Introduction

Choosing an optimal location for any business is a difficult decision every entrepreneur faces. As studies show, the location has potentially tremendous effect on revenue. This is particularly important in restaurant industry, where potential customers can be easily tempted by attractive-looking interior or simply proximity to workplace or home.

In this paper I study the factors driving restaurants locations, specifically in Warsaw market. I study the influence of two factors in depth. First, I check whether **restaurants location is dependent on other businesses locations** in particular area. Second, **if it is dependent on number of inhabitants in the area**. I also assess the importance of communication infrastructure surrounding the restaurants, specifically number of bus stops and roads density.

There are just a few papers addressing restaurants location specifics in particular. Most of the existing works are also as old as 40 years, and thus are possibly outdated due to transformations in the industry. The specifics of restaurant industry is different in each city, thus extrapolating the results from other cities on Warsaw should be done carefully.

Warsaw is still an immature market when it comes to restaurants. The growth of the whole sector is steady, and every month new restaurants are opened. Also, average expenditure for restaurants in Poland is constantly growing (ca. 9% y/y).[1]

From a technical point of view, this study is conducted using Algorithmic Modeling methods (as specified by Breiman and others 2001). I use state-of-the-art Machine Learning modeling with Variable Importance (VI) assessment methods. Specifically, I estimate Random Forest and Logistic Regression models. For assessing Variable Importance, I use Model Class Reliance algorithm (Fisher, Rudin, and Dominici 2018). To make the analysis more robust, I also employ other VI method, that is Mean Decrease Gini measure computed from Random Forest model results. Modeling is a widely acclaimed method for inference when the relationships are highly complicated and assuming a specific type of data model is impossible. Recent advances in the field of Explainable Artificial Intelligence (Gunning 2017) enable researchers to draw conclusions from black-box models, which was not possible before, at least not to such extent.

[1] https://stat.gov.pl/obszary-tematyczne/warunki-zycia/dochody-wydatki-i-warunki-zycia-ludnosci/budzety-gospodarstw-domowych-w-2017-r-,9,12.html

As the competitiveness of the market is raising, restaurateurs should seek for new ways to stand out of the crowd. Results of this study can help them understand what creates the biggest impact on the success of a restaurant.

Literature overview

Businesses locations studies

The location dimension was neglected in mainstream economics for a long time. As Krugman said: *"How did the mainstream cope with spatial issues? By ignoring them."* (1997). Despite that, various theories of location have been developed through the years.

First approaches in the stream of classical economy concentrated on industry and agriculture. The earliest theory concerning location is by Von Thünen (1875). His model of agricultural land laid foundations for later works. Theory of industrial location made by Weber (1929) concentrated on transportation costs of raw materials and final products. According to the theory, entrepreneurs create their industrial sites in places where the cost of transportation was the lowest.

Works of Walter Christaller should also be mentioned. He developed a central theory model, in which he tried to explain the location of cities and villages across the space. Similar to von Thunen model, a village has one function, that is to create space for exchange of goods produced somewhere else.

Hotelling's linear city model (1990) is one of classical game theory models. Every firm wants to achieve the best location and attract as many customers as possible. The novelty of this model is that firms take their competitors' locations into account. As a result, similar firms are getting very close to each other, and in their interest is to have similar product to the competitors'. This phenomenon is visible in retail market, especially bars, restaurants and pharmacies.

These few models served as a basis for later empirical works in the field of location concerning businesses of various types. Important factors for choosing a site for a factory and service-based businesses are fundamentally different, and thus are usually studied separately. For example, in industry transportation costs of raw materials and final products must be taken under serious consideration. Availability of a big pool of skilled workforce specialized in a particular industry also plays an important role. On the other side, the demand for retail

stores and services is often location-bound and is bigger in the cities. Van Noort and Reijmer (1999) notice that not only sector in which a business operates matters, but also small and big businesses should be treated separately in location studies. Motivations and available resources for these two segments vary considerably. Their study focuses on smaller businesses and is rather qualitative in nature. They claim that SMEs location decision is a short-term and is not a result of an extensive consideration. Also, the smaller the business, the bigger chance is that it will be established in proximity of its owners private house.

Restaurant industry shares some of the specifics with retail industry in general, and thus studies in this broader sector are analyzed. Also, retail industry (same as restaurants), consists mostly of small businesses, so studies in this area should be generalizable.

There are two streams of studies in businesses location. One is determining the factors that drive entrepreneurs to opening the business in particular area. An example is an early study made by Rolph (1932). He shows that retail stores location is highly correlated with population density, average income in the area and other factors. This study strives to find factors that particular area's entrepreneurial landscape consists of and thus determine businesses locations.

Second stream of studies concentrates on spatial agglomeration of businesses. There is an assumption that previously existing businesses of the same type could help the performance of the new firm. An example could be a restaurant district that is popular among customers. There is a big chance that customers will be interested in new restaurant in the area just because they have seen it when visiting other places.

Contrary to Rolph (1932), Lee and Koutsopoulos (1976) tried to prove that population density does not have a significant influence on stores locations. They suggested that spatial agglomeration may be a more important factor than various socio-economic factors in the area, when making a decision about opening a store. Dubé, Brunelle, and Legros (2016) showed that, in accordance with the classical location theories, businesses in primary sectors tend to be isolated and far from agglomeration center. In contrast, highly advanced manufacturing and services showed high clustering tendency in the cities.

Concerning restaurants clustering tendency, there are two studies that were meant to assess that. Pillsbury (1987) studied the area of Atlanta, USA. He did not classify restaurants by their types (fast food, family etc.), but rather the customers' needs they serve. This is based on the

fact that for some types of restaurants (e.g. soul food) there is no need for good availability, and *journey to dine* becomes an integral part of the dining experience. As Pillsbury claimed, "*Today, virtually no new restaurant is found outside a cluster of its competitors.*". Moreover, restaurants clustered by non-spatial criteria (socio-economics, ambiance and accessibility) corresponded with their geographical locations. This means that similar restaurants tend to be closer to each other.

