CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Spatial Analysis of Czech Parliamentary Election: Comparison of Spatial Econometrics and Machine Learning

Master's thesis

Author: Bc. Jakub Černý

Study program: Economics

Supervisor: PhDr. Lenka Šťastná Ph.D.

Year of defense: 2022

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, August 1, 2022

Jakub Černý

Abstract

This thesis analyses the results of the Czech Parliamentary election in 2021 and attempts to explain the voting support of major political subjects by using aggregate data from Czech municipalities. Since the data evince spatial autocorrelation, it is necessary to specify a suitable spatial model. The thesis provides both empirical and economic evidence for the Spatial Durbin Error Model, which enables distinguishing the direct and indirect effects of particular independent variables and accounts for the spatial dependence of error terms. This method shows that variables describing the socio-economic characteristics of inhabitants, such as the share of entrepreneurs or people with university education, play the most significant role in explaining voting results and evince mostly the direct effects. On the contrary, variables describing municipalities, such as public spending or infrastructure, are more likely to impact the election result indirectly. Subsequently, the analysis is replicated using two tree-based machine learning algorithms and all models are evaluated based on their ability to predict the election results from unseen data. Even though machine learning methods estimate only relative variable importance instead of standard coefficients, this approach represents a perspective complement to the established field of spatial analyses.

JEL Classification	C01, C31, B55, R19
Keywords	spatial analysis, parliamentary election, spatial durbin error, machine learning, random forest
Title	Spatial Analysis of Czech Parliamentary Election: Comparison of Spatial Econometrics and Machine Learning

Abstrakt

Práce analyzuje výsledky voleb do Poslanecké sněmovny Parlamentu ČR z roku 2021 a snaží se vysvětlit voličskou podporu významných politických subjektů za použití agregovaných dat z obcí ČR. Protože data vykazují prostorovou autokorelaci, je nutné k analýze použít vhodný prostorový model. Práce poskytuje empirické i ekonomické důkazy ve prospěch Spatial Durbin Error modelu, který umožňuje rozlišovat přímé a nepřímé efekty jednotlivých proměnných a bere v potaz i prostorovou závislost reziduí. Tato metoda ukazuje, že proměnné popisující socio-ekonomickou úroveň obyvatelstva, jako např. podíl podnikatelů nebo lidí s vysokoškolským vzděláním, hrají důležitou roli při vysvětlování volebních výsledků a převážně vykazují přímé efekty. Naopak, proměnné obecně popisující obce, jako např. veřejné výdaje nebo úroveň infrastruktury, ovlivňují volební výsledky spíše nepřímo. Následně je analýza replikována pomocí dvou algoritmů strojového učení na principu rozhodovacích stromů a všechny modely jsou porovnány na základě jejich schopnosti předpovídat volební výsledek z neznámých dat. Navzdory skutečnosti, že metody strojového učení neodhadují koeficienty k jednotlivým proměnným, ale pouze jejich relativní důležitost, představuje tento přístup perspektivní doplněk k oboru prostorových analýz.

Klasifikace JEL	C01, C31, B55, R19
Klíčová slova	prostorová analýza, parlamentní volby, spatial durbin error, strojové učení, náhodné lesy
Název práce	Prostorová analýza parlamentních voleb v ČR: Srovnání prostorové ekonometrie a strojového učení

Acknowledgments

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Černý, Jakub: *Spatial Analysis of Czech Parliamentary Election: Comparison of Spatial Econometrics and Machine Learning.* Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2022, pages 107. Advisor: PhDr. Lenka Šťastná Ph.D.

Contents

List of Tables

List of Figures

Acronyms

AIC	Akaike Information Criterion
BMA	Bayesian Model Averaging
CART	Classification and Regression Trees
CZSO	Czech Statistical Office
DT	Decision Trees
GWR	Geographically Weighted Regression
GRF	Geographical Random Forest
HC	Hierarchical Clustering
kNN	k-nearest neighbours
LISA	Local Indicators of Spatial Association
LM	Lagrange Multiplier
LR	Logistic Regression
LL	log-likelihood
ML	machine learning
MLE	maximum likelihood estimation
MSE	mean squared error
NN	Neural Networks
OLS	Ordinary Least Squares
OOB	out-of-bag
PIP	posterior inclusion probabilities
RF	Random Forest
SDEM	Spatial Durbin Error Model
SDM	Spatial Durbin Model
SEM	Spatial Error Model

SLM	Spatial Lag Model
SLX	Spatially Lagged X Model
SVM	Supported Vector Machines
VIF	variance inflation factor
WLS	Weighted Least Squares

ANO	ANO 2011 - Movement of Andrej Babiš
CSSD	Czech Social Democratic Party
KDU-CSL	Christian and Democratic Union – Czechoslovak People's Party
KSCM	Communist Party of Bohemia and Moravia
ODS	Civic Democratic Party
Pirati	Czech Pirate Party
Pirati+STAN	Pirates and Mayors
PRISAHA	Robert Šlachta's Civic Movement
STAN	Mayors and Independents
SPD	Freedom and Direct Democracy
SPOLU	Together – Civic Democracy
TOP 09	Tradition, Responsibility, Prosperity

Master's Thesis Proposal

Author	Bc. Jakub Černý
Supervisor	PhDr. Lenka Šťastná Ph.D.
Proposed topic	Spatial Analysis of Czech Parliamentary Election: Comparison of Spatial Econometrics and Machine Learning

Motivation Elections are an interesting subject of research and there exist various branches of their study, for instance trying to explain electoral support of political parties or uncovering trends in election results. This kind of analyses can be done using aggregated data on municipal or regional level, considering that there are no micro data covering all voters and their characteristics. Moreover, related literature assumes that the overall level of observed variables can be also important. Becker *et al.* (2017), Lysek *et al.* (2020), Maškarinec (2019) argue that the socio-economic level of given region might be a significant determinant of voting decisions which can even partially overshadow personal characteristics of voters. This corresponds with Rodríguez-Pose (2020) who states that regions suffering from persistent economic decline tend to systematically revolt against traditional political parties (by voting for populist and anti-establishment parties) since they feel left-behind.

Some of the papers in this field use spatial econometrics to uncover possible geographical relationships between observed units. Those relationships can either stem from similarities in (in)dependent variables or from unobserved sources. Nevertheless, the spatial dependencies might play an important role in model selection since they influence the dependent variable but cannot be detected by methods such as OLS or correlation coefficients. According to Maškarinec (2017), this phenomenon of spatial dependence is still present in the Czech Republic and therefore it is necessary to consider spatial methods when performing detailed econometric analyses.

In my master's thesis I would like to extend my work from Černý (2019), where I performed general analysis of voting results in the Czech Parliamentary elections in 2017. Using the weighted least squares model and the spatial error model, I was trying to find the determinants of support of political parties and factors influencing voter turnout. The results showed that aggregated data on municipal level can provide reasonable estimation of voting behaviour and that there are indeed spatial dependencies within the data. Finally, there were a couple of possible modifications that might improve the analysis and those will be now incorporated in my master's thesis.

I would like to work with data from the Czech Parliamentary election in 2021 and utilize the fact that there should be the results of population census at the beginning of 2022. Many important variables, such as the level of education, are available only once in ten years and now they have been collected only few months before the election. Regarding the set of independent variables from Černý (2019), it will be slightly enhanced and enlarged by additional factors reflecting the overall socio-economic level. This will be inspected in further detail to determine and analyse (clusters of) municipalities that economically or socially lag behind. It is interesting to investigate whether those regions exist in the Czech Republic and to what extent is their separation related to geographical location, i.e., whether peripheral regions might be more separated when sources of education, employment, etc. (towns and cities) are more distant. Those areas might also be the most susceptible in terms of supporting populists or extremists, who play a significant role in the recent destabilization of election results as discussed by Havlík & Voda (2016). Importantly, peripheral areas do not appear only close to country borders but also inside, for instance due to very bad infrastructure or low reachability of services, education, etc.

Overall, there are two significant extensions of my bachelor thesis. Firstly, I intend to estimate the spatial Durbin model as the main spatial method. This model is more complex and as Glass *et al.* (2012) mention, it is primarily used when looking for spillovers, i.e., investigating whether the explained variable in a municipality might be influenced by independent variables of neighbouring units. In our case, it is possible that certain independent variables from the neighbourhood, representing the socio-economic level, might influence the voting decision as mentioned in the first paragraph.

Secondly, I would like to experiment and perform the same analysis using various machine learning methods. This idea arose from the fact that the estimation of spatial model is computationally very heavy, and the methodology is relatively complicated. Machine learning offers wide scale of methods and primarily, it should be considerably faster to estimate. Therefore, it will be very interesting to see, whether some of the ML methods can find similar results as the spatial econometrics, which was specifically developed for this type of data. The ML is sometimes used in this field, however, rather to provide predictions of election results using for instance data from social networks etc. Nevertheless, Richardson (2020) try to explain elections results within U.S. districts using their aggregated demographic data. This is quite similar to basic idea of this analysis, however, I would like to use more detailed set of population characteristics and, primarily, include the spatial aspect of the data.

Hypotheses

> Hypothesis #1: Electoral results of Czech political parties or coalitions are spatially correlated. This effect might be explained by common observed characteristics of neighbouring municipalities that affect voting behaviour, but also by spatial error dependence, i.e., when unobserved variables are likely to be spatially correlated.
>
> Hypothesis #2: Support of political parties or coalitions is associated with population characteristics. Most importantly, with those that describe socio-economic situation of the municipality such as unemployment, distraints or higher education. Results of the spatial Durbin model uncover that besides own characteristics of a municipality, also the characteristics of its neighbours play a certain role.
>
> Hypothesis #3: Clusters of municipalities that evince lower social or economic level appear within the data and their voting results evince similar trends.

Methodology The whole analysis will be performed using aggregated data on municipal level, which implies more than 6 250 observations representing all municipalities in the Czech Republic. Even though we do not know the election results yet, it is probable that every model will be estimated for all parties or coalitions that reach sufficient percentage of votes to get to the Parliament. First, the weighted least squares model will be estimated as in Černý (2019) to provide benchmark values for the following methods and to study dynamics of voting behaviour, which can change rapidly within 4-year voting cycle as discussed by Maškarinec (2019). The weights are used to reflect the size of municipalities and for this purpose I will use the number of inhabitants who participated in the election.

Using Moran's I, the presence of spatial autocorrelation within the election results will be tested. Subsequently, a set of Lagrange Multiplier tests will be implemented to determine whether the data evince signs of spatially lagged explained variable or dependence within errors. The former provides evidence for the spatial lag model, the latter for the spatial error model. Amara & El Lahga (2016) mention that if both types of tests favour spatial alternative, it is appropriate to use the spatial Durbin model that incorporates spatially lagged explanatory variables and allows to measure direct and indirect effects on the dependent variable. Moreover, authors state that this model provides unbiased estimates also in the case that only one type of tests supports the spatial dependence. As the weight matrix I will probably use the inverse-distance matrix or, if the data are available, the commuting matrix.

Finally, I will introduce a couple of machine learning methods such as supported vector machines, decision trees, neural networks, etc., and discuss whether they are appropriate for the analysis. Some of them offer interesting properties, for instance ability to uncover non-linear relationships between variables, that might provide advantages over OLS or WLS. I will incorporate the location of municipalities as independent variable and to compare the results with the spatial Durbin model, I might also include variables describing (in)dependent variables in nearby municipalities. This will attempt to explain the spatial effects and to estimate the same models as in the previous part.

Expected Contribution As I have partially mentioned, trends in the results of the Czech Parliamentary elections in the last decade are significantly more volatile than in the past. It is primarily due to the rise of new parties, many of them labelled as populist, who can attract voters within all parts of the country, including areas that used to unreservedly support well-established parties. Moreover, actual situation in the country is heavily affected by the Covid-19 pandemic and therefore it is even more interesting to analyse actual determinants of voting support only few months after the election. Simultaneously, it will be interesting to monitor situation within peripheral areas and compare it with the rest of observations.

Regarding the methodological point of view, local literature (Havlík & Voda 2016; Lysek *et al.* 2020; Maškarinec 2017; 2019) deals with the spatial dependence within election results by using maps, correlation coefficients, LISA indicators or spatial error models and is usually, at least partially, based on political science or social context. This thesis should provide rather technical analysis of the election results and the use of spatial Durbin model should enrich the current literature in this field. Moreover, implementing the machine learning point of view should provide a unique comparison of the method specifically designed for spatial data and the methods that are quite universal and applicable in many areas.

Generally, there are sophisticated ML methods developed for spatial data, however, in the context of elections results there are not many literature sources. On the other hand, for instance Praciano *et al.* (2018) perform spatio-temporal sentiment analysis to predict outcome of Brazilian elections by using a couple of supervised learning techniques on data from social networks. This thesis should combine similar frameworks with the aggregated data and enrich local analyses of voting behaviour and election results.

Outline

1. Introduction
2. Literature review: I will summarize Czech papers analysing local political scene, international papers that are devoted to spatial econometrics, especially those using the spatial Durbin model, and papers that implement machine learning methods into their analyses.
3. Data: after introduction of election results and describing the chosen set of independent variables, I will perform exploratory analysis to get better insight into the data and to uncover potential clusters of municipalities that socially or economically lag behind.
4. Methodology: I will describe methodological frameworks of all methods used for the analysis.
5. Results: estimates of spatial Durbin model and corresponding machine learning models will be presented and compared.
6. Conclusion: final remarks evaluating the implementation of the spatial Durbin model and performance of the machine learning methods when used for spatially dependent data.

Core bibliography

Amara, M., El Lahga, A. (2016): Tunisian constituent assembly elections: how does spatial proximity matter? Quality Quantity, 50(1), 65-88.

Becker, S.O., Fetzer, O., Novy, D. (2017): Who voted for Brexit? A comprehensive district-level analysis. Economic policy 32(92), 601-650.

Černý, J. (2019): Population Characteristics of Voters: Evidence from the Czech Parliamentary Election. Charles University, Bachelor thesis, 1-40.

Glass, A. J., Kenjegalieva, K., Sickles, R. (2012): The Economic Case for the Spatial Error Model with an Application to State Vehicle Usage in the U.S. Rice University manuscript.

Havlík, V., Voda, P. (2016): The Rise of New Political Parties and Re-Alignment of Party Politics in the Czech Republic. Acta Politologica, 8(2), 119-144.

Lysek, J., Pánek, J., Lebeda, T. (2020): Who are the voters and where are they? Using spatial statistics to analyse voting patterns in the parliamentary elections of the Czech Republic. Journal of Maps, 17(1), 33-38.

Maškarinec, P. (2019): The rise of new populist political parties in Czech parliamentary elections between 2010 and 2017: the geography of party replacement. Eurasian Geography and Economics, 60(5), 511-547.

Maškarinec, P. (2017): A Spatial Analysis of Czech Parliamentary Elections, 2006–2013. Europe-Asia Studies, 69(3), 426-457.

Praciano, B. J. G., da Costa, J. P. C. L., Maranhão, J. P. A., de Mendonça, F. L. L., de Sousa Júnior, R. T., Prettz, J. B. (2018): Spatio-temporal trend analysis of the Brazilian elections based on Twitter data. IEEE International Conference on Data Mining Workshops, IEEE, 1355-1360.

Richardson, B., Hougen, D. F. (2020): Districts by demographics: Predicting US house of representative elections using machine learning and demographic data. 19th IEEE International Conference on Machine Learning and Applications, IEEE, 833-838.

Rodríguez-Pose, A. (2020): The Rise of Populism and the Revenge of the Places That Don't Matter. LSE Public Policy Review, 1(1), 1-9.

Chapter 1

Introduction

Many academic papers analyse the outcomes of elections, attempting to find their determinants or learn how to predict them. Except for the political science context, there are several interesting perspectives to inspect election results. Specifically, academic literature often examines whether provided data evince spatial trends, and subsequently aims to explain them using spatial econometric methods. This framework plays a significant role in the field of election analyses since it enables to capture various forms of spatial effects, which are completely disregarded by standard non-spatial methods.

According to Cook *et al.* (2020), the last decade is associated with a more frequent use of spatial econometric models in the political science field, however, many papers usually do not pay enough attention to model specification and therefore might not correctly address spatial processes within data. In connection with that, LeSage (2014) argues that most researchers estimate similar models as their predecessors and under-utilize methods such as the Spatial Durbin Model (SDM), which might be, in many cases, more suitable for the explanation of election results.

Apart from the spatial econometric framework, this field might also be significantly influenced by the recent expansion of machine learning (ML). Nikparvar & Thill (2021) summarize numerous algorithms that are able to analyse spatial data and show that the ML has the potential to complement the established methodology of spatial analyses. Currently, the algorithms are utilized more for prediction than for inference, which might be related to the problem of low interpretability. Nevertheless, some academic papers interpret the election results by estimating the importance of particular variables and aim to overcome the challenge.

This thesis examines the results of the Czech Parliamentary election in 2021 and attempts to find the determinants of the support of major political subjects. Primarily, it focuses on addressing spatial trends within data by utilizing the established spatial econometric framework. Secondarily, it aims to replicate the analysis using several ML algorithms and to evaluate the (dis)advantages of both approaches.

The baseline analysis, which uses standard linear models, highlights the importance of variables related to the socio-economic level of municipalities and shows a significant polarity between political subjects. Subsequently, the thesis uncovers the autocorrelation within data and thoroughly evaluates various spatial econometric methods by performing numerous specification tests. This procedure, which also includes the selection of a weight matrix describing the relationships between particular observations, suggests that the Spatial Durbin Error Model (SDEM) is the most suitable method for the analysis.

This model accounts for the spatial lag of independent variables and spatial dependence among error terms, which means it enables distinguishing the direct and indirect effects of particular variables or handling unobservable trends. Those properties are considered to be very useful since both phenomena might appear in the election result. Therefore, the model is supported both by empirical evidence and economic reasoning.

The SDEM results confirm most trends observed in the baseline analysis and further examine their direct and indirect effects. The former describes the impact of municipality characteristics on its voting results, whereas the latter accounts for the impact of neighbouring municipalities. On average, variables describing the characteristics of inhabitants, such as the share of entrepreneurs or people with university education, evince more significant direct effects. On the contrary, the characteristics of municipalities, such as the ones referring to infrastructure or public spending, seem to impact the observations indirectly.

Finally, in order to replicate the analysis using ML algorithms, the geographical coordinates of observations are added as independent variables and the Random Forest (RF) model is estimated. Subsequently, the thesis also utilizes the spatial extension of this algorithm, the Geographical Random Forest (GRF). Both methods produce similar results, estimating the importance of particular variables and showing that the socio-economic characteristics of inhabitants play the most significant role. Simultaneously, when testing the ability to predict the election result from unseen data, the RF algorithm outperforms all methods used in the analysis.

Chapter 2

Literature Review

Academic literature that is concerned with election results incorporates a wide spectrum of analyses and methodological frameworks. In almost every country, there can be found political-science papers that study the outcomes of elections and discuss a country-specific political context. For instance, Szabo & Tatrai (2016) inspect the social cleavages of voters in Slovakia, or Hanley & Vachudova (2018) describe the recent decline of democracy in the Czech Republic. Such studies might bring valuable political insights that help other academics to construct theoretical models explaining the determinants of election results.

The analyses of elections have a long history, already Rattinger (1981) or Abramowitz (1988) attempt to explain voting results by examining population characteristics or political context. Similar papers can be found also in the last decade, such as Goodwin & Heath (2016) and Becker *et al.* (2017), who analyse the results of the Brexit referendum by using aggregate socio-demographic data. Similarly, Deniz *et al.* (2021) use socio-economic variables to explain the long-term voting support of a political party in Turkey and point out that the party is more supported at the time of economic prosperity and vice versa.

2.1 Determinants of Election Results

Among studies devoted to election results, there are numerous papers discussing the interconnection between the support of incumbent parties and the situation in society, so-called economic retrospective voting. Akarca & Tansel (2006) analyse tens of Turkish elections and show that voters reflect the economic development, which they consider to be the responsibility of major incumbent parties. However, they take into consideration only the events that happened within the last year and forget about the history. Similarly, **?**, who examines the subjective well-being of European citizens, shows that when explaining the support of incumbent parties, the (dis)satisfaction of voters explains more variance than standard macroeconomic indicators.

An unconventional approach is presented by Burnett & Kogan (2017), who use the quality of local roadways as a proxy for the overall satisfaction of inhabitants. Remarkably, they show that the number of complaints about potholes in roadways indeed relates to the support of local incumbents. Finally, an interesting insight into the topic is provided by Bojar & Vlandas (2021), who argue that different social groups evince specific forms of retrospective voting. As an example, they show that low-skilled workers tend to penalize the incumbents for the rise of unemployment, whereas older people rather take revenge for higher inflation. This suggests that the problem might be more complex and a simple analysis might not uncover the trends if they have the opposite directions by different groups of voters.

In general, it is very important, and usually quite complicated, to choose an appropriate set of variables for a model. Geys (2006) provides a meta-analysis of eighty-three studies that inspect voter turnout and points out that there is a high variability of used variables and hardly any of them are being used across all analyses. However, there are certain factors, such as population size or election closeness,[1] that seem to influence the turnout and should be, according to the author, consistently used and verified in all studies in this field. Even though academic literature does not have an explicit benchmark model for examining the outcomes of elections, there are multiple variables that appear to be serious candidates.

Many papers explaining election results argue that the overall socio-economic level might be a significant determinant. Becker *et al.* (2017) consider it to be an important factor in the Brexit referendum and discuss other potential variables, such as economic performance, the share of people in state administration, or wage level. Lasoń & Torój (2019) provide a thorough analysis of the Polish parliamentary election in 2015 and state that socio-economic or demographic variables, namely unemployment, education, the field of employment, divorce rate, etc., explain the most variance in the support of political parties. According to Amara & El Lahga (2016), age cohorts and the level of education are the

[1] There are multiple candidates to win the election and the result is expected to be close. This phenomenon is observed primarily in the case of two-round ballots (Fauvelle-Aymar & Francois 2006).

most important variables in the Tunisian election in 2011. In connection with that, Fiorino *et al.* (2021) analyse the turnout of several European Parliament elections and argue that GDP per capita, unemployment rate, and age, are the most important determinants.

Goodwin & Heath (2016) show that regions supporting the Brexit seem to be socially weaker since they evince significantly higher shares of low-skilled, older, or less educated people. They argue that the resistance to the EU is now observable in more places across the whole country, however, it is related to a narrower group of inhabitants, namely to those who are socially 'left-behind'. Rodríguez-Pose (2020) argues that people living in economically declining regions might get an impression that their surroundings have no development potential and might want to take revenge on authorities by voting for anti-establishment (populist) parties. Similarly, Pagliacci & Bonacini (2021) state that socially weaker regions, which do not economically benefit from the recent globalization, are more susceptible to voting for extremist parties.

In order to determine the weaker regions, demographic statistics, such as divorce rates, death rates, or migration balance, should be analysed. Especially the last variable might be very useful since it can be used as a proxy for the 'attractiveness' of a region. Apart from measuring the overall emigration and immigration, it is possible to inspect the share of people in productive age or the numbers of people in various age cohorts. If those data are available, they might bring valuable insights into the population structure and reveal negative trends such as high depopulation or the significant outflow of economically active inhabitants.

Nevertheless, the disadvantaged regions might not be separated only in terms of socio-economic factors, but also in terms of their geographical location. A lower development potential might in fact stem from the worse accessibility of jobs, goods, and services, which is connected to high commuting distances or the bad quality of infrastructure. This might amplify the differences between urban and rural areas and originate (even inland) peripheries. Constantino *et al.* (2021) investigate the determinants of voter turnout and in order to incorporate the influence of large cities, they monitor the commuting distance to the nearest city, which is considered to serve as the centre of education, administration, etc.

According to Lysek *et al.* (2020), the worse socio-economic situation and greater differences between cities and the countryside might be sometimes related to specific historical contexts in given countries. In particular, they argue that this phenomenon is observable in Czechia, Hungary, and Poland, however, not in the case of Slovakia, where the situation is more complex and cannot be defined as an urban-rural cleavage.

When discussing the overall characteristics of inhabitants, it is important to also consider their racial composition and religion. As Kinsella (2013) shows, the support of Republicans in the U.S. presidential election in 2008 is highly associated with the share of white people, and the same holds for Democrats and the share of people in minorities. The religion in highly developed European countries might not be an as important factor as for instance in the Middle East, where the voters are divided into groups of secularists and pro-Islamists(Carkoglu & Hinich 2016). Nevertheless, Joppke (2015) argues that religion still has its place in society even in highly secular countries in western Europe, which sometimes evince a considerable inflow of new inhabitants that are believers.

In connection with that, Otto & Steinhardt (2014) inspect the number of immigrants and asylum seekers in the city of Hamburg. They discuss a positive relationship between the level of immigration and the support of extreme right-wing parties, who usually enforce policies against minorities and foreigners. On the contrary, Pagliacci & Bonacini (2021), who study the same phenomenon in Italy, argue that higher numbers of immigrants within municipalities are associated with lower support of the Italian extreme right-wing party Lega. Nevertheless, the authors also inspect corresponding numbers in neighbouring municipalities and show that there might be the opposite trend. This suggests that the voters do not mind the actual presence of immigrants but, possibly also due to the extremists, are afraid of immigrants living somewhere else.

Considering all possible determinants of election results, it is necessary to also discuss the pandemic of Covid-19. There are not many related papers yet, however, for instance, Constantino *et al.* (2021) use Covid-19 incidence and the number of deaths to explain voter turnout in a recent municipality election in Brazil. According to them, high values of the indicators one month before the election have a significant negative effect on the turnout, whereas high Covid-19 values three or more months earlier evince the opposite trend. Therefore, as in the case of Bojar & Vlandas (2021), a naive analysis using all data would not uncover this relationship. Since the unprecedented global pandemic is expected to influence not only the turnout but also election outcomes, statistics such as incidence, death rates, or vaccination should be incorporated into related studies.

On the one hand, it has been previously mentioned that an unpromising environment might partially overshadow the personal characteristics of voters. On the other hand, there is also a problem called an ecological fallacy, which warns about the inference of individual characteristics based on data of the entire group the individual belongs to. According to Burnett & Lacombe (2012), the problem is still present in this field, however, it is much less problematic when using a lower level of aggregation. In the past, the analyses were performed on a country or regional level, whereas nowadays it is possible to collect the data for municipalities or sometimes even for electoral districts. The authors also mention that those districts are the smallest units that enable the use of overall statistics such as unemployment, the level of education, etc.

On the contrary, Maškarinec (2017) argues that in the Czech Republic there is a significant number of small municipalities whose statistics might be quite volatile and disrupt the inference. Therefore, he instead uses municipalities with extended powers which are assumed to represent smaller administrative centres and their natural catchment areas.

2.2 Elections in the Czech Republic

Considering the literature devoted to the Czech political scene, many authors use a sociological or political science approach to describe political subjects and their development. Havlík & Voda (2016) pay attention to a so-called theory of cleavages[2] and compare it to the actual situation in the country. They argue that relatively soon after the establishment of the Czech Republic in 1993, there was formed a stable political scene which included two strong parties, the Civic Democratic Party (ODS) and the Czech Social Democratic Party (CSSD), which were on the opposite sides of the left-right political spectrum and represented the socioeconomic cleavage among voters.

