PFI rank

Figure 2 presents ALE plots for the number of infant places in nurseries, distance to Warsaw as a straight line, relative income, percentage of votes for the liberal opponent, number of worship sites and population density. It is clear to see that assuming a linear relationship between the number of migrants and these features is infeasible.

The relationship between the number of migrants and the number of infant places in nurseries is positively linear for all algorithms, except for XGB and RF (where it appears constant after 200 (XGB) and 380 (RF)). The largest positive slope is for OLS and Lasso, but there is a clear positive tendency for Ridge and SVR as well.

The number of nurseries is an institutional measure. Nurseries are desirable especially by families with children and young couples – the parents of both groups usually work in Warsaw and it is comfortable to leave an offspring safely in a nursery for a long work day. A nursery can be possibly replaced by grandparents who take care of infants, while parents are at work, but in case of Warsaw, there has been an ongoing influx of people from other parts of Poland to the whole agglomeration in the last 30 years. Migrants from other parts of Poland wouldn't have parents in Warsaw, who could take care of their children. Hence, the often need for nurseries. Furthermore, the relationship between the number of migrants and distance is clearly negative for all algorithms, which is conforming with the existing literature on

suburbanisation, as well as the gravity model of migration. The number of migrants drops in both, RF and XGB around 20 km, after which value is remains almost constant. Income is another important, strong pulling factor. The number of migrants rises from 0 to around 80 for income equal 0.75 (which can be interpreted as the average income in a borough being 75% of the average income in all municipalities, weighted by population) in both, RF and XGB. It can be seen that the relationship between the number of migrants and the percentage of votes obtained by the largest liberal opponent in 2019 parliamentary election is clearly positive for all algorithms. In RF, the number of migrants rises delicately at the percentage of votes equal 40%. The boroughs, where KO got more than 35% are also those with highest relative income. Approval for the largest liberal party in Poland is likely a proxy for age and progressiveness in boroughs.

Migrants are usually young, affluent middle-class representatives who support novel, progressive ideas and choose leaders accordingly. Both income and preference for liberal party depict social affiliations' role in migration in our study. The number of worship sites is another pulling factor. It can be seen that the number of migrants rises at value 5 for both RF and XGB. Surprisingly, the slope is slightly negative for SVR, but positive for other algorithms where this feature was used. It is worth to mention, that the preference to choose liberal rulers does not necessarily contradict a need for religiousness. Although the vast majority of Poles is formally Catholic, not all of them are devout and especially not in agglomerations. Services provided by Catholic church (such as weddings, funerals) are yet, still desirable as religion is strongly related to tradition in Poland. Finally, population density, with respect to the number of migrants, has a constant slope for Ridge Regression, RF and XGB, while it's negative for Lasso, OLS and SVR. This result is the most surprising one out of all obtained in this study, as it is not in line with the theory of the gravity model of migration.

However, it conforms to the "natural evolution theory" described here in the Literature Review. It turns out that, in case of suburban migration, migrants prefer less densely populated municipalities, of more living space.

After visual inspection of the plots and taking the average values of PFI measured as percentage change in RMSE into consideration, we can conclude that migrants choose boroughs of more infant places available in nurseries (PFI = 59%), especially those with more than 200 places. They settle in municipalities located closer to Warsaw (PFI = 40%), in particular those located under 20 km to the center of Warsaw and prefer those of higher relative income (PFI = 26%) - primarily above 0.75. Migrants choose more liberal boroughs (PFi = 14%), but at the same time those of greater number of temples (PFI = 13%). Finally, they prefer sparsely populated municipalities (PFI = 9%) of more living space.

