SAPIENZA
UNIVERSITÀ DI ROMA

Three essays on Political Economy

Gabriele Pinto

PhD in markets and institutions for innovation and development (XXXV cycle)
Sapienza University – Department of Economics and Social Sciences (DISSE)
Supervisor: Prof. Giampaolo Garzarelli
Tutor: Prof. Emma Galli

January 2023

Index

Preface

Until the beginning of my PhD I was strictly focused on studying international economics and development. My bachelor's thesis was about the central-bank interchange system, and my master's thesis discussed some issues of intra-euro trade relationships. I never had one single class in politics. So, what made me interested in elections - the topic of the 3 articles of this thesis - ?

What led me to study elections is a specific moment in my life. It was 2019. I was working in a polling station as a poll watcher. At one point, a middle-aged woman approached me with her ballot paper in hand to ask me how to fill out the ballot paper to vote for her preferred candidate. The operation was relatively trivial. It was simply a question of placing a cross on the symbol of her favorite party, yet, that was neither the first nor the last time I received similar questions.

This brief and frugal experience led me to reflect on an almost obsessively debated topic in contemporary economics, i.e. *Institutions*, leading me to ask how much we really know about how institutions are created, propagated, and destroyed. And, since many (formal) contemporary institutions are governed through elective mechanisms, the choice to study elections came almost naturally to me.

Ok, so what is this thesis about?

In the first chapter of my thesis I study the composition of the political class by looking at which individual characteristics of politicians affect their electoral success. In other words, I look at certain characteristics such as age, education, experience, or presence in social media to see if those affect the likelihood that a candidate will be elected. Concerning the literature, the contribution of this paper is two-fold. First, I have built a new dataset of politicians by text-mining all candidates' curriculum vitae for the European Election of 2019 and collected information on politicians' activities on social media (retrieved by scraping social media and API). Second, I assess how the valence characteristics of candidates (such as gender, education, job, political experience, etc.) interact with the influence on candidates' social media. I assess the importance of those characteristics using a battery of econometric models and predict candidates' electoral performance using state-of-the-art machine learning models.

My results suggest that some individual characteristics are relevant predictors of politicians' electoral success. However, I also find a certain degree of heterogeneity within the political spectrum (some characteristics are more/less relevant for some parties).

The second chapter of the thesis is strictly connected to the first one, here I study how revealing individual characteristics of candidates influences voting behavior. In other words, I want to know if people would vote differently if they were informed about who politicians are. Would voters prefer educated politicians if they knew who the educated ones are?

I answer this and similar questions by running a randomized online experiment to evaluate the effect of providing voters with a set of information about candidates' curricula. I created a website (www.elezionitrasparentiroma.it) with a collection of transparency information regarding candidates for the recent municipal election of Rome, Italy, in October 2021. The collection of transparency information published on the website derived from a law to curb corruption in Italy (adopted by the Five Star Movement in 2019) that required parties to publish their candidates' criminal record certificates, curriculum vitae, and previous political experience. I evaluate the campaign's effectiveness by inviting a randomly selected group of voters to visit the website just a few days before the election. I find that voters who visited the website were likelier to vote for the incumbent mayor Virginia Raggi, the Five Star Movement candidate. I do not see a statistically significant treatment effect on other measures of voting behavior (turnout, preference vote, etc.). Overall, my results suggest that making voters better informed on who the politicians are would not influence their attitudes. However, making transparency more salient can have partisan effects, as treated voters vote more for the Five Star Movement. These results suggest that politicization of transparency might be an important channel through which transparency campaigns can affect voter attitudes in a non-trivial way that pave future avenue of research.

The third chapter of the thesis is instead part of a separate research project that saw me very engaged in these years: the construction of a new geocoded dataset of electoral results at the precinct (*sezione elettorale*) level for Italy. The project comprises three outputs: a dataset, a dashboard to visualize electoral results at the precinct level, and a methodological paper that explain the process.

Obtaining geocoded electoral results at the precinct level can be challenging in many countries. In the paper, I present two validated methodologies for overcoming these difficulties and introduce new geocoded electoral results for several Italian cities. The dataset covers the last 20 years (1999-2022) and includes data for various types of elections, including national, regional, municipal, and referendum. After introducing the methodology, I provide an overview of some notable patterns in voting trends in major Italian cities. These include a high level of heterogeneity in voting within cities, an increase in spatial polarization of voting behavior, and an increase in the concentration of left-wing voters in central and wealthier areas of metropolitan cities. These trends may be influenced by a range of factors and can have significant implications for political representation and policy-making. The dataset provides a valuable resource for understanding these trends and exploring their underlying causes.

And that's all!

Gabriele Pinto, January 2022

Ringraziamenti

Questo lavoro non sarebbe stato possibile senza il supporto del mio supervisor prof. Giampaolo Garzarelli e il mio tutor prof.ssa Emma Galli che mi hanno sempre supportato, che mi hanno indicato la via e corretto i miei errori, ma che mi hanno anche dato tanta fiducia e libertà durante la mia attività di ricerca. In questi anni mi è capitato spesso di parlare con i miei colleghi dottorandi del comportamento dei supervisor, e mi sono reso conto come tutto questo non sia affatto scontato e della fortuna che ho avuto. Per questo motivo voglio esprimere i miei più sentiti e sinceri ringraziamenti a entrambi.

Ringrazio mio padre Guglielmo Pinto e mia madre Sophie Rossigneux che hanno sempre creduto in me, anche quando nessuno ci credeva.

Ringrazio il mio amore Livia che mi è stata accanto durante tutta quest'avventura, che si è sentita prima di tutti le mie presentazioni, che ha sopportato le mie notturne davanti ai codici, e che mi ha consolato quando le cose non andavano.

… e tutti gli altri

What characteristics of politicians affect electoral success? Estimates from a new dataset of European candidates' CV

ABSTRACT

We exploit the recent introduction of a law for political transparency to create a new dataset of all the 1076 candidacies for the European Parliamentary elections held in Italy in 2019. The dataset – dubbed the "Open-Candidati-Europee" – assembles a rich set of characteristics regarding the candidates' profiles as extracted from CV and other sources. This dataset provides a solution to overcome two important limitations for the study of political selection: we cover all the population of candidates (including non-winning candidates) and we include important omitted factors of candidates' success (such as the presence on the web). To provide a first exploratory analysis, we leverage on the expected electoral results predicted by the position in the list to provide descriptive evidence that political experience, education, and presence on the web positively predict candidates' success. We confirm previous findings of a gender bias towards female candidates. Finally, we show how to use these characteristics in a machine learning framework.

KEYWORDS: Elections, Political Selection, Quality of Politicians, Media, Politics, Machine Learning, Predicting electoral performance

The replication package for this paper is available here:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9PJDL9

1 Introduction

On February 2nd 2021, during the COVID-19 pandemic, Italian President Sergio Mattarella urged the Parliament to form a new, highly competent government to replace the second cabinet of Prime Minister Giuseppe Conte (09.05.2019–02.13.2021). Mario Draghi, former ECB Governor and an economist with a long track record in private and public management (and a PhD in Economics from MIT), was called upon by President Mattarella to form and lead a "government of the best" (*governo dei migliori*) with the mandate of saving Italy from the economic shock of COVID-19.

As a result, for the fourth time in 30 years, Italy had a technocratic government at the helm without elections.[1] The issue ultimately still seems to boil down to the need to have competent politicians at the helm. As Italian legal scholar Cassese (2018, our translation) put it, *"if the plumber and the doctor are expected to know their profession, is it not appropriate to ask someone who has to perform a more socially important task, such as that of being a member of parliament or of the executive, to have a certain degree of competence?"* [2]

Whether or not more competent politicians are desirable is a debatable[3], perhaps even a partisan question. Voters might have a different concept of representation, for instance they may disregard competence and prefer representatives more like themselves (e.g. *descriptive representation*), or interpret the vote as a judgment on the performance of incumbent (Pitkin, 1967), or simply vote for better looking candidate (Todorov et al., 2005) or those they would prefer to have a beer with.[4] At the same time, there is some sense that competent politicians correlate positively with economic performance[5] and that

[1] Though the notion is intuitive, the formal definition of a technocratic government can be nebulous. In Italy, it is widely recognized that the first case of technocratic government was led by Ciampi (1993-1994), followed by Dini (1995-1996) and Monti (2011-2013). For a review of definitions and classifications of technocratic governments in Europe, see Mcdonnel and Valbruzzi (2014).

[2] Contrary to other professions, such as judges and physicians, politicians are not required to have a higher-level degree or pass any competence exam to be elected or appointed. The fundamental (and only) barrier that political candidates need to pass is the vote. In virtually all cases, requirements boil down to minimum age and a clean criminal record. Simply put, the democratic political market has few entry barriers (e.g., Kapoor & Magesan, 2018; Stratmann, 2005; Tullock, 1965).

[3] Political philosophy widely agrees that competence should be a recognized quality of politicians. For instance, Plato: *"philosophers [must] become kings…or those now called kings [must]…genuinely and adequately philosophize"* (Plato The Republic, 5.473.d). More recently (200 years ago) James Maddison: *"..The aim of every political Constitution, is or ought to be, first to obtain for rulers men who possess most wisdom to discern, and most virtue to pursue, the common good of society…"* (Federalist Paper #57, James Maddison).

However, economists and political scientists do not always share the same view. For instance, Buchannan (1989) "*To improve politics, it is necessary to improve or reform rules, the framework within which the game of politics is played. There is no suggestion that improvement in politics lies in the selection of morally superior agents*".

On a opposite view, Besley (2005) *"Almost every major episode of economic change over the past 200 years of political history has been associated with key personalities coming to power with a commitment to these changes. How those personalities are selected ? Who become politicians ?it is essential to understand how political selection works"*.

We can summarize these two views in two starks statements: 1) all politicians are all the same 2) all politicians are NOT the same.

[4] The basic idea of the "beer test" is that voters – especially the undecided ones – vote for the candidate that they would be more willing to go out for beers with (Yascha Mounk, 2020).

[5] Several empirical works find that politicians and decision-makers with higher education and work experience produce a positive impact on economic performance. Cross-country comparisons show that national leaders and their educational

voters prefer more skilled and more experienced politicians (Dal Bó & Finan, 2018).

Essential characteristics such as age, education, gender, or experience of each candidate might matter for success and can be used as shortcuts by voters to evaluate candidates' competence. These characteristics can be relevant when parties select candidacies (before elections), and voters evaluate candidacies (in the polls).

Some studies empirically investigate the characteristics that render candidates more successful (Besley et al., 2017; Christensen et al., 2020; Dal Bó et al., 2017; Fedeli et al., 2014; Freier & Thomasius, 2016; Gavoille, 2015; Hobolt & Høyland, 2011; Le Barbanchon & Sauvagnat, 2019; Mechtel, 2014; Portmann, 2022; Schwarz & Coppock, 2022). However, these studies share, fully or partially, the same two limitations. A first limitation is the nature of the sample studied, which is often limited only to winning candidates, excluding those who did not pass the election barrier (selection bias). [6] A second limitation is the omission of some new relevant factors, such as the presence and activity of candidate on the web, that might be correlated with candidate characteristics (e.g. experience, gender and age) (Erikson et al., 2021; Gorrell et al., 2020; Mechkova & Wilson, 2021; Ward & McLoughlin, 2020).

Considering these limitations, the purpose of this paper is two-fold: first, to introduce a new, open-access dataset of candidates' Curricula Vitae (CVs) with rich information about candidates; second, to explore whether other, previously not considered, factors can predict a candidate's success. In our exploratory analysis we consider how these individual characteristics correlate with the candidates' electoral success measured by the share of preferences obtained by each candidate within their party.

Our original dataset – which we call "Open-Candidati-Europee" –describes all of the 1076 candidacies for the 2019 European elections in Italy. As for our knowledge, this is the first dataset that covers the entire population of candidates for an Italian election.[7] There are two primary motivations behind the selection of the Italian case study: (*a*) we face a Proportional Representation election with Preference Voting, allowing us to disentangle the preferences for parties from those for single candidates; and (*b*) we have access to the CVs of all candidates.

attainment are positively correlated with the rate of growth of GDP (Besley et al., 2011; Jones & Olken, 2005). Within country, Gagliarducci and Nannicni (2013) show that a higher wage improves economic performance by attracting better educated and better skilled politicians. Freier and Thomasius (2016), using education and experience level of local mayors in Germany, find that only experience matters for fiscal performance (in terms of reducing debt, expenditure and taxes), while education does not seem to exert any significant effect. Meriläinen (2022) finds that competent politicians improve fiscal sustainability but do not decrease the size of the public sector (Meriläinen, 2022).

[6] Swedish and Finnish data used by Dal Bó et al. (2017) and Meriläinen (2022) constitute a notable exception as they cover the entire population of politicians.

[7] Available here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9PJDL9

To construct the dataset, we exploit the introduction of new anti-corruption rules that require all candidates to fill their CVs in a public archive (Law 3, 9 January 2019). We construct the dataset by text-mining those CVs of the candidates and other official administrative sources. [8] To provide a first exploratory analysis, we estimate a battery of Vote-Popularity-Functions (VPF) to evaluate which characteristics increase the share of preferences obtained by each candidate. We use the same information to predict electoral success using Machine Learning tools. We consider education, gender, previous administrative experience, campaign spending, and the candidates' social media presence.

Our findings suggest that previous political experience strongly predicts a candidate's success, while education has a limited role in explaining candidates' success and is rarely statistically significant. Consistent with previous research, we find evidence of a solid negative gender bias towards female candidates. [9]

We also contribute by delving into the influence of popularity in social media (Facebook, Twitter, and Google Trends), where we find, as expected, that a candidate's exposure on the web correlates positively with a higher number of preferences obtained. We also find that accounting for social media reduces the effect of education, political experience, and gender bias. However, being a female candidate reduces or even inverts the positive impact of being famous on the web. The negative gender bias is more pronounced among right-wing parties (except for parties run by a female leader), while education is more appreciated among left-wing parties.

Lastly, we show the predictive power of candidates' characteristics in a Machine Learning (ML) framework. We find that employing more advanced and non-linear models can significantly improve predictions about candidates' success by as much as 40% compared to the baseline linear OLS model. Thus, we show how combining traditional characteristics of candidates with data from social networks can increase the performance of ML models for electoral predictions. These results add to the literature on employing ML for electoral predictions that can be leveraged for evaluating electoral results in Machine Learning Control Models (MCLM) (Cerqua et al., 2021; Gayo-Avello et al., 2011; Grimmer et al., 2021; Schoen et al., 2013; Skoric et al., 2020; Tsakalidis et al., 2015; Varian, 2016).

Our paper contributes to a strand of literature that devoted attention to the determinants that shape the quality of the political class by providing a new rich data-source of politicians characteristics and new descriptive evidence of which

[8] The text-mining procedure (together with a validation procedure to ensure quality of the) is carefully described in online appendix and the scripting code is stored in the public replication package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9PJDL9

[9] While the presence of gender bias is largely shared in the literature, some studies disagree. Broockman and Soltas (2020) found no presence of gender bias; Christensen et al (2020) found a positive effect for women candidates in Finland; McDermott (1998) claims that women and black candidates are stereotyped as more-liberal than the average white male. In a meta-analysis of vignette experiments Schwarz and Coppock (2022) show that on average the gender bias is positive.

individual characteristics of politicians matters for their electoral success (Besley, 2005; Besley et al., 2017; Besley & Reynal-Querol, 2011; Braendle & Stutzer, 2019; Buisseret & Prato, 2020; Caselli & Morelli, 2004; Christensen et al., 2020; Dal Bó & Finan, 2018; Fedeli et al., 2014; Freier & Thomasius, 2016; Gulzar, 2021; Hobolt & Høyland, 2011; Mechtel, 2014; Meriläinen, 2022; Stokes, 1963).

2 Data Overview

2.1 Data and sample description

The opportunity to assemble this dataset comes from a political transparency regulation introduced by Law n. 3/9 January 2019 that requires all parties competing in an electoral race to publish their candidates' CVs and criminal records online. CVs and criminal records must be published online on the party's website and on a dedicated page of the website of Ministry of Interior (for National elections) or Municipality (for Local elections) called "Transparent Elections" (*Elezioni Trasparenti*). The regulation was part of a law package to fight corruption, literally named "sweeper of the corrupt" (*spazzacorrotti*). It is not a coincidence that the objective of the regulation is to nudge voters to consider the candidate's qualifications, with the final goal to produce a more competent political class.

Since the information on these CVs comes in an "unstructured" form (scanned pages) that is not immediately usable for analysis, we applied text mining techniques to extract the information of interest.[10] We then matched the candidates with other information originating from different sources (*Anagrafe degli amministratori, archivio camera dei deputati e del senato*) including the preferences obtained by each candidate in the elections (*collected from the Ministry of Interior*). We merged this information in a new database with over 80 features for all the candidates running for the Italian-European Elections of 2019 named the "Open-Candidati-Europee." In the exploratory analysis of this paper we will only make use of a limited set of these variables. Other features include the electronic text of the CV, and detailed description of the criminal record of each candidate. A full list of the features can be found in the online appendix.

Table 2 shows the summary statistics of some selected features we will use in the analysis presented in the next section. We have 1076 candidacies, a unique combination of 983 candidates, 18 party lists, and 5 districts. We measure the results of each candidacy at the provincial level (107 provinces), making a total of 24.052 observations.[11]

[10] In the online appendix we describe how we collected and processed the CVs to convert those from an unstructured form to a structured dataset that is immediately usable for analysis, and how we collected additional features and reconciled them from a variety of sources.

[11] Note that: (*a*) some candidates have multiple candidacies in different districts (*b*) each district where the candidacy applies has a different number of provinces.

The supply of candidates is qualitatively high compared to the Italian population, with a significant share of the candidates being professional politicians. In our dataset, over 30 percent of candidates have at least one political experience, and 70 percent are "freshmen" (no previous political experience). Just 4.6 percent of them are incumbent. Of our total candidacies, 69 percent have a university degree (20 percent of the population), and 6 percent also have a PhD (4 per cent of the population).

For each candidate, we observe *Degree* that is a dummy equal to 1 if the candidate has at least a University degree, *PhD* is a dummy equal to 1 if the candidate has a PhD. *Female* is a dummy equal to 1 if the candidate is female (0 otherwise). *Political experience* is a dummy on whether the candidate has at least one previous administrative experience in the Italian Parliament, European Parliament, or a local assembly (region, province, or municipality). *MEP* is a dummy on whether the candidate is incumbent (i.e., is a member of the European Parliament) (4.6 percent of them). *Regional experience, Provincial experience, Municipal experience* are the number of years of experience respectively in the Region, Province, and Municipality. *Parliament experience (Chamber)* and *Parliament experience (Senate)* are the numbers of legislative mandates served in National Parliament either in the Chamber of Deputies or the Senate of the Republic (*Camera dei Deputati* and *Senato Della Repubblica*). Concerning experience in the Local Assembly, if the candidate had multiple assignments troughout the legislature (such as mayor and councilor, or mayor and assessor[12]), the years of experience are counted two times.

We measure the *age* of the candidates, with the youngest candidate being 25 years old and the oldest candidate being 90 years old.[13] We collect information on the position of the candidate in the list and the influence on social media. *Position list* is the initial position of the candidates on the ballot list, and *Rank Change* is the difference between the initial position in the list and the rank of the candidate according to the preference received within the list. *Google Index* is the cumulative sum of the Daily Google Trend Index calculated over the two months before the elections (Between March 26th, 2019 and May 26th, 2019)[14]. *Twitter* is the mean number of retweets calculated over the candidate-related tweets in

[12] In Italy an assessor (*assessore* in Italian) is a member of the executive body of local government. It is a figure similar to that of a Minister in the national government. Assessors are appointed by the chief executive of local government (Mayor in municipalities and President in regions, respectively) and are assigned a portfolio for a specific aspect of municipal or regional affairs (environment, health, etc.) and the supervision of the corresponding branch of local government, called *assessorato* (department).). They serve until the chief executive ends his term or resigns, unless dismissed beforehand.

[13] Francesco Alberoni, candidate for Fratelli D'Italia, was born on 31st December 1929.

[14] The google trend index measures the intensity of research of a term on google in a given period (https://newsinitiative.withgoogle.com/resources/lessons/basics-of-google-trends/). The raw index returned by Google is a normalized value measured in each day that ranges between 0 (if not enough data are available, e.g. too few searches) and 100 (maximum value of searches in a given period). We retrieve the daily google trend index for the name and surname of each candidate for the whole period of the two months before the election (between March 26th and May 26th 2019). After having retrieved this value, we compute the sum of the daily values for each candidate to construct our ***Google Index.*** The interpretation of our index is that candidates with a constant presence on the web will have a higher value of the google trend index, while candidates who are not popular on the web or have a non-constant presence will have a low value of the google trend index.

the week before the election. *Facebook Ads* is the amount spent on Facebook advertising between March 2019 and June 2019 (thus including the post-election period). The primary dependent variable, *Preferences*, is the total number of preference votes obtained by the candidate. *Preferences (%)* are the number of preferences obtained by the candidacy over the total number of preferences of the candidate list in the same district.

As is clear from Figure 1 and Table 2, the distribution of the *Preferences* is highly skewed to the left, meaning that we have many observations with "small" values and very few observations with values above the average. This observation is valid both in the relative share of preferences (*Preferences (%)*, right plot) and with more intensity on the absolute number of preferences (*Preferences*, left plot).[15]

The boxplots in Figure 2 already reveal some significant differences between a candidate's attributes and electoral results. The candidates with *Degree* have higher values in mean and in the whole 75th percentile than those without any degree. *PhD* candidates do not seem to differ in the electoral outcome compared to candidates without a *PhD*. In contrast, male candidates have significantly better results than *Females*, suggesting a gender bias towards female candidates. The distribution within parties is skewed to the left and often comprises some "champion" candidates (*campioni delle preferenze*) who obtain a very high share of the preferences (the black dots in the second-last plot of Figure 2).[16] There are no significant differences in the distribution of preferences among electoral districts with the exception of the district of *Sicilia* and *Sardegna* (*District IV – Italia Insulare*).

[15] The presence of very "extreme" values "far" from the median and the mean of the distribution might create issues in our model estimation (outlier issues). We report model estimates that exclude outliers and log transformations to tackle these issues in our exploratory analysis (Table A-1 in the Appendix).

[16] This is true except for few smaller parties such as *Partito Animalista, Autonomie Per l'Europa, PPA* and SVP.

