Politics, Institutions and (Random) Forests? The drivers of COVID-19 vaccination rates in Poland

Tymoteusz Mętrak [1]
Radosław Rybakowski [2]
Dominik Stempień [1]
Wojciech Szymczak [1]

Abstract:

This paper studies the drivers influencing the COVID-19 vaccination rates in Poland, emphasizing the role of spatial, developmental, and political factors. Despite Poland's economic and social position, its vaccination rates lag behind CEE countries. Using various analytical techniques, including spatial regressions, unsupervised learning, instrumental variables, and machine learning, we investigate the differences in the COVID-19 intake in the dimensions of demographics, education, politics and institutions. The research reveals significant regional disparities, with Eastern Poland and certain rural areas exhibiting lower vaccination rates, largely unaffected by historical partitions. Surprisingly, political factors, particularly left-wing voter enthusiasm, are critical drivers of vaccination decisions, overshadowing concerns associated with right-wing hesitancy. Our results shed new light on the possible roles of the political parties in influencing the decisions in relatively underdeveloped regions.

JEL classification: I14, C26, C52

[1] University of Warsaw, Faculty of Economic Sciences, Warsaw, Poland.
[2] Warsaw School of Economics, Warsaw, Poland.
This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Introduction

Vaccinations are one of the most effective methods to prevent population from the risk of infection. In his popular science book "*Fake Medicine*", Brad McKay has emphasized that "*all children should be vaccinated so that they have the opportunity to become adults*".[3] However, even adults sometimes behave less mature than children and decide to not take up the vaccine. In our study, we explore the possible drivers that hindered the vaccination rates among Polish population during the pandemic of COVID-19.

In this paper, we study the impact of several factors – institutions, public services, social composition and political preferences – on the vaccination rates. We analyze the case of Poland, which, as of 2023, is the 5th largest EU country and the 21st largest economy in the world. Still, Poland stands out of the Central Eastern Europe region in terms of vaccination rates. Despite similar economic and social indicators, in Czech Republic, Hungary, Lithuania the share of fully vaccinated individuals was almost 10 percentage points higher than in Poland[4]. Since the probability of the outbreak of new health hazards remains high, it is crucial to find the drivers of the COVID-19 hesitancy to provide evidence-based effective solutions.

The heterogeneity in the population's needs demands flexible solutions, as uniform policies, treating all vulnerable groups symmetrically, usually requires some difficult trade-offs such as extending lockdowns (Acemoglu et al., 2021). The possible consequences of low vaccination rates go far beyond the health outcomes, as the external effects significantly influence labor market (see Altig et al., 2023; Bloom et al., 2023) and education (see Jakubowski et al., 2023). Since the role of jobs and education is significant in driving the outcomes of the economy, we explore the possible solutions in the paper. In this study we focus on determining the differences in vaccination patterns between rural and urban areas, as well as sociodemographic characteristics (such as age and gender). Knowing the history of the partitions in Poland, we explore their role in driving the trends. By exploring the differences in patterns and the relationships between vaccination and public services provision and other factors, we obtain valuable information on the heterogeneities in the effectiveness of selected policies in increasing the health outcomes.

[3] Brad McKay (2021) Fake Medicine
[4] Source: Our world in data

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team "Maximum (Victory) Likelihood EsTeamation".

Our study makes several contributions to scientific literature. Firstly, we use novel machine learning techniques to support the selection of the variables to the final model. By doing so, we are able to make use of a large number of variables and detect possibly non-linear relationships between the data. It should be emphasized that despite the increasing popularity of machine learning in studies on health outcomes, we find little evidence on its use for determining the drivers of the COVID-19 vaccinations. In addition, we are able to answer the above-mentioned research questions by comparing its impact on the predictive power of the model. Secondly, we use spatial modeling to obtain information on the geographical patterns in the vaccination uptake in polish municipalities.

Our work is built-up on the recent advances in empirical literature in health economics and related fields. First, there is a large strand of literature focused on access to public services and inequalities associated with health deprivation (e.g. Carrieri & Jones, 2018; Brown et al., 2021; Lillebråten et al., 2023). Advanced access to goods such as healthcare and education may ease access to vaccinations and build knowledge about the advantages of vaccinations. However, large inequalities may lower the trust towards the public institutions, decreasing the probability of the COVID-19 vaccination intake. Indeed, there are several papers emphasizing the primary role of the inequalities and vulnerability in increasing the vaccination rates (Bilal et al., 2022; Clouston et al.; 2023). Relative to these papers, using mainly the inequalities features, we study the impact of the already available public services. The central advantage of our approach is evaluation of possible solutions to the already studied problems.

Second, our paper relies on the evidence on the impact of political preferences on the vaccination. There are mixed results on the impact of the populism on the hesitancy to uptake the vaccinations. While there is a strand of literature showing null effects of the populism on the vaccinations (Juen et al., 2023;), there are also other research showing that support for democratic, rather left-wing oriented parties increases the intake of the vaccinations. In our study, we do not use any indices to summarise the political preferences, but rather compare the results between the regions. We also explore the role of the political engagement, by studying the influence of the political turnout on the uptake of the vaccinations.

We explore the possible social learning of the decision to uptake the vaccine. Meng et al. (2023) used k-means clustering to compare the low- and high-vaccinations regions and

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

compared their characteristics to understand the possible social and behavioral drivers of the vaccinations in the US. They found that the cluster characterized by the lowest vaccination rates were mainly males, from rural areas. Interestingly, they showed that the individuals who were diagnosed with COVID-19 in the past were less likely in general to vaccinate against the COVID-19. It contradicts the argument that the exposure to the disease should reinform the population to get vaccinated.

Our results show that there are significant differences in vaccination patterns between the regions in Poland. However, these differences are not related to the historical partitions background of Poland. Instead, we find that the Eastern Poland and some rural areas across the countries are characterized by the high vaccination rates. Both causal and machine learning methods emphasize the role of politics in driving the trends in vaccination. However, we find that it may be the enthusiasm (of the left-wing voters) to take up the vaccines rather the hesitancy (of the righ-wing voters) that drives the vaccination decsions. We find no evidence on the role of the exposure to the increased mortality due to COVID-19 to increase the vaccination rates.

Data

Final dataset created on purpose of our paper is constructed basing on dataset provided by Warsaw Economic Challenge organizers and our own additional data. Detailed information can be found in the Appendix.

BDL GUS

The Local Data Bank of the Polish National Statistical Office (BDL GUS) serves as the primary source of statistical data in our research endeavors. A significant portion of the variables initially provided by the Warsaw Economic Challenge organizers originates from this dataset.

Furthermore, the authors opted to enhance the BDL database by incorporating additional variables. Among these, variables related to agricultural areas' characteristics were deemed essential for inclusion in our model. Variables such as information on rural taxes or the average size of farms could provide valuable insights. Regarding rural taxes, we computed

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

them as the ratio of original rural taxes per capita divided by the total revenue generated variable, 'rolny_revenue_ratio'.

Moreover, data about unemployment was downloaded, enabling us to calculate the unemployment level in age groups as a percentage of the population of municipalities. Importantly, the dataset based on counties was also merged with municipalities, merging according to municipality code. This ensured that every municipality in a county has the same level of several variables, due to the lack of a better disaggregation method.

PKW

The National Electoral Commission (PKW,)[5] plays a pivotal role as the governmental entity responsible for orchestrating and supervising elections in Poland, thus serving as a cornerstone data source for electoral information. We leveraged data from the 2019 (PKW) parliamentary elections, which provides valuable insights into the popularity of Polish political parties. This popularity can significantly influence vaccination trends among the populace. Accordingly, support for so-called populist parties may correlate with our dependent variables, a relationship to be examined across various models. Additionally, this dataset contains information on voter turnout and the percentage of invalid votes**.**

Transformation and own variables

Additionally, numerous other variables, including our own, were integrated, along with transformations based on this dataset.

Partitions

In 1918, Poland regained independence after 123 years of occupation by three foreign powers: Prussia, Russia, and Austria. These partitions continue to significantly influence many phenomena in Poland even today, underscoring the importance of including this historical impact in our models. Utilizing data provided by WEC organizers, we constructed binary variables indicating the predominant partition to which most municipalities belonged. Subsequently, we created variables for the Prussian and Russian partitions, with the Austrian partition serving as the reference category.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Latitude and longitude

One of our research hypotheses involves examining the influence of geographical coordinates. These variables serve as representations of geographical and historical conditions, such as climate or proximity to foreign economies. For each municipality, latitude and longitude were calculated based on the centroid coordinates of each area, which can be described as a polygon.

Urbanization categories

Another crucial question addressed by the outcomes of our models is the degree of urbanization in each municipality, wherein the size of cities must be distinguished. To meet this condition, a categorical variable 'type' was provided, which determines the municipality category. An algorithm was devised as follows, demonstrating the table along with information about the fraction of each category in the total.

Table 1. Description of the categories of city sizes

City category 'type'	***Count***	***Population***	***Population (%)***	***Description***
500k+	5	4 556 677	12,0%	Population larger than 500k
100_500k	32	6 060 472	15,9%	Population between 100k & 500k
50_100k	53	3 491 576	9,2%	Population between 50k & 100k
20_50k	219	6 626 182	17,4%	Population between 20k & 50k
0_20k	635	6 494 112	17,1%	Population smaller than 20k
agr	1533	10 859 545	28,5%	Type of municipality is rural
Total	*2477*	*38 088 564*	*1*	

Source: Own elaboration

Other demographic factors

One of the most significant demographic factors stratifying society is age groups and the declared sex factor of individuals. Generations and sex groups can vary based on numerous social and psychological factors, potentially influencing vaccination tendencies. To gauge the impact of these variables, we introduced new variables representing the fraction of the total population in each municipality. Additionally, these variables were aggregated compared to the original groups, with the following breaks assumed:

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

- 0_19: individuals between 0 and 19 years old,
- 20_29: individuals between 20 and 29 years old,
- 40_49: individuals between 40 and 49 years old,
- 50_59: individuals between 50 and 59 years old,
- 60_84: individuals between 60 and 84 years old.

In this scheme, the 30-39 age group was retained as the reference. Furthermore, we created the variable 'pop_f_all_perc,' indicating the fraction of females in municipalities, while other population segments were maintained as the reference.

Exposure to COVID-19 related mortality

The spread of COVID-19 was characterized by spread of misinformation, which may have caused concern towards taking up the vaccines. Indeed, the study of Masters et al. (2022) revealed that the persisting concern was one of the main drivers of the vaccine hesitancy. However, the population might experience the so-called social learning, which refers to the philosophy of communities learning via observations. In particular, the groups that were significantly exposed to the COVID-19 risk could be the most into getting vaccinated as soon as possible. Therefore, we argue that vaccination intake might have been dependent on the exposure to abnormal mortality resulting from COVID-19, as the spread of fear on the municipal level could have motiviated individuals to vaccinate against the virus. In other words, municipalities which experienced significantly higher deaths (caused by COVID-19) could be more prone to get vaccinated. Therefore, we construct an indicator which captures the increased mortality by averaging the mortality between 2010 and 2019 and subtracting it from the 2020 value. By doing this we obtain information on the excess deaths, which could be caused by COVID-19. We can also express it mathematically as:

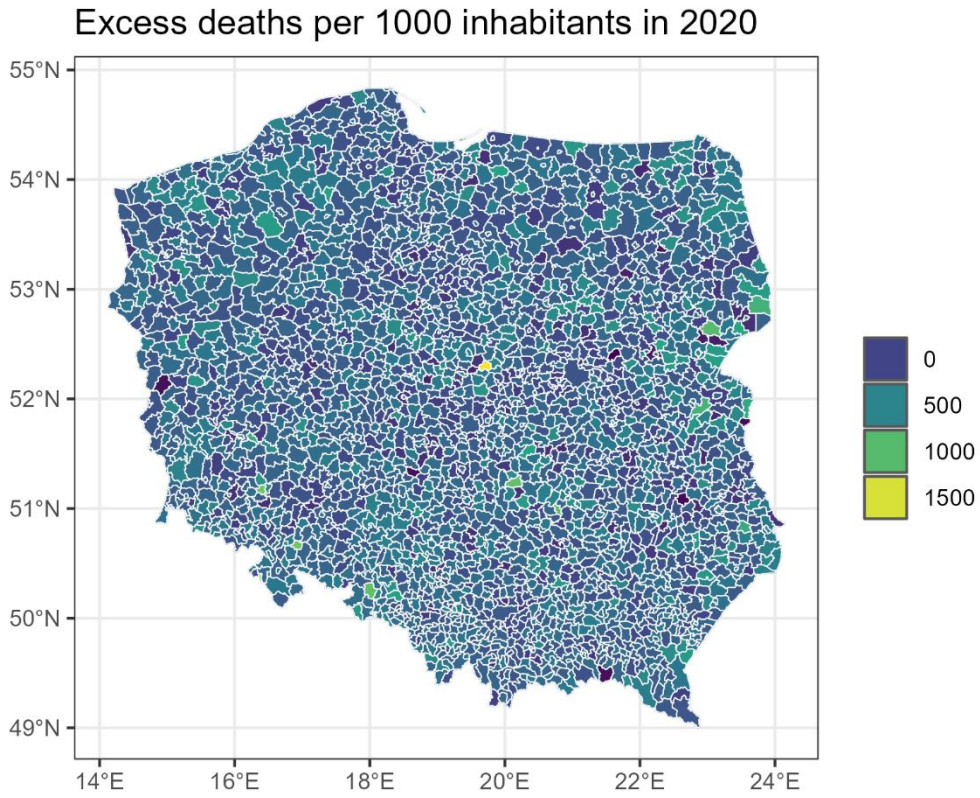
$$DeathExposure_i = \frac{1}{10} \sum_{t=2010}^{2019} Deaths_{i,t} - Deaths_{i,2020}$$

This approach was widely applied in the recent literature concerning for instance the effectiveness of vaccinations. Lewandowski & Madoń (2022) used the deviation from the trend between the so-called "*third wave*" and the "*fourth wave*" to show the high effectiveness of COVID-19 vaccinations in Poland. Figure 1 presents our measure on the level of the

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team "Maximum (Victory) Likelihood EsTeamation".

municipality. We find no clear patterns suggesting clustering of the municipalities into areas by abnormal deaths caused by the pandemic.

Figure 1. Excess deaths per 1000 inhabitants in 2020



Source: Own elaboration based on GUS data

Dimensionality reduction

Increasing access to public services, such as schools and healthcare is found to increase the physical and mental health of the population. However, the provision of public services is multidimensional in its nature, which leads to several issues from the econometric perspective. As our model incorporates machine learning techniques, it is important to reduce the dimensions of the data to speed up the estimation process. Moreover, it allows for generalizing the studied issue, which is crucial from the perspective of public services provision. At the same time, visualizing the data is easier when several features are recalculated as single variable.

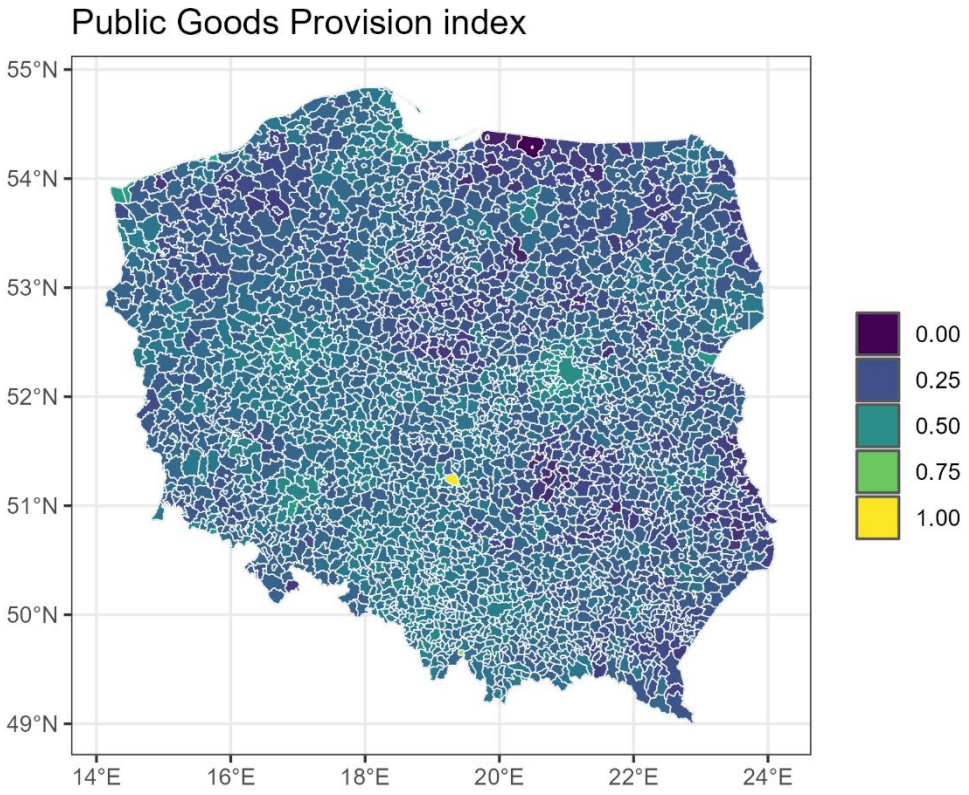
This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

World Health Organization proposed 20 (overlapping) solutions to increase the effectiveness of the COVID-19 vaccination. They found that good distribution of incomes systems are one of the features predicting high vaccination to the COVID-19. In our paper, we focus on six dimensions of the public services provision – general development (water supply installations), housing market (apartment area per person), labor market (unemployment rate), healthcare (population per pharmacy), education (net scholarization). We also take into account the income in the municipal units (revenues per capita). We focus on these dimensions as these can be the responses to the pandemic.

We build an index using the scores from the principal component analysis (PCA). We standardize and centralize the variables before the analysis. We use only the first component, which explains 25% of the total variance of the data. We observe large and moderate correlations between the index and the variables used to generate it. Figure 2 presents the spatial distribution of the index, which is in the range between 0 and 1. We find the higher provision of the public goods is centered around the largest urban areas in Poland (Warsaw, Wroclaw) and the most affluent municipalities (e.g Kleszczów). On the other hand, the index is low in regions in the Eastern Poland, especially in the Mazurian region, which is in line our intuition concerning the regional differences. In addition, our index is compliant with the work of Kopczewska (2013), who showed that the provision of public services is centered at large urban districts in Poland.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Figure 2. Public goods provision index



Source: Own elaboration based on GUS data

Methods

In the article we decided to utilize a great array of methods to fully and reliably answers as many questions stemming from the variance of vaccination levels in Poland as possible. We divided our research into two main parts: econometric approach and Machine Learning (ML) one. Econometrics allowed us to use our expertise in selecting variables, apply spatial methods and test the stability of estimations in different settings. This approach ensures the easy interpretability of the results as well. Machine Learning, however, allowed us to identify undoubtedly the most import factors of vaccination level. It is also more useful in predicting vaccination level because it adapts to the data very well and captures non-linear dependencies at ease.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Linear model

In econometrics, our baseline model to start with was OLS (Ordinary Least Squares). We used this model to get the general understanding of the data, existing collinearities, high VIF values etc. At last we eliminated some variables to obtain relevant estimates, avoiding unnecessary lack of efficiency of OLS estimation caused by stochastic collinearity.

Based on the OLS model we turned to spatial methods to address the spatial aspect of the analyzed data. We decided to create models on the subsamples as well, in order to check whether the coefficients are stable. We decided to limit our analysis to variables accounting for political beliefs (where votes for PiS where a baseline level), general health (eg. doctors), wealth and income (eg. average wage), social conditions, urbanization level (income from agriculture tax and forest area), education (only level of higher education was included in a final sample due to insignificant values and high correlation with other factors), tourism, and spatial location of the municipality. To account for heteroskedasticity, which we detected *inter alia* using Breusch-Pagan test, we employed the robust standard errors, clustered at the level of counties. Knowing, that different municipalities differ in population, we weighted linear model with population in a municiplaity.

We suspected the spatial dependence as well. So, we conducted a test for spatial dependence (I-Moran). The I-Moran statistic was 0.7095696143 with p-value almost zero. Therefore, there is a spatial dependence, so we decided to use spatial methods to handle this problem. We created weights matrix stating who is the neighbour of the municipality, normalized by row.

Spatial regression

The primary method devised for our research purposes is spatial regression modeling, allowing us to tackle most of our research questions. The model, specified on municipalities, comprises an appropriate set of control and test variables, enabling us to address the majority of our research inquiries. Incorporating spatial effects is considered crucial, as indicated by tests like the Moran test. GNS being too much computationally demanding, we decided to run SEM and SER models. Neglecting spatial effects could introduce bias in estimating the covariance-variance matrix. At first we introduced Spatial Error Model (SEM):

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

$$y = X\beta + \epsilon,$$

$$\epsilon = W\lambda\epsilon + u$$

The other model we used is generalized spatial two stage least squares (SAR). In other words it is a spatial lag model, that accounts for the autocorrelation of explained variable (vaccination rate):

$$y = y\rho + X\beta + \epsilon$$

Where y is vaccination rate, W is matrix of spatial weights, accounting for neigbourhood effects and rho is the spatial autoregression coefficient.

The results of the model are not easily interpretable, but their transformation allows to disseminate direct and indirect (spillover) effect from neighoring municipalities.

Instrumental Variables

Despite the possible significance of the political preferences and public goods, these variables are endogenous, as these may be related to other variables. Therefore, we use instrumental variables estimation to obtain causal effects. We use the so-called shift-share design in our model, by assuming that political preferences in one municipality can be closely related to the political preferences of others. In the same sense, public provision of services is geographically dependent as individuals can enjoy some services in close neighbourhood. Therefore, we use an instrument which summarizes the political preferences for the left-wing parties and public services provision in the county, except for the studied municipality. In mathematical terms:

$$Left_m^{IV} = \sum_{i \in I,\, i \neq m} \omega_i \cdot Left_i$$

where corresponds to the share of population in the municipality in the reference to the county. We follow the same procedure in terms of the public goods provision index. We use Kleibergeen-Paap First Stage F-Statistic to evaluate the quality of the instrument.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Machine learning approaches

Machine learning methods have gained significant popularity since the moment of their creation. They are widely used in various areas of research including economic and social studies. Machine learning algorithms are able to uncover linear and, most importantly, non-linear patterns presented in data, offering valuable insights into the understanding of dependencies between variables. In contrary to econometric models, machine learning methods are not obliged to meet strict assumptions regarding data distribution (e.g. OLS assumptions for linear regression model). Because of their high predictive power and ability to function in high-dimensional space, we decided to employ a bunch of machine learning algorithms in order to offer more complex and robust approach to identify main drivers of the level of COVID-19 vaccination in Poland.

K-nearest neighbors algorithm

The K-nearest neighbors (KNN) algorithm is one of the simplest machine learning methods belonging to the category of supervised learning. It is widely used in classification and regression problems providing an applicable solution across many research areas. KNN algorithm relies on the proximity of observable training data and prediction of the model depends on the values of k closest observations. In regression problems where the dependent variable is continuous, output of the model is calculated as an average value of k nearest observations. In order to make this model work optimally, it is necessary to standardize data before providing it as an input. The most important step of the hyperparameter tuning process is the identification of the number of neighbors. Other significant parameters are the metric used to calculate distances (Euclidean, Manhattan or Minkowski with various values of p) and weights assigned to the observations (uniform or as an inversed value of distance between them).

Support Vector Machines algorithm

Support Vector Machines (SVM) algorithm also falls into the category of supervised learning algorithms being able to solve classification and regression problems. The main goal of the basic form of the SVM algorithm is to find an optimal hyperplane in high-dimensional space which separates observed data into different groups. Chosen hyperplane should maximize the margin between itself and closest observations from various classes. Regression version of Support Vection Machines is more complicated, however it is based on analogical

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

assumptions. Thanks to so called “kernel trick” it is possible to transform not linearly separable data into another dimension where such division is applicable. This procedure makes the SVM algorithm a very powerful tool and allows to obtain better results. Main hyperparameters of Support Vector Machine model consist of: kernel function (most popular are linear, polynomial and radial) and the value of regularization parameter (term C). Similarly to KNN model, input data for SVM algorithm requires standardization.

Random Forest algorithm

Random Forest algorithm presents another popular and successful approach used in various specifications. This model falls into the category of ensemble training methods that combine multiple weak learners to provide more accurate and robust predictions. Random Forest algorithm is based on idea of bootstrap aggregating (so called bagging): random samples of training data with replacement are chosen and passed to decision trees algorithms. Random Forest model reduces the risk of overfitting and provides feature importance measures which might offer valuable insights for other models. However, comparing to KNN and SVM algorithms number of hyperparameters which requires tuning is bigger. Random Forest algorithm was used by Hasan et al. (2021) in classification task regarding underlying factors of measles vaccine uptake in Bangladesh.

Extreme Gradient Boosting algorithm

Extreme Gradient Boosting (XGBoost) algorithm presents the most advanced machine learning model used in our research. Similarly to Random Forest algorithm, XGBoost is an ensemble learning algorithm which is based on simple decision trees. However, in the case of XGBoost the bagging is replaced by the boosting approach where the aim of each weak leaner is to correct the errors of the previous one using gradient descent method. Since the moment of its creation, XGBoost algorithm gained significant popularity due to its high predictive power and fast computation time. Cheong et al. (2021) employed XGBoost model to predict COVID-19 vaccination uptake in US counties.

Methodology

We proposed 4 machine learning models to analyze drivers of COVID-19 vaccination in Poland: K-Nearest-Neighbors, Support Vector Machines, Random Forest and XGBoost algorithm. Our data is split into the training and testing sample in order to provide a final

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

models' performances based on the unseen data. Selection of the best performing model with optimized values of hyperparameters is based on K-Fold cross-validation approach performed on training sample. Number of folds was chosen to be equal to 5. Adoption of cross-validation technique reduces the risk of overfitting and provides more robust insights into the model's performace criteria. To prevent the risk of any data leaking, standardization of the continuous variables for KNN and SVM algorithms is executed during cross-validation stage. Process of hyperparameters tuning is applied with the use of two algorithms: random search (Random Forest and XGBoost) or grid search (KNN and SVM). Lastly, we decided to apply root mean squared error (RMSE) metric as the loss function which is used to assess the quality of the predictions of chosen models. The primary benefit of this metric is that its value is expressed in the same units as a target variable, enabling straightforward interpretation. Moreover, the same evaluation criteria was used by Cheong et al. (2021).

Explainable machine learning

After conducting cross-validation, we chose one best performing model with the lowest RMSE metric. However, obtained value of RMSE informs us only about accuracy of the predictions, not telling us anything about the impact of specific variables. In order to overcome this obstacle, we applied explainable machine learning methods to understand reasoning of our model. Precisely, we employed SHAP values technique which is based on game theory by assigning each feature its impact on the change of the dependent variable in the model. Owing to the fact, our machine learning method doesn't fall any longer into the "black box" category because its interpretability has become similar to classical econometric models.

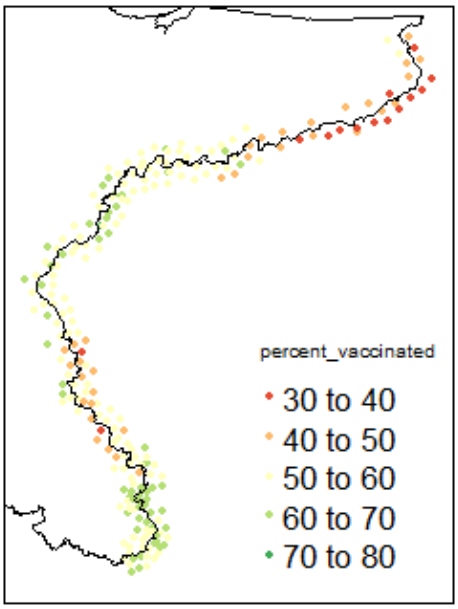
Regression discontinuity design

To explore the hypothesis regarding the influence of partitions, we have decided to employ a spatial regression discontinuity design, which we believe will effectively address this question. We make the assumption that neighboring municipalities have similar characteristics around the partition border, and therefore, should exhibit similar vaccination rates. Any statistically significant differences found in vaccination rates could serve as compelling evidence of the relevance of partitions on vaccination rates.

Based on our previous results, we have decided to designate the Prussian partition as the treatment group, as we believe its significance will be best explored through the research This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team "Maximum (Victory) Likelihood EsTeamation".

discontinuity design. To carry out this operation, it is essential to identify the nearest municipalities to the border according to a chosen cutoff point. The cutoff point will serve as a hyperparameter. If the cutoff point is large, the bias of the result may increase, but the efficiency should also improve due to the larger number of municipalities analyzed. Consequently, an analysis was conducted across a range of hyperparameters to determine the optimal cutoff point.

Figure 3. Figure shows municipalities in neighbourhood of 15 km of the boards and values of dependent variable - `percent_vaccinated`.



Source: Own elaboration

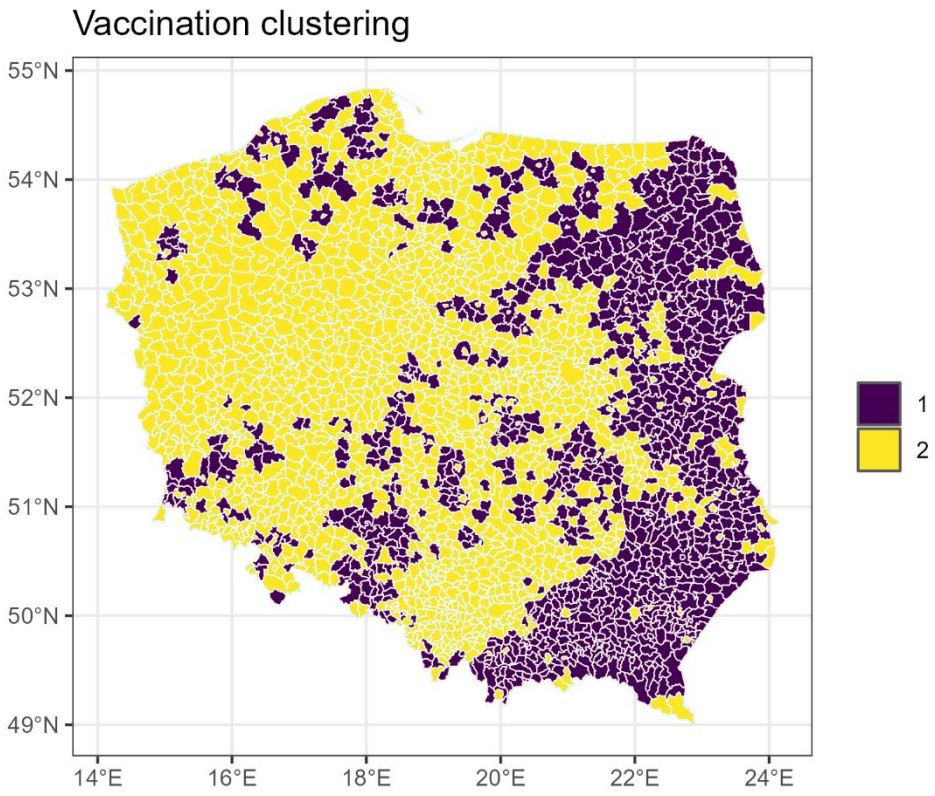
Results and Discussion

Unsupervised Learning

We start the analysis by clustering the vaccinations rates using simple k-means method. Although simple, k-means works well for linear data and is very often found in the literature on epidemic. Based on standard metrics, such as silhouette index and weighted sum of squares, we detected two clusters of the vaccination rates. We find significant geographical differences in the distribution of the clusters (Figure 4). We find that the first cluster (average vaccination: This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

44.2%) is located mainly in the eastern Poland. Still there are some discontinuities – large urban districts such as Krakow, Lublin or Rzeszow were assigned to cluster 2 (average vaccination: 57%). Therefore, we further explore the differences in the characteristics of the clusters.

Figure 4. K-means clustering of the vaccination rates



Source: Own elaboration based on GUS data

Table 2 presents the comparison of the characteristics between cluster 1 and cluster 2. Columns 1 and 2 present the mean characteristics, while column 3 presents the difference and its statistical significance using weighted t-test[6]. We find large and statistically significant differences between the clusters in terms of the political preferences and urbanisation. On average, municipalities characterized by higher share of mature (60+ population), with poorer access to public services in small and rural areas were less likely to get fully vaccinated against COVID-19. Interestingly, there are small, yet significant, differences in terms of voting for Konfederacja, which was the main opponet of the vaccination during COVID-19 pandemic. In

[6] We use the weights package in R

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

particular, since in cluster 2, characterized by higher vaccination, there was high support for Konfederacja, it may suggest its little role in driving the hesistancy towards vaccinations. On the other hand, there are large differences in the support of the left-wing party – SLD, which may suggest that it is not the hesitancy that has driven the trend, but rather the enthusiasm towards vaccinations.

Table 2. Differences in the characteristics of the vaccinations clusters

Variable	***Cluster 1***	***Cluster 2***	***Difference (t-value)***
Vaccination rates	44.2%	57%	-69.73***
Political Preferences			
Votes PiS	61.3%	40.2%	43.3***
Votes KO	13.7%	29.9%	-44.2***
Votes SLD	6.4%	13.7%	-46.7***
Votes Konferderacja	6.5%	6.8%	-3.46***
Urbanisation			
0-20 thousands ihabitants	21.4%	16%	3.47***
20-50 thousands ihabitants	74.7%	38.5%	
50-100 thousands ihabitants	3.95%	10.5%	-6.49***
100-500 thousands ihabitants	0%	20%	-19.6***
More than 500 thousands	0%	15%	-16.47***
Social factors			
Exposure to abnormal COVID-19 mortality	244.5	237.4	1.11
Public goods provision index	*0.31*	*0.39*	*-24.6****

Source: Own elaboration based on GUS data. Note: we used frequency weights to obtain the mean values and the differences between the clusters. * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

Linear regression

From the Table 3 it may be seen, that there are many important factors impacting the vaccination level. Political variables, reflecting the votes in 2019 parliamentary election are

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

significant The share of population supporting right PiS is associated with lower vaccine uptake (our baseline level) and KO, PSL and SLD being associated with higher vaccine level. the share of higly-educated people and the number of cars per 1000 people increasing the vaccination level. Surprisingly, the revenue from agricultural tax increaces vaccination level, while share of forest decreases, therefore the effect of being an agricultural municipality is not fully straightforward. Interestingly, the location of the municipality matters. The municipalities with higher longitude are characterized by lower vaccination level, while with higher latitude, they have higher vaccination uptake as well. Speaking of cardinal directions, the more to North, the higher vaccination rate, the more to East, the lower. The simple OLS model suggests that the lowest vaccination level shall be in the South-East, and the highest in the North-South.

SEM model provided the information that Lambda value is equal to 0.78522, therefore the spatial dependence of error is highly relevant. While OLS disregards the area of municipality, SEM shows that is has a weak but significant effect on vaccination. Birthrate is much more significant and stronger in SEM than in OLS, same for Prussian and Russian partitions. We show that partitions are important up-to-date for the behaviours of individuals concerning their health decisions.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Table 3. OLS and Spatial Error Model Regressions results

OLS estimates

	Dependent variable:		
	Vaccinated population in municipality (%)		percent_vaccinated
	OLS	OLS with robust s.e.	Spatial Error Model
	(1)	(2)	(3)
area_km2	-0.001	-0.001	-0.003***
	(0.001)	(0.002)	(0.001)
healthcare_advices	0.00000***	0.00000**	0.00000
	(0.00000)	(0.00000)	(0.00000)
persons_per_appartment	-2.488***	-2.488***	-2.420***
	(0.328)	(0.529)	(0.270)
forests_area	-0.0005***	-0.0005***	-0.0005***
	(0.0001)	(0.0002)	(0.0001)
bicycle_paths_per_10k_persons	-0.053***	-0.053***	-0.021**
	(0.014)	(0.014)	(0.008)
revenues_per_capita_PIT	0.001**	0.001	0.001***
	(0.0004)	(0.001)	(0.0003)
children_3_5_in_kindergartens	0.003***	0.003***	0.004***
	(0.001)	(0.001)	(0.0005)
birthrate_per_1000_persons	-0.052	-0.052	-0.116***
	(0.039)	(0.060)	(0.032)
Prussian	0.180	0.180	1.294*
	(0.342)	(0.767)	(0.754)
Russian	1.501***	1.501**	2.750***
	(0.298)	(0.642)	(0.638)
frekwencja	0.407***	0.407***	0.250***
	(0.022)	(0.027)	(0.021)
glosy_KO	0.252***	0.252***	0.178***
	(0.016)	(0.024)	(0.017)
glosy_KONF	-0.037	-0.037	-0.026
	(0.063)	(0.089)	(0.054)
glosy_PSL	0.342***	0.342***	0.158***
	(0.022)	(0.026)	(0.018)
glosy_SLD	0.294***	0.294***	0.216***
	(0.026)	(0.034)	(0.029)
high_exposure	-0.0004	-0.0004	-0.0001
	(0.001)	(0.001)	(0.0005)
index_ineq	-3.017	-3.017	1.147
	(1.904)	(2.349)	(1.434)
average_wage_relative	0.019**	0.019	-0.018
	(0.009)	(0.013)	(0.013)
doctors_per_1000_persons	-0.046***	-0.046***	-0.013
	(0.007)	(0.010)	(0.011)
cars_per_1000_persons	0.009***	0.009***	0.005***
	(0.001)	(0.002)	(0.002)
education_share_higher	0.171***	0.171***	0.100***
	(0.024)	(0.040)	(0.033)
tourits_per_1000_persons	-0.001***	-0.001***	-0.0003***
	(0.0001)	(0.0001)	(0.0001)
rolny_revenue_ratio	78.190***	78.190***	35.464***
	(9.561)	(9.698)	(7.153)
long	-0.732***	-0.732***	-0.939***
	(0.062)	(0.090)	(0.148)
lat	0.227***	0.227*	0.668***
	(0.078)	(0.120)	(0.234)
Constant	15.235***	15.235**	12.740
	(4.583)	(7.727)	(12.607)
Observations	2,477		2,477
R^2	0.869		
Adjusted R^2	0.868		
Log Likelihood			-6,447.863
$sigma^2$			9.216
Akaike Inf. Crit.			12,951.730
Residual Std. Error	435.139 (df = 2451)		
F Statistic	649.948*** (df = 25; 2451)		
Wald Test			2,514.174*** (df = 1)
LR Test			1,362.426*** (df = 1)
Note:			*p<0.1; **p<0.05; ***p<0.01

Source: Own elaboration

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Our next spatial model was SAR, estimated as generalized spatial two stage least squares, is more reliable model than both OLS, and SEM. It allows for interpretation of direct and indirect effects. While SEM suggested the significant impact of Russian partition, SAR indicated the significant impact of Prussian one. Again, agricultural tax, share of forests, bicycle paths, children per 1000 people were significant. The SAR model shows that not only social and economical variables inside the municipality are important, but also those in the nearest vicinity of them.

Table 4 Direct, indirect and total effects from the SAR model

	Direct Effects	Indirect Effects	Total Effects
area_km2	-0.003***	-0.003**	-0.007***
healthcare_advices	0**	0**	0**
persons_per_appartment	-2.322***	-2.141***	-4.463***
forests_area	-0.001***	-0.001***	-0.001***
bicycle_paths_per_10k_persons	-0.025**	-0.023**	-0.049**
revenues_per_capita_PIT	0.002***	0.001***	0.003***
children_3_5_in_kindergartens	0.006***	0.005***	0.011***
birthrate_per_1000_persons	-0.004	-0.004	-0.009
Prussian	-0.906**	-0.835**	-1.74**
Russian	0.571	0.527	1.098
frekwencja	0.232***	0.214***	0.447***
glosy_KO	0.132***	0.122***	0.253***
glosy_KONF	-0.167***	-0.154***	-0.321***
glosy_PSL	0.177***	0.163***	0.341***
glosy_SLD	0.207***	0.191***	0.398***
high_exposure	0.001	0.001	0.002
index_ineq	-3.047**	-2.809**	-5.857**
average_wage_relative	-0.022**	-0.02*	-0.042**
doctors_per_1000_persons	0.013	0.012	0.024
cars_per_1000_persons	0.002	0.002	0.004
education_share_higher	-0.017	-0.015	-0.032
tourits_per_1000_persons	0***	0***	-0.001***
rolny_revenue_ratio	36.862***	33.979***	70.841***
long	-0.381***	-0.352***	-0.733***
lat	0.527***	0.486***	1.013***

Source: Own elaboration based on GUS data. * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Instrumental Variables

Table 5 presents the results of OLS and GMM-IV regressions. We find that there was significant impact of the left-wing party's popularity on the vaccination rates. On average, an increase in left-wing popularity results in a 1.14 percentage points increase in the vaccination rate. On the other hand, when controling for education, demographics and spatial variables, we find no significant effect of the public goods provision. It needs to be noted that using the results from OLS would underestimate the role of public goods and the role of the left-wing party's popularity.

Table 5. OLS and GMM-IV results

	(1)	(2)	(3)	(4)
	OLS	OLS	OLS	OLS
Public Goods Provision	13.81***	16.47***	20.59***	8.39***
	(2.61)	(2.42)	(2.38)	(2.17)
Votes for SLD	0.94***	0.62***	0.61***	0.21***
	(0.03)	(0.04)	(0.04)	(0.04)
	2SLS	2SLS	2SLS	2SLS
Public Goods Provision	10.01**	22.42***	22.77***	2.37
	(15.4)	(2.80)	(4.10)	(3.03)
Votes for SLD	1.08***	0.88***	0.77***	0.25***
	(0.06)	(0.05)	(0.06)	(0.06)
Education	Yes	No	No	Yes
Demographic Structure (age, gender)	No	Yes	No	Yes
Longitude and Lattitude	No	No	Yes	Yes
First Stage Kleibergen-Paap F-Statistic	689	813	829	595
Observations	2477	2477	2477	2477

Source: Own elaboration based on GUS data. * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team "Maximum (Victory) Likelihood EsTeamation".

Machine learning methods

Another approach to identify the drivers of COVID-19 vaccination in Poland relied on machine learning models. In table 6, we present the performance of our trained and optimized models on validation and testing sets calculated with the use of root mean squared error metric. The XGBoost algorithm has outperformed other methods on both the validation and testing samples. The RMSE value indicates that the XGBoost made a prediction error mistake of approximately 3,47% on the unseen data. Next in order were: Random Forest, Support Vector Machines and K-Nearest Neighbors algorithms. The lower value of RMSE for the SVM model on the testing sample compared to validation set is caused by a significant variance of the results during the cross-validation stage.

Table 6. Comparision of the machine learning models performance

RMSE	*KNN*	*SVM*	*Random Forest*	*XGBoost*
Validation	3,78	3,92	3,72	3,35
Testing	3,95	3,81	3,76	3,47

Source: Own elaboration

Figure 5. presents calculated SHAP values for a group of variables with the most significant effect on the dependent variable with the use of optimized XGBoost model. The impact on the dependent variable is sorted in a descending order. Variables describing the longitude, latitude, percent of votes for leftist parties, voter turnout and percent of the eldest part of population in municipalities appeared to be the main drivers of XGBoost model's predictions. SHAP value on x-axis bigger than zero means that given level of the feature positively impacts the predictions (yellow color is equivalent to high values of a variable and dark blue to low).

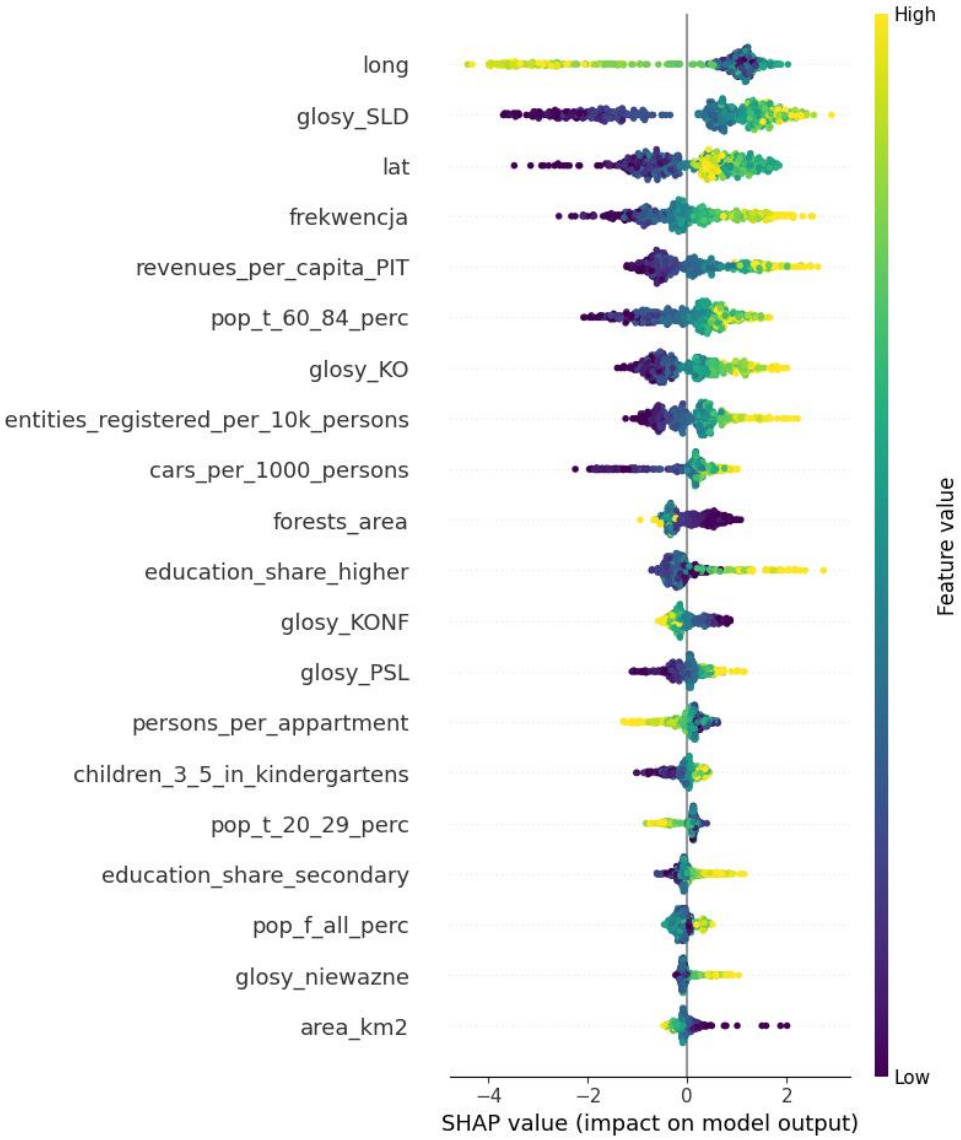
Figure 6. is composed of the set of two SHAP dependence plots for arbitrary chosen most significant variables in terms of the aim of our research. They how the impact of political preferences in the society on the percentage of vaccinated people. It is clearly visible that the higher percent of votes for the SLD (Polish leftist party), the higher the values of the dependent variable. Contrary, significant support for Konfederacja (Polish rightist party) is characterized

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team "Maximum (Victory) Likelihood EsTeamation".

by negative relationship with the popularity of vaccines. Next, let's compare the two most powerful variables connected with the age structure of population (Figure 7.). A higher percentage of elderly individuals, aged between 60 and 84 years, in the specific municipality demonstrates nearly a linear and positive correlation with vaccination uptake. On the other hand, a significant proportion of young adults in a society presents a negative impact on the percentage of vaccinated citizens. Interesting, non-linear patterns are visible during examination of the Figure 8 focused on a geographical location of Polish municipalities. Small and medium values of longitude (representing western and central parts of Poland) demonstrate consistently positive impact on vaccine uptake. There is a notable shift in the effect for eastern regions of Poland, however border areas present a higher propensity to vaccination. The relationship between latitude and the percentage of vaccinated individuals exhibits a polynomial shape, but in general the visible pattern is further north regions demonstrate higher values of dependent variable.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team "Maximum (Victory) Likelihood EsTeamation".

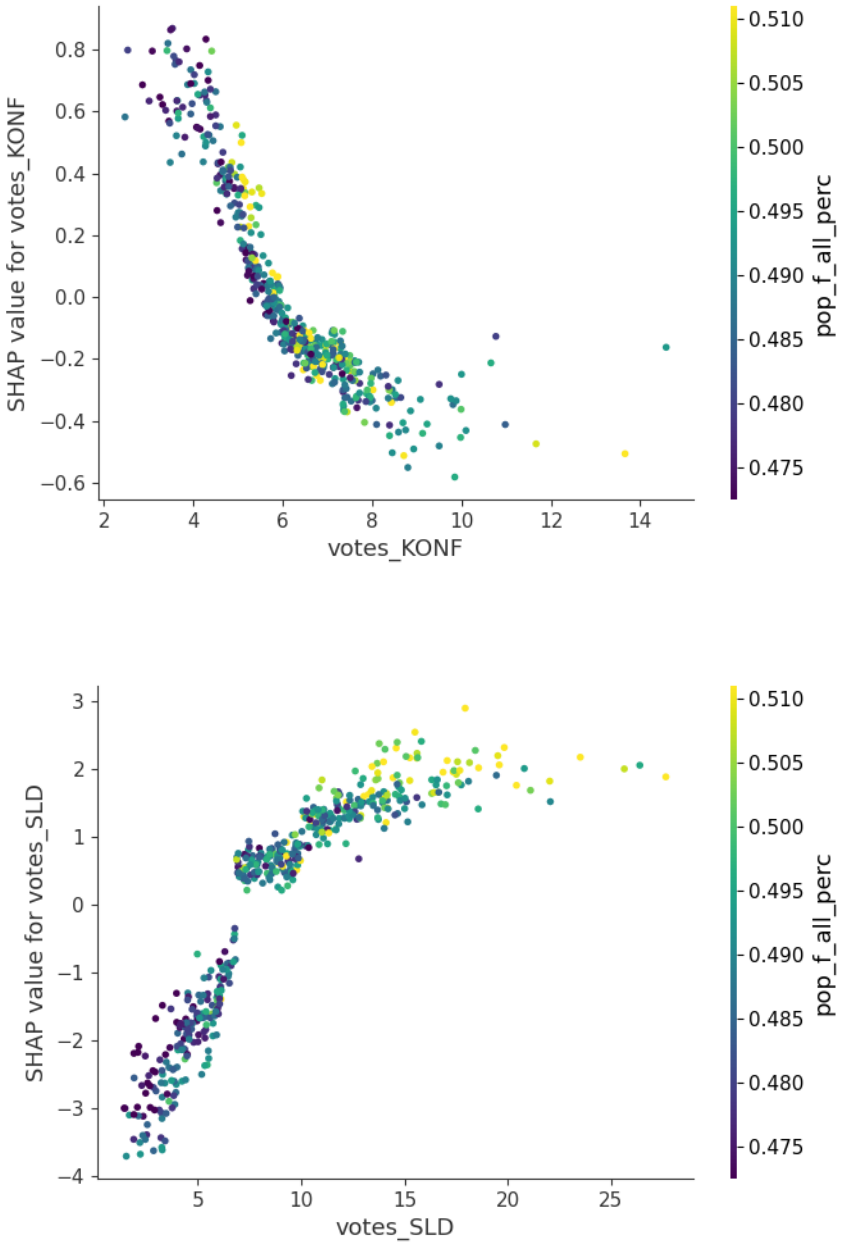
Figure 5. SHAP values for the most significant variables



Source: Own elaboration

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

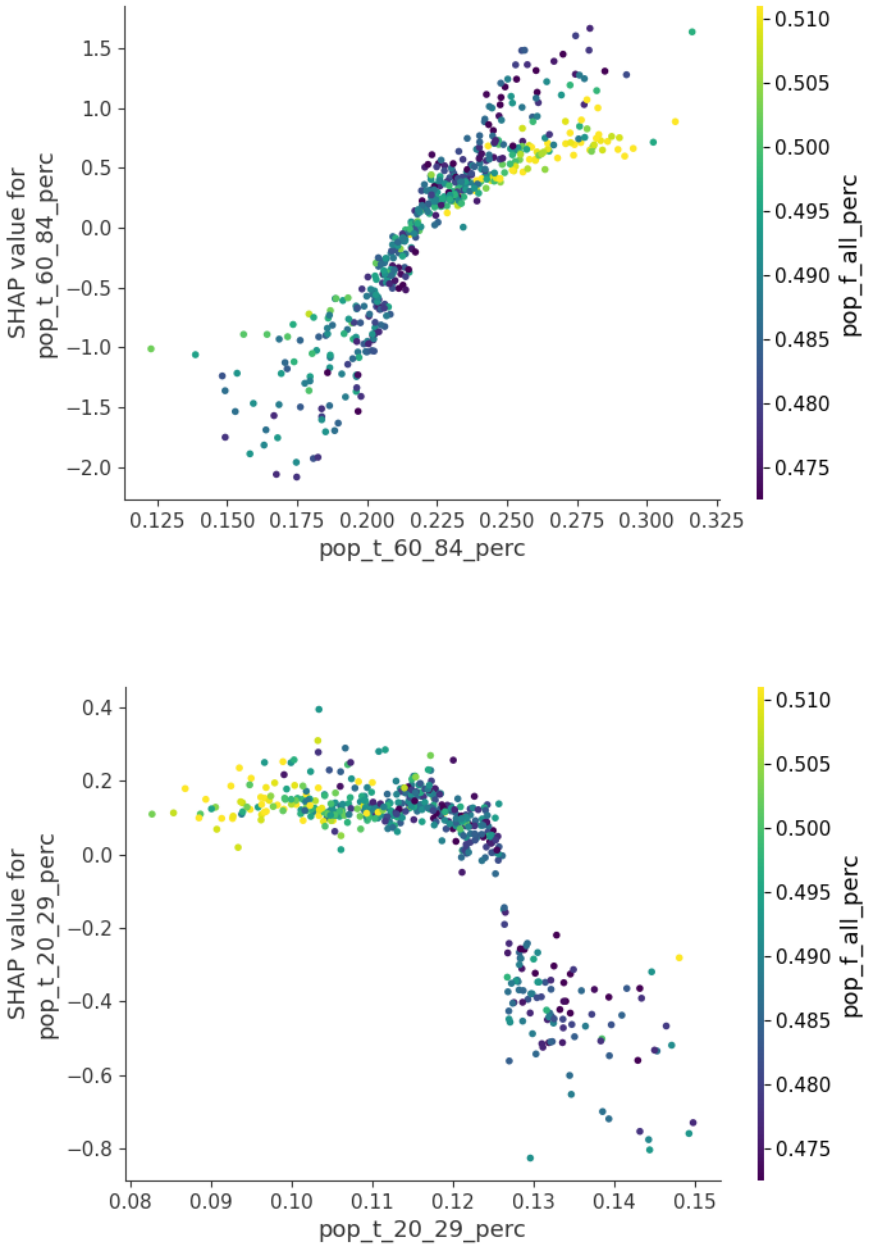
Figure 6. SHAP values for the variables concerning political views



Source: Own elaboration

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

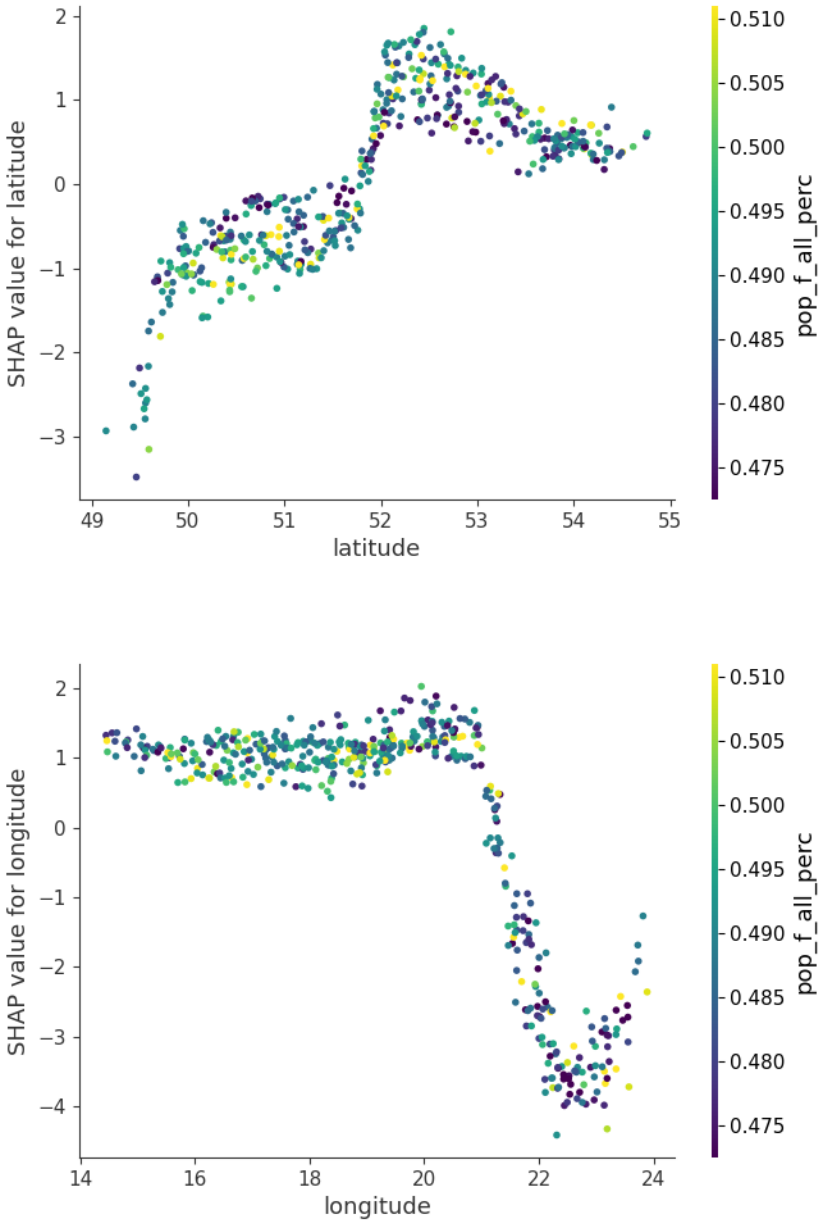
Figure 7. SHAP values for the variables concerning political age groups



Source: Own elaboration

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Figure 8. SHAP values for the variables concerning geographical location



Source: Own elaboration

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Regression discontinuity design

Table 7. Results of regression discontinuity design analysis

Coeffcient	*estimate*	*std.error*	*t.statistic*	*p.value*	*Cutoff (meters)*	*df*
(Intercept)	63,000	2,404	26,212	3,01E-08	2000	*7*
Prussian	-3,350	2,944	-1,138	0,293	2000	*7*
(Intercept)	57,424	1,093	52,514	3,10E-51	4000	*59*
Prussian	-1,968	1,644	-1,197	0,236	4000	*59*
(Intercept)	54,881	0,867	63,297	1,33E-94	6000	*121*
Prussian	-0,453	1,374	-0,329	0,742	6000	*121*
(Intercept)	54,579	0,805	67,840	1,51E-115	8000	*152*
Prussian	0,141	1,238	0,114	0,909	8000	*152*
(Intercept)	54,291	0,773	70,200	1,57E-128	10000	*172*
Prussian	0,973	1,186	0,820	0,413	10000	*172*
(Intercept)	54,130	0,739	73,276	8,04E-143	12000	*193*
Prussian	0,787	1,126	0,699	0,485	12000	*193*
(Intercept)	54,342	0,666	81,634	2,09E-165	14000	*218*
Prussian	0,721	1,029	0,700	0,485	14000	*218*
(Intercept)	54,450	0,610	89,316	2,25E-188	16000	*244*
Prussian	0,693	0,947	0,732	0,465	16000	*244*
(Intercept)	54,508	0,591	92,228	3,81E-206	18000	*270*
Prussian	0,185	0,905	0,204	0,839	18000	*270*
(Intercept)	54,092	0,566	95,651	3,56E-225	20000	*297*
Prussian	0,807	0,851	0,948	0,344	20000	*297*

Source: Own elaboration

The results of the RDD contradict our investigated hypotheses. Regardless of the cutoff hyperparameter, the 'Prussian' binary variable appears to be insignificant. Furthermore, in cases where the cutoff is relatively low, the point estimation falls below zero, undermining our previous assumption of the effect of partitions.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Conclusions

The vaccinations against COVID-19 have been an important factor in decreasing the abnormal mortality rates in Poland and beyond. Lewandowski and Madoń (2022) indicated that only in 2021, the vaccines have saved more than 61 thousands individuals. However, their effectiveness is strongly related to the share of individuals vaccinated in the close proximity. We have found similar results regardless of the applied methods – spatial regressions, unsupervised learning, instrumental variables and machine learning techniques. Our results show the large significant importance of the spatial, developmental and political factors that influence the COVID-19 uptake. We find that while historical partitions do not play the major role in driving the trends, we find that there are large differences between Eastern Poland (Lubelskie, Podlaskie and Podkarpackie) and the remaining regions. In particular these regions are more often characterized by lower urbanisation and higher popularity of the PiS political party. We find that in regions characterised by higher share of elderly, there is a larger share of the COVID-19 intake in reference to the youth. Further exploration using instrumental variables, we find that provision of public services and support for the left-wing parties is associated with higher intake of the vaccines. Interestingly, despite of the many information spread on the COVID-19 hesitancy among the far right-wing politicians, it is rather the enthusiasm of the left-wing voters that drives the trends.