6. Conclusions

In this paper, we investigated the suburbanisation phenomenon in the Warsaw agglomeration. We aimed to identify the features of boroughs which are key pulling factors for migrants and constitute choosing one borough instead of another. Basing on the extended gravity model of migration, we built several predictive models and assessed their performance by the common benchmarks of RMSE, MAE and R2. Extreme Gradient Boosting turned out to yield the most accurate predictions. Permutation based Feature Importance was calculated for each chosen feature, for each model and Accumulated Local Effects were plotted for the top 6 most important variable as indicated by the average PFI for all models. We identified 4 pulling factors

by mean PFI: the number of infant places in nurseries, relative income, percentage of votes obtained by KO in 2019 parliamentary election and the number of worship sites. While the mentioned income and percentage of votes are likely proxies for social affiliation preferences, the two latter features are institutional amenities. Our finding with respect to these four measures are especially valuable in terms of spacial planning and can be used by local authorities. We also determined 2 pushing factors: migrants settle in municipalities located closer to Warsaw, in terms of distance in km and prefer sparsely populated places. In order to attract migrants to the further located boroughs, attractive means of transport, such as suburban train can be built.

Our findings are (in some extent) in contrast with the previous, quantitative research on the topic. For example, population density turned out to be a pushing, rather than a pulling factor. Nonetheless, previous studies were based solely on OLS, which we proved to be ineffective when non-linear relationships are observed in the data and having lower predictive power than Extreme Gradient Boosting. It appears that, even though OLS is pretty accurate in predicting the number of suburban migrants, it can still be beaten by a more sophisticated algorithm. In addition, other algorithms let us identify more relevant features than OLS, where only four were significant on 10% level. Finally, the results conforming with the previous findings is the negative relationship between the number of migrants and the distance from the city center and positive influence of the average income on the number of migrants.

In addition to shedding light on the local context of Warsaw, we have also filled gaps not addressed in the previous studies: considering a wide variety of possible pulling factors and identifying non-linear relationships between the dependent variable and the regressors. We believe that our work can be valuable for spacial planners not only in Poland, but in other countries of similar suburbanisation patterns.

Several possible extensions of this paper are possible. A natural follow-up is investigating the same problem in a panel setting. In our work, some of the features were scrapped from the Internet sources, such as Google Maps, where historical data is not available. Yet, taking a 30 year time frame (in case of Poland) could possibly lead to a deeper insight into the outflux from Warsaw to the suburbs, for example by taking into consideration time effects. An individual-level analysis could also yield valuable conclusions about the mechanisms behind people's decisions. Finally, a comparable study of different metropolies in Poland and the region can be conducted to verify if the processes are indeed similar in all post-communist countries. Thus, further research on these issues is highly anticipated.

7. References

Apley, D., & Zhu, J. (2019, 12). Visualizing the effects of predictor variables in black box supervised learning models.

Beine, M., Bertoli, S., & Fernandez-Huertas Moraga, J. (2016). A practitioners' guide to gravity models of international migration. The World Economy, 39(4), 496-512.

Belot, M., & Ederveen, S. (2012). Cultural barriers in migration between OECD countries. Journal of Population Economics, 25(3), 1077-1105.

Boyle, P., Flowerdew, R., & Shen, J. (1998). Modelling inter-ward migration in Hereford and Worcester: The importance of housing growth and tenure. Regional Studies, 32(2), 113-132.

Breiman, L. (2001). Machine learning, volume 45, number 1 - Springerlink. Machine Learning, 45, 5-32.

Brueckner, J. (1987). The structure of urban equilibria: A unified treatment of the Muth-Mills model. In E. S. Mills (Ed.), Handbook of regional and urban economics (1st ed., Vol. 2, p. 821-845). Elsevier.

Caves, R. (2004). Encyclopaedia of the city. Taylor and Francis.

Downs, A., McCann, B., & Mukherji, S. (2005). Sprawl costs: Economic impacts of unchecked development. Island Press.

Duany, A., Plater-Zyberk, E., & Speck, J. (2010). Suburban nation: the rise of sprawl and the decline of the American dream (10th ed.).

Farrar, Straus and Giroux.Fan, C. (2005). Modelling interprovincial migration in china, 1985-2000. Eurasian Geography and Economics, 46(3), 165-184.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20, 1-81.

Fouberg, E., & Murphy, A. (2020). Human geography: people, place, and culture. Wiley.

Friedman, J. (2001, 11). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29.

