The drivers of the level of COVID-19 vaccination in Poland

Tymoteusz Barciński
Ziemowit Głowaczewski
Jakub Bazyluk
Antonina Ślubowska, MD

Warsaw University of Technology,
Faculty of Mathematics and Information Science

May 12, 2024

Abstract

This study investigates the influence of various independent variables on vaccination rates across municipalities, employing both econometric and machine learning methodologies. Emphasizing statistical inference, our research aims to validate hypotheses pertaining to the vaccination process.

In the econometric approach, linear regression models are initially utilized, integrating feature engineering and selection techniques to handle a multitude of predictors. Addressing nonlinearities, additive models with spline basis expansions are considered, ensuring robustness in model specification. Spatial correlation structures are later incorporated using SARAR models, validated through diagnostic tests such as Moran's I.

Concurrently, the machine learning approach serves to corroborate findings from the econometric model, employing explainable AI to ensure interpretability. Notably, the entire dataset is utilized for model estimation to uphold statistical rigor.

Findings indicate significant correlations between vaccination rates and political views, demographic factors, and spatial context. However, discrepancies between expected and observed correlations prompt further exploration, suggesting nuanced relationships requiring deeper analysis.

Overall, our study provides valuable insights into the complex interplay of factors influencing vaccination rates, offering a comprehensive understanding of the vaccination process across municipalities.

1 Background

The COVID-19 pandemic, officially declared by the WHO in March 2020, has deeply affected global society, impacting public health, economics, and social dynamics. Healthcare systems worldwide faced shortages of vital resources, leading to the adoption of restrictive measures like lockdowns and social distancing protocols. While these measures aimed to control the spread of infections, they also disrupted social interactions and economic activities, causing widespread financial losses and job insecurity. By May 2023, collaborative efforts between scientists and public health campaigners led to the development and deployment of effective vaccines, signaling the official end of the pandemic. Despite this milestone, the lessons learned from addressing COVID-19 vaccine hesitancy remain relevant. Vaccine hesitancy, once a significant challenge, may resurface in future, underscoring the need to understand its determinants for preparedness and response efforts in the years to come. Positioned in Central and Eastern Europe, Poland offers a unique context for studying vaccination behavior, often overlooked compared to more extensively studied regions like the US and UK. Poland's historical experiences, socio-cultural dynamics, and political landscape significantly shape public health policies and vaccination strategies. Furthermore, analyzing regional disparities in vaccination rates within Poland can provide insights into the impact of socio-economic factors, healthcare infrastructure, and access to healthcare services. Despite the availability of COVID-19 vaccines, Poland has shown lower vaccination rates compared to the EU average, underscoring the importance of understanding the underlying reasons for this hesitancy. [1] By thoroughly examining these factors, policymakers and healthcare professionals can develop targeted interventions to effectively address vaccine hesitancy and enhance public health outcomes.

2 Literature review

To contextualize our study, we conducted a literature review to explore existing research on COVID-19 vaccination acceptance and hesitancy, aiming to understand the various approaches and findings in this area.

The study [6] aimed to assess the factors influencing COVID-19 vaccination

acceptance among the Polish population, analyzing sociodemographic factors and physical and mental health status. The evaluation was conducted through a survey involving 200,000 participants from Poland. Logistic regression analysis was employed to identify significant associations between these factors and vaccine acceptance or refusal. The research identified fear of post-vaccination complications and safety concerns as primary reasons for vaccine hesitancy. Male respondents with lower education levels exhibited more negative attitudes, while older individuals, those with higher education, and residents of larger cities were more likely to accept vaccination. Good physical and mental health conditions were also associated with vaccine acceptance.

Another research [11] examined the predictors of delayed COVID-19 vaccine uptake in Polish sub-regions using regression models. Regression analysis was employed to explain the increase in COVID-19 vaccination rates across 378 Polish sub-regions. The analysis compared vaccination rates for age groups of 20 years and older between June 30, 2020, and January 31, 2021. The results indicate that initial high vaccination rates did not reduce willingness to vaccinate, but rather increased it, leading to greater disparities between regions. Support for Eurosceptic and anti-establishment parties strongly correlated with persistent vaccine hesitancy. Social inclusion markers like voter turnout and employment rate remained significant predictors, while higher education levels showed mixed effects on vaccination rates across different age groups. These findings suggest that vaccine hesitancy in Poland is influenced by political views, social exclusion, and regional historical context, rather than anti-vaccine movements.

One approach to assess the various factors contributing to vaccine hesitancy is through the application of psychological concepts like the 5C framework. This model identifies five key factors—confidence, complacency, convenience, risk calculation, and collective responsibility—that influence individuals' decisions regarding vaccination. [2], [8] The study [10] explored COVID-19 vaccine hesitancy using the 5C model across a national and South Carolina sample. Analyzing data from a national sample (n = 1634) and a South Carolina sample (n = 784), the study investigates how the 5C drivers of vaccine behavior influence early adoption and intentions. Results reveal that the South Carolina sample exhibits lower vaccine intentions and higher levels of 5C barriers compared to the national sample. Moreover, demographic factors such as race and key drivers like confidence and collective responsibility significantly impact vaccine trust and intentions across both groups.

In our solution, we incorporate insights drawn from the literature review to tackle COVID-19 vaccine hesitancy and boost vaccination rates.

3 Aim of the study

Our study aims to investigate the factors influencing COVID-19 vaccination rates at the municipal level in Poland. By analyzing vaccination data alongside socio-demographic, economic, and healthcare-related characteristics of municipalities, we seek to adapt the 5C model to formulate hypotheses about the municipality-level drivers of vaccination uptake. Additionally, we have developed hypotheses not directly related to the 5C components, considering factors such as population density, urban/rural classification, historical partition terrain, and income per capita, as well as spatial dependencies.