Nevertheless, new anti-establishment parties emerged and gained considerable support in the elections in 2010 and 2013. This decreased the stability of political system and simultaneously weakened the cleavage theory since the electorates of new parties were not unequivocally definable by the socio-demographic characteristics of voters. The authors also observe that the results

[2] A theory developed by S. M. Lipset and S. Rokkan, which states that social discrepancies, and possible conflicts, between groups of inhabitants lead to a political conflict. Simultaneously, it assumes that the cleavages affect the composition of political parties within a given country or region (Lipset & Rokkan 1967).

of new parties were usually not strongly associated with the established ones, which might imply certain changes in voting patterns rather than a substitution of political subjects.

Lysek *et al.* (2021) analyse the dynamics of recent parliamentary elections, focusing on a decline of social democrats (the CSSD), and confirm a dramatic transformation of voting results in the last decade. Considering the populist parties, the overall socio-economic environment of the territory seems to be a significant determinant of their voting support. This implies that regions suffering from higher unemployment, higher divorce rates, more inhabitants facing distraints, etc. might be more susceptible to vote for populists, whose recent expansion is, according to the authors, a key factor causing the decline of the traditional left. Moreover, Hanley & Vachudova (2018) argue that populists play a significant role in the recent decline of democracy.

Maškarinec (2017) analyses parliamentary elections in 2006 – 2013 and argues that the Czech political scene used to be quite stable and until 2010 it was defined by the cleavage theory as Havlík & Voda (2016) discuss. Then, a transformation began and weakened the previously strengthening importance of social class stratification. This was primarily associated with the ability of new political parties to attract voters both across the country and various social groups. Previously, the support of right- or left-wing political parties rather corresponded to areas with a high or low potential for development, respectively.

Nevertheless, Maškarinec (2019) emphasizes that it is still very important to study the socio-economic level of regions when inspecting political preferences. The dynamics of voters' support in the last decade shows that despite the significant changes in election results, spatial patterns remain relatively the same. According to him, established parties rather lose some part of the votes at the expense of their new alternatives, and the division related to the development potential is being preserved. Even though there can be found political subjects with unclear or changing spatial patterns, they usually quickly rise and fall or change their political orientation.

2.3 Spatial and Machine Learning Approach

Many papers that perform the analysis of election results do not take into consideration a possible geographical dependence between observations and they primarily focus on the influence of explanatory variables. However, academic literature argues that the aggregate data frequently evince spatial patterns and ignoring this aspect might lead to a biased and/or inefficient inference (Lasoń & Torój 2019; Mansley & Demšar 2015). Nevertheless, non-spatial analyses might provide sources of inspiration in terms of the selection of variables.

Already O'Loughlin *et al.* (1994) state that the main way to improve election analyses is to use a proper set of explanatory variables and to account for the spatial aspect of data. This also documents that spatial econometrics is a well-established methodological framework. In fact, Anselin (2010), who thoroughly summarizes its scopes and development over the past decades, considers the year 1979 as the beginning of spatial econometrics.[3]

The spatial aspects of data might be incorporated into analysis in various ways. Some papers only detect their presence and discuss political context, such as Nwankwo (2019), who examines the political situation in Nigeria, or Maškarinec (2019), who pays attention to the rise of populism in Czechia and the relationship between party's support and socio-economic situation in regions. Other studies provide a benchmark non-spatial model and replicate it for various territories in order to inspect whether the results are place specific. For instance, O'Loughlin *et al.* (1994) analyse old voting results of the Nazi party NSDAP and examine the differences between its support in the whole of Germany and in federal states.

Finally, many papers apply specific econometric methods that account for the spatial distribution of data and show that they explain significantly more variance (Amara & El Lahga 2016; Fotheringham *et al.* 2021; Mansley & Demšar 2015). Cook *et al.* (2020) confirm that the spatial approach is an appropriate solution, however, they point out that there are numerous frameworks, each accounting for a different type of spatial dependence, and the model specification might be a challenging task.

Considering the literature related to election analyses, there are also multiple papers incorporating ML algorithms. Those approaches, which are becoming increasingly popular in many fields, are often used to predict the results of elections based on data from social networks. For instance, Paul *et al.* (2017) and Liu *et al.* (2021) forecast the results of the U.S. presidential elections by performing sentiment analysis of Tweets in corresponding U.S. states. Simultaneously, there are few papers that analyse aggregate data and primarily focus

[3] Luc Anselin published numerous papers in this field, including the famous book of spatial econometrics (Anselin 1988).

on finding the determinants of voting support, such as Richardson (2020), who examines the House of Representatives election in the U.S.

The majority of papers, however, do not incorporate spatial aspects of data and therefore might suffer from the problems that have been previously mentioned. Nevertheless, the ML can also handle the spatial distribution of data, as it is shown by Li *et al.* (2019), who predict the results of the Australian federal election in 2016. On the one hand, this paper clearly shows that implementing such frameworks into election analysis might provide significant improvements in the accuracy of predictions and fitting of data. On the other hand, this might be associated with the lower interpretability and comparability of results since those algorithms work differently than standard econometric models.

In general, as the ML frameworks take over the initiative in many fields, a quick development is observed also in the domain of spatial analyses. Kopczewska (2021) summarizes current ML techniques and discusses both their potential and pitfalls. The author primarily argues that the algorithms are significantly more effective in computation terms, since they usually do not work with large weight/distance matrices, which are used by standard spatial econometric methods.

At the same time, the algorithms enable to incorporate multiple statistic and econometric methods at the same time, which might help to address more problems simultaneously, or they extensively use bootstrapping and boosting techniques, which significantly improve their predicting power on out-of-sample observations. Nikparvar & Thill (2021), who also argue that the ML provides very promising results, describe typical characteristics of spatial data and thoroughly discuss available ML algorithms and their benefits or limitations.

As it has been shown, election results can be analysed from many different perspectives and academic papers usually focus only on a single aspect at the time. In the Czech Republic, authors usually address the spatial distribution of data, however, they either only detect its presence (Maškarinec 2019) or elaborate on the result of a specific party, discussing its political context (Lysek *et al.* 2021). Providing a general spatial analysis of the last election in the Czech Republic should uncover current trends in voting support of Czech voters and the application of ML algorithms should provide a new approach to local studies.

Chapter 3

Data

This chapter describes all independent and dependent variables that appear in the analysis. Section 3.1 introduces the results of the Czech Parliamentary election in 2021 and all relevant political subjects in the Czech political scene. Even though the election results might be aggregated to the level of electoral counties (which implies 14,775 observations), it is not possible to collect other variables on the same level and therefore the thesis works with data from Czech municipalities, incorporating 6,254 observations in total.[1]

Section 3.2 is concerned with independent variables and, apart from introducing various types of data that can be collected about Czech municipalities, it also describes the process of data pre-processing and variable selection. Section 3.3 aims to perform exploratory analysis and provide supplementary material for the empirical analysis. It discusses descriptive statistics of particular variables and examines whether they might be spatially correlated.

The procedure of collecting and processing data is performed using the programming language Python, in particular, standard packages such as *numpy* or *pandas* for data manipulation, *selenium* for web scraping, *scikit-learn* for substituting missing values, *matplotlib* for creating figures, or *geopandas* for handling spatial data. In order to perform the variable selection, the package *BMS* in the statistical software R is utilized. Subsequently, the complete dataset is exported into the Geoda software, which enables displaying variables in maps and performing exploratory analysis.

[1] As of 1 January 2021, there are situated 6,258 municipalities in the Czech Republic. Nevertheless, four of them serve as a military district and are not inhabited.

3.1 Election Results

The parliamentary election in the Czech Republic took place on October 8-9, 2021, and reached an above-average voter turnout of 65.43 percent. Similarly as in recent elections (Lysek et al., 2021; Maškarinec, 2019), there appeared relatively significant changes in the local political scene. Firstly, stable members of the Parliament, the Czech Social Democratic Party (CSSD) and the Communist Party of Bohemia and Moravia (KSCM), did not manage to overcome the lawful threshold of five percent of votes and did not nominate to the Chamber of Deputies. Secondly, the governing party ANO 2011 - Movement of Andrej Babiš (ANO), which used to cooperate with the CSSD and the KSCM, neither reached the overall majority of mandates nor formed a new government, and therefore became an opposition party after 8 years of governance.

Apart from altering the force ratio on the local political scene, the election also resulted in a substantial decrease in the number of political subjects in the Chamber of Deputies, which dropped from nine to four. Nevertheless, this is primarily due to the establishment of two coalitions incorporating five political parties in total. This implies that it might be challenging to compare the results from different elections since there are different subjects gaining votes.

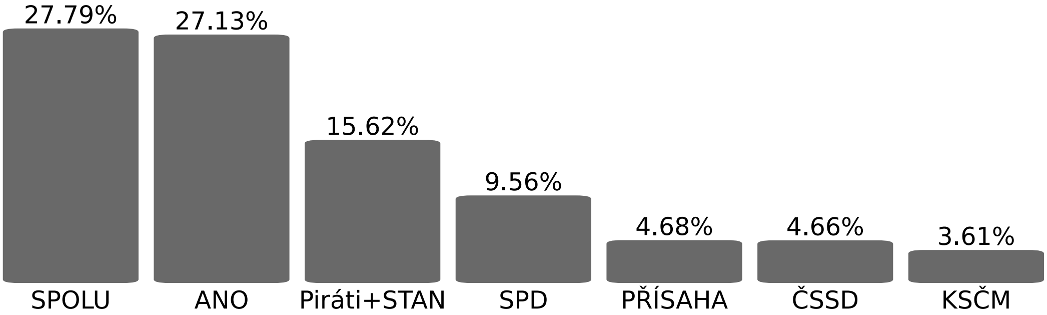


Figure 3.1: The results of the Czech Parliamentary Election in 2021

As it can be seen in Figure 3.1, first place in the election belonged to the coalition Together – Civic Democracy (SPOLU), which was formed by the Civic Democratic Party (ODS), the Christian and Democratic Union – Czechoslovak People's Party (KDU-CSL) and the Tradition, Responsibility, Prosperity (TOP 09). Closely behind the coalition finished the ANO, followed by the other coalition, Pirates and Mayors (Pirati+STAN), incorporating the Czech Pirate Party (Pirati) and the Mayors and Independents (STAN). The last subject that was nominated into the Chamber of Deputies was the Freedom and Direct

Democracy (SPD). Both coalitions received in total 108 mandates out of 200 and formed a new government, leaving the ANO and the SPD in opposition. Even though other parties displayed in Figure 3.1 did not succeed, the thesis incorporates them into the analysis since they used to play an important role in the Czech political scene. A five-number summary of election results is available in Appendix A in Table A.1.

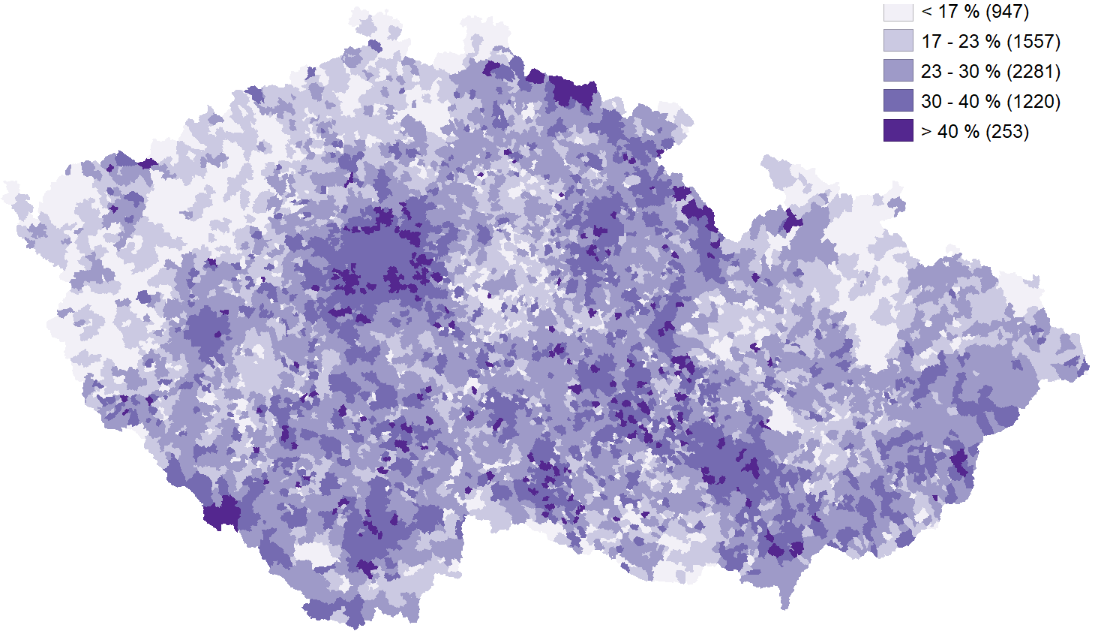


Figure 3.2: Election result - the coalition SPOLU

Figure 3.2, which displays the election result of the coalition SPOLU, suggests that there are regions, often including large cities and their surroundings, where the winner of election reached a higher voting support and vice versa. Interestingly, those regions contrast with the election result of the SPD, which can be observed in Figure 3.3, and this implies that there might be certain spatial patterns within the data. The phenomenon of possible spatial correlation is further examined in Section 3.3.

3.2 Independent Variables

Most variables, which characterize Czech municipalities and can be related to election results, are obtained from the Czech Statistical Office (CZSO), which regularly monitors a lot of population characteristics and performs a census every ten years. Apart from detailed election results, it provides actual data describing the numbers of inhabitants (including their gender or age structure)

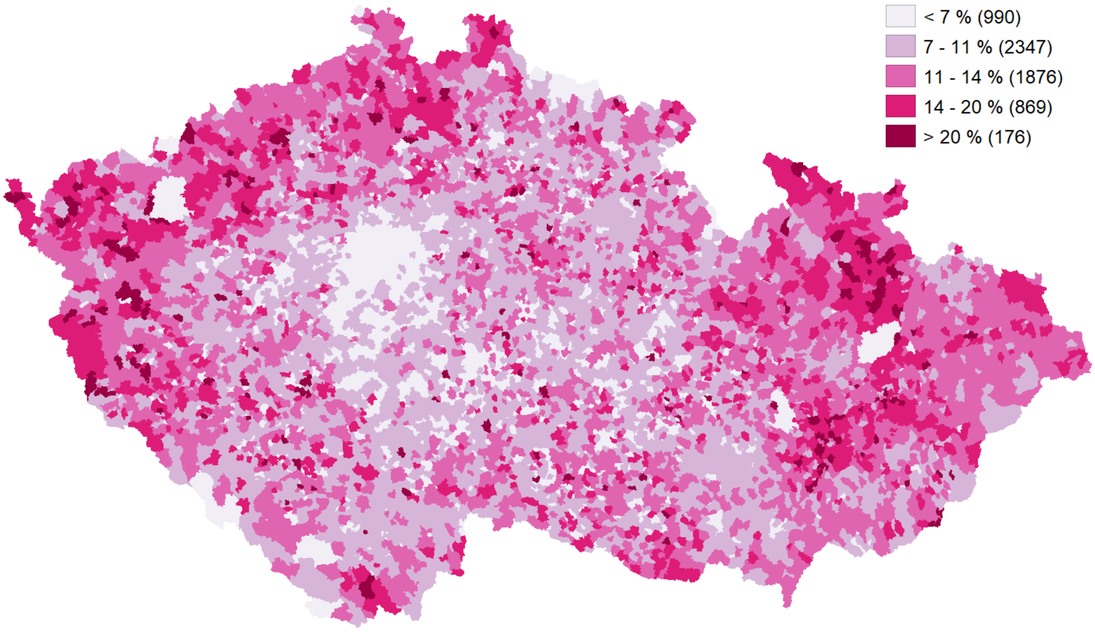


Figure 3.3: Election result - the SPD

and also demographic data monitoring migration, natural population growth, marriages, divorces, etc. In addition, it offers data about infrastructure (hospitals, schools, installation of gas/water piping, etc.), which might help to estimate the standard of living within municipalities. Regarding the census, there can be found more specific characteristics of inhabitants, such as their education, religion, race, or the field of employment.[2]

In addition to the CZSO, there are multiple sources providing valuable information about Czech municipalities. For instance, the Ministry of Labour and Social Affairs reports monthly statistics about unemployment, the Institute of Health Information and Statistics reports data about the Covid-19 pandemic (the numbers of Covid-19 cases and vaccinated people), and the Open Society examines issues such as bankruptcies and distraints. In order to evaluate the financial situation of municipalities, it is possible to analyse the database of the Ministry of Finance (so-called Monitor of Public Finance), which reports detailed municipal budgets, balance sheets, etc. A complete list of used data sources and corresponding links is available in Appendix A in Table A.2.

All independent variables, that can be gathered about municipalities, have to be further examined in order to depict an appropriate combination for empirical model. However, before the selection process, it is necessary to pre-process

[2] It monitors the number of economically active inhabitants and the number of entrepreneurs.

collected data. Firstly, there appear a few missing values, primarily by former military regions or by municipalities established in the last decade (14 observations in total). Since it is not desirable to remove whole observations and lose a part of available information, the missing values are substituted using the k-nearest neighbours (kNN) algorithm.[3]

Secondly, significant differences in magnitudes of data are observed since some of them report the shares of population and other absolute numbers. Therefore, all variables referring to the numbers of inhabitants are transformed to represent the shares of population and other variables with high magnitude are reported per capita and transformed by natural logarithm. Lastly, several dummy variables are created to describe the geographical locations of municipalities. They determine whether the municipality is situated within the reach of Prague or other large cities, which should serve as centres of education, employment, or services. The dummy variables might thus reveal whether the commuting (or at least aerial) distance to a city influences the standard of living and simultaneously the preferences of voters. The final list of all independent variables considered in the analysis is available in Appendix A in Table A.3.

Considering the variable selection, the thesis, similarly as Lasoń & Torój (2019), uses the Bayesian Model Averaging (BMA) to determine relevant variables and estimate a suitable model size. The BMA with uniform priors is applied to seven models (one for each political party that is considered in the analysis) and its average results are shown in Appendix A in Table A.4. Even though every political party might be associated with a different set of variables and thus report different posterior inclusion probabilities (PIP), the average of those values should provide an approximate guideline for selecting important variables. Simultaneously, there are specific cases, such as the *distraint*, that are chosen in spite of lower PIP values, since they are considered to be the significant determinants of voting support (Černý 2019).

3.3 Exploratory Analysis

In order to explore the collected data, Table 3.1 provides the basic descriptive statistics of all independent variables selected according to the BMA. It sug-

[3] In particular, kNN Regressor and kNN Classifier are applied to estimate missing numerical and categorical variables, respectively. Both algorithms use 9 nearest neighbours and are applied to modified dataset which includes the rescaled number of inhabitants and the level of unemployment (algorithms thus look for municipalities with similar size and social level).

gests, for instance, that the distribution of data might be relatively skewed since the mean value often significantly differs from the median. This is probably connected to the large number of small municipalities that are situated in the Czech Republic and whose characteristics might strongly influence the value of the median.[4] In connection with that, variables often evince a wide range of values since it is more likely to observe outliers in smaller municipalities.[5]

Table 3.1: Descriptive statistics of explanatory variables

	Mean	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
turnout	0.6836	0.0753	0.2645	0.6400	0.6867	0.7317	1.0000
inhabitants_log	6.2095	1.2064	2.7081	5.3891	6.0958	6.8635	14.1045
inhabitants_over_64	0.1986	0.0437	0.0429	0.1731	0.1962	0.2206	0.7857
average_age	42.4222	2.6793	30.7300	40.8600	42.3300	43.7900	66.4300
unemployment	0.0326	0.0181	0.0000	0.021	0.0296	0.0404	0.1966
distraint	0.0764	0.0580	0.0000	0.0395	0.0619	0.0945	0.6951
bankruptcy	0.0119	0.0101	0.0000	0.0052	0.0101	0.0165	0.0976
covid_vaccination	0.6240	0.0728	0.1700	0.5800	0.6300	0.6700	1.0000
covid_cases	0.1590	0.0459	0.0000	0.1320	0.1572	0.1837	1.1471
population_density_log	4.0592	0.9201	0.6388	3.433	3.9937	4.6089	7.8562
believers	0.1739	0.1276	0.0000	0.0755	0.1367	0.2371	0.8182
economically_active	0.6644	0.0523	0.1337	0.6400	0.6705	0.6953	0.9375
entrepreneurs	0.1194	0.0349	0.0183	0.0974	0.1156	0.1367	0.5000
roma_people	0.0003	0.0020	0.0000	0.0000	0.0000	0.0000	0.0625
primary_education	0.2120	0.0529	0.0294	0.1757	0.2062	0.2427	0.5380
highschool_education	0.3988	0.0569	0.1008	0.3667	0.4015	0.4345	0.6349
university_education	0.0737	0.0396	0.0000	0.0477	0.0674	0.0909	0.3236
immigrated	0.0306	0.0205	0.0000	0.0181	0.0271	0.0387	0.2556
emigrated	0.0239	0.0151	0.0000	0.0148	0.0219	0.0301	0.2191
born	0.0103	0.0103	0.0000	0.0069	0.0098	0.0131	0.0556
died	0.0122	0.0121	0.0000	0.0076	0.0112	0.0152	0.1893
near_Prague	0.0508	0.2197	0.0000	0.0000	0.0000	0.0000	1.0000
near_city	0.5819	0.4933	0.0000	0.0000	1.0000	1.0000	1.0000
regular_expenditures_pc_log	9.9017	0.4275	0.0000	9.6320	9.8581	10.1186	12.8563
total_expenses_pc_log	9.8298	0.4081	7.2802	9.5671	9.7879	10.0434	12.781
non_tax_income_pc_log	7.8667	0.9805	0.0000	7.2617	7.8349	8.4068	12.6503
capital_income_pc_log	4.1354	3.0490	0.0000	0.0000	4.5125	6.5249	12.3144
balance_sheet_brutto_pc_log	13.5338	0.4628	11.771	13.2348	13.5144	13.7991	17.3702
gas_piping	0.6364	0.4811	0.0000	0.0000	1.0000	1.0000	1.0000
water_piping	0.8826	0.3219	0.0000	1.0000	1.0000	1.0000	1.0000
sewerage	0.7761	0.4169	0.0000	1.0000	1.0000	1.0000	1.0000

Note: $N = 6,254$

Sources: A complete list of the sources is available in Table A.2.

A natural way to explore spatial data is to display them on a map and observe whether they evince some trends. Figure 3.4 depicts the average rate of unemployment and demonstrates several regions, such as north Moravia or

[4] The impact of this phenomenon on the empirical analysis is further discussed in Chapter 5.

[5] All suspicious values, such as full participation in the election or the majority of population facing distraints, are solely reported in small villages.

northwest Bohemia, that evince high values. In addition, a very similar phenomenon can be observed in Appendix B in Figures B.2 and B.3, which display the shares of people facing distraints and bankruptcies, respectively. Firstly, this implies that there might be certain spatial trends by those variables. Secondly, it shows that those regions, which conspicuously resemble the former Sudetenland, seem to be in worse socio-economic condition than the rest of the country. Rodríguez-Pose (2020) denotes such regions as *'places, that don't matter'* and studies whether they systematically resist traditional political parties by voting for populists or extremists.

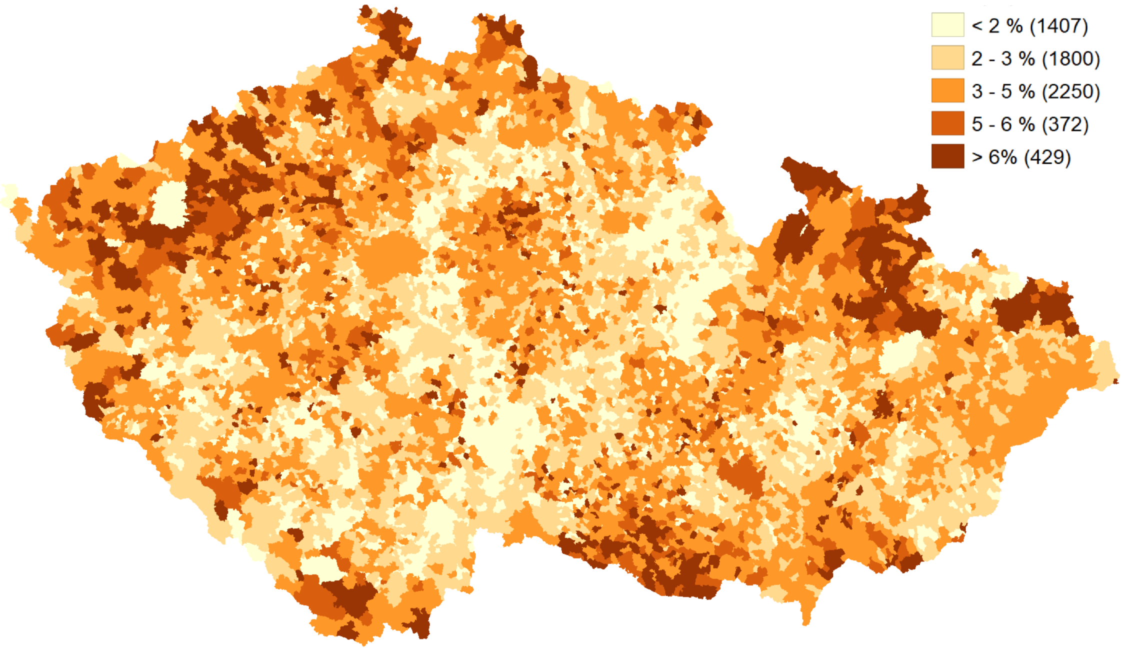


Figure 3.4: Unemployment rate in Czech municipalities

Considering Czech political subjects, the SPD might be considered to target anti-establishment voters. As Figure 3.3 shows, it reaches higher support exactly in the socially 'weaker' regions with the high rates of unemployment, bankruptcy, etc., and its political support clearly contrasts with the SPOLU.

Regarding the variables that describe the level of education, Figure 3.5 (and Figure B.4 in Appendix B) demonstrate an evident relationship between education and municipality size. Large cities and their neighbourhoods evince a significantly higher share of inhabitants with university education and naturally also a lower share of people with primary education. The latter variable rather corresponds with the socio-economically weaker regions and supports the assumption that those regions also suffer from lower human capital.

Simultaneously, in order to examine a possible urban-rural cleavage (Lysek *et al.* 2020), it might be very useful to inspect the variable *near_city*. From Table 3.1, it can be derived that 42% of Czech municipalities are not situated up to 15 km from the nearest city with more than 15 thousand inhabitants. This might be reflected in voting decisions since their inhabitants might feel more isolated and more neglected by authorities (Rodríguez-Pose 2020).

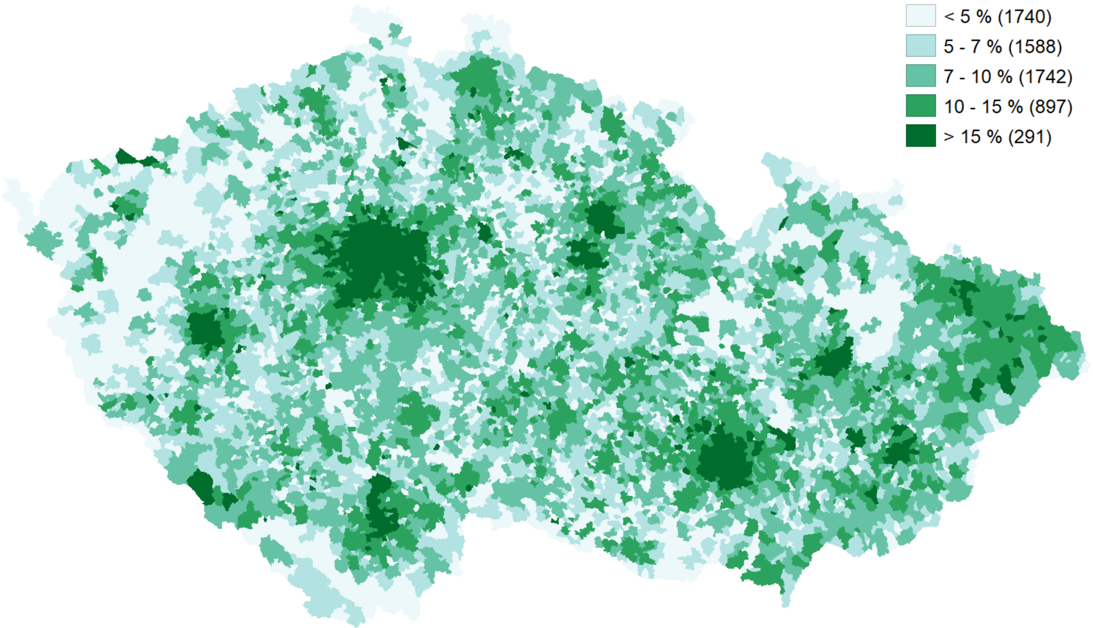


Figure 3.5: Inhabitants with university education

Among all variables in the analysis, the most distinct geographical pattern is observed in the case of religion. Figure 3.6 depicts the share of believers within Czech municipalities and shows that the highest shares can be found in certain parts of Moravia or Silesia. The share of believers is considered to be closely related to the voting support of the KDU-CSL (Černý 2019), however, this party is a member of the coalition SPOLU in the 2021 election and therefore the relationship might be less evident or slightly different than in previous elections.

The last step of exploratory analysis further examines the assumption of spatial dependence of data. As Kopczewska (2021) summarizes, one possibility to inspect spatial patterns is to cluster observations based on the values of independent variables, which do not include any spatial information. Therefore, a standard unsupervised ML method, the Hierarchical Clustering (HC), is applied in order to group observations into a specified number of clusters.[6]

[6] The algorithm merges observations using Euclidean metric (to compute distance between them) and Ward linkage criterion (to minimize the variance of clusters to be merged). The process stops when there are eight clusters in total.

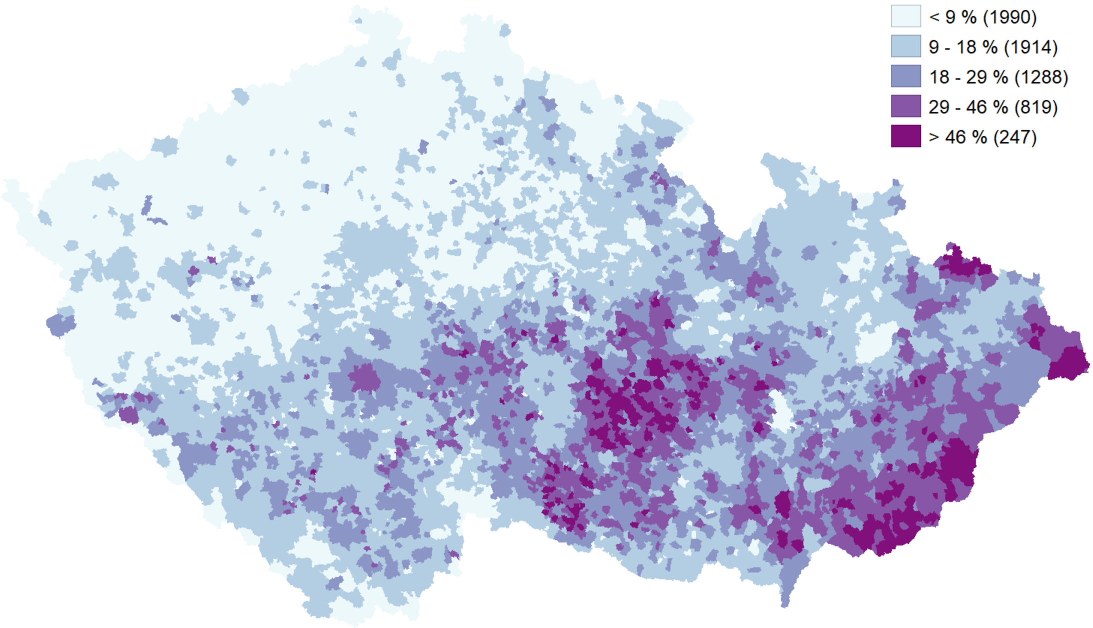


Figure 3.6: The share of believers

Considering all independent variables in the analysis, two of them (*near_city* and *near_Prague*) possess certain geographical information and therefore are not used in this case. Figure 3.7 displays the results of the HC for eight clusters and a closer look suggests there are indeed some groups of neighbouring municipalities within the same clusters. Interestingly, this trend appears even though the dataset is comprised solely from non-spatial variables.

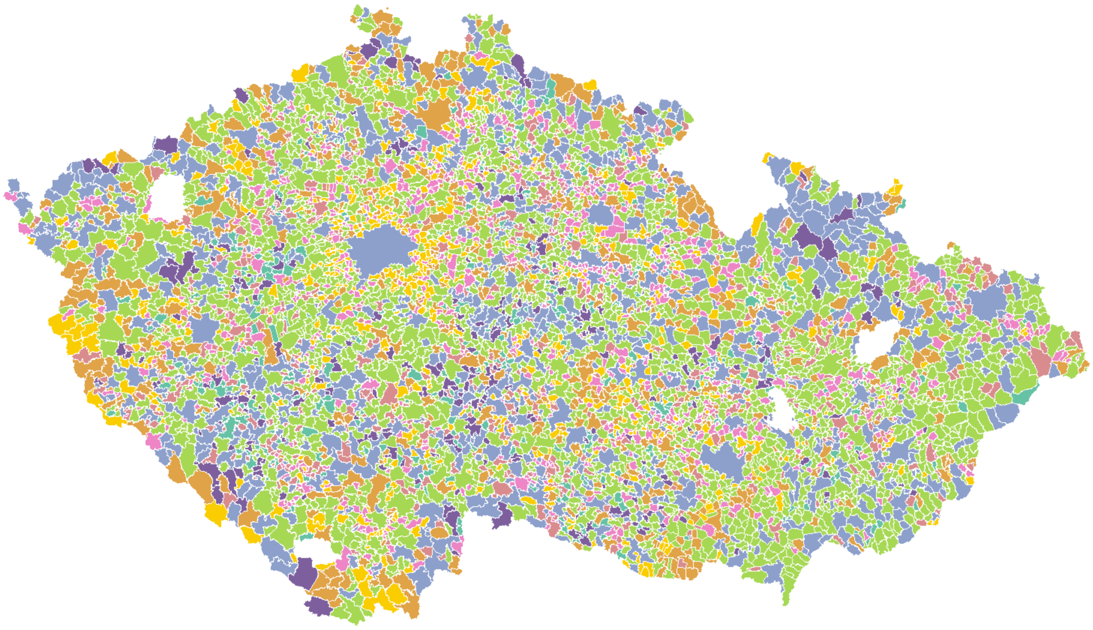


Figure 3.7: Hierarchical clustering results

Chapter 4

Methodology

This chapter provides a theoretical background to all empirical methods applied in the analysis. Section 4.1 describes the standard Ordinary Least Squares (OLS) method and discusses its possible generalization to the Weighted Least Squares (WLS) method. Section 4.2 explains the necessity to examine a possible spatial aspect of the data and introduces various methods that are able to account for it. Simultaneously, it provides a description of the SDEM model, which is, based on the empirical results from Chapter 5, chosen for the analysis. Section 4.3 introduces the possibility of using ML methods for spatial data and describes two algorithms, the RF and the GRF, which are utilized in the analysis.

4.1 Baseline Methods

The basic approach to explain the voting result of various political subjects is to estimate the standard linear regression. Even though this method has several limitations, which are discussed in the next section, it provides a benchmark model for the whole analysis. It is represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \tag{4.1}$$

which includes the dependent variable y, the intercept β_0, the error term u, and the independent variables $x_1, \ldots, x_k$ with corresponding coefficients $\beta_1, \ldots, \beta_k$. The coefficients are estimated by minimizing the residual sum of squares:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y_i})^2, \tag{4.2}$$

where the y_i and the $\hat{y}_i$ represent the true and the estimated values, respectively. The optimization process thus provides the equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k, \tag{4.3}$$

which enables to interpret partial effects of particular regressors on the dependent variable. The OLS coefficients have ceteris paribus interpretation, which implies that when holding other variables fixed, the change of x_1 is associated with the change of y:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1. \tag{4.4}$$

Econometric theory (Wooldridge 2013) defines so-called Gauss-Markov assumptions, MLR.1 - MLR.6, which enable to evaluate the quality of linear regression estimates.[1] According to Gauss-Markov Theorem, if the assumptions MLR.1 - MLR.5 hold, the estimated $\hat{\beta}$ coefficients are the best linear unbiased estimates.

A very important assumption is the MLR.4 (zero conditional mean), which requires the zero expected value of the error term, given all independent variables:

$$\mathbb{E}(u|x_1, \ldots, x_k) = 0. \tag{4.5}$$

The violation of this assumption implies a problem with multicollinearity, which is further examined and discussed in Chapter 5. The assumption MLR.5 (homoskedasticity) requires the same variance of the error term, given all independent variables:

$$Var(u|x_1, \ldots, x_k) = \sigma^2. \tag{4.6}$$

Therefore, if the variance is not constant, data are said to be heteroskedastic and Wooldridge (2013) suggests the WLS as a possible solution. The method enables to assign various weights to particular observations, possibly accounting for the changing variance. If the correct form of the heteroskedasticity is specified, the WLS results become the best linear unbiased estimates. Otherwise, standard errors are not valid, and the method might not be more efficient than the OLS.

Generally, the WLS might be applied also for economic reasons, for instance, when certain observations are more important than others. In this particular case, municipalities in the Czech Republic significantly differ in their size and, as it is further discussed in Section 5.1, the population is very unequally distributed among them. This might significantly influence the OLS results since

[1] The assumptions are listed in Appendix C.

this method treats all observations equally and numerous small municipalities might have a greater impact than a single large city. Therefore, it might be appropriate to assign greater weights to municipalities with more inhabitants, for instance by using corresponding numbers of voters as the weights.

The estimation process of the WLS is similar to the OLS, however, the minimization of residual sum of squares incorporates the weights as w_i:

$$WSS(\boldsymbol{\beta}, \mathbf{w}) = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2. \tag{4.7}$$

Most importantly, the interpretation of estimated coefficients remains the same, which implies that it is possible to study the ceteris paribus effect of particular variables.

4.2 Spatial Methods

As it has been mentioned in Chapter 2, election results and corresponding data about municipalities often evince spatial autocorrelation. If this phenomenon is present, classical non-spatial methods, such as the OLS, might be inefficient and biased (Anselin 1988). In order to test the presence of autocorrelation, most academic papers employ the Moran's I statistic, which is defined as:

$$I = \frac{n}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}} \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}, \tag{4.8}$$

where n stands for the number of observations, x_i represents the variable value at location i ($\bar{x}$ being the mean value), and w_{ij} describes the weight between observations i and j. Most frequently, the weight is related to the mutual distance of observations or it takes binary values, describing whether the observations are neighbours or not.

The value of Moran's I has to be interpreted with respect to the null hypothesis, which essentially assumes no autocorrelation in data.[2] In this case, instead of testing autocorrelation in particular variables, it is more appropriate to employ the Moran's I test, which enables to test autocorrelation in residuals of a linear model.

[2] If the statistic is positive and the p-value is sufficiently low, data are said to be spatially clustered (or dispersed - in the case of the negative statistic).

Many related papers (Amara & El Lahga 2016; Maškarinec 2017; Ozen & Kalkan 2017) also utilize the Local Indicators of Spatial Association (LISA) indicator, which enables to decompose the Moran's I to individual observations and inspect their contribution to the overall statistic. This might show the clusters of above- or below-average values and further describe spatial patterns within data. Another way to uncover the spatial dependence is the Geographically Weighted Regression (GWR), which is an extension of the OLS based on performing regressions for individual observations. Many authors (including Fotheringham *et al.* 2021; Mansley & Demšar 2015) implement this method and argue that it fits data significantly better than standard non-spatial methods.

Nevertheless, the most common approach is to use one of the spatial econometric models. In this case, the appropriate method has to be specified since autocorrelation can either stem from the dependent variable, residuals, independent variables, or the arbitrary combination of those. Therefore, multiple spatial models have been developed and each of them accounts for a specific type of autocorrelation. The most general framework is the General Nesting Model, sometimes called the Manski Model, which assumes all forms of spatial dependence at the same time:

$$y = \rho W y + X\beta + WX\theta + u \qquad u = \lambda W u + \epsilon, \tag{4.9}$$

where $\epsilon \sim \mathbb{N}(0, \sigma^2 I)$, W stands for the matrix of the mutual weights of all observations, and parameters ρ, θ, and λ denote the spatial autoregressive coefficient, the coefficient of fixed parameters, and the spatial autocorrelation coefficient, respectively. As it can be seen in Table 4.1, when restricting some of the parameters to zero, other spatial econometric models arise (leading to the standard linear regression in the case of restricting all parameters).

Table 4.1: Spatial Econometrics Models, based on Cook *et al.* (2020)

$y = X\beta + WX\theta + u$	$u = \lambda W u + \epsilon$	Spatial Durbin Error Model
$y = \rho W y + X\beta + u$	$u = \lambda W u + \epsilon$	Kelejian-Prucha Model
$y = \rho W y + X\beta + WX\theta + \epsilon$		Spatial Durbin Model
$y = X\beta + u$	$u = \lambda W u + \epsilon$	Spatial Error Model
$y = X\beta + WX\theta + \epsilon$		Spatially Lagged X Model
$y = \rho W y + X\beta + \epsilon$		Spatial Lag Model
$y = X\beta + \epsilon$		OLS Model

Cook *et al.* (2020) argue that even though it is appropriate to apply spatial econometrics in many fields, an incorrectly chosen spatial model does not perform well. Therefore, they provide a thorough description of particular models, serving as a guide for model specification, and highlighting that it is necessary to properly test the data before choosing a model.

Burnett & Lacombe (2012) extensively discuss multiple types of spatial specification tests and argue that the Lagrange Multiplier (LM) test is commonly used across academic literature. However, this test examines only one source of autocorrelation at a time (either residuals or the dependent variable) and disregards the rest, which implies that it can provide the evidence for the Spatial Lag Model (SLM) or the Spatial Error Model (SEM). This might not be sufficient and further tests are necessary.

LeSage & Pace (2009) use the spatial Hausman test, which assumes that coefficients from the OLS are the same as those from the SEM. If the hypothesis is not rejected, the OLS provides unbiased estimates with incorrect standard errors. Otherwise, there might appear the omitted variable bias and another method might be desirable. Authors argue that a possible candidate might be the SDM, which is not utilized enough in academic literature. Burnett & Lacombe (2012) show that if the SDM is implemented as the baseline model, a set of likelihood ratio tests (Elhorst 2010) can be applied in order to test whether the model might be reduced to the SLM.

In related academic literature (Fiorino *et al.* 2021; Kim *et al.* 2003; Lasoń & Torój 2019; O'Loughlin *et al.* 1994), it is common to detect autocorrelation in OLS residuals and subsequently estimate the SLM or the SEM. Nevertheless, multiple papers (Amara & El Lahga 2016; Jensen *et al.* 2013; Burnett & Lacombe 2012) propose that the SDM might be more appropriate since it accounts for spatial dependence in explanatory variables and might be the best model when the omitted variable bias is suspected (LeSage & Pace 2009). Authors also highlight that the SDM provides the direct and the indirect effects of particular variables and thus enables to study spillover effects.

Constantino *et al.* (2021) explain voter turnout using the SDEM model, which incorporates autocorrelation in explanatory variables and residuals. From the economic point of view, this approach seems to be the most appropriate also for this analysis. Firstly, there are likely certain unobservable factors, such as a strong political campaign within the region, which affect the election result. Secondly, the characteristics of nearby observations, describing the socio-economic level of the neighbourhood, are also expected to impact the election result.

On the contrary, the SLM model is considered to be inappropriate since it examines whether a voting result might be influenced by other voting results in the neighbourhood. The economic reasoning is further discussed in Section 5.2, which also provides empirical evidence in favour of the SDEM model and determines that this model is utilized in the spatial analysis of the thesis.

As it has been previously shown, the SDEM model is represented by the following equation:

$$y = X\beta + WX\theta + u \qquad u = \lambda Wu + \epsilon, \tag{4.10}$$

including the dependent variable y, the independent variables X with corresponding β coefficients, spatial error term u, model error term ϵ, and the terms WX and Wu, representing the exogenous effects of independent variables and the effects of error terms, respectively. W denotes a $n \times n$ weight matrix describing the spatial relationship of particular observations, either by expressing the distance between them or by stating whether they are adjacent to each other.

According to Elhorst (2010), spatial models can be estimated for instance by methods based on maximum likelihood estimation (MLE), instrumental variables/generalized method of moments[3], or Bayesian Markov Chain Monte Carlo approach. The thesis utilizes the MLE approach, which incorporates the following log-likelihood function (LeSage 2008):

$$ln\ L = -\frac{n}{2} ln(\pi\sigma^2) + ln|\mathbf{I} - \lambda\mathbf{W}| - \frac{e^T e}{2\sigma^2}. \tag{4.11}$$

Elhorst (2010) also discusses the interpretation of direct and indirect effects in various spatial models. When examining the partial derivatives of y with respect to an explanatory variable x, the author shows that the interpretation significantly differs for various types of models. Nevertheless, in the case of the SDEM model, the interpretation is straightforward since the direct and indirect effects are represented by β and θ coefficients, respectively. Therefore, the coefficients might be theoretically compared with those from linear models.

[3] The author argues that this case should be more studied in academic literature, since it has several useful properties. For instance, it helps to overcome computational difficulties which appear in the case of a large number of observations.

4.3 Machine Learning Methods

Academic papers incorporating ML algorithms usually employ multiple methods simultaneously (Liu *et al.* 2021) and primarily evaluate the accuracy of their predictions. This implies that a single paper might provide the Decision Trees (DT), the Logistic Regression (LR), the Supported Vector Machines (SVM), the Random Forest (RF), and other methods at the same time. Those algorithms represent possible alternatives to standard econometric models and since they work on different principles, they can provide useful properties, such as, for instance, the ability to uncover non-linear relationships within data.

On the one hand, ML methods are often applied straightforwardly and evince more accurate predictions than their standard counterparts. On the other hand, the interpretation of their results might be in certain cases quite challenging. Richardson (2020), who performs the DT, the LR, and the RF, aims to provide the relative importance of particular variables, which can be seen as an analogy to estimating the OLS coefficients. On the contrary, Li *et al.* (2019), who train the Hierarchical Graph Convolutional Neural Network and manage to fit data very well, do not examine explanatory variables at all since their meaning is quite ambiguous.

As it has been introduced in Chapter 2, multiple academic papers discuss the characteristics of spatial data and provide an overview of available ML algorithms in this field. According to Nikparvar & Thill (2021), the most straightforward approach to analysing spatial data is to incorporate the spatial characteristics of particular observations as independent variables (for instance, by adding their geographical coordinates), and subsequently apply standard ML tools.

A typical method that is used to analyse such data is the RF. Nevertheless, in order to introduce its theoretical foundations, it is necessary to first define the DT algorithm, whose idea is based on partitioning the data space into smaller regions and solving those regions with simpler models. The most common approach is the Classification and Regression Trees (CART) described by Breiman *et al.* (2017).

In the case of the regression DT, all observations at the beginning belong to a node τ. For each node, it is possible to compute its target value, which is in fact the average value of all included observations:

$$\hat{y}_\tau = \frac{1}{|I_\tau|} \sum_{i \in I_\tau} y_i, \tag{4.12}$$

where I_τ represents the indices of observations belonging to the node τ. Simultaneously, each node has its criterion c_τ, which enables to determine the degree of homogeneity of included observations. In the case of regression, the criterion is computed as:

$$c_{SE}(\tau) = \sum_{i \in I_\tau} (y_i - \hat{y_\tau})^2, \tag{4.13}$$

which is in fact the same squared error as in the linear regression. The tree is constructed by the repeated splitting of the node τ and the value of the criterion plays a significant role in this procedure. The goal is to determine a variable (and a specific value), which is able to split the node τ into nodes τ_L and τ_R, such the quantity $c_{\tau_L} + c_{\tau_R} - c_\tau$ is minimized. In other words, there is a maximum decrease of the overall criterion value and the new nodes are as homogeneous as possible.

The process of tree construction is usually constrained by several hyperparameters. The most common examples are the *maximum_depth*, which prohibits to further split the nodes that reach a specified depth, or the *minimum_samples*, which does not allow to split the nodes that have too few observations.

In practice, finding a suitable set of hyperparameters is called the hyperparameter tuning and this procedure requires to split the data into training, validation, and test samples. Using the training data, the algorithm is trained for many different combinations of hyperparameters. Simultaneously, trained models are evaluated on the validation sample in order to find the best set of hyperparameters. Finally, the best model is estimated and its performance is measured using the test data.

When considering more complex datasets with an increasing number of observations and variables, the DT algorithm might cease to be a sufficient tool. Nevertheless, it can be used to build up a more powerful algorithm, the RF. This method uses several principles to train a high number of independent decision trees and to make final predictions based on all of them.

The first principle is called bagging (from bootstrap and aggregating) and it is an ensemble algorithm which aims to increase the stability and accuracy of the final model. It includes creating m random samples of data, training a model for each of them, and making a final prediction, which is in this case the average prediction of the m trained models. The samples are drawn uniformly and with replacement, which ensures that they are independent.

The second principle is called the random subsetting of features and its purpose is to influence the process of building the trees. For each node, the algorithm is allowed to use only a random sample of available variables, which might significantly affect the tree structure and increase its variability. Similarly to the case of the DT algorithm, the hyperparameter tuning is an important part of utilizing the RF method, however, the most hyperparameters are the same since the construction of particular trees is a crucial part of the algorithm.

In general, the RF algorithm might be able to analyse spatial data (to some extent) and provide a different approach than the standard linear regression. Nevertheless, the method is in fact non-spatial and might not be sufficient to account for the spatial heterogeneity. Due to this fact, Georganos *et al.* (2021) introduce an extension of the RF algorithm, which is called the GRF. It is based on the same idea as the GWR, which incorporates the estimation of many local models in order to increase the overall flexibility. In this case, the RF is estimated for every observation and each of those models, which include also nearby observations, has its own feature importance and performance.

To illustrate the difference between the RF and the GRF method, Georganos *et al.* (2021) use the regression equation:

$$y_i = ax_i + \epsilon, \tag{4.14}$$

where ax_i is the standard (non-linear) prediction of the RF algorithm. This implies that all observations are used to make the prediction and their geographical location is not taken into consideration. On the contrary, the GRF equation is described as:

$$y_i = a(u_i, v_i)x_i + \epsilon, \tag{4.15}$$

where the prediction $a(u_i, v_i)x_i$ incorporates the coordinates u_i, v_i and the local model is estimated for every observation.

In order to select the observations that are used in local models, it is necessary to specify a kernel.[4] In this case, the kernel might be *adaptive*, which implies it includes a specified number of nearest observations, or *fixed*, which means it uses all observations that are within a specified radius. The former

[4] The nearby observations of location i are said to form its neighbourhood, also called a kernel.

might be useful especially when the observations significantly differ in their size.

When using the GRF algorithm to predict the values of new observations, it is necessary to specify a weight parameter which determines whether the prediction is more influenced by the local or the global estimates. The idea is that the former enables to monitor local heterogeneity and provide less biased estimates, whereas the latter uses a larger amount of data and thus evinces a lower variance of estimates. In general, this implies that the weight parameter and the kernel (including the number of neighbours or the radius) are crucial hyperparameters of the GRF algorithm, together with those inherited from the RF.

Generally, ML algorithms frequently evince different properties and requirements than standard econometric methods. In certain cases, such as the neural networks, it is necessary to scale or transform particular features[5] and some algorithms might not work at all if there are any missing values. Nevertheless, neither of those problems is relevant for this analysis, since the RF and the GRF are tree-based methods, which do not require feature scaling, and the missing values in data are substituted by the kNN algorithm, as discussed in Section 3.2.

[5] In machine learning, the variables are more frequently called features.

Chapter 5

Results

This chapter follows the structure of Chapter 4 and discusses many phenomena that appear in the results of the Czech Parliamentary election in 2021. Section 5.1 provides the results of baseline methods and explains that, in this case, it might be useful to estimate both the OLS and the WLS. Section 5.2 points out the possible limitations of the baseline analysis and thoroughly discusses the implementation of the spatial econometric framework, including the process of choosing the appropriate spatial model and weight matrix. Section 5.3 presents the results of the SDEM model and, except for the interpretation of direct and indirect effects, compares them with the baseline analysis. Section 5.4 provides the results of two ML algorithms, in particular, the RF and its spatial extension, the GRF. Finally, Section 5.5 endeavours to compare all methods by evaluating their ability to predict the results of the election.

The whole analysis is performed in the statistical software R. The baseline results are obtained using the built-in methods of the software and further analyses are done by utilizing several packages. In particular, *spdep* for spatial specification, *spatialreg* for the methods of spatial econometrics, *randomForest* for the RF algorithm, *reprtree* for visualizing a representative DT, *spatialML* for the estimation of the GRF and *stargazer* for exporting the results into LaTeX.

5.1 Baseline Analysis

Before estimating empirical models, independent variables, that have been selected for the analysis in Section 3.2, need to be tested for the presence of multicollinearity. In Appendix D there can be found the correlation matrix of all independent variables, which suggests that some of them seem to be relatively strongly correlated to each other. In this situation, academic literature (Franke 2010) recommends using the variance inflation factor (VIF) method to quantify the risk of multicollinearity and states that variables with the VIF ≥ 5 might be problematic. In this case, variables *inhabitants_over_64* and *average_age* exceed the threshold. Since the former is considered to be more relevant when explaining voting behaviour (the share of pensioners directly describes a specific part of voters that might evince the same electoral behaviour), the latter is removed from the analysis. Similarly, slightly higher VIF values are reported by the variables related to education and because it is more appropriate to control for the lower (higher) level of education, the *highschool_education* is also removed from the analysis.

In order to provide a benchmark model, most academic papers in this field usually estimate the OLS (Fotheringham *et al.* 2021; Pagliacci & Bonacini 2021, and many others). Nevertheless, this might be slightly problematic in the case of the Czech Republic, since the population is very unevenly distributed among municipalities. As it can be seen in Table 5.1, there are situated thousands of small villages which, however, represent only a small share of the total population. The thesis thus applies the WLS as the benchmark model[1] and simultaneously provides the standard OLS for comparison. The WLS results are available in the following Tables 5.2 and 5.3, whereas the OLS results can be found in Tables D.1 and D.2 in Appendix D.

Table 5.1: Distribution of population by municipality size

	<=100	**101-500**	**501-1000**	**1001-5000**	**5001-25000**	**>25000**	**Total**
Municipalities	438	2 981	1 364	1 194	236	41	6 254
Inhabitants	32 082	814 777	979 046	2 373 643	2 422 167	4 080 062	10 701 777
Share of inh.	0.30 %	7.61 %	9.15 %	22.18 %	22.63 %	38.13 %	100 %

According to the WLS results, the coalition SPOLU is more supported in municipalities with higher voter turnout, more inhabitants with university education and the higher share of entrepreneurs or born children. This suggests that the coalition achieves better voting results in regions with greater social capital, which corresponds with previous findings (Černý 2019) and confirms that parties such as the ODS or the TOP 09 are expected to attract educated voters living in more perspective regions. This assumption is further supported by negative coefficients on unemployment, bankruptcy, or primary education, which might all indicate certain social problems.

[1] The method uses the numbers of valid votes in the election as weights.

Considering the variables describing municipal budgets (per capita), the SPOLU receives more votes in municipalities with greater balance sheets and higher capital or non-tax incomes. Both factors indicate that a municipality is wealthier and better off, which is in line with previous statements. The coalition is also more supported in municipalities with a greater share of believers, which is undoubtedly related to the presence of the KDU-CSL in the coalition, however, the relationship is not as strong as in the case of the KDU-CSL in past elections. Interestingly, the voting result of the coalition is strongly associated with the share of Roma people. Nevertheless, this might be influenced by the fact that Roma people often live in towns or cities, which are in general expected to more support the coalition.

When considering all statistically significant variables, the movement ANO mostly evinces exactly the opposite coefficients than the coalition SPOLU, which implies it is considered to attract rather socially weaker voters. It receives greater support in municipalities with high rates of unemployment and bankruptcy, more pensioners, and low shares of born children and entrepreneurs. The opposite coefficients are observed also in the case of university education and municipal budgets, which further supports the polarity between the most important political subjects of the election.

Very similar differences can be observed between the coalition Pirati+STAN and the SPD, where the former strongly resembles the SPOLU and the latter is rather similar to the ANO. The Pirati+STAN is more supported in municipalities with higher shares of entrepreneurs, higher voter turnout and lower bankruptcy rates. On the contrary, the SPD gains more votes in municipalities with the high rates of unemployment and bankruptcy, lower voter turnout and the low share of born children. Interestingly, the coalition Pirati+STAN is associated with a higher share of people facing distraints and a lower share of economically active, which suggests that it can also attract some voters from weaker regions.

Data related to Covid-19 show that there is a relatively strong negative association of Covid-19 vaccination and the voting results of the SPD, which is probably related to its strong campaign against compulsory vaccination and Covid-19 restrictions imposed by the government. The ANO, which was governing the Czech Republic before the Parliamentary election, evinces a moderate increase of support in municipalities with a higher share of vaccinated people and a moderate decrease in the case of higher number of Covid-19 cases. This might suggest that voters from regions more affected by the pandemic tend to penalize the governing party.

Other political subjects, the CSSD, and the KSCM do not evince any strong relationships to examined variables and do not fit the data very well. Their voting support is associated with the high shares of pensioners or deceased inhabitants and also with the low shares of entrepreneurs or people with university education. As it has been mentioned in Chapter 3, those established parties used to play an important role in the Czech political scene, however, in last years they experience a constant decline in electoral support, as examined for instance by Lysek *et al.* (2021). In connection with the low fit to data, the last subject considered in the analysis, the Robert Šlachta's Civic Movement (PRISAHA), evinces extremely low goodness of fit. Its OLS and WLS results can be found in Appendix D in Table D.3.

The differences between both coalitions and the opposition parties are further supported by coefficients on variables referring to geographical location. As it is expected, the coalitions reach significantly higher voting support in Prague and its surroundings, whereas the ANO and the SPD lag behind in this region. Nevertheless, the other geographical variable, which examines whether the municipality is located close to a city, evinces exactly the opposite trends. This might suggest that the opposition parties receive higher support in the neighbourhood of cities and towns, or that the coalitions are able to also attract voters from small municipalities that are geographically separated and thus the total effect is ambiguous. Nevertheless, the trend is not clear since the results also suggest that the coalitions have slightly higher support in the places with higher population density, which should rather imply towns, cities, and their close surroundings. This ambiguity also corresponds with many insignificant coefficients on the number of inhabitants.

Regarding the demographic variables, the movement SPD reaches a higher support in regions with a lower share of born children and a higher share of deceased inhabitants. Those effects are not surprising, however, they do not correspond with the negative coefficient on the variable *inhabitants_over_64*[2] and this further suggests that this party is more successful in socially weaker regions.

The ANO evinces significantly lower support in municipalities with the higher share of born children, which might imply that the movement is not supported in places with many inhabitants in productive age, who are expected

[2] Municipalities with the higher share of pensioners are assumed to evince more deceased inhabitants. According to Figure D.1, there is indeed a positive relationship between those variables.

to start families. The results of the Pirati+STAN are associated with higher immigration, which implies that the coalition receives a greater support in regions that are attractive for new inhabitants. Surprisingly, the variables describing municipal infrastructure suggest that the Pirati+STAN, which includes a typical urban party the Pirati, is more supported in regions with no gas piping and no sewerage system, i.e. in small villages with only few inhabitants.

Table 5.2: WLS results, part 1

	Dependent variable:		
	SPOLU	ANO	Pirati+STAN
Constant	0.016 (0.033)	0.378*** (0.031)	0.019 (0.028)
turnout	0.228*** (0.017)	−0.247*** (0.016)	0.161*** (0.014)
inhabitants_log	0.0002 (0.001)	−0.001 (0.001)	0.002** (0.001)
inhabitants_over_64	−0.166*** (0.023)	0.227*** (0.022)	−0.095*** (0.020)
unemployment	−0.545*** (0.042)	0.465*** (0.040)	−0.109*** (0.036)
distraint	0.067*** (0.018)	−0.047*** (0.017)	0.139*** (0.015)
bankruptcy	−0.441*** (0.112)	0.583*** (0.108)	−0.291*** (0.095)
covid_vaccination	0.091*** (0.012)	0.077*** (0.012)	0.085*** (0.010)
covid_cases	0.019 (0.019)	−0.065*** (0.018)	0.189*** (0.016)
population_density_log	0.004*** (0.001)	−0.002 (0.001)	0.004*** (0.001)
believers	0.142*** (0.007)	−0.040*** (0.007)	−0.075*** (0.006)
economically_active	0.055*** (0.019)	−0.062*** (0.018)	−0.063*** (0.016)
entrepreneurs	0.620*** (0.026)	−0.483*** (0.025)	0.281*** (0.022)
roma_people	1.221*** (0.444)	−0.754* (0.427)	−0.262 (0.379)
primary_education	−0.089*** (0.024)	0.128*** (0.023)	−0.027 (0.020)
university_education	0.293*** (0.022)	−0.234*** (0.021)	0.115*** (0.019)
immigrated	0.055 (0.049)	−0.030 (0.047)	0.251*** (0.042)
emigrated	0.151** (0.061)	0.015 (0.059)	0.072 (0.052)
died	−0.130 (0.134)	−0.214* (0.129)	−0.108 (0.114)
born	0.353** (0.173)	−0.681*** (0.167)	0.212 (0.148)
near_Prague	0.021*** (0.002)	−0.016*** (0.002)	0.015*** (0.002)
near_city	−0.010*** (0.001)	0.012*** (0.001)	−0.005*** (0.001)
regular_exp_pc_log	−0.023*** (0.002)	0.019*** (0.002)	−0.008*** (0.002)
total_exp_pc_log	0.003 (0.002)	−0.003 (0.002)	−0.002 (0.002)
non_tax_inc_pc_log	0.002*** (0.001)	−0.002** (0.001)	0.001 (0.001)
capital_inc_pc_log	0.001*** (0.0002)	−0.002*** (0.0002)	0.001*** (0.0002)
bal_sheet_br_pc_log	0.005*** (0.002)	−0.003* (0.002)	0.002 (0.002)
gas_piping	0.002 (0.002)	0.005** (0.002)	−0.011*** (0.002)
water_piping	0.008** (0.004)	−0.003 (0.004)	−0.003 (0.003)
sewerage	0.012*** (0.003)	−0.004 (0.002)	−0.005** (0.002)
Observations	6,254	6,254	6,254
R^2	0.781	0.721	0.602
Adjusted R^2	0.780	0.720	0.600
Res. Std. Error (df = 6224)	1.066	1.025	0.909
F Statistic (df = 29; 6224)	765.954***	554.844***	324.886***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5.3: WLS results, part 2

	Dependent variable:		
	SPD	CSSD	KSCM
Constant	0.375*** (0.017)	0.0005 (0.013)	0.009 (0.012)
turnout	−0.083*** (0.009)	−0.043*** (0.006)	−0.029*** (0.006)
inhabitants_log	−0.002*** (0.0004)	0.001*** (0.0003)	−0.0003 (0.0003)
inhabitants_over_64	−0.049*** (0.012)	0.058*** (0.009)	0.076*** (0.008)
unemployment	0.210*** (0.022)	−0.030* (0.016)	0.050*** (0.015)
distraint	−0.003 (0.009)	−0.071*** (0.007)	−0.053*** (0.006)
bankruptcy	0.181*** (0.059)	−0.0003 (0.043)	0.095** (0.040)
covid_vaccination	−0.214*** (0.006)	0.045*** (0.005)	0.004 (0.004)
covid_cases	−0.015 (0.010)	0.013* (0.007)	−0.052*** (0.007)
population_density_log	−0.001** (0.001)	−0.002*** (0.0004)	−0.003*** (0.0004)
believers	−0.016*** (0.004)	0.038*** (0.003)	−0.023*** (0.002)
economically_active	0.006 (0.010)	0.004 (0.007)	0.006 (0.007)
entrepreneurs	−0.155*** (0.014)	−0.106*** (0.010)	−0.107*** (0.009)
roma_people	−0.554** (0.235)	0.069 (0.170)	−0.075 (0.157)
primary_education	0.024* (0.013)	−0.072*** (0.009)	0.023*** (0.008)
university_education	−0.056*** (0.012)	−0.055*** (0.008)	−0.023*** (0.008)
immigrated	−0.139*** (0.026)	−0.045** (0.019)	−0.035** (0.017)
emigrated	−0.081** (0.033)	−0.012 (0.024)	−0.018 (0.022)
died	0.150** (0.071)	0.108** (0.051)	0.100** (0.047)
born	−0.152* (0.091)	0.031 (0.066)	0.018 (0.061)
near_Prague	−0.008*** (0.001)	0.0005 (0.001)	0.002** (0.001)
near_city	0.008*** (0.001)	−0.001 (0.001)	−0.001*** (0.0005)
regular_exp_pc_log	0.006*** (0.001)	0.002*** (0.001)	0.003*** (0.001)
total_exp_pc_log	−0.004*** (0.001)	0.002** (0.001)	0.006*** (0.001)
non_tax_inc_pc_log	0.0001 (0.0004)	0.001 (0.0003)	−0.001*** (0.0003)
capital_inc_pc_log	−0.0003*** (0.0001)	−0.0002** (0.0001)	−0.0002** (0.0001)
bal_sheet_br_pc_log	−0.005*** (0.001)	0.002*** (0.001)	−0.00002 (0.001)
gas_piping	0.003*** (0.001)	−0.002*** (0.001)	0.002** (0.001)
water_piping	0.002 (0.002)	−0.005*** (0.001)	−0.001 (0.001)
sewerage	−0.001 (0.001)	−0.001 (0.001)	−0.002** (0.001)
Observations	6,254	6,254	6,254
R^2	0.685	0.217	0.365
Adjusted R^2	0.684	0.213	0.362
Res. Std. Error (df = 6224)	0.563	0.408	0.378
F Statistic (df = 29; 6224)	467.265***	59.513***	123.152***
Note:			*p<0.1; **p<0.05; ***p<0.01

When considering all WLS models in general, it is possible to state that the method provides results that roughly correspond with previous findings in this field. Simultaneously, the OLS provides relatively similar results as the benchmark method and the estimated coefficients mostly differ only in their magnitude. It is assumed that those differences are primarily caused by the fact that small municipalities can influence the OLS estimates because of their count.[3]

Surprisingly, both the WLS and the OLS method show that one of the greatest impacts on voting results is observed by the share of Roma people. Simultaneously, there can be seen a peculiar result by the coalition SPOLU, which reports exactly the opposite coefficients for both methods. Theoretically, this might be caused by the fact that this variable frequently evinces zero values and the several remaining observations are thus more likely to distinctively change the result when assigned various weights. Generally, the low values of *roma_people* suggest that the coefficients on this variable, which evince remarkably high magnitudes, are not so peculiar since the changes of this variable are expected to be rather negligible.

Regarding the measure for goodness of fit, the R^2, there are observed significant differences between political parties and also between the WLS and the OLS methods. However, the former might artificially increase the R^2 statistic and therefore cannot be used as a relevant metric, whereas the coefficients of determination from OLS models should provide reasonable measures.

In the case of political parties in the Chamber of Deputies, the R^2 ranges between 0.23 and 0.43. Other established parties report lower statistics, around 0.13, and the newly established civic movement PRISAHA only 0.03. This corresponds with Maškarinec (2019), who argues that in the Czech Republic there exist political parties whose support can hardly be empirically described and usually does not last a long time. Due to the impossibility to find the determinants of voting support, the movement PRISAHA is not included in further analysis.

[3] This is further supported by estimating the WLS using the logarithm or the square root of weights. Stronger transformations (which mitigate the differences in population sizes and assign relatively smaller weights to cities) imply that the coefficients are more similar to OLS results.

5.2 Spatial Specification

As it has been discussed in Section 3.3, voting results and corresponding population or municipal characteristics are expected to evince spatial trends. Nevertheless, in order to test for the presence of spatial autocorrelation and in order to perform spatial analysis, it is necessary to determine a spatial weight matrix. In this case, the queen contiguity matrix and the inverse distance matrix might be appropriate candidates. The former determines which observations share a common border with a particular municipality (or with neighbours of the municipality in the case of higher order) and assigns them value one. The latter includes all observations that are situated within a given distance, in this case the aerial distance, and assigns them weights which correspond to their inverse distance from the municipality.

On the one hand, the use of the aerial distance might be less accurate than using for instance the commuting time. On the other hand, the Czech Republic has highly developed transport infrastructure, including very dense railway or road networks, and there are not situated any natural obstacles complicating the transport, such as mountains, seas, etc. Therefore, the same aerial distances should be associated with similar commuting times and there should be no significant bias. Table 5.4 presents several variants of weight matrices and shows that they significantly differ in terms of the average number of neighbours. The queen contiguity matrix includes only a close neighbourhood of a municipality, whereas the inverse distance matrix incorporates much more observations and assigns them lower weights if they are situated farther away.

Table 5.4: Overview of weight matrices and corresponding numbers of neighbours

Type	**Mean**	**Min**	**Median**	**Max**
queen contiguity (order 2)	21.93	2	21	102
queen contiguity (order 3)	50.02	5	48	172
inverse distance (up to 20 km)	110.17	8	115	190
inverse distance (up to 30 km)	238.10	15	251	367

Generally, all weight matrices evince a wide range of the number of neighbours, which is primarily caused by the fact that municipalities located near the borderline do not have many Czech neighbours. This might be slightly problematic for the analysis since it aims to study the impact of the close neighbourhood but it is not able to reflect municipalities from other countries.

Those might provide important sources of employment, services, etc., and thus significantly influence the standard of living in a given region.

In order to select the most appropriate spatial weights, all types of weight matrices are combined with linear models and subsequently tested both for the presence of spatial autocorrelation and the spatial dependence of residuals or explained variables. Table 5.5 presents the values of Moran's I tests for various political subjects or weight matrices and confirms that autocorrelation is present in all tested cases. This implies that it is necessary to specify the correct form of a spatial model using the LM tests.

Table 5.5: Moran's I test for various political subjects and spatial weight matrices

	SPOLU	**ANO**	**Pirati+STAN**	**SPD**	**CSSD**	**KSCM**
QC (order 2)	0.210***	0.237***	0.270***	0.135***	0.142***	0.128***
QC (order 3)	0.174***	0.206***	0.227***	0.114***	0.113***	0.100***
ID (20 km)	0.143***	0.168***	0.201***	0.111***	0.107***	0.097***
ID (30 km)	0.112***	0.136***	0.164***	0.093***	0.088***	0.076***

Note: *p<0.1; **p<0.05; ***p<0.01

When evaluating the LM test statistics for various types of weight matrices (an example is presented below in Table 5.6), results suggest that queen contiguity matrices provide in general weaker evidence of spatial dependence than their counterparts using inverse distances. The latter group also shows that the higher the distance threshold, the stronger evidence it provides. Nevertheless, when increasing the distance up to 50 kilometres, there is only a negligible increase of the test statistic. Therefore, the inverse distance matrices using the 20 and the 30 kilometres threshold are considered to be suitable candidates for the estimation of several spatial models and their subsequent evaluation.

Selecting those matrices is appropriate also from the economic point of view. Firstly, the used distances roughly correspond to the size of a district or a smaller region, respectively.[4] The surroundings that has approximately the size of a district is considered to be optimal since most inhabitants are not expected to commute very long distances to work or to spend most of the time far away from their residence. Secondly, the voting result of a municipality should by primarily influenced by its closest neighbours, whose impact should diminish with increasing distance. Therefore, it is appropriate to use the inverse distance matrix which takes those facts into consideration.

[4] In the Czech Republic, the average distance to the closest district town is 18.18 km and the closest region city is situated on average 35.45 km from a municipality.

As it has been already mentioned, autocorrelation in data implies that it is necessary to specify a spatial model using the LM tests. When performing the tests for various weight matrices and political subjects, the results are mostly consistent and provide significant evidence in favour of the spatial dependence of residuals. In certain cases, the tests also suggest that there is a spatial dependence in the explained variable, however, this does not hold for all political subjects. Simultaneously, spatial models that include the lag of y do not seem to be the most appropriate for this analysis, because the actual interpretation of those effects is rather ambiguous. Since all tested cases evince similar LM test statistics, Table 5.6 provides only an example of the results, in this case for the movement ANO and the inverse distance matrix using the 20 kilometres threshold.

Table 5.6: Spatial specification tests (the ANO, the ID matrix with 20 km threshold)

Test	**Statistic**
LM Error Test	1,711.0***
Robust LM Error Test	1,720.1***
LM Lag Test	3.7*
Robust LM Lag Test	12.8***
Note:	*p<0.1; **p<0.05; ***p<0.01

After performing spatial specification tests, several spatial models are estimated and evaluated in order to select the most appropriate combination of an econometric method and a weight matrix. Table 5.7 presents eight different models (analyzing the movement ANO) and reports their log-likelihood (LL) and Akaike Information Criterion (AIC), assuming that the preferred model should maximize the LL and minimize the AIC. In this case, the results suggest that the SDEM model is the most preferred option and that the inverse distance matrix with the 20 kilometres threshold consistently outperforms the other matrix with the 30 kilometres threshold. When estimating the same models also for the coalition SPOLU, the results prefer exactly the same combination.

On the one hand, it might be slightly problematic to directly compare the LL and AIC indicators.[5] On the other hand, Burkey (2018) argues that there is no universal technique which would be able to compare all types of spatial

[5] Even though all models use the same set of independent variables, estimating for instance the SDM model implies creating the spatial lags of regressors and thus increasing the number of variables, which subsequently affects the values of both indicators.

Table 5.7: Evaluation of several spatial models, the ANO

Model	Weights	LL	AIC
Spatial Lag	ID (20km)	9,018.3	-17,913.0
	ID (30km)	9,012.9	-17,902.0
Spatial Error	ID (20km)	8,924.8	-17,782.0
	ID (30km)	8,900.1	-17,732.0
Spatial Durbin	ID (20km)	9,036.3	-17,941.0
	ID (30km)	9,034.8	-17,938.0
Spatial Durbin Error	ID (20km)	9,052.0	-17,972.0
	ID (30km)	9,038.1	-17,944.0

models and unequivocally determine the best one. According to the author, this implies that researchers have to primarily incorporate their expert judgment to select the correct method and, secondarily, to utilize various statistical measures providing at least certain information about the goodness of fit of the models.

From an economic point of view, the SDEM model is considered to be the most suitable method for this analysis. On the one hand, it takes into account population (or general) characteristics of neighbouring municipalities, which are expected to determine the socio-economic situation. On the other hand, this method also accounts for the unobservable effects of neighbouring municipalities that cannot be described by any variables.

According to LeSage (2014), the most important step to select the appropriate spatial model is to consider the type of spillovers that is expected to appear in the analysis. If there are present endogenous interactions and feedback effects, i.e., a change in the characteristics of observation is associated with a set of adjustments of many (possibly all) observations, leading to creating a new equilibrium, the spillovers are said to be global. This scenario requires incorporating the spatial lag of y, which means using either the SLM, the SDM, the Kelejian-Prucha Model, or the Manski Model.

The second case mentioned by LeSage (2014) does not include endogenous feedback effects and assumes that the interactions between observations do not take place in the whole space. Therefore, this situation is called the scenario of local spillovers and implies using either the Spatially Lagged X Model (SLX) or the SDEM Model. The author argues that methods working with global spillovers are in general more used in academic literature, even though this scenario is in practice less probable and also more complicated to interpret.

In this particular case, the spillovers are not expected to be global. Technically, it is impossible that voting results in neighbouring municipalities could affect each other, since the election takes place at one time and the results are presented afterwards. Theoretically, previous election results might play a certain role, nevertheless, this scenario is highly improbable when considering an impact of neighbouring municipalities. Similarly, a possible influence of the incumbents of particular municipality is much more likely to affect only the municipality itself and not its surroundings. In addition, this phenomenon might not play such an important role in the Parliamentary election, since many incumbents in Czech municipalities, primarily in small villages, have no connection to political subjects in the Chamber of Deputies.

Generally, it is much more likely that (un)observed characteristics of Czech municipalities have a certain local impact on their neighbourhood, since they attempt to describe their inhabitants and the standard of living. Economic reasoning thus corresponds with the results of statistical tests presented above, which are, furthermore, supported also by the Likelihood Ratio tests presented in Table 5.8. All tests provide enough evidence for rejecting the null hypothesis which states that the SDEM model should be reduced. Finally, this implies that the spatial analysis of the thesis is performed using the SDEM model and the inverse distance matrix with the 20 kilometres threshold.

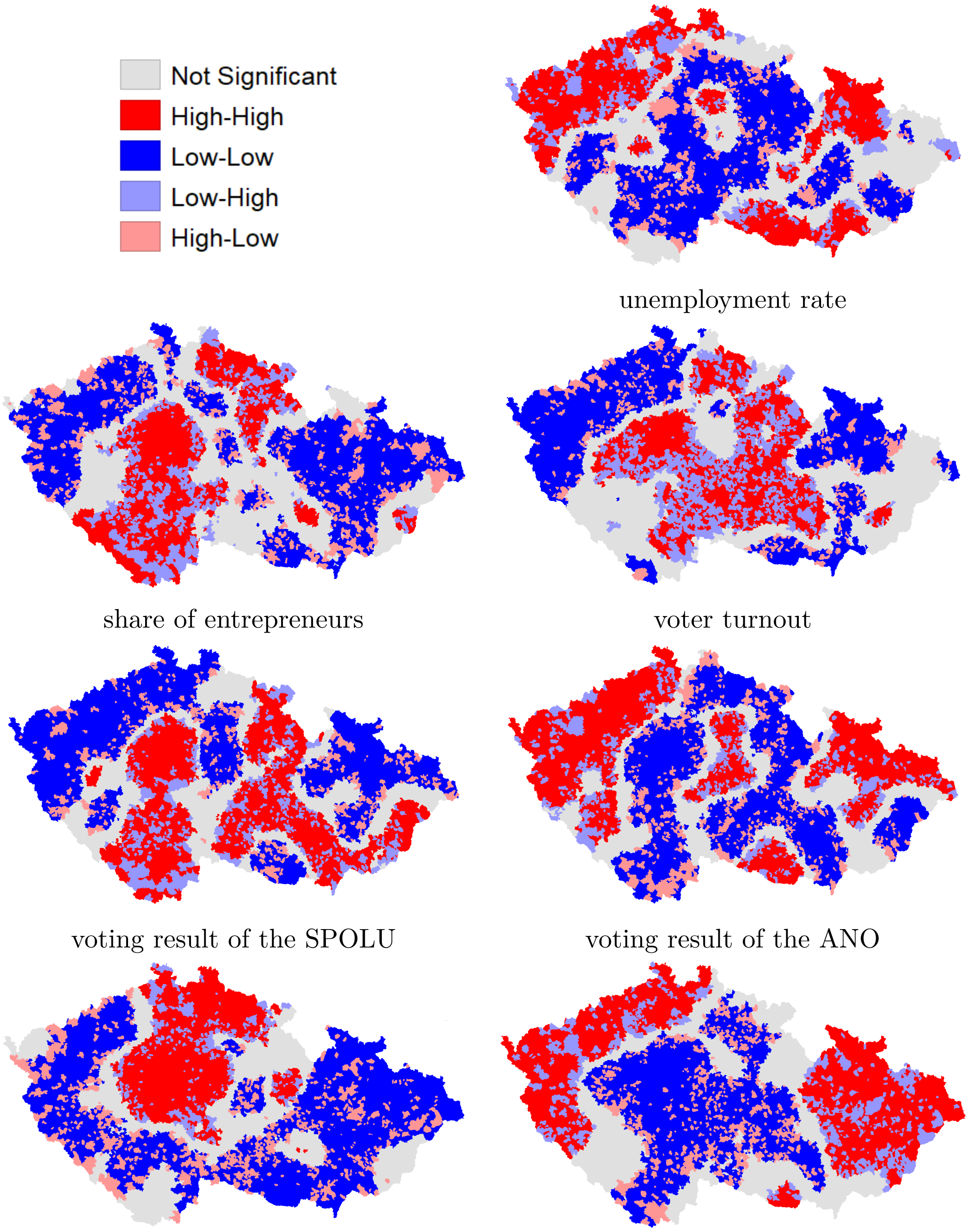
Table 5.8: LR tests of reducing the SDEM model (the ANO, the ID matrix with 20 km threshold)

Reduced model	**Likelihood ratio**
WLS Model	1,091.3***
Spatially Lagged X Model	367.2***
SEM Model	218.1***

Note: *p<0.1; **p<0.05; ***p<0.01

Nevertheless, before estimating the set of SDEM models, it is possible to further inspect the spatial trends, which have been discussed previously. Computing the LISA indicators and depicting their values in maps provides a more specific overview of regions that are associated with unusually high or low values. Figure 5.1 shows the LISA indicators of several independent variables and major political subjects. Not surprisingly, it confirms many trends which have been suggested by the WLS results or discussed in Chapter 3.

In the vast majority of regions where the coalition SPOLU receives higher support of voters, the ANO lags behind and vice versa. A similar difference can be observed between the coalition Pirati+STAN and the SPD, which evince the opposite values mainly in the eastern part of the Czech Republic and in



unemployment rate

share of entrepreneurs

voter turnout

voting result of the SPOLU

voting result of the ANO

voting result of the Pirati+STAN

voting result of the SPD

Figure 5.1: LISA indicators for various variables (the ID matrix with 20 km threshold)

northwest Bohemia. Nevertheless, the support of both political subjects partially overlaps in the Usti nad Labem region, which is considered to be socially weaker.

Simultaneously, the voting results of all political subjects indeed correspond with the values of independent variables. The most evident examples are northwest Bohemia and north Moravia, which are associated with the higher support of the ANO and the SPD, the low share of entrepreneurs, low voter turnout high unemployment rate.

5.3 Spatial Analysis

According to spatial specification tests and economic reasoning provided in Section 5.2, the SDEM model is considered to be the most suitable method for the spatial analysis of the Czech Parliamentary election. Tables 5.9 - 5.11 present the estimated impacts of direct, indirect, and total effects of six SDEM models, including all relevant political subjects (without the movement PRISAHA, which has been excluded from the analysis in Section 5.1). In Appendix D, there can be found the complete results of all models, presented in Tables D.4 and D.5.

In general, the estimated impacts show that the total effects of SDEM results correspond with the baseline analysis, in particular, with the OLS results. This is not surprising since this spatial method does not account for the municipal size and is thus more similar to the OLS. Nevertheless, the estimation of direct and indirect effects enables to study the structure of total effects in greater detail.

The results show that the voting support of the coalition SPOLU is primarily determined by several direct effects. All of them suggest that the coalition is more successful in municipalities that are socially and economically better off, since they possess more assets per capita, there are more entrepreneurs, educated people, etc. Interestingly, variables *unemployment* and *immigrated* seem to be irrelevant within the municipality but evince significant indirect effects, suggesting that the coalition is more supported if there are fewer immigrants and unemployed inhabitants in neighbouring municipalities. Simultaneously, the SPOLU seems to be successful in municipalities with more inhabitants, whereas it evinces lower support if there are more neighbouring municipalities located close to a city. This might indicate that the SPOLU is able to approach voters from various locations, including both crowded cities and isolated municipalities in the countryside.

Similarly as in the baseline analysis, the movement ANO significantly differs from the winner of the election in many coefficients, which implies that it rather attracts voters in municipalities with lower human capital. Nevertheless, the variable *died* suggests that the movement is more supported in the places with a lower share of deceased inhabitants, which is rather unexpected, especially when observing the positive coefficient on the number of pensioners within the municipality.

The indirect effects of the unemployment rate and the share of immigrated inhabitants are again more statistically significant and evince higher magnitudes than corresponding direct effects (as expected, there are the opposite signs than in the case of the coalition SPOLU). Those variables might thus have certain impact on the neighbourhood of a given municipality and theoretically influence the voting behaviour of local voters. Similarly, there are significant indirect effects of regular expenditures, suggesting that the coalition SPOLU evinces lower support if the neighbouring municipalities spend more resources on their administration, schooling, local services, etc. (and vice versa for the movement ANO).

When inspecting the coalition Pirati+STAN, the SDEM model shows that its support is primarily associated (both directly and indirectly) with lower bankruptcy rate and a higher share of entrepreneurs. The results also suggest, that the coalition receives more votes in municipalities with more educated inhabitants and a lower share of pensioners. Nevertheless, those variables simultaneously evince the opposite (not very high) indirect effects and there is observed also an indirect positive effect of the share of inhabitants facing distraints.

A possible explanation of those results, which are rather unexpected and counter-intuitive, can be the fact that, apart from Prague and its close neighbourhood, the coalition succeeds also in regions which are socially or economically weaker, and it is thus more complicated to find the determinants of its voting support. This claim is further supported by the variables describing migration, which suggest the Pirati+STAN reaches higher support if the neighbouring municipalities evince the high share of emigrants and the low share of immigrants, implying that the regions are not very attractive for any inhabitants.

In connection with that, dummy variables describing the infrastructure of a municipality suggest that the Pirati+STAN is more supported if there is worse technical infrastructure in the neighbourhood, in particular, no gas piping and sewerage system installed. Theoretically, when considering also the positive direct effect of the number of inhabitants, the results might suggest that the coalition Pirati+STAN is more supported in towns that are surrounded by small municipalities with worse infrastructure and lower social capital.

The last political subject that nominated to the Chamber of Deputies evinces slightly weaker relationships to provided data than the previous cases. The SPD is assumed to be more successful in municipalities that are economically and socially worse off, since its support is primarily associated with higher bankruptcy rates, fewer inhabitants with university education, or higher unemployment.

Considering the variables *inhabitants_log* and *near_city*, the SDEM results suggest that the SPD might be more supported in smaller municipalities that are located in the neighbourhood of larger towns and cities. Simultaneously, it attracts more voters if the municipality is located in a region that is characterized by more deceased inhabitants and a lower amount of public possession. This further supports the assumption that the SPD receives greater support in regions that somehow lag behind the rest of the country.

The results of the remaining political parties, the CSSD and the KSCM, do not provide much information about corresponding voters. The greatest impact on their voting results is observed in the case of bankruptcy rates and the share of entrepreneurs, however, the remaining variables mostly evince only negligible coefficients. This might be related to the fact that those parties experience a constant decline in political preferences and no longer attract specific groups of voters. For instance, even though both parties might be expected to gain votes from older people, their coefficients on *inhabitants_over_64* are lower than in all previous models.

Generally, the majority of described SDEM models show that the variable *roma_people* evinces the highest coefficient among all independent variables, often suggesting that a 1% increase of Roma inhabitants implies more than a 1% change in voting results. In this case, it is primarily due to strong direct effects observed by almost all political subjects. Nevertheless, as it has been previously explained, those coefficients might be slightly misleading since the variable is very unlikely to significantly change its value.[6]

[6] Table 3.1 shows that the mean share of Roma people is only 0.03%.

Table 5.9: Estimated direct, indirect, and total effects of the SDEM, part 1

	SPOLU			**ANO**		
	Direct	Indirect	Total	Direct	Indirect	Total
turnout	0.14***	0.006	0.146***	-0.118***	-0.024*	-0.142***
inhabitants_log	0.006***	0.002	0.008***	-0.006***	-0.003***	-0.009***
inhabitants_over_64	-0.248***	-0.004	-0.253***	0.349***	-0.034**	0.315***
unemployment	-0.051	-0.078**	-0.129**	0.022	0.119***	0.141**
distraint	0.017	0.005	0.022	-0.012	-0.01	-0.022
bankruptcy	-0.163*	-0.109	-0.273**	0.384***	0.219**	0.603***
covid_vaccination	0.068***	-0.006	0.062***	0.117***	0.018**	0.135***
covid_cases	0.003	-0.007	-0.004	-0.029*	0.024*	-0.006
population_density_log	0.002	-0.002	0.000	0.004**	0.003**	0.007***
believers	0.25***	-0.022***	0.227***	-0.091***	0.012***	-0.08***
economically_active	-0.061***	0.004	-0.056**	0.069***	-0.013	0.056**
entrepreneurs	0.471***	0.085***	0.556***	-0.425***	-0.068***	-0.493***
roma_people	-1.19***	-0.341	-1.531***	-1.19***	0.066	-1.124**
primary_education	-0.04*	0.01	-0.029	0.053***	0.033**	0.087***
university_education	0.326***	0.019	0.345***	-0.286***	0.039**	-0.246***
immigrated	-0.024	-0.124***	-0.148**	0.07*	0.165***	0.235***
emigrated	0.068	0.003	0.071	-0.029	-0.104*	-0.133
died	0.127	-0.163	-0.036	-0.273***	0.092	-0.182
born	0.18	0.033	0.213	-0.072	-0.201	-0.273
near_Prague	-0.012	0.001	-0.011	0.003	0.001	0.004
near_city	0.001	-0.002**	-0.001	-0.003	0.002*	-0.001
regular_exp_pc_log	-0.003	-0.008***	-0.011***	0.000	0.008***	0.008**
total_exp_pc_log	-0.002	0.006**	0.004	-0.001	-0.004	-0.005
non_tax_inc_pc_log	-0.001	-0.001	-0.001	0.002	0.000	0.002
capital_inc_pc_log	0.000	0.000	0.001	0.000	-0.001***	-0.001**
bal_sheet_br_pc_log	0.004**	0.002	0.007***	-0.002	-0.003**	-0.005**
gas_piping	-0.001	0.000	-0.001	0.000	0.002	0.002
water_piping	0.003	-0.002	0.001	-0.002	0.001	-0.001
sewerage	0.003	0.005***	0.009***	0.000	-0.001	-0.001
Observations	6,254			6,254		
Lambda	0.065			0.065		
Log Likelihood	8,907.843			9,033.811		
σ^2	0.003			0.003		
Akaike Inf. Crit.	−17,691.690			−17,943.620		
Wald Test	3,253.170*** (df = 1)			5,717.104*** (df = 1)		
LR Test	339.481*** (df = 1)			367.218*** (df = 1)		

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.10: Estimated direct, indirect, and total effects of the SDEM, part 2

	Pirati+STAN			**SPD**		
	Direct	Indirect	Total	Direct	Indirect	Total
turnout	0.043***	0.03***	0.073***	-0.044***	-0.007	-0.051***
inhabitants_log	0.002**	0.001	0.004**	-0.003***	-0.001	-0.004***
inhabitants_over_64	-0.166***	0.033**	-0.133***	-0.035***	0.019*	-0.017
unemployment	-0.026	-0.015	-0.041	0.068**	0.015	0.083**
distraint	-0.02	0.033**	0.013	0.012	0.012	0.024
bankruptcy	-0.186***	-0.268***	-0.454***	0.139**	0.074	0.213***
covid_vaccination	0.046***	0.003	0.05***	-0.142***	-0.017***	-0.159***
covid_cases	0.019	0.01	0.029*	0.013	-0.007	0.006
population_density_log	-0.001	0.001	0.001	0.000	0.001	0.001
believers	-0.029***	-0.001	-0.030***	-0.063***	0.009***	-0.055***
economically_active	-0.077***	0.011	-0.066***	0.037***	0.007	0.044***
entrepreneurs	0.176***	0.026*	0.202***	-0.057***	-0.032***	-0.090***
roma_people	0.65**	0.249	0.899**	-0.061	0.168	0.107
primary_education	0.027*	-0.011	0.016	-0.003	-0.025***	-0.028*
university_education	0.225***	-0.052***	0.173***	-0.118***	0.001	-0.116***
immigrated	0.008	-0.054*	-0.046	-0.018	0.035	0.017
emigrated	0.036	0.077*	0.112*	0.03	0.015	0.044
died	-0.058	-0.074	-0.132	0.003	0.102*	0.105
born	0.007	-0.009	-0.003	-0.02	0.061	0.041
near_Prague	-0.011	0.003	-0.009	0.006	-0.001	0.005
near_city	-0.003	0.000	-0.002	0.002	0.001**	0.003**
regular_exp_pc_log	-0.001	-0.001	-0.002	0.001	-0.001	0.000
total_exp_pc_log	0.005***	0.002	0.008**	-0.002	0.001	-0.002
non_tax_inc_pc_log	0.000	-0.001	-0.001	0.001	0.001	0.001
capital_inc_pc_log	0.000	0.000**	0.001**	0.000	0.000	0.000
bal_sheet_br_pc_log	0.001	0.001	0.002	-0.003***	-0.002**	-0.005***
gas_piping	0.002	-0.003***	-0.001	0.001	0.000	0.001
water_piping	-0.002	0.000	-0.002	0.002	-0.001	0.001
sewerage	-0.001	-0.005***	-0.006***	-0.001	0.000	-0.001
Observations	6,254			6,254		
Lambda	0.066			0.051		
Log Likelihood	10,862.860			11,836.570		
σ^2	0.002			0.001		
Akaike Inf. Crit.	−21,601.710			−23,549.130		
Wald Test	12,636.440*** (df = 1)			209.948*** (df = 1)		
LR Test	508.519*** (df = 1)			107.819*** (df = 1)		

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.11: Estimated direct, indirect, and total effects of the SDEM, part 3

	CSSD			**KSCM**		
	Direct	Indirect	Total	Direct	Indirect	Total
turnout	-0.007	-0.009	-0.016*	-0.012*	-0.004	-0.016*
inhabitants_log	0.000	-0.001	-0.001	0.000	0.000	0.000
inhabitants_over_64	0.045***	0.004	0.049***	0.08***	0.007	0.086***
unemployment	-0.007	-0.025*	-0.032	0.019	0.015	0.034
distraint	0.002	-0.007	-0.004	-0.01	-0.015**	-0.026**
bankruptcy	-0.104***	-0.007	-0.11*	0.111***	0.091**	0.201***
covid_vaccination	0.028***	0.003	0.031***	-0.022***	0.000	-0.022***
covid_cases	0.018**	0.008	0.026***	0.003	-0.011**	-0.007
population_density_log	0.000	0.000	0.000	-0.003***	-0.001**	-0.005***
believers	0.011**	0.005***	0.016***	-0.047***	0.003*	-0.044***
economically_active	0.005	0.001	0.006	0.004	0.002	0.006
entrepreneurs	-0.093***	-0.009	-0.102***	-0.089***	-0.017**	-0.106***
roma_people	1.087***	0.306*	1.393***	-0.128	-0.179	-0.307
primary_education	0.000	-0.013*	-0.013	0.014	-0.003	0.011
university_education	-0.012	0.003	-0.009	-0.042***	0.018**	-0.025*
immigrated	-0.024	-0.03	-0.054*	-0.005	-0.005	-0.01
emigrated	-0.077***	0.017	-0.059	0.005	0.02	0.026
died	0.004	0.103**	0.107	0.014	0.002	0.016
born	-0.093*	0.092	-0.002	-0.046	0.062	0.015
near_Prague	0.005	0.000	0.005	-0.001	0.000	-0.001
near_city	0.000	0.000	0.000	-0.001	0.000	-0.001
regular_exp_pc_log	0.001	0.000	0.001	0.000	0.000	0.000
total_exp_pc_log	0.000	-0.001	-0.001	0.003***	-0.001	0.002
non_tax_inc_pc_log	0.000	0.000	0.000	0.000	0.000	0.000
capital_inc_pc_log	0.000***	0.000	0.000**	0.000	0.000	0.000
bal_sheet_br_pc_log	0.000	0.001	0.001	-0.002***	0.000	-0.002**
gas_piping	-0.001	-0.001**	-0.002**	0.001	0.001	0.001
water_piping	-0.002*	0.000	-0.002	0.001	-0.001	0.000
sewerage	0.000	0.000	0.000	-0.001	0.000	-0.001
Observations		6,254			6,254	
Lambda		0.058			0.055	
Log Likelihood		13,914.110			14,146.100	
σ^2		0.001			0.001	
Akaike Inf. Crit.		−27,704.230			−28,168.190	
Wald Test		499.037*** (df = 1)			299.735*** (df = 1)	
LR Test		176.782*** (df = 1)			117.674*** (df = 1)	

Note: *p<0.1; **p<0.05; ***p<0.01

In general, the estimated total effects of SDEM models are primarily composed of direct effects, whose coefficients are on average more statistically significant and evince higher magnitudes. Theoretically, this might be associated to the fact that the majority of variables directly describe the inhabitants, reporting for instance the share of entrepreneurs or pensioners, and those characteristics might be less likely to influence other municipalities (nevertheless, this does not hold for *unemployment* or the variables related to migration). On the contrary, when inspecting the variables describing the municipalities (either their budget, infrastructure, or location), the indirect effects seem to be on average more relevant to explain the voting support of particular political subjects.

5.4 Machine Learning Analysis

As it has been discussed in Section 4.3, the basic approach to apply ML methods to spatial data is to include the geographical characteristics of particular observations and use the algorithms in a standard way. In this case, the coordinates of each municipality are added into the dataset as independent variables and the RF model is estimated.

The first step to implement the RF method is to find the best set of hyperparameters. Nevertheless, this might be a challenging task since there are many possible values of particular hyperparameters, which in total represent a huge number of combinations. Therefore, it is necessary to select a feasible sample of values. In Table 5.12, there can be seen an example of hyperparameter tuning,[7] showing the models that evince the lowest mean squared error (MSE) values when evaluated on validation data.

In this analysis, the tuning procedure primarily focuses on the construction of trees, and therefore it experiments with the minimum size of terminal nodes and various numbers of features which are used during a node split. The former plays an important role since it indirectly limits the maximum depth of each tree. In this case, the best models incorporate at least 15 observations in every terminal node. Simultaneously, the best models use 10 or 15 features during each node split, which is similar to the default value of the hyperparameter, usually a third of the total number of features.

The RF algorithm also enables to incorporate the weights of particular ob-

[7] The procedure primarily uses the model for the movement ANO.

servations, which represents another way to influence tree construction. When using the number of valid votes in each municipality as the weight (the same approach as in the WLS method), the models seem to perform worse. However, taking a logarithm of the values significantly decreases the differences between observations and slightly improves the overall performance of the models. Lastly, hyperparameter tuning aims to approximate the optimal number of trees which should be estimated. Typically, academic literature uses several hundred or thousand of trees, which is also the case of this analysis.

Table 5.12: The example of hyperparameter tuning, the RF

weights	**features**	**min node size**	**estimated trees**	**MSE**
log(votes)	15	15	700	0.003827
log(votes)	10	15	1000	0.003830
log(votes)	10	15	200	0.003831
log(votes)	10	25	700	0.003837
log(votes)	10	3	1000	0.003839
log(votes)	15	3	1000	0.003839
none	10	3	700	0.003840
none	10	25	200	0.003841

Before selecting the best set of hyperparameters, it is possible to replicate the approach of spatial econometrics by adding several independent variables describing the neighbourhood of each observation. In this case, the neighbourhood is approximated by 20 nearest observations and for each independent variable, the average values of the neighbours are taken as a 'neighbourhood effect'. Nevertheless, this approach, which uses twice as many variables as the initial model, does not imply any improvement in the MSE statistic and is therefore not developed any further.

As it can be seen in Table 5.12, the results of hyperparameter tuning do not show an unequivocal combination of values that outperforms the rest. Nonetheless, they provide the idea of suitable hyperparameters, which is further verified by repeating the whole process for several political parties. Since the results are approximately the same, the final RF model is estimated using the logarithm of valid votes as the weights, 10 features during each node split, terminal nodes with at least 15 observations and 1000 estimated trees.

The problem of many ML algorithms lies in the complicated interpretation of their results, which often significantly differ from standard econometric methods. In this case, the estimated RF models provide only two statistics, which describe the relative importance of particular variables.[8] Nevertheless, according to Parr *et al.* (2018), measuring the importance of variables is an efficient, reliable, and universal technique, which is applicable to many different algorithms.

The first statistic, the % increase in the MSE, is based on measuring the MSE on the out-of-bag (OOB) data and comparing it with the MSE after the permutation of particular features.[9] The second statistic, the increase in node purity, describes the average drop of the residual sum of squares when a feature is used to split a node. In general, Parr *et al.* (2018) argue that the % increase in the MSE is a more reliable indicator of the variable importance since the increase in node purity might in certain cases artificially increase the importance of some types of variables.

In Tables 5.13 and 5.14, there can be seen the values of both statistics for all estimated models. When inspecting the % increase in the MSE, it is evident that the results of many political subjects are primarily determined by the share of entrepreneurs, the level of education, or the geographical location. The socio-economic situation within the municipality also seems to be a relevant factor since all variables such as *unemployment*, *distraint*, and *bankruptcy* evince relatively high values, as well as the variable describing the number of inhabitants.

Similarly as in the previous results, the share of believers impacts the election results, and this trend is by far the most obvious in the case of the coalition SPOLU. Another example of a wide gap between political subjects is observed by the movement ANO and its strong connection to the share of older people. This corresponds with the previous analysis, however, a similar trend is expected also by other parties, such as the CSSD and the KSCM.

Since the thesis thoroughly discusses the spatial dimension of the analysed data, it is not surprising that the geographical coordinates seem to be important variables in the RF models. The highest values are observed by the coalition Pirati+STAN, which might suggest that its voting support is frequently associated with specific locations. Interestingly, the SPD evinces much higher importance of the *longitude* than the *latitude*. Theoretically, this might be

[8] In both cases, higher numbers imply higher importance of variables.

[9] The RF creates a bootstrap sample (with replacement) for each tree, which usually leads to omitting certain observations. Those are called the OOB data and serve as a validation dataset for computing the MSE of the trained model. Subsequently, individual features are randomly shuffled and the MSE is computed again. The difference is recorded and then averaged and normalized for all trees, which results in the overall statistic.

Table 5.13: Variable importance of RF models - the % increase in the MSE

	Dependent variable:					
	SPOLU	ANO	PirSTAN	SPD	CSSD	KSCM
longitude	12.81	17.00	36.71	19.13	5.46	0.79
latitude	12.68	12.90	41.38	2.22	6.59	2.34
turnout	43.07	41.86	7.41	7.78	0.76	1.87
inhabitants_log	7.11	7.43	4.37	3.42	2.54	1.29
inhabitants_over_64	12.63	29.02	2.63	1.58	1.74	2.60
unemployment	8.82	9.52	4.33	1.58	1.08	0.71
distraint	13.38	8.27	3.61	4.26	1.20	0.77
bankruptcy	10.82	7.25	3.81	3.36	0.91	1.52
covid_vaccination	2.87	5.76	5.18	29.29	2.26	0.17
covid_cases	1.53	0.59	0.88	0.44	0.69	0.61
population_density_log	7.77	5.31	3.25	1.40	1.06	0.85
believers	53.78	13.60	11.64	4.63	5.88	2.00
economically_active	4.17	3.37	1.84	1.03	0.50	0.30
entrepreneurs	75.41	59.76	17.51	5.91	2.40	4.28
roma_people	-0.07	0.03	0.04	0.03	0.02	-0.03
primary_education	26.12	19.34	15.71	3.86	0.99	2.31
university_education	60.56	31.54	14.08	5.93	1.07	2.41
immigrated	1.70	2.13	0.94	0.97	0.78	0.31
emmigrated	1.77	2.09	1.07	0.82	1.31	0.35
died	2.14	2.83	0.91	0.72	0.52	0.48
born	1.75	2.04	1.06	0.73	0.40	0.26
near_Prague	5.30	3.66	5.45	0.84	0.12	0.06
near_city	0.20	0.28	0.34	0.39	0.06	0.02
regular_exp_pc_log	3.29	1.73	1.17	0.43	0.27	0.36
total_exp_pc_log	2.12	1.72	1.26	0.78	0.37	0.44
non_tax_inc_pc_log	1.15	1.05	1.42	0.56	0.53	0.35
capital_inc_pc_log	1.22	0.45	0.38	0.32	0.00	0.07
bal_sheet_br_pc_log	1.96	1.75	0.91	0.85	0.21	0.16
gas_piping	0.16	0.16	0.70	0.09	0.03	0.03
water_piping	0.18	0.13	0.44	0.05	0.02	0.03
sewerage	0.88	0.22	0.24	0.06	0.03	0.05
Mean of sq. residuals	0.00399	0.00399	0.00206	0.00145	0.00076	0.00069
% Var explained	41.42	31.97	30.90	27.98	13.74	12.37

Note: Due to better readability, all values are multiplied by *E+05*.

Table 5.14: Variable importance of RF models - the increase in node purity

	Dependent variable:					
	SPOLU	ANO	PirSTAN	SPD	CSSD	KSCM
longitude	0.582	0.647	1.258	0.343	0.107	0.038
latitude	0.522	0.490	1.053	0.098	0.167	0.068
turnout	3.476	3.065	0.289	0.408	0.036	0.041
inhabitants_log	1.481	0.780	0.638	0.210	0.051	0.067
inhabitants_over_64	0.642	0.575	0.182	0.084	0.059	0.061
unemployment	0.684	0.742	0.259	0.116	0.043	0.035
distraint	1.039	1.103	0.177	0.157	0.043	0.030
bankruptcy	2.100	1.232	0.209	0.158	0.037	0.045
covid_vaccination	0.298	0.216	0.335	1.049	0.047	0.021
covid_cases	0.287	0.241	0.247	0.076	0.046	0.042
population_density_log	0.682	0.482	0.387	0.145	0.034	0.088
believers	0.902	0.424	0.494	0.109	0.110	0.051
economically_active	0.492	0.477	0.195	0.078	0.037	0.031
entrepreneurs	5.372	4.394	1.530	0.921	0.069	0.146
roma_people	0.084	0.072	0.072	0.018	0.020	0.009
primary_education	5.934	3.710	2.130	1.222	0.042	0.114
university_education	6.100	3.335	1.323	0.589	0.058	0.201
immigrated	0.386	0.273	0.213	0.088	0.043	0.031
emmigrated	0.235	0.249	0.145	0.075	0.042	0.031
died	0.469	0.295	0.141	0.077	0.040	0.039
born	0.241	0.266	0.138	0.078	0.034	0.031
near_Prague	2.207	1.402	1.853	0.319	0.003	0.023
near_city	0.021	0.019	0.016	0.009	0.004	0.006
regular_exp_pc_log	0.577	0.390	0.273	0.105	0.029	0.030
total_exp_pc_log	0.377	0.320	0.215	0.098	0.030	0.034
non_tax_inc_pc_log	0.224	0.187	0.175	0.061	0.035	0.030
capital_inc_pc_log	0.195	0.174	0.126	0.052	0.027	0.025
bal_sheet_br_pc_log	0.359	0.294	0.187	0.109	0.030	0.029
gas_piping	0.018	0.014	0.016	0.005	0.002	0.002
water_piping	0.013	0.011	0.012	0.004	0.002	0.002
sewerage	0.022	0.016	0.010	0.007	0.002	0.002
Mean of sq. residuals	0.00399	0.00399	0.00206	0.00145	0.00076	0.00069
% Var explained	41.42	31.97	30.90	27.98	13.74	12.37

related to the fact that the movement succeeds in both the western and the eastern edge of the country and thus the east-west dimension is more important.

In general, all RF models evince similar goodness of fit as their OLS counterparts. Simultaneously, the average magnitudes of the variable importance decrease with the percentage of explained variance. Surprisingly, the movement SPD reports a very high value on *covid_vaccination*, which is by far the most important variable for the party and also the highest value among other models. The increased impact of the vaccination is reported also in previous results, however, it is clearly not as strong as in this case.

When considering the increase in the node purity, the values are (in relative terms) similar to the % increase in the MSE. This implies that the variables such as *entrepreneurs*, *university_education*, or *primary_education* decrease the most of the variance when used to split a tree node. Similarly as in the previous case, voter turnout seems to be very relevant for the results of the SPOLU or the ANO, and there is an increased importance of the variable describing whether a municipality is located close to the city of Prague.

On the one hand, interpreting the RF results is quite challenging since the algorithm does not uncover specific relationships between the independent variables and the voting results. On the other hand, it estimates the importance of particular variables and aims to determine which features play an important role when constructing the trees.

Even though the RF is an ensemble method, and it is meaningless to interpret its individual decision trees, it is possible to generate a representative tree which aims to approximate the whole ensemble. Despite the fact that it is feasible to visualize only a limited depth, it presents the decision rules and illustrates the process of the tree construction. In Appendix D, there can be seen two examples of the representative trees in Figures D.2 - D.5.

The next step of the ML analysis is the estimation of the GRF model. This method, which has been developed exclusively for the analysis of spatial data, also requires hyperparameter tuning in order to find the correct setting of the model. In this case, it is necessary to determine a suitable kernel (including the corresponding radius or the number of neighbours), and the way to estimate the feature importance of the dependent variable. Simultaneously, the procedure optimizes standard hyperparameters specifying the number of features used in a node split or the number of estimated trees. Lastly, when using the model to predict values for unseen data, it is necessary to determine the weight of the local model, which is used for the prediction.

Table 5.15 presents the most successful combinations of hyperparameters, reaching the lowest MSE values. Since the GRF is far more computationally demanding than the RF algorithm, the whole procedure is performed on a subset of data. Nevertheless, when cross-validating the process on several random samples, the results are similar and thus the best model from Table 5.15 is selected for the analysis. Unfortunately, the computation time of the GRF increasingly grows with the rising number of observations and the chosen setup is not feasible. Therefore, the estimated number of trees has to be decreased from 400 to 100.

Table 5.15: The example of hyperparameter tuning, the GRF

kernel	distance	importance	features	est. trees	loc. imp.	MSE
fixed	20	permutation	7	400	1	0.00204
fixed	20	impurity	7	400	1	0.002041
fixed	20	impurity	7	400	0.75	0.002041
fixed	20	impurity	7	400	0.5	0.002044
fixed	30	impurity	15	400	0.5	0.002048
fixed	30	impurity	15	400	0.75	0.002049
fixed	20	permutation	7	400	0.75	0.002051
fixed	20	permutation	15	400	1	0.002053
fixed	30	impurity	15	400	1	0.002053
fixed	20	permutation	15	400	0.75	0.002057

As it has been previously mentioned, the GRF extends the RF algorithm by estimating also a local model for every observation. Therefore, apart from the variable importance from the global model, it is possible to report the average variable importance estimated from the local models. Not surprisingly, Table 5.16 shows that the global model results strongly correspond with the % increase in the MSE reported by the RF method. In this case, particular variables only slightly differ in magnitudes, which is probably caused by omitting the geographical coordinates.

When evaluating the average variable importance from local models, presented in Table 5.17, it is obvious that the results do not differ significantly. There can be observed minor distinctions, such as that voter turnout and university education are more important for the coalition SPOLU in local terms, as well as the share of pensioners for the movement ANO. However, neither political subjects nor independent variables significantly deviate from the rest.

According to Georganos *et al.* (2021), the alternative option to present the GRF results is to visualize them in a map, either by plotting the importance of particular variables or by showing the goodness of fit of local models. Figure 5.2

Table 5.16: Variable importance of the GRF - global model

	Dependent variable:					
	SPOLU	ANO	PirSTAN	SPD	CSSD	KSCM
turnout	58.3	53.42	16.11	16.68	1.64	3.53
inhabitants_log	18.14	12.78	9.92	7.62	3.85	3.21
inhabitants_over_64	20.72	38.74	7.42	4.61	2.02	3.01
unemployment	14.37	11.6	6.04	2.16	1.09	1.89
distraint	17.57	13.73	4.73	7.65	2.93	1.95
bankruptcy	13.40	9.97	5.98	5.95	1.48	3.26
covid_vaccination	3.42	8.77	5.73	37.51	2.88	0.92
covid_cases	1.21	0.81	2.98	1.49	0.44	1.50
population_density_log	15.39	12.93	10.68	3.69	2.11	4.59
believers	61.49	19.83	29.79	5.08	6.08	1.64
economically_active	6.28	6.72	3.57	2.65	0.79	0.54
enterpreneurs	77.82	64.66	24.93	8.64	3.97	4.40
roma_people	0.17	0.25	0.03	0.09	0.05	0.03
primary_education	28.57	29.46	24.03	4.96	2.04	3.15
university_education	76.85	45.09	22.94	9.91	2.41	3.71
immigrated	1.97	2.91	2.79	1.41	0.91	0.61
emmigrated	2.27	2.95	2.13	2.24	0.94	0.54
died	3.28	4.85	1.98	1.16	0.67	0.62
born	0.20	3.99	1.10	1.42	0.65	0.70
near_Prague	6.43	4.26	5.72	1.68	0.52	0.07
near_city	0.67	0.58	0.90	1.52	0.14	-0.01
regular_exp_pc_log	4.88	5.05	2.55	2.15	0.57	1.47
total_exp_pc_log	3.98	5.81	4.95	1.51	1.11	1.70
non_tax_inc_pc_log	2.50	4.43	2.13	1.31	0.86	0.40
capital_inc_pc_log	2.25	0.74	1.21	0.05	0.44	0.30
bal_sheet_br_pc_log	4.85	2.20	1.72	1.13	0.21	0.48
gas_piping	0.94	0.59	2.38	0.35	0.15	0.50
water_piping	-0.16	0.32	0.44	-0.01	-0.12	0.20
sewerage	1.05	0.05	0.24	0.20	0.14	0.07
R^2	0.3964	0.2951	0.2068	0.2341	0.0907	0.0985

Note: Due to better readability, all values are multiplied by *E+05*.

depicts the importance of *entrepreneurs* and shows that, even though there are small groups of municipalities with the same value, the overall situation is ambiguous and does not suggest that this variable is more important in specific regions. Similarly, Figure 5.3 presents the R^2 of local models for the movement ANO and does not evince any regions where the models perform better.

Table 5.17: Mean variable importance of the GRF - local models

	Dependent variable:					
	SPOLU	ANO	PirSTAN	SPD	CSSD	KSCM
turnout	62.06	50.39	16.40	17.36	2.13	4.24
inhabitants_log	17.09	15.02	11.24	6.95	3.73	2.96
inhabitants_over_64	21.77	43.42	6.00	3.87	2.37	3.50
unemployment	12.05	12.05	5.60	3.52	1.35	1.46
distraint	21.12	15.65	5.73	9.84	2.51	1.84
bankruptcy	12.80	12.00	5.84	5.12	1.09	2.76
covid_vaccination	5.29	10.10	6.99	37.63	2.72	0.60
covid_cases	2.44	1.79	2.56	1.10	0.58	1.10
population_density_log	15.01	11.12	11.88	4.39	2.40	3.90
believers	62.03	19.89	31.30	5.21	6.65	1.66
economically_active	5.98	4.60	3.57	2.55	0.94	0.61
enterpreneurs	77.70	67.23	23.56	7.86	3.22	4.68
roma_people	0.13	0.18	0.14	0.06	0.01	0.01
primary_education	28.70	28.31	21.76	6.91	1.95	3.87
university_education	85.62	46.88	24.56	10.38	2.66	3.44
immigrated	3.86	3.74	2.68	1.91	1.07	0.51
emmigrated	3.42	3.90	2.17	1.91	1.25	0.76
died	4.40	6.07	2.45	1.49	0.89	0.81
born	1.86	3.37	1.75	1.40	0.48	0.60
near_Prague	6.53	4.41	6.78	1.65	0.44	0.13
near_city	0.54	0.54	0.75	1.38	0.11	0.08
regular_exp_pc_log	7.01	4.70	2.80	1.77	1.04	1.26
total_exp_pc_log	5.21	4.38	4.15	1.98	1.16	1.17
non_tax_inc_pc_log	3.07	2.86	2.52	1.23	1.28	0.63
capital_inc_pc_log	2.00	1.12	0.80	0.38	0.36	0.18
bal_sheet_br_pc_log	4.14	3.43	2.24	1.76	0.77	0.33
gas_piping	0.67	0.83	3.81	0.55	0.17	0.27
water_piping	0.14	0.52	0.59	0.21	-0.01	0.06
sewerage	1.21	0.32	0.67	0.19	0.05	0.11
OOB MSE	0.00411	0.00413	0.00236	0.00152	0.00077	0.00070

Note: Due to better readability, all values are multiplied by *E+05*.

Nevertheless, the absence of spatial patterns in variable importance or R^2 does not reflect the quality of estimated models. On the contrary, a correctly specified model is expected to perform well in all regions and any spatial trends might indicate unobserved factors or other pitfalls. Similarly, the resemblance of the RF and the GRF results does not imply anything about their goodness of fit. Therefore, it is desirable to evaluate the performance of both models and compare their ability to handle spatial data.

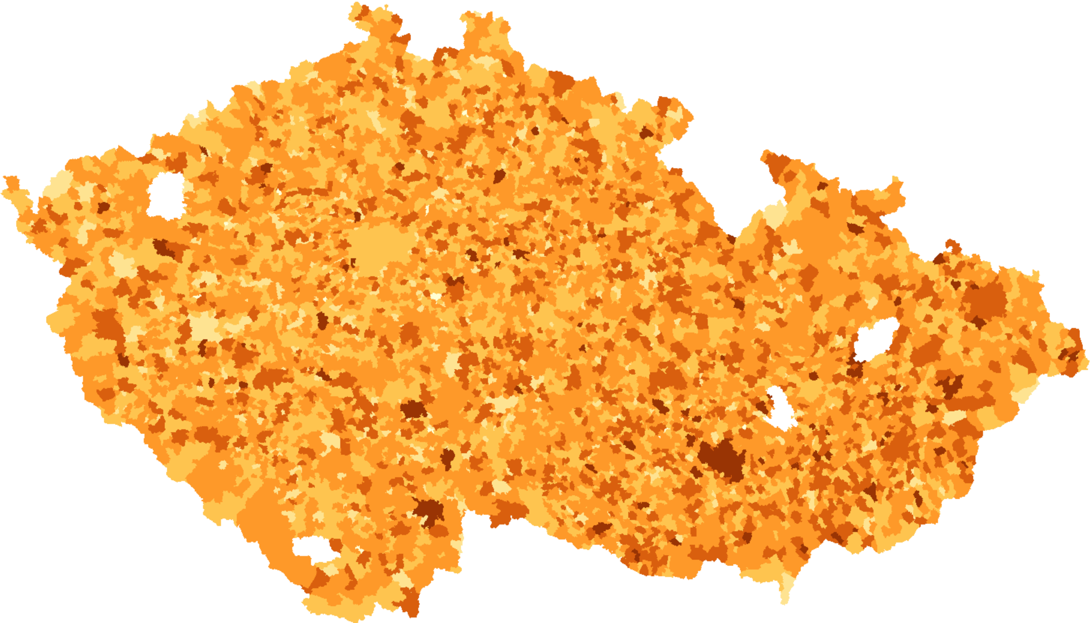


Figure 5.2: Estimated variable importance of *entrepreneurs* (the SPOLU)

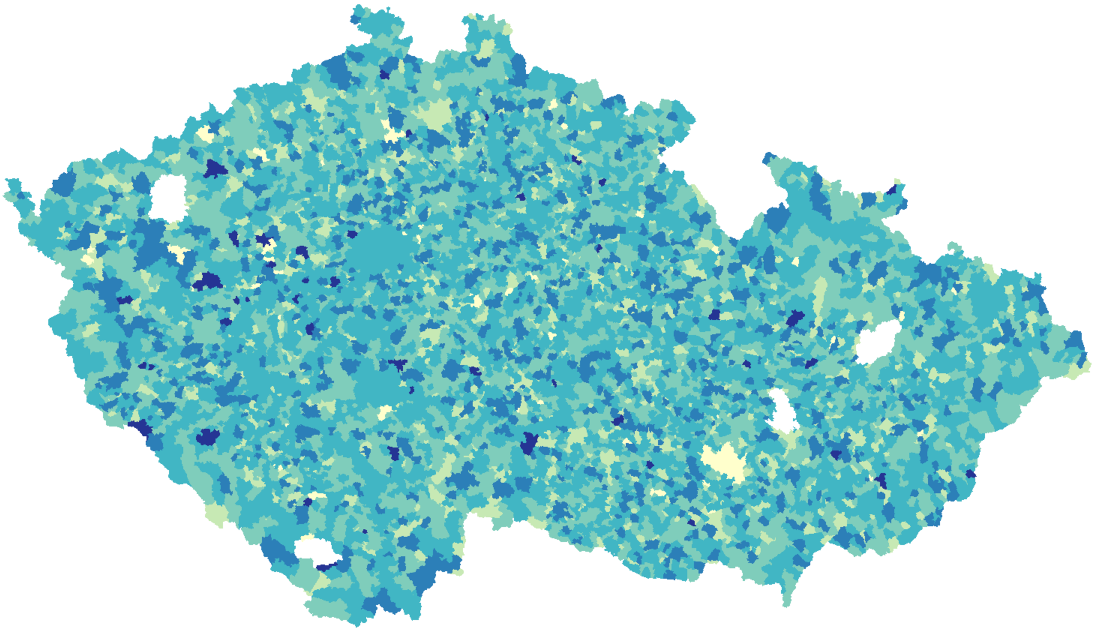


Figure 5.3: R^2 of local models (the ANO)

5.5 Comparison of Used Methods

As it can be seen in previous sections, ML models provide different types of results, which cannot be directly compared to other methods. Simultaneously, both ML algorithms produce similar outcomes and it is not clear whether the GRF, which is designed for the analysis of spatial data, outperforms its non-spatial counterpart.

In order to directly compare all approaches, it is possible to test their ability to predict the election result for unseen data. For this purpose, 6,000 observations are randomly selected as train data and remaining 254 observations serve for testing. The average prediction error (MSE) of all models, estimated for the coalition SPOLU, is presented in Table 5.18.

Firstly, the performance on train data suggests that ML methods manage to fit data better than linear models, especially the GRF algorithm, which evinces a significantly lower MSE value than other methods. Nevertheless, this performance is not relevant unless confirmed or rejected based on test data.

Secondly, the result of the SDEM model, which is in this case the initial model trained on complete data, shows that this method performs the worst when predicting the outcome of election. Therefore, since its training error is significantly higher than the test errors of other methods, and also due to high computational requirements, the SDEM model is not estimated further to evaluate the performance on unseen observations.

Finally, the results for test data express the overall performance of particular methods. Linear models show an even increase of the MSE and the OLS still performs slightly better than the WLS. In the case of the GRF, there is observed the most significant increase of the MSE, which ranks the model behind the OLS. The RF also evinces a decent increase of prediction errors, however, it reaches the lowest MSE value among all methods. Importantly, the results stay the same when repeating the procedure for different samples of data.

Table 5.18: MSE of all estimated models (the SPOLU)

	OLS	**WLS**	**SDEM**	**RF**	**GRF**
train data	0.00384	0.00404	0.01036	0.00294	0.00003
test data	0.00505	0.00532	-	0.00501	0.00527

In general, the results show that tree-based methods fit train data better,[10] but evince a relatively higher increase of the MSE for the testing sample. Surprisingly, linear methods do not perform significantly worse and the OLS ranks closely behind the RF method. On the contrary, the SDEM model cannot predict election results very well. Theoretically, this might imply that the spatial lag of error terms, which is not used to compute fitted values, plays a significant role when explaining the results. Therefore, the thesis shows that methods with greater explanatory power evince worse predictive power and vice versa.

[10] The case of the GRF evinces a severe overfitting problem, which, however, stems from the definition of the algorithm.

Chapter 6

Discussion

As it has been previously mentioned, many academic papers analysing election results evince certain shortcomings, for instance, the disregard of spatial trends within data or the insufficient model specification leading to the use of sub-optimal methods. The thesis attempts to address many of those pitfalls and confirms that the concerns of some authors (Cook *et al.* 2020; LeSage 2014) are relevant also in the case of the Czech Parliamentary election in 2021.

Firstly, the analysis provides evidence of spatial trends within election results and corresponding municipal characteristics, which implies that spatial econometric methods should be applied to complement the baseline analysis. In order to specify a suitable model, the thesis thoroughly compares and evaluates various types of models and weight matrices. This procedure provides both empirical evidence and economic reasoning for the SDEM, indicating that this type of model might be indeed under-utilized in this field.

Therefore, the use of the SDEM and the interpretation of direct and indirect effects might enrich related academic literature, showing that the spatial lags of independent variables are probably far more relevant in terms of explaining election results than the autoregressive processes examined by numerous papers. Considering particular results, it is evident that the support of political subjects is relatively polarized and strongly corresponds with the socio-economic level of regions. Therefore, municipalities that lag behind in many aspects are associated with different political subjects than municipalities that are better off. In accordance with related literature, the 'worse' regions are usually associated with increased support of movements that are sometimes considered to be populist.

The incorporation of ML methods attempts to provide an alternative approach to the field of election analyses. The algorithms work on different principles and might thus provide various (dis)advantages when compared to the spatial econometric framework. The naive approach, based on including the coordinates of observations and estimating the RF, performs surprisingly well. Despite the fact that it does not provide similar coefficients as standard econometric methods, it can estimate the importance of particular variables (using various approaches). Simultaneously, this algorithm performs the best when predicting test values and requires a significantly lower amount of estimation time when compared to the GRF or spatial econometric models.

Even though the thesis attempts to thoroughly analyse the outcome of election and address many related problems, it faces several limitations and offers the space for improvement. Firstly, the ability to explain the election result is limited by the availability of data. In this case, numerous variables that might significantly help to describe particular municipalities are often not publicly available or not collected at all. For instance, it might be very useful to monitor more characteristics of inhabitants, such as their field of employment and wage, or to describe particular municipalities by inspecting the availability of goods and services or by incorporating other phenomena such as crime rate or housing prices.

As it has been already discussed, the socio-economic variables are very relevant in terms of explaining the election results. Unfortunately, some of them are available only from the census, which is performed once in ten years, and the CZSO publishes them with a significant delay. This means that more than one year after the last census, the results on the municipal level are not available and the thesis thus has to use some data from 2011. Similarly, research organizations working on interesting projects almost never share their data in a usable format, despite the fact they present all of them online, and complicate the consequent analyses.

Another possible improvement of the analysis is based on increasing the computing power dedicated to model training. Due to the use of weight matrices, spatial econometric models are computationally very demanding for higher numbers of observations. Similarly, ML algorithms are based on hyperparameter tuning and therefore might require high computing capacity in order to estimate the model for every desired combination of hyperparameters. In addition, the GRF method creates a local model for every observation, which slows down the estimation process significantly. Therefore, the thesis has to use smaller samples of data to make the hyperparameter tuning feasible, which is obviously not the optimal scenario.

In general, the ML analysis performed in the thesis might be extended in several ways. Firstly, when utilizing non-spatial algorithms, there might be a more suitable way to incorporate the spatial dimension of data than including the geographical coordinates. Secondly, there might be a better approach to interpret the results of methods based on random forests, possibly trying to estimate the impacts of particular variables. Lastly, the analysis might be replicated by other algorithms that are able to account for the spatial dimension of the data, such as the SVM, or the Self-Organizing Maps.

Chapter 7

Conclusion

The thesis performs the analysis of the Czech Parliamentary election in 2021 and contributes to related academic literature in two ways. Firstly, it thoroughly specifies a suitable spatial method and provides both empirical and economic evidence for the Spatial Durbin Error Model (SDEM) model. This method, which is not often utilized in election analyses, enables the examination of the direct and indirect effects of particular variables, i.e., distinguishes the impact of municipality characteristics from the effect of neighbouring units, and also accounts for unobservable effects. Both phenomena are considered to be relevant in terms of the election result and thus should be addressed by the model.

Secondly, the thesis replicates the spatial analysis using two ML algorithms. Those methods, which are primarily used in other fields, provide the relative importance of particular variables and complement the spatial econometric framework. Since the algorithms differ from standard econometric methods, it is possible to incorporate the coordinates of observations into data and apply non-spatial methods, such as the Random Forest (RF). This approach is relatively straightforward and provides similar results as its spatial extension, which is, similarly to spatial econometric models, significantly more demanding in computing terms.

In general, particular results of the analysis confirm many hypotheses provided by academic literature, assuming that voters from disadvantaged regions tend to support anti-establishment parties, or that variables describing the socio-economic characteristics of inhabitants play the most significant role when explaining election results. In the spatial analysis, variables describing the inhabitants, such as the share of entrepreneurs or university graduates, evince stronger direct effects than variables characterizing municipalities, which are more likely to affect the election results indirectly.

Bibliography

Abramowitz, A. I. (1988): "Explaining senate election outcomes." *American Political science review* **82(2)**: pp. 385–403.

Akarca, A. T. & A. Tansel (2006): "Economic performance and political outcomes: An analysis of the turkish parliamentary and local election results between 1950 and 2004." *Public Choice* **129(1)**: pp. 77–105.

Amara, M. & A. El Lahga (2016): "Tunisian constituent assembly elections: how does spatial proximity matter?" *Quality & Quantity* **50(1)**: pp. 65–88.

Anselin, L. (1988): *Spatial econometrics: methods and models (Vol. 4).* Springer Science & Business Media.

Anselin, L. (2010): "Thirty years of spatial econometrics." *Papers in regional science* **89(1)**: pp. 3–25.

Becker, S. O., O. Fetzer, & D. Novy (2017): "Who voted for brexit? A comprehensive district-level analysis." *Economic policy* **32(92)**: pp. 601–650.

Bojar, A. & T. Vlandas (2021): "Group-Specific Responses to Retrospective Economic Performance: A Multilevel Analysis of Parliamentary Elections." *Politics & Society* **49(4)**: pp. 518–548.

Breiman, L., J. H. Friedman, R. A. Olshen, & C. J. Stone (2017): *Classification and regression trees.* Routledge.

Burkey, M. L. (2018): "A short course on spatial econometrics and GIS." *MPRA* **5(3)**: pp. 13–18.

Burnett, C. M. & V. Kogan (2017): "The politics of potholes: Service quality and retrospective voting in local elections." *The Journal of Politics* **79(1)**: pp. 302–314.

Burnett, J. W. & D. J. Lacombe (2012): "Accounting for spatial autocorrelation in the 2004 presidential popular vote: A reassessment of the evidence." *Review of Regional Studies* **42(1)**: pp. 75–89.

Carkoglu, A. & M. J. Hinich (2016): "A spatial analysis of turkish party preferences." *Electoral Studies* **25(2)**: pp. 369–392.

Constantino, S. M., A. D. Cooperman, & T. M. Moreira (2021): "Voting in a global pandemic: Assessing dueling influences of Covid-19 on turnout." *Social Science Quarterly* **102(5)**: pp. 2210–2235.

Cook, S. J., J. C. Hays, & R. Franzese (2020): "Model specification and spatial interdependence." *The SAGE Handbook of Research Methods in Political Science and International Relations* pp. 730–747.

Deniz, P., B. C. Karahasan, & M. Pinar (2021): "Determinants of regional distribution of akp votes: Analysis of post-2002 parliamentary elections." *Regional Science Policy & Practice* **13(2)**: pp. 323–352.

Elhorst, J. P. (2010): "Applied Spatial Econometrics: Raising the Bar." *Spatial economic analysis* **5(1)**: pp. 9–28.

Fauvelle-Aymar, C. & A. Francois (2006): "The impact of closeness on turnout: An empirical relation based on a study of a two-round ballot." *Public Choice* **107(3/4)**: pp. 469–491.

Fiorino, N., N. Pontarollo, & R. Ricciuti (2021): "Spatial links in the analysis of voter turnout in european parliamentary elections." *Letters in Spatial and Resource Sciences* **14(1)**: pp. 65–78.

Fotheringham, A. S., Z. Li, & L. J. Wolf (2021): "Scale, Context, and Heterogeneity: A Spatial Analytical Perspective on the 2016 U.S. Presidential Election." *Annals of the American Association of Geographers* **111(6)**: pp. 1602–1621.

Franke, G. R. (2010): *Multicollinearity.* Wiley international encyclopedia of marketing.

Georganos, S., T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff, & S. Kalogirou (2021): "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling." *Geocarto International* **36(2)**: pp. 121–136.

Geys, B. (2006): "Explaining voter turnout: A review of aggregate-level research." *Electoral studies* **25(4)**: pp. 637–663.

Glass, A. J., K. Kenjegalieva, & R. Sickles (2012): "The economic case for the spatial error model with an application to state vehicle usage in the us." *Rice University manuscript* pp. 1–36.

Goodwin, M. J. & O. Heath (2016): "The 2016 referendum, Brexit and the left behind: An aggregate-level analysis of the result." *The Political Quarterly* **87(3)**: pp. 323–332.

Hanley, S. & M. A. Vachudova (2018): "Understanding the illiberal turn: democratic backsliding in the Czech Republic." *East European Politics* **34(3)**: pp. 276–296.

Havlík, M. & P. Voda (2016): "The rise of new political parties and realignment of party politics in the czech republic." *Acta Politologica* **8(2)**: pp. 119–144.

Jensen, C. D., D. J. Lacombe, & S. G. McIntyre (2013): "A bayesian spatial econometric analysis of the 2010 uk general election." *Papers in regional science* **92(3)**: pp. 651–666.

Joppke, C. (2015): *The secular state under siege: Religion and politics in Europe and America.* John Wiley & Sons.

Kim, J., E. Elliott, & D. M. Wang (2003): "A spatial analysis of county-level outcomes in US Presidential elections: 1988–2000." *Electoral Studies* **22(4)**: pp. 741–761.

Kinsella, C. J. (2013): "Political geography of the south: A spatial analysis of the 2008 presidential election." *American Review of Politics* **13(1)**: pp. 227–240.

Kopczewska, K. (2021): "Spatial machine learning: new opportunities for regional science." *The Annals of Regional Science* **68(3)**: pp. 1–43.

LASOŃ, A. & A. TORÓJ (2019): "Anti-liberal, anti-establishment, or constituency-driven spatial econometric analysis of polish parliamentary election results in 2015." *European Spatial Research and Policy* **2(1)**: pp. 199–236.

LESAGE, J. & R. K. PACE (2009): *Introduction to Spatial Econometrics.* Chapman and Hall/CRC.

LESAGE, J. P. (2008): "An introduction to spatial econometrics." *Revue d'économie industrielle* **123(3)**: pp. 19–44.

LESAGE, J. P. (2014): "What regional scientists need to know about spatial econometrics." *The Review of Regional Studies* **44(1)**: pp. 13–32.

LI, M., E. PERRIER, & C. XU (2019): "Deep hierarchical graph convolution for election prediction from geospatial census data." *In Proceedings of the AAAI conference on artificial intelligence* **33(1)**: pp. 647–654.

LIPSET, S. & S. ROKKAN (1967): *Cleavage structures, party systems, and voter alignments: an introduction.* Toronto: The Free Press.

LIU, R., X. YAO, C. GUO, & X. WEI (2021): "Can we forecast presidential election using twitter data? an integrative modelling approach." *Annals of GIS* **27(1)**: pp. 43–56.

LYSEK, J., J. PÁNEK, & T. LEBEDA (2020): "Mapping the 2020 slovak parliamentary election." *Czech Journal of Political Science* **27(3)**: pp. 278–302.

LYSEK, J., J. PÁNEK, & T. LEBEDA (2021): "Who are the voters and where are they? using spatial statistics to analyse voting patterns in the parliamentary elections of the czech republic." *Journal of Maps* **17(1)**: pp. 33–38.

MANSLEY, E. & U. DEMŠAR (2015): "Space matters: Geographic variability of electoral turnout determinants in the 2012 London mayoral election." *Electoral Studies* **40(1)**: pp. 322–334.

MAŠKARINEC, P. (2017): "A spatial analysis of czech parliamentary elections, 2006–2013." *Europe-Asia Studies* **69(3)**: pp. 426–457.

MAŠKARINEC, P. (2019): "The rise of new populist political parties in czech parliamentary elections between 2010 and 2017: the geography of party replacement." *Eurasian Geography and Economics* **60(5)**: pp. 511–547.

Nikparvar, B. & J.-C. Thill (2021): "Machine learning of spatial data." *ISPRS International Journal of Geo-Information* **10(9)**: pp. 1–32.

Nwankwo, C. F. (2019): "The spatial pattern of voter choice homogeneity in the nigerian presidential elections in the fourth republic." *Bulletin of Geography* **43(1)**: pp. 143–165.

O'Loughlin, J., C. Flint, & L. Anselin (1994): "The geography of the nazi vote: Context, confession, and class in the reichstag election of 1930." *Annals of the association of American geographers* **84(3)**: pp. 351–380.

Otto, A. H. & M. F. Steinhardt (2014): "Immigration and election outcomes—Evidence from city districts in Hamburg." *Regional Science and Urban Economics* **45(1)**: pp. 67–79.

Ozen, I. C. & K. O. Kalkan (2017): "Spatial analysis of contemporary turkish elections: a comprehensive approach." *Turkish Studies* **18(2)**: pp. 358–377.

Pagliacci, F. & L. Bonacini (2021): "Explaining anti-immigrant sentiment through spatial analysis: a study of the 2019 European elections in Italy." *DEMB Working Paper Series* **192(1)**: pp. 1–40.

Parr, T., K. Turgutlu, C. Csiszar, & J. Howard (2018): *Beware Default Random Forest Importances.* explained.ai.

Paul, D., F. Li, M. K. Teja, X. Yu, & R. Frost (2017): "Compass: Spatio temporal sentiment analysis of us election what twitter says!" *In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* **1(1)**: pp. 1585–1594.

Praciano, B. J. G., J. P. C. L. da Costa, J. P. A. Maranhão, F. L. L. de Mendonça, R. T. de Sousa Júnior, & J. B. Prettz (2018): *Spatio-temporal trend analysis of the Brazilian elections based on Twitter data.* IEEE International Conference on Data Mining Workshops.

Rattinger, H. (1981): *Unemployment and the 1976 election in Germany: Some findings at the aggregate and the individual level of analysis.* Contemporary Political Economy.

Richardson, B., H. D. F. (2020): "Districts by demographics: Predicting US house of representative elections using machine learning and demographic data." **40(1)**: pp. 833–838.

RODRÍGUEZ-POSE, A. (2020): "The rise of populism and the revenge of the places that don't matter." *LSE Public Policy Review* pp. 1–9.

SZABO, B. & P. TATRAI (2016): "Regional and social cleavages in the slovak elections after the change of the regime." *Geografický časopis* **68(3)**: pp. 195–212.

WOOLDRIDGE, J. M. (2013): *Introductory Econometrics 5th ed.* Cengage Learning.

ČERNÝ, J. (2019): "Population Characteristics of Voters: Evidence from the Czech Parliamentary Election." *Charles University* pp. 1–40.

Appendix A

Variables

Table A.1: Descriptive statistics of the Czech Parliamentary Election in 2021

	Mean	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
number of voters	1 320	13 016	21	177	353	748	903 239
valid votes	857	8 823	17	122	239	511	627 399
turnout	0.6836	0.0753	0.2645	0.6400	0.6867	0.7317	1.0000
SPOLU	0.2495	0.0826	0.0000	0.1926	0.2450	0.3011	0.6571
ANO	0.2906	0.0767	0.0435	0.2400	0.2857	0.3351	0.6364
Pirati + STAN	0.1348	0.0546	0.0000	0.0991	0.1286	0.1624	0.5333
SPD	0.1061	0.0449	0.0000	0.0769	0.1017	0.1303	0.4396
PRISAHA	0.0505	0.0275	0.0000	0.0351	0.0472	0.0614	0.3810
CSSD	0.0509	0.0292	0.0000	0.0327	0.0466	0.0625	0.3636
KSCM	0.0448	0.0281	0.0000	0.0265	0.0400	0.0571	0.2439

Table A.2: Data sources and corresponding reference dates

Data	**Source**	**Reference Date**
code list of municipalities	czso.cz	January 2021
election results	volby.cz	October 2021
unemployment	data.mpsv.cz	November 2020 - October 2021
population	czso.cz	January 2021
gender and average age	czso.cz	January 2021
demography	czso.cz	January 2020 - December 2020
covid-19	uzis.cz	March 2020 - October 2021
distraint	mapaexekuci.cz	December 2017
bankruptcy	mapabankrotu.cz	December 2020
vaccination covid-19	uzis.cz	August 2021
population characteristics	czso.cz	March 2011
infrastructure	czso.cz	December 2010
overview of public finance	monitor.statnipokladna.cz	January 2020 - December 2020

Table A.3: Description of all independent variables collected for the analysis

Variable	**Description**
turnout	share of voters who participated in the election
inhabitants_log	number of inhabitants (logarithm)
inhabitants_over_64	share of people over 64 years
average_age	average age
unemployment	share of unemployed people (average value over last year)
distraint	share of people facing distraint(s)
bankruptcy	share of people facing bankruptcy
covid_vaccination	share of people vaccinated against covid-19
covid_cases	share of people infected by covid-19 (since March 2020)
population_density_log	population density (logarithm)
believers	share of believers
economically_active	share of economically active inhabitants
entrepreneurs	share of entrepreneurs
roma_people	share of Roma inhabitants
primary_education	share of inhabitants with primary education
highschool_education	share of inhabitants with high school education
university_education	share of inhabitants with university education
immigrated	proportional increase of population due to immigration
emigrated	proportional decrease of population due to emigration
born	proportional increase of population due to born children
died	proportional decrease of population due to passing
near_Prague	up to 30 km from Prague (dummy)
near_city	up to 15 km from a city with > 15k inhabitants (dummy)
regular_exp_pc_log	financing operation of municip. (in CZK per capita, log.)
total_exp_pc_log	reduction in municipal assets (in CZK p. c., log.)
non_tax_inc_pc_log	non-tax income acquired by activities of mun. (in CZK p. c., log.)
capital_inc_pc_log	resources to finance municip. investments (in CZK p. c., log.)
bal_sheet_br_pc_log	gross value of LT assets and liab. (in CZK p. c., log.)
gas_piping	gas piping installed (dummy)
water_piping	water piping installed (dummy)
sewerage	sewerage system installed (dummy)
tax_income_log	tax revenues of mun., imposed by law (in CZK p. c., log.)
total_profit_per_capita_log	increase in municipal assets (in CZK p. c., log.)
updated_budget_per_capita_log	updated municipal budget (in CZK p. c., log.)
capital_expenditures_per_capita_log	expenditures on financing investments (in CZK p. c., log.)
received_transfers_per_capita_log	subsidies/transfers from pub. budgets (in CZK p. c., log.)
inhabitants_under_15	share of inhabitants under 15 years
marriage	number of marriages with respect to population over 15 years
women_share	share of women
post_office	post office is situated within municipality (dummy)
divorce	number of divorces with respect to population over 15 years
kindergarten_school	kindergarten or school is situated within municipality (dummy)
health_service	healthcare center is situated within municipality (dummy)
abortion	number of abortions with respect to population over 15 years

Table A.4: Average results of the BMA

Variable	PIP	Post Mean	Post SD	Cond. Pos. Sign
believers	0.8652	0.0038	0.0047	0.3958
entrepreneurs	0.8571	-0.0027	0.0165	0.3333
primary_education	0.8185	0.0051	0.0161	0.4286
highschool_education	0.7659	-0.0055	0.0138	0.7126
gas_piping	0.7606	0.0007	0.0014	0.5714
covid_vaccination	0.7353	0.0154	0.0074	0.6667
average_age	0.7129	-0.0001	0.0004	0.5714
turnout	0.6917	0.0027	0.0103	0.5
near_Prague	0.5835	0.0021	0.0022	0.3333
near_city	0.5627	0.0002	0.001	0.2857
unemployment	0.5231	0.0055	0.0314	0.5
covid_cases	0.4801	0.0059	0.0092	0.4
inhabitants_over_64	0.4139	0.0132	0.0235	0.4286
bankruptcy	0.4124	0.0565	0.0734	0.4286
inhabitants_log	0.412	0.0006	0.0006	0.5
university_education	0.3913	0.0016	0.0212	0.2
regular_exp_pc_log	0.3284	-0.0005	0.001	0.5
population_density_log	0.297	-0.001	0.0004	0.4873
sewerage	0.2821	-0.0001	0.0008	0.5431
roma_people	0.2626	0.0067	0.1721	0.3333
non_tax_inc_pc_log	0.2189	-0.0001	0.0002	0.6403
economically_active	0.1739	0.0013	0.0051	0.5
bal_sheet_br_pc_log	0.1612	-0.0006	0.0006	0.5
emigrated	0.1464	-0.0153	0.006	0.75
water_piping	0.1369	-0.0006	0.0003	0.75
died	0.1119	-0.0322	0.0341	0.3333
capital_inc_pc_log	0.0875	0	0.0001	0.6443
total_exp_pc_log	0.0733	0.0002	0.0006	0.5758
distraint	0.0668	0.0016	0.0045	0.4
tax_income_per_capita_log	0.0587	0.0001	0.0011	0.5
marriage	0.0408	0.0111	0.0395	0.5
immigrated	0.037	0.0007	0.0077	0.2857
abortion	0.0306	-0.0032	0.0297	0.5
women_share	0.0224	-0.0013	0.0049	0.25
inhabitants_under_15	0.0218	-0.0001	0.0066	0.3267
born	0.0217	0.0002	0.0134	0.25
kindergarten_school	0.0205	-0.0001	0.0002	0.5
received_transfers_per_capita_log	0.0183	0	0.0001	0.7404
total_profit_per_capita_log	0.0162	0	0.0003	0.8077
divorce	0.0114	-0.0032	0.0179	0
capital_expenditures_per_capita_log	0.0084	0	0	1
health_service	0.0055	0	0.0001	0.5
post_office	0.0014	0	0.0001	0.5
updated_budget_per_capita_log	0	0	0	NaN

Appendix B

Maps

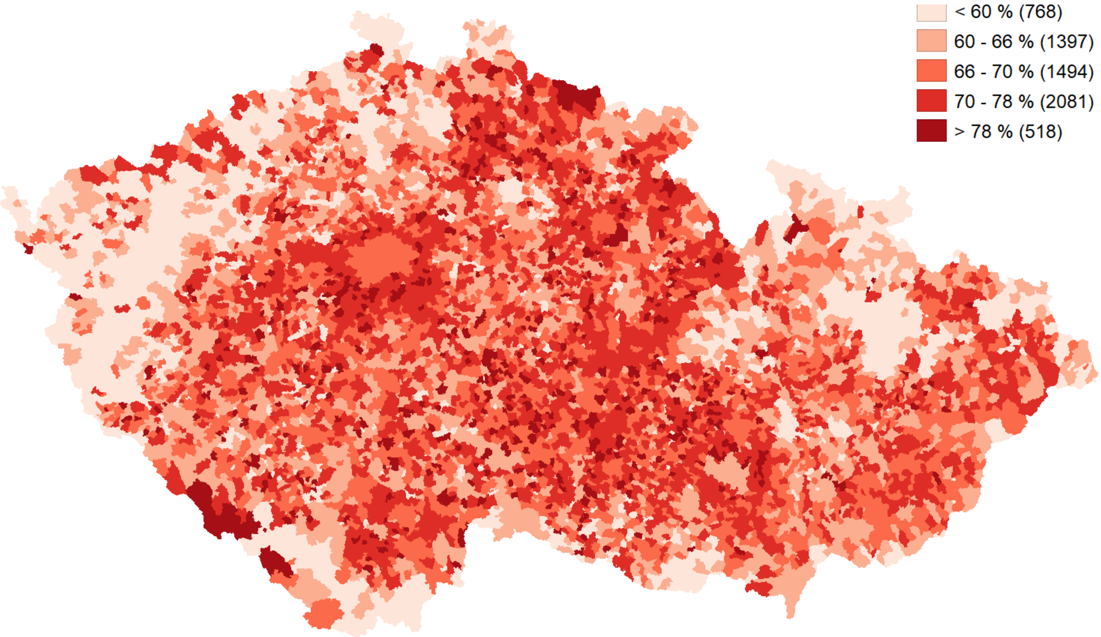


Figure B.1: Voter turnout in the Czech Parliamentary election in 2021

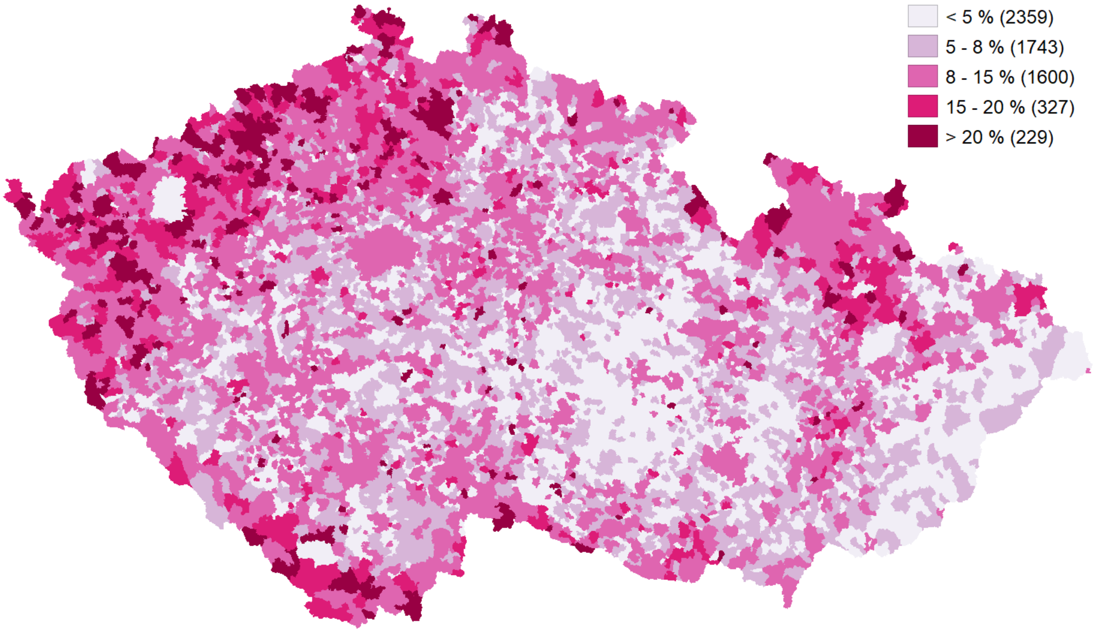


Figure B.2: The share of inhabitants facing distraints

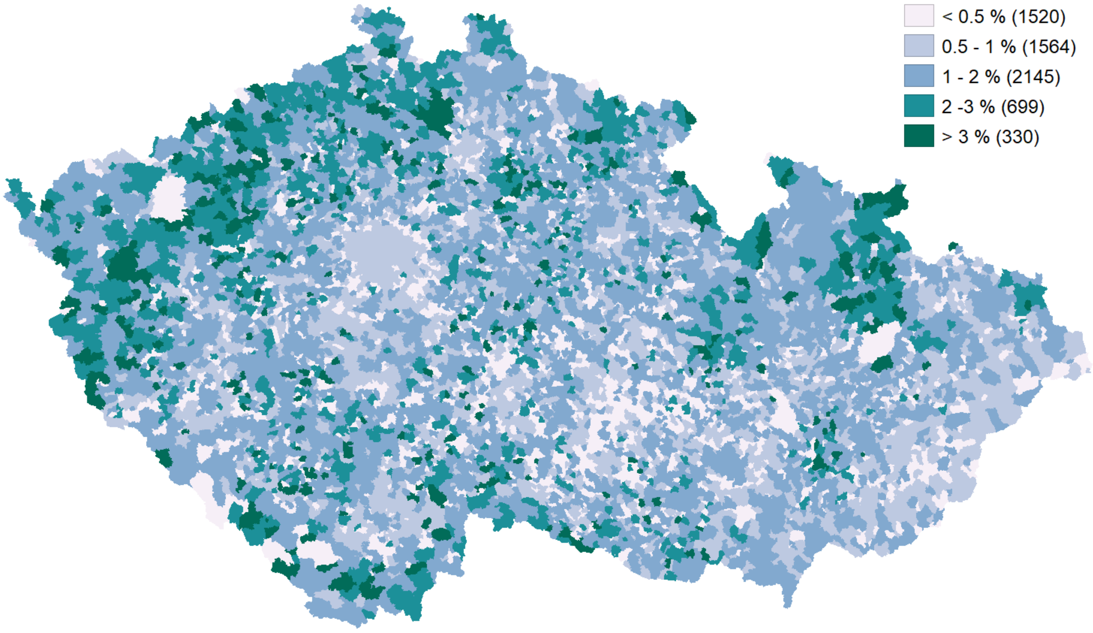


Figure B.3: The share of inhabitants facing bankruptcy

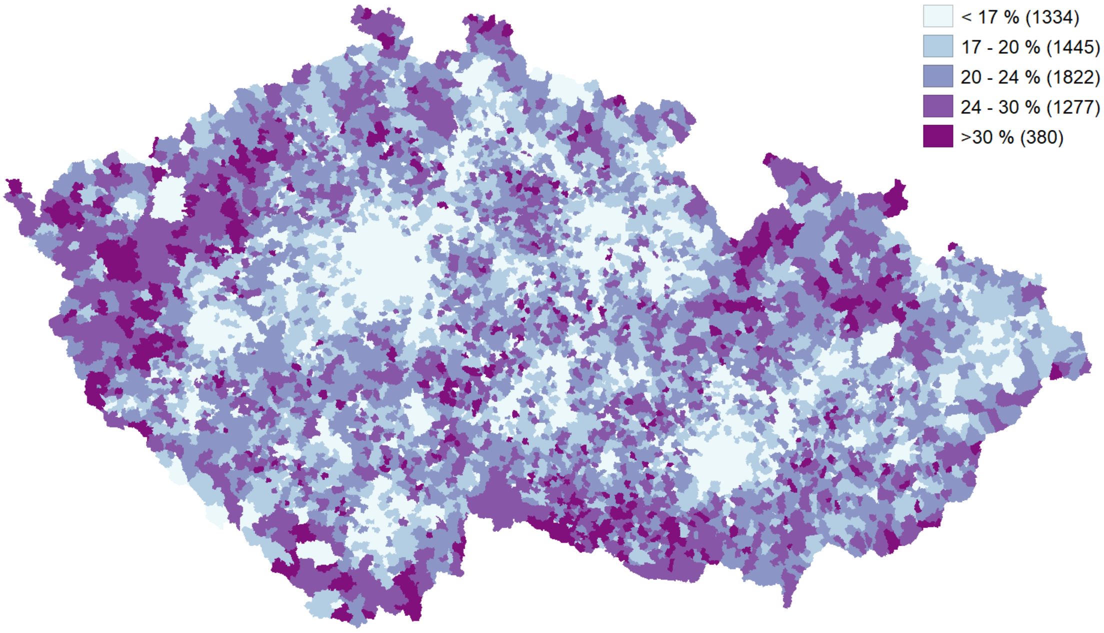


Figure B.4: The share of inhabitants with primary education

Appendix C

Methodology

OLS Assumptions

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term. Assumption MLR.1 describes the population relationship we hope to estimate, and explicitly sets out the β_j-the ceteris paribus population effects of the x_j on y-as the parameters of interest.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n$, following the population model in Assumption MLR.1. This random sampling assumption means that we have data that can be used to estimate the β_j, and that the data have been chosen to be representative of the population described in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables. Once we have a sample of data, we need to know that we can use the data to compute the OLS estimates, the $\hat{\beta}_j$. This is the role of Assumption MLR.3: if we have sample variation in each independent variable and no exact linear relationships among the independent variables, we can compute the $\hat{\beta}_j$.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the explanatory variables. In other words, $\mathbb{E}(u|x_1, x_2, \ldots, x_k) = 0$. As we discussed in the text, assuming that the unobserved factors are, on average, unrelated to the explanatory variables is key to deriving the first statistical property of each OLS estimator: its unbiasedness for the corresponding population parameter. Of course, all of the previous assumptions are used to show unbiasedness.

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any values of the explanatory variables. In other words,

$$Var(u|x_1, x_2, \ldots, x_k) = \sigma^2.$$

Compared with Assumption MLR.4, the homoskedasticity assumption is of secondary importance; in particular, Assumption MLR.5 has no bearing on the unbiasedness of the $\hat{\beta}_j$. Still, homoskedasticity has two important implications: (1) We can derive formulas for the sampling variances whose components are easy to characterize; (2) We can conclude, under the GaussMarkov assumptions MLR.1 to MLR.5, that the OLS estimators have smallest variance among all linear, unbiased estimators.

Assumption MLR.6 (Normality)

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance $\sigma^2 : u \sim (0, \sigma^2)$.

Citation from: Wooldridge (2013)

Appendix D

Results

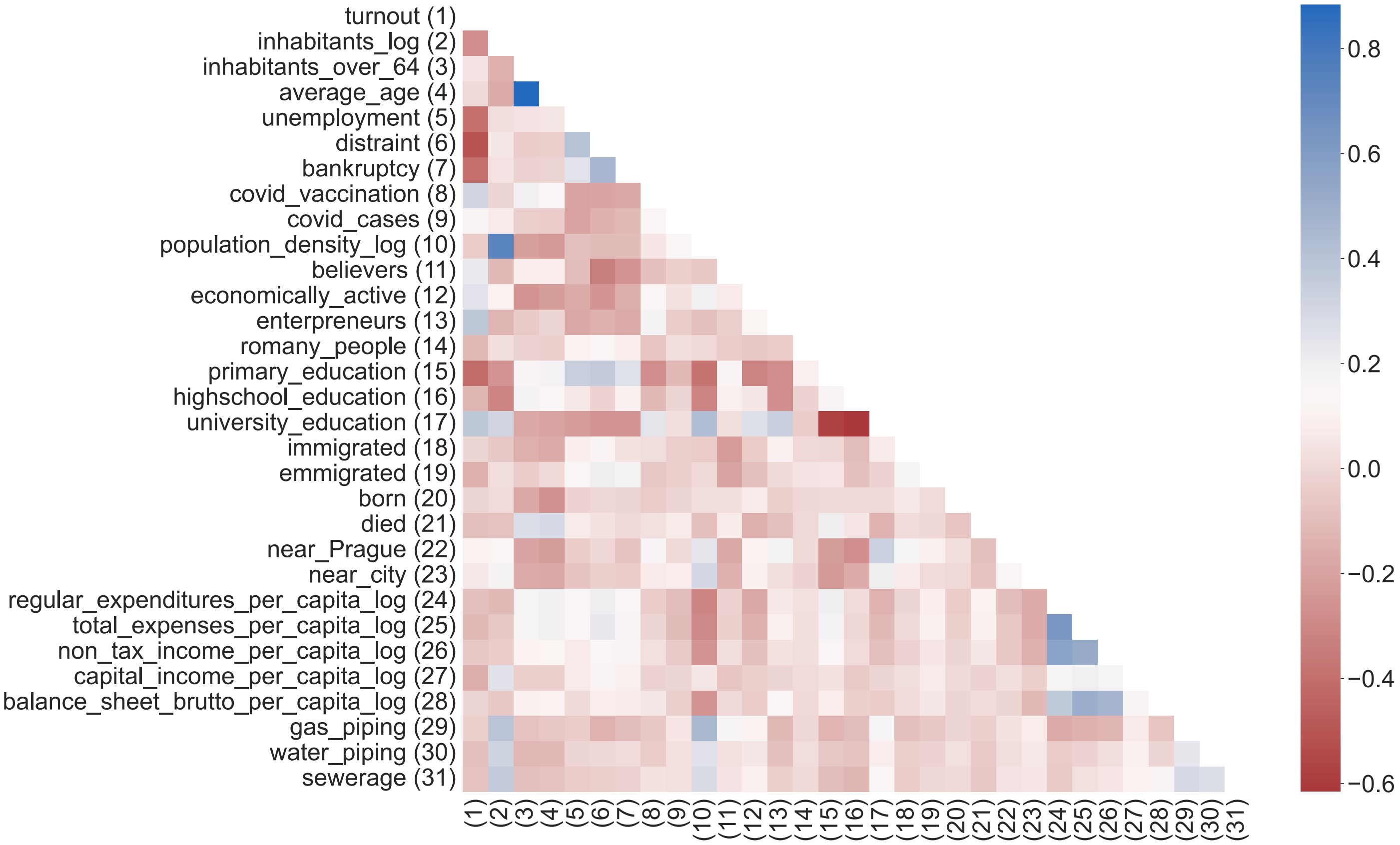


Figure D.1: Correlation matrix of all independent variables

Table D.1: OLS results, part 1

	Dependent variable:		
	SPOLU	ANO	Pirati + STAN
Constant	0.043 (0.037)	0.314*** (0.037)	0.033 (0.029)
turnout	0.173*** (0.016)	−0.174*** (0.016)	0.086*** (0.012)
inhabitants_log	0.005*** (0.001)	−0.005*** (0.001)	0.004*** (0.001)
inhabitants_over_64	−0.269*** (0.022)	0.352*** (0.022)	−0.147*** (0.017)
unemployment	−0.240*** (0.051)	0.238*** (0.051)	−0.037 (0.039)
distraint	0.016 (0.019)	0.012 (0.019)	−0.010 (0.014)
bankruptcy	−0.211** (0.093)	0.485*** (0.093)	−0.252*** (0.071)
covid_vaccination	0.086*** (0.012)	0.098*** (0.013)	0.073*** (0.010)
covid_cases	0.021 (0.018)	−0.031* (0.018)	0.067*** (0.014)
population_density_log	0.003* (0.002)	0.0004 (0.002)	0.003** (0.001)
believers	0.178*** (0.008)	−0.083*** (0.008)	−0.057*** (0.006)
economically_active	−0.048*** (0.017)	0.057*** (0.017)	−0.079*** (0.013)
enterpreneurs	0.576*** (0.027)	−0.499*** (0.027)	0.250*** (0.021)
roma_people	−0.960** (0.402)	−1.359*** (0.403)	0.533* (0.309)
primary_education	−0.070*** (0.022)	0.126*** (0.022)	−0.014 (0.017)
university_education	0.349*** (0.029)	−0.240*** (0.029)	0.148*** (0.022)
immigrated	−0.012 (0.042)	0.067 (0.042)	0.053* (0.032)
emigrated	0.046 (0.055)	0.008 (0.055)	0.057 (0.042)
died	0.137 (0.099)	−0.329*** (0.099)	−0.108 (0.076)
born	0.243* (0.132)	−0.244* (0.132)	0.051 (0.101)
near_Prague	0.041*** (0.004)	−0.030*** (0.004)	0.021*** (0.003)
near_city	−0.008*** (0.002)	0.006*** (0.002)	−0.002 (0.001)
regular_exp_pc_log	−0.009*** (0.003)	0.007*** (0.003)	−0.005** (0.002)
total_exp_pc_log	−0.004 (0.003)	0.0002 (0.003)	0.004* (0.002)
non_tax_inc_pc_log	−0.0001 (0.001)	0.001 (0.001)	0.0001 (0.001)
capital_inc_pc_log	0.0001 (0.0003)	−0.00002 (0.0003)	0.0004* (0.0002)
bal_sheet_br_pc_log	0.007*** (0.002)	−0.005** (0.002)	0.002 (0.002)
gas_piping	−0.005*** (0.002)	0.010*** (0.002)	−0.012*** (0.001)
water_piping	0.0004 (0.003)	0.001 (0.003)	−0.003 (0.002)
sewerage	0.008*** (0.002)	−0.0003 (0.002)	−0.008*** (0.002)
Observations	6,254	6,254	6,254
R^2	0.429	0.335	0.231
Adjusted R^2	0.427	0.332	0.228
Res. Std. Error (df = 6224)	0.063	0.063	0.048
F Statistic (df = 29; 6224)	161.559***	108.101***	64.582***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table D.2: OLS results, part 2

	Dependent variable:		
	SPD	CSSD	KSCM
Constant	0.301*** (0.023)	0.035** (0.016)	0.082*** (0.016)
turnout	−0.064*** (0.010)	−0.014** (0.007)	−0.023*** (0.007)
inhabitants_log	−0.002*** (0.001)	−0.001* (0.001)	−0.001** (0.0005)
inhabitants_over_64	−0.033** (0.013)	0.057*** (0.010)	0.087*** (0.009)
unemployment	0.105*** (0.031)	−0.066*** (0.022)	0.053** (0.021)
distraint	0.021* (0.011)	−0.007 (0.008)	−0.025*** (0.008)
bankruptcy	0.163*** (0.057)	−0.119*** (0.041)	0.114*** (0.039)
covid_vaccination	−0.194*** (0.008)	0.048*** (0.005)	−0.011** (0.005)
covid_cases	−0.007 (0.011)	0.024*** (0.008)	−0.020*** (0.008)
population_density_log	−0.001 (0.001)	−0.002** (0.001)	−0.004*** (0.001)
believers	−0.030*** (0.005)	0.037*** (0.003)	−0.021*** (0.003)
economically_active	0.033*** (0.010)	0.006 (0.007)	0.004 (0.007)
enterpreneurs	−0.112*** (0.016)	−0.106*** (0.012)	−0.118*** (0.011)
roma_people	−0.043 (0.247)	1.199*** (0.177)	−0.160 (0.168)
primary_education	0.012 (0.013)	−0.025*** (0.009)	0.013 (0.009)
university_education	−0.095*** (0.018)	−0.031** (0.013)	−0.035*** (0.012)
immigrated	−0.034 (0.026)	−0.030 (0.018)	−0.014 (0.017)
emigrated	0.043 (0.034)	−0.093*** (0.024)	−0.013 (0.023)
died	0.041 (0.061)	0.037 (0.043)	0.045 (0.041)
born	−0.068 (0.081)	−0.037 (0.058)	−0.038 (0.055)
near_Prague	−0.017*** (0.002)	−0.003* (0.002)	−0.002 (0.002)
near_city	0.008*** (0.001)	−0.002** (0.001)	−0.002** (0.001)
regular_exp_pc_log	0.005*** (0.002)	−0.00004 (0.001)	−0.0003 (0.001)
total_exp_pc_log	−0.0003 (0.002)	−0.00004 (0.001)	0.004*** (0.001)
non_tax_inc_pc_log	−0.0002 (0.001)	0.001*** (0.0005)	0.00003 (0.0004)
capital_inc_pc_log	0.0001 (0.0002)	−0.0004*** (0.0001)	0.00004 (0.0001)
bal_sheet_br_pc_log	−0.005*** (0.001)	0.001 (0.001)	−0.002** (0.001)
gas_piping	0.005*** (0.001)	−0.003*** (0.001)	0.003*** (0.001)
water_piping	0.003** (0.002)	−0.004*** (0.001)	−0.0002 (0.001)
sewerage	0.0001 (0.001)	0.001 (0.001)	−0.0001 (0.001)
Observations	6,254	6,254	6,254
R^2	0.273	0.122	0.141
Adjusted R^2	0.269	0.118	0.137
Res. Std. Error (df = 6224)	0.038	0.027	0.026
F Statistic (df = 29; 6224)	80.500***	29.858***	35.372***

Note: *p<0.1; **p<0.05; ***p<0.01

Table D.3: Model results of the PRISAHA

	Dependent variable: *PRISAHA*	
	OLS	WLS
Constant	0.051*** (0.016)	0.023 (0.014)
turnout	0.020*** (0.007)	0.022*** (0.007)
inhabitants_log	0.0001 (0.001)	0.0001 (0.0003)
inhabitants_over_64	−0.025*** (0.009)	−0.035*** (0.010)
unemployment	0.006 (0.022)	0.064*** (0.018)
distraint	−0.010 (0.008)	−0.035*** (0.008)
bankruptcy	−0.096** (0.040)	−0.135*** (0.048)
covid_vaccination	0.002 (0.005)	0.011** (0.005)
covid_cases	−0.040*** (0.008)	−0.072*** (0.008)
population_density_log	0.001 (0.001)	0.001 (0.0005)
believers	−0.005 (0.003)	−0.008*** (0.003)
economically_active	0.024*** (0.007)	0.039*** (0.008)
enterpreneurs	−0.027** (0.012)	−0.060*** (0.011)
roma_people	0.306* (0.174)	0.123 (0.190)
primary_education	−0.028*** (0.009)	0.004 (0.010)
university_education	−0.082*** (0.012)	−0.056*** (0.009)
immigrated	−0.009 (0.018)	0.001 (0.021)
emigrated	0.013 (0.024)	−0.060** (0.026)
died	−0.018 (0.043)	−0.111* (0.057)
born	0.035 (0.057)	0.166** (0.074)
near_Prague	−0.004** (0.002)	−0.007*** (0.001)
near_city	−0.004*** (0.001)	−0.004*** (0.001)
regular_exp_pc_log	0.002 (0.001)	0.001 (0.001)
total_exp_pc_log	−0.002* (0.001)	0.001 (0.001)
non_tax_inc_pc_log	−0.001*** (0.0005)	−0.001*** (0.0004)
capital_inc_pc_log	0.00001 (0.0001)	0.0002* (0.0001)
bal_sheet_br_pc_log	0.001 (0.001)	−0.0001 (0.001)
gas_piping	0.002*** (0.001)	0.002* (0.001)
water_piping	0.001 (0.001)	0.0004 (0.002)
sewerage	0.001 (0.001)	0.001 (0.001)
Observations	6,254	6,254
R^{2}	0.032	0.145
Adjusted R^{2}	0.028	0.141
Res. Std. Error (df = 6224)	0.027	0.456
F Statistic (df = 29; 6224)	7.157***	36.343***
Note:		*p<0.1; **p<0.05; ***p<0.01

Table D.4: SDEM results, part 1

	Dependent variable:		
	SPOLU	ANO	Pirati+STAN
Constant	0.004 (0.036)	0.351*** (0.036)	0.054** (0.027)
turnout	0.140*** (0.015)	−0.118*** (0.015)	0.043*** (0.011)
inhabitants_log	0.006*** (0.001)	−0.006*** (0.001)	0.002** (0.001)
inhabitants_over_64	−0.248*** (0.021)	0.349*** (0.021)	−0.166*** (0.015)
unemployment	−0.051 (0.053)	0.022 (0.052)	−0.026 (0.038)
distraint	0.017 (0.018)	−0.012 (0.017)	−0.020 (0.013)
bankruptcy	−0.163* (0.087)	0.384*** (0.086)	−0.186*** (0.064)
covid_vaccination	0.068*** (0.013)	0.117*** (0.013)	0.046*** (0.010)
covid_cases	0.003 (0.018)	−0.029* (0.018)	0.019 (0.013)
population_density_log	0.002 (0.002)	0.004** (0.002)	−0.001 (0.001)
believers	0.250*** (0.012)	−0.091*** (0.011)	−0.029*** (0.009)
economically_active	−0.061*** (0.016)	0.069*** (0.016)	−0.077*** (0.012)
enterpreneurs	0.471*** (0.026)	−0.425*** (0.026)	0.176*** (0.019)
romany_people	−1.190*** (0.377)	−1.190*** (0.369)	0.650** (0.276)
primary_education	−0.040* (0.021)	0.053*** (0.020)	0.027* (0.015)
university_education	0.326*** (0.029)	−0.286*** (0.028)	0.225*** (0.021)
immigrated	−0.024 (0.040)	0.070* (0.039)	0.008 (0.029)
emigrated	0.068 (0.052)	−0.029 (0.051)	0.036 (0.038)
died	0.127 (0.094)	−0.273*** (0.092)	−0.058 (0.069)
born	0.180 (0.125)	−0.072 (0.123)	0.007 (0.092)
near_Prague	−0.012 (0.011)	0.003 (0.011)	−0.011 (0.008)
near_city	0.001 (0.003)	−0.003 (0.003)	−0.003 (0.002)
regular_exp_pc_log	−0.003 (0.002)	−0.0004 (0.002)	−0.001 (0.002)
total_exp_pc_log	−0.002 (0.003)	−0.001 (0.003)	0.005*** (0.002)
non_tax_inc_pc_log	−0.001 (0.001)	0.002 (0.001)	−0.0001 (0.001)
capital_inc_pc_log	0.0003 (0.0003)	−0.00005 (0.0003)	0.0002 (0.0002)
bal_sheet_br_pc_log	0.004** (0.002)	−0.002 (0.002)	0.001 (0.002)
gas_piping	−0.001 (0.002)	−0.0003 (0.002)	0.002 (0.002)
water_piping	0.003 (0.003)	−0.002 (0.003)	−0.002 (0.002)
sewerage	0.003 (0.002)	0.0001 (0.002)	−0.001 (0.002)
lag.(Intercept)	−0.011 (0.035)	0.011 (0.034)	−0.060** (0.026)
lag.turnout	0.006 (0.014)	−0.024* (0.014)	0.030*** (0.010)
lag.inhabitants_log	0.002 (0.001)	−0.003*** (0.001)	0.001 (0.001)
lag.inhabitants_over_64	−0.004 (0.018)	−0.034** (0.017)	0.033** (0.013)
lag.unemployment	−0.078** (0.032)	0.119*** (0.031)	−0.015 (0.024)

Table D.4: continued from previous page

lag.distraint	0.005 (0.018)	−0.010 (0.018)	0.033** (0.013)
lag.bankruptcy	−0.109 (0.090)	0.219** (0.089)	−0.268*** (0.066)
lag.covid_vaccination	−0.006 (0.008)	0.018** (0.008)	0.003 (0.006)
lag.covid_cases	−0.007 (0.013)	0.024* (0.013)	0.010 (0.010)
lag.population_density_log	−0.002 (0.002)	0.003** (0.001)	0.001 (0.001)
lag.believers	−0.022*** (0.004)	0.012*** (0.004)	−0.001 (0.003)
lag.economically_active	0.004 (0.015)	−0.013 (0.015)	0.011 (0.011)
lag.enterpreneurs	0.085*** (0.020)	−0.068*** (0.020)	0.026* (0.015)
lag.romany_people	−0.341 (0.414)	0.066 (0.407)	0.249 (0.305)
lag.primary_education	0.010 (0.017)	0.033** (0.017)	−0.011 (0.013)
lag.university_education	0.019 (0.020)	0.039** (0.020)	−0.052*** (0.015)
lag.immigrated	−0.124*** (0.044)	0.165*** (0.043)	−0.054* (0.032)
lag.emigrated	0.003 (0.061)	−0.104* (0.060)	0.077* (0.045)
lag.died	−0.163 (0.106)	0.092 (0.105)	−0.074 (0.078)
lag.born	0.033 (0.139)	−0.201 (0.136)	−0.009 (0.102)
lag.near_Prague	0.001 (0.002)	0.001 (0.002)	0.003 (0.002)
lag.near_city	−0.002** (0.001)	0.002* (0.001)	0.0004 (0.001)
lag.regular_exp_pc_log	−0.008*** (0.002)	0.008*** (0.002)	−0.001 (0.002)
lag.total_exp_pc_log	0.006** (0.003)	−0.004 (0.003)	0.002 (0.002)
lag.non_tax_inc_pc_log	−0.001 (0.001)	0.0005 (0.001)	−0.001 (0.001)
lag.capital_inc_pc_log	0.0003 (0.0003)	−0.001*** (0.0003)	0.0005** (0.0002)
lag.bal_sheet_br_pc_log	0.002 (0.002)	−0.003** (0.002)	0.001 (0.001)
lag.gas_piping	−0.0003 (0.001)	0.002 (0.001)	−0.003*** (0.001)
lag.water_piping	−0.002 (0.002)	0.001 (0.002)	0.0004 (0.002)
lag.sewerage	0.005*** (0.002)	−0.001 (0.002)	−0.005*** (0.001)
Observations	6,254	6,254	6,254
Log Likelihood	8,907.843	9,033.811	10,862.860
σ^2	0.003	0.003	0.002
Akaike Inf. Crit.	−17,691.690	−17,943.620	−21,601.710
Wald Test (df = 1)	3,253.170***	5,717.104***	12,636.440***
LR Test (df = 1)	339.481***	367.218***	508.519***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table D.5: SDEM results, part 2

	Dependent variable:		
	SPD	CSSD	KSCM
Constant	0.308*** (0.023)	0.037** (0.016)	0.076*** (0.016)
turnout	−0.044*** (0.009)	−0.007 (0.007)	−0.012* (0.006)
inhabitants_log	−0.003*** (0.001)	−0.0003 (0.001)	0.0001 (0.001)
inhabitants_over_64	−0.035*** (0.013)	0.045*** (0.009)	0.080*** (0.009)
unemployment	0.068** (0.033)	−0.007 (0.024)	0.019 (0.023)
distraint	0.012 (0.011)	0.002 (0.008)	−0.010 (0.008)
bankruptcy	0.139** (0.055)	−0.104*** (0.039)	0.111*** (0.038)
covid_vaccination	−0.142*** (0.008)	0.028*** (0.006)	−0.022*** (0.006)
covid_cases	0.013 (0.011)	0.018** (0.008)	0.003 (0.008)
population_density_log	0.0001 (0.001)	0.0002 (0.001)	−0.003*** (0.001)
believers	−0.063*** (0.007)	0.011** (0.005)	−0.047*** (0.005)
economically_active	0.037*** (0.010)	0.005 (0.007)	0.004 (0.007)
enterpreneurs	−0.057*** (0.017)	−0.093*** (0.012)	−0.089*** (0.011)
romany_people	−0.061 (0.236)	1.087*** (0.169)	−0.128 (0.163)
primary_education	−0.003 (0.013)	0.0002 (0.009)	0.014 (0.009)
university_education	−0.118*** (0.018)	−0.012 (0.013)	−0.042*** (0.013)
immigrated	−0.018 (0.025)	−0.024 (0.018)	−0.005 (0.017)
emigrated	0.030 (0.033)	−0.077*** (0.023)	0.005 (0.023)
died	0.003 (0.058)	0.004 (0.042)	0.014 (0.040)
born	−0.020 (0.078)	−0.093* (0.056)	−0.046 (0.054)
near_Prague	0.006 (0.007)	0.005 (0.005)	−0.001 (0.005)
near_city	0.002 (0.002)	0.0003 (0.001)	−0.001 (0.001)
regular_exp_pc_log	0.001 (0.002)	0.001 (0.001)	−0.0001 (0.001)
total_exp_pc_log	−0.002 (0.002)	0.0003 (0.001)	0.003*** (0.001)
non_tax_inc_pc_log	0.001 (0.001)	−0.00000 (0.0005)	0.00001 (0.0005)
capital_inc_pc_log	0.00001 (0.0002)	−0.0003*** (0.0001)	0.00004 (0.0001)
bal_sheet_br_pc_log	−0.003*** (0.001)	−0.0004 (0.001)	−0.002*** (0.001)
gas_piping	0.001 (0.001)	−0.001 (0.001)	0.001 (0.001)
water_piping	0.002 (0.002)	−0.002* (0.001)	0.001 (0.001)
sewerage	−0.001 (0.001)	−0.0001 (0.001)	−0.001 (0.001)
lag.(Intercept)	0.037* (0.019)	0.006 (0.015)	0.010 (0.014)
lag.turnout	−0.007 (0.008)	−0.009 (0.006)	−0.004 (0.006)
lag.inhabitants_log	−0.001 (0.001)	−0.001 (0.001)	0.0003 (0.001)
lag.inhabitants_over_64	0.019* (0.010)	0.004 (0.007)	0.007 (0.007)

Table D.5: continued from previous page

lag.unemployment	0.015 (0.017)	−0.025* (0.013)	0.015 (0.012)
lag.distraint	0.012 (0.010)	−0.007 (0.008)	−0.015** (0.007)
lag.bankruptcy	0.074 (0.050)	−0.007 (0.038)	0.091** (0.035)
lag.covid_vaccination	−0.017*** (0.004)	0.003 (0.003)	0.0004 (0.003)
lag.covid_cases	−0.007 (0.007)	0.008 (0.005)	−0.011** (0.005)
lag.population_density_log	0.001 (0.001)	−0.0004 (0.001)	−0.001** (0.001)
lag.believers	0.009*** (0.002)	0.005*** (0.002)	0.003* (0.002)
lag.economically_active	0.007 (0.008)	0.001 (0.006)	0.002 (0.006)
lag.enterpreneurs	−0.032*** (0.010)	−0.009 (0.008)	−0.017** (0.008)
lag.romany_people	0.168 (0.238)	0.306* (0.179)	−0.179 (0.168)
lag.primary_education	−0.025*** (0.009)	−0.013* (0.007)	−0.003 (0.007)
lag.university_education	0.001 (0.011)	0.003 (0.009)	0.018** (0.008)
lag.immigrated	0.035 (0.025)	−0.030 (0.019)	−0.005 (0.018)
lag.emigrated	0.015 (0.035)	0.017 (0.026)	0.020 (0.025)
lag.died	0.102* (0.062)	0.103** (0.046)	0.002 (0.043)
lag.born	0.061 (0.078)	0.092 (0.059)	0.062 (0.055)
lag.near_Prague	−0.001 (0.001)	0.0004 (0.001)	0.0004 (0.001)
lag.near_city	0.001** (0.0004)	0.0001 (0.0004)	0.0002 (0.0003)
lag.regular_exp_pc_log	−0.001 (0.001)	0.0003 (0.001)	0.0004 (0.001)
lag.total_exp_pc_log	0.001 (0.002)	−0.001 (0.001)	−0.001 (0.001)
lag.non_tax_inc_pc_log	0.001 (0.0005)	−0.0001 (0.0004)	0.0002 (0.0003)
lag.capital_inc_pc_log	0.0001 (0.0001)	−0.00002 (0.0001)	−0.00003 (0.0001)
lag.bal_sheet_br_pc_log	−0.002** (0.001)	0.001 (0.001)	0.0002 (0.001)
lag.gas_piping	0.0002 (0.001)	−0.001** (0.001)	0.001 (0.0005)
lag.water_piping	−0.001 (0.001)	0.0002 (0.001)	−0.001 (0.001)
lag.sewerage	0.0004 (0.001)	0.0004 (0.001)	0.0001 (0.001)
Observations	6,254	6,254	6,254
Log Likelihood	11,836.570	13,914.110	14,146.100
σ^2	0.001	0.001	0.001
Akaike Inf. Crit.	−23,549.130	−27,704.230	−28,168.190
Wald Test (df = 1)	209.948***	499.037***	299.735***
LR Test (df = 1)	107.819***	176.782***	117.674***

Note: *p<0.1; **p<0.05; ***p<0.01

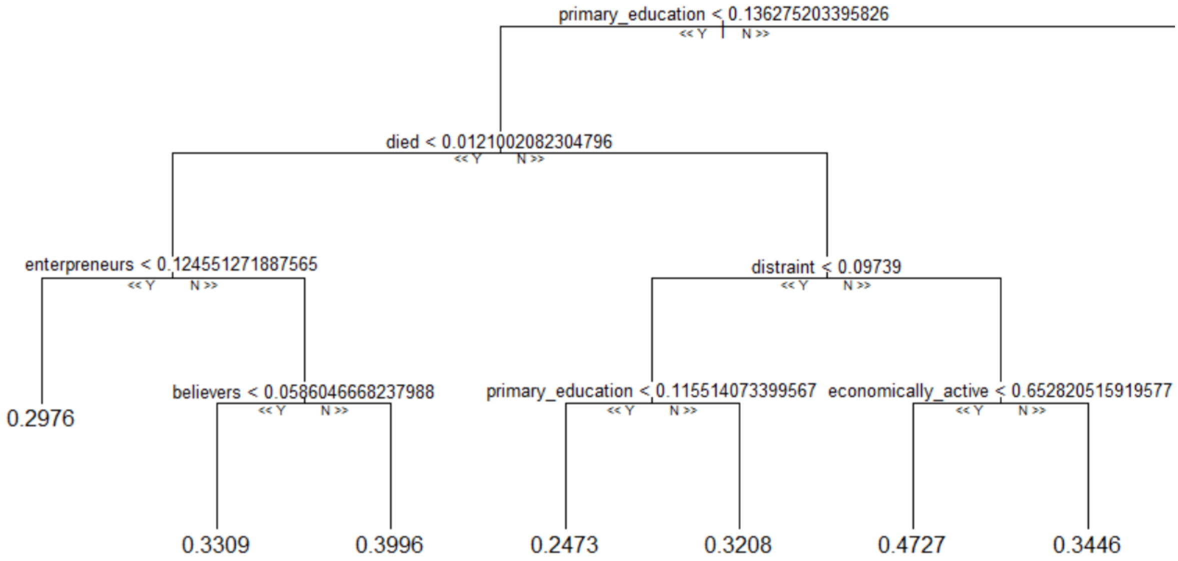

primary_education < 0.136275203395826
<< Y
N >>
died < 0.0121002082304796
<< Y
N >>
enterpreneurs < 0.124551271887565
<< Y
N >>
distraint < 0.09739
<< Y
N >>
0.2976
believers < 0.0586046668237988
<< Y
N >>
primary_education < 0.115514073399567
<< Y
N >>
economically_active < 0.652820515919577
<< Y
N >>
0.3309
0.3996
0.2473
0.3208
0.4727
0.3446

Figure D.2: A representative tree of the RF model (the SPOLU), part 1

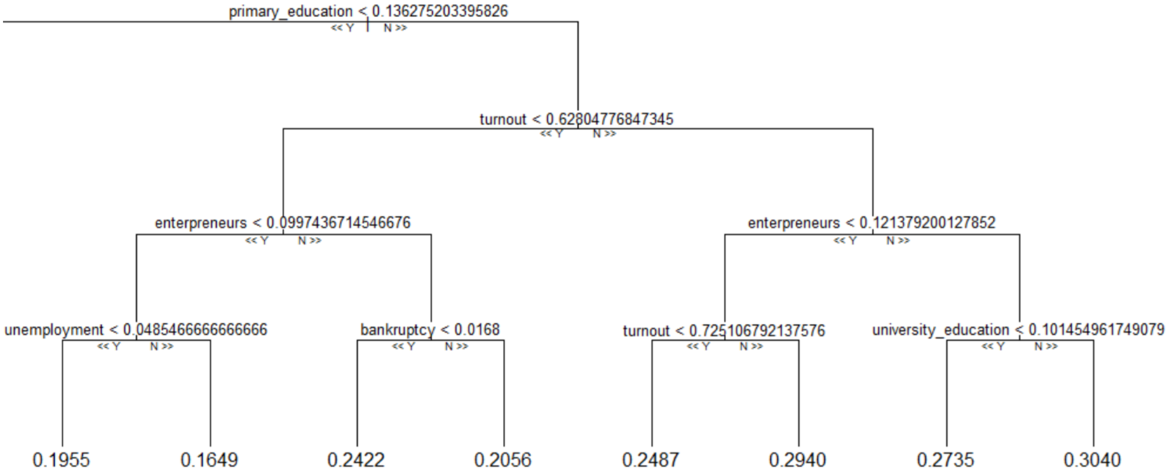

primary_education < 0.136275203395826
<< Y
N >>
turnout < 0.62804776847345
<< Y
N >>
enterpreneurs < 0.0997436714546676
<< Y
N >>
enterpreneurs < 0.121379200127852
<< Y
N >>
unemployment < 0.0485466666666666
<< Y
N >>
bankruptcy < 0.0168
<< Y
N >>
turnout < 0.725106792137576
<< Y
N >>
university_education < 0.101454961749079
<< Y
N >>
0.1955
0.1649
0.2422
0.2056
0.2487
0.2940
0.2735
0.3040

Figure D.3: A representative tree of the RF model (the SPOLU), part 2

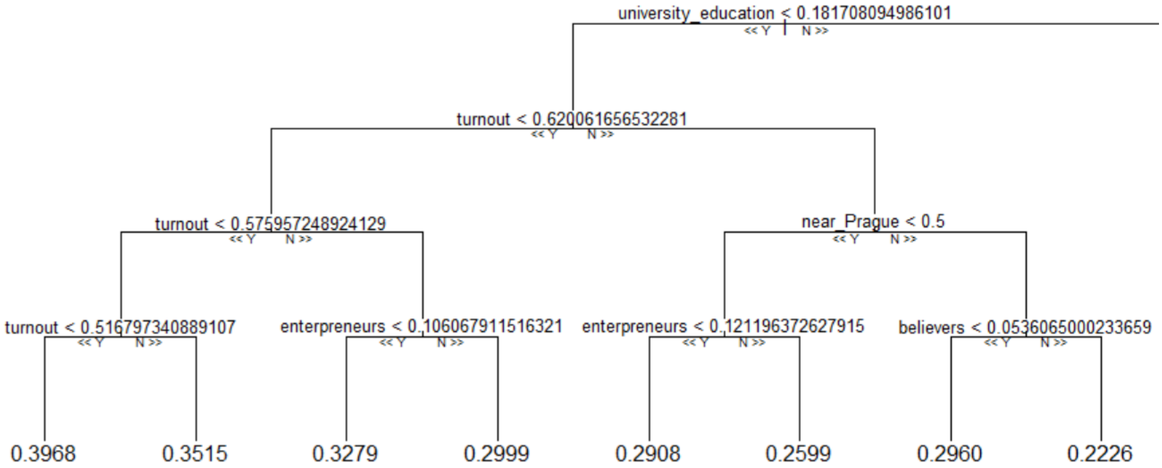

university_education < 0.181708094986101
<< Y N >>
turnout < 0.62006165653281
<< Y N >>
turnout < 0.575957248924129
<< Y N >>
near_Prague < 0.5
<< Y N >>
turnout < 0.516797340889107
<< Y N >>
enterpreneurs < 0.106067911516321
<< Y N >>
enterpreneurs < 0.121196372627915
<< Y N >>
believers < 0.0536065000233659
<< Y N >>
0.3968
0.3515
0.3279
0.2999
0.2908
0.2599
0.2960
0.2226

Figure D.4: A representative tree of the RF model (the ANO), part 1

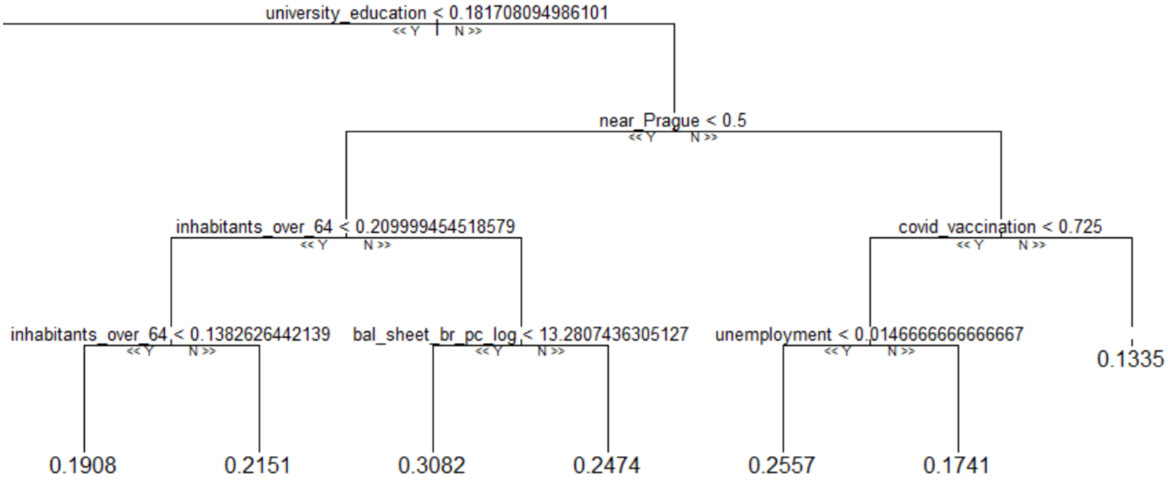

university_education < 0.181708094986101
<< Y N >>
near_Prague < 0.5
<< Y N >>
inhabitants_over_64 < 0.209999454518579
<< Y N >>
covid_vaccination < 0.725
<< Y N >>
inhabitants_over_64 < 0.1382626442139
<< Y N >>
bal_sheet_br_pc_log < 13.2807436305127
<< Y N >>
unemployment < 0.0146666666666667
<< Y N >>
0.1335
0.1908
0.2151
0.3082
0.2474
0.2557
0.1741

Figure D.5: A representative tree of the RF model (the ANO), part 2