Table 2: Descriptive statistics

	count	mean	std	min	25%	50%	75%	max
Preferences	24052	541.07	3252.12	0	3	17	115	208448
Preferences (%)	24031	6.95	12.83	0	0.39	1.75	6.9	100
Degree	24052	0.69	0.46	0	0	1	1	1
PhD	24052	0.06	0.24	0	0	0	0	1
Female	24052	0.49	0.5	0	0	0	1	1
Political experience	24052	0.32	0.46	0	0	0	1	1
Regional experience	24052	0.5	2.44	0	0	0	0	27
Provincial xperience	24052	0.32	1.57	0	0	0	0	19
Municipality experience	24052	2.46	5.88	0	0	0	1	48
Parliament experience (Camera)	24052	0.14	0.7	0	0	0	0	9
Parliament experience (Senate)	24052	0.03	0.21	0	0	0	0	4
European experience	24052	0.05	0.21	0	0	0	0	1
Age	24052	52.35	12.29	28	44	52	61	92
Position list	24052	6.98	4.3	1	3	6	11	15
Rank change	24052	0.29	4.77	-14	-2	0	3	14
Google index (log)	24052	2.78	2.79	0	0	4.11	5.29	8
Twitter (log)	24052	0.37	0.7	0	0	0	0.51	4.81
Facebok Ads (log)	24052	1.27	2.8	0	0	0	0	11.77
Provinces	107							
District	5							

Figure 1

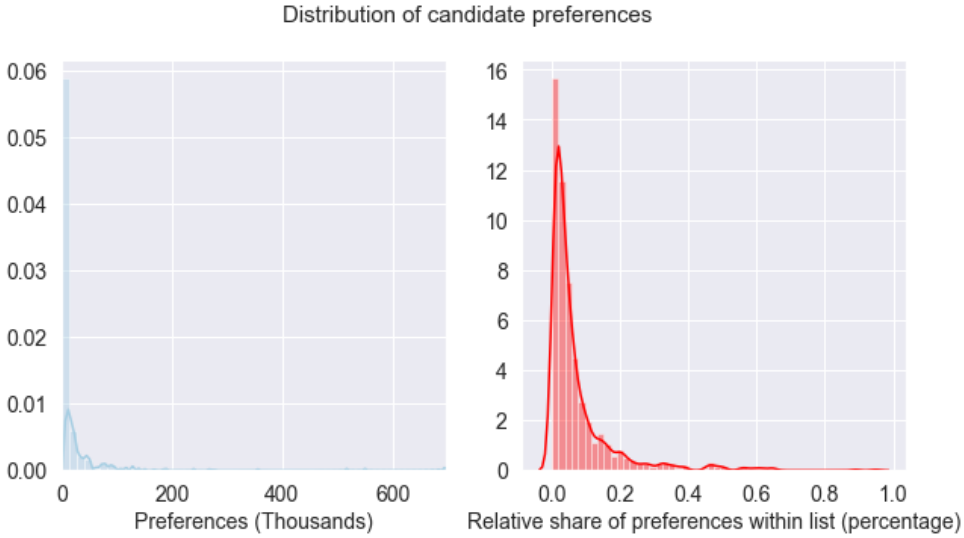
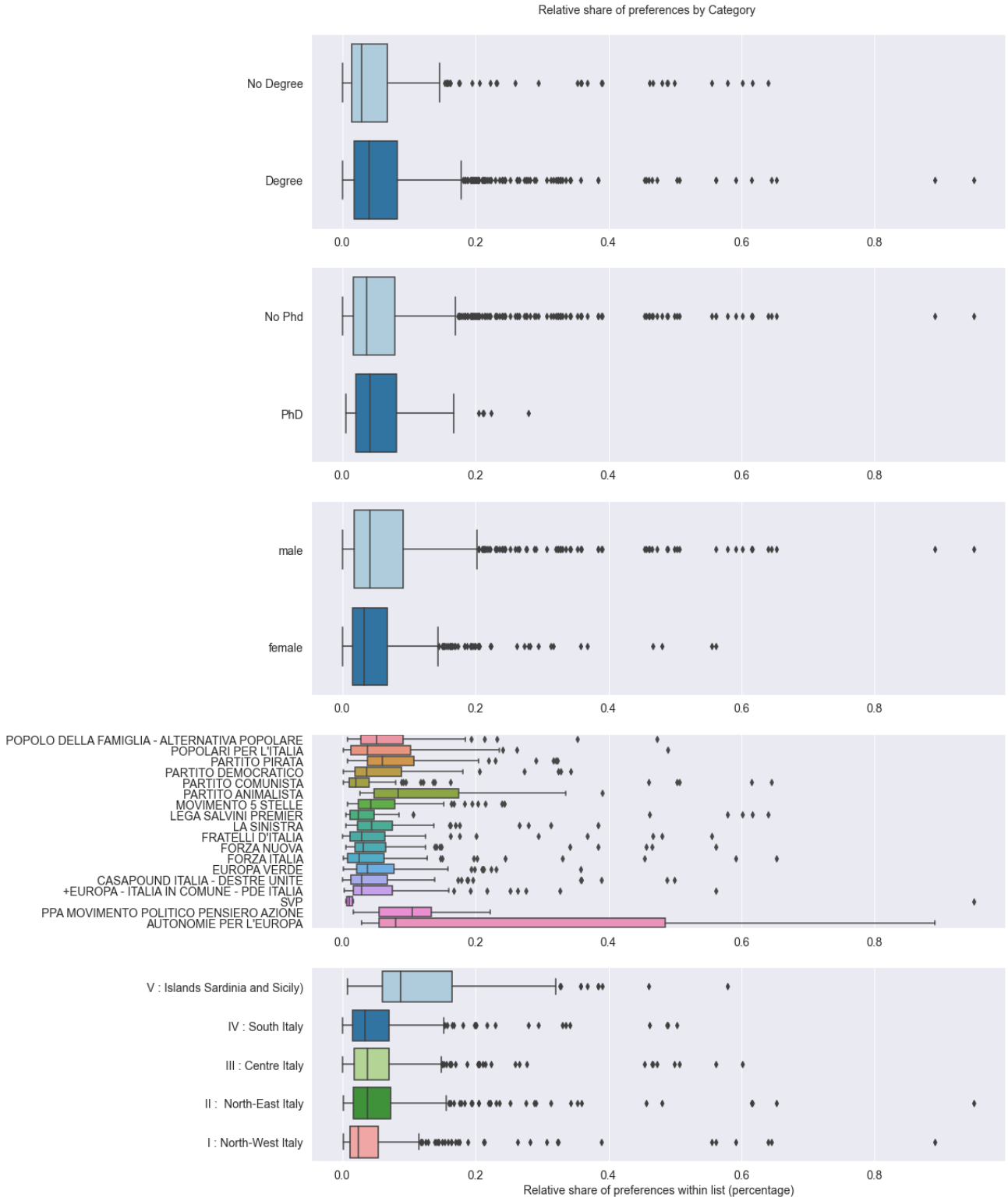


Figure 2



Notes: The length of the box contains all the values between the 25th the 75th percentile of the distribution for each category (Labels on the left). The length of the whiskers (line) reach the values corresponding to 1.5 times the interquartile range (difference between third and first quartile). Dots are the observation that do not fall within this range (i.e. observations that are small or larger than 1.5*(75th Percentile – 25th Percentile). Those heuristic values are taken from the default values of the matplotlib package.

2.2 The European elections in Italy

The European Elections represent an interesting case study as they have an electoral Proportional Representation with Preference Voting (André et al., 2012; De Luca, 2001; Katz, 1985). In the Preference Voting system, voters are free to express a preference for a specific candidate[17] , while candidates have incentives to "personalize" the electoral campaign (Renwick & Pilet, 2016). We look specifically at the round of the 2019 elections because *a)* it is the first election after introducing the law for political transparency[18] *b)* it is the first election for which we have information on all candidates –and not only for those elected – *c)* it allows us to provide a significant improvement compared to other elections for which information is only available for elected candidates.

European Elections are "second-order elections" but have relevant political value at the national level (Pasquino & Valbruzzi, 2019; Willermain, 2014), overcoming another limitation of other studies mainly investigating low-level municipal elections. Indeed, the mechanisms at play in a second-order election are different from that of a first-order election (less participation, less media attention) and radically different from elections where preference voting is not present. For instance, European Elections might attract a sample of voters that is not representative of voters participating in other type of elections. We acknowledge that the results we present in this paper should be read within the context of the specific case study and not generalized to another electoral context where "personalization" incentives and the deal at stake are radically different.

3 Empirical Strategy

3.1 Empirical model

We estimate a Vote-Popularity function (Nannestad & Paldam, 1994; Paldam, 2008) to evaluate the association between the candidate's electoral result and each candidate's characteristics. Our dependent variable in equation (1) is the candidate's electoral success. While our independent variables are a set of factors that should explain a candidate's success. We can think of our dependent variable as the result of voters' evaluation of the independent variables (the candidacy features).[19] In other words, we will try to estimate the coefficients of the demand function of the voters, given the "supply" of candidacies.

[17] In fact, the European Elections are the only election at national scale in Italy where voters can express a preference for a specific candidate. The preference voting mechanism is also present in municipal elections, but nationally recognized leader do not usually run for this type of election.

[18] Law n. 3/9 January 2019

[19] This approach has also been used in related literature aimed at explaining how voters price and infer information on candidates. For example, Mechtel (2014) investigates how voters react to the provision of occupational information listed in the ballot paper (e.g. when the ballot paper provides the "occupational" profile of the candidate), and finds that there is a significant "occupational" effect. A pioneering paper from Mueller (1970) also finds that when candidates have an occupation related to high-education profiles, the success of those candidates increases significantly.

Note that we measure preferences at the province level, while the supply of candidates is fixed at the district level. Thus, we keep the supply of candidacies invariant while the demand varies among provinces.[20]

In the simplest case, we estimate the following equation:

$$P_{c,p} = \left[\sum_j \beta_j * X_{jcp}\right] + u_{c,p}, \qquad (1)$$

where $P_{c,p}$ is the success of candidacy *c* in province *p*, which will be measured using the *Preferences (%)* of the Candidate in the list.[21] On the right-hand side of the equation, we have a matrix of j independent variables X_{jcp}. This matrix includes: two dummies for *Degree* and *PhD* (the corresponding coefficients β_i can be considered the *education premiums*); a *Political experience* dummy vector equal to 1 if the candidate has occupied any post in elected offices in the period before the elections (either at provincial, regional, municipal, and parliament level), equal to 0 otherwise. The corresponding coefficient will estimate the *experience premium*. Next, we have two vectors representing our 2 indexes for popularity (*Google Index* and *Twitter*), [22] with the corresponding coefficients being the *Popularity premia*. We also add a battery of control vectors, including *Age*, Party fixed effects, and Province-fixed effects. Since characteristics of the candidate replicate in different provinces where we evaluate results, we opt to cluster standard errors at candidacy level *c*.

The stochastic component $u_{c,p}$ captures all those factors that affect a candidate's success and are not observable or not included in the regression. To satisfy the OLS assumptions for unbiasedness of the β estimators, we need orthogonality between $u_{c,p}$ and our independent variables X_{jcp}.

A first concern is that parties might decide to allocate candidates with different strategies, causing bias due to supply-side effects. Province and district fixed effects let us capture this source of endogeneity. We thus leverage the fact that while candidates are supplied at the district level, our observations vary at a lower level (province level). The use of province and district fixed effects also controls the supply-side impact caused by the different degrees of political competition (Galasso and Nannicini 2011). We rule out ideology and party affiliation by using Party fixed effect.

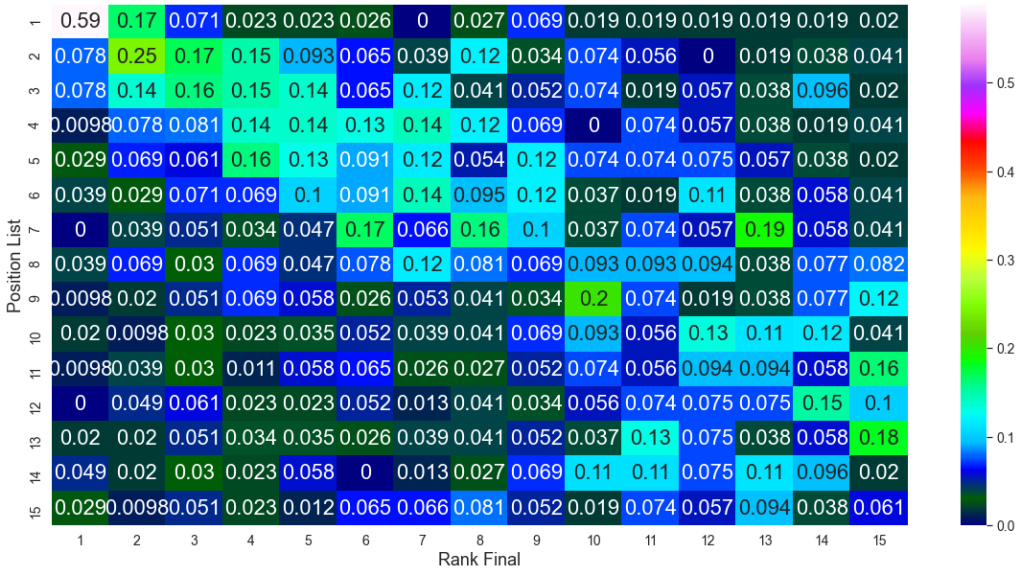
[20] To avoid that repeated candidacies inflate our estimators, we cluster standard errors at candidacy level. In the appendix we also report baseline results by measuring electoral success at district level.

[21] Alternatively, we also show the results using as dependent variable the *Rank Change* (results are reported in the Appendix, Table A-2) that is the difference between the initial position of the candidate in the list and the final ranking of the candidate in the list according to the number of preferences received. In other robustness tests we use as dependent variable the total number of preferences a candidate got in a certain province over total number of preferences in that province (Table A-4 in the appendix).

[22] The average Google Trend Index in the two months prior to the election and the total number of retweets of candidate-related tweets in the week before the election

Among other factors that can invalidate our orthogonality assumption, we can mention the budget invested in the campaign (campaign spending). An extensive literature established the presence of an effect of campaign spending on election outcome (e.g., Jacobson 1990).[23] To control the impact of campaign spending to reduce this bias, we provide two different strategies. First, we include a variable that measures the candidate's campaign spending for advertising on Facebook platforms.[24] Second, we have a measure of popularity on social media. Our measures of *popularity* should capture part of the effect of campaign spending. A candidate who spent more on the campaign can hire professional staff to manage the press office management of social media communication (such as Twitter, etc.) to increase popularity.[25]

Figure 3



[23] This is particularly relevant since it can of course be correlated with almost all our predictors and thus invalidate orthogonality. Consider a candidate that is expected to be a "winning horse," on whose campaign the party (or the candidate herself) could decide to invest heavily. Since it is likely that campaign spending is positively correlated with candidate success, the bias will have a positive sign on the coefficients of the variables included in the regression that are positively correlated with campaign spending. For example, if parties decide to invest more on the candidates with more experience, this will inflate positively the experience premium. Following the same reasoning, a candidate with higher education, might have a higher income (for which we have no information), or a higher access to fund its campaign budget, the coefficient of the education premium would then be positively biased.

[24] Available only for those candidates that actually invested in advertising on the platform. We are also aware that this latter measure might be not reliable and could suffer from mismeasurement (Ansolabehere & Gerber, 1994), as it might also be associated with other features (such as how much the candidate relies on digital media instead of other forms of political campaign). Furthermore, the same campaign spending variable might be endogenous. We can argue that Voter Success and campaign spending cause each other through the channel of expectations of the results (i.e. if a candidate expects to be "behind target" it will increase its budget). Other endogeneity issues with estimating the effect of campaign spending are widely discussed in Jacobson (1990, 2006) and Gerber (1998).

[25] If our popularity indexes are highly correlated with the omitted variable *campaign spending*, this will partially capture an endogeneity problem. In the "extreme" case in which our popularity measures (that basically measure the presence of the candidate in the web), are perfectly correlated with campaign spending, there will not be any residual variation and thus the omission of the variable is not relevant for the unbiasedness of the other estimators. But if campaign spending is used also for other campaign channels (such as organization of dinners, street posters, street rally), the omission of campaign spending would still be a source of bias for our estimates.

The rank position of a candidate in the ballot list can explain the candidate's success. Parties know the importance of the rank position and play a role in deciding the order of the candidates on the ballot list. For this reason, parties might strategically allocate candidates based on external conditions (competition), electoral goals, and specific characteristics (Berz & Jankowski, 2022; Buisseret & Prato, 2020). Thus, the inclusion of the variable ***Position list*** should allow us to capture quite a significant component of the parties' strategies. Figure 3 shows a co-occurrence matrix of the initial position on the ballot list against the final rank, as we see candidates tend to place near their initial ballot position.[26]

There can be two reasons to explain this ranking. The "name-order effect" is powerful,[27] or – more plausibly – the initial position on the ballot is a revelation of the party's strategy and pre-election chances of the candidate. Including *Position list* in the regression or using *Rank Change* as a dependent variable will account for all the factors anticipated by the ranking of the candidate on the ballot. In other words, it disentangles the actual results from the expected results of the candidate.

The whole discussion on the identification of the model refrains us from claiming causality; we are aware of potential issues, such as those just discussed, that can bias our estimates. Any evaluation of the results must then be read with caution due to the possible failure of the underlying assumptions described. Despite those evident limitations, we can treat our estimates as significant associations between the considered variables. We try to control for a wide range of variables we consider relevant and in line with the literature and we provide a series of robustness tests with different samples and measures, comparing our results with previous studies.

4 Results

4.1 Baseline Model

Table 3 and Figure 4 show the results from the estimation of Equation 1. Model-1 includes only the education dummies (*Degree* and *PhD*); from Model-2 to Model-5 we sequentially include *Political experience*, *Female*, *Age*, and *Position list*, popularity indexes (*Google Index* and *Twitter*), and campaign spending (*Facebook Ads*). In Model-6 we also include Party and district fixed effects (a matrix of dummy vectors for each of the 18 parties and 5 districts[28]).

[26] The high values among the diagonal line of the matrix in Figure 3 provide evidence for this. An additional piece of evidence is provided in figure A-7 in the Appendix, where we show that Position in the List has strong predictive power for candidates' vote share.

[27] For "name-order effect" we are referring to the effect caused from being in a position on the list independently from any other variable (such as the effect we would register if candidates would be randomly ordered). For a brief explanation, see Taebel (1975) and Koppell and Steen (2004).

[28] The two matrixes have c*(n-1) dimension since we exclude reference categories to avoid multicollinearity.

We see that the coefficient of *Degree* and *PhD* are rarely statistically significant at conventional confidence levels, but while having only a *Degree* has a positive effect, having a *PhD* has a negative one, suggesting a non-monotonic relationship between education and electoral success. If we exclude top and bottom outliers, the coefficient for having a university *Degree* is significant at 95 per cent confidence level (Table A-1 in the Appendix). Different type of Degree (law, econ and medical) do not appear to affect electoral success (Table A-X in the appendix).

Having previous political experience has a strong effect on a candidate's success. The coefficient of *Political experience* has a relatively stable and statistically significant positive sign. In terms of magnitude, having at least one previous political experience increases the relative share of votes of a candidate by 2.89 percentage points (≈1/4 standard deviation).

The coefficient of *Age* is positive but never statistically significant at conventional confidence interval.[29]

Our results acknowledge the presence of a gender bias in line with other studies (e.g., Barbanchon and Sauvagnat 2018). The gender coefficient (*Female*) has a negative sign of around -2.4 (MODEL-4), which is statistically significant in all specifications and suggests the presence of a strong gender bias that leads to an average reduction of 2.4 percentage points (≈1/5 standard deviation) of the relative share of votes for female candidates.

The coefficient of the position of the candidate in the list (*Position list*) is statistically significant in all specifications, with (as expected) a negative sign (-0.71) (model 6). A downgrade of one position in the rank is associated with a reduction of 0.71 percentage points in the relative share of preferences. This effect is indeed quite strong if we consider that the Position list variable varies between 1 and 15.[30] A candidate that is 10th in the list has on average 7.81 percentage points less in the relative share of votes compared to a candidate that is positioned in the first position of the list.[31] This strong effect might measure the strategy of the Party in supporting certain candidates (reflected in the initial position assigned on the list).

Popularity in social media is strongly correlated with the electoral success of a candidate. The Popularity indexes (*Google Index, Twitter,* and *Facebook Ads*) are all statistically significant. Since those variables are expressed in logs, while the dependent variable is expressed in level, the interpretation of the coefficient is expressed by the formula $\Delta y = \frac{\beta}{100}\%\Delta X$. For example, an increase of 10

[29] We insert ***Age*** in log transformation and quadratic form (Table A-4 in the appendix). In both cases we do not see statistical significance

[30] Some of the candidate lists have less than 15 candidates. We could have normalized the position of the list. We decided to keep the original number as the coefficient is still statistically significant and the original number helps us with the interpretation of the coefficient.

[31] -0.71*10-0.71=-7.81

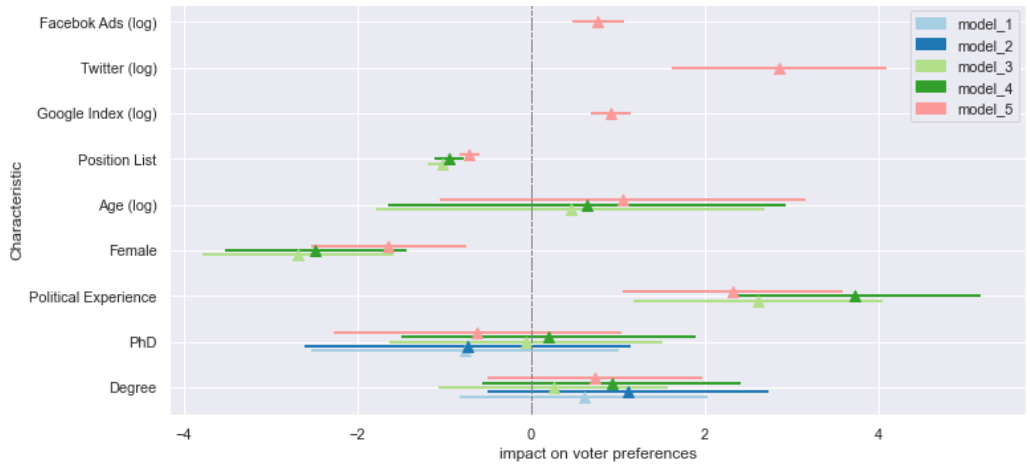
percent of spending on Facebook advertising increases the relative share of votes by 0.077 percentage points (≈1/10 standard deviation) (model 6).[32] Apart from looking at the effect of popularity per se, we can also highlight that including social media indexes in the regression (model 5) reduces the other coefficients.

Table 3: Main Results

Dep.Var. Preference Votes (%)	I	II	III	IV	V
Degree	0.61	1.11	0.26	0.93	0.73
	(0.72)	(0.82)	(0.67)	(0.76)	(0.63)
PhD	-0.76	-0.73	-0.06	0.20	-0.61
	(0.90)	(0.96)	(0.80)	(0.86)	(0.84)
Political Experience			2.62***	3.73***	2.33***
			(0.73)	(0.74)	(0.65)
Female			-2.68***	-2.48***	-1.64***
			(0.56)	(0.53)	(0.45)
Age (log)			0.46	0.65	1.04
			(1.15)	(1.18)	(1.08)
Position List			-1.02***	-0.94***	-0.71***
			(0.08)	(0.08)	(0.06)
Google Index (log)					0.92***
					(0.12)
Twitter (log)					2.85***
					(0.63)
Facebok Ads (log)					0.77***
					(0.15)
N	*24052*	*24052*	*24052*	*24052*	*24052*
CONTROLS: Province and Party fixed effects in model II, IV and V					

Notes: *,**,***; p-value at 1,5 and 10 per cent; standard errors are clustered at candidate level. Dependent variable is expressed in percentage (min=1, max=100). Magnitude interpretation: A coefficient of 2.03 for a dummy variable means an increase of 2.03 percentage points for the relative share of preferences. The mean value for the dep variable is 6.9, with 12.8 standard deviation.

Figure 4: Regression coefficient plot of Table 3



[32] $\frac{0.77}{100} * 10 = 0.077$

At the same time, the null effect of having a PhD becomes negative and significant (from 0.20 to -0.62). The change in the coefficients suggests that omitting popularity in social media might cause an upward bias of education and experience premium.

Indeed, the use of social media might be endogenous to omitted characteristics or interact with known ones. For instance, women might receive more confrontation and abuse speech on social platforms (Erikson et al., 2021; Gorrell et al., 2020; Mechkova & Wilson, 2021; Ward & McLoughlin, 2020). We find that being female has a negative interaction effect with being popular on the web (e.g. the interaction coefficient is negative, see Table A-7 in the appendix), suggesting that being popular on the web might be detrimental for negative candidate.

4.2 Selection bias

As we mentioned at the outset, one of the main novelties of this dataset is that we can cover both elected and non-elected politicians, overcoming the selection bias issue caused by looking only at winner candidates. But to what extent is the selection bias relevant? To answer this question in Table 4 and Figure 5 we compare the results of estimating the same model of equation 1 using either the whole sample of 1076 candidates or a reduced sample with only the 76 elected candidates.[33]

As we can see, the selection bias appears to be relevant. The estimated effects of Age, Female and education in the model using only the sample of elected candidates (red estimates in Figure 5) are significantly biased upwards (higher effect). However, the other coefficients (Popularity index, Political experience, and PhD) do not appear to be affected depending on the selection of the sample.

4.3 Political experience

In the model described so far, we treated the political experience as binary. The information we have on different types of political experience let us further investigate the effect of Political experience. Table 5 and Figure 6 report the coefficients from estimating a model like Equation 1.

[33] For this exercise, we estimate the same models of Table 3. With one exception, we do not include Province Fixed Effect in both full sample and the elected sample estimations. We took this decision because we do not have enough observations for some combination of Party and Province (the matrix X^1X does not have full rank) in the Elected Sample. The confidence interval are clearly larger in the model using only the Elected Sample for a trivial reason: smaller number of observations (75 compared to 1076 in the full sample)

Table 4: Selection Bias

Dep.Var. Preference Votes (%)	I	II	III	IV	V	VI
Sample	full	full	full	elected	elected	elected
Degree	0.26	-0.39	0.74	5.96**	3.44	-1.53
	(0.67)	(0.59)	(0.63)	(2.35)	(2.11)	(1.83)
PhD	-0.06	-1.17	-0.62	-0.99	-1.01	-2.00
	(0.80)	(0.86)	(0.82)	(3.13)	(3.20)	(2.60)
Political Experience	2.62***	0.45	2.28***	2.80	1.87	-1.28
	(0.73)	(0.64)	(0.66)	(2.02)	(1.94)	(1.52)
Female	-2.68***	-2.21***	-1.63***	-5.78***	-4.93**	-1.15
	(0.56)	(0.51)	(0.46)	(2.13)	(1.99)	(1.55)
Age (log)	0.46	0.94	1.00	19.46***	15.69**	12.12**
	(1.15)	(1.10)	(1.10)	(6.95)	(6.35)	(4.74)
Position List	-1.02***	-0.87***	-0.74***	-1.53***	-1.51***	-0.79***
	(0.08)	(0.07)	(0.06)	(0.35)	(0.41)	(0.25)
Google Index (log)		0.66***	0.94***		0.55	0.34
		(0.10)	(0.12)		(0.40)	(0.32)
Twitter (log)		2.68***	2.82***		1.74	2.99***
		(0.64)	(0.63)		(1.64)	(1.01)
Facebok Ads (log)		0.47***	0.76***		0.54	0.58**
		(0.16)	(0.15)		(0.34)	(0.27)
N	*24052*	*24052*	*24052*	*1786*	*1786*	*1786*

CONTROLS: Province and Party fixed effects in model III and VI

*,**,***; p-value at 1,5 and 10 per cent; standard errors are clustered at candidate level. Dependent variable is expressed in percentage (min=1, max=100).