Another example of assessing clustering tendency is the study of Smith (1985). He showed that this phenomenon is highly visible in fast-food restaurants. Another finding was that population density is exponentially related to car traffic volume in the area. Higher traffic was also correlated with presence of restaurants.

There is a possible reason why taking account of agglomeration phenomenon is widely present in location studies. Data concerning location is usually very easy to acquire, for instance compared to sales data in different locations. No matter how valuable insights one would gather from such information, data of such kind is usually unavailable to independent researchers (Smith 1983).

It should be stated that two approaches widely applied in business location studies are highly dependent on each other. Spatial clustering can be present from the reason stated above, that is other businesses presence. The second reason for agglomeration is because there are good conditions for particular business types in the area. Thus, spatial clustering is present, but is driven by other factors than competitor's locations.

There is little publicly available research on restaurants locations in particular. According to Smith, most of the previous research *"... has been done under contract for particular restaurant franchises.."*, and thus is unavailable for academic researchers (1985).

Restaurants locations can be analyzed from both supply and demand site. Despite the fact that the decision about the location is made by the owner in the short term, it is consumer's force and decisions that influence whether particular restaurant will withstand the test of time. Consumer needs and habits are constantly changing and new behavioral patterns can be observed.

One approach to assessing customers choices in restaurants is using a multi-attribute value theory. In this framework, customers have a set of attributes that they perceive as important for making a decision. These attributes and their personal values are then compared against

traits of a business. If the assessment is positive, then a purchase is made. This approach was used in a study by Auty (1992), who applied segmentation framework to analyze restaurants choices among customers. Various studies have been made and all of them show that importance of various attributes is highly consumer-dependent. Johns and Pine (2002) provided a review of studies concerning consumer behavior in restaurant industry.

There are several studies of factors directly driving restaurants locations. Ayatac and Dokmeci (2017) examined spatial distribution of restaurants in Istanbul. In this study, data from 1997 and 2013 was analyzed. Thus, it was possible to analyze temporal dynamics. The influence of GNP per area, population density and distance from sea shore was investigated. First two factors were proven to be significant in both analyzed years. As Istanbul was rapidly developing throughout the years, some changes in spatial structure were observed, e.g. restaurants *sprawled* from CBD and historical center to less inhabited, suburban areas.

Smith (1983) analyzed the location of restaurants in Kitchener-Waterloo. He showed that restaurants locations do not depend on land values in macro scale. However, various strategies are utilized to minimize influence of high average rent in particular area- for example restaurants are located on smaller and less visited streets downtown. Also, restaurants tend to be smaller in high-rent areas compared to similar restaurants in other parts of the town. There is an evidence that regular restaurants are mainly located in CBD area, to take advantage of high daytime traffic. The decision of renting a place in a commercial building may be leveraged two way- one by avoiding big cost of owning a place, and second by attracting employees from that particular building to have lunch there. Smith also emphasizes the importance of zoning regulations as the driving factor of restaurants location decisions in Kitchener-Waterloo.

Binkley and Bales (1998) estimated average expenditure for fast food restaurants across American cities using linear model. Among the best predictors were: average fast food price, average grocery price, unemployment rate and number of fast food restaurant in the area. It should also be mentioned that population density was not found to be significant. A study of Morland et al. (2002) provides different possible reason for specific restaurants patterns. The main concern was to inspect relation between average income in the neighborhood and racial structure, and location of food stores and restaurants. They found that in lower-income areas, in south-eastern part of the USA, availability to high-quality food services is lower. The same was apparent in mostly African-American neighborhoods. Also, the quality of restaurants was

bound to average income in the areas. The results were the same also for high-quality food stores.

Studies concerning restaurants locations have lots of differences when it comes to methods and hypotheses tested. Thus, the results are rather incomparable and have a high degree of uncertainty as no verifying studies were performed. Also, as some of the above authors stated, the results obtained in one city or region should be carefully extrapolated to other areas. Each city has its own specifics, not to mention country's overall culture and its inhabitants habits.

There are few studies concerning Warsaw and Poland restaurant market. The most complex is the one made by Głuchowski, Rasińska, and Czarniecka-Skubina (2017). They show that the number of restaurants in Warsaw is constantly growing. 3 groups of customers visiting restaurants are most visible - people doing this for entertainment purposes (e.g. meeting with friends or experiencing new cuisines), tourists visiting Warsaw, and people deciding to eat outside during the workday, rather than preparing meal at home. This tendencies could be reflected in spatial distribution of restaurants in Warsaw. One could make assumptions that most restaurants will be situated in touristic district (Stare Miasto) and in business districts (Centrum, Mokotów and Wola). Similar to Pillsbury (1987), restaurants mainly for entertainment purposes will not be in one specific district, as the "journey to dine" plays a role in customer's decision. However, as Atlanta is a two times smaller city than Warsaw, one can expect that high-quality restaurants will be less likely be located in suburban areas of Warsaw.

Spatial Models Estimation

Early studies in the area of spatial phenomena did not account for spatial dependence. The usage of OLS method has been common. Later works accounted for spatial lag, however these models were still overly simplistic. In modern studies, more complex models were developed and are in use.

Even though studies on spatial regression models are advanced (for overview see LeSage 2008), there is a big gap in studies concerning spatial classification models. There are only few publications dedicated to this area. In their study Frank, Ester, and Knobbe (2009) developed spatial classification algorithm based on the concept of Voronoi tessellation. Koperski, Han, and Stefanovic (1998) improved decision tree classification algorithm to take into account spatial relations. The main novelty of this study was implementation of an existing algorithm using GIS software-specific spatial predicates. This was to improve

efficiency and velocity of model fitting and predictions. Also, some solutions were proposed to take into account spatial objects of various types (lines, points, polygons). This algorithm was also capable of using information on different levels of aggregation and feed them into decision tree estimation.