Our results shed new light on the possible role of political parties in driving the COVID-19 intake. In Poland, only Konfederacja was openly against the vaccines during the pandemic. However, regions characterised by high and low intakes of the vaccines had similar share in the voting for the far right-wing parties in Poland. Therefore, the opportunity political cost in keeping up with the anti-vaccinations views was relatively low, as there were possibly different factors that driven the political preferences. At the same time, regions characterised by low intake of the vaccination, were mostly the supporters of PiS – the governing party at that moment. Despite some political actions towards increasing the vaccinations rates, solutions focused on the Eastern part of the country could have played a large role in increasing the vaccination rates.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

References

Acemoglu, D., Chernozhukov, V., Werning, I., & Whinston, M. D. (2021). Optimal targeted lockdowns in a multigroup SIR model. *American Economic Review: Insights*, *3*(4), 487-502.

Altig, D., Baker, S., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., ... & Thwaites, G. (2020). Economic uncertainty before and during the COVID-19 pandemic. *Journal of public economics*, *191*, 104274.

Bilal, U., Mullachery, P. H., Schnake-Mahl, A., Rollins, H., McCulley, E., Kolker, J., ... & Diez Roux, A. V. (2022). Heterogeneity in spatial inequities in COVID-19 vaccination across 16 large US cities. *American journal of epidemiology*, *191*(9), 1546-1556.

Bloom, N., Bunn, P., Mizen, P., Smietanka, P., & Thwaites, G. (2023). The impact of COVID-19 on productivity. *Review of Economics and Statistics*, 1-45.

Brown, C. C., Young, S. G., & Pro, G. C. (2021). COVID-19 vaccination rates vary by community vulnerability: A county-level analysis. *Vaccine*, *39*(31), 4245-4249.

Cheong, Q., Au-Yeung, M., Quon, S., Concepcion, K., & Kong, J. D. (2021). Predictive modeling of vaccination uptake in US counties: A machine learning–based approach. *Journal of medical Internet research*, *23*(11), e33231.

Jakubowski, M., Gajderowicz, T., & Patrinos, H. A. (2023). Global learning loss in student achievement: First estimates using comparable reading scores. *Economics Letters*, *232*, 111313.

Hasan, M. K., Jawad, M. T., Dutta, A., Awal, M. A., Islam, M. A., Masud, M., & Al-Amri, J. F. (2021). Associating measles vaccine uptake classification and its underlying factors using an ensemble of machine learning models. *IEEE Access*, *9*, 119613-119628.

Juen, C. M., Jankowski, M., Huber, R. A., Frank, T., Maaß, L., & Tepe, M. (2023). Who wants COVID-19 vaccination to be compulsory? The impact of party cues, left-right ideology, and populism. *Politics*, *43*(3), 330-350.

Kopczewska, K. (2013). The spatial range of local governments: does geographical distance affect governance and public service?. *The Annals of Regional Science*, *51*, 793-810.

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Lewandowski, P., & Madoń, K. (2022). Skuteczność Szczepień Przeciw COVID-19 w Polsce. *IBS Policy Paper. Od Badań do Polityki Publicznej*, *1*, 1-14.

Lillebråten, A., Todd, M., Dimka, J., Bakkeli, N. Z., & Mamelund, S. E. (2023). Socioeconomic status and disparities in COVID-19 vaccine uptake in Eastern Oslo, Norway. *Public Health in Practice*, *5*, 100391.

Meng, L., Masters, N. B., Lu, P. J., Singleton, J. A., Kriss, J. L., Zhou, T., ... & Black, C. L. (2023). Cluster analysis of adults unvaccinated for COVID-19 based on behavioral and social factors, National Immunization Survey-Adult COVID Module, United States. *Preventive Medicine*, *167*, 107415.

PKW. (2019). Dane w arkuszach—Wybory do Sejmu i Senatu Rzeczypospolitej Polskiej 2019 r. https://sejmsenat2019.pkw.gov.pl/sejmsenat2019/pl/dane_w_arkuszach?fbclid=IwZXh0bgNhZW0CMTAAAR2IezodiA2UR4NRcn14TP14seJRLCJeApHi0zMjvV1DqrNORnibOoOVsXU_aem_AaFEwcdb9oIwKgQIGJBuiBogu57pXjE54cP9ghvVnMKlEjtglSxrf963lNenSIroDlhLatJtGVuGrxZkpfGEAWqN

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

Appendix

source	*new_label*	*variable_description*
wec	municipality_code	unique municipality code (TERYT) - 8 digit number. The first two digits represent the code of the voivodeship (region), and the last digit represents a type of the municipality (1 = urban, 2 = rural, 3 = mixed urban-rural)
wec	municipality_name	name of municipality
wec	percent_vaccinated	percent_vaccinated
wec	county_code	code of the county in which the municipality is located
wec	area_km2	area in square kilometeres area in square kilometeres
wec	urbanization_rate	urbanization rate (share of the urban area within the municipality)
wec	healthcare_advices	primary health care - total number of advices
wec	installations_watersupply	apartments equipped with installations, % of apartments with water supply
wec	installations_toilet	apartments equipped with installations, % of apartments with flushable toilet
wec	installations_central_heating	apartments equipped with installations, % of apartments with central heating
wec	installations_network_gas	apartments equipped with installations, % of apartments with network gas
wec	persons_per_appartment	average number of people per apartment
wec	persons_per_library	population per 1 library facility
wec	library_books_per_1000_persons	library book collection per 1 inhabitants
wec	library_readers_per_1000_persons	public library readers per 1 inhabitants
wec	library_loans_per_reader	loan of book collections per reader in volumes
wec	forests_area	total area of forests (in hectars)
wec	bicycle_paths_per_100km2	bicycle paths per 100 km2
wec	bicycle_paths_per_10k_persons	bicycle paths per 10 thousand population
wec	revenues_per_capita_PIT	total revenues of municipality budget from personal income tax per inhabitant
wec	revenues_per_capita_CIT	total own revenues of municipality budget from corporate income tax per inhabitant
wec	investment_expenditures_per_capita	total investment expenditures of municipality budget per inhabitant
wec	children_3_5_in_kindergartens	children in kindergartens and other forms of pre-school education per 1 children aged 45356
wec	marriages_per_1000_persons	marriages per 1000 inhabitants
wec	birthrate_per_1000_persons	birthrate per 1000 inhabitants

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.

wec	net_migrations_per_1000_persons	net migrations per 1000 inhabitants
wec	entities_registered_per_10k_persons	number of entities registered per 10 thousand inhabitants
wec-pow	average_wage_relative	-
own	Prussian	Binary
own	Russian	Binary
dane.gov	województwo	-
dane.gov	frekwencja	-
dane.gov	glosy_niewazne	-
dane.gov	glosy_wazne	-
dane.gov	glosy_KO	Votes for KO
dane.gov	glosy_KONF	Votes for KONF
dane.gov	glosy_PSL	Votes for PSL
dane.gov	glosy_PIS	Votes for PIS
dane.gov	glosy_SLD	Votes for SLD
transformation	pop_t_0_19_perc	Age group
transformation	pop_t_40_49_perc	Age group
transformation	pop_t_20_29_perc	Age group
transformation	pop_t_50_59_perc	Age group
transformation	pop_t_60_84_perc	Age group
transformation	pop_f_all_perc	Females
bdl	avg_gosp_wiejskie	Average size of farm
bdl	unemployment_50	Unemployment above 50 years old
bdl	unemployment_25	Unemployment below 25 years old
transformation	high_exposure	Index of exposure to COVID-19
transformation	index_ineq	Index of inequality
wec_pow	average_wage_relative	-
wec_pow	doctors_per_1000_persons	-
wec_pow	beds_in_hospitals	-
wec_pow	cars_per_1000_persons	-
wec_pow	motorcycles_per_1000_persons	-
wec_pow	education_share_higher	-
wec_pow	education_share_secondary	-
wec_pow	education_share_vocational	-
wec_pow	tourits_per_1000_persons	-
wec_pow	population_total_log	-
transformation	rolny_revenue_ratio	Taxes from rural tax per inhabitant vs all revenues
own	type_agr	-
own	type_0_20k	-
own	type_50_100k	-
own	type_500k+	-
own	type_100_500k	-

This research is the outcome of the Warsaw Econometric Challenge 2024. All the authors are the members of the team “Maximum (Victory) Likelihood EsTeamation”.