Spatial statistics and machine learning
term project

MSc Applied Data Science

Spatial variability of Municipal Council election turnout in the Netherlands

Authors:
Max van den Elsen
2590611

Thom Venema
1157485

Sander Engelberts
1422138

April 14, 2022

Abstract

With decreasing municipal council election turnout over the past decades, it is important to research what factors are related to turnout percentages and how this varies for different neighbourhoods in the Netherlands. Global, semi-local, and local models have been applied, to account for spatial autocorrelation and spatial heterogeneity. A combination of socio-economic, socio-demographic, and location variables was most effective, but still only 30% of the variance could be explained by the models. Most efforts of policy makers should go to attracting lower income households and decreasing the distance to polling stations, as these were shown to have highest influence on the turnout percentages.

1 Introduction

Municipal Council election turnout has been decreasing since the eighties [1]. This year, a historical low turnout was reached [2]. Out of all the Dutch citizens eligible to vote, only 50,3% casted their votes. The election turnouts significantly differ per city (Rotterdam: 39%; Rozendaal: 39%) [2]. Therefore, The objective of this research is to identify external factors that could have influence on the election turnout based on open source data[1].

Voter turnout has been studied in several regions, with different resolution, models, and variables, as can be seen in Table 1. Eskov (2013) showed with exploratory analysis that aggregation of polling station data to different resolutions in Canada all exhibit spatial patterns [3]. We are not aware of Dutch elections being studied like Eskovs (2013) method.

As shown in Table 1, there is a complex spatial interaction between socio-demographic and socio-economic factors and election turnout. For several countries it has been shown that varying degrees of different factors[2] relate to turnout due to inequalities and differences in political participation [5, 9, 10, 11]. These influences can be explained by the Rational Choice Theory [12] and Civic Voluntarism [13] models, which respectively relate electoral behaviour to individual utility and to socio-economic resources that are available to someone [5, 10]. Specifically in the Netherlands, a higher age, education level, and income are shown to be related to higher turnout, and a higher population density, and number of parties to a lower turnout [14].

[1] The code, data and full set of results of this study can be found on the project GitHub.

[2] E.g. level of education [4], gender [5], age [6], ethnicity [5], home ownership [7], marriage [7], population density [8], and economic factors like unemployment rate [9].

Article	Fiorino, Pontarollo & Ricciuti (2021) [9]	Mansley & Demšar (2015) [10]	Manoel, Costa & Cabral (2022) [11]
Research	Investigating turnout in European Parliamentary elections in 155 EU-12 regions. Significant effects are found for GDP per capita, unemployment, age, and institutional and electoral variables that determine how the government and voting types vary over countries. Heterogeneity needs to be accounted for still, but spill-over effects from neighbouring countries are discovered.	Analysing the geographic socio-demographic and socio-economic variability on voter turnout at the 2012 London mayoral election at a detailed spatial level (625 wards). Localised patterns have been found where in certain regions a variable had more negative/ positive effects on the turnout, especially when comparing the city centre and peripheral boroughs.	Researching socio-demographic and socio-economic variables, but also density of residential buildings, that could explain voter turnout in Portugal at a municipal level. Some of these influence turnout differently over the country while others are constant in space.
Model	Spatial lag model after verifying the existence of spatial autocorrelation with ordinary least squares model.	Geographically weighted regression to look into local variations while accounting for spatial autocorrelation and heterogeneity. A global model has been checked to perform worse than this local one.	Semiparametric geographically weighted regression with global and local coefficients got compared with standard geographically weighted regression and ordinary least squares models. The first has less complexity than the second one.

Article	Fiorino, Pontarollo & Ricciuti (2021) [9]	Mansley & Demšar (2015) [10]	Manoel, Costa & Cabral (2022) [11]
Data	Turnout, GDP per capita, unemployment, population density, education, impartiality of government, if election is during weekday.	Turnout, age, percentage of students and professional/ managerial occupations, percentage newly living in London, population density, unemployment, home ownership, difference between winner and second place in previous election.	Turnout, percentage of house owners, families with young children, education, distance to big cities, distance to city centre, density of residential homes, gender, age.

Table 1: Overview of research about spatial effects on election turnouts.

2 Study area

This research has been carried out for the Municipal Council elections of 2022 in The Netherlands. This was done on the detailed neighborhood level. Important to note about these elections is that all residents who are over 18 years old, including ones without a Dutch nationality, can vote within the municipality borders they live [1]. Further, local independent parties are significantly represented in election outcomes next to local branches of national parties [1].

3 Data and methods

3.1 Data sources

Firstly, a governement dataset containing highly detailed voting results of each of the polling stations (n=*9241*) was used [15], as can be seen in the flowchart of Figure 1. Its data is however highly inconsistent, with in many cases absent postal codes, unstructured naming, duplicates and inconsistent counting. The second dataset contains the coordinates of the polling stations [16]. The two datasets were matched which resulted in a dataset (n=*7850*) containing both turnout and coordinates. Data loss occurred due to inability to match all stations. The data were then aggregated to neighborhood level, and appended with information on a broad range of population statistics from CBS[3] [17].

[3]Dutch bureau for statistics

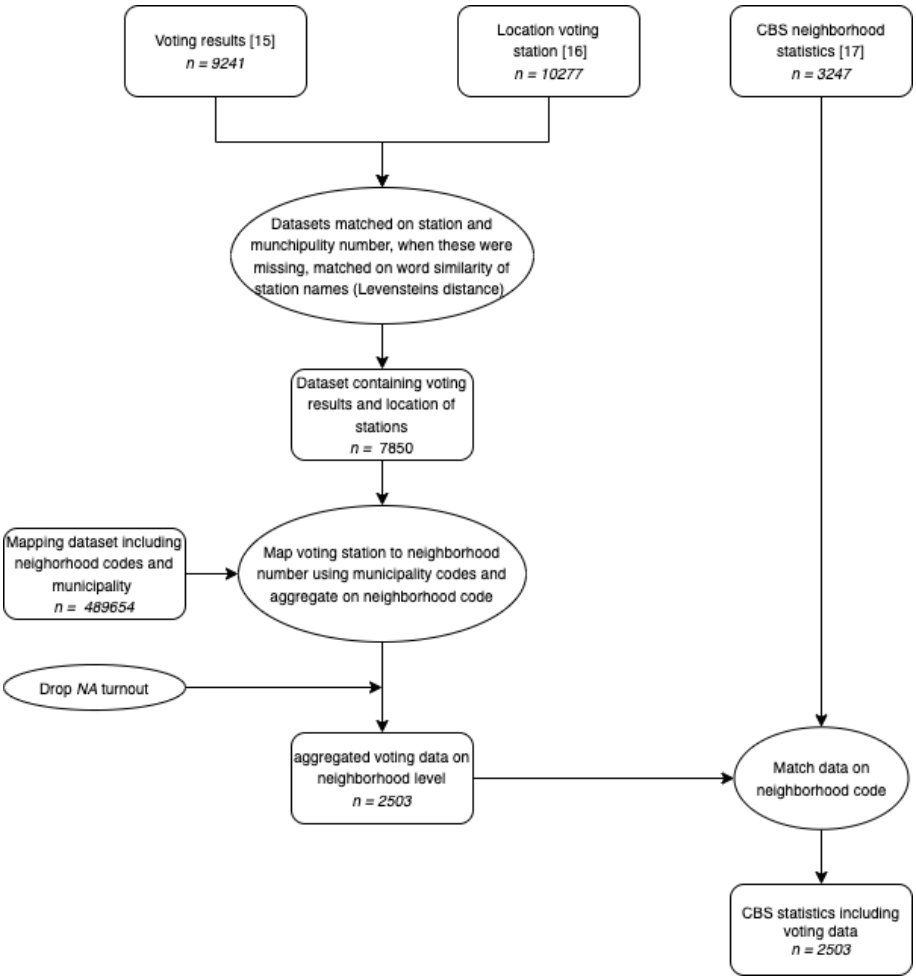
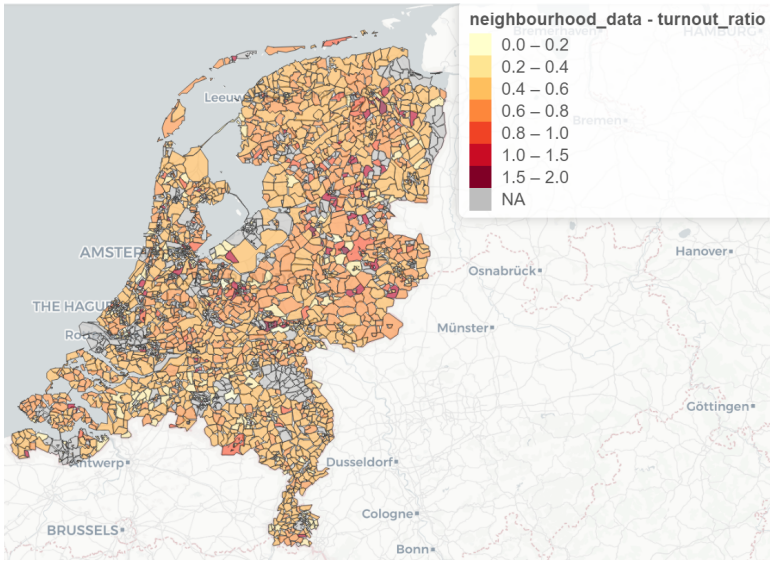


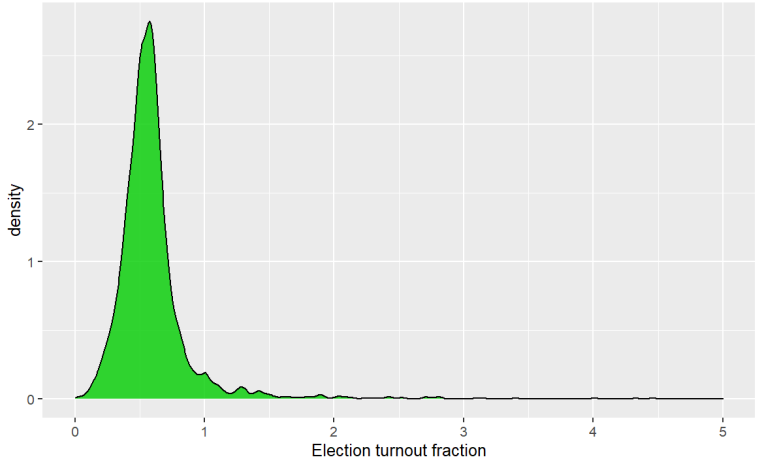
Figure 1: Overview of the pre-processing and matching of datasets. This resulted in the neighbourhood dataset used for training models.

Because of the inconsistency in the data of the polling stations, the assumptions were made that every citizen above 18 years within a certain neighborhood is eligible to vote [18], and that the age groups have a uniform distribution. For computing the turnout fraction, CBS data [17] was used to divide the total turnout value per neighbourhood by its number of adults. This distribution of turnout can be seen in Figure 2.

For more robust training, outliers were given a maximum value for each variable (e.g. for turnout this was set to 100%), and these were z-normalized and scaled [20]. Also, neighbourhoods with *NA* values in any variable were removed because imputing these would give additional assumptions that can bias the results. This resulted in 2503 of 3247 neighbourhoods.



(a) Spatial variation of election turnout fraction.



(b) Density plot of election turnout fraction.

Figure 2: Map and density plot of the dependent variable election turnout as a fraction of people who are allowed to vote in each neighbourhood in the Netherlands. To note, neighbourhoods with *NA* values mainly did not have municipal elections in 2022 [19], or there was no polling station in that area. Here can be seen that there is a spatially varying pattern of turnout, mostly between 40% and 80%. However, also large outliers exist where more votes are cast in a neighbourhood than elegible voters live. This may be due to adjacent neighbourhoods not having polling stations, as visual inspection hypothesized, or the convenience of voting near work, shops, or stations.

3.2 Methods

To check for presence of spatial autocorrelation in the data, Moran's I metric with Monte Carlo approach, in combination with different types of weight matrices, was used on a linear regression model. Additionally, spatial lag and error models were trained as a semi-local approach. An inverse distance weight matrix was used to relax the assumption of the error-terms being independent.

To also be able to explain local variations with maps, a GWR[4] model was fit [21, 22]. An adaptive kernel was utilized with a Gaussian and bisquare distribution to weight the N closest polling stations based on their distance from a regression point. Furthermore, cross-validation was used to find the optimal distribution function and neighborhood density for the kernel [21, 22]. One of the pitfalls of GWR is the effect of multicollinearity [23]. The global multicollinearity can be distinguished by calculating a variables' variance inflation factor (VIF) [24]. The VIFs were repeatedly calculated, each time dropping a variable having a score above 5, until all VIFs were below this threshold[5]. Table 2 displays the final variables used, as well as their VIF values.

Additionally, coefficients were mapped to explore the spatial non-stationarity of predictors. Additionally, its local significance was checked with a t-test, and the local explained variance R^2 by the model was plotted [22]. This way conclusions could be drawn about the effects of the independent variables on the election turnout for different areas.

Moreover, a random forest regression was fit on the data ($n\_tree = 2000$), which allows for inspection of the feature importance [25]. Using spatial cross-validation and grid search, the optimal hyperparameter values were selected for the number of predictors to consider and the tree complexity [25].

All models were trained for different combinations of predictor variables: *socio-economic, socio-demograhic, and spatial related predictors.*

[4] Geographically weighted regression

[5] The variables that were dropped during the procedure are: percentage of population married/unmarried/widowed, percentage of people in high income class, average income, social welfare benefits, percentage 65+ years old, and lastly the distance to the nearest supermarket.

[6] The selected socio-economic variables were: % Low income, No. income earners, and Avg. house value; socio-demographic variables were: % Western Immigrant, % Non-Western Immigrant, % Divorced, % Female, % 0-15 y/o, % 15-25, % 25-45, % 45-65; spatial location variables were: Distance to voting station (meters), High School within 3km (meters), No. Companies, Pop. Density per km^2.

Variable	VIF score	Variable	VIF score
Pop. Density per km^2	3.185994	% Western Immigrant	3.342604
% Female	1.889696	No. Companies	1.328345
% 0-15 y/o	2.120951	Avg. house value	2.179000
% 15-25 y/o	1.900657	No. income earners	1.562464
% 25-45 y/o	4.756518	% Low income	2.230633
% 45-65 y/o	3.216237	High School within 3km	3.274812
% Non-Western Immigrant	1.756704	Distance to voting station	1.205926
% Divorced	2.625683		

Table 2: Overview of used variables and their VIF-scores. A higher VIF score indicates a higher multicollinearity with another variable. These final selected predictor variables were all used in the models without stratification[6].

4 Results

The Moran's I MC simulation resulted in significant tests, as can be seen in Table 3, which rejects the *0-hypothesis* of autocorrelation not being present in the data. The spatial pattern of residuals can also be inspected in Figure 3.

GWR models deal with this spatial auto-correlation, of which their optimal parameters and model statistics are shown in Table 4. Also, a map with local R^2 distribution can be seen in Figure 4. Moreover, maps were created with coefficient values and its significance distribution, for which some can be seen in Figure 5. For the other predictors and model variants, their significance and medium coefficient values are mentioned in Table 5.

Spatial cross-validation used for grid search resulted in the optimal parameter values and model performance for the random forest models as can be seen in Table 6. Furthermore, the feature importance was calculated which is shown in Figure 6.

Model	**Stratum**	**Moran's I MC sim. (p-value)**	**AIC**
Linear Model	Demographic	0.00099	-1331
Spatial Error Model	Demographic	0.1508	-1914
Spatial Lag Model	Demographic	0.00099	-1830
Linear Model	Economic	0.00099	-1209
Spatial Error Model	Economic	0.2777	-1849
Spatial Lag Model	Economic	0.00099	-1765
Linear Model	Location	0.00099	-1168
Spatial Error Model	Location	0.05994	-1669
Spatial Lag Model	Location	0.00099	-1643
Linear Model	None	0.00099	-1567
Spatial Error Model	None	0.6314	-2133
Spatial Lag Model	None	0.07293	-2124

Table 3: Overview of results of linear regression, spatial error, and spatial lag models. A significant Monte-Carlo simulation of Moran's I p-value rejects the *0-hypothesis* of autocorrelation not being present in the data. This is only not the case for the spatial error models (with $p \leq 0.05$). The model with minimum AIC value is preferred, which is for each stratum also the spatial error model, and the least preferred are the linear regression models.

Stratum	Kernel function	Density parameter	R^2	RSME	AIC
Demographic	Bisquare	0.99	0.17	70.62	-1810
Economic	Bisquare	0.99	0.1	76.03	-1635
Location	Bisquare	0.99	0.08	78.35	-1558
None	Bisquare	0.99	0.27	62.13	-2118
None	Gaussian	0.98	0.26	62.82	-2101

Table 4: Overview of statistics of the GWR models. Here almost all neighbourhoods are considered by the adaptive kernels. With respect to the statistics, R^2 should be maximized, RMSE minimized, and AIC also minimized. The best model is thus the GWR without stratification on variables and with a bisquare kernel shape, closely followed by the similar model with a gaussian kernel shape.

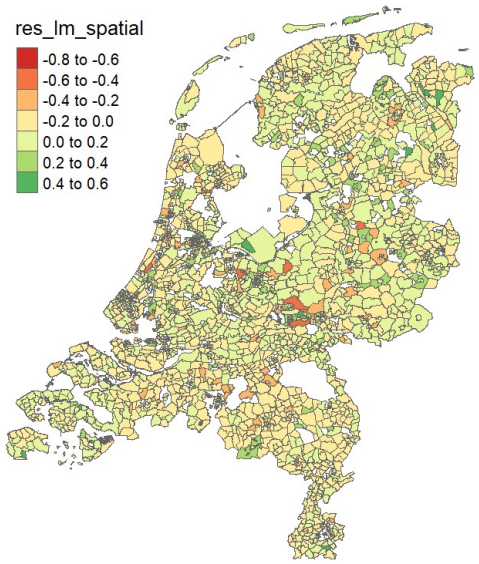


Figure 3: Residuals map of the linear regression model trained without variable stratification, with coefficient tuning using spatial cross validation, and using an inverse distance weight matrix. Here can be visually seen that the residuals do not have a constant value over space, with some neighbourhoods having large negative or positive values.

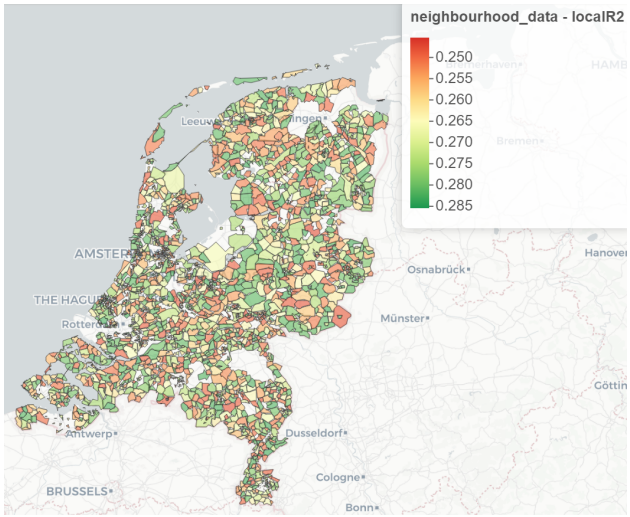
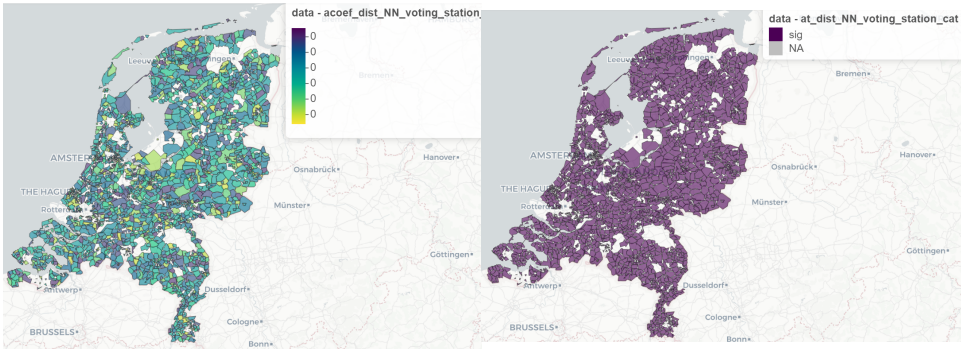
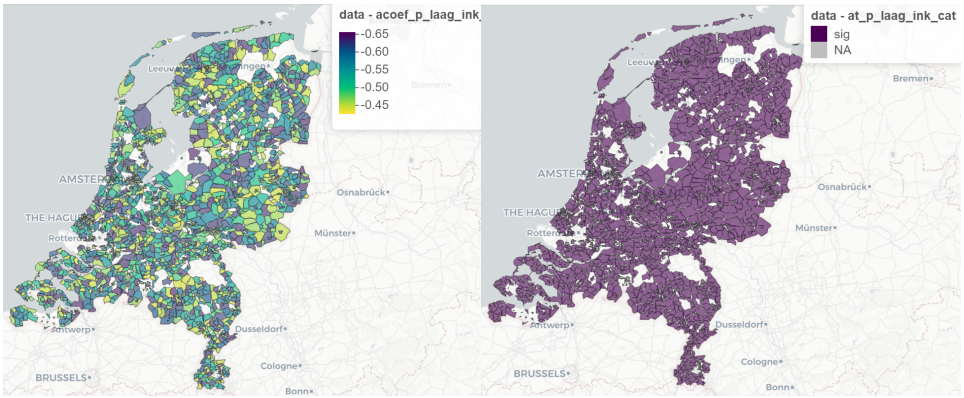


Figure 4: Local R^2 map of the GWR model trained without variable stratification and with a bisquare kernel shape. A range of values between 0.250 and 0.285 can be seen, which vary over space but often adjacent neighbourhourhoods have similar values.



(a) Coefficient map of distance to nearest voting station.

(b) Significance map of distance to nearest voting station.



(c) Coefficient map of percentage of residents with a low income.

(d) Significance map of percentage of residents with a low income.

Figure 5: Coefficient and significance of variable distribution in the GWR model trained without stratification and with a bisquare kernel shape. This is shown for the two most important features from the random forest model (see Figure 6) that have a significant relation with turnout in most neighbourhoods, as can be seen in their significance maps. The coefficient maps show variation in values over space, with adjacent neighbourhoods often containing relatively similar values.

Variable	All; Gaussian	All; Bi-square	Demographic; Bi-square	Economic; Bi-square	Location; Bi-square
Pop. Density per km^2	-1.57 E-6	-1.45 E-6	-	-	-2.34 E-5
% Female	-4.52 E-1	-4.28 E-1	-1.48 E-1	-	-
% 0-15 y/o	-5.22 E-1	-5.24 E-1	-4.23 E-1	-	-
% 15-25 y/o	5.67 E-1	5.67 E-1	3.12 E-2	-	-
% 25-45 y/o	-3.74 E-1	-3.76 E-1	-3.46 E-1	-	-
% 45-65 y/o	-1.10	-1.09	-6.51 E-1	-	-
% Divorced	4.03 E-1	3.85 E-1	-1.25	-	-
% Non-Western Immigrant	-4.81 E-1	-4.83 E-1	-8.15 E-1	-	-
% Western Immigrant	-9.56 E-1	-9.54 E-1	-7.26 E-2	-	-
No. Companies	-3.89 E-5	-3.85 E-5	-	-	-3.37 E-5
Avg. house value	5.31 E-4	5.30 E-4	-	6.38 E-4	-
No. income earners	3.96 E-2	3.74 E-2	-	2.89 E-1	-
% Low income	-5.45 E-1	-5.47 E-1	-	-2.29 E-1	-
High School within 3km	5.37 E-4	5.18 E-4	-	-	1.67 E-3
Distance to voting station	-3.95 E-5	-3.95 E-5	-	-	-3.38 E-5

Table 5: Overview of the median estimates of the GWR-models per different stratum of variables, and different kernel shapes. The estimates have been tested for significance and plotted on the maps. If an estimate is coloured red, the number is not significant, a green number shows a significant estimate and an orange estimate varies in significance over space. As can be seen, the effect of following variables becomes insignificant when taking all predictors into account instead of only socio-economic, socio-demographic or location variables: population density, % divorced, number of income earners, and distance to highschool. In contrast, % non-western immigrants becomes significant, and the following ones significant in a part of the previously insignificant neighbourhoods: % female, % 15-25 y/o, and % low income. Most effects of variables on the voting turnout are negative, with only some being positively related and significant.

Cross validation	Stratum	mtry	min_n	R^2	MSE
Spatial	Demographic	3	100	0.25	0.025
Spatial	Economic	3	100	0.14	0.029
Spatial	Location	3	100	0.14	0.029
Spatial	None	6	100	0.31	0.023
Stratified on turnout	None	9	100	0.31	0.023

Table 6: Overview of results of Random Forest models. Interesting to note is that the results are equal of the random forest models without stratum, which were trained on- and had parameter selection with spatial cross validation and cross validation using stratification on the turnout variable. These models also have the best performance, with maximum R^2 and minimum MSE values. Their number of predictors to consider at each split (mtry) is lower for the model with spatial cross validation, which is thus preferred due to lower required computational power. Lastly, the minimum number of observations in leaf nodes is the same for each model variant.

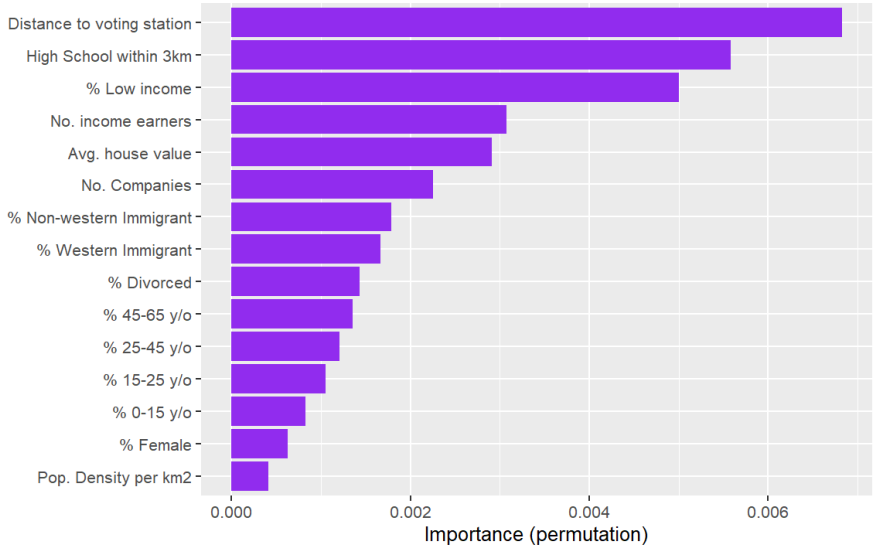


Figure 6: Feature importance based on permutation for the random forest model without stratification and trained with spatial cross validation. Location and socio-economic variables can be seen as the most important variables, with an exception of the population density. Further, socio-demographic variables are deemed less important by the model.

5 Discussion

As mentioned in 3.1, some neighbourhoods contained missing values, which may have influenced the results in adjacent neighbourhoods. Another interesting insight is that some neighbourhoods had more votes than residents, which is likely due to citizens being allowed to vote everywhere within their municipality so doing that near home, train station, or work. Indeed, distance to polling station and number of companies showed a significant effect.

Furthermore, multicollinearity was handled globally using the VIF method. However, local multicollinearity may be a problem. This is recommended to take into account by applying PCA in future studies [24].

Moreover, for each model, the variant trained on all variables performed best, followed by the ones based on only socio-demographic factors. Still, the latter are less valued by the random forest. Also, lag models didn't properly account for spatial autocorrelation. This is not solely due to the turnout variable itself. Further, GWR models do not improve upon the spatial error models. The random forest models do however explain the turnout variance best (31%). This indicates that voter turnout is a complex system that is difficult to accurately predict on a neighbourhood level in the Netherlands. It would be interesting to research how the models perform with different resolutions, applying SGWR [11], and include additional (closed-source) variables.

Furthermore, in the best performing GWR, only the average house value had a significant positive effect on the turnout in (almost) all neighbourhoods, partly in % 15-25 y/o, and the other effects being negative or insignificant. This first effect is in line with the study of CBS [14], where house price is related with income and in turn with education level. The second effect is contradicting their global relation of higher age being related with higher turnout [14].

Moreover, the effect of population density is not significant and no similar coefficient clusters around cities were found, like in Fiorino, Pontarollo & Riciuti (2021) [9], so no direct difference between rural and urban areas could be distinguished.

Lastly, for policy-makers it is important to stimulate people with low income and ones with an immigration background to vote, as their coefficients were highly negative.

6 Conclusion

To conclude, Dutch municipal council 2022 election turnout behaviour is a complex system for which this study could not well describe the variance on neighbourhood level. Still, this can best be done using a combination of socio-economic, socio-demographic and location predictors. As most impor-

tant factors to increase the voting turnout, the models showed that decreasing the distance to the nearest voting station, and focusing on low income and immigrant households may have a positive effect.

7 References

[1] Congress of Local and Regional Authorities, "Information report on municipal elections in the Netherlands (21 march 2018)," 09 2018. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016808e4a9d.

[2] Politieke redactie AD, "Historisch lage opkomst verkiezingen, minister onderzoekt oorzaak," 03 2022. https://www.ad.nl/politiek/historisch-lage-opkomst-verkiezingen-minister-onderzoekt-oorzaak a491fac6/?referrer=https%3A%2F%2Fwww.google.com%2F.

[3] A. Eskov, "Spatial patterns and irregularities of the electoral data: general elections in canada," 2013. https://run.unl.pt/bitstream/10362/11682/1/TGEO0122.pdf.

[4] A. Gallego, "Understanding unequal turnout: Education and voting in comparative perspective," *Electoral Studies*, vol. 29, no. 2, pp. 239–248, 2010.

[5] K. Smets and C. Ham, "The embarrassment of riches? a meta-analysis of individual-level research on voter turnout," *Electoral Studies*, vol. 32, p. 344–359, 06 2013.

[6] Y. Bhatti, K. M. Hansen, and H. Wass, "The relationship between age and turnout: A roller-coaster ride," *Electoral Studies*, vol. 31, no. 3, pp. 588–593, 2012.

[7] B. Highton and R. E. Wolfinger, "The first seven years of the political life cycle," *American Journal of Political Science*, vol. 45, no. 1, pp. 202–209, 2001.

[8] A. F. Tavares and R. Raudla, "Size, density and small scale elections: A multi-level analysis of voter turnout in sub-municipal governments," *Electoral Studies*, vol. 56, pp. 1–13, 2018.

[9] N. Fiorino, N. Pontarollo, and R. Ricciuti, "Spatial links in the analysis of voter turnout in european parliamentary elections," *Letters in Spatial and Resource Sciences*, vol. 14, 04 2021.

[10] E. Mansley and U. Demšar, "Space matters: Geographic variability of electoral turnout determinants in the 2012 london mayoral election," *Electoral Studies*, vol. 40, pp. 322–334, 2015.

[11] L. Manoel, A. C. Costa, and P. Cabral, "Voter turnout in portugal: A geographical perspective," *Papers in Applied Geography*, vol. 8, no. 1, pp. 88–111, 2022.

[12] A. Downs *et al.*, "An economic theory of democracy," 1957.

[13] S. Verba and N. H. Nie, *Participation in America: Political democracy and social equality*. University of Chicago Press, 1987.

[14] D. Koerntjes, "Het verschil in opkomst tussen tweede kamerverkiezingen en gemeenteraadsverkiezingen," 03 2022. https://www.cbs.nl/nl-nl/longread/statistische-trends/2022/het-verschil-in-opkomst-tussen-tweede-kamerverkiezingen-en-gemeenteraadsverkiezingen.

[15] K. (Rijk), "Verkiezingsuitslagen gemeenteraad 2022," 2022. https://data.overheid.nl/dataset/verkiezingsuitslagen-gemeenteraad-2022#panel-resources.

[16] ckan.dataplatform, "Stembureaus tweede kamerverkiezingen 2021," 2021. https://ckan.dataplatform.nl/dataset/stembureaus-tweede-kamerverkiezingen-2021/resource/eb2c1546-7f8d-41d4-9719-61b53b6d2111.

[17] PDok, "Cbs wijken en buurten 2019 (wfs)," 2019. https://www.pdok.nl/geo-services/-/article/cbs-wijken-en-buurten#68b8f7e2848a93f685cdf399e3fb2ef2.

[18] K. (Rijk), "Stemmen," n.d. https://www.kiesraad.nl/verkiezingen/gemeenteraden/stemmen.

[19] D. de Groot, "Gemeenteraadsverkiezing 2022: Alle gemeenten en partijen op een rij," 02 2022. https://www.bnr.nl/nieuws/politiek/10467061/gemeenteraadsverkiezing-2022-alle-gemeenten-en-partijen-op-een-rij.

[20] J. Brownlee, "How to scale data with outliers for machine learning," 05 2020. https://machinelearningmastery.com/robust-scaler-transforms-for-machine-learning/.

[21] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically weighted regression: A method for exploring spatial nonstationarity," *Geographical Analysis*, vol. 28, no. 4, pp. 281–298, 1996.

[22] C. D. Lloyd, *Local Models for Spatial Analysis*. Boca Raton: CRC Press, 2 ed., 2010.

[23] F. Tusell, P. Menéndez, M. B. Palacios, and M. Bárcena, "Alleviating the effect of collinearity in geographically weighted regression," *Journal of Geographical Systems*, vol. 16, pp. 441–466, 2014.

[24] Y. Li, X. Liu, Z. Han, and J. Dou, "Spatial proximity-based geographically weighted regression model for landslide susceptibility assessment: A case study of qingchuan area, china," *Applied Sciences*, vol. 10, 2020.

[25] B. Boehmke and B. Greenwell, "Random forests," in *Hands-On Machine Learning with R*, ch. 11, Taylor & Francis Group, 2020. https://bradleyboehmke.github.io/HOML/random-forest.html.