Area of advanced modeling methods is rapidly growing in recent years. Algorithms like Gradient Boosted Models, Random Forest and Support Vector Machines are state-of-the art solutions when it comes to various prediction tasks. These algorithms, however, are neglected, when it is important to understand specific process, not only making the best predictions. When it comes to explaining the decisions of algorithms, classic modeling methods like OLS and Logistic Regression are still in large use. Their main advantage, compared to more complex methods, is possibility to quantitatively assess which predictors drive particular decision.

Because of the fact that complex algorithms cope very well in real-world tasks, efforts are made to create solutions for assessing process of algorithmic decision-making. Another reason for rapid development of Explainable Artificial Intelligence (XAI) is companies' need to adjust to European GDPR regulation, specifically right to explanation of algorithm's decision (Voigt and Bussche 2017). Some of the most important frameworks and algorithms are Local Interpretable Model-Agnostic Explanations (Ribeiro, Singh, and Guestrin 2016), Partial Dependency Plots (Friedman 2001) and model-agnostic variable importance assessment (Fisher, Rudin, and Dominici 2018).

Most of the practical studies that used spatial classification are standard classification algorithms, fit to spatial data. Some of the studies do not take into account spatial dimension. An example is a geological study of landslide probability made by Goetz et al. (2015). Others do, however spatial information is assessed by a primitive method of using geographical coordinates in the model (Mascaro et al. 2014).

One of such studies is the one by Kanevski et al. (2004). It was based on environmental data, however methods developed in the paper are general and can be used in other areas. A hybrid approach using classical geostatistical tools and 2 machine learning algorithms was used. Main advantage of this method over classic statistics framework is capability of taking into account complex, non-linear relationships. At the same time, the results are still easy to interpret compared to algorithms of which this method consists, that is Artificial Neural Network and Support Vector Machine.

Khan, Ding, and Perrizo (2002) develop an efficient spatial algorithm for classification. Similar to Koperski, Han, and Stefanovic (1998), the main improvement of this study is making a already existing algorithm efficient for spatial data sets. In this approach, a problem of streaming the data and classification *on the fly* is explored.

The usage of Random Forest for spatial modeling is not widely populated. Various studies were conducted in natural sciences. Mascaro et al. (2014) analyzed the usage of Random Forest in comparison with Multiple Linear Regression for prediction of carbon mapping in Amazon Forest. They showed that using spatial context with Random Forest improved explained variance. Similarly, Čeh et al. (2018) used Random Forest and Multiple Regression for apartments prices prediction. Improvement in predictive power was also substantial. In this study, however, spatial dimension was not taken into account.

Variable Importance assessment

Variable importance in context of modeling is defined as measure to what extent is target variable dependent on considered variable. In the case of measuring influence of business and population location on presence of restaurants, variable importance can be thought of as a measure of that influence.

One broad class of assessing importance of variables is through using modeling (Wei, Lu, and Song 2015). This way it is possible to asses influence on target variable in a complex way to mimic true relationships in the data. Also these are non-parametric methods that do not require any assumptions about the underlying process (normal distribution etc.). Fulfilling these requirements are hard in real-world people' decision processes, as the decision criteria of customers are usually way more complex.

Ishwaran and others (2007) Provide a theoretical assessment of variable importance measures concerning binary regression trees and Random Forest. Louppe et al. (2013) is a large extension of work of Ishwaran. The study shows that Mean Decrease in Impurity, a standard logarithm in Random Forest assessment, is a reliable source of information about the importance of variables. Specifically, the authors prove that the algorithm satisfies basic requirement for variable importance method, that is *"[Variable importance] is equal to zero if and only if the variable is irrelevant and it depends only on the relevant variables"*.

These measures, however, have been proven to be potentially biased. Calle and Urrea (2010) provided a comparison of two Random-Forest-specific variable importance measures- Mean

Decrease Accuracy and Mean Decrease Gini. They show that the first measure is highly sensitive to small perturbations in the data set and generally should be used with caution. They prove that measure based on Gini coefficient is much more robust. Strobl et al. (2007) show that some characteristics of independent variables are favored by an algorithm, and thus a sub-optimal subset of features is chosen in the training process. This becomes an issue when there is a mix of categorical and continuous features, or when the nominal predictors vary in the number of categories. The authors propose an improved version of random forest algorithm to mitigate that problem.

Janitza, Strobl, and Boulesteix (2013) Introduce an Area-Under-Curve-based variable importance for using in random forest settings. They show that the method outperforms the two standard variable importance measures in case of highly imbalanced data sets. The results obtained in balanced classification problems are similar with all 3 methods.

The research done on the variable importance in logistic regression is not particularly broad. The existing works are only extensions of VI measures in the setting of ordinary least squares. Moreover, there is no dominating method among researchers, as it is in Mean Decrease Impurity in Random Forest setting. However, some measures have been proposed. Azen and Traxel (2009) Extended a framework of dominance analysis previously developed for linear regression by Budescu (1993). Tonidandel and LeBreton (2010) Use a concept of Relative Weights also firstly developed as a OLS tool. These works are not as widely used in practical settings as the Random Forest methods.

Grömping (2009) Show a comparison of variable importance assessment in a regression task by two algorithms, Random Forest and Linear Regression.

Recently, as more and more new algorithms are being created, there is a need to provide a model-agnostic method for assessing variable importance. Ideal for such algorithm would be to serve as a wrapper over any modeling process. A *Model Class Reliance* (MCR) algorithm provided by Fisher, Rudin, and Dominici (2018) meets these requirements. Using this method, one can easily compare the variable importances of two or more algorithms in a meaningful way. Also, a advantage compared to other metrics is that there is no need to fit the model to the data more than once. This is especially important in big data sets with many observations and variables, where model training alone, not mention cross-validation procedure, can be a resource consuming task.

The inner workings of the algorithm is as follows: first, the model is fit on all variables and a goodness-of-fit measure (AUC for example) is checked. Then, one independent variable is randomly shuffled and the goodness-of-fit is measured using perturbed variable. The process is the repeated for each variable in the data set. The Variable Importances can be perceived as the difference between the goodness-of-fit of original data set and the data sets with perturbed one variable.

