

University of Warsaw
Faculty of Economic Sciences

Kamil Matuszelański
Student's book no.: 387078

**Business and its neighbourhood – factors driving
restaurants locations in Warsaw**

First cycle degree thesis
specialty: Interdisciplinary Economic-Managerial Studies

The thesis written under the supervision of
dr. hab. Katarzyna Kopczewska, prof.ucz.
Department of Statistics and Econometrics
WNE UW

Warsaw, July 2019

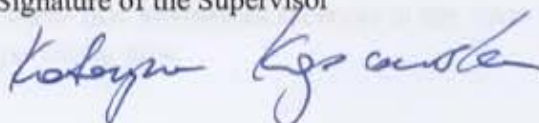
Declaration of the supervisor

I declare the following thesis project was written under my supervision and I state that the project meets all submission criteria for the procedure of academic degree award.

Date

05.06.2019

Signature of the Supervisor



Declaration of the author of the project

Aware of legal responsibility, I declare that I am the sole author of the following thesis project and that the project I submit is entirely free from any content that constitutes copyright infringement or has been acquired contrary to applicable laws and regulations.

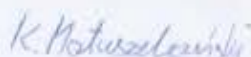
I also declare that the below project has never been subject of degree-awarding procedures in any school of higher education.

Moreover I declare that the attached version of the thesis project is identical with the enclosed electronic version.

Date

05.06.2019

Signature of the Author



Summary

This study is an attempt to find the factors mostly driving restaurants locations in the Warsaw agglomeration. Importance of population density, other businesses locations and infrastructural factors is tested. The results show that businesses location is the main factor driving restaurateurs to open a business in a particular area.

Key words

location theory, restaurants locations, variable importance, random forest

Area of study (codes according to Erasmus Subject Area Codes List)

Economics (14300)

The title of the thesis in Polish

Przedsiębiorstwo i jego otoczenie – czynniki wpływające na lokalizację restauracji w Warszawie

Table of contents

INTRODUCTION	7
CHAPTER I. Literature overview.....	8
1.1. Business location studies	8
1.2. Restaurants location studies	10
CHAPTER II. Dataset overview.....	13
CHAPTER III. Methods description.....	17
3.1. Introduction	17
3.2. Spatial classification models estimation.....	17
3.3. Variable Importance assesment.....	18
3.4. Modeling details.....	20
CHAPTER IV. Results.....	23
SUMMARY.....	27
REFERENCES	28

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.

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.

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.

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. 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.

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

CHAPTER I

Literature overview

1.1 Business location 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 (1933) 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 (1929) 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.

There are two main types of agglomeration phenomena, Marshallian and Jacobian (Panne 2004). Marshall (1890) claims that as knowledge and workforce spillovers are happening mostly inside the same or similar industry, agglomeration containing similar companies can be mostly observed. Also, strongest innovation externalities should be observed in location where this type of agglomeration is present. The second approach is a work of Jacobs (1969). She claims that the strongest knowledge spillover externalities are happening when knowledge from one industry is transformed into the other. Thus, diversity of industries in the same region promotes innovation.

Lee and Koutsopoulos (1976) tried to prove that population density does not have a significant influence on stores locations. This hypothesis is contrary to the one stated by Rolph (1932). 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. 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 related to presence of restaurants in the area.

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.

1.2. Restaurants location studies

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).

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. There is an evidence that regular restaurants are mainly located in CBD area, to take advantage of high daytime traffic. Smith 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. 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. Three 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. This effect may be compensated by good availability to a specific suburban area, both by car and public transport.

CHAPTER II

Dataset overview

This study is restricted to the analysis of the Warsaw metropolis only. There are few variables included in the dataset, coming from various sources: **Restaurants locations** - obtained from Zomato website; there were 2341 observations in total, but due to incorrect addresses, 72 restaurants were excluded., **Businesses locations** – obtained from AMADEUS database for 2017, there were 319773 observations, **Population density** - coming from 2011 GUS National Census², **Bus stops locations** and **Roads locations** - obtained using Open Street Map service.

Pillsbury (1987) shows that accessibility to an restaurant for potential customers may play an important role. However, this is not always the case, as 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. To not neglect this potential effect, two *availability features* were used, that is number of bus stops and length of roads in the area.

Results concerning population density are mixed. Both Ayatac and Dokmeci (2017) and Rolph (1932) showed a positive correlation with restaurants presence. However, Lee and Koutsopoulos (1976) proved a lack of relation.

Although restaurants, businesses and infrastructural features (bus stops and roads) are points data, population density is in a form of an 1 km x 1 km 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. This way, 808 grid cells were obtained.

Table 1 shows basic descriptive statistics of features present in the dataset. The most important finding is that all variables are highly skewed. For example, for restaurants count, more than 50% of observations are equal to 0.

The map of restaurants locations shown on figure 2 proves 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. Similarly, as shown on figure 2, population and business presence are also highly concentrated in the city center. On the figure business count is log-transformed to improve readability, as the values are highly skewed.

² <https://geo.stat.gov.pl/nsp-2011>

Table 1-Basic descriptive statistics of used variables.

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

Source: own work

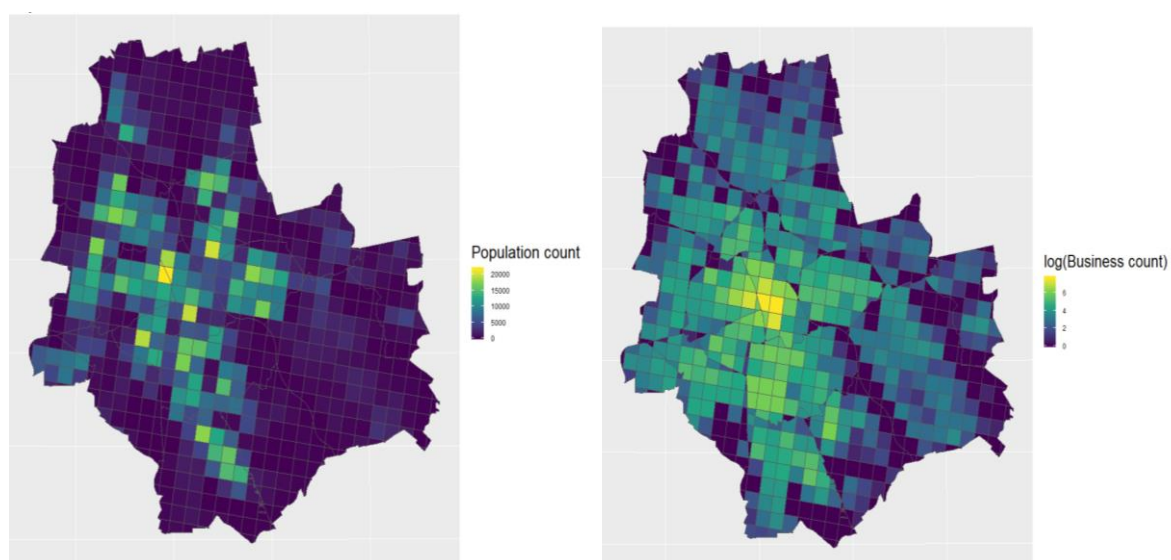


Figure 1. Population count and logarithm of business count in the areas of Warsaw.

Source: own work

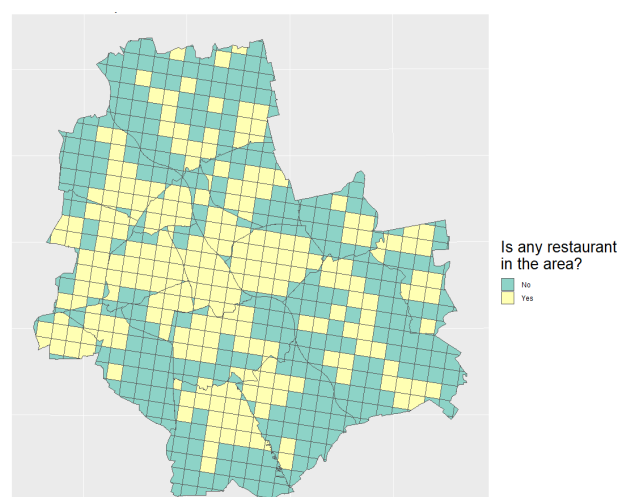


Figure 2. Presence of restaurants in the areas of Warsaw.

Source: own work

On figure 3, population and business count distribution in a binned grid is shown. As a smoothing method *Kernel Density Estimation* was used, with *gaussian* kernel and *bandwidth* equal to 1. 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.

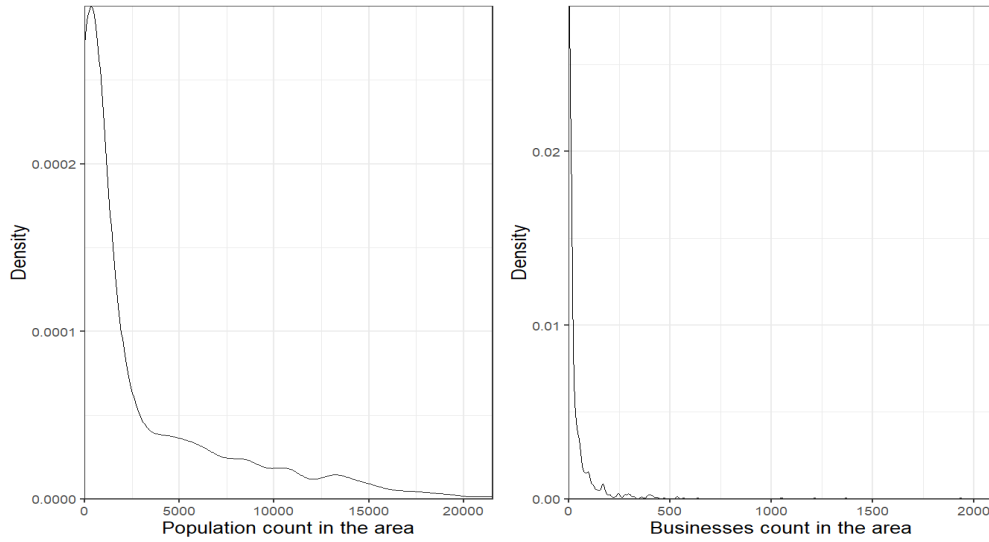


Figure 3. *Kernel Density Estimation* of population count and business count in the areas of Warsaw.

Source: own work

As shown on the figure 4, the subsamples containing and not containing any restaurants are significantly different in terms of business and population locations. The boxplots were obtained from 808 grid cells. 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.

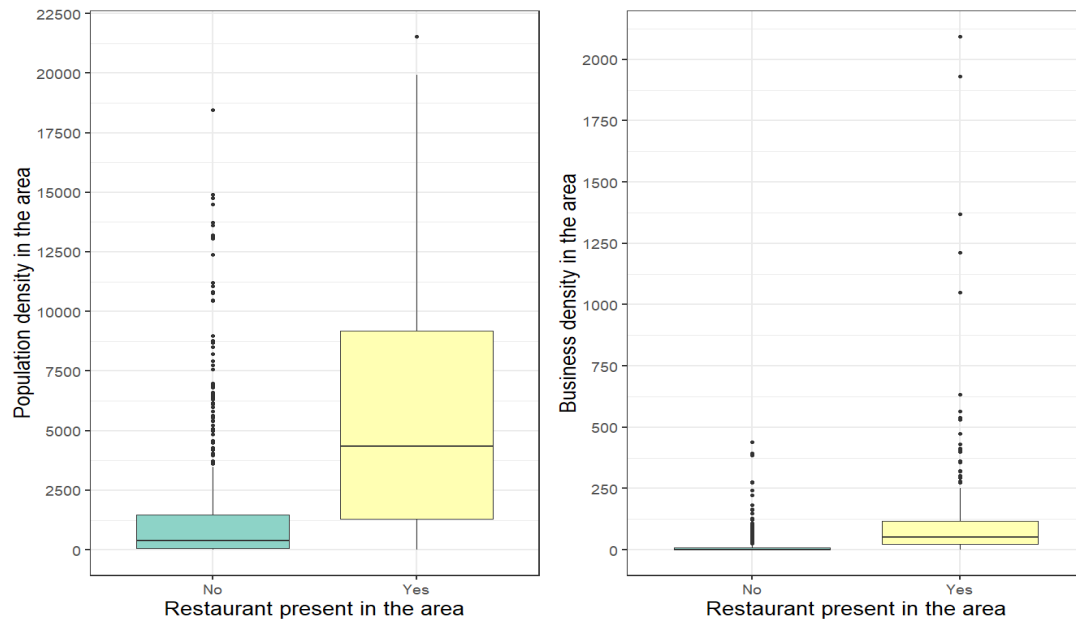


Figure 4. Population and business count in the areas of Warsaw grouped by restaurant presence.