Greenwood, M. (1993). Internal migration in developed countries. In M. R. Rosenzweig & O. Stark (Eds.), Handbook of population and family economics (1st ed., Vols. 1, Part B, p. 647-720).

Grogger, J., & Hanson, G. (2011). Income maximization and the selection and sorting of international migrants. Journal of Development Economics, 95(1), 42-57.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

Hobbs, F., & Stoops, N. (2002). Demographic trends in the twentieth century. census 2000 special reports.

Jordan, S., Ross, J., & Usowski, K. (1998). U.S. suburbanization in the 1980s. Regional Science and Urban Economics, 28(5), 611-627.

Kahn, M. E. (2000). The environmental impact of suburbanization. Journal of Policy Analysis and Management, 19(4), 569-586.

Kok, H. (1999). Migration from the city to the countryside in hungary and poland. GeoJournal, 49(1), 53-62.

Lisowski, A. (2004). Social aspects of the suburbanisation stage in the agglomeration of Warsaw. Dela.

Loibl, W. (2004). Simulation of suburban migration: driving forces, socio-economic characteristics, migration behaviour and resulting land-use patterns. In (p. 201-223).

Lowry, I. S. (1966). Migration and metropolitan growth: two analytical models. Chandler Pub.

Mieszkowski, P., & Mills, E. (1993). The causes of metropolitan suburbanization. Journal of Economic Perspectives, 7(3), 135-147.

Millock, K. (2015, 11). Migration and environment. Annual Review of Resource Economics,

7, 35-60.

Molnar, C. (2019). Interpretable machine learning.

Mulder, C. (1993). Migration dynamics: A life course approach. Amsterdam: Thesis Publishers.

Murray, P., & Szelenyi, I. (2009). The city in the transition to socialism. International Journal of Urban and Regional Research, 8, 90 - 107.

Newton, I. (1687). Philosophiæ naturalis principia mathematica. Benjamin Motte.

Nuissl, H., & Rink, D. (2005). The "production" of urban sprawl in eastern Germany as a phenomenon of post-socialist transformation. Cities, 22(2).

Pietrzak, M., Wilk, J., & Matusik, S. (2013). Gravity model as the tool for internal migration analysis in Poland in 2004-2010 (Working Papers No. 28/2013). Institute of Economic Research.

Poot, J., Alimi, O., Cameron, M., & Mar´e, D. (2016). The gravity model of migration: The successful comeback of an ageing superstar in regional science., 2016, 63-86.

Ravenstein, E. (1885). The laws of migration, part 1. Journal of the Statistical Society of London, 48(2), 167-235.

Ravenstein, E. (1889). The laws of migration, part 2. Journal of the Royal Statistical Society, 52(2), 241-305.

Stewart, J. (1950). The development of social physics. American Journal of Physics, 18(5), 239-253.

Sturm, R., & Cohen, D. (2004). Suburban sprawl and physical and mental health. Public health, 118, 488-96.

Tiebout, C. (1956). A pure theory of local expenditures. Journal of Political Economy, 64.

Timar, J., & Varadi, M. (2001, 10). The uneven development of suburbanization during transition in hungary. European Urban and Regional Studies, 8, 349-360.

Tung, A. (2001). Preserving the world's great cities: The destruction and renewal of the historic metropolis. Three Rivers Press.

Vapnik, V. N. (1995). The nature of statistical learning theory. Berlin, Heidelberg: Springer-Verlag.

Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. Journal of Business and Economic Statistics, 1-19.