Dataset description

In my study I have restricted the analysis to the Warsaw metropolis. The variables included in the dataset are:

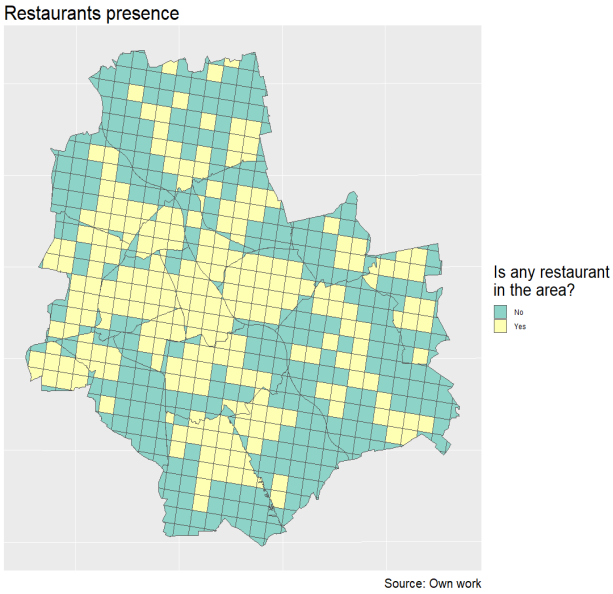
- Restaurants locations
- Businesses locations
- Population density
- Bus stops locations
- Roads locations.

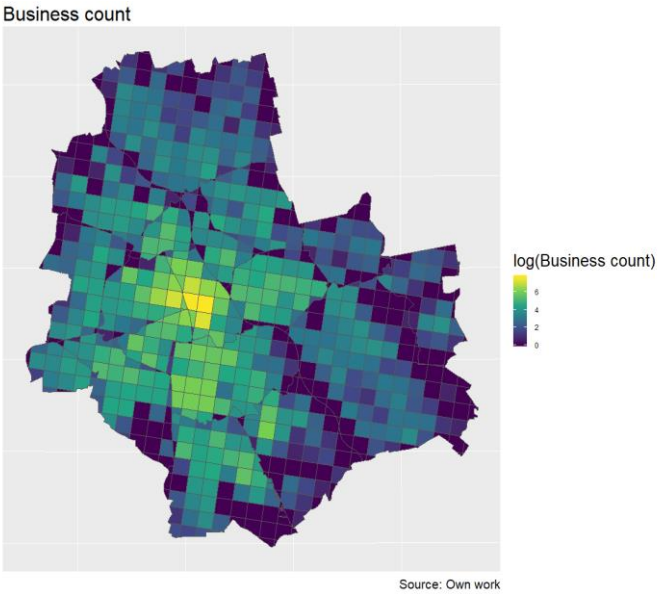
These features come from various sources. Restaurants' locations were obtained from Zomato website. There were 2341 observations total, but due to incorrect addresses, 72 restaurants were excluded. Population density comes from 2011 GUS National Census[2]. The data about businesses was gathered from (…). Location of bus stops and roads was obtained using Open Street Map service.

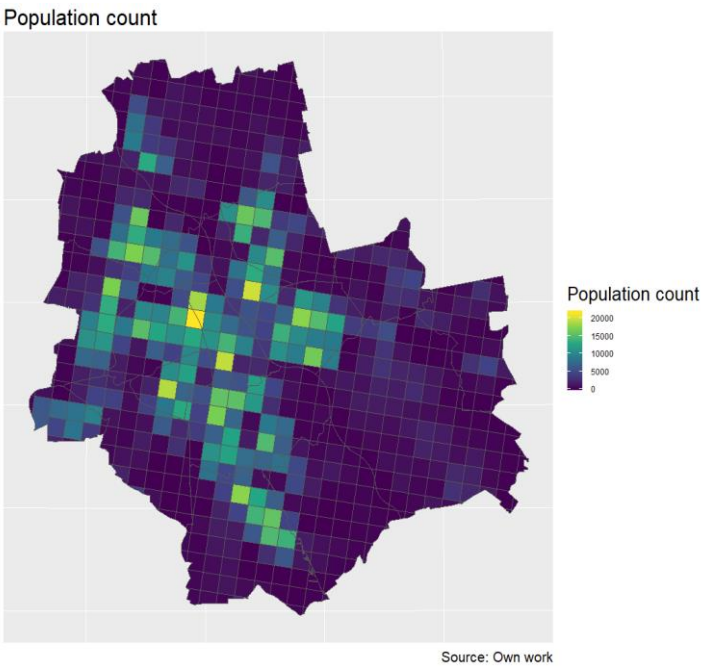
Although restaurants, businesses and infrastructural features (bus stops and roads) are points data, population density is in a form of an 1km x 1km aggregated grid. Thus, to asses population influence on the presence of restaurants, it was necessary to convert all variables to the same format. To do this, all variables were binned to a grid in the same resolution as population density data.

The map of restaurants locations shows that there exists high centrality. Also, in regions far from city center it is visible that restaurants are located in proximity to the largest streets, some of which are exit roads. Population and business presence are also highly concentrated in the city center.

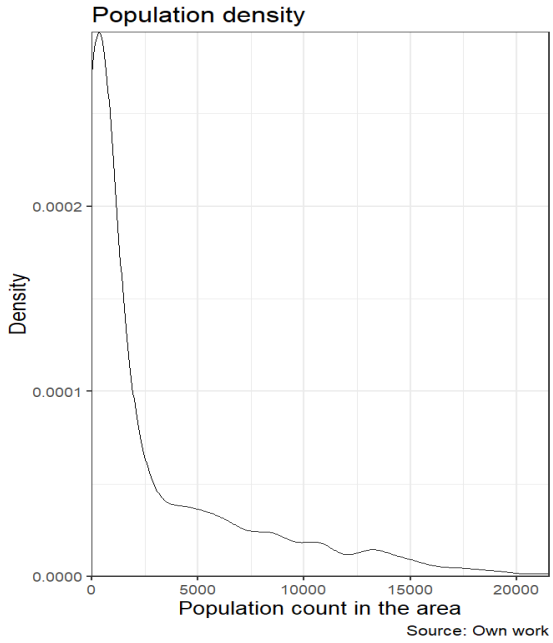
[2] https://geo.stat.gov.pl/nsp-2011

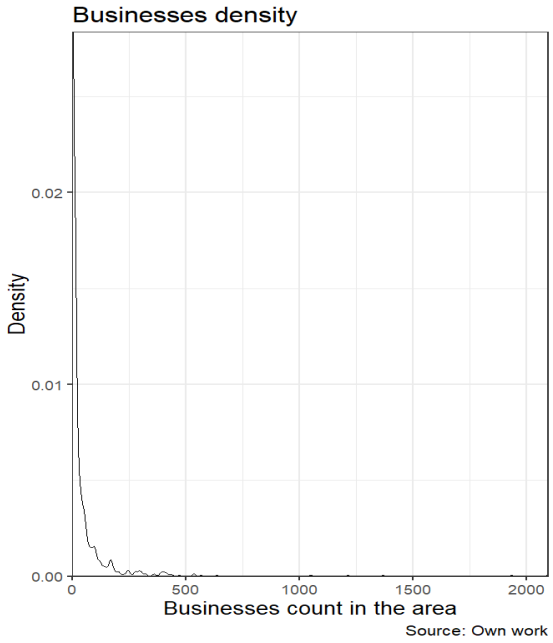




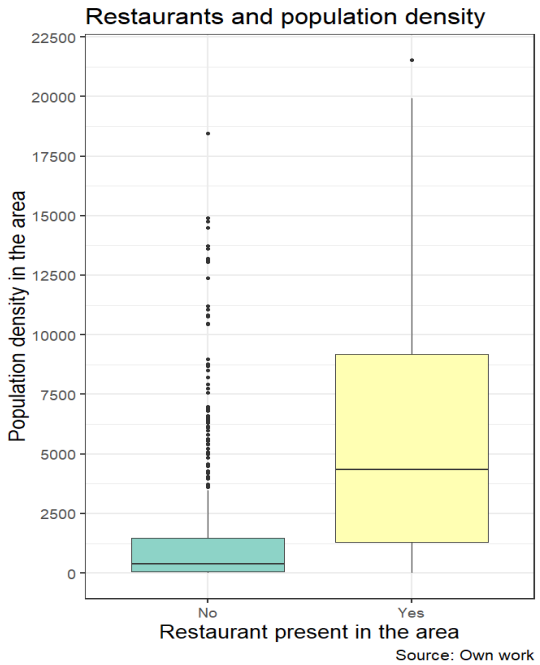


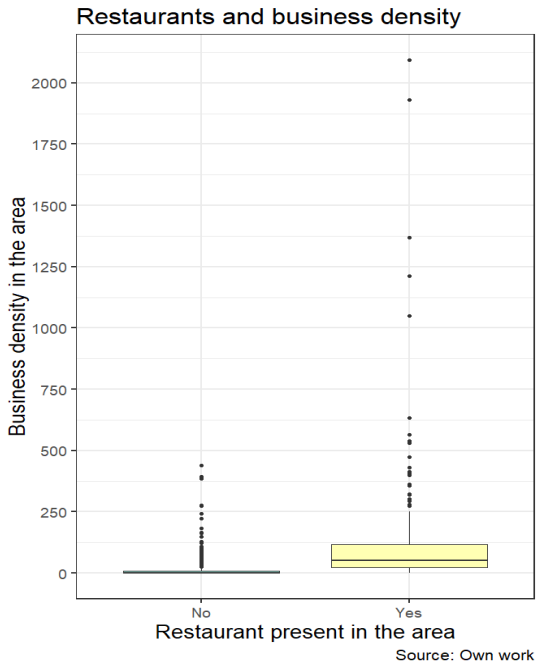
After binning the points data to a grid it can be seen that both restaurants and businesses locations distributions are highly skewed. Typical power law distribution is observed, with majority of values close to 0 and few observations with extreme values. The population density data is also highly right-skewed, but to a way lower extent than the other two variables.





As shown on the boxplots below, the subsamples containing and not containing any restaurants are significantly different in terms of business and population locations. Average business count in a grid cell in which the restaurants was present was 116.62, while in regions without restaurants was only 14.53. Similarly, average population density in restaurants' regions was 5784.74, compared to 1613.98 in regions without restaurants presence.





The join-count statistic was performed on restaurants presence data. With p-value< 0.0001, there is evidence that spatial autocorrelation in target variable exists. This means that estimates using non-spatial modelling will be biased, and there is necessity to take spatial dimension into account.

Variable	min	Q1	median	mean	Q3	max
Population count	0,00	142,75	738,00	2976,70	4295,00	21531,00
Restaurants count	0,00	0,00	0,00	2,81	1,00	228,00
Business count	0,00	0,00	7,00	47,89	39,25	2093,00
Bus stops count	0,00	0,00	3,00	5,66	8,00	54,00
Roads length	0,00	711,47	3594,25	4887,42	8151,69	20111,20

Table 1- basic descriptive statistics of used variables

Methods description

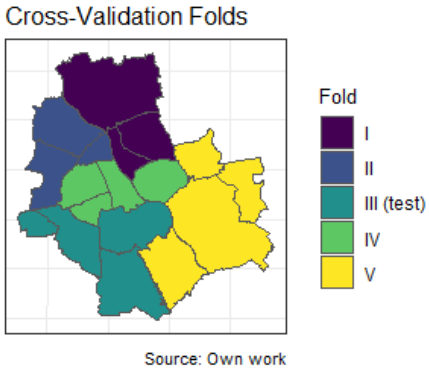
As join-count analysis on presence of restaurants shows that there exists positive spatial auto-correlation, I have taken spatial dimension into account. Neighborhood was defined with queen criterion (two areas are neighbors if they share at least on edge or vertices). For each variable (including target variable), its spatial equivalent defined as sum of this variable across neighbors was computed. This process is similar to using spatial weights matrix in Geographically Weighted Regression.[3]

I have tried 2 methods to assess whether analyzed factors are important. In the first method I have estimated a Random Forest model. To assess variable importance I have used a model-specific measure defined in an introducing publication of this model (Breiman 2001), that is Mean Decrease Gini. In the paper also Mean Decrease Accuracy is presented, however, as Calle and Urrea (2010) proved, this measure is highly dependent on small perturbations in the data set and is not stable.

[3] General data manipulation was performed using *tidyverse* (Wickham 2017) and *sf* (Pebesma 2018) packages in R CRAN software (R Core Team 2018).

In Random Forest model there is one parameter that has to be set manually, that is number of variables to randomly select during each tree fitting (*ntry*). A long-established practice for selecting the best model parameters set is using cross-validation. However, this procedure assumes that subsequent folds are independent from each other. For spatial data this possesses a problem, as choosing completely random observations could lead to leakage of information from other folds.

In my work I have used a solution proposed by Baddeley et al. (2005). He suggests that observations chosen to one fold should be densely located to minimize leakage of information from other folds. One possible implementation of this rule is dividing the space into a grid and using all observation inside one cell as one cross-validation fold. To simplify the process of splitting the space, I have used Warsaw' districts as aggregating units. Each fold consists of 3-4 districts, as shown on the map.



Source: Own work

I have created 5 folds. One of them was not included in model training and served as a test set to assess model performance on previously unseen data. Final accuracy of Random Forest model was assessed using Area Under Curve (AUC) criterion.[4]

Second method I have used is based on work of Fisher, Rudin, and Dominici (2018). The inner workings *Model Class Reliance* algorithm is as follows: first, an arbitrary model is fit on all variables and a goodness-of-fit measure (for example AUC) is checked. Then, one independent variable is randomly shuffled and the goodness-of-fit is measured using perturbed variable. The process is then repeated for each variable in the data set. The variable importances can be measured as the difference between the goodness-of-fit of original data set and the data sets with perturbed one variable. If perturbed model performs significantly worse than the original one, that means that the variable which was perturbed is important.[5]

In this setting I have used two models- one is Random Forest, the same as in the first method. I have also estimated a Logistic Regression model, as it is widely used and not overly

[4] Data modeling was performed using *caret* (Max Kuhn 2018), *randomForest* (Liaw and Wiener 2002) and *pROC* (Robin et al. 2011) packages.

[5] MCR method is implemented in *DALEX* package (Biecek 2018).

complicated. I have assessed performance of the model when each of the independent variables are perturbed. Also using the same mechanism as in the algorithm, I have tested two special situations, which are meant not to assess importance of one variable, but rather one category of variables. In first one, I have perturbed all spatially lagged variables (that is sum of particular variable in the neighboring areas). This was to check how the model is performing when spatial dimension is not included. I have also tested joint variable importance on one category of predictors. That is, in each round, I have perturbed two variables at once, non-spatial and spatially-lagged.

Results

Logistic Regression and Random Forest models both performed very well in the classification task. AUC measure on the test set was 0.825 and 0.847, respectively. This shows that presence of restaurants in the area can be easily predicted. Cross-Validation showed that the best value of *mtry* parameter for Random Forest is 3. This is consistent with Breiman (2001), who suggested setting this parameter to $\sqrt{number\ of\ variables}$.

Although both models performed equally well, the factors they took under consideration when making predictions varied. Results of Model Class Reliance for single variable and multiple variables are shown in tables 1 and 3, respectively. VI assessed by Random Forest method are show in table 2.

Population density is shown as insignificant both using Logistic Regression and Random Forest. Also spatially lagged variable is marked as not important.

Business count in the area was assessed as the most important using Random Forest with both MCR and MDG. Results concerning business count using Logistic Regression were different from Random Forest after simply using the same variables in both. However, after the business count variable was log-transformed, the results were consistent with Random Forest-based method.

Spatially lagged predictor variable, that is **count of neighboring areas in which there is a restaurant**, was an important variable. Logistic Regression use it as the most important one, and both methods with Random Forest assess it as second and third most important (using MCR and MDG approaches, respectively).

Results concerning length of roads were inconsistent among used methods. Using logistic regression with MCR, it was second best predictor. Same results were obtained from Mean Decrease Gini used with Random Forest. However, using MCR with Random Forest showed that this predictor was completely irrelevant, and excluding it from the model even improved performance. Bus stops was the least important variable using all methods.

I have also tested how two models perform with *blindfolded* spatial variables. Both random forest and logistic regression perform significantly worse. This is due to excluding spatially lagged restaurant indicator.

Both methods used showed that population density in the area is an unimportant factor when it comes to predicting presence of restaurant. It is apparent that business count in the area is the most important predictor for presence of restaurants in the area.

Variable	**Logistic Regression**	**Logistic Regression (with business count logarithm)**	**Random Forest**
	AUC (% of the full model performance)		
Full model performance	*0,825 (-)*	*0,818 (-)*	*0,847 (-)*
Businesses	0,801 (97,2%)	0,698 (85,4%)	0,737 (87,0%)
Businesses in neighbouring areas	0,825 (100,0%)	0,818 (100,0%)	0,845 (99,8%)
Population	0,822 (99,7%)	0,81 (99,0%)	0,823 (97,3%)
Population in neighbouring areas	0,824 (99,9%)	0,811 (99,2%)	0,834 (98,5%)
Roads	0,737 (89,4%)	0,815 (99,7%)	0,833 (98,4%)
Roads in neighbouring areas	0,823 (99,8%)	0,816 (99,7%)	0,837 (98,8%)
Bus stops	0,812 (98,5%)	0,812 (99,3%)	0,843 (99,6%)
Bus stops in neighbouring areas	0,828 (100,4%)	0,814 (99,6%)	0,852 (100,6%)
If restaurant in neighbouring areas	0,689 (83,6%)	0,742 (90,7%)	0,762 (90,0%)

Table 1- Single Variable Importance using MCR method

Variable	**Mean Decrease Gini (% of the most important variable)**
Businesses	87,605 (100,0%)
Businesses in neighbouring areas	16,154 (18,4%)
Population	21,221 (24,2%)
Population in neighbouring areas	16,14 (18,4%)
Roads	50,865 (58,1%)
Roads in neighbouring areas	18,7 (21,3%)
Bus stops	24,858 (28,4%)
Bus stops in neighbouring areas	13,684 (15,6%)
If restaurant in neighbouring areas	25,963 (29,6%)

Table 2- Single Variable Importance using Mean Decrease Gini method

Variable	Logistic Regression	Logistic Regression (with business count logarithm)	Random Forest
	AUC (% of the full model performance)		
Full model performance	*0,825 (-)*	*0,818 (-)*	*0,847 (-)*
Businesses	0,811 (98,3%)	0,687 (84,0%)	0,712 (84,1%)
Population	0,817 (99,1%)	0,807 (98,7%)	0,835 (98,7%)
Roads	0,753 (91,3%)	0,824 (100,7%)	0,834 (98,5%)
Bus stops	0,814 (98,7%)	0,818 (100,0%)	0,843 (99,6%)
Spatially lagged variables	0,709 (86,0%)	0,764 (93,5%)	0,69 (81,5%)

Table 3- Joint Variable Importance using MCR method

Summary

The aim of this study was to assess importance of various factors influencing restaurants locations in Warsaw. Specifically I have checked the influence of businesses locations and population density in the area. Importance of infrastructural factors, that is availability to bus stops and length of roads was assessed, too. To make the results complete, I have also included spatial dimension by including spatially lagged independent and dependent variables. To assess importance of all these factors I have used advanced Machine Learning modelling in combination with Variable Importance (VI) assessment methods.

I have used a *Model Class Reliance* (MCR) method to assess Variable Importance in model-agnostic way. Up to date there are just a few studies using this method, however it gives promising results. In addition to MCR I have used Random Forest-specific VI assessment. I have shown a comparison of these two methods. Results obtained are consistent for most variables. **There is a consensus that business location plays the most important role from factors studied. The population density was proven to be unimportant.** Also, as the presence of restaurants in the neighboring areas was proven to be an important factor, clustering tendency should be assessed more carefully. All spatially lagged independent variables were unimportant.

References

Archer, Kellie J, and Ryan V Kimes. 2008. “Empirical Characterization of Random Forest Variable Importance Measures.” *Computational Statistics & Data Analysis* 52 (4). Elsevier: 2249–60.

Auty, Susan. 1992. “Consumer Choice and Segmentation in the Restaurant Industry.” *Service Industries Journal* 12 (3). Taylor & Francis: 324–39.

Ayatac, Hatice, and Vedia Dokmeci. 2017. “Location Patterns of Restaurants in Istanbul.” *Current Urban Studies* 5 (02). Scientific Research Publishing: 202.

Azen, Razia, and Nicole Traxel. 2009. “Using Dominance Analysis to Determine Predictor Importance in Logistic Regression.” *Journal of Educational and Behavioral Statistics* 34 (3). Sage Publications Sage CA: Los Angeles, CA: 319–47.

Baddeley, Adrian, Rolf Turner, Jesper Møller, and Martin Hazelton. 2005. “Residual Analysis for Spatial Point Processes (with Discussion).” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (5). Wiley Online Library: 617–66.

Biecek, Przemyslaw. 2018. “DALEX: Explainers for Complex Predictive Models in R.” *Journal of Machine Learning Research* 19 (84): 1–5. http://jmlr.org/papers/v19/18-416.html.

Binkley, James K, and James Bales. 1998. “Demand for Fast Food Across Metropolitan Areas.” *Journal of Restaurant & Foodservice Marketing* 3 (1). Taylor & Francis: 37–50.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1). Springer: 5–32.

Breiman, Leo, and others. 2001. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” *Statistical Science* 16 (3). Institute of Mathematical Statistics: 199–231.

Budescu, David V. 1993. “Dominance Analysis: A New Approach to the Problem of Relative Importance of Predictors in Multiple Regression.” *Psychological Bulletin* 114 (3). American Psychological Association: 542.

Calle, M Luz, and Víctor Urrea. 2010. “Letter to the Editor: Stability of Random Forest Importance Measures.” *Briefings in Bioinformatics* 12 (1). Oxford University Press: 86–89.

Čeh, Marjan, Milan Kilibarda, Anka Lisec, and Branislav Bajat. 2018. “Estimating the Performance of Random Forest Versus Multiple Regression for Predicting Prices of the Apartments.” *ISPRS International Journal of Geo-Information* 7 (5). Multidisciplinary Digital Publishing Institute: 168.

Dubé, Jean, Cédric Brunelle, and Diègo Legros. 2016. “Location Theories and Business Location Decision: A Micro-Spatial Investigation in Canada.” *The Review of Regional Studies* 46 (2): 143–70.

Esteban-Bravo, Mercedes, José M Múgica, and Jose M Vidal-Sanz. 2006. “Do Business Density and Variety Determine Retail Performance?”

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2018. “All Models Are Wrong but Many Are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, Using Model Class Reliance.” *arXiv Preprint arXiv:1801.01489*.

Frank, Richard, Martin Ester, and Arno Knobbe. 2009. “A Multi-Relational Approach to Spatial Classification.” In *Proceedings of the 15th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 309–18. ACM.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*. JSTOR, 1189–1232.

Goetz, JN, Alexander Brenning, H Petschko, and P Leopold. 2015. “Evaluating Machine Learning and Statistical Prediction Techniques for Landslide Susceptibility Modeling.” *Computers & Geosciences* 81. Elsevier: 1–11.

Grömping, Ulrike. 2009. “Variable Importance Assessment in Regression: Linear Regression Versus Random Forest.” *The American Statistician* 63 (4). Taylor & Francis: 308–19.

Gunning, David. 2017. “Explainable Artificial Intelligence (Xai).” *Defense Advanced Research Projects Agency (DARPA), Nd Web*.

Głuchowski, Artur, Ewa Rasińska, and Ewa Czarniecka-Skubina. 2017. “Rynek Usług Gastronomicznych W Polsce Na Przykładzie Warszawy.” *Handel Wewnetrzny*, no. 4 (369) Tom II. Instytut Badań Rynku, Konsumpcji i Koniunktur: 118–33.

Hotelling, Harold. 1990. “Stability in Competition.” In *The Collected Economics Articles of Harold Hotelling*, 50–63. Springer.

Ishwaran, Hemant, and others. 2007. “Variable Importance in Binary Regression Trees and Forests.” *Electronic Journal of Statistics* 1. The Institute of Mathematical Statistics and the Bernoulli Society: 519–37.

Janitza, Silke, Carolin Strobl, and Anne-Laure Boulesteix. 2013. “An Auc-Based Permutation Variable Importance Measure for Random Forests.” *BMC Bioinformatics* 14 (1). BioMed Central: 119.

Johns, Nick, and Ray Pine. 2002. “Consumer Behaviour in the Food Service Industry: A Review.” *International Journal of Hospitality Management* 21 (2). Elsevier: 119–34.

Kanevski, M, Roman Parkin, Aleksey Pozdnukhov, Vadim Timonin, Michel Maignan, V Demyanov, and Stéphane Canu. 2004. “Environmental Data Mining and Modeling Based on Machine Learning Algorithms and Geostatistics.” *Environmental Modelling & Software* 19 (9). Elsevier: 845–55.

Khan, Maleq, Qin Ding, and William Perrizo. 2002. “K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees.” In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 517–28. Springer.

Koperski, Krzysztof, Jiawei Han, and Nebojsa Stefanovic. 1998. “An Efficient Two-Step Method for Classification of Spatial Data.” In *Proceedings of International Symposium on Spatial Data Handling (Sdh’98)*, 45–54.

Krugman, Paul R. 1997. *Development, Geography, and Economic Theory*. Vol. 6. MIT press.

Lee, Y, and K Koutsopoulos. 1976. “A Locational Analysis of Convenience Food Stores in Metropolitan Denver.” *The Annals of Regional Science* 10 (1). Springer: 104–17.

LeSage, James P. 2008. “An Introduction to Spatial Econometrics.” *Revue d’économie Industrielle*, no. 123. De Boeck Supérieur: 19–44.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. “Understanding Variable Importances in Forests of Randomized Trees.” In *Advances in Neural Information Processing Systems*, 431–39.

Mascaro, Joseph, Gregory P Asner, David E Knapp, Ty Kennedy-Bowdoin, Roberta E Martin, Christopher Anderson, Mark Higgins, and K Dana Chadwick. 2014. “A Tale of Two ‘Forests’: Random Forest Machine Learning Aids Tropical Forest Carbon Mapping.” *PloS One* 9 (1). Public Library of Science: e85993.

Max Kuhn, et al. 2018. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Mitchell, Lisle S, and Paul E Lovingood Jr. 1975. “Some Spatial Aspects of Public Urban Recreation in Columbia, South Carolina.” *Southeastern Geographer* 15 (2). The University of North Carolina Press: 93–101.

Morland, Kimberly, Steve Wing, Ana Diez Roux, and Charles Poole. 2002. “Neighborhood Characteristics Associated with the Location of Food Stores and Food Service Places.” *American Journal of Preventive Medicine* 22 (1). Elsevier: 23–29.

Pebesma, Edzer. 2018. “Simple Features for R: Standardized Support for Spatial Vector Data.” *The R Journal*. https://journal.r-project.org/archive/2018/RJ-2018-009/index.html.

Pillsbury, Richard. 1987. “From Hamburger Alley to Hedgerose Heights: Toward a Model of Restaurant Location Dynamics.” *The Professional Geographer* 39 (3). Taylor & Francis: 326–44.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, ca, Usa, August 13-17, 2016*, 1135–44.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. “PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves.” *BMC Bioinformatics* 12: 77.

Rolph, Inez K. 1932. “The Population Pattern in Relation to Retail Buying: As Exemplified in Baltimore.” *American Journal of Sociology* 38 (3). University of Chicago Press: 368–76.

Sadahiro, Yukio. 2000. “A Pdf-Based Analysis of the Spatial Structure of Retailing.” *GeoJournal* 52 (3). Springer: 237–52.

Smith, Stephen LJ. 1983. “Restaurants and Dining Out: Geography of a Tourism Business.” *Annals of Tourism Research* 10 (4). Elsevier: 515–49.

Smith, Stephen LJ. 1985. “Location Patterns of Urban Restaurants.” *Annals of Tourism Research* 12 (4). Elsevier: 581–602.

Stasiak, Andrzej. 2015. “Rozwój Turystyki Kulinarnej W Polsce.” In *Kultura I Turystyka–wokół Wspólnego Stołu;* Regionalna Organizacja Turystyczna Województwa Łódzkiego.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8 (1). BioMed Central: 25.

Tonidandel, Scott, and James M LeBreton. 2010. “Determining the Relative Importance of Predictors in Logistic Regression: An Extension of Relative Weight Analysis.” *Organizational Research Methods* 13 (4). Sage Publications Sage CA: Los Angeles, CA: 767–81.

Van Noort, EA, and I Reijmer. 1999. “Location Choice of Smes.” *Blees, J*.

Voigt, Paul, and Axel Von dem Bussche. 2017. “The Eu General Data Protection Regulation (Gdpr).” *A Practical Guide, 1st Ed., Cham: Springer International Publishing*. Springer.

Von Thünen, Johann Heinrich. 1875. *Der Isolirte Staat in Beziehung Auf Landwirtschaft Und Nationalökonomie*. Vol. 1. Wiegant, Hempel & Parey.

Weber, Alfred. 1929. *Theory of the Location of Industries*. University of Chicago Press.

Wei, Pengfei, Zhenzhou Lu, and Jingwen Song. 2015. “Variable Importance Analysis: A Comprehensive Review.” *Reliability Engineering & System Safety* 142. Elsevier: 399–432.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the ’Tidyverse’*. https://CRAN.R-project.org/package=tidyverse.