3.1 Hypotheses

3.1.1 5C-based hypotheses

Confidence This component refers to trust in the effectiveness and safety of vaccines, as well as in the healthcare system and authorities that promote vaccination. Factors that can affect confidence include misinformation, mistrust in healthcare providers or pharmaceutical companies, and concerns about vaccine side effects.
As a potential measure of general trust in authorities, we considered the percentage of participants in parliamentary elections. Additionally, we investigated the voting patterns of residents in each municipality.
We hypothesized that a higher voter turnout correlates with a higher vaccination rate, and that political affiliation impacts vaccination acceptance rates.

Complacency Complacency refers to the perception of the risk posed by vaccine-preventable diseases. When individuals perceive these diseases as low-risk or non-threatening, they may become less inclined to get vaccinated. Due to the increased risk that elderly individuals face from developing severe forms of COVID-19 and related complications, they exhibit greater willingness to receive vaccinations. Moreover, vaccination programs often prioritize this demographic.
Consequently, our hypothesis suggests that municipalities with a higher proportion of seniors will exhibit higher vaccination rates, while those with a higher proportion of individuals under the age of 20 will likely demonstrate lower vaccination rates.

Convenience Convenience refers to the ease of access to vaccination services. Barriers to vaccination, such as long wait times, inconvenient clinic hours, or lack of transportation to vaccination sites, can reduce vaccine uptake.
Based on this, we hypothesize that the number of cars per 1000 inhabitants will positively correlate with vaccination rates.
Additionally, we obtained data on the number of vaccination sites situated within a 10-kilometer radius of the municipality's center and minimal distances from the center of the municipality to the nearest vaccination point in March 2021, at the outset of the vaccination campaign. We regarded these factors as indicators of the intensity of vaccination efforts within each municipality.
Our hypothesis posited that these factors would positively correlate with the final vaccination rate in 2021.

Risk calculation Calculation refers to the process individuals use to weigh the risks and benefits of vaccination. Factors influencing this calculation include perceived vaccine efficacy, perceived severity of vaccine-preventable diseases, and perceived risk of vaccine side effects. Individuals may be more likely to accept vaccination if they perceive the benefits of vaccination to outweigh the risks.
Based on this understanding, we hypothesize that a higher percentage of residents with higher education levels will positively correlate with vaccination rates. This assumption stems from the notion that education fosters a better understanding of the benefits of vaccination, thereby increasing acceptance rates.

Collective responsibility Collective responsibility refers to the sense of duty individuals feel towards protecting the health of their community through vaccination. Factors influencing collective responsibility include social norms surrounding vaccination, perceived social pressure to vaccinate, and the belief that vaccination is a civic duty. Strengthening collective responsibility can help foster a culture of vaccination acceptance.
This concept intersects with the hypotheses outlined in section 3.1.1.

3.1.2 Other factors

In addition to examining the 5C components, our study aimed to explore how the diverse characteristics of Polish municipalities relate to vaccination rates. These characteristics include population density, urban or rural classification, historical partitions of Poland's terrain, and income per capita,

reflecting the economic diversity across Poland's regions.
Our hypothesis suggests that all of these factors play a role in shaping vaccination rates. Particularly, we anticipate a positive correlation between income per capita and vaccination rates, as higher income levels often improve access to vaccination services, thereby enhancing convenience aspect mentioned in 3.1.1.3.

Last but not least, we decided to verify whether the vaccine acceptance in a municipality's neighborhood affects the vaccination rate in the municipality itself. Intuitively, we formulated the hypothesis that these factors should be related, even after controlling for all other significant variables.

3.2 Summary

All hypotheses are summarized in the table below.

Table 1: Summary of Hypotheses

Hypothesis	**Description**
H1	Higher voter turnout correlates with a higher vaccination rate.
H2	Political affiliation impacts vaccination acceptance rates.
H3	Municipalities with a higher proportion of seniors exhibits higher vaccination rates, while those with a higher proportion of individuals under the age of 20 likely demonstrates lower vaccination rates.
H4	Municipalities with a higher proportion of seniors exhibits higher vaccination rates.
H5	Intensity of vaccination efforts at the outset of the vaccination campaign positively correlates with the final vaccination rate in 2021.
H6	Higher percentage of residents with higher education levels will positively correlate with vaccination rates.
H7	Population density, urban or rural classification, historical partitions of Poland's terrain, and income per capita play a role in shaping vaccination rates. There is a positive correlation between income per capita and vaccination rates.
H8	Vaccine acceptance in a municipality's neighborhood affects the vaccination rate in the municipality itself, even after controlling for all other significant variables.

4 Data preparation

4.1 Data provided by the Warsaw Econometric Challenge organizers

The organizers provided datasets with information about polish municipalities (describing the situation in 2021) and counties (describing the situation in 2020). The data revolves around the percentages of vaccinated population and socio-economic situation of given locations. Moreover, spatial mapping of said municipalities and counties was provided, with an addition of potential historical / geographical divisions – the map of partitioned Poland and east-west split.

Name of data file	Year of measurement
data_municipalities.csv	2021
data_counties.csv	2020

4.2 Additionally acquired data

Shape files for municipalities and historical partitions of Poland Dataset about the 2019 parliamentary election results by municipality from the National Elections Committee. Dataset about vaccination points in Poland from 09-03-2021. [9]

4.3 Dependent variable

The `percent_vaccinated` is a variable whose dependencies are studied in this paper. It is showing the percentage of vaccinated population per municipality. On the national level, we can observe that the variable displays a significant variance depending on location of the measurement. Furthermore, basing on papers on the topic, [5] it is observed that political affiliation is a strong predictor of a person's willingness to vaccinate.

4.4 Variables description

In our municipality-wise analysis, we used the following variables:

- `voters_turnout` - Per municipality percentage voter turnout

- `SLD_percent`, `PO_percent`, `Konfederacja_percent`, `PSL_percent` - Percentage results of Polish 2019 Parliamentary elections for the opposition parties per municipality
- Logarithm of `revenues_per_capita_PIT` - Logarithm of national budget in income tax revenues from the given municipality
- `percent_over_60` - Percentage of municipality population above 60 years old
- Square root of `min_dist` - Square root of minimum distance to a vaccination point from the given municipality in kilometers
- `no_loc_in10km` - Number of vaccination points within a 10-kilometer radius from the municipality
- `education_share_higher` - Percentage share of population with higher education
- `healthcare_advices_ratio_total` - Number of healthcare advices provided per capita
- Second degree polynomial transformation of `cars_per_1000_persons` - Number of cars per 1000 capita

4.5 Features Understanding

- `political_variables` - From available scientific sources, we deduced that political views and local political engagement are important factors in personal attitudes towards vaccination.
- Logarithm of `revenues_per_capita` - We decided that income tax revenue is a good metric of a municipality's economic situation, which correlates with willingness to vaccinate.
- Similar correlation occurs with `education_share_higher` and `healthcare_advices_ratio_total` - Predictors of belief in scientific and medical methods, resulting in an inclination towards vaccination.
- Variables relating to spatial relations of municipalities to vaccination points - `no_loc_in10km`, `min_dist`, and `cars_per_1000_persons` are metrics of vaccination availability in given locations, and capabilities to reach said locations.

5 Methodology

The effect of independent variables on the `percent vaccinated` for each municipality was modelled by two different approaches. Firstly, the econometric approach based on the linear models, which was later extended to take into account the spatial correlation structure between observations, and secondly the machine learning approach. Note, that the emphasis in this study is on the statistical inference to draw conclusions related to the hypothesis list. Therefore, the machine learning approach is included to confirm or reject the conclusions drawn from the linear model based on the statistical tests, via the explainability artificial intelligence. Moreover, since the goal was statistical inference, the dataset was not divided into the training and testing parts, but the entire dataset of all municipalities was used for model estimation.

The dependent variable `percent vaccinated` proposed by the organizers was bounded between zero and one hundred. We decided to transform the variable via the logit function after dividing it one hundred. This allows the new dependent variable to be unbounded which doesn't violate the assumption of the linear model.

6 Econometric approaches

6.1 Linear model

6.1.1 Approach

Linear models are widely used in many different areas of modelling due to interpretability and inference. In our study we used linear regression modelling approach to allow for the statistical inference on coefficients. Due to the high number of potential predictors after extensive feature engineering, feature selection methods were used. The forward and backward feature selection methods were utilized, which allowed for initial diagnostics of significant predictions. Moreover, linear Pearson correlation between independent variables and the dependent variable was considered. After a subset of predictor was chosen, the modelling approach focused on adding new variables to the linear model in an interative fashion. The nonlineartites were dignozed and accounted for by the usage of the partial residual plots. Moreover, the additive models with spline basis expansions were considered to detect potential nonlinearities, [12] which were later parameterized and included in the linear model (such as via the square root, logarithm or quadratic expansion). Importantly, the variables that related to our hypothesis were chosen. Note that those variables were not dropped in the model reduction phase,

which was performed via the F tests, because of their significance for stated hypothesis.

6.1.2 Results

Details of the estimated linear regression model are presented in 1. The variables related to the political views were found to be significant. The PIS_percent variable was excluded from the model due to high correlation with the variable PO_percent (-0.88). Spatial variables related to partitions and municipality type are also significant, and almost all of them (except for mixed municipality type) contribute positively to the vaccinated percent.

The function of form of some independent variables included in the model are seen in the partial residual plots in the figure 4. It can be visually identified that the coefficients estimated by the linear model are sensible.

The assumptions of the linear model were verified. The figure 2 shows the standard diagnostics plots. By visual diagnostics, we see the heavy tails of the residuals from the QQ plot, but apart from that no serious validations of the linear models assumptions are reported. The lack of multicolinearity between the independent variables were confirmed by the Varaince Inflation Factor shown in figure 3. The model does not take the spatial dependence into account so the spatial correlation of the residuals were investigated, initally by the map shown in figure 5. We clearly see the spatial dependence of residuals in certain parts of Poland. This dependence was confirmed by the Moran's I test with the $pvalues \leq 0.001$. Therefore the assumptions of the linear model are not fully satisfied and it motivates us to consider tha spatial regression models in the subsequent section.

```
Call:
lm(formula = y ~ factor(partitions) + factor(type_of_municipality) +
    sqrt(SLD_percent) + PO_percent + frekwencja_wyborcza + PSL_percent +
    Konfederacja_percent + revenues_per_capita_PIT_log + percent_over_60 +
    sqrt(min_dist) + no_loc_in10km + education_share_higher +
    poly(cars_per_1000_persons, 2) + healthcare_advices_ratio_total,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79138 -0.11485 -0.00213  0.12031  0.87480

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.292234   0.109988 -20.841  < 2e-16 ***
factor(partitions)2              0.036094   0.012274   2.941 0.003305 **
factor(partitions)3             -0.073719   0.012943  -5.696 1.37e-08 ***
factor(type_of_municipality)2    0.020530   0.018214   1.127 0.259774
factor(type_of_municipality)3    0.033433   0.016721   1.999 0.045671 *
sqrt(SLD_percent)                1.296565   0.082300  15.754  < 2e-16 ***
PO_percent                       0.787015   0.070429  11.175  < 2e-16 ***
frekwencja_wyborcza              1.402662   0.091279  15.367  < 2e-16 ***
PSL_percent                      1.235108   0.086294  14.313  < 2e-16 ***
Konfederacja_percent            -1.549119   0.284885  -5.438 5.93e-08 ***
revenues_per_capita_PIT_log      0.081831   0.018221   4.491 7.41e-06 ***
percent_over_60                  0.828644   0.160702   5.156 2.72e-07 ***
sqrt(min_dist)                   0.015344   0.004034   3.804 0.000146 ***
no_loc_in10km                    0.003298   0.001345   2.452 0.014287 *
education_share_higher           0.006999   0.001245   5.623 2.09e-08 ***
poly(cars_per_1000_persons, 2)1  2.224090   0.225669   9.856  < 2e-16 ***
poly(cars_per_1000_persons, 2)2 -1.078310   0.193191  -5.582 2.64e-08 ***
healthcare_advices_ratio_total   0.024238   0.002856   8.488  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.186 on 2459 degrees of freedom
Multiple R-squared:  0.6731,    Adjusted R-squared:  0.6709
F-statistic: 297.9 on 17 and 2459 DF,  p-value: < 2.2e-16
```

Figure 1: The estimated linear regression model with summary statistics and p values of statistical t tests on the coefficients

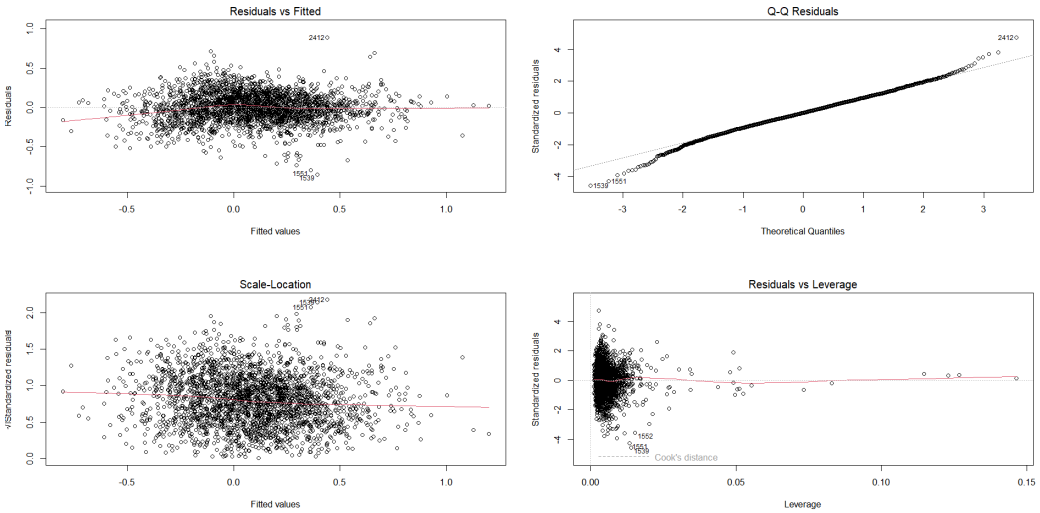

Figure 2: The diagnostics plot for the estimated linear regression model, indicating no significant violation of the model assumptions

	GVIF	Df	GVIF^(1/(2*Df))
factor(partitions)	3.356628	2	1.353555
factor(type_of_municipality)	2.619614	2	1.272212
sqrt(SLD_percent)	2.867842	1	1.693470
PO_percent	3.831755	1	1.957487
frekwencja_wyborcza	2.733252	1	1.653255
PSL_percent	1.472092	1	1.213298
Konfederacja_percent	1.387384	1	1.177873
revenues_per_capita_PIT_log	3.973669	1	1.993406
percent_over_60	1.870610	1	1.367702
sqrt(min_dist)	1.609691	1	1.268736
no_loc_in10km	1.264410	1	1.124460
education_share_higher	1.929025	1	1.388893
poly(cars_per_1000_persons, 2)	1.563637	2	1.118237
healthcare_advices_ratio_total	1.443463	1	1.201442

Figure 3: The Variance Inflation factor for included variables in the linear model

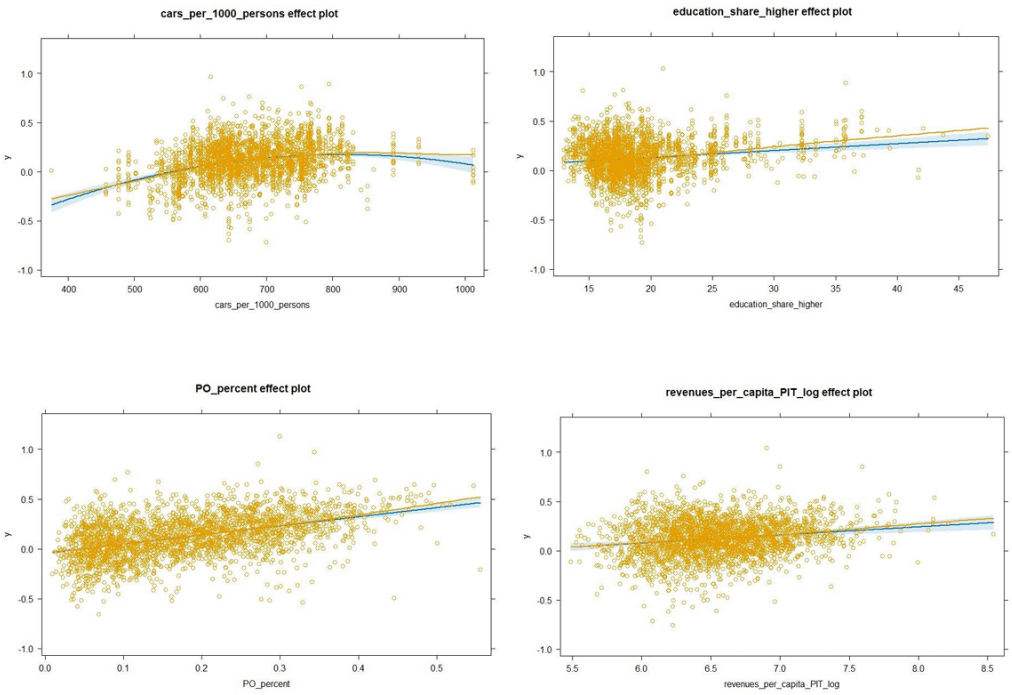

Figure 4: Partial residual plots between the selected independent variables and the partial residuals, which indicated the validity of the function form of selected independent variables in the regression model

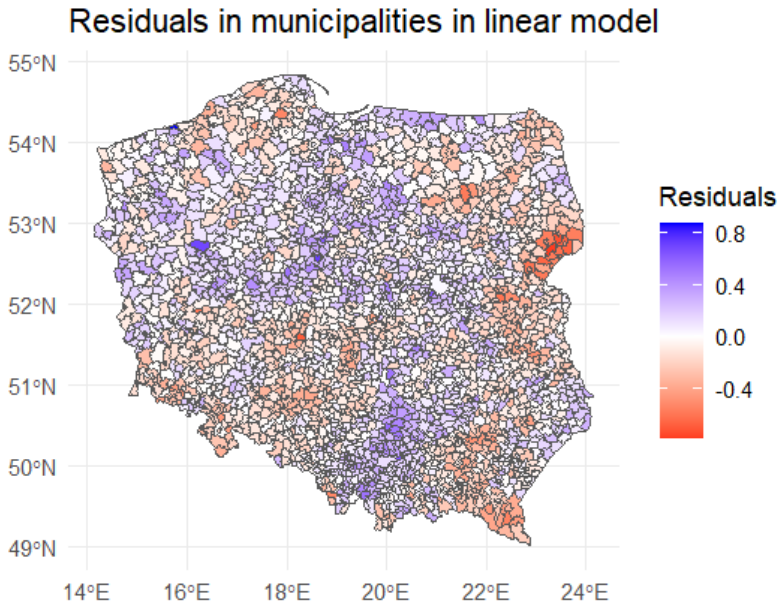

Figure 5: Residuals of the linear model across space, clear indication of spatial dependence structure are present

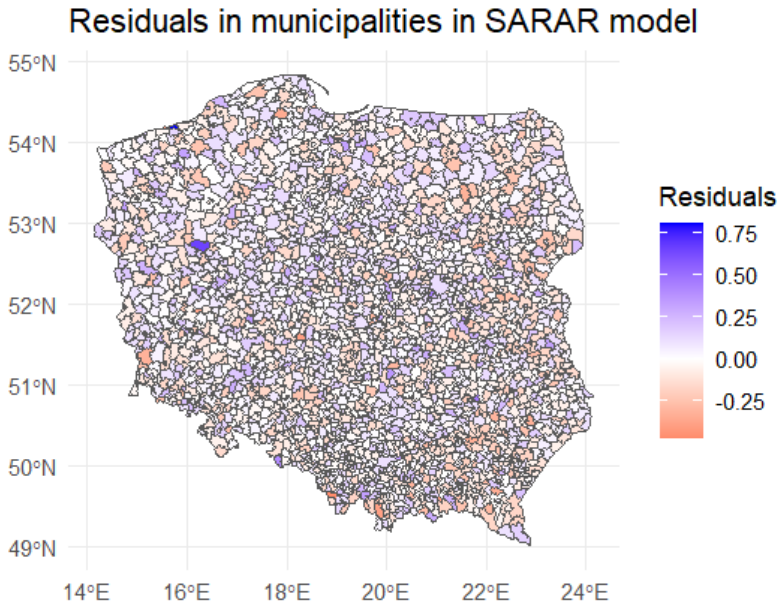

Figure 6: Residuals of the spatial linear model across space, indicating that no spatial dependence structure

6.2 Econometric approach with spatial dependence

6.2.1 Approach

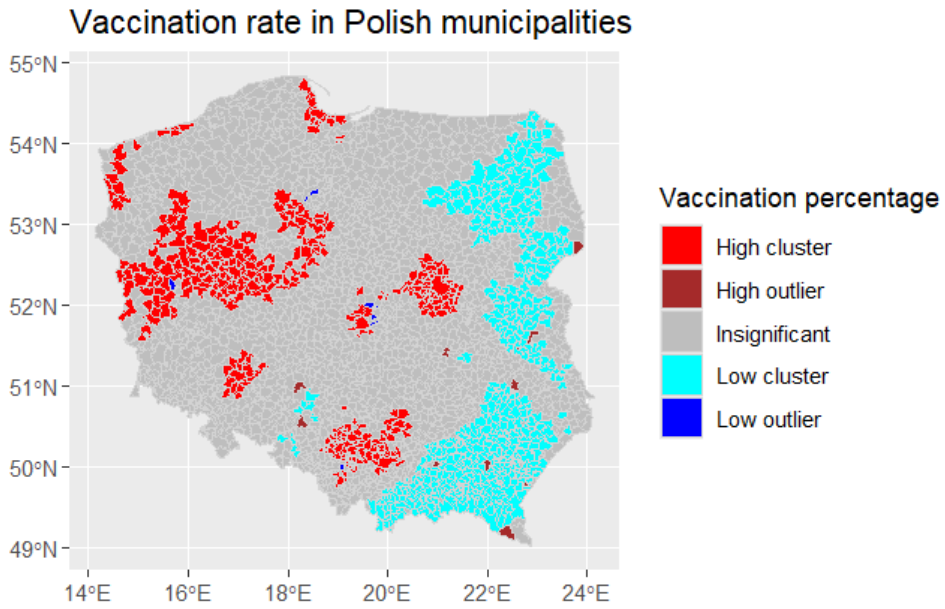

Figure 7: The diagnostics plot for the estimated linear regression model, indicating no significant violation of the model assumptions

SAC/SARAR models are used in cases where the spatial dependence of dependent variable is significant. Based on previous models it is a sensible hypothesis. First, we applied LISA method to visually analyze map of dependence 7. In order to formally check the presence of spatial dependence and thus proving our approach reasonable we performed Moran's I test. We obtained Moran I statistic 0.71, which was enough to reject the hypothesis of independence with p-value $<$ 2e-16. We decided to incorporate spatial component into our analysys using Spatial Autoregressive with Autoregressive Conditional Heteroskedasticity (SARAR) model. [3]

6.2.2 Model formulation

Model is given by following equation:

$$y = \rho_{Lag} W y + X\beta + u, \quad u = \lambda_{Err} W u + \epsilon$$

Where:

y - dependent variable percent vaccined

ρ_{Lag} - spatial autoregressive coefficient

W - matrix of weights

λ_{Err} - spatial moving average coefficient

$\epsilon \sim N(0, \sigma^2)$

Parameters ρ_{Lag}, λ_{Err} and β are estimated using MLE and $W[i, j]$ is reversed distance between municipalities i and j for $i \neq j$ and $W[i, i] = 0$. Distance is calculated using Hervesine formula for exact real-life distances between municipalities. We then proceeded to perform another Moran I test for residuals obtaining Moran I statistic $= -0.05$ and $p - value = 0.9999$. Therefore there is no reason to assume that residuals in our model are spatially dependent, consequently proving that our model was properly chosen.

6.2.3 Results

The conditional component accounts for additional predictors that may influence vaccination percent, such as population demographics, healthcare infrastructure, and socioeconomic factors. The error component captures unobserved factors and measurement error that contribute to variability in vaccination percent. Spatial autocorrelation in the errors is accounted for, indicating that errors in vaccination percent for neighboring municipalities may be correlated. The SAC/SARAR model allows for the interpretation of coefficients representing the direct effects of predictors on vaccination percent, excluding spatial effects. These were covered by the model and are not present in coefficients due to model architecture. This was confirmed by the model that included factors of partitions (relating to geopolitical partitions of Poland during XIX century) and municipality factor (rural, urban and mixed). All of them turn out to be insignificant, which is astounding result, once again proving that our approach is well-chosen.

Finally we prepared map of residuals similar to previous models 6 that is clearly interpretable. Residuals are actually random in space. Residual in linear model map have clear dependence, see 5 for comparison.

```
Coefficients: (asymptotic standard errors)
                                    Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)                      -1.15503297  0.10911773 -10.5852 < 2.2e-16
PO_percent                       -0.91013017  0.25890927  -3.5152 0.0004393
sqrt(SLD_percent)                 0.63334277  0.07153444   8.8537 < 2.2e-16
frekwencja_wyborcza               0.38771081  0.11952885   3.2437 0.0011800
PSL_percent                       0.56884887  0.06752224   8.4246 < 2.2e-16
log(revenues_per_capita_PIT)      0.05828763  0.01270469   4.5879 4.478e-06
poly(percent_over_60, 2)1         2.73715615  0.19323698  14.1648 < 2.2e-16
poly(percent_over_60, 2)2        -0.65295395  0.13701464  -4.7656 1.883e-06
healthcare_advices_ratio_total    0.01197998  0.00169683   7.0602 1.663e-12
education_share_higher_neighbours 0.00941282  0.00242605   3.8799 0.0001045
cars_per_1_persons                0.16912663  0.05972142   2.8319 0.0046269
min_dist                          0.00141591  0.00054875   2.5802 0.0098734
population_density_log            0.00250469  0.00400191   0.6259 0.5313983
no_loc_in10km                     0.00043792  0.00081904   0.5347 0.5928722
PO_percent:frekwencja_wyborcza    2.76340804  0.43455456   6.3592 2.028e-10

Rho: -0.3377
Asymptotic standard error: 0.037602
    z-value: -8.9808, p-value: < 2.22e-16
Lambda: 0.91842
Asymptotic standard error: 0.010728
    z-value: 85.613, p-value: < 2.22e-16

LR test value: 1696.2, p-value: < 2.22e-16

Log likelihood: 1508.757 for sac model
ML residual variance (sigma squared): 0.013342, (sigma: 0.11551)
Number of observations: 2477
Number of parameters estimated: 20
AIC: -2977.5, (AIC for lm: -1285.3)
```

Figure 8: The estimated linear regression with spatial component model with summary statistics and p values of statistical t tests on the coefficients.

7 Machine Learning Approach

7.1 Motivation and approach

We have trained a machine learning model to see if potential complex non-linear dependencies change the impact of predictions on our target variable. We have used XGBoost model, because it proves to be effective solution for tabular data. [4] We have trained it on the same data that we used with linear model, however we did not apply any transformations to it (apart from one-hot-encoding of categorical variables), because we wanted the model to figure it out by itself. We do not compare performance of ML model with linear model, because we do this analysis only for explainability purposes.

Explanation of this model was done using SHAP analysis: SHapley Additive exPlanations. SHAP values are a method used in machine learning to explain the prediction of a model by quantifying the contribution of each feature to the prediction. They provide insights into how each feature affects the model's output, helping understanding model behavior and feature importance. [7]

SHAP analysis provides us with a SHAP value for each feature of each observation in our dataset. SHAP value measures how value of given feature in given observation "moves" our prediction away from baseline prediction (which is usually average of all predictions in our dataset). This can be represented on a "beeswarm plot", where SHAP values are represented on the x-axis, while the y-axis represents the features. Each observation in the dataset is then represented as a point on the plot, with the color of the points representing the value of the corresponding feature. This visualization helps to understand how the SHAP values vary across different features for each observation in the dataset.

7.2 Results

Beeswarm plot of our XGBoost model is presented in Figure 9 below.

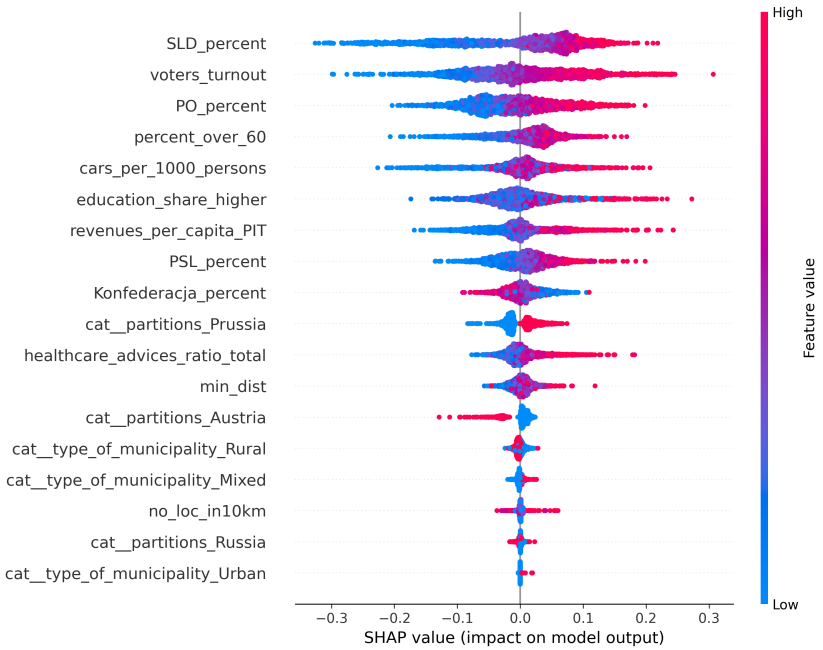

Figure 9: Beeswar plot of SHAP analysis for XGBoost model.

As we can see, explain ability results from XGBoost model are similar to ones we have observed in the linear model. The beeswarm plot analysis reveals a consistent alignment between the interpretability results obtained from the XGBoost model and the observations derived from the linear model. Notably, key factors driving vaccination rates in Polish municipalities predominantly stem from political and educational variables.

Leading the pack are variables associated with political affiliations and participation, such as voters turnout, and percentages of votes for parties indicating their substantial influence on vaccination outcomes. We can also see that political affiliations are in line with expectations: votes for parties that are considered progressive (SLD and PO) have positive impact on vaccination rates, while for conservative party (Konfederacja) has negative impact. PSL party, generally conservative (but as of 2024 in coalition with SLD and PO) also has positive impact on vaccination rates.

Interestingly, we can see that the relationship between each feature and its SHAP value appears more condensed compared to the typical scatterplot distribution observed between features and predictions. Instead of a scattered cloud of points, the SHAP values exhibit a more concentrated pattern

resembling a function plot, where for a given X, the associated SHAP values cluster closely around a singular Y value, which is especially visible for political parties as seen in the Figure 10 below.

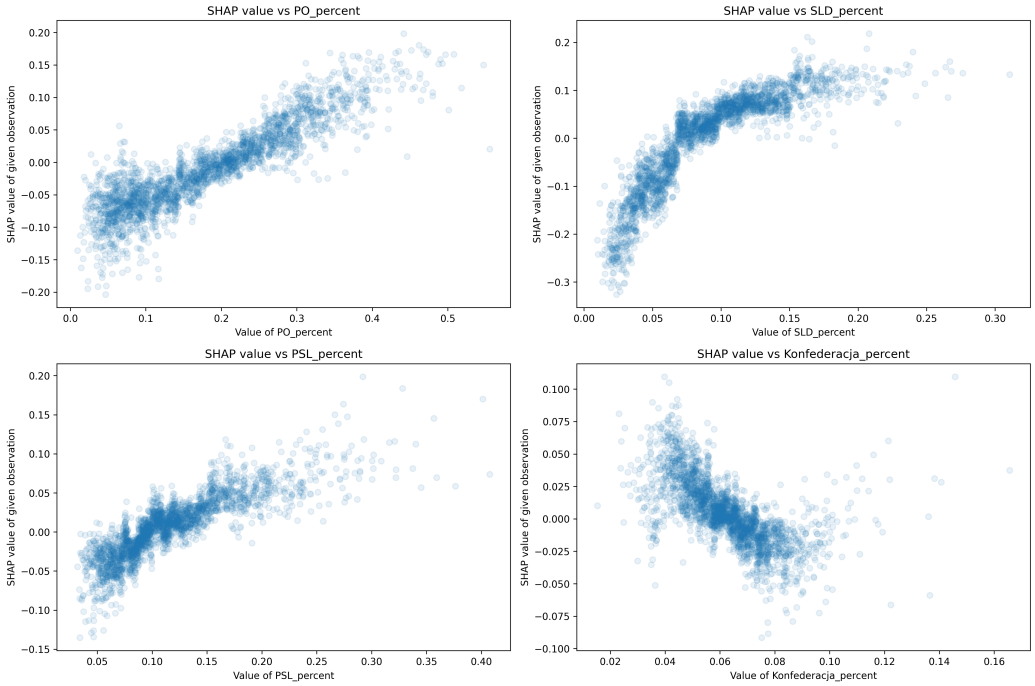

Figure 10: Political party vote percentage vs SHAP value for that feature.

Demographic characteristics also play a discernible role, with variables like percentage of people over 60 and cars per 1000 persons highlighting the influence of age distribution and accessibility on vaccination uptake. However, certain features exhibit relatively lower importance, such as type of municipality, suggesting their minor impact on vaccination rates within the studied context.

We can see that historical partitions of Poland have generally have impact in line with our expectations (negative for Austrian and positive on Prussian), however that impact is low. We believe that this is caused by high correlation between historical partitions and other variables that are already included in the model.

The consistent alignment between the XGBoost and linear model results validates the pivotal role of political and educational variables in shaping vaccination rates across Polish municipalities. These findings affirm the reliability of our analytical framework and provide further evidence supporting the effectiveness of our approach in capturing the nuances of vaccination

dynamics.

8 Conclusions

H1. The findings from our study support the validity of H1. Voter turnout, both independently and in interaction with the percentage of votes received by the political party (PO) in the municipality, demonstrates a significant and positive relationship.

H2. H2 is further supported by our study, as the coefficients of features associated with the percentages of votes received by a specific party, such as SLD, are found to correlate with vaccination rates. Although some of these features were highly correlated with each other, preventing their simultaneous inclusion in the model, our model preparation process revealed the diverse directions of these relationships.

H3. Our research suggests the validity of Hypothesis 3. The coefficient of the feature 'percent of inhabitants over 60 years old' is significant and positive in our model. However, a partial residual plot indicates a 'bend' or 'inflection point,' suggesting a non-linear relationship. Additionally, the square of this factor included in the model is also significant and has a negative coefficient, which indicates that the impact of an elderly demographic of a municipality population is more complex. This complexity may be related to other factors such as political views and requires further study.

H4. The situation with H4 mirrors that of H3. We observe a significant positive coefficient for the feature 'number of cars per 1000 inhabitants,' but the partial residual plot indicates a more nuanced relationship. It may be beneficial to include additional details about the vehicles to better understand this complexity and resolve any potential discrepancies.

H5. Our research findings do not support H5. Contrary to expectations, our model does not demonstrate a positive correlation between the initial intensity of vaccination efforts and the final vaccination rate in 2021. This suggests that the distribution of vaccination points in March, when only certain segments of society were eligible for vaccination, may not accurately reflect their distribution during the peak of the vaccination campaign.

H6. H6 has not been directly supported in our research, as the percentage of higher educated residents of the municipality was found not to be a statistically significant feature in the model. However, the mean percentage in the municipalities located around the analyzed one showed a statistically significant, slightly positive correlation with the vaccination rate.

H7 & H8. The detailed hypothesis within the complex H7 concerning income per capita is supported by our findings. The logarithm of this feature exhibits a statistically significant and positive coefficient in the final model. While other features may not appear to directly impact the vaccination rate, it is noteworthy that they are all related to the municipality's location, and thus, were factored into the spatial component of the model. Even after controlling for all other significant variables, the localization of the municipality, understood as the vaccination acceptance level of its neighbors, remained a significant factor influencing the vaccination rate in this municipality.

9 Discussion

The dynamic nature of the pandemic necessitates the inclusion of a temporal component in our analysis. This would allow us to assess whether the coefficients in the model remain consistent over time or if specific events, such as governmental interventions, impact people's motivations and the significance of factors. One limitation of our study is that we solely examined the overall vaccination rate in 2021.

References

[1] Covid-19 vaccine tracker, 2024. Accessed: 2024-05-11.

[2] Cornelia Betsch, Philipp Schmid, Dorothee Heinemeier, Lars Korn, Cindy Holtmann, and Robert Böhm. Beyond confidence: Development of a measure assessing the 5c psychological antecedents of vaccination. *PLOS ONE*, 13(12):1–32, 12 2018.

[3] Roger S. Bivand, Edzer Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R.* Springer, UseR! Series, 2nd edition, 2013. Softcover.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[5] A. J. Dolman, T. Fraser, C. Panagopoulos, D. P. Aldrich, and D. Kim. Opposing views: associations of political polarization, political party

affiliation, and social trust with covid-19 vaccination intent and receipt. *J Public Health (Oxf)*, 45(1):36–39, 2023.

[6] Justyna Gołębiowska, Anna Zimny-Zając, Mateusz Dróżdż, Sebastian Makuch, Krzysztof Dudek, Grzegorz Mazur, and Siddarth Agrawal. Evaluation of the approach towards vaccination against covid-19 among the polish population—in relation to sociodemographic factors and physical and mental health. *Vaccines*, 11(3), 2023.

[7] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*, 2017.

[8] Shingai Machingaidze and Charles Shey Wiysonge. Understanding covid-19 vaccine hesitancy. *Nature Medicine*, 27(8):1338–1339, Aug 2021.

[9] National Science Center. Vaccination points data. `https://polon.nauka.gov.pl/pomoc/wp-content/uploads/2021/03/Punkty_szczepien_09_03_2021_publikacja_13.20.xlsx`, 2021. Accessed: 2024-05-11.

[10] Caitlin Rancher, Angela D. Moreland, Daniel W. Smith, Vickey Cornelison, Michael G. Schmidt, John Boyle, James Dayton, and Dean G. Kilpatrick. Using the 5c model to understand covid-19 vaccine hesitancy across a national and south carolina sample. *Journal of Psychiatric Research*, 160:180–186, 2023.

[11] Marcin Piotr Walkowiak, Jan Domaradzki, and Dariusz Walkowiak. Better late than never: Predictors of delayed covid-19 vaccine uptake in poland. *Vaccines*, 10(4), 2022.

[12] S.N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2nd edition, 2017.