Appendix A

Model hyper-parameters found by one-leave-out cross validation:

(1) Ordinary Least Squares: none (2) Ridge Regression:

•α=0

• λ = 278.86

(3) Lasso:

•α=1

• λ=9.57

(4) Elastic Net:

•α=1

• λ=9.57

(5) Random Forest:

• number of trees = 17

• max number of variables considered by each split = 10

• min size of a node=12

• max number of terminal nodes = 7

(6) Support Vector Regression:

• kernel: radial

• γ = 0.002

• ε=0.13

• cost = 207

(7) Extreme Gradient Boosting:

• η=0.36

•γ=0

• α = 40

•λ=0

• max number of terminal nodes = 3

• min size of a node=7

• subsample = 0.8

• column sample by tree = 0.4

Table A1. Basic summary statistics for all variables used in the analysis

Variable	Mean	Std. Dev.	Min	Max	Scale
check_in	146.14	182.96	0	909	# of migrants
pop_dens	5.59	9.38	0.25	39.93	people/km^2
dist	29.04	12.16	9.07	59.04	km
income	0.86	0.45	0.26	2.33	ratio
unempl	0.03	0.01	0.014	0.065	percentage
pis	0.47	0.12	0.30	0.73	percentage
ko	0.24	0.08	0.07	0.42	percentage
bu	0.2	0.40	0	1	dummy
bur	0.27	0.45	0	1	dummy
br	0.53	0.50	0	1	dummy
area	8001.75	4129.56	576	20577	ha
forest	0.22	0.16	0.00	0.76	ratio
greenery	55.69	63.67	2.33	285.98	ha
kinder	9.50	9.84	0	58	# of kindergartens
nursery	91.69	109.45	0	494	# of nurseries
shops	140.43	170.54	1	781	# of shops
tourist	53.27	153.22	0	1278	# of tourist sites
leisure	660.20	866.04	8	5761	# of leisure sites
sport	103.40	113.10	0	469	# of sport sites
restaurant	11.03	13.52	0	64	# of restaurants
worship	4.93	3.44	1	19	# of worship sites
bldgs_dist_mean	25.21	6.06	13.29	48.16	m
bldgs_hghws_dist_mean	370.39	164.02	105.45	859.75	m
md	129.54	164.76	2.79	738.08	dwelling units/km^2
train	0.51	0.50	0	1	dummy
dist_waw_drive	38.16	13.67	13.19	72.06	km
time_waw_drive	42.97	11.92	21.45	74.52	min
min_dur	46.33	22.05	16	105	min
price_m2	4712.64	484.59	3889	5717	PLN
parcel_n	46.74	50.13	1	247	# of parcels
parcel_mean	327.45	670.78	16	5569.49	PLN