Source: own work

The join-count statistic was performed on restaurants presence data. With $p\text{-value} < 0.0001$, 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. As of independent variables used, there is no empirical study that show a spatial interdependence between restaurants presence and such features. However, including spatially lagged dependent variable is in accordance with works of Pillsbury (1987), Smith (1985) and Dubé, Brunelle, and Legros (2016), who all showed a clustering tendency in restaurants.

CHAPTER III

Methods description

3.1. Introduction

There are various methods to check if one variable is dependent on the other. Probably the most common ones are Pearson's correlation and tests for variable significance in Linear Regression modeling. These approaches are useful only when specific and rigorous requirements are made, for example linearity of a relationship and continuous dependent variable.

One broad class of assessing importance of one variable on another is through using modeling (Wei, Lu, and Song 2015). This way it is possible to assess 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's decision processes, as the decision criteria of customers are usually way more complex.

The data used in this study has two important characteristics that should determine the usage of a concrete method. First, spatially lagged variables should be taken into account. Second, the task is classification, not regression. This prevents from using standard spatial methods like *Geographically Weighted Regression*.

3.2. Spatial classification models estimation

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. Some of the proposed methods are meant not to improve comprehension of spatial phenomena, but rather to improve efficiency or enable usage of spatial data in some contexts.

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.

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

As of approaches taking into account phenomena like *spatial lag*, there are few algorithms created to fill that need. In their study Frank, Ester and Knobbe (2009) developed spatial classification algorithm based on the concept of *Voronoi tessellation*. Kanevski et al. (2004) used a hybrid approach using classical geostatistical tools and two machine learning algorithms. Main advantage of this method over classic statistics framework is capability of taking into account complex, non-linear spatial 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* (ANN) and *Support Vector Machine* (SVM).

In empirical studies, a common practice is to use standard classification algorithms, and fit them to spatial data. Some of the studies do not take into account spatial dimension (Goetz et al., 2015).

Concerning usage of specific algorithms for spatial modeling, *Random Forest* is 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.

The usage of *Random Forest* in context of restaurants presence prediction is justified from various reasons. First, it is suitable for classification tasks. Second, as stated above, this algorithm was previously successfully used in spatial context. Third, it is able to take into account complex nonlinear relations. Also, there exists a widely used method for assessment of *Variable Importance* in this algorithm, that is *Mean Decrease Gini* defined in an introducing publication of this model (Breiman 2001).

3.3. Variable importance assessment

As none of the methods previously presented are used in specific setting like in this study (classification task based on location theory), the method of *Random Forest* should be assessed carefully.

Ishwaran et al. (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 et al. (2007). The study shows that *Mean Decrease Impurity*, a standard Variable Importance assessment measure in *Random Forest* method, is a reliable source of information about the importance of variables.

These measures have been proven to be potentially biased in some specific settings. Calle and Urrea (2010) provided a comparison of two *Random-Forest*-specific variable importance measures - *Mean Decrease Accuracy* and *Mean Decrease Gini (Impurity)*. 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. However, in the settings of this study, there are no categorical independent variables, so the usage of an improved method is not justified.

As there are no comparable studies in area of restaurants locations using algorithmic data modeling, results obtained from *Random Forest* method with *Mean Decrease Gini* method should be checked somehow. I have decided to also use different algorithm to predict restaurants presence and compare the results of VI assessment. Specifically, I have used *Logistic Regression*, as it is a vastly different model than RF.

The problem with *logistic regression* is that the research done on the variable importance with this algorithm 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.

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 obtain and compare the Variable Importances of two or more algorithms in a meaningful way. An advantage of this algorithm compared to other metrics is that there is no need to fit the model to the data more than once. This is however not so important in this study as the data is relatively small and model training is not so time-consuming.

The inner procedure of *Model Class Reliance* algorithm is as follows: first, an arbitrary model f is fitted on all variables x_1, \dots, x_i to predict $Pr(y = 1|x_1, \dots, x_i)$. Then, goodness-of-fit measure (for example AUC) is calculated using the same set of variables.

In the next step of an algorithm, independent variable x_1 is randomly shuffled (randomisation of order) to obtain variable x'_1 and the goodness-of-fit AUC_{x_1} is measured using a model f on variables x'_1, \dots, x_i . The process is then repeated for each variable in the data set.

When the variable x_i is being shuffled, the model still uses this variable. However, for each observation this variable will be wrong. If the model firmly relies on this variable, the results obtained will be also completely wrong. Thus, if the AUC measure drops significantly using perturbed variable x_i , it is clear that this variable is important.

In this setting, the smaller AUC_{x_i} , the more important x_i is. This is somewhat non-intuitive, as usually the goal is to maximize AUC . The same is valid for variable importance measures, that is - the more important the variable is, the bigger the measure (for example Mean Decrease Gini in Random Forest method) will be. Thus, it is more convenient to use a measure $AUC - AUC_{x_1}$, as the direction of changes is than the same as in other VI assessment methods.³

3.4 Modeling details

As join-count analysis on presence of restaurants shows that there exists positive spatial auto-correlation, study takes 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*.⁴

³ MCR method is implemented in *DALEX* package (Biecek 2018).

⁴ General data manipulation was performed using *tidyverse* (Wickham 2017) and *sf* (Pebesma 2018) packages in R CRAN software (R Core Team 2018).

As the *Logistic Regression* method does not handle non-linear relationships very well, and the business count variable was shown to be most highly skewed, I have estimated two models, one with all variables included as-is, and another one, in which I have log-transformed business count variable

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 (*mtry*). 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.

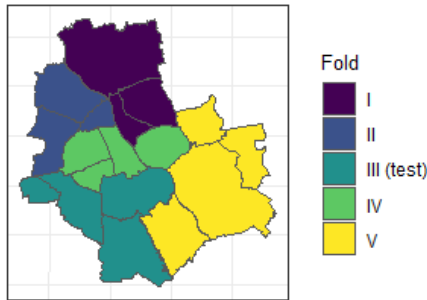


Figure 5. Cross- validation folds.
Source: own work

This work applies 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 figure 5.

For cross-validation purposes, five folds were created. 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.⁵ *AUC* is a widely used measure of goodness-of-fit for classification tasks. Specifically, it is an area under the *Receiver Operating Statistic (ROC)* curve. It is created by plotting *true positive rate (TPR)* versus *false positive rate (FPR)* of predictions. *AUC* has its values in a range between 0 and 1. A model is better than random guessing if $AUC > 0,5$, thus, for model perfectly fit to the data $AUC = 1$.

In the setting of *Model Class Reliance* algorithm 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

⁵ Data modeling was performed using *caret* (Max Kuhn 2018), *randomForest* (Liaw and Wiener 2002) and *pROC* (Robin et al. 2011) packages for R CRAN software.

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.

CHAPTER IV

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 $\sqrt{\text{number of variables}}$.

Table 2- Single Variable Importance using MCR method

Dependent variable: Is restaurant present in the area?	Logistic Regression	Logistic Regression with business count logarithm	Random Forest
Explanatory variables	AUC decrease compared to full model (% of the full model performance)		
Businesses	0,024 (97,2%)	0,12 (85,4%)	0,11 (87,0%)
Businesses in neighbouring areas	0 (100,0%)	0 (100,0%)	0,002 (99,8%)
Population	0,003 (99,7%)	0,008 99,0%)	0,024 (97,3%)
Population in heighbouring areas	0,001 (99,9%)	0,007 (99,2%)	0,013 (98,5%)
Roads	0,088 (89,4%)	0,003 (99,7%)	0,014 (98,4%)
Roads in heighbouring areas	0,002 (99,8%)	0,002 (99,7%)	0,01 (98,8%)
Bus stops	0,013 (98,5%)	0,006 (99,3%)	0,004 (99,6%)
Bus stops in heighbouring areas	-0,003 (100,4%)	0,004 (99,6%)	-0,005 (100,6%)
If restaurant in neighbouring areas	0,136 (83,6%)	0,076 (90,7%)	0,085 (90,0%)

Source: own work

In table 2, *Model Class Reliance* method results are presented. The value for businesses count explanatory variable for *Logistic Regression* (second column) is interpreted as follows: when the model was not able to take into consideration businesses count variable, accuracy measured by AUC dropped by 0,024. This means that model with removed businesses count variable achieves 97,2% performance of the full model (this value is shown in parentheses). In third column there is also *Logistic Regression* model, but with log-transformed businesses count variable. As the percentage of full model performance drops significantly from standard Logistic Regression (85,4% compared to 97,2%), this means that logistic regression was not able to handle highly skewed businesses count variable. However, after log-transformation, this variable is shown to be the most important one. Value for businesses count for *Random Forest* model is 0,11. This is 87% of the full model performance. From that, one can conclude that *Random Forest* method uses businesses count extensively, as this variable achieves biggest drop for all variables in *Random Forest* method (fourth column). The drop in predictive power is bigger than in *Logistic Regression* without any variables transformations.

However, when businesses count variable is log-transformed, results are in line with *Random Forest* method.

Table 3 shows *Model Class Reliance* method results for grouped variables. The interpretation is similar to the one for single variable (as explained above). For example, for businesses variables group, the models were not able to use businesses count for specific area and businesses count for neighbouring areas. This is the case with first 4 groups of variables (businesses, population, roads and bus stops). The group “spatially lagged variables” is specifically created to assess whether there is a need to take into account spatial dimension.

Table 3- *Joint Variable Importance* using *MCR* method

Dependent variable: Is restaurant present in the area?	Logistic Regression	Logistic Regression with business count logarithm	Random Forest
Explanatory variables	AUC decrease compared to full model (% of the full model performance)		
Businesses	0,014 (98,3%)	0,131 (84,0%)	0,135 (84,1%)
Population	0,008 (99,1%)	0,011 (98,7%)	0,012 (98,7%)
Roads	0,072 (91,3%)	-0,006 (100,7%)	0,013 (98,5%)
Bus stops	0,011 (98,7%)	0 (100,0%)	0,004 (99,6%)
Spatially lagged variables	0,116 (86,0%)	0,054 (93,5%)	0,157 81,5%)

Source: own work

Table 4 presents results of standard *Variable Importance* measure for *Random Forest*, that is *Mean Decrease Gini*. The bigger the value for specific variable, the more important the variable is. For example, businesses count variable has biggest *MDG*, which means that is the most important one. For easier comparison between variables, percentage of *MDG* compared to the one with biggest value are shown in parentheses. For example, population density variable has importance of only 24,2% when compared to businesses count variable.

Table 4- *Single Variable Importance* using *Mean Decrease Gini* method

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

Source: own work

Although both models performed equally well, the factors they took under consideration when making predictions varied.

Business count in the area was assessed as the most important using *Random Forest* with both *MCR* and *MDG* methods. 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.

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

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 which were 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.

Proven importance of spatially lagged independent variable is in accordance with the results of studies made by Pillsbury (1987), Lee and Koutsopoulos (1976) and Smith (1985). They all showed that spatial agglomeration plays an important role in restaurants location. Results concerning population density in the area are however not so well aligned with previous research. Research made by Rolph (1932), Smith (1985) and Ayatac and Dokmeci (2017) showed an importance of this factor. The differences in the results may come from two sources. One possibility is that as previous studies did not take businesses location into account, the results may suffer from omitted-variable bias. Another is a specification of the data in the studies of Rolph and Ayatac and Dokmeci. Large aggregation (to city districts)

was present. Thus, results presented may be biased, as stated by Briant (2010). Study presented in this paper has much lower level of aggregation, and thus should provide more accurate results.

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 businesses location plays the most important role from factors studied.** Up to date, there is no other study meant to assess this factor for restaurants locations. Similar factor, that is proximity to Central Business District was however shown to be important by Smith (1985) and Ayatac and Dokmeci (2017). **The population density was proven to be unimportant.** This is contrary to results of Rolph (1932), Smith (1985) and Ayatac and Dokmeci (2017). Also, the **presence of restaurants in the neighboring areas was proven to be an important factor**, as previously shown in the works of Pillsbury (1987), Lee and Koutsopoulos (1976) and Smith (1985).

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.
- Briant, Anthony, P-P. Combes, and Miren Lafourcade. "Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?." *Journal of Urban Economics* 67.3 (2010): 287-302.
- 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.
- Christaller, Walter. 1933. "Die Zentralen Orte in Süddeutschland (the Central Places in Southern Germany)." Jena: Gustav Fischer.

- 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 Wewnętrzny*, no. 4 (369) Tom II. Instytut Badań Rynku, Konsumpcji i Koniunktur: 118–33.
- Harold, Hotelling. 1929. "Stability in Competition." *Economic Journal* 39 (153): 41–57.
- 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.
- Jacobs, J., 1969. *The Economies of Cities*. Random House, New York.
- 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.
- Marshall, A., 1890. *Principles of Economics*. MacMillan, London.
- 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.” Bles, 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.