Figure 5: Regression coefficient plot of Table 4

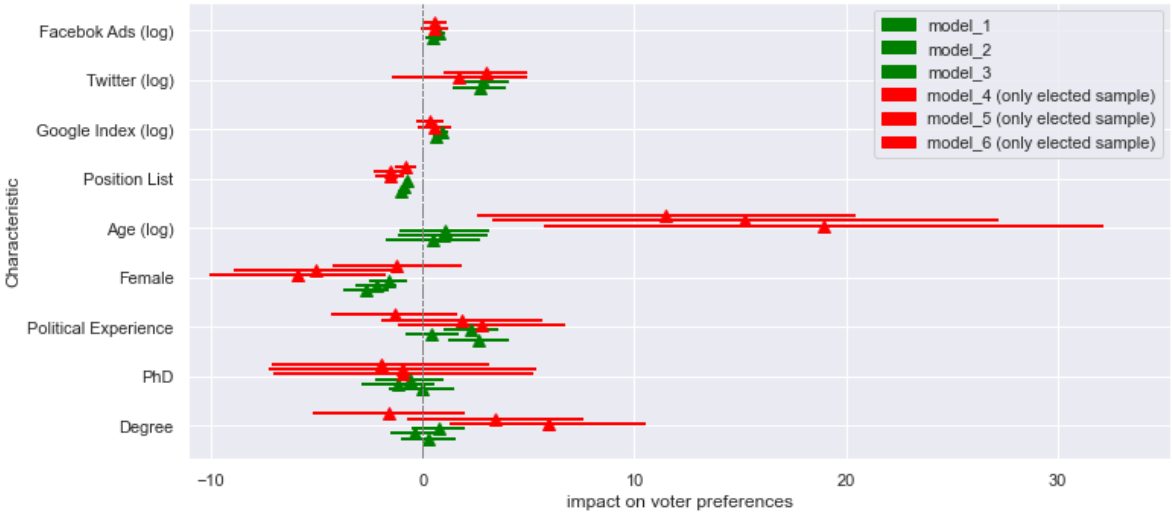


Figure 6: Regression coefficient plot of Table 5

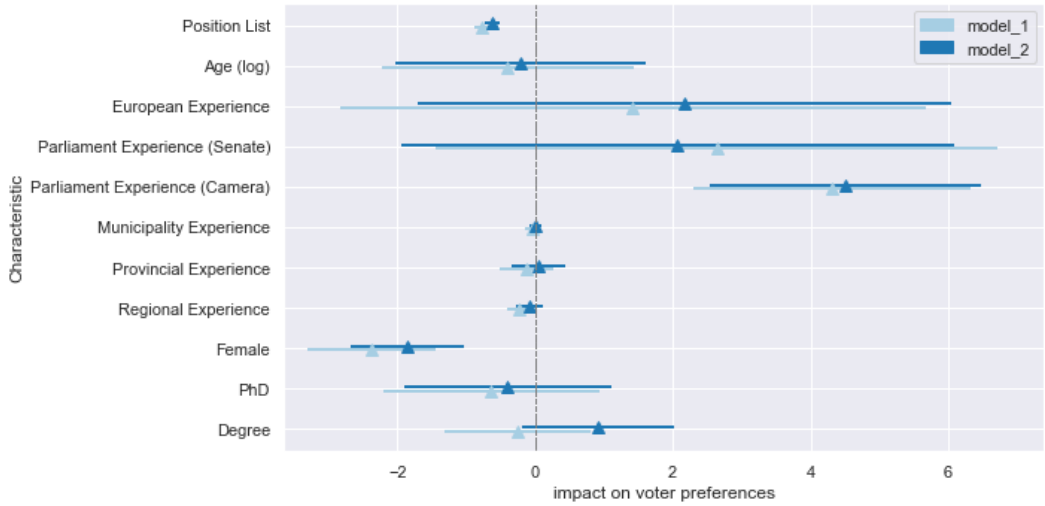


Table 5: Type of Political Experience

Dep.Var. Preference Votes (%)	I	II
Degree	-0.26	0.91
	(0.54)	(0.56)
PhD	-0.64	-0.40
	(0.80)	(0.76)
Female	-2.38***	-1.85***
	(0.47)	(0.42)
Regional experience	-0.22**	-0.08
	(0.10)	(0.10)
Provincial experience	-0.13	0.04
	(0.20)	(0.20)
Municipality experience	-0.05	0.00
	(0.05)	(0.05)
Parliament experience (Camera)	4.32***	4.51***
	(1.03)	(1.01)
Parliament experience (Senate)	2.64	2.07
	(2.08)	(2.05)
European experience	1.42	2.17
	(2.17)	(1.98)
Age	-0.46	-0.28
	(0.94)	(0.93)
Position list	-0.77***	-0.62***
	(0.06)	(0.05)
	(2.17)	(1.98)
N	*24052*	*24052*
CONTROLS: In model **II**, *Province and Party fixed effects*		

Notes: *,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. Dependent variable is expressed in percentage (min=1, max=100). Magnitude interpretation: Political experience (Regional, Province, Comune, Camera and Senate)is now expressed in number of years. Thus, a coefficient of 4.47 for Parliament experience implies an increase of 4.47 percentage points in the relative share of votes for each number of years spent in Parliament. European experience (European Parliament experience or "Incumbent" is a dummy variable). The dependent variable has mean 6.9, with 12.8 standard deviation

We replace Political experience with several categorical variables measuring different types of political experience. The benchmark condition remains to have no previous experience, compared to having experience in local, national, and European level assembly. As we see, only parliamentary experiences in the *Camera dei Deputati* (*Parliamentary experience (Camera)*) continue to have a positive and statistically significant sign. In contrast, other types of experience have "inverted" signs and are not statistically significant. *European experience*, which is a dummy on whether the candidate is "incumbent" or not, has a positive sign but is not strongly statistically significant. This "incumbent effect" coefficient tells us that incumbent candidates have an advantage of 2.17 percentage points compared to not having any experience. Having experience at the higher level (National or supra-national Parliament) has a positive effect, while being experienced only at the local level does not matter for being successful in this higher-order election.

4.4 Heterogeneous effects

Finally, we investigate how the influence of those characteristics varies among parties. We focus on the six main parties in Figure 7 from the more right-wing to the more left-wing (from top to bottom). Namely: *Fratelli d'Italia* (Populist – Far right - Eurosceptic), *Lega Salvini Premier* (Populist – Far right - Eurosceptic), *Forza Italia* (Centre-Right), *Movimento 5 Stelle* (Populist – Eurosceptic), *Partito Democratico* (Centre-Left) and *La Sinistra* (Left).[34]

Looking at the coefficients in Figure 7, we acknowledge that most coefficients are not significantly different from zero at the 95 per cent confidence interval. Despite the feeble significance, we can depict some trends over the political spectrum (from the right to the left-wing). For instance, we can observe a declining role of Political experience (except *La Sinistra*), a decreasing magnitude of Gender Bias (except for *Fratelli d'Italia*), an increasingly positive role of education in explaining candidate's success.

The exception of the coefficient of Female**Fratelli d'Italia* is worthy of discussion. Despite being the Party with the more extreme right-wing position in the Italian political spectrum, women candidates of this Party appear to be less discriminated than those of all other parties, including parties with a traditionally strong position on feminist issues (such as *Partito Democratico* and *La Sinistra*). This finding might be partly explained by the presence, among the candidates, of the party leader Giorgia Meloni, a 44 years old female politician who has run the Party since its foundation in 2014.

From a statistical perspective, excluding Giorgia Meloni from the analysis (results reported in Appendix Table A-3) reduces the magnitude of the effect but not the main results – namely, that the coefficient is larger than those of all the

[34] We adopt simple definitions given by the PopuList Version 1.0 https://popu-list.org/about/

other parties. We see here the potential of a role-model effect caused by a charismatic female leader on other female candidates and the choice of their voters. This role-model effect seems to trump the ideological stances taken from the left-wing parties.

Figure 7: Regression coefficient plot of Table 6

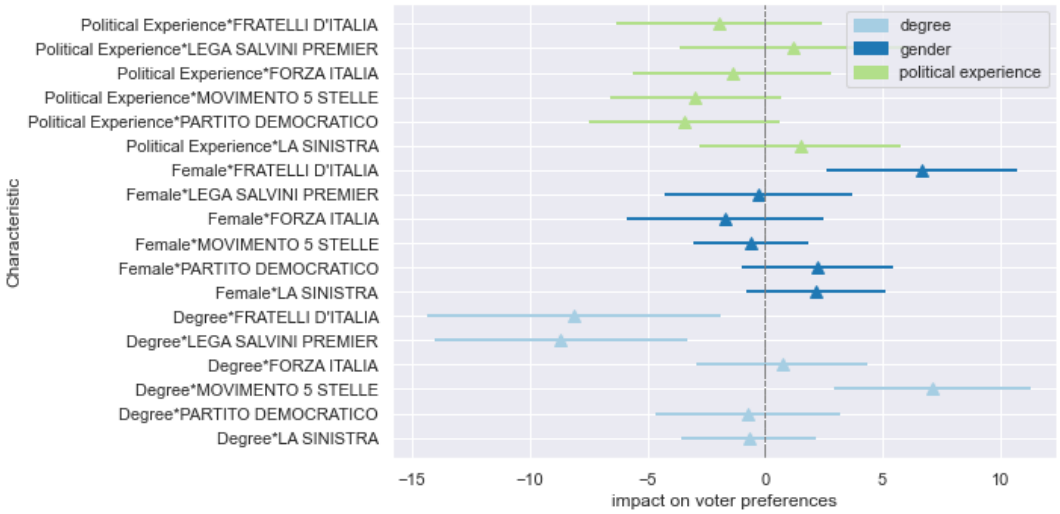


Table 6: Interaction between characteristics and party

Dep.Var. Preference Votes (%)	I	II	III
Party Interaction with	***Degree**	***Political experience**	***Female**
La Sinistra	-0.69	1.50	2.20
	(1.47)	*(2.18)*	*(1.51)*
Partito Democratico	-0.73	-3.43*	2.21
	(2.01)	*(2.08)*	*(1.64)*
Movimento 5 Stelle	7.11***	-2.94	-0.58
	(2.13)	*(1.84)*	*(1.25)*
Forza Italia	0.75	-1.39	-1.68
	(1.85)	*(2.15)*	*(2.13)*
Lega Salvini Premier	-8.69***	1.19	-0.30
	(2.74)	*(2.45)*	*(2.03)*
Fratelli D'italia	-8.12**	-1.94	6.69***
	(3.19)	*(2.22)*	*(2.07)*
N	*24052*	*24052*	*24052*
CONTROLS: Popularity (Twitter and Facebook) Party, Province, Age and Political Experience			

Notes: *,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. Dependent variable is expressed in percentage (min=1, max=100). The table reports the value of the interaction coefficients between the Party and the dimension listed in the first row. (e.g. in the example of Degree, the delta δ coefficient of the following regression: $P_{c,p} = \alpha * Degree_{c,p} + \gamma * Party_{c,p} + \delta * Party_{c,p} * Degree_{c,p} + Controls_{c,p} + u_{c,p}$)

4.5 Machine Learning Predictions

The previous sections have analyzed the impact of characteristics following a traditional linear form of model estimations. Those types of analysis have the great advantage of being easily interpretable. We can interpret the sign and the size of each character and their relationship – *ceteris paribus* – with the candidate's electoral success. However, the advantage of interpretability comes with the cost of not capturing interactions and non-linear relationships: both *within* the characteristics and *between* the characteristics and the candidate's electoral success.

Thus, this section provides an alternative approach employing models designed to tackle these issues. By doing this, we shift our lens from a descriptive task to a predictive task. We assess the predictive power of the characteristics of a candidate on her electoral success. To do so, we estimate a battery of the most popular models in the Machine Learning framework, namely Lasso, Ridge, Random Forest, and Neural Network.

We randomly split our candidates' dataset into a training set (80%) and a test set (20%). We use the training set to train and cross-validate the models while we use the test set to evaluate the predictive power of each pre-trained model. We evaluate the predictive power of each model using the Mean Squared Error (MSE) of the share of preferences votes gained by each candidate within their list. Where the MSE is computed as the difference between the value predicted by the model and the actual value.

Figure 8 and Table 7 show the results and the comparison of the MSE between a baseline "mean" model,[35] the OLS model (the model we performed in the previous section), and the more advanced ML models. To test the statistical significance of the ML models, we report results averaged over 100 repetitions of the random split of the train and test set. We note in the Appendix the details on the training and test phase, the features included for each model, and the Nadeau and Bengio statistical tests (Nadeau & Bengio, 1999).

As we can see from Table 7, using Random Forest and Neural Network models, we can reduce the MSE (and thus increase the predictive performance) of the characteristics of the candidate by ≅60% and ≅40% compared to the "mean" model and the OLS model.[36] These results show that characteristics such as gender, political experience, age, presence on the web, and all the other characteristics used in these models are good predictors of candidates' electoral performance. Employing ML models that account for non-linear and interdependent relationships, we can reasonably increase our capability of predicting electoral performances of candidates (compared to the baseline linear

[35] The “mean” model is nothing else than a model whose predicted value is simply the mean value of the actual values of the training set (e.g., the predicted value is simply equal to the mean of “past” values).

[36] 42.94 compared to 109.1 and 71.96 (see table 6).

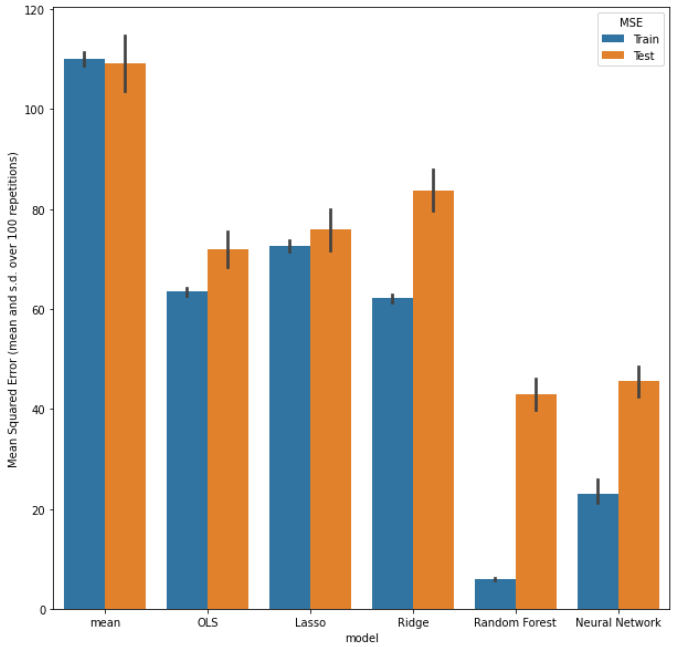
models). These performance improvements can be relevant in the application of evaluation of electoral results that use Machine Learning Control Method (MLCM) (Cerqua et al., 2021; Grimmer et al., 2021; Varian, 2016). We show that linear models fail to capture the full non-linear dependencies that characterize the political realm. While previous studies have shown the potential and limitations of the usage of Internet and social network data for predicting elections with ML models (Gayo-Avello et al., 2011; Schoen et al., 2013; Skoric et al., 2020; Tsakalidis et al., 2015), this is the first attempt that combines these types of data with actual individual characteristics of politicians.

Some limitations apply to the results of this section. First, the relatively small size of the sample (1072 candidates) might not be enough to capture the full potential of ML models. Second, by using these models, we lose the interpretability of which characteristic is more relevant and significant.[37]

Table 7

	MSE test	MSE train	MSE test - % change to OLS -	MSE train - % change to OLS -
mean	109.11	110.11	-51.6 %	-73.5 %
OLS	71.96	63.47	0.0 %	0.0 %
Lasso	75.89	72.62	-5.5 %	-14.4 %
Ridge	83.61	62.23	-16.2 %	1.9 %
Random Forest	42.94	5.97	40.3 %	90.6 %
Neural Network	45.56	23.08	36.7 %	63.6 %

Figure 8



[37] In the Appendix we report the "Feature Importance" of the Random Forest model (Figure A-10). The results clearly show that "position in the list" is the most important predictor of the electoral performance of candidates. This finding is coherent with the anedoctal evidence we show in Figure 3. However, Feature Importance of Random forest should be treated carefully when making inference, as they do not give a clear indication of the sign (plus-minus) of the relationship.

5 Conclusions

Research and public opinion seems to acknowledge that certain individual characteristics of politicians might have an impact on their electoral success and the decisions they make when in office. Although the empirical evidence in this field is growing considerably (Berz & Jankowski, 2022; Dal Bó et al., 2017; Dal Bó & Finan, 2018; Meriläinen, 2022; Portmann, 2022; Schwarz & Coppock, 2022), there are still important limits placed on access to microdata with detailed and rich information on the characteristics of politicians.[38]

Leveraging on the adoption of a new law for political transparency in Italy we created a new openly available dataset with rich and detailed information of all candidates for the European Election 2019 in Italy (note that the exact same procedure could also be applied to subsequent elections).

Using this new dataset, we showed how essential individual characteristics such as age, education, gender, and experience of each candidate might matter for the electoral success of the candidates. We also considers the positive role of social media in spreading candidates' popularity and how taking into account this role together with other traditional candidate's characteristics can lead to significantly heterogeneous results of the social media effect. For instance, being woman and being popular on the web correlates negatively with electoral success. We also show that the omission of these important factors can cause an overestimation of the education and experience premia often found in the literature.

Finally, we showed the predictive performance of these characteristics using Machine Learning (ML) models that account for interdependencies and non-linear relationships. Our ML prediction exercises show that these characteristics can be employed to refine predictions of candidates' electoral success that could be useful to predict counterfactuals.

We attach to this paper a technical appendix with a detailed description of all the steps made to create the dataset.

[38] Especially because the availability of this data is mainly circumscribed to Scandinavian countries (Dal Bó et al., 2017; Meriläinen, 2022)

6 REFERENCES

André, A., Wauters, B., & Pilet, J.-B. (2012). It's not only about lists: explaining preference voting in Belgium. *Journal of Elections, Public Opinion \& Parties*, *22*(3), 293–313.

Ansolabehere, S., & Gerber, A. (1994). The mismeasure of campaign spending: Evidence from the 1990 US House elections. *The Journal of Politics*, *56*(4), 1106–1118.

Berz, J., & Jankowski, M. (2022). Local preferences in candidate selection. Evidence from a conjoint experiment among party leaders in Germany. *Party Politics*, 13540688211041770.

Besley, T. (2005). Political Selection. *Journal of Economic Perspectives*, *19*(3), 43–60. https://doi.org/10.1257/089533005774357761

Besley, T., Folke, O., Persson, T., & Rickne, J. (2017). Gender quotas and the crisis of the mediocre man: Theory and evidence from Sweden. *American Economic Review*, *107*(8), 2204–2242.

Besley, T., Montalvo, J. G., & Reynal-Querol, M. (2011). Do educated leaders matter? *The Economic Journal*, *121*(554), F205--227.

Besley, T., & Reynal-Querol, M. (2011). Do democracies select more educated leaders? *American Political Science Review*, *105*(3), 552–566.

Braendle, T., & Stutzer, A. (2019). *Voters and Representatives: How Should Representatives Be Selected?*

Buisseret, P., & Prato, C. (2020). Competing principals? Legislative representation in list proportional representation systems. *American Journal of Political Science*.

Caselli, F., & Morelli, M. (2004). Bad politicians. *Journal of Public Economics*, *88*(3–4), 759–782.

Cassese, S. (2018). Che fare se è la democrazia a minnaciare se stessa. *Luiss Open - Corriere Della Sera*. http://open.luiss.it/2018/02/21/che-fare-se-e-la-democrazia-a-minacciare-se-stessa/

Cerqua, A., Di Stefano, R., Letta, M., & Miccoli, S. (2021). Local mortality estimates during the COVID-19 pandemic in Italy. *Journal of Population Economics*, *34*(4), 1189–1217.

Christensen, H. S., Rosa, M. S. La, & Grönlund, K. (2020). How candidate characteristics affect favorability in European Parliament elections: Evidence from a conjoint experiment in Finland. *European Union Politics*, *21*(3), 519–540.

Dal Bó, E., & Finan, F. (2018). Progress and perspectives in the study of political selection. *Annual Review of Economics*, *10*, 541–575.

Dal Bó, E., Finan, F., Folke, O., Persson, T., & Rickne, J. (2017). Who becomes a politician? *The Quarterly Journal of Economics*, *132*(4), 1877–1914.

De Luca, R. (2001). Il ritorno dei" campioni delle preferenze" nelle elezioni regionali. *Polis*, *15*(2), 227–248.

Erikson, J., Håkansson, S., & Josefsson, C. (2021). Three Dimensions of Gendered Online Abuse: Analyzing Swedish MPs' Experiences of Social Media. *Perspectives on Politics*, 1–17.

Fedeli, S., Forte, F., & Leonida, L. (2014). The law of survival of the political class: An analysis of the Italian parliament (1946--2013). *European Journal of Political Economy*, *35*, 102–121.

Freier, R., & Thomasius, S. (2016). Voters prefer more qualified mayors, but does it matter for public finances? Evidence for Germany. *International Tax and Public Finance*, *23*(5), 875–910.

Gagliarducci, S., & Nannicini, T. (2013). Do better paid politicians perform better? Disentangling incentives from selection. *Journal of the European Economic Association*, *11*(2), 369–398.

Gavoille, N. (2015). *Individuals matter: three essays on French politicians*. Université Rennes 1.

Gayo-Avello, D., Metaxas, P., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 490–493.

Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., & Bontcheva, K. (2020). Which politicians receive abuse? Four factors illuminated in the UK general election 2019. *EPJ Data Science*, *9*(1), 18.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395–419.

Gulzar, S. (2021). Who enters politics and why? *Annual Review of Political Science*, *24*, 253–275.

Hobolt, S. B., & Høyland, B. (2011). Selection and sanctioning in European parliamentary elections. *British Journal of Political Science*, *41*(3), 477–498.

Jones, B. F., & Olken, B. A. (2005). Do leaders matter? National leadership and growth since World War II. *The Quarterly Journal of Economics*, *120*(3), 835–864.

Kapoor, S., & Magesan, A. (2018). Independent candidates and political representation in India. *American Political Science Review*, *112*(3), 678–697.

Katz, R. S. (1985). Preference voting in Italy: votes of opinion, belonging, or exchange. *Comparative Political Studies*, *18*(2), 229–249.

Le Barbanchon, T., & Sauvagnat, J. (2019). Electoral Competition, Voter Bias, and Women in Politics. *Journal of the European Economic Association*.

McDonnell, D., & Valbruzzi, M. (2014). Defining and classifying technocrat-led and technocratic governments. *European Journal of Political Research*, *53*(4), 654–671.

Mechkova, V., & Wilson, S. L. (2021). Norms and rage: Gender and social media in the 2018 US mid-term elections. *Electoral Studies*, *69*, 102268.

Mechtel, M. (2014). It's the occupation, stupid! Explaining candidates' success in low-information elections. *European Journal of Political Economy*, *33*, 53–70.

Meriläinen, J. (2022). Political Selection and Economic Policy. *The Economic Journal*, *132*(648), 3020–3046. https://doi.org/10.1093/ej/ueac035

Mueller, J. E. (1970). Choosing among 133 candidates. *The Public Opinion Quarterly*, *34*(3), 395–402.

Nadeau, C., & Bengio, Y. (1999). Inference for the generalization error. *Advances in Neural Information Processing Systems*, *12*.

Nannestad, P., & Paldam, M. (1994). The VP-function: A survey of the literature on vote and popularity functions after 25 years. *Public Choice*, *79*(3), 213–245.

Paldam, M. (2008). Vote and popularity functions. In *Readings in public choice and constitutional political economy* (pp. 533–550). Springer.

Pasquino, G., & Valbruzzi, M. (2019). The 2019 European elections: A 'second-order'vote with 'first-order'effects. *Journal of Modern Italian Studies*, *24*(5), 736–756.

Pinto, G. (n.d.). *Open Candidati Europee* (DRAFT VERS). Harvard Dataverse. https://doi.org/doi:10.7910/DVN/04RIGS

Pitkin, H. F. (1967). *The concept of representation*. University of California Press.

Portmann, L. (2022). *What Makes a Successful Candidate? Political Experience and Low-Information Cues in Elections*.

Renwick, A., & Pilet, J.-B. (2016). *Faces on the ballot: The personalization of electoral systems in Europe*. Oxford University Press.

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*.

Schwarz, S., & Coppock, A. (2022). What Have We Learned about Gender from Candidate Choice Experiments? A Meta-Analysis of Sixty-Seven Factorial Survey Experiments. *The Journal of Politics*, *84*(2), 655–668.

Skoric, M. M., Liu, J., & Jaidka, K. (2020). Electoral and public opinion forecasts with social media data: A meta-analysis. *Information*, *11*(4), 187.

Stokes, D. E. (1963). Spatial models of party competition. *American Political Science Review*, *57*(2), 368–377.

Stratmann, T. (2005). Ballot access restrictions and candidate entry in elections. *European Journal of Political Economy*, *21*(1), 59–71.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626.

Tsakalidis, A., Papadopoulos, S., Cristea, A. I., & Kompatsiaris, Y. (2015). Predicting elections for multiple countries using Twitter and polls. *IEEE Intelligent Systems*, *30*(2), 10–17.

Tullock, G. (1965). Entry barriers in politics. *The American Economic Review*, *55*(1/2), 458–466.

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, *113*(27), 7310–7315.

Ward, S., & McLoughlin, L. (2020). Turds, traitors and tossers: the abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, *26*(1), 47–73.

Willermain, F. (2014). From ‘second-order’elections to ‘first-order supranational’elections: A state of play after the 2014 EU elections. *Studia Diplomatica*, *67*(3), 93–107.

Yascha Mounk. (2020). The Inverted Likability Test. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2020/01/inverted-beer-test/604282/

Does political transparency influence voting behavior? Experimental evidence from Italy

Gabriele Pinto

ABSTRACT

Does political transparency influence voting behavior? We created a website (www.elezionitrasparentiroma.it) with a set of transparency information regarding candidates for the recent municipal election of Rome, Italy, in October 2021. The collection of transparency information published on the website derived from a law to curb corruption in Italy that required parties to publish their candidates' criminal record certificates, curriculum vitae, and previous political experience. We evaluate the campaign's effectiveness by inviting a randomly selected group of voters to visit the website just a few days before the election. We find that voters who saw the website - compared to another randomly selected group who were not invited and did not visit the website - voted like their counterparts. We do not find effects on the preferences for councilors nor the probability of turnout and expression of preferences. However, voters who visited the website were significantly more likely to vote for the incumbent mayor Virginia Raggi, the Five Star Movement candidate. We read these results in light of the Five Star Movement's role in promoting political transparency in the last decade. We discuss the limitation of these results and interpretations.

The pre-registration and the whole pre-analysis plan are available here:

https://www.socialscienceregistry.org/trials/7763

1 Introduction

The participation of citizens in *political life* is the quintessential ingredient of modern democracies. Citizens exercise their power by delegating to politicians. But politicians might have divergent interests from citizens, hence, the need for political accountability. Citizens should substantially influence, monitor, and evaluate politicians where political accountability is strong.

Several challenges can prevent full political accountability. One of the main obstacles is the influence of behavioral and psychological factors. For example, some citizens may not be interested in politics or have the cognitive ability to provide unbiased and rational opinions on political and economic issues. These barriers can make it difficult for citizens to hold their elected officials accountable effectively. (Brennan, 2016; Caplan, 2011; Leiser & Shemesh, 2018; Simon, 1978).

Voters often have very poor information about politicians (Brennan, 2016; Redlawsk, 2004). At the same time, politicians might have a strong incentive to not fully disclose and manipulate information that regards them and the policies they implement. This asymmetry in the disposal of information has boosted worldwide demand for greater *political transparency*.[1]

Ideally, if we provide voters with more information about representatives, politicians could be held more accountable (Besley, 2005; Dunning et al., 2019; Fearon, 1999; Persson & Tabellini, 2012)[2].

For this reason, policy initiatives to increase political transparency have been spurred all over the world by governments and NGOs.[3] And it is with this spirit that in 2019 Italy – one of the most corrupt countries among the developed ones[4] – has adopted its first Law for political transparency to prevent and curb corruption.[5]

The new Italian Law requires all politicians that run for elections to disclose their criminal record certificate and their Curriculum Vitae (CV). The rationale behind the

[1] In principle, transparency was initially limited to the field of the battle against corruption in public administration. In recent decades, the *transparency* paradigm spilled over to all aspects of the public sphere, including the process that governs the relationship between politicians and citizenship (McGee & Gaventa, 2011).
[2] There is a general perception that more transparency, and thus more information, is always beneficial for accountability relationship. However, there are several theoretical reasoning that highlight that this might not be always the case, see for instance (Mattozzi & Merlo, 2007; Prat, 2005).
[3] See for example the Transparency and Accountability Initiative (https://www.transparency-initiative.org/) or the list of studies reported in (Banerjee et al., 2011; Bank, 2016; Pande, 2011). In section 2 we portait some examples of political transparency campaign conducted in Italy.
[4] Italy ranks 32 out of 38 OECD countries in the Corruption Perception Index, only ahead of Mexico, Colombia, Turkey, Slovakia, Hungary and Greece. (International, 2021)
[5] The whole law is literally named “sweep of the corrupt” (*spazzacorrotti*).

Law is that a higher level of political transparency should reduce corruption through a change in the political class composition.

Political transparency can impact the selection of political leaders through two main channels. The first is a "deterrence" effect, in which transparency discourages parties from nominating unqualified candidates. The second is a direct influence on voter decision-making, by providing information that updates their prior beliefs and reduces biases (Bank, 2016; Dunning et al., 2019; Kendall et al., 2015). This study will focus on the second mechanism and examine whether political transparency affects voter choices.

To make this evaluation, we organized a fictional campaign based on the transparency information requirement of recently introduced Italian Law. We exposed a randomly selected group of voters to the informational campaign (the treatment group), and we compared the vote choices of treated respondents with another randomly selected group of voters not exposed to the campaign (the control group).

Our experiment follows a vast literature that has adduced evidence about the effect of political transparency on accountability and political engagement (Bhandari et al., 2021; Boas & Hidalgo, 2011; Cruz et al., 2021; Dunning et al., 2019; Enikolopov et al., 2011; Grossman et al., 2020; Kronick & Marshall, 2018; Larreguy et al., 2018). The results hitherto are far from encouraging: in most cases, the impact of *political transparency* on voter behavior is either null or small (Bank, 2016; Dunning et al., 2019). This evidence contradicts the emphasis that Government and NGO put on the value of information and transparency, thus leaving a space for further inspection.

We identify two crucial aspects of this literature from which we deviate. First, most of this literature has focused its attention on the capability of sanctioning incumbent performance. The typical intervention provides voters with information on incumbent performance (legislative activity, efforts, etc…) to estimate the effect on performance or re-election probability.[6] Second, most experiments on the effect of information campaigns focused on developing countries. In these countries, a general lack of freedom of expression (and transparency) often limits information's baseline availability (Platas & Raffler, 2021).

We follow this literature, but we differ on these two main aspects: the type of information we provide to voters for the intervention and the context of the experiment. First, we focus on the capability of voters to evaluate ex-ante the

[6] This is the case for example in the majority of the studies in the Metaketa I initiative (Dunning et al., 2019), as well in other (Bhandari et al., 2021; Chong et al., 2015; Dunning et al., 2019; Malesky & Schuler, 2020; Pande, 2011; Pande et al., 2012, 2014). Table A-1 and Figure A-1 in the appendix report the complete list of studies taken into consideration and their geographical coverage.

potential of all the candidates (not focusing on the incumbent), adapting to a context where there is a large number of candidates.[7]

Second, we move the focus of the investigation to the municipal elections of a Western democracy: Rome, Italy, in a context where the availability of baseline information about candidates is supposedly less scarce than in developing countries.

Specifically, we evaluate the effect of providing voters with information about candidates' curricula vitae. The information set comprises basic information deducted from candidates' curriculum vitae (education, work, and political experience). While this type of information might be less relevant than performance reports, we motivate our choice with some considerations.

First, these characteristics (education, work, and political experience) can impact voter behavior since voters can use these traits as shortcuts to evaluate candidates (Campbell & Cowley, 2014; Mechtel, 2014; Portmann, 2022). *Second*, the information we provide is very basic: it does not require any kind of political sophistication. *Third*, we can evaluate the impact on candidates with no experience (and no performance..). *Fourth*, in our case study, we use information that is, in principle, already available to the public. However, this information is difficult to collect and process. The high number of candidates (more than 1600) makes the use of this information quite difficult to summarize for the average voter. We increase the usability of this information, and we reduce the search cost by delivering voters an easy-readable summary (Redlawsk, 2004). *Fifth*, the information set is taken from the recently introduced compulsory transparency requirements introduced by a Law against corruption ("*Legge spazzacorrotti*"). This latter point ensures the non-partisanship of information. *Sixth*, the informational campaign we organized mimics some real examples of the campaign organized by civil society in Italy. We thus replicate previous realistic cases where civil society organizations used *political transparency* to influence elections.

Our experimental results show that the two groups of voters do not significantly differ in vote behavior in any measure of political engagement (turnout, expression of a vote preference) or valence characteristics of voted candidates (sex, experience, and education). Even when we consider different characteristics of voters, the results remain insignificant.

Instead, we find that treated respondents appear to be more likely to vote for the incumbent mayor: Virginia Raggi (candidate for the populist Five Star Movement). Voters in the treated group are five percentage points more likely (32 percent more than the baseline) to vote for her. While we do not have a clear mechanical explanation for this effect, we try to interpret this result with two anecdotes. First,

[7] We do not focus only on the mechanisms of "sanctioning" incumbent performance. Instead, we look at the informational effect of essential characteristics that describe the ex-ante candidate's potential. When we use this approach, we account that voters might look at elections as a selection mechanism rather than a sanctioning mechanism (Ashworth, 2012; Fearon, 1999; Mansbridge, 2009)

the Five Star Movement has played an important role in promoting political transparency in Italy. Second, we have some evidence that the page of Virginia Raggi was the most visited URL of the website.

Eventually, we report tracking statistics of the website, showing how much time respondents spent on the website and how the length of the visit correlates with the revealed evaluation on the usefulness of transparency.

All the hypotheses we tested and the design of the experiment were pre-registered in a public repository before the experiment took place.[8]

[8] https://www.socialscienceregistry.org/trials/7763

2 Background

2.1 Transparency in Italy

Back at the beginning of the 19th century, 90 years before Transparency International was born[9], Filippo Turati (socialist and anti-fascist) addressed the parliament about the importance of giving the right to access secreted documents. In his acclaimed speech, Turati formalized the idea of Transparency with the example of the glasshouse: "…*where a superior, public interest does not impose a momentary secret, the administration house should be made of glass….*"[10]

But, despite the ancient roots, it is only after one century that transparency has received systematic attention in Italy (Di Mascio et al., 2019; Galli et al., 2019). Between 2009 and 2016, in a period of increasing political competition[11], Italy adopted a number of steps toward increasing transparency in the Public Administration to reduce systematic corruption.[12] The Italian *wave of transparency* culminated with the adoption of the Italian Freedom of Information Act[13] , which has enlarged access to information further by allowing the generalized dissemination of information upon request (Galli et al., 2019).

Most of those interventions fall under the umbrella of *administrative transparency*, targeting predominantly public officials and the public administration. However, *political transparency* also became salient in the Italian public discourse in the same years, targeting representative institutions, politicians, and parties.

We could not explain the increased attention to *political transparency* without considering the role of the new entry of Italian Politics – "The Five Star Movement" – which put the issue of *political transparency* at the forefront of its populist stance against corrupted political elites. Demand for transparency touched on several aspects, including but not limited to: the activity of politicians in the parliament, party funding, lobbying, selections of key roles in government and public utilities, and even

[9] The word "Transparency" became famous thanks to the homonym anti-corrpution foundation Transparency International in 1993 (Ball, 2009) .
[10] Translated from original: "*Dove un superiore, pubblico interesse non imponga un segreto momentaneo, la casa dell'Amministrazione dovrebb'essere di vetro*" (*Atti Parlamentari - Legislatura 22 - Prima Sessione, 17 Giugno 1908*, 1908)
[11] (Di Mascio et al., 2019) provide a detailed history of the political origins of Italian transparency.
[12] Legislative Decree no. 150/2009 containing "*Provision on optimization of the productivity of public employees and efficiency and transparency of public administrations*.";
Law no. 190/2012 containing "*Provisions for the prevention and repression of corruption and illegality in public administration.*"
Legislative decree no. 33/2013 containing "*Rules about publicity, transparency and information provision of public administrations.*"
[13] Legislative decree no. 97/2016, containing "*revision and simplification of rules on the prevention of corruption, publicity and transparency,*"

Figure 1: campaign for political transparency in Italy 2013-2019

The figure reports banners from campaigns organized by civic organizations advocating political transparency in Italy between 2013 and 2019.**Top-Left:** "*La carta della candidata e del candidato Trasparente*", an initiative from three civic organizations (*Carte in Regola*, *Laboratiorio per una politica trasparente e democratica* and *Open Polis*) demanding all candidates to make transparent their biography, conflict of interests, financing of the campaign and manifesto. https://www.carteinregola.it/index.php/perunapoliticatrasparentedemocratica/carta-della-candidata-e-del-candidato-trasparente/ **Top-Right:** *"open politici"* an initiative from *Open Polis* (largest organization for transparency in Italy) launched in 2009. *"open politici"* is a large database of Italian Politicians where citizien can access freely information on career, declarations and activities of their representatives https://politici.openpolis.it/ **Bottom-Left and Right:** *"Sai chi voti"* and "Candidati Trasparenti" organized by Riparte il Futuro (now The Good Lobby) for the first time in 2013. An initiative to demand politicians to make public their curruiculum vitae, criminal records, and declaration on conflict of interests: https://www.thegoodlobby.it/campagne/sai-chi-voti/; https://www.ilfattoquotidiano.it/2018/02/09/elezioni-2018-riparte-il-futuro-lancia-la-piattaforma-che-certifica-la-trasparenza-dei-candidati/4148828/

the formation of the government.[14] All those points were keys in the political platform of the original "Five Star Movement".

At the same time, between 2013 and 2019, several civic organizations started advocating transparency issues. These campaigns, which took the form of information campaigns, targeted politicians to publicize their curriculum vitae,

[14] When the "Five Star Movement" entered first in the parliament afte the election of 2013 required (for the first and last time of Italian history) to publicly stream on the internet the consultations for the formation of the new government https://www.youtube.com/watch?v=FBftBmRaQ4M

criminal records, declaration of interests, and income statement (see Figure 1 for some examples).

This demand for political transparency culminated in adopting the first regulation establishing norms and conduct to make parties and politicians transparent: the *corrupt-sweeper* Law (*spazzacorrotti*, Law no. 3/2019). The Law introduced a set of norms that disciplined the transparency of parties[15] (and foundations behind them) together with norms requiring transparency of candidacies at all levels of government.

The Law required parties to publish all candidates' curriculum vitae and criminal records on the internet, mimicking civil society organizations' campaigns.[16]

These initiatives clearly aim to influence voter choices to increase the quality of the political class. As for our knowledge, an evaluation of these campaigns' efficacy and impact does not exist yet. Our experiment aims to empirically evaluate if such campaigns can influence voter choices and to what extent.

2.2 The context of the experiment

Our case study is the election of the Mayor (*Sindaco*) and the 48 Members of the Council Assembly (councilors, *consiglieri comunali*) of the Rome Municipality (*Assemblea Capitolina*) that took place in October 2021.

The electoral rules (Law 81 of 1993) foresee a two-round system for the mayor candidates and an open-list preferential system for the councilor candidates. In the first round, voters can cast a preference for a mayor and a list. Within the selected lists, voters can express a maximum of two preferences (of different sex) for specific candidates at the council assembly. If no mayor candidate obtains more than 50% of the votes, the two most-voted candidates go to the run-off election in the second round. Once the winner mayor candidate is determined, the lists supporting them get at least 60% percent of the seats available for councilors.

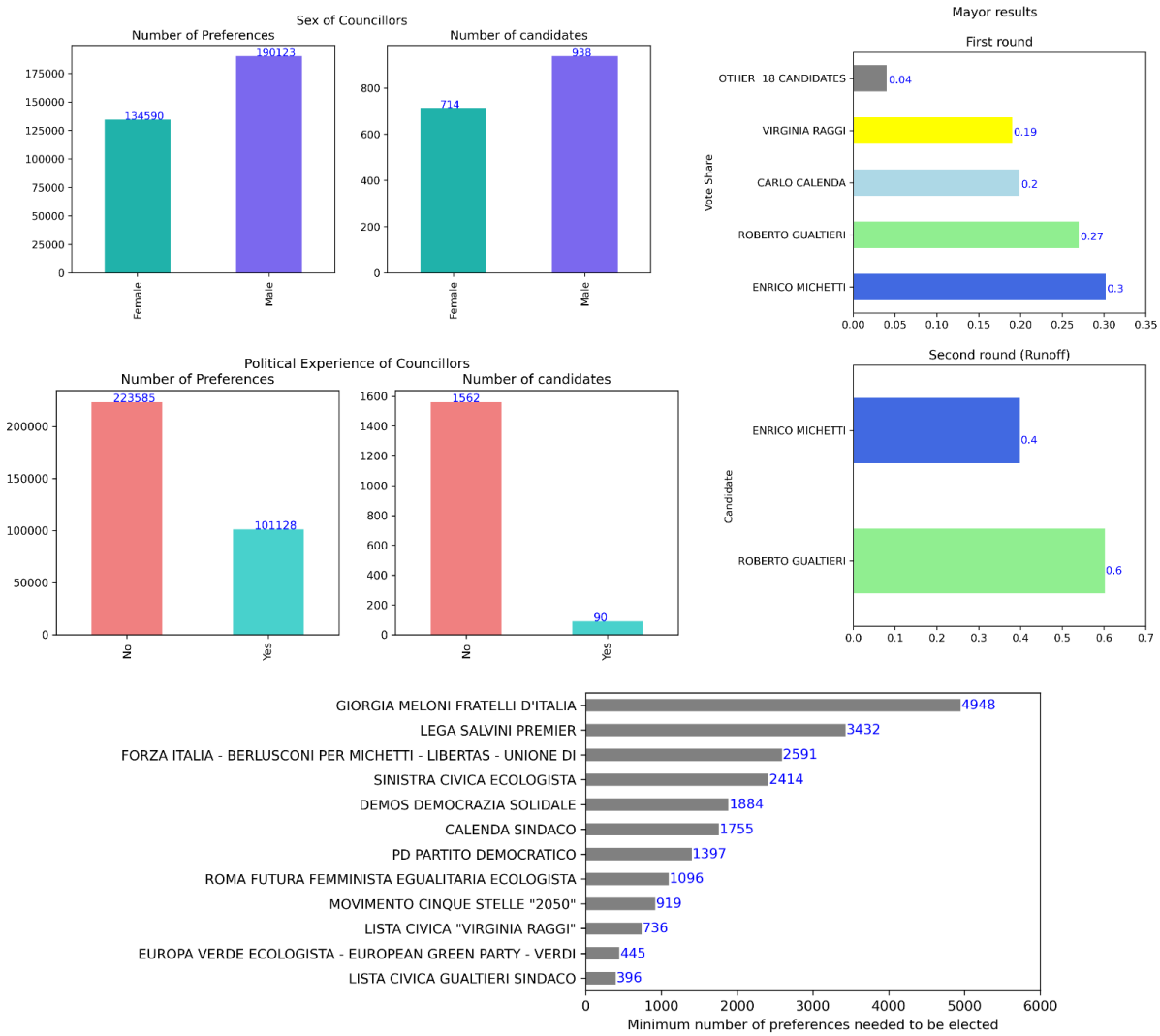
In total, there were 22 candidates for mayor (31% female), 38 lists, and 1646 councilor candidates (43% female) for the Council Assembly (48 seats in total). Roberto Gualtieri, former Minister of Economics and Finance, and candidate of a broad coalition of left-wing parties won the elections in the second round with 60.51 % of the votes. Turnout was only 48.83% at the first round and 40.6% at the second round (out of a total of 2.35 million eligible voters).

Apart from choosing the mayor, the voter could express a maximum of two preferences for the councilors. In total, voters cast 324,713 preferences for

[15] https://temi.camera.it/leg18/temi/tl18_disciplina_dei_partiti_politici.html

[16] Curriculum Vitae shall be published on the party website, and on the website of the administration where the election takes place (municipality or ministry of interior). However, conflict of interests disclosure and income statement were excluded from the transparency requirements. Note that Member of the parliament and/or of the executive are obliged (after being elected) to publish their income and patrimonial statement according another law approved in the same years: *Decreto Legge* 149/2013.

Figure 2: Rome 2021 election results



councilors (0.29 preferences per voter on average), of which 134,1590 went to women candidates (41%).

The main highlights of the election are:

- The battle for the run-off round was fought between four (out of 22) potential candidates supported by 38 lists: *Roberto Gualtieri* (the left-wing coalition, 8 lists), *Enrico Michetti* (right-wing coalition, 6 lists), Virginia Raggi (*the populist Five-Star-Movement,* 6 lists), and Carlo Calenda (*supported by a small coalition of liberal parties,* 1 list). Altogether, those four mayor candidates got 96% of the votes
- For the first time since the introduction of the run-off, the incumbent mayor (Virginia Raggi) did not achieve the second round.
- A record-low turnout in the history of the city's election (48% at the first round, 40% at the run-off)

- The threshold to enter the municipality assembly varied greatly between lists. For instance, the last candidate of the Lista Civica Gualtieri only needed 396 votes to enter the Municipal Assembly. In contrast, the last candidate of the Fratelli D'Italia (Italian Brotherhood) needed more than 4948 preferences (12 times more).
- In total, 43% of the preferences went to women candidates.
- The candidates who got the highest performance in terms of preferences among the list coalition supporting the mayor candidate were *rewarded* with an appointment as alderman (*Assessore*).[17]
-

3 Study Design

To evaluate the impact of political transparency on voting behavior, we organized a dedicated campaign on the transparency of the candidates. The campaign reconstructs information from all Curricula Vitae presented by all the candidates for transparency requirements. We retrieved the CVs from either the party website or the municipality's website.[18] Since the information comes in a long CV, we extrapolated the more salient characteristics (education, job experience, political experience) and structured this information in tabular format. We published this information into a lean website where visitors can see information about the candidates (www.elezionitrasparentiroma.). To evaluate the campaign's effectiveness, we organized a panel survey experiment where we exposed treatment group individuals to the campaign while covering the campaign to control group individuals[19].

Our panel survey experiment sample comprises eligible voters aged over 18 in the Rome area. An independent survey company administrated the questionnaire using proprietary software.[20]

[17] For instance Alfonsi (Environment), Veloccia (Urban Planning), Pratelli (Education), Funari (Social Policy) and Lucarelli (Economy and Gender Equality).

[18] The website of the Rome Municipality made available the link to download all curricula in a zip file, or, alternatively, the link redirecting the user to the website page of parties where curriculum were stored. Some minor parties did not present the curriculum of their candidates. In principle, this should result in infringement of the law and a sanction between €12.000 and €120.000 (art.1, comma 23, Law 9 Janaury, 2019, n. 3).
https://www.comune.roma.it/web/it/scheda-servizi.page?contentId=INF814326&pagina=19

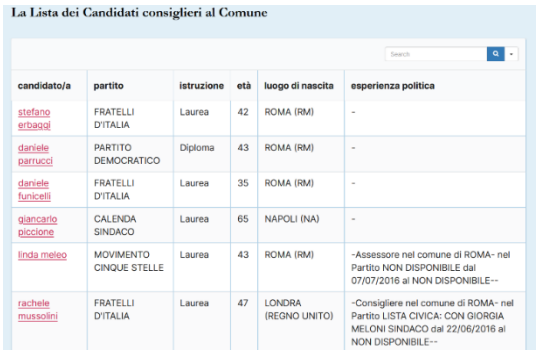
[19] Note that within our design we are estimating the effect of providing information at the individual level. Because our sample represent only a tiny fraction of the population (1000 out of 3 million) we will not have such a thing as a network effect of information (e.g. people discussing about the content of information and "spreading the word").

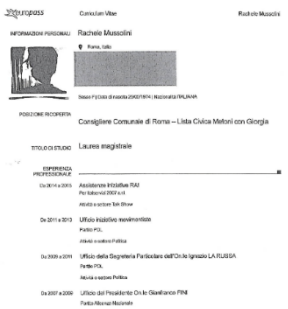
[20] Demetra opinioni srl, here the references to the recruitment contract: https://web.uniroma1.it/gareappalti/sites/default/files/TD1618287_Offerta_0.pdf; The software used is LimeSurvey. The respondents are part of a online panel managed by the survey company. The recruitment of panelists takes place via the web and landline or mobile phone. In the first case it happens through social media campaigns in the second case at the end of the telephone interview they conducted for others surveys they ask for any interest in participating in web surveys. If the answer is positive, we ask for an email address and send it the invitation. The incentive is calculated

Figure 4: Screenshot of the message inviting respondents to visit the website

Dear user, thank you for completing the survey. Now we invite you to visit this site: https://www.elezionitrasparentiroma.it/ On the site you will find neutral information on candidates regarding their curriculum vitae, level of education, their political experience, and whether or not they have a criminal record. The information was produced by the candidates themselves under Law 3 of 9 January 2019. It will not be possible to complete the questionnaire until the site is visited. Thanks again and good vote!

Figure 3: Screenshots of the website





The figure shows four screenshots of the website that treatment respondents had to visit before the election. The landing homepage (top-left) where visitors could explore the interactive table with the list of the candidates together with their party membership, education, age, place of birth, political experience (top-right) and a link to access the single candidate page (bottom-left) where they could also see their original criminal record and curriculum vitae (bottom-right). The website is still active and can be visited here www.elezionitrasparentiroma.it .

on the duration of the questionnaire: from 10 to 20 cents per minute. The median value of completion of the baseline questionnaire is 4 minutes, for the endline questionnaire is 40 seconds.

We asked respondents to fill out two online questionnaires (CAWI) before and after the election. In the first questionnaire (3-4 days before the election day), we asked about the respondents' demographics, attitude characteristics, prior beliefs, and voting intention.[21] In the second questionnaire (a week after the election day), we asked all respondents to report their vote choices'. The core of the experiment consists of submitting an informational treatment to a sub-sample of treated units that we randomly selected from the primary sample.

The information treatment is an invitation to visit a website with information about candidates through a link. This treatment is submitted just after the last question of the first survey. At that moment, the treated units saw on their screen the following message (original message in Figure 3):

By clicking on the link, the respondents entered this website www.elezionitrasparentiroma.it, where they could explore the list of candidates, together with a summary of their education profile, experience, and a link to the full CVs.

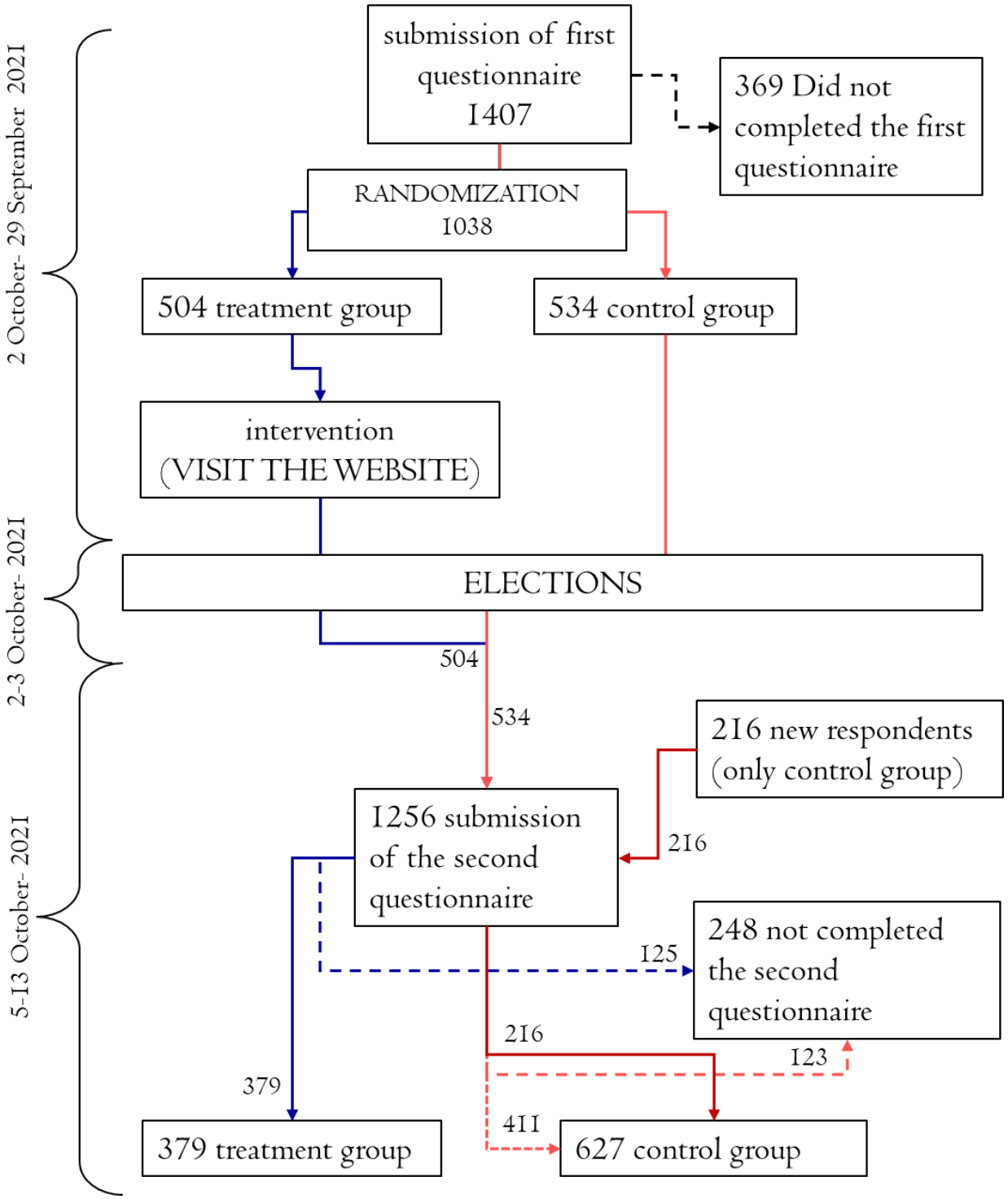
Opening the link was compulsory to successfully complete the questionnaire.

In the first round, we interviewed 1040 respondents, of which 792 also answered in the second round. We added 216 respondents interviewed only in the second round to increase the sample size.[22]

We end up with a final sample of 1008 respondents, of which 379 make up the treatment group (they received in the link in the first round), and 629 make up the control group (because they either did not receive the link in the first round, 413; or they were interviewed only in the second round, 216). Figure 2 summarizes these statistics.

[21] The full questionnaire is reported in the appendix
[22] And for controlling for the effect of being interviewed in the first round.

Figure 5: Flow-Chart of the Experiment



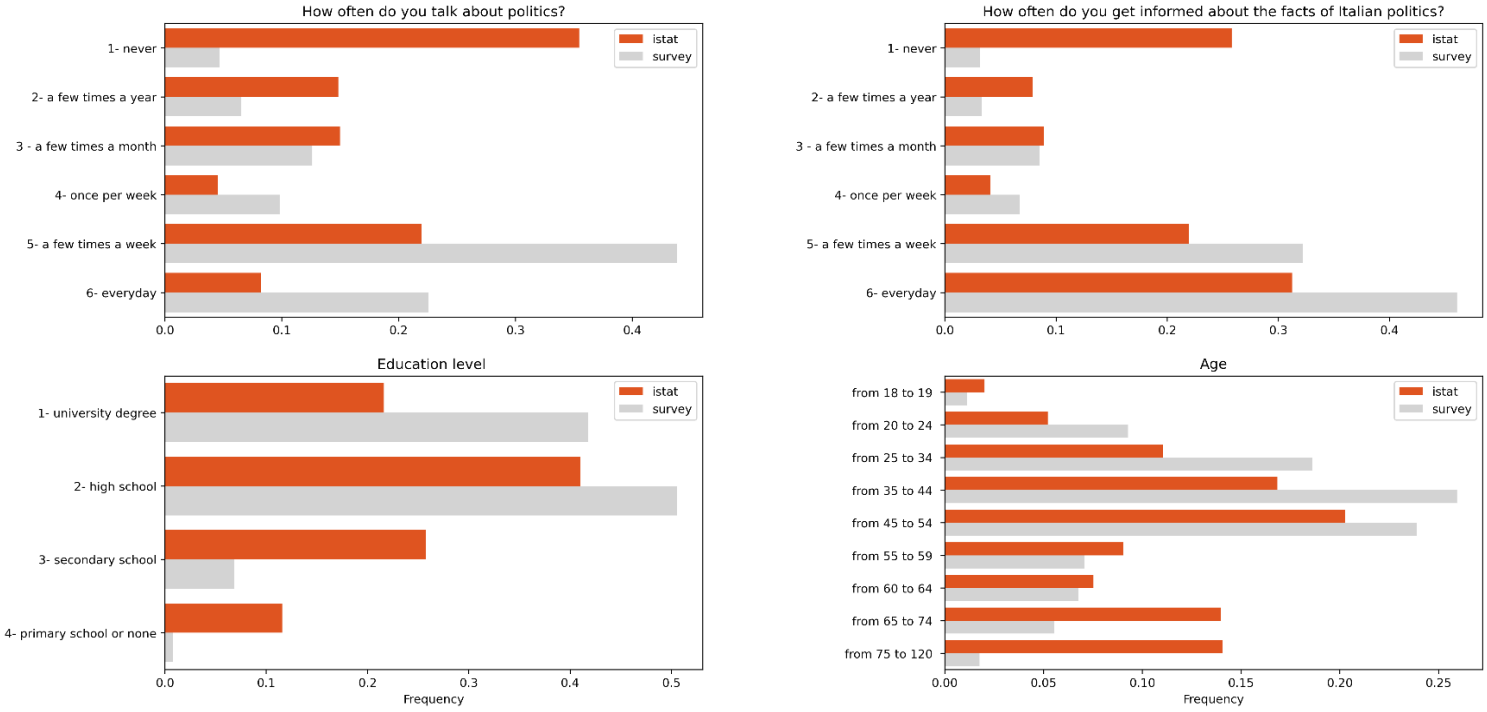
4 Data

4.1 Representativeness of the sample

One of the major concerns of online political survey sample is the related difficulty of interviewing people who are not interested in politics, do not vote, and do not use the internet (Berinsky et al., 2012; Coppock & McClellan, 2019; Enamorado & Imai, 2019; Twyman, 2008; Van Selm & Jankowski, 2006). To test the representativeness of our sample, we can compare our sample to another (more) representative sample of the general population compiled by the national institute of statistics (ISTAT). In Figure 6, we run such an exercise by comparing a number of relevant characteristics.

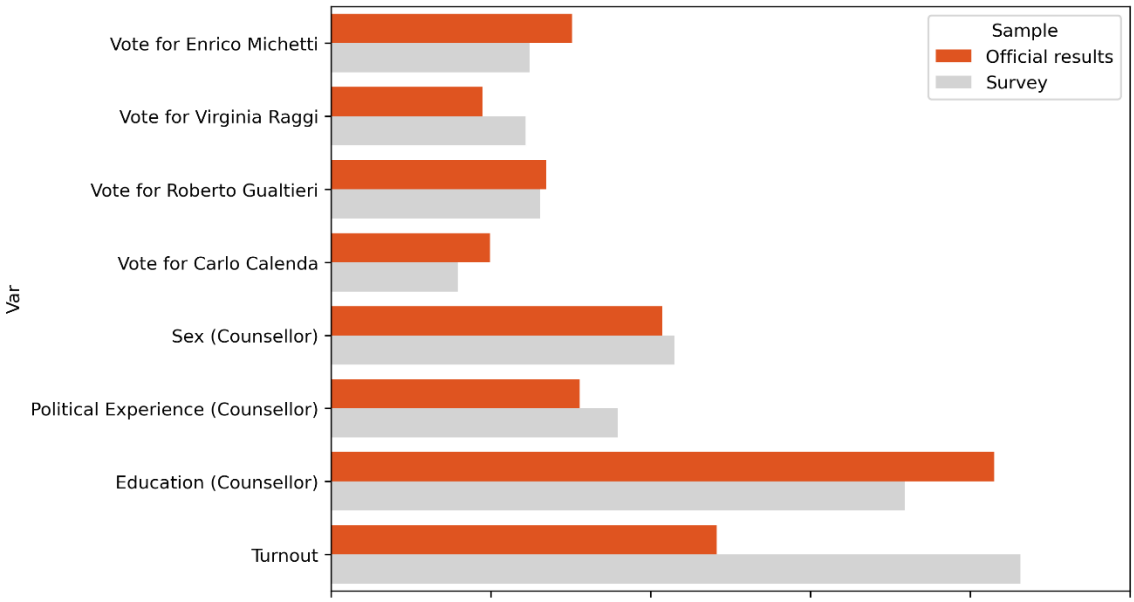
We find substantial differences between our sample and the population. The respondents of our sample get informed and speak more often about politics. Also, they are better educated and younger than the general population. Specifically, our sample significantly underestimates the presence of people over 65 years old (Figure 6 bottom-right panel).

Figure 6: Representativness of the survey sample



The figure reports answers over a set of questions measuring political participation, education level and age in the survey sample used for the experiment compared to the sample of citizens living in the Rome municipality taken from the survey of National Institute of Statistics (ISTAT) - Multipurpose survey on households: aspects of daily life - *Indagine Multiscopo sugli Aspetti della Vita Quotidiana*. https://www.istat.it/it/archivio/91926 - English version: https://www.istat.it/en/archivio/129934

Figure 7: Representativness of the vote outcome in the survey sample compared to official election results



The figure report the level of turnout, the share of votes obtained by the four main mayor candidates, female councellors, experienced counsellors, educated counsellors, respectively for the respondents of our survey compared to the official results for the whole city of Rome taken from administrative source. https://www.elezioni.comune.roma.it/elezioni/2021/amministrative/a102021/down.shtml.

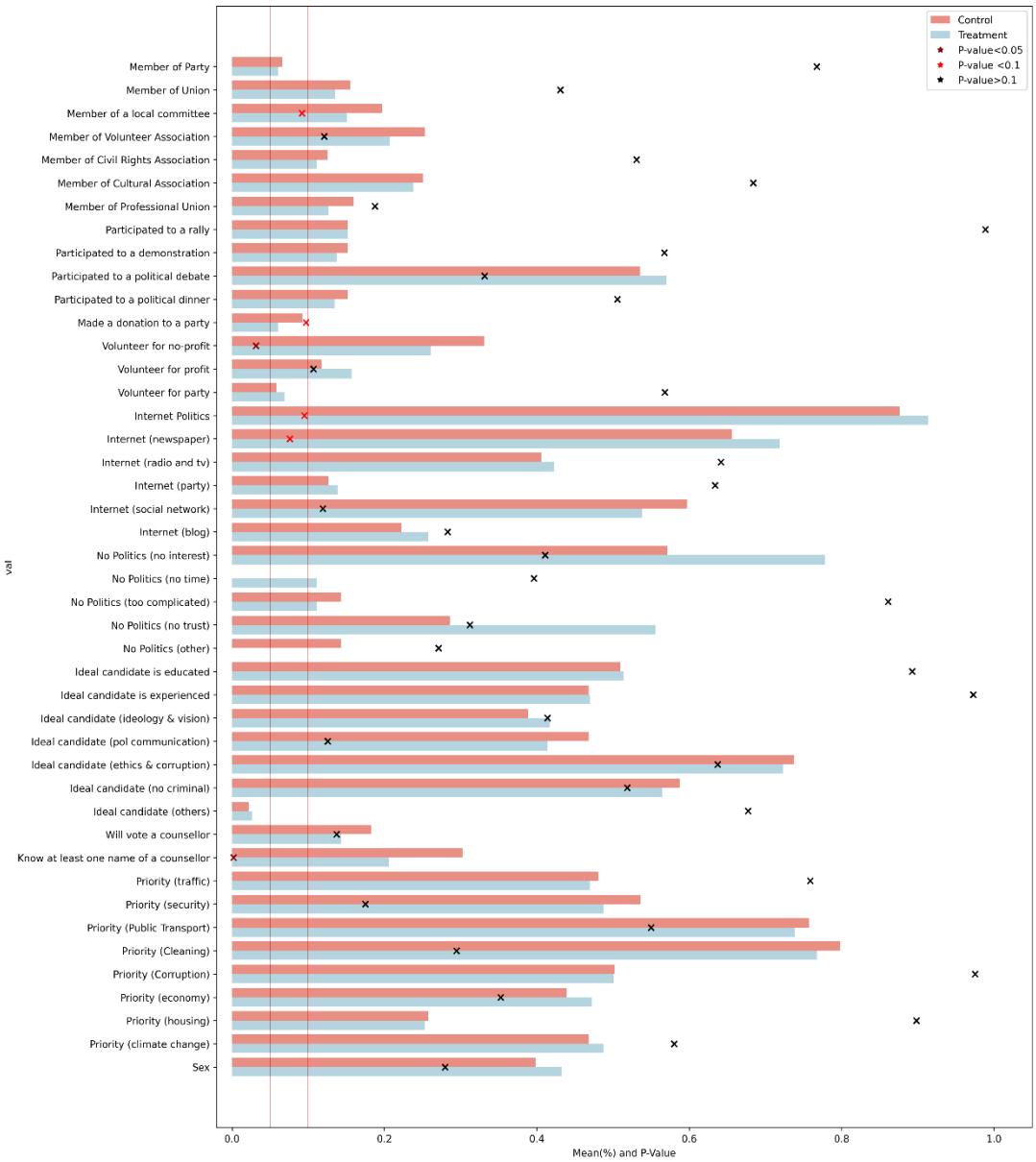
Concerning voting behavior, we can compare the vote choices reported in our survey with the actual election results coming from administrative sources. While vote choices appear to not differ much from the actual results, in line with previous findings on the validity of online surveys (Twyman, 2008), we find, not surprisingly (Enamorado & Imai, 2019), notable differences in the reported level of turnout.

These figures highlight the scarce representativeness of our sample with respect to the whole voting population, requiring caution in interpreting the external validity of our experiment.

Still, we want to highlight two factors that alleviate this limitation. First, the specific intervention we are testing is not intended to target the whole voting population. The Law foresees that the information on transparency should be published on the internet, de-facto excluding from its potential reach all citizens that do not use the internet, that, in most cases, belong to the elderly group of the population. In Figure A-4 in the appendix, we run the same exercise by only considering as the benchmark population a reduced sample of the population who use the internet every day. Compared to this reduced population sample, the differences are significantly smaller.

The second factor we want to highlight is specifically related to the differences we observe in political participation. Notably, in a period of elections, people discuss and get informed more often about politics and are more likely to be involved in political activities. We conducted our survey just a few days before the elections took place. Thus, the timing of the survey can partially explain the high level of political participation we observe in our sample.

Figure 8: Balance test

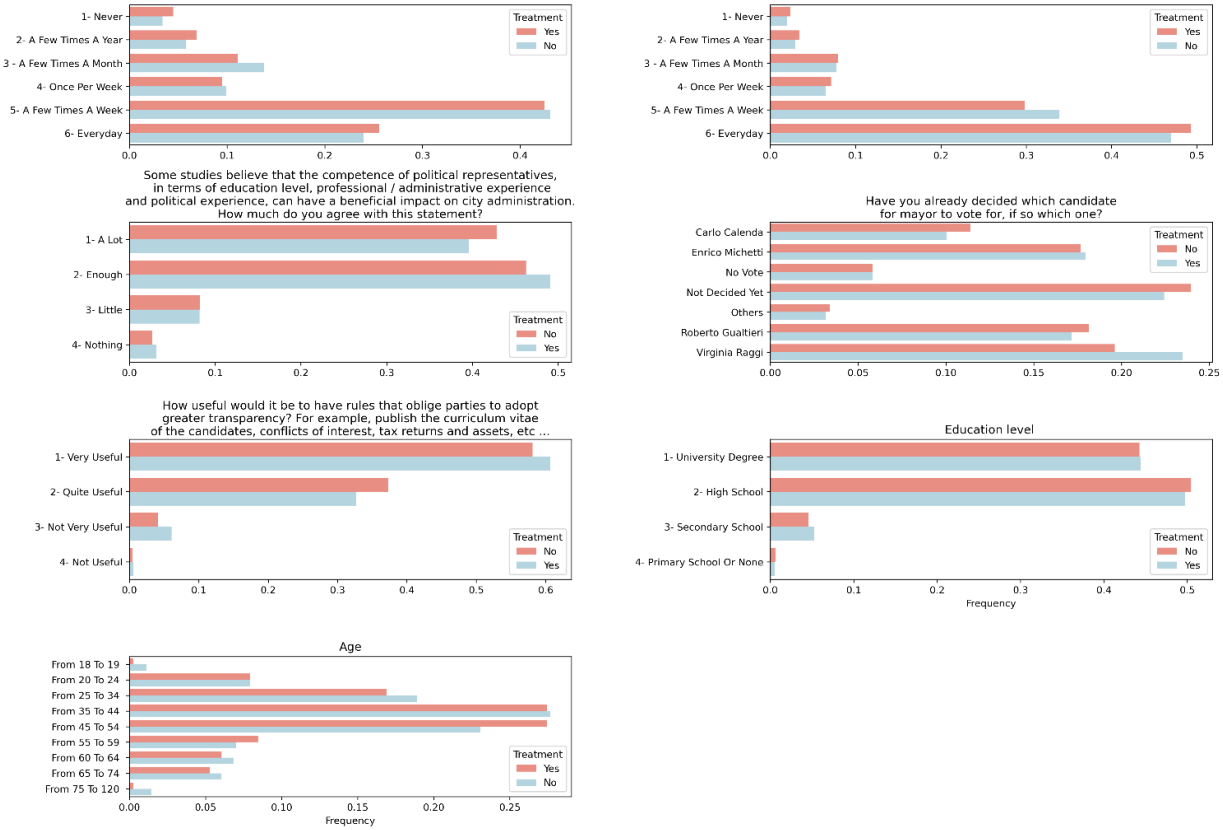


The bar in the figure report the mean value of characteristics measured in the baseline questionnaire for the treatment and the control group. All the variables have a dummy response (Yes/No). The markers report the estimate of the P-value computed with a T-Test for the significance of the difference of the mean in the two groups.

4.2 Balance test between treatment and control group

Before the election, we submitted a questionnaire to all respondents to measure how they get informed about politics, the degree and type of political participation, vote intention, priority issues, and demographic characteristics (age, sex, and education). While our randomization procedure is, in theory, independent from these characteristics, we check if that is the case empirically. Figures 8 (dummy variables) and 9 (categorical variables) show the frequency of characteristics of respondents belonging to the treatment and control group, with the p-value reporting the significance of the differences between the two groups. We fail to reject the hypothesis the two groups are not different for 6 out of 44 characteristics (13 per cent, at 10% confidence interval).[23] Because some of these variables[24] are likely predictors of some of the outcomes of interest (expression of preferences and mayor choice) we decide to include those variables as covariates in the robustness test

Figure 9: Balance test (categorical variables)



The bar in the figure report the mean value of multiple-choice answers measuring characteristics of respondents in the baseline questionnaire for the treatment and the control group. The title of each plot report the question. The labels of the y-axis report all possible answers for each question.

[23] Note that these results are partially driven by the fact that we have many hypothesis (e.g. multiple testing problem)

[24] Namely: " Member of a local committee", "Made a donation to a party", “Volunteer for no-profit”, “Interest Politics”, "Internet (newspaper)", “Know at least one name of a counsellor”, "Not Decided yet".

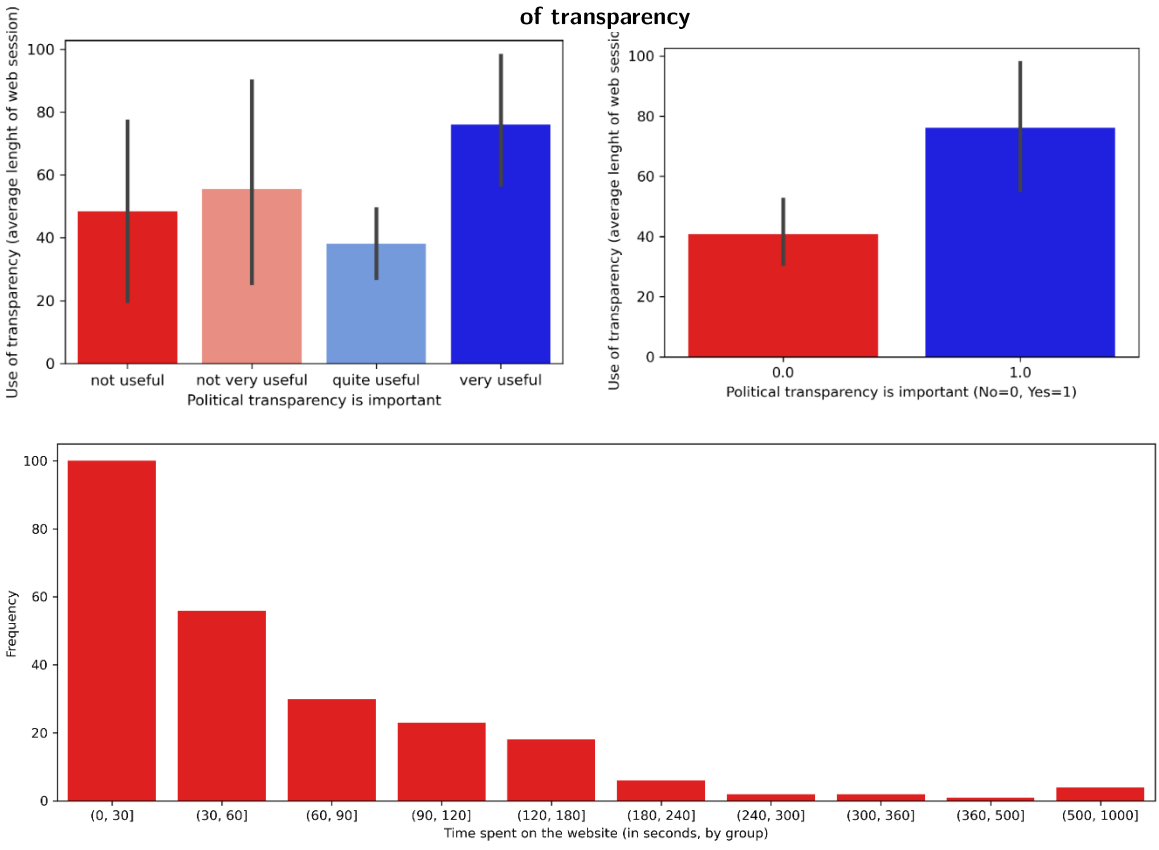
(reported in Appendix Table 2-5). Overall, we confirm that these variables predict some of the outcomes of interest, but they do not affect the main findings of the experiment.

.

4.3 Tracking and use of transparency

To check the effective delivery of the treatment, we measured if and how respondents visited the website (after opening it). To do so, we tracked the activity on the website of each (treated) respondent by using the unique link to the homepage of the website tracked by Google analytics report.[25]

Figure 10: Correspondence between revealed interest for transparency and actuale use of transparency



[25] Each respondent received a unique id and a unique link:
For example, respondents with id 1,2,3 received three different URLs: www.elezionitrasparenti.com/homepage-1/, www.elezionitrasparenti.com/homepage-2/, www.elezionitrasparenti.com/homepage-3/ While in fact these three pages are the same "copied" homepage, this trick let us being able to track each individual activity on google analytics. In fact, when looking at the traffic of the page www.elezionitrasparenti.com/homepage-2/ we can attribute this traffic only to respondent with id "2".

Figure 10 (bottom panel) shows that respondents spent 72 (average) and 41 (median) seconds on the website. The most visited pages are those of the candidates that appear on the landing page and those of mayor candidates.[26]

We also investigate if the revealed demand for transparency predicts the actual use of transparency (time spent on the website). We measure the need for transparency by asking respondents if they believe transparency is useful.[27] As figure 10 shows (top panels), people who revealed more demand for transparency also spent more time on the website. In the appendix (Figure A-6) we also show the correlation between time spend on the website and our outcome variables. We find that people that spent more time on the website have higher likelihood to vote and express a preference vote.

[26] Figure A-3 in the appendix report the top 20 most visited pages (except the homepage). Note that the candidates reported in the landing page have been selected randomly, but it does also include the mayor candidate (by chanche) Virginia Raggi (32 visits). Other mayor candidates not present in the landing page have also recorded visit of their personal page but with significant lower records (Gualtieri 7 visits, Michetti 10 visits, Calenda, 3 visits). In total, we recorded 634 clicks (excluding the landing page).

[27] The full question is: *How useful would it be to have rules that oblige parties to adopt greater transparency? For example, publish the curriculum vitae of the candidates, conflicts of interest, tax returns and assets, etc ...*the answer to the question are: "*not useful*"," *not very useful*"," *quite useful*", "*very useful*"

5 Results

We evaluate the effect of the information campaign by comparing the behavior of voters belonging to the treatment group (who visited the website) to that of the control group (who did not visit the website).

We evaluate voting behavior along with these measures: turnout, expression of a preference vote for at least one councilor, and characteristics of the voted mayor and counselor (education, political experience, and sex).

After running the baseline results, we investigate if the effect interacts with respondents' characteristics that correlate with political participation. Respondent's characteristics include education, party membership, civic participation, and political interest.

5.1 Voting behavior: turnout and expression of a preference (Hypothesis A1)[28]

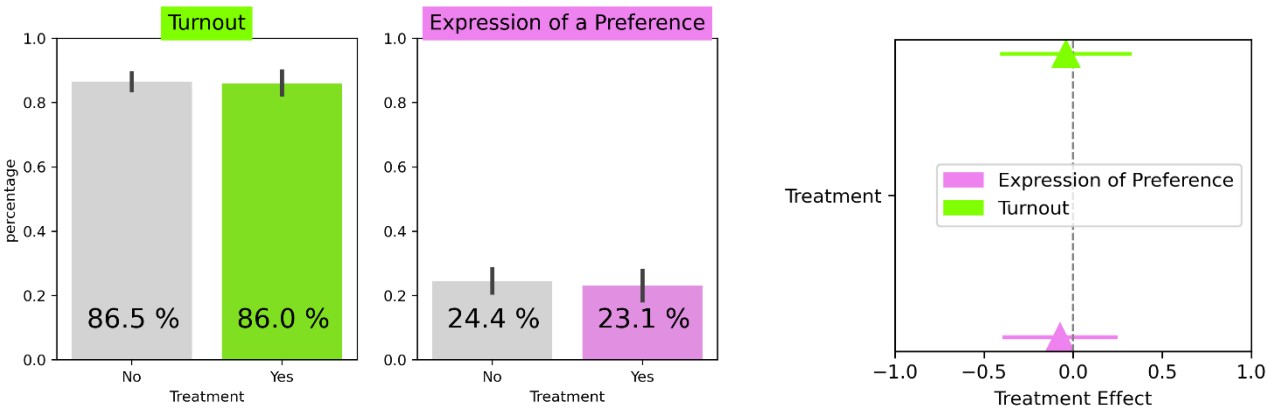
We evaluate if individuals of the treatment group statistically differ from the control group for *Turnout* and *Expression of a vote preference*.[29] Figure 11 reports the mean and standard deviation for the two groups, along with the estimated impact of treatment and the associated statistical test.

As we see, the intervention does not cause any changing behavior of voters when looking at whether they reported to go to the polls (*Turnout*) or if they opted to give a more "sophisticated" vote by casting a preference (*Expression of a preference vote*). The same results hold after controlling for prior intention to vote (Table A-3 in the appendix), suggesting that the information treatment does not affect prior beliefs.[30]

Contrary to what we expect, giving voters more resources (information) does not affect their engagement in the election (André et al., 2012; Katz, 1985; Wauters et al., 2012). These results partially contradict a strand of literature that provides evidence that political engagement increases with transparency and can increase voter turnout (Aker et al., 2017; Banerjee et al., 2011; Bank, 2016; Gine & Mansuri, 2011; Guan & Green, 2006). However, our results are consistent with mixing and contradicting evidence on the general effect of information on political engagement (Bank, 2016; Chong et al., 2015; Gentzkow, 2006).

[28] The code reported in the title of the sub-paragraphs that follow (A1, A2, etc…) refers to the corresponding section of the pre-registration.
[29] Voters can cast till two preference votes. We only consider the case where respondents report at least one preference.
[30] Because our outcome variable measuring turnout is based on reported turnout, we might be concerned that being part of the experiment might have affected the response bias (e.g. the difference between reported and actual turnout). To rule out that is the case, we compare reported turnout between the fraction of the control group which compiled the baseline survey before the election and the fraction of the control group who only did the endline survey. Results are reported in Appendix Table A-. We show that being part of the experiment does not affect reported turnout.

Figure 11: treatment effect on turnout and expression of a preference vote



	Turnout	Expression of a Preference
Treatment	-0.0397	-0.0731
	(0.1885)	(0.1654)
constant	1.8563***	-1.1309***
	(0.1166)	(0.1001)
N	1008	866

Top-Left: The height of the bar is the mean of the outcome in the group (grey bar= control group, colored bar=treatment group). Vertical line reports 95% confidence interval. Outcome variables are always dummy variable.Turnout is equal to 1 if the respondent has reported to vote. Expression of a Preference is equal to 1 if respondent has casted at least one preference vote. **Top-Right** The triangle markers show the value of the estimated coefficient of Treatment (having visited the website) on the outcome variable. The equation is a logit model ($logit(\pi_i) = a_1 + a_2 * T_i + u_i$) measured at individual respondent level *i*, where T_i is treatment, and π_i is a dummy variable measuring the outcome. The color of the marker and the legend report the outcome variable. For each outcome variable we run a separate model. **Bottom:** Regression output table where column names are the dependent variables, row names are independent variables. The second model has a lower N (866<1008) because respondent who did not voted at all are not included in the model. **Robustness results are reported in the appendix Table A-2 and A-3.**

Voting behavior: candidates' characteristics (Hypothesis A3)

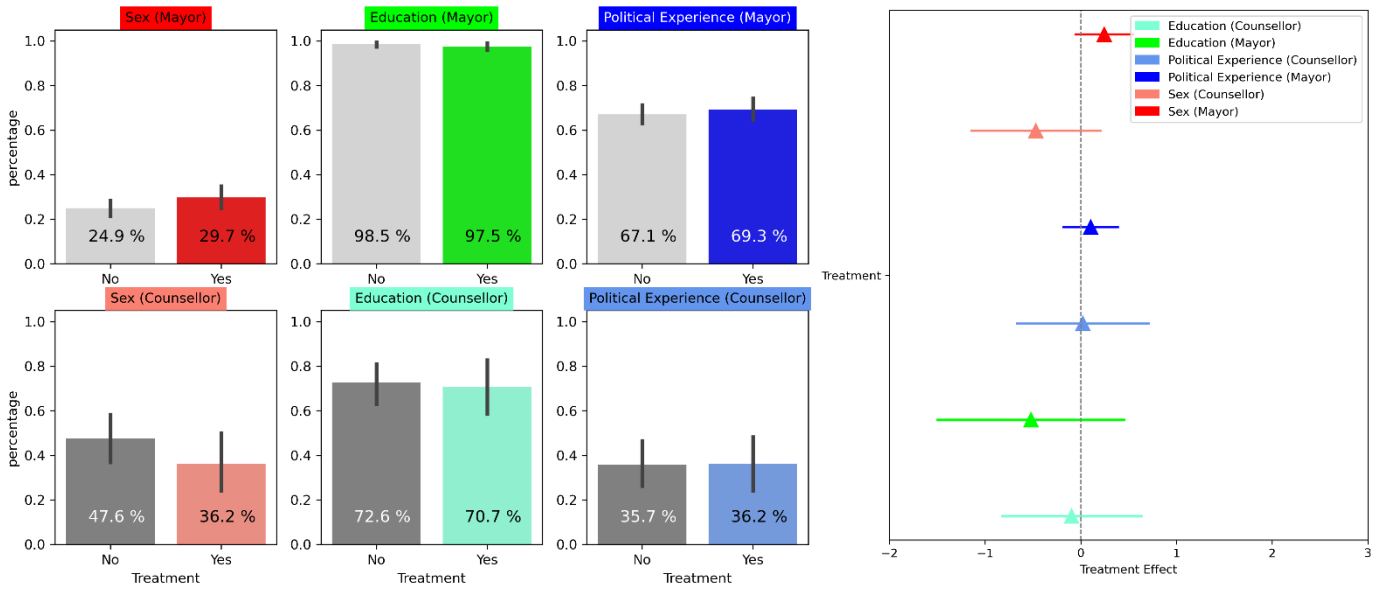
We evaluate if treatment has any relevant effect on voting for candidates with certain characteristics compared to others. We look at the following valence characteristics:

- Sex (Female=1; Male=0)[31]
- Education (University Degree=1; Else =0)
- Political Experience (At least one previous political experience =1; Else=0)

Figure 12 reports the estimation of the treatment effect size on any of these characteristics, together with the associated test of significance, for both counsellor and mayors candidates.

[31] We label the sex (male or female) of the candidate inferred from their name.

Figure 12: treatment effect on characteristics of voted candidate



	Sex (Mayor)	Education (Mayor)	Political Experience (Mayor)	Sex (Counsellor)	Education (Counsellor)	Political Experience (Counsellor)
Treatment	0.2443	-0.5199	0.1027	-0.4711	-0.0950	0.0214
	(0.1591)	(0.5052)	(0.1510)	(0.3498)	(0.3783)	(0.3557)
constant	-1.1037***	4.1705***	0.7126***	-0.0953	0.9754***	-0.5878***
	(0.1008)	(0.3563)	(0.0915)	(0.2185)	(0.2447)	(0.2277)
N	842	842	867	142	142	142

Top-Left: The height of the bar is the mean of the outcome in the group (grey bar= control group, colored bar=treatment group). Vertical line reports 95% confidence interval. The outcome variable are all dummy variables (0 or 1). Education is equal to 1 if a candidate has at least a university degree. Political Experience is equal to 1 if the candidate has at least one previous experience in running public offices. Sex is equal to 1 if the candidate is female. **Top-Right** The triangle markers show the value of the estimated coefficient of Treatment (having visited the website) on the outcome variable. The equation is a logit model ($logit(\pi_i) = a_1 + a_2 * T_i + u_i$) measured at individual respondent level *i*, where T_i is treatment, and π_i is a dummy variable measuring the outcome. The color of the marker and the legend report the outcome variable. For each outcome variable we run a separate model. **Bottom:** Regression output table where column names are the dependent variables, row names are independent variables. **Notes:** The statistical test of coefficients related to counsellor vote is less powerful because of smaller sample size (only a small fraction of voters casted a preference vote).

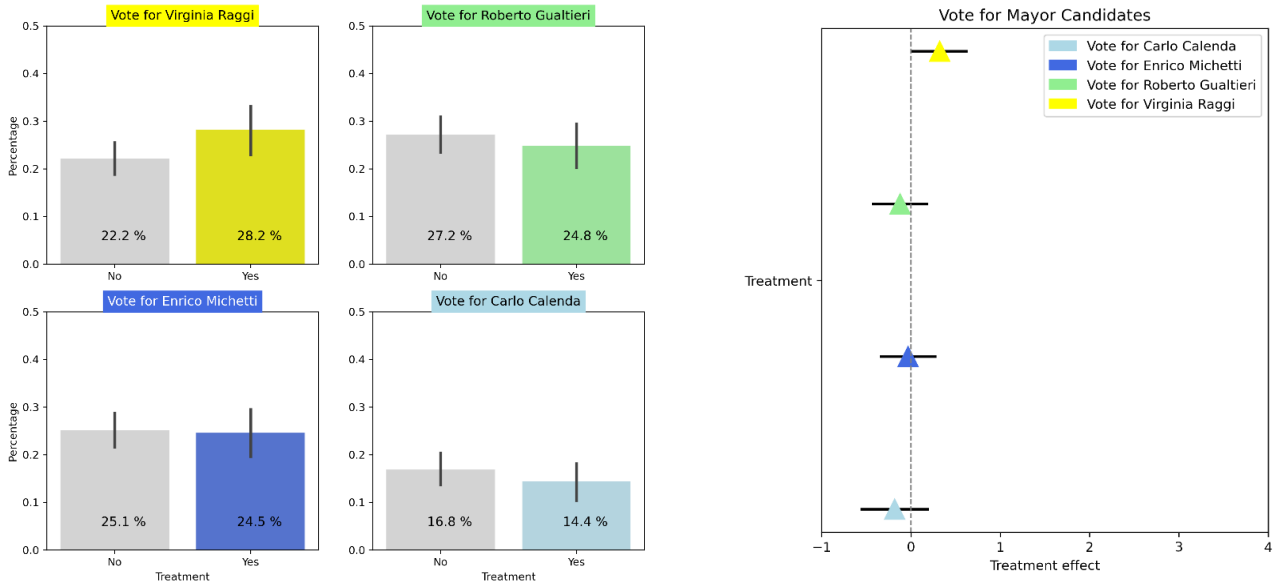
Our results are contradictory, especially when we compare the same characteristics for counsellor or mayor candidates. For instance, the coefficient of the treatment effect on the sex and the level of experience of candidates suggest that while respondents prefer female and better-experienced mayor candidates, the contrary applies to counsellor candidates (adverse effect on female and not experienced counsellor candidates). Similarly, treated respondents tend to vote more for better-educated counsellors but less for better-educated mayors. The statistical tests point towards no effect at conventional confidence intervals (e.g., p-value>0.1) in all cases.

The reading of these results suggests that the type of information provided (the curriculum vitae of candidates) does not affect voter choices.

Voting behavior: Mayor candidates (Hypothesis A2)

We estimate the effect of treatment on the probability of voting for a particular mayoral candidate. We focus on the four main mayor candidates.[32] We also include the intention to vote - measured in the baseline questionnaire - as a covariate to account for a change in prior beliefs (results reported in the appendix).[33]

Figure 13:treatment effect on mayor preference vote



	Vote for Enrico Michetti	Vote for Carlo Calenda	Vote for Virginia Raggi	Vote for Roberto Gualtieri
Treatment	-0.0321	-0.1827	0.3216**	-0.1209
	(0.1624)	(0.1951)	(0.1608)	(0.1605)
constant	-1.0912***	-1.5984***	-1.2551***	-0.9859***
	(0.0991)	(0.1149)	(0.1035)	(0.0966)
N	867	867	867	867

Top-Left: The height of the bar is the mean of the outcome in the group (grey bar= control group, colored bar=treatment group). Vertical line reports 95% confidence interval. The outcome variable is dummy variables (0 or 1) measuring if respondent voted or not for that candidate. **Top-Right** The triangle markers show the value of the estimated coefficient of Treatment (having visited the website) on the outcome variable. The equation is a logit model ($logit(\pi_i) = a_1 + a_2 * T_i + u_i$) measured at individual respondent level *i*, where T_i is treatment, and π_i is a dummy variable measuring the outcome. The color of the marker and the legend report the outcome variable. For each outcome variable we run a separate model. **Bottom:** Regression output table where column names are the dependent variables (voted for that candidate), row names are independent variables. **Notes**: Robustness test that control for the intention to vote are reported in the appendix. Sample only include respondents that voted.

[32] We exclude all other mayor candidates because the information provided has weak statistical power (only 48 respondents did not vote for the 4 main Mayor candidates)

[33] Because in our balance tests we observe that – despite the randomization of the intervention – respondents whom reported in the first round the intention to vote for Virginia Raggi are more frequent in the treatment group (but not statistically different), we decide to also include prior intention to vote for Virginia Raggi as a covariate. The coefficient appears to remain significant and positive.

People who received the intervention are more likely to vote for the incumbent mayor Virginia Raggi. The effect is statistically and economically significant ($\cong$+30% probability). Apart from being the only candidate of the Five Star Movement, Virginia Raggi was also, at the same time, the incumbent and the only female who gained a sizable share of votes. Thus, we have an overlapping (non-separable) effect of being a female mayor, incumbent, and candidate of the "Five Star Movement". For this reason, we do not have a clear-cut explanation to interpret this effect.

We mention two anecdotes that might be relevant for interpreting the specific case. First, Raggi's page was the most visited URL on the website[34], supporting the assumption that the treatment drives the effect.

Second, we want to stimulate a reflection on the perception of transparency as a non-partisan issue. Indeed, we presented the campaign to visitors as a non-partisan initiative with clearly neutral content. Still, nothing excluded that visitors might have interpreted it differently or that the campaign had a subtle partisan interpretation (*"whistling"* visitors about the importance of transparency and those parties who support it). Because the Five Star Movement has played a recognized role in promoting political transparency in Italy (as described at the beginning of this paper). We could use this to explain why the intervention has benefited only their candidate. We are cautious in saying that we do not have clear evidence of the mechanism. We leave this reflection to future investigations.

5.2 Mediator characteristics (Hypothesis A3)

Prior level of information of respondents might reduce or amplify the treatment effect. We thus study what role characteristics that describe the profile of voters (such as level of education and political participation) play when interacting with the treatment effect. For this purpose, we estimate an interaction model of the type expressed in Equation 2:

Equation 2) $$logit(\pi_i) = a_1 + a_2 * T_i + +a_3 * C_i + a_4 * C_i * T_i + u_i$$

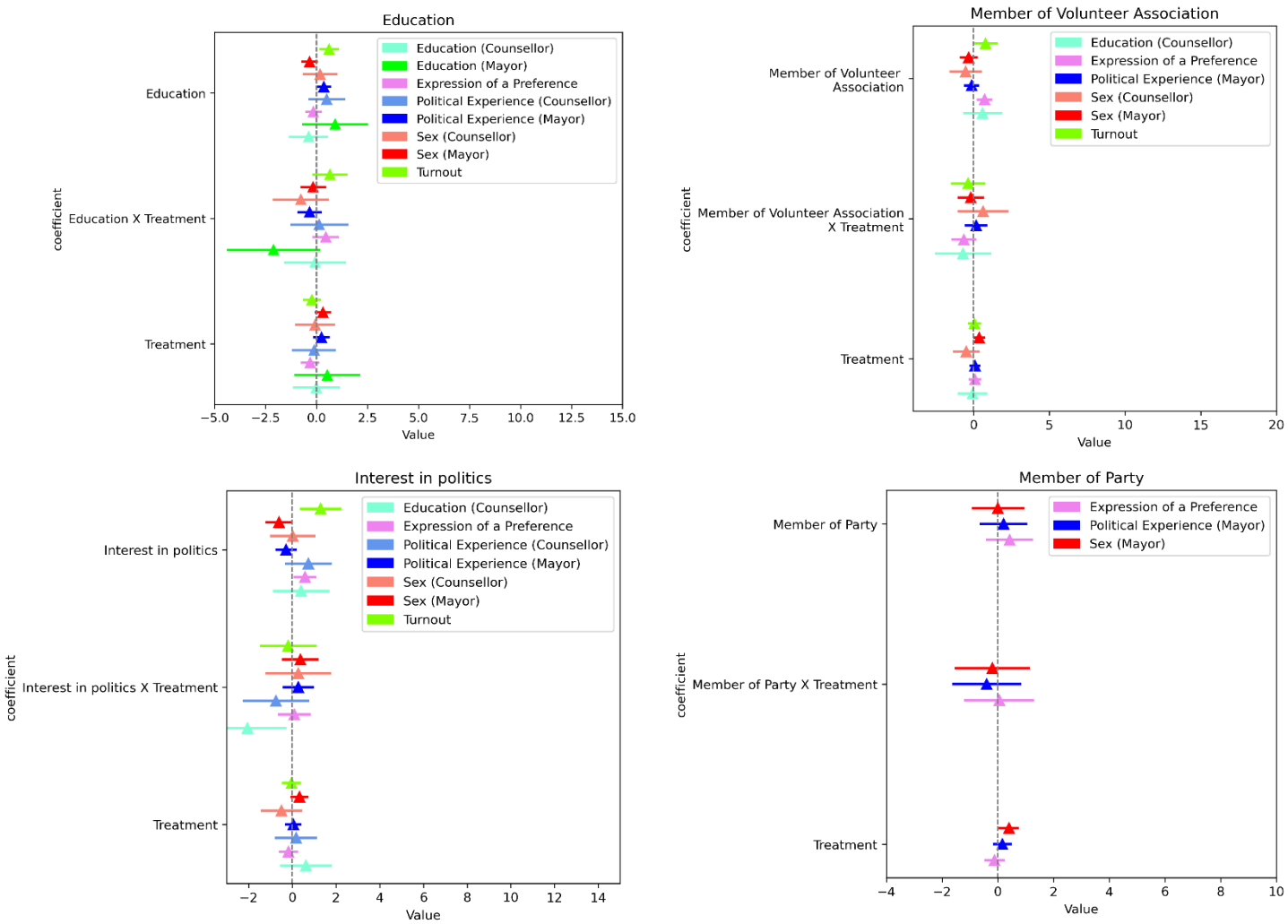
Where the coefficient a_4 gives an estimate of the size of the mediating effect of characteristic C_i on T_i .

We estimate a_4 for the following characteristic C_i of voters i.

- Member of a Political Party ("*are you a member of a party?"*)
- Interest in politics ("*how frequently you discuss politics?"*)

[34] List of most visited pages in the website is reported in Figure A-3 in the appendix

Figure 14: Interaction between voter characteristics and Treatment



The triangle markers show the value of the estimated coefficient for 1) Treatment (having visited the website) 2) the interaction between Treatment and a characteristic of the respondent, and 3) the characteristic of respondent on the outcome variable. The bar show the 95 per cent confidence interval. The equation is a logit model ($logit(\pi_i) = a_1 + a_2 * T_i + +a_3 * C_i + a_4 * C_i * T_i + u_i$) measured at individual respondent level *i*, where T_i is treatment, π_i is a dummy variable measuring the outcome, C_i is the variable measuring the characteristic of the respondent. The color of the marker and the legend report the outcome variable. For each outcome variable we run a separate model.

- Civic Participation (*"are you a member of a volunteer association?"*)
- Education (*"what is the maximum educational degree you obtained?"*)

Figure 14 shows the results of this exercise. The interaction coefficient is insignificant in most cases and points towards a null effect, with few exceptions. For instance, we observe (Figure 14, top-left panel) that while better-educated people tend to vote for better-educated mayors, the intervention vanishes their propensity (the interaction coefficient has an opposite negative sign). The same mechanism is at play when

looking at people involved in volunteer activity (Figure 14, top-right panel). Being interested in politics, or being a member of a party, does not have any relevant interaction effect on our treatment intervention (Figure 14, bottom-panels). Overall, we can state that respondents' characteristics do not significantly impact the magnitude or direction of the treatment effect.[35] Because of the limited size of the sample and the low variance of some characteristics, it is not always possible to estimate the effect of the interaction (See figure A-1 in the appendix). Low variance, and small sample size bias our significance towards the null effect.

[35] Galasso et al. (2022) show that response to information treatment might depend on respondent being populist or not. For robustness, we report in the appendix (Figure A-5) results from estimating the interaction between the treatment coefficient and a variable measuring if the respondent is populist or not. We label respondents as populist if they reveal intention to vote for a populist candidate (Virginia Raggi or Enrico Michetti). The results show that there is no interaction between treatment and being populist.

Figure 15: Interaction between voter characteristics and Treatment

	Sex (Mayor)	Political Experience (Mayor)	Expression of a Preference
constant	-1.2609***	0.6611***	-1.1149***
	(0.1354)	(0.1169)	(0.1280)
Treatment	0.4038**	0.1642	-0.1209
	(0.1863)	(0.1711)	(0.1881)
Member of Party	0.0081	0.2039	0.4218
	(0.4823)	(0.4374)	(0.4279)
Member of Party X Treatment	-0.1925	-0.4006	0.0519
	(0.6888)	(0.6313)	(0.6415)
N	659	678	679

	Sex (Mayor)	Political Experience (Mayor)	Sex (Counsellor)	Education (Counsellor)	Turnout	Expression of a Preference
constant	-1.1642***	0.7050***	-0.1112	0.9555**	1.6740***	-1.3193***
	(0.1491)	(0.1334)	(0.3338)	(0.3721)	(0.1573)	(0.1532)
Treatment	0.3841*	0.0904	-0.4835	-0.0547	0.0567	0.1016
	(0.2029)	(0.1902)	(0.4565)	(0.4966)	(0.2256)	(0.2140)
Member of Volunteer Association	-0.3201	-0.1205	-0.5174	0.6026	0.8005**	0.7348***
	(0.3072)	(0.2522)	(0.5506)	(0.6641)	(0.4003)	(0.2632)
Member of Volunteer Association X	-0.1807	0.1724	0.6421	-0.6925	-0.3621	-0.6351
	(0.4456)	(0.3873)	(0.8515)	(0.9541)	(0.5707)	(0.4119)
N	654	673	117	117	784	674

	Sex (Mayor)	Education (Mayor)	Political Experience (Mayor)	Sex (Counsellor)	Education (Counsellor)	Political Experience (Counsellor)	Turnout	Expression of a Preference
constant	-0.9555***	3.8501***	0.5569***	-0.1744	1.1575***	-0.8267***	1.6163***	-1.0578***
	(0.1316)	(0.4126)	(0.1218)	(0.2960)	(0.3457)	(0.3204)	(0.1438)	(0.1340)
Treatment	0.3174	0.5319	0.2454	-0.0668	-0.0048	-0.1178	-0.2300	-0.3137
	(0.2112)	(0.8225)	(0.2066)	(0.5000)	(0.5820)	(0.5487)	(0.2246)	(0.2342)
Education	-0.3384*	0.9163	0.3622*	0.1744	-0.3843	0.5082	0.6133**	-0.1567
	(0.2056)	(0.8213)	(0.1859)	(0.4392)	(0.4912)	(0.4589)	(0.2484)	(0.2018)
Education X Treatment	-0.1662	-2.0996*	-0.3444	-0.7661	-0.0753	0.1309	0.6588	0.4634
	(0.3244)	(1.1637)	(0.3040)	(0.7062)	(0.7726)	(0.7301)	(0.4344)	(0.3335)
N	840	840	865	142	142	142	1006	864

	Sex (Mayor)	Political Experience (Mayor)	Sex (Counsellor)	Education (Counsellor)	Political Experience (Counsellor)	Turnout	Expression of a Preference
constant	-1.1251***	0.7629***	-0.2007	1.0986***	-0.7309**	1.6403***	-1.2406***
	(0.1462)	(0.1329)	(0.3178)	(0.3651)	(0.3376)	(0.1530)	(0.1478)
Treatment	0.3302	0.0541	-0.4925	0.6242	0.1713	-0.0309	-0.1725
	(0.2052)	(0.1941)	(0.4872)	(0.6075)	(0.4949)	(0.2212)	(0.2211)
Interest in politics	-0.6053*	-0.2860	0.0183	0.4055	0.7309	1.2935***	0.5792**
	(0.3247)	(0.2504)	(0.5332)	(0.6625)	(0.5439)	(0.4838)	(0.2631)
Interest in politics X Treatment	0.3735	0.2799	0.2693	-2.0482**	-0.7466	-0.1839	0.0903
	(0.4275)	(0.3665)	(0.7664)	(0.9137)	(0.7748)	(0.6612)	(0.3836)
N	662	681	120	120	120	792	682

6 Discussion and conclusions

We evaluated the effect of a policy establishing transparency of candidates that run for election to fight corruption. We run an online experiment that uses the information that a National Law made compulsory public. Using a setting that mimics the actual deployment of the Law (a website with information about politicians), we showed that the policy does not have any impact on voter behavior.

The only significant result we found is an increase in the vote share for the incumbent mayor Virginia Raggi, a candidate of the Five Star Movement. We find some evidence that the effect is driven effectively by the treatment, given that Raggi's page was the most visited URL on the website. However, our understanding of the true mechanisms at play is not clear-cut.

We are concerned that respondents have not interpreted the initiative as neutral and non-partisan. We believe the politicization of transparency during the last decade might explain our results. Our experiment and the recent development of transparency in Italy suggest that the politicization of transparency is relevant for understanding how transparency impacts political engagement.

Indeed, previous literature has underlined that the reliability of sources and information is an important aspect when evaluating the effect of this type of intervention (Bank, 2016). With this respect, we make our information treatment as neutral as possible. The candidate's curriculum vitae have been presented *as they are*[36], and the website did not contain any evaluation or judgment of politicians. Still, respondents might have interpreted it differently. Our takeaway is that the politics behind transparency initiatives (Berliner, 2014; Di Mascio et al., 2019) should have a higher weight when evaluating its impact.

Besides, our survey shows that voters appear to be very interested[37] and likely to consume transparency information, despite being not likely to change their vote behavior. These stylized facts are consistent with the widespread adoption of initiatives aimed at increasing political transparency despite its little or no influence on voting decisions.

This study primarily deals with the effect of transparency on voting behavior. Still, we acknowledged that there might be other channels through which transparency might affect the political class selection (such as the decision to run as a candidate or the selection of candidates made by parties).

[36] and as they have been provided directly by candidates

[37] 60 per cent of respondents answered that transparency is "very useful", 35 per cent "quite useful", and only 5 per cent of them "not very useful"

7 References

Aker, J. C., Collier, P., & Vicente, P. C. (2017). Is information power? Using mobile phones and free newspapers during an election in Mozambique. *Review of Economics and Statistics*, *99*(2), 185–200.

André, A., Wauters, B., & Pilet, J.-B. (2012). It's not only about lists: explaining preference voting in Belgium. *Journal of Elections, Public Opinion \& Parties*, *22*(3), 293–313.

Ashworth, S. (2012). Electoral Accountability: Recent Theoretical and Empirical Work. *Annual Review of Political Science*, *15*(1), 183–201. https://doi.org/10.1146/annurev-polisci-031710-103823

Ball, C. (2009). What is transparency? *Public Integrity*, *11*(4), 293–308.

Banerjee, A., Kumar, S., Pande, R., & Su, F. (2011). Do informed voters make better choices? Experimental evidence from urban India. *Unpublished Manuscript*.

Bank, W. (2016). *Making Politics Work for Development: Harnessing Transparency and Citizen Engagement*. The World Bank.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.

Berliner, D. (2014). The political origins of transparency. *The Journal of Politics*, *76*(2), 479–491.

Besley, T. (2005). Political Selection. *Journal of Economic Perspectives*, *19*(3), 43–60. https://doi.org/10.1257/089533005774357761

Bhandari, A., Larreguy, H., & Marshall, J. (2021). Able and Mostly Willing: An Empirical Anatomy of Information's Effect on Voter-Driven Accountability in Senegal. *American Journal of Political Science*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/ajps.12591

Bidwell, K., Casey, K., & Glennerster, R. (2015). *Debates: Voter and political response to political communication in Sierra Leone*.

Boas, T. C., & Hidalgo, F. D. (2011). Controlling the airwaves: Incumbency advantage and community radio in Brazil. *American Journal of Political Science*, *55*(4), 869–885.

Brennan, J. (2016). *Against democracy*. Princeton University Press.

Campbell, R., & Cowley, P. (2014). What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment. *Political Studies*, *62*(4), 745–765. https://doi.org/10.1111/1467-9248.12048

Caplan, B. (2011). *The Myth of the Rational Voter: Why Democracies Choose Bad Policies-New Edition*. Princeton University Press.

Casey, K. (2015). Crossing party lines: The effects of information on redistributive politics.

American Economic Review, *105*(8), 2410–2448.

Chong, A., De La O, A. L., Karlan, D., & Wantchekon, L. (2015). Does corruption information inspire the fight or quash the hope? A field experiment in Mexico on voter turnout, choice, and party identification. *The Journal of Politics*, *77*(1), 55–71.

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research \& Politics*, *6*(1), 2053168018822174.

Cruz, C., Keefer, P., & Labonne, J. (2021). Buying informed voters: New effects of information on voters and candidates. *The Economic Journal*, *131*(635), 1105–1134.

Atti Parlamentari - Legislatura 22 - Prima sessione, 17 giugno 1908, 22962 (1908) (testimony of Camera Dei Deputati). https://storia.camera.it/regno/lavori/leg22/sed529.pdf

Di Mascio, F., Natalini, A., & Cacciatore, F. (2019). The political origins of transparency reform: Insights from the Italian case. *Italian Political Science Review/Rivista Italiana Di Scienza Politica*, *49*(3), 211–227.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., Nellis, G., Adida, C. L., Arias, E., Bicalho, C., Boas, T. C., & others. (2019). Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*, *5*(7), eaaw2612.

Enamorado, T., & Imai, K. (2019). Validating self-reported turnout by linking public opinion surveys with administrative records. *Public Opinion Quarterly*, *83*(4), 723–748.

Enikolopov, R., Petrova, M., & Zhuravskaya, E. (2011). Media and political persuasion: Evidence from Russia. *American Economic Review*, *101*(7), 3253–3285.

Fearon, J. D. (1999). Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance. In A. Przeworski, S. C. Stokes, & B. Manin (Eds.), *Democracy, Accountability, and Representation* (pp. 55–97). Cambridge University Press. https://doi.org/10.1017/CBO9781139175104.003

Galli, E., Rizzo, I., & Scaglioni, C. (2019). Transparency and socio-political environment in Italy. In *Italian Studies on Quality of Life* (pp. 97–109). Springer.

Gentzkow, M. (2006). Television and voter turnout. *The Quarterly Journal of Economics*, *121*(3), 931–972.

Gine, X., & Mansuri, G. (2011). Together we will: evidence from a field experiment on female voter turnout in Pakistan. *World Bank Policy Research Working Paper*, *5692*(7.1), 1.

Gottlieb, J. (2016). Greater expectations: A field experiment to improve accountability in Mali. *American Journal of Political Science*, *60*(1), 143–157.

Grossman, G., Michelitch, K., & Prato, C. (2020). *The Effect of Sustained Transparency on Electoral Accountability*.

Guan, M., & Green, D. P. (2006). Noncoercive mobilization in state-controlled elections: an experimental study in Beijing. *Comparative Political Studies*, *39*(10), 1175–1193.

International, T. (2021). *Corruption Perceptions Index 2020*. https://images.transparencycdn.org/images/CPI2020_Report_EN_0802-WEB-1_2021-02-08-103053.pdf

Katz, R. S. (1985). Preference voting in Italy: votes of opinion, belonging, or exchange. *Comparative Political Studies*, *18*(2), 229–249.

Kendall, C., Nannicini, T., & Trebbi, F. (2015). How do voters respond to information? Evidence from a randomized campaign. *American Economic Review*, *105*(1), 322–353.

Kronick, D., & Marshall, J. (2018). *Opposition media and political accountability: Evidence from Venezuela*.

Larreguy, H. A., Marshall, J., & Snyder Jr, J. M. (2018). Leveling the playing field: How campaign advertising can help non-dominant parties. *Journal of the European Economic Association*, *16*(6), 1812–1849.

Leiser, D., & Shemesh, Y. (2018). *How we misunderstand economics and why it matters: The psychology of bias, distortion and conspiracy*. Routledge London and New York.

Malesky, E., & Schuler, P. (2020). Single-party incumbency advantage in Vietnam: A conjoint survey analysis of public electoral support. *Journal of East Asian Studies*, *20*(1), 25–52.

Mansbridge, J. (2009). A “selection model” of political representation. *Journal of Political Philosophy*, *17*(4), 369–398.

Mattozzi, A., & Merlo, A. (2007). The transparency of politics and the quality of politicians. *American Economic Review*, *97*(2), 311–315.

McGee, R., & Gaventa, J. (2011). Shifting Power? Assessing the Impact of Transparency and Accountability Initiatives. *IDS Working Papers*, *2011*(383), 1–39. https://doi.org/https://doi.org/10.1111/j.2040-0209.2011.00383_2.x

Mechtel, M. (2014). It’s the occupation, stupid! Explaining candidates’ success in low-information elections. *European Journal of Political Economy*, *33*, 53–70.

Pande, R. (2011). Can Informed Voters Enforce Better Governance? Experiments in Low Income Democracies. *Annual Review of Economics*, *3*, 215–237.

Pande, R., Banerjee, A., Green, D., & McManus, J. (2014). Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India. *Political Communications*, *31*, 391–407.

Pande, R., Banerjee, A., & Walton, M. (2012). *Delhi’s Slum-Dwellers: Deprivation, Preferences, and Political Engagement among the Urban Poor*. International Growth Center.

Persson, T., & Tabellini, G. (2012). *Political economics: explaining economic policy*. MIT

Press Cambridge, MA.

Platas, M. R., & Raffler, P. J. (2021). Closing the gap: Information and mass support in a dominant party regime. *The Journal of Politics*, *83*(4), 1619–1634.

Portmann, L. (2022). *What Makes a Successful Candidate? Political Experience and Low-Information Cues in Elections*.

Prat, A. (2005). The wrong kind of transparency. *American Economic Review*, *95*(3), 862–877.

Redlawsk, D. P. (2004). What voters do: Information search during election campaigns. *Political Psychology*, *25*(4), 595–610.

Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review*, *68*(2), 1–16.

Twyman, J. (2008). Getting it right: YouGov and online survey research in Britain. *Journal of Elections, Public Opinion and Parties*, *18*(4), 343–354.

Van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. *Quality and Quantity*, *40*(3), 435–456.

Wauters, B., Verlet, D., & Ackaert, J. (2012). Giving more weight to preferential votes: welcome or superfluous reform? The case of the local elections in Flanders (Belgium). *Local Government Studies*, *38*(1), 91–111.

Sezioni Elettorali Italiane (SEI):
A new database of Italian electoral results geocoded at the precinct level.

ABSTRACT

Obtaining geocoded electoral results at the precinct level can be a challenging task in many countries. In this study, we present two validated methodologies for overcoming these difficulties and building a new geocoded electoral results for several Italian cities. Our dataset covers the last 20 years (1999-2022) and includes data for various types of elections, including national, regional, municipal, and referendum. After introducing our methodology, we provide an overview of some notable patterns in voting trends in major Italian cities. These include: a high level of heterogeneity in voting within cities, an increase in spatial polarization of voting behavior, and an increase in the concentration of left-wing voters in central and wealthier areas of metropolitan cities. These trends may be influenced by a range of factors and can have significant implications for political representation and policy-making. Our dataset provides a valuable resource for understanding these trends and exploring their underlying causes.

KEYWORDS: GIS, ELECTIONS, SPATIAL ANALYSIS, POLITICAL GEOGRAPHY

1. Introduction

Electoral results at the national and regional level are today relatively easy to retrieve for many countries thanks to decades of collecting efforts.[1] However, things get complicated when we need to access detailed and granular-level electoral results. Precinct-level electoral results and the associated GIS (Geographic Information System) files – which generally constitute the most granular data– are still difficult to obtain for many countries, including Italy.[2]

The main contribution of this paper is a new openly available dataset of Italian elections, which we dubbed *Sezioni Elettorali Italiane* (SEI), which comes together with a set of methodological solutions to overcome the obstacles to collecting these types of data.[3]

Precinct-level electoral results are valuable for different reasons.[4] First, a lower level of aggregation reduces the risk of incurring ecological fallacy (Alabrese et al., 2019; Robinson, 2009; Russo, 2017; Schwartz, 1994). Second, they provide an excellent alternative to using individual surveys notably affected by reporting bias (Amos, McDonald, et al., 2017; Bernstein et al., 2001; Selb & Munzert, 2013). Third, having information on the geographic residence of voters (often absent or scarce in polls) opens up the possibility of carrying out very detailed spatial analysis (Beck et al., 2006; Bivand et al., 2008; Darmofal, 2015; Mobasheri et al., 2020; Sinton & Lund, 2007; Weidmann & Schutte, 2017).

The difficulty of collecting precinct-level and geolocated electoral results in Italy (as in many other countries[5]) is caused by two distinctive characteristics. First, precinct-level data are stored in (often) poor format at the municipality level. Second, the dataset of voters' address lists that link voters to precincts - used by electoral offices to allocate voters - does not generally provide geo-encoded information.

---

[1] For International dataset of elections see the CLEA project (Kollman et al., 2019). For Italy, see the Eligendo Project (Ministero Dell'Interno, 2021).

[2] In Italy precinct correspond to the "*sezione elettorale*" and they are the smallest unit into which electoral districts are divided. Also in the United States, precinct-level electoral data were scarce, at least until few years ago. However there have been significant contributions to collect these type of data, see for example (Baltz et al., 2022; Voting and Election Science Team, 2020)

[3] We also provide:

1) A repository with all the dataset and a Python package that can be used to reproduce the study and to replicate the methodology for other cities and countries (https://github.com/gabrielepinto/dati-sezioni-elettorali) .
2) A user-friendly dashboard with interactive maps to explore the dataset (https://sezionielettorali.herokuapp.com/).

[4] This list of reasons is not exhaustive. For instance, poor data availability also affects the choice of case studies. Notably, the huge coverage of the literature of US elections – compared to other countries - might partly be explained by the rich availability of electoral results at a granular level, not to mention the availability of the voter's list file (Cooper et al., 2009) that are specific to the US and not available for many other countries.

[5] In the next section we will discuss circumstances that are specific to Italy, however Willis et al report very similar experiences for the United States (Willis et al., 2021).

As far as we know, this is the first contribution that provides a systematic methodology to build precinct boundaries and the first dataset of geocoded electoral precincts in Italy.[6]

After having presented the institutional background and the methodology to build the dataset, we provide a first exploratory analysis to provide empirical evidence of a bunch of stylized facts of the voting pattern in the three sizeable Italian metropoles (Rome, Milan, and Turin). Namely: *a)* an high level of heterogeneity of voting within cities, *b)* an increase in the spatial inequality and polarization of voting behavior, *c)* an increase in the concentration of left-wing voters in towards the center and wealthier areas of metropolitan cities (Cini et al., 2021; Dijkstra et al., 2020; Pratschke et al., 2021).

2. Background of the electoral process in Italy

The very reason why electoral precincts are not readily available needs traces back to the governance scheme of the Electoral process in Italy (similar to that of other countries[7]). Precinct-level data is collected and archived in municipal local electoral offices and not transmitted to the central archive.[8]

The governance design of the electoral process is based on a hierarchical framework where the Ministry of Interior (*Ministero dell'Interno)* shall oversee the entire electoral process implemented and organised by the local offices of municipalities (*Ufficio elettorali dei comuni*).

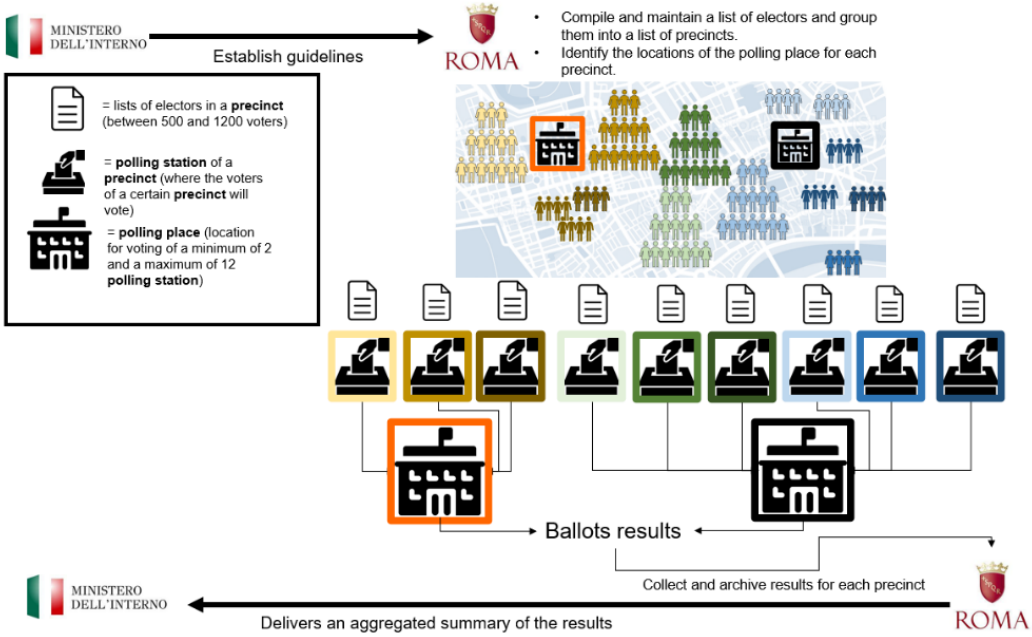
The local offices of municipalities compile the voter's lists (*liste elettorali)* and classify voters in groups of precincts (*sezione elettorale*). Voters of a precinct can only vote in the polling station (*seggio elettorale*) of their precinct located in a predetermined

[6] Indeed, there are some contributions on this issue for the United States, mainly coming from a series of publications from Brian Amos and Michael McDonald. However, none of this specifically address the methodology to reconstruct precinct boundaries (Amos, McDonald, et al., 2017; Amos, Smith, et al., 2017; Baltz et al., 2022; Voting and Election Science Team, 2020). See also: https://openprecincts.org/about/. In Italy, The most important contribution on this type of data in Italy is from Keti, Monni and Tomassi with the project "mapparoma" (https://www.mapparoma.info/). They provide electoral data (only for the city of Rome) at urban area level ("zone urbanistiche") (Lelo et al., 2021). Apart from expanding the coverage to other cities we also significantly increase the level of granularity of data that is 16 times higher (the city of Rome is divided in 155 "zone urbanistiche" and 2600 "sezioni elettorali"). Another important contribution to mention is the work from Corbetta and Piretti (Corbetta & Piretti, 2009) that have created a historical atlas of elections in italy from 1860 to 2008, however their data are generally aggregated at the city level.

[7] For instance see the obstacle to the collection of precinct data in the US (Willis et al., 2021)

[8] According to a request for clarification (*accesso civico*) the Ministry of Interior (*Dipartimento affari interni e territoriali)* is currently running an experimental programme to check the feasibility of collecting precinct results from municipalities. Insofar, no data have yet been made available from the Ministry. In absence of a centralized archive, and a lack of coordinated standards, collecting all precinct data requires inspecting 7980 archives (that's approximately the number of municipalities in Italy as of today).

Figure 1: organization of the electoral process in Italy – example with municipality of rome



polling place (*locali del seggio elettorale)*.[9] There can be from 2 to 12 polling stations in each polling place.[10]

The precincts are the most granular level for which electoral results are available. The size of the precincts ranges between 500 and 1200 voters, but it can also vary over those limits. The law foresees (very few) exceptions when the condition of distance and viability poses difficulties to the right to vote.[11] Voters belong to a precinct based on their place of residence, organized into lists maintained every six months by municipal offices following the guidelines of the Ministry of Interior.[12]

According to Italian Law, it is mandatory to cast a vote to physically going to a pre-determined polling place where the elector's polling station is located (e.g. voters cannot choose the place where they want to vote). As a norm, the polling place is set up in some selected public schools. Some exceptions relate to the particular polling station (*seggio speciale*) for prisons and hospitals, for people unable to go to

[9] To clarify, each voter is assigned to a precinct (*sezione*), and can only vote to a uniquely corresponding polling stations (*seggio elettorale*) that is located in a polling place (*locale del seggio elettorale*) where there are multiple polling station. The polling station is a unique room where only voters from a certain precinct can vote. The polling station is located in a building that constitutes the polling place, where there are multiple polling station.

[10] Art. 34 del D.P.R. del 20 marzo 1967, n. 223

[11] Art. 34 del D.P.R. del 20 marzo 1967, n. 223

[12] Circolare 2600 del Ministero dell'Interno del 1 febbraio 1986

their polling station, and for people working in the polling station (that have the faculty of voting in the section where they are working).[13]

Each polling station is administered by a small commission headed by the polling station president (*Presidente del seggio elettorale*) assisted by at least three poll clerks.[14] At the end of the vote count- held at the polling station level - the president of each polling station delivers a summary of the results to a delegate of the mayor (*delegato del sindaco*) at the electoral office of the municipality.

The results at the precinct level are stored in the electoral office of the municipality archive. Only a summary of the results is delivered to the competent department of the Ministry of Interior. Figure 1 summarize the process we just described.

3. Methodology to construct the GIS data

Data regarding the location of the precinct are not immediately available as a GIS shapefile. The Municipal Electoral Office only provides the Electoral Road Dataset (*Viario Elettorale*). [15] The electoral road dataset is a dataset that provides the ranges of addresses delimited by civic numbers that belong to a specific precinct. The Electoral Road Dataset aggregates the street/civic number level of the voter's list file stored in the municipality archive.[16] Figure 2 shows a preview of the electoral road dataset for the municipality of Rome. As we see, the electoral Road Dataset only provides the address in the form of a string and not in the actual coordinates. Thus, our first step is to get the coordinates of the addresses.

To do so, we geo-encoded the addresses strings for each address using a geocoder.[17] After collecting all the geocoordinates of all the over 30,000 addresses

[13] Art. 40 del D.P.R. 16 Maggio 1960 n.570; art.48 del D.P.R. 30 marzo 1957, n.361; art. 51 del D.P.R. 30 marzo 1957, n. 361

[14] Both the president and the poll clerks are nominated among citizens enrolled in a register ("albo degli scrutatori" and "albo dei presidenti di seggio")

[15] Some municipalities do offer the shape file of electoral precinct. Among the largest, Firenze and Rimini.
Firenze: https://opendata.comune.fi.it/metarepo/categorylist?q=metarepo/datasetinfo&id=3870c875-d92d-41dd-b211-2f3433da8dfd
Rimini: https://opendata.comune.rimini.it/dataset/sezioni-elettorali1/resource/8a0ca36e-d001-4ecb-a0be-deb30bd3c0f7.

[16] For this reason, the same procedure we present here could be reproduced by aggregating voter's list file (*liste elettorali)*. The voter's list file of Italian municipalities should be accessible for research purposes according to the Italian law. However, in practice is difficult to collect these files in a digital format. We tried to request the voter's list file to the electoral office of Rome without success.

[17] Using as an example the address reported in the Figure 4, for the first observation, we send the query with the name of the street to the API of the geocoder "Via Abano Terme, Roma, Italy", and the API returned us the location of the street with latitude and longitude. In some cases, a unique street correspond to multiple precinct depending on the street number. For instance, in the example above (Figure 2) , "Via Abbiate Grasso" correspond to multiple precincts: odd street number from 1 to 61 vote and even number from 2 to 68 vote in precinct 2392, while all others residents of "Via Abbiate Grasso" vote in precinct 2391. In those cases, we sent four different queries to the API: "Via Abbiate Grasso 1", "Via Abbiate Grasso 2", "Via Abbiate Grasso 68", "Via Abbiate Grasso 61" for precinct 2392, while "Via Abbiate Grasso 70" and "Via Abbiate Grasso 63" for precinct 2391.

Figure 2: The electoral Road Dataset

	Municipio (Sub-Municipality)	Sezione (Precinct)	Descrizione Strada (Address)	da_civico (from street number)	a_civico (till street number)	tipo (street number type: odd/even)
0	14	2357	VIA ABANO TERME	all	all	tutti i civici
1	7	1126	VIA ABASCANTO	all	all	tutti i civici
2	12	2530	VIA ABATE UGONE	all	all	tutti i civici
3	15	2440	VIA ABBADIA SAN SALVATORE	all	all	tutti i civici
4	14	2392	VIA ABBIATEGRASSO	1	61	civici dispari
5	14	2391	VIA ABBIATEGRASSO	63	end	civici dispari
6	14	2392	VIA ABBIATEGRASSO	2	68	civici pari
7	14	2391	VIA ABBIATEGRASSO	70	end	civici pari
8	14	2391	VIA ABBIATEGRASSO	all	all	senza numeri
9	14	2391	VIA ABBIATEGRASSO	all	all	tutti i lotti
10	5	535	VIA DELLE ABELIE	all	all	tutti i civici
11	5	501	VIA DEGLI ABETI	1	49	civici dispari
12	5	528	VIA DEGLI ABETI	51	end	civici dispari
13	5	501	VIA DEGLI ABETI	2	48	civici pari
14	5	528	VIA DEGLI ABETI	50	end	civici pari

in the electoral road dataset, we are left with a nebula of geo-encoded points (latitude and longitude). Each point-address we obtain belongs to a unique precinct.

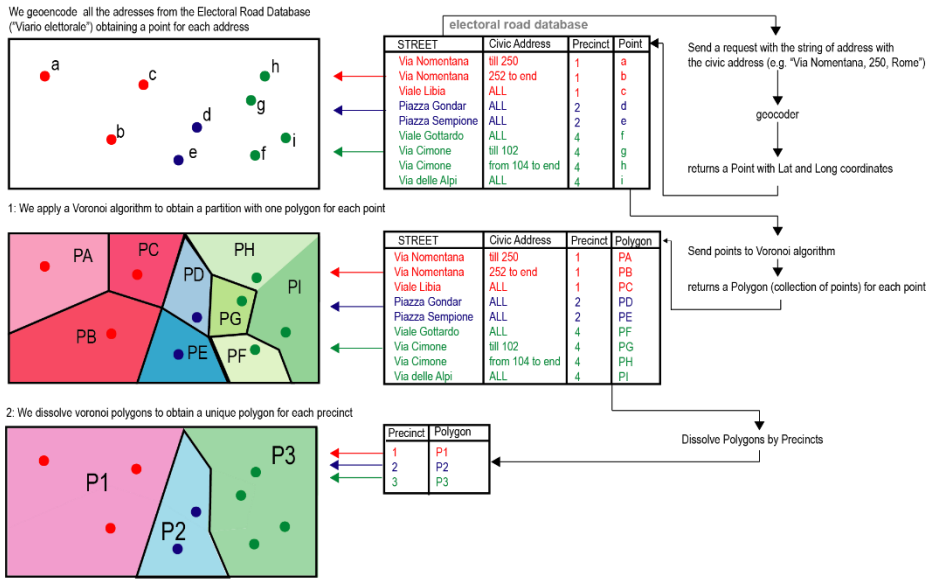
After obtaining the nebula of geo-encoded address-points, we consider two different options to reconstruct the area of precincts. The first option relies on using the Voronoi Diagram (Burrough et al., 2015, p. 160). The second option relies on the use of census blocks.

Voronoi Precinct

The algorithm of the first option (that we call *Voronoi precinct*), illustrated extensively in Figure 3, is the following:

1. For each point, compute the Voronoi area to get a set of polygons: one polygon for each point.
2. Dissolve all the Voronoi polygons among precincts and obtain one unique polygon for each precinct.[18]

Figure 3: Voronoi Procedure



[18] This procedure is partly the result of a discussion with Prof. Filippo Celata to whom I am grateful.

Census Precinct

The algorithm of the second option (we call it *Census precinct*) is the following:

1. Overlap the shapefile of census blocks on the nebula of points
2. For each census block, compute the frequency of points-per-precinct it contains
3. Assign each census block to the precinct with the highest frequency (calculated in the previous step)
4. After applying steps 1-3, we will have that some precincts remain unassigned (because of the low frequency of their points), and some blocks will remain unassigned (because they do not have points on their surface).
   To assign all precincts to at least one census block, we run the following loop:
   4a. Select all census blocks that have been assigned to a precinct that has already been assigned to more than one census block (>1)
   4b*.* Within the list census obtained from step 1, we look up if any of those census blocks contains at least a point that belong to an unassigned precinct we assign it to that precinct.
   4c. Repeat from ***4a*** till when all precincts have been assigned to at least one census block
5. To assign remaining unassigned census blocks to at least one precinct, run the following loop:
   5a. For each census block, get a list of all the precincts assigned to the neighbouring census blocks.
   5b. Assign the census block to a randomly chosen precinct from the list obtained in step 1
   5c. Repeat till when all census blocks have been assigned (we need this because there might be census blocks that are isolated).
6. Dissolve all census blocks by precinct

In both cases (*Voronoi* and *Census* precinct), we are left with a collection of polygons (one for each precinct). In Figure 4, we show an example of applying the two procedures to actual data.

Figure 4: Census procedure

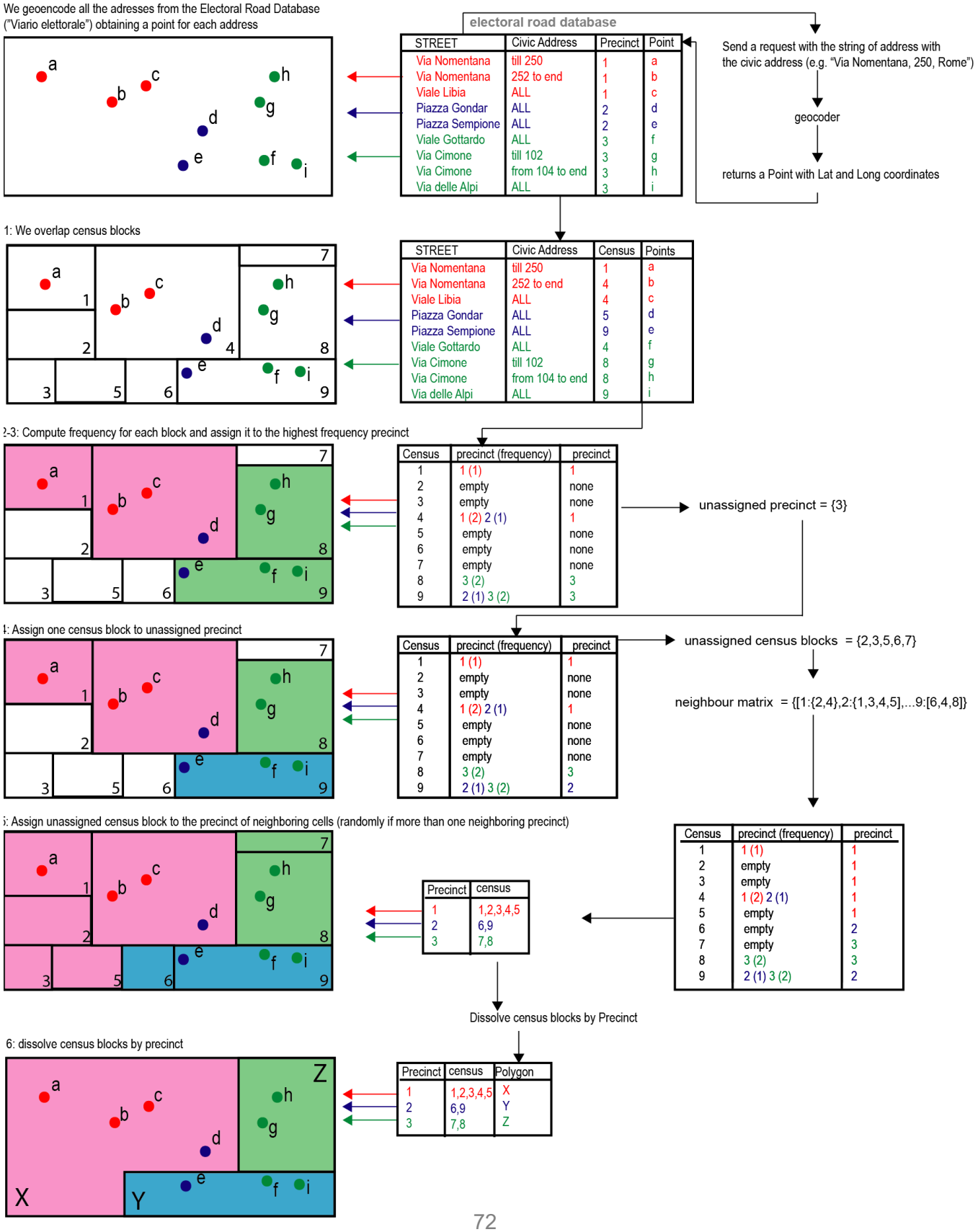
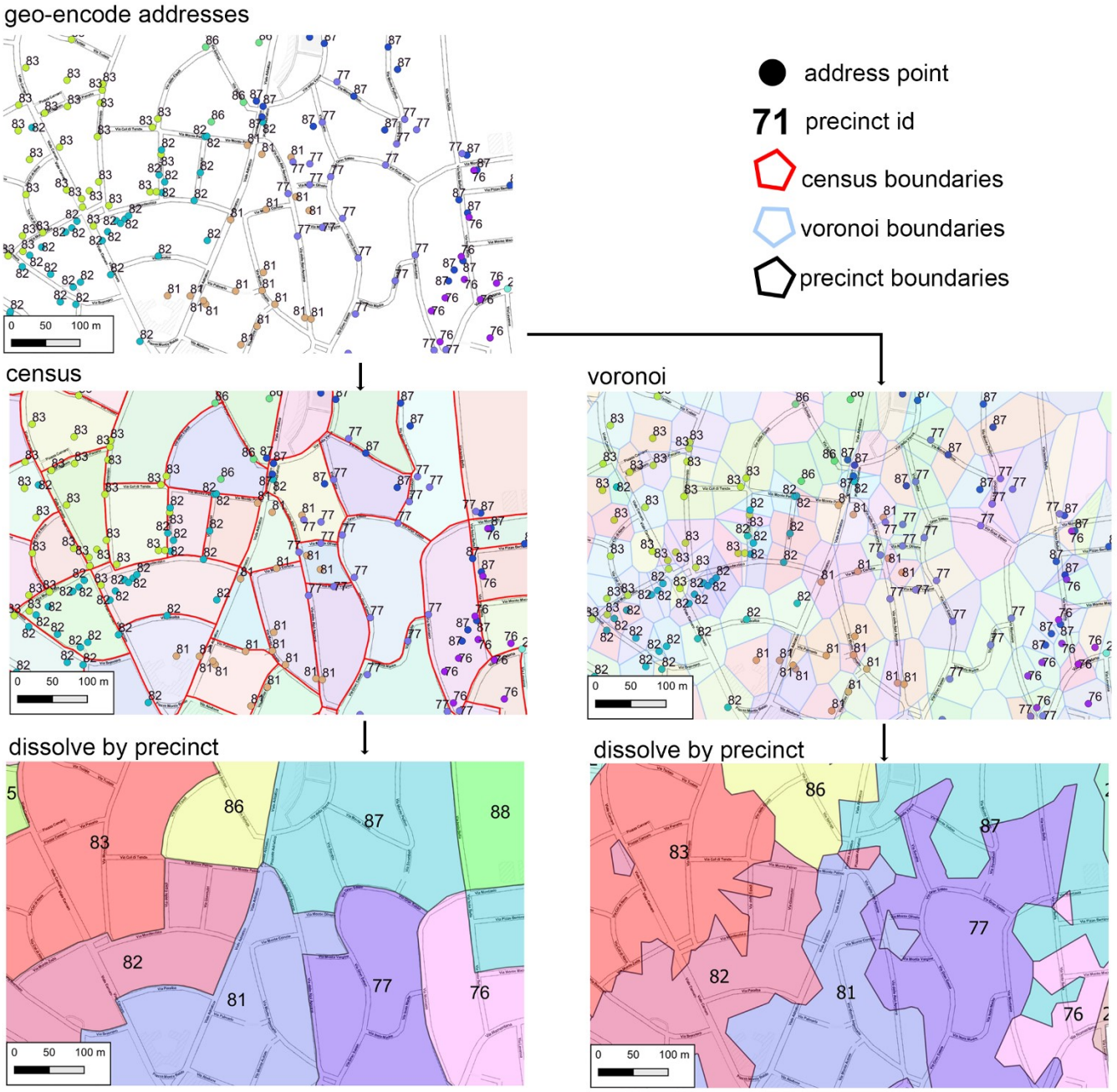


Figure 4: a comparison between the two procedures (Voronoi vs Census)



The figure shows how the two procedures behave when applied on real data. In the top panel the dots represent the coordinates of each address/civic number (colored by precinct). The number (81, 82, 77..etc) are the IDs of the precinct. The central panels, shows on the left the census blocks (delimited by the red line) while on the right the computed Voronoi polygons for each point (delimited by the blue line). The bottom panels show the final polygons obtained for each precinct (delimited by the black line). As we can see, the census precinct overlay smoothly the actual streets (because of the way census blocks are designed). Instead, Voronoi precinct do not strictly follow underlying streets but are better able to capture points that fall over different census blocks (see for example the points that fall over the border between precinct 87 and 77).

4. Limitations

As we see, the two procedures produce two different collections of polygons. Which should be preferred? What are the pros and cons?

The *census precinct* overlaps the actual network of streets and roads, while the voronoi precinct does not (always). *Census precincts* are also generally contiguous areas, which is not often the case for *Voronoi precinct*.

On the cons, *Voronoi precincts* can be applicable and scalable to other countries, even in the absence of census block shapefile (or when census blocks are larger than precincts). In fact, *census precinct* relies on a relatively old shapefile (2011 for Italy) that might not sufficiently reflect the current urban structure of the territory.

Some other limitations apply to both procedures and are relevant for researchers who wish to use these data or are willing to replicate the exercise. First, the geo-encoding API might sometimes return blatantly wrong coordinates; we exclude the points of those addresses when we believe that is the case using a rule-of-thumb routine.[19] Second, sometimes multiple precincts might have a shared address. When that is the case, we split the resulting geometry into multiple pieces equal to the number of overlapping precincts (luckily, just a few of the total).[20] Thirdly, it is important to acknowledge that the polygons we produced do not represent the *exact* area and borders of the precinct but they are only a *proxy* of the area where the voters live.[21] Fourth, although census blocks and electoral precincts are constructed with a toponymic approach (e.g., street and civic numbers), there is no unique correspondence.[22]

[19] Clearly wrong geo-encoding occur when 1) the coordinates received are outside the administrative boundaries of the municipality 2) the coordinates are extremely far from other points of the same precinct (we set the bar at 3 times larger than the mean distance between all other points of the same precinct). In the appendix 1 we give a full explanation of what we do.

[20] For example, in the Electoral Road Dataset of the municipality of Rome there are 5 precincts (with ID from 2413 to 2417) that all correspond to only one unique address: Largo dell'Olgiata 15.

[21] The address we geo-encoded is the "postal address" of a voter, this might not coincides with the voter residency if the entrance of the street is far from where the house is located.

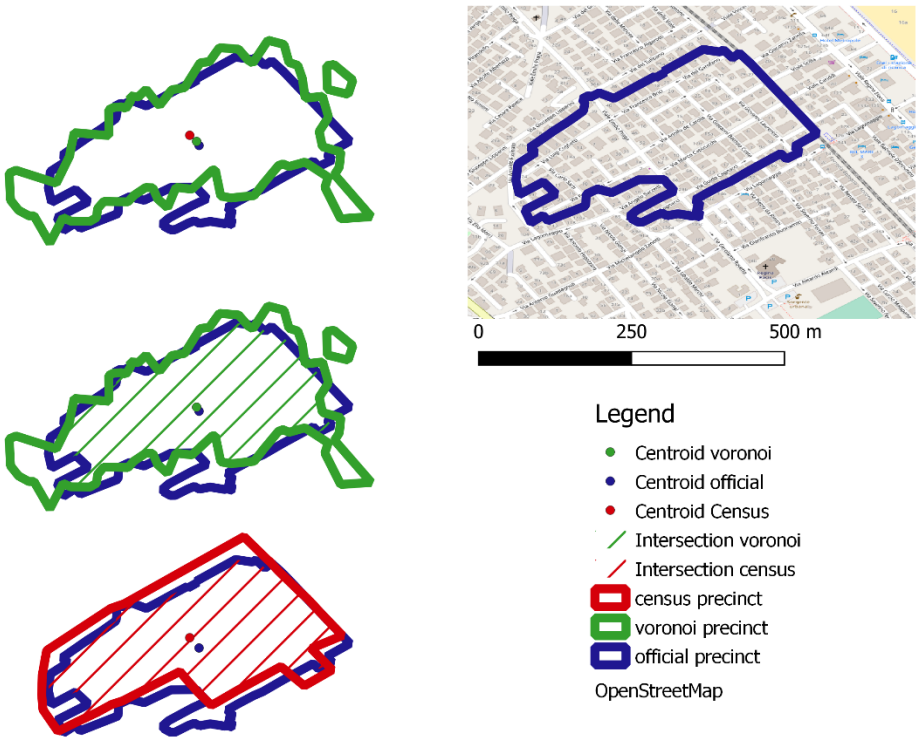
[22] In principle, both electoral precincts and census blocks are created from the dataset of all addresses and civic numbers. In theory, they should be reconcilable. However, the two procedures are carried with clearly different objectives: census blocks are created for statistical purposes, while electoral precincts are only created to assign voters to polling station compliant with the limit posed by existing norms (discussed in the first chapter of this paper). On a private conversation with officers of the municipality they confirmed us that the two procedures are carried independently. The only exception being the boundaries of electoral constituency ("collegi elettorali"). In fact, when the new electoral law for the election of parliament came into force in 2018, the National Institute of Statistics (ISTAT) had to design the new boundaries for the constituencies (Commissione Collegi Elettorali, 2020). When the commission had to assign multiple constituencies to one municipality, they used statistical sub-units of the municipality. For instance, in the case of Rome the commission used the boundaries of the "Zone Urbanistiche", that are an aggregation of census blocks. For this reason, the boundaries of electoral constituency will match boundaries of census blocks. But, still, there might be cases under which an electoral precinct could overlaps different electoral constituency, in these cases, the law foresee that the precinct should be assigned to the constituency where the polling place is located (comma 2,art. 3, Legislative Decree n° 189/2017)..

These limitations are relevant and should be considered when using these data. However, as we will show in the next section through a validation procedure, we are confident that they represent a valuable source and a *reasonably good* approximation.

5. Validation

Some Italian municipalities provide a GIS shapefile of precincts.[23] We can look at these municipalities to validate our proposed methodologies. Our validation procedure will compare the GIS shapefile we obtained by employing our two methodologies (*Census* and *Voronoi*), with the *official* GIS shapefile provided by the municipality. As a case study, we examine the *official* GIS shapefile provided by the municipality of Rimini in the Emilia Romagna.[24] Polygons in the official GIS shapefile have been constructed by hand drawing the boundaries from the local electoral office in the electoral road dataset. In particular, the *official* GIS shapefile does not have

Figure 5: Validation Metrics



The Figure shows the boundaries of one precinct of the municipality of Rimini (precinct 38) taken from the official dataset (*official precinct* – in blue) compared to the two boundaries produced with our methodologies (*census precinct* – in red, and *Voronoi precinct* in green). The line area show the intersection between the different boundaries, while the dot show the centroids.

[23] Rimini, Firenze and Genova.

[24] Available here: https://opendata.comune.rimini.it/dataset/sezioni-elettorali1
We choose the municipality of Rimini (and not Firenze or Genova) because they made available both the GIS file and the Electoral Road dataset. Also, we had several personal contacts with the local office of the municipality which gave us helpful explanation about the data. That was not the case for the municipality of Firenze which never responded to our requests for clarification.

legal force, whereas the electoral road dataset does. Since the boundaries have been handwritten, they cannot have absolute precision. Still, we believe that they can represent a valuable source to validate our methodologies externally.[25]

For validation, we apply our methodology to the Electoral Road Dataset of the municipality of Rimini, following precisely the same steps explained in the previous section. After having obtained two GIS shapefiles from the two methodologies, we compare those with the *official* shapefile provided by the municipality of Rimini.

Our validation procedure relies on two intuitive metrics, *coverage* and *centroid distance* – computed for each of the 142 precincts of the city of Rimini.

Coverage is the percentage of area that overlaps between the polygon of the *Official precinct* and the corresponding polygon of the same precinct we produced with our methodology (*Census precinct* and *Voronoi Precinct)*. Formally (Equation 1), for each precinct p, we will have that $Coverage_p$ is equal to the ratio between:1) the size of the area of the intersection of polygon Z_p (the polygon of *Official precinct* p) and polygon X_p (the polygon produced by our methodology, either the *Census precinct* or the *Voronoi precinct*) and 2) the area of polygon X_p.

Equation 1: $$Coverage_p = \frac{Area(Z_p \cap X_p)}{Area_p^Z}$$

$Coverage_p$ ranges between 0 and 1. It will be equal to 0 if the two polygons X_p and Z_p do not have any area in common, while equal to 1 if they perfectly overlap.

The second metric we use is the $Centroid\ Distance_p$ (Equation 2) that is the geodesic distance (expressed in kilometers) between the centroid Z_p^c of polygon Z_p (the "official" polygon of precinct p) and the centroid X_p^c of polygon X_p. $Centroid\ Distance_p$ has a positive value and can take the minimum of 0 (if the two centroids perfectly overlap) and a maximum distance equal to the *diameter*[26] of the municipality.

Equation 2: $$Centroid\ Distance_p = Distance\ (Z_p^c, X_p^c)$$

Figure 5 reports an example for one precinct.

Table 1 reports the descriptive statistics of the two metrics computed for all *Voronoi* and *Census Precinct* of the municipality of Rimini compared to the *Official Precinct.* As we see, the two procedures perform similarly. On average, *Census precinct* covers 67 percent of the area of *Official precinct* (70 percent for *Voronoi precinct*).

[25] This is the clarification we received from the office *"The shapefile was drawn manually following the electoral roadmap, where the streets and civic numbers belonging to each section are indicated. Because the borders do not pass through the center line they do not have absolute precision, and must be periodically revised, as in the map new constructions could be visually shown in another precinct"* (original translation: *Lo shapefile è stato disegnato manualmente seguendo lo stradario elettorale, dove sono indicate le vie e civici appartenenti ad ogni sezione. Il confine non passando per la mezzeria non ha una precisione assoluta, e va periodicamente rivisto, in quanto in piantina nuove costruzioni potrebbero risultare visivamente in altra sezione*9

[26] e.g. the diameter of the circle that contains all the area of the municipality

The median value being 75 and 70 percent respectively. The average distance between centroids is 230 meters for *Census precinct*, compared to 160 meters for *Voronoi precinct*. But, also in this case, the median value is much lower (100 meters and 70 meters). Manual inspection of the results shows that larger precincts drive the right-skewness of the distribution.[27]

Table 1: Validation metrics results

	Coverage		Centroid Distance (km)	
	Census	Voronoi	Census	Voronoi
mean	0.67	0.70	0.23	0.16
std	0.26	0.15	0.34	0.26
min	0.00	0.00	0.00	0.01
1%	0.00	0.33	0.00	0.01
2.5%	0.01	0.36	0.01	0.01
5%	0.07	0.42	0.01	0.02
10%	0.22	0.51	0.02	0.02
25%	0.59	0.62	0.04	0.04
50%	0.75	0.70	0.10	0.07
75%	0.85	0.80	0.19	0.14
90%	0.92	0.88	0.80	0.37
95%	0.96	0.90	0.93	0.63
97.5%	0.97	0.92	1.25	0.87
99%	0.99	0.94	1.34	1.08
max	1.00	0.95	1.98	2.14

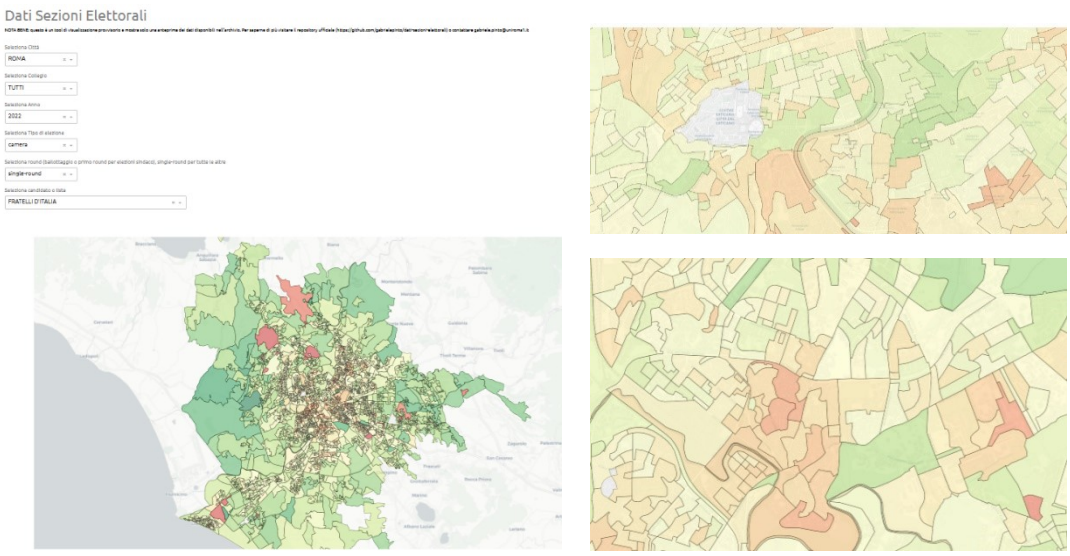
6. The Dataset

We applied the methodologies described in the previous sections to those cities that made the electoral road dataset available. At the time of this writing, the dataset includes 10 of the largest Italian cities, covering 8 million people (13 % of the Italian Population). We make the resulting geoencoded dataset in the form of Esri shapefile that can be downloaded from a public repository (https://github.com/-gabrielepinto/dati-sezioni-elettorali) in both formats (Voronoi and Census precinct). To enhance the usability of the data, we attach to each dataset of each city: 1) the list of geocoded polling places together with their geocoded location, 2) interpolated census data from the National Institute of Statistics (ISTAT), 3) a Polish version of the electoral results at the precinct level. Furthermore, we constructed a web application to explore and visualize the dataset interactively (https://gabrielepinto-dashboard-sezioni-elettorali-appstreamlite-4t5mbj.streamlit.app/).

[27] In large area precincts the metrics Centroid Distance is higher. The distribution of the metrics is right-skewed, wit mean larger than the median because of the presence of large extreme values.

7. Exploratory analysis

Figure 6:Screenshot of the Web-App



This section presents an initial analysis of the data with the aim of providing a descriptive overview of the main patterns observed in the data and demonstrating the potential uses of the dataset.

We explore the topic of spatial inequalities in voting behavior, specifically examining how voting patterns in cities may have become more unequal and polarized over time. To do so, we use three different indexes: a) the Gini inequality index, b) the Moran I index of Spatial Inequality c) the ratio between the 80th and 20th percentiles of the vote. These indexes are calculated for both turnout and voting for left- and right-wing[28] parties in parliamentary elections between 2006 and 2022. The three indexes are plotted in Figure 7, 8 and 9, overall, there has been a consistent increase in spatial inequality and polarization in voting behavior over this time period. This trend is particularly pronounced for left-wing parties, which saw a sustained increase until reaching a peak in 2018 before declining slightly. In Figure 10 we plot, for each election, a collection of maps that report the first group of coalitions and parties in each precinct. By visually inspecting the maps, we can provide evidence that voting for left-wing parties shows a trend toward more central areas of cities that - in the case of Milan and Rome - correspond to wealthier areas. At the same time, we account for the emergence of the M5S (*Five Star Movement*) in the peripheric area of Rome and Turin.

These findings suggest that spatial inequalities in voting behavior may be a significant factor influencing the political landscape in cities and should be further examined in order to understand the underlying causes and potential consequences.

[28] We classify all parties competing in national elections in four main groups: left-wing, right-wing, M5S (Five Star Movement) and Third Pole. In the appendix we report the full list of parties under each group. The Third Pole comprises the parties linked to the former prime minister Mario Monti (*Scelta Civica and Scelta per l'Europa*), and the recently formed alliance between *Azione* and *Italia Viva*.

Figure 7: Inequality in voting behaviour – Gini Index