Figure A1. Correlation matrix between all variables used in the analysis

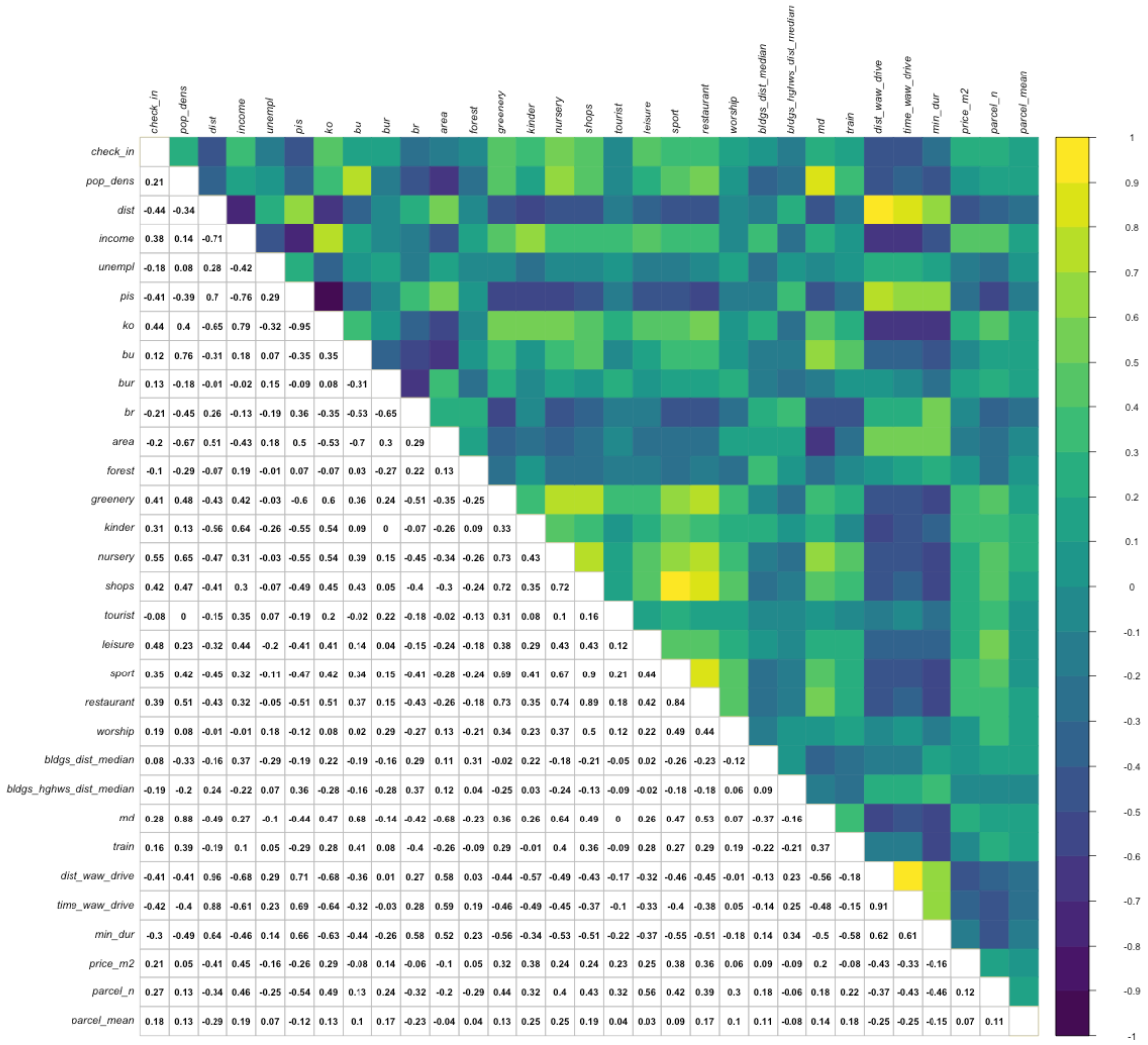


Figure A2: Visual representation of population density in suburban boroughs (2019).

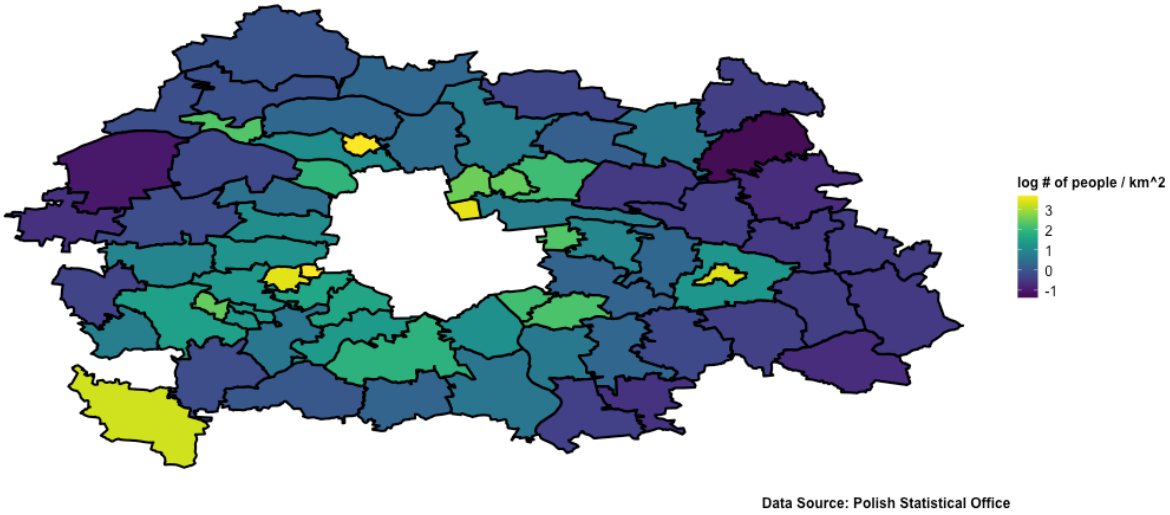


Figure A3: Visual representation of the straight-line distance to Warsaw in suburban boroughs (2019).

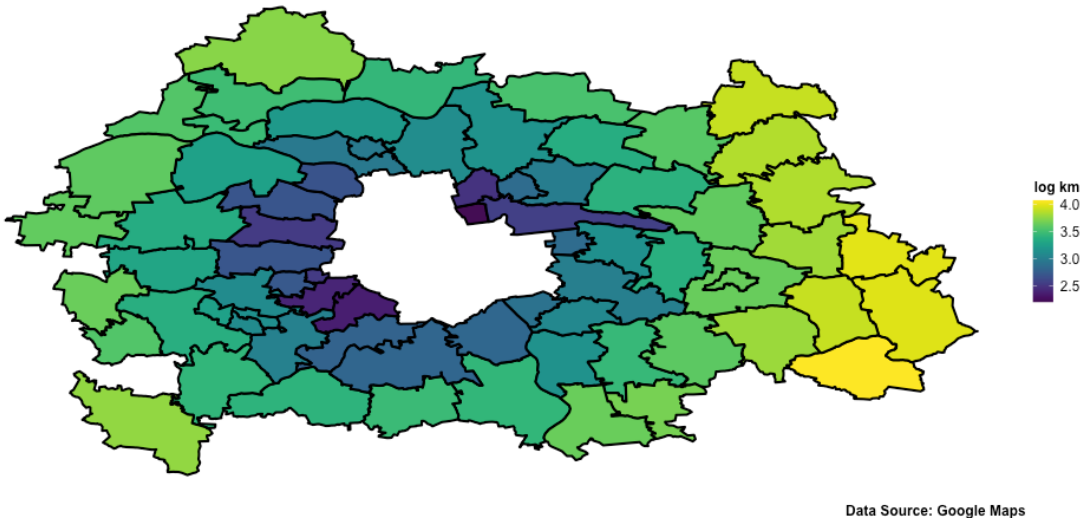
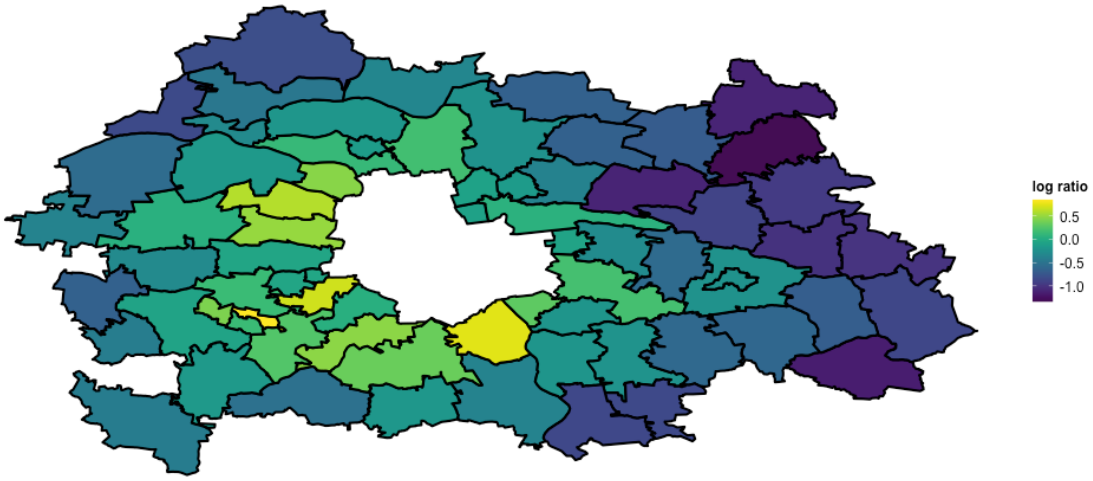


Figure A4: Visual representation of the relative income in suburban boroughs (2019).



Data Source: Polish Statistical Office

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl