Thesis

#### 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”* (Krugman 1997). However, some theories of location were developed through the years. First approaches in the stream of classical economy concentrated on industry and agricuture. The earliest theory concerning location is by von Thunen. (…) His model of agricultural land layed (?) foundations for later works. Theory of industrial location made by Weber (…) concentrated on transportation costs of raw materials and final products. According to the theory, entrepreneurs created their industrial sites in places where the cost of transportation was the lowest.

Later 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. This assumption was valid in pred- industrial era.

Hotelling’s linear city model (…) is on 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 as the competitors. This phenomenon is visible in retail market, especially bars, restaurants and pharmacies (…).

Hotelling’s theory was then expanded by Salop (…). Main difference was that instead of a strainght line, the street along which businesses competed was a circle with no-end points.

*empirycznie czynniki wpływające na industry i retail*

#### Restaurants location studies

*tutaj cały wstęp teoretyczny*

There is little publically avaliable research on restaurants location. According to Smith, most of the previous research *… has been done under contract for particular restaurant franchises..*, and thus is unavaliable for academic researchers (Smith 1985).

Retail location theory is particularly broad. …..

There are 2 main streams of restaruants locations studies. One is mainly concentrated on clustering tendencies in restaurants locations. As proven in numerous studies, (…), various spatial phenomena have been shown to cluster well. This is particularly apparent in locations of some bussinesses categories, as shown by (…).

(Pillsbury 1987) Showed that restaurants locations have high clustering tendency. As claimed, *Today, virtually no new restaurant is found outside a cluster of its competitors*. Moreover, restaurants clustering criteria used (socio-economics, ambiance and accessibility) were reflected in restaurants locations. Novelty of this study was not to 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 (eg. soul food) there is no need for good avaliability, and *journey to dine* becomes an integral part of the dining experience.

(Smith 1985) Claim that some restaurants categories (mostly fast food) are highly clustered. Another finding was that population density is exponentialy related to traffic volume in the area. Higher raffic was also correlated with presence of some types of restaurants.

(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.

(Morland et al. 2002) Examinated location of food stores and services in south-eastern part of the USA. The main concern of this study was to inspect relation between average income in the neighbourhood and racial structure, and location of food stores and restaurants. They found that in lower-income areas avalioability to high-quality food stores is lower. The same was apparent in mostly black neighbourhoods. Also, the quality of restaurants was bound to average income in the areas.

(Ayatac and Dokmeci 2017) Examinated spatial distribution of restaurants in Istanbul. In this study, data from 1997 and 2013 were 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, eg. restaurants *sprawled* from CBD and historical center to less habitated, suburban areas.

#### Estimation of Complex Spatial Models

Early studies in the area of spatial phenomenas did not account for spatial dependence. The usage of OLS 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 (…), there is a big gap in studies concerning development of spatial classification models. In a canonical book in terms of spatial econometrix, LeSage did not even mention classification task. There are only few publications dedicated to this area. In his study <http://www.cs.sfu.ca/~ester/papers/KDD-2009-SpatialClassification.final.pdf> (…) developed spatial classification algorithm based on the concept of Voronoi tasselation (…). In his work (<https://pdfs.semanticscholar.org/c9e1/0cf4006690e6f3a3c05a151515d0c5a8ca6d.pdf>) improved decision tree classification algorithm to take into account spatial relations. The main novelty of this study was implementation of this 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 the data of various types (lines, points, polygons). This algorithm is also capable of using information on different levels of aggregation and feed them into decision tree estimation.

Area of advanced non-parametric and machine learning methods is rapidly growing in recent years. Algorithms like Gradient Boosted Models, Random Forest and Support Vacor Machines are state-of-the art solutions when it comes to various prediction tasks. Second concern is usage of spatial dimension in

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 quantitavely assess which predictors drive particular decision.

However, because of the fact that complex algorithms cope very well in real-world tasks, efforts are made to create solutions for assesing process of algorithmic decision-making. Another reason for rapid development of Explainable Artificial Intelligence is companies’ need to adjust to European GDPR regulation, specifically right to explanation of algorithm’s decision (<https://www.privacy-regulation.eu/en/r71.htm>). 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 assesment(Fisher, Rudin, and Dominici 2018).

This trend creates opportunities in research. As new tools are developed, relying only on linear models is getting less reasonable.

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 assesed by a primitive method of using geographical coordinates in the model (… lasy?).

However, the usage of Random Forest for spatial modeling is not widely populated. Various studies were conducted in natural sciences. (…) 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 variation by 16%.

Similarly, a (..) study, (Čeh et al. 2018) used Ranom Forest and Multiple Regression for apartaments prices prediction. Using the first method, improvement in prediction measured by R^2 was 0.34. In this study, however, spatial dimension was not taken into account. i

#### Ogólny opis zadania

Hipotezy

The goal of this study is to find out what are main factors that drive restaurators’ decisions about whether to open a business in a given location. Specifically, I have tested 2 main hypotheses. One is the influence of business proximity. Importance of this factor would mean that main clients of the restaurants are mainly people during their lunch breaks. The other hypothesis is the influence of population density. This would mean that people go to restaurants in their homes’ proximity in the evenings.

To asses thoroughly the importance of the to factors stated above, I have also included other variables of interest, that can also indirectly influence the restaurants locations. These are proximity of bus stops and total length of roads. These two can be seen as measure of how well is the potential location connected to other parts of the city.

One broad class of assesing importance of variables … is through using modeling. 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 distributions etc.). Fulfilling these requirements are hard in real-world people’ decission processes, as the decision criteria are usually way more complex.

Variable importance

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 f restaurants, variable importance can be thought of as a measure of that influence.

#### Opis 3 metod

In my work I have tried 3 methods to asses whether these predictiors are imporant.

Random Forest is a long-established model in classification tasks. During model training, various decision trees are created. Each decision tree is fitted using 1. different subset of observations (obtained from full data set using bootstraping) and 2. different subset of features (variables). This way model overfitting compared to classic decision tree is largely reduced.

I have chosen random forest for two reasons. First, it is a well established model performing well in various prediction tasks (…). Secondly, assesing variable importance is straightforward and model dependent.

Variable importance in RF models was defined in a introducing publication of this model (…). At each of the splits in training phase, the variable on which to make a split is chosen using Gini Impurity criterion. Importance of given variable is defined as sum of decrease in Gini Impurity in all the splits, in which the variable was used.

I have estimated the model on training data described in section (…). I have used *caret* package as wrapper over randomForest method. In each cross validation phase I have tried different number of predictiors used during each tree fitting. Final accuracy of the model was assesed using AUC criterion, as the target variable is slightly imbalanced (In the training set the share of observations without restaurant is ~67%).

In addition to a model with all variables, 3 other Random Forest models were fitted. First one did not include information about business locations. This means that 2 variables were ommited, business count in the area and business count in surrounding areas (containing spatial dimension). Second model was similar, but did not include information about population density. Similarly, 2 variables were ommited. The third model served as a baseline. All variables that are of interest in this study (business count and population density) were ommited. The variables selected were bus stops count and length of roads and their spatial equvalents (these variables summed across all neighbours of target area). To adress spatial autocorrelation fully, spatially lagged target variable was also included (defined as sum of neighbouring areas, in which restaurants were present).

4 models described above were then used to make predictions on previously held out data. Accuracy measured by AUC was reported. Models with all variables and the one without the most interesting variables were treated as upper- and lower-bound for assesing prediction accuracy. Assesing variable importance in this setting is straightforward.

The method used is as follows. Let us consider a classification task, with a set of i (i>2) predictors and x\_i and target variable y. The goal is to asses which variable, x\_1 or x\_2, has bigger influence on the target variable y. We can assume that the best model will be the one containing all variables. Also, the poorest prediction will be given by the model containing all variables except x\_1 and x\_2. Then, models including x\_1 only and x\_2 only will have accuracy somewhere between model 1 and 2.

I have performed similar procedure, but using Logistic Regression as a classifier. The reason is that it is a linear model simple to estimate and use. The predictions are also highly explainable, and thus it is possible to achieve reasoning for every classification easily.

Finally, the last method to asses variable performance is a modified version of my second approach. Similarly, I have estimated 8 models total (4 using Random Forest algorithm and 4 using logistic regression). The difference is in assesing performance of the models. Instead of straightforward AUC comparison on held-out data, I have used resampling method described by (…) and (…caret). First, the models are estimated on training data. Then, instead of using full test set, the observations are sampled with replacement and AUC is computed for each sample set. This way, instead of point estimate of AUC for a given model, one can get multiple values of AUC and obtain estimate of distribution for the results. This method is a improvement compared to simple AUC estimation because it enables to perform statistical inferences on the results. For example, knowing the vector of AUC estimates for two models, one can use t-test to asses equality of estimates between given models, and thus infer about significance of the difference.

#### Dataset description

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

* Restaruants locations
* Businesses locations
* Population density
* Bus stops locations
* Roads loactions.

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 (<https://geo.stat.gov.pl/nsp-2011>). 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 neccesary 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 centre 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 concetrated in the city centre. 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 (…), the subsamples containig 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 modeling will be biased, and there is neccessity to take spatial dimension into account.

#### Spatial weights matrix

As join-count analysis on presence of restaurants shows that there exists poistive spatial autocorrelation, spatial dimension was taken into account. Neighbourhood was defined with queen criterion, which means that two areas are neighbours if they share at least on edge or vartice. For each variable (including target variable), its spatial equivalent defined as sum of this variable across neighbours was computed. This process is similar to using spatial weights matrix in Geographically Weighted Regression. The reason to include spatial dimension explicitly by adding new variables to the model is due to specification of implementation of modeling algorithms in R software. Aspecially in more advanced models (like Random Forest), adding spatial weights matrix would require rewriting the whole method.

#### Cross validation

Good practice for estimating machine mearning models is using cross-validation. However, this procedure assumes that subsequent folds are independent from each other. For spatial data this posesses a problem, as choosing completely random observations could lead to leakage of information from other folds.

The solution was proposed by (…). He suggests that observations chosen to one fold should be densely located to minimise leakage of information from other folds.

I have used similar approach. I have created (…) folds. 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 map (…)

The choice of districts to include in one fold was arbitrary, the main criterion was to ensure that folds have similar area and the districts inside share borders with each other.

#### Dump

This process can negatively influence the predictions, but as the grid resolution is good enough, this probably will not have big effect on the results of analysis.

The way I approached this was through model estimation- more important factor is the one, on which basis one can make best prediction about presence of restaurants. Because of the fact that in most grid cells the number of restaurants was 0 (~67%), I have decided to perform classification rather than regression. The target variable is then presence of at least one restaurant in given area.

#### References

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

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.

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.

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*.

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.

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

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.

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.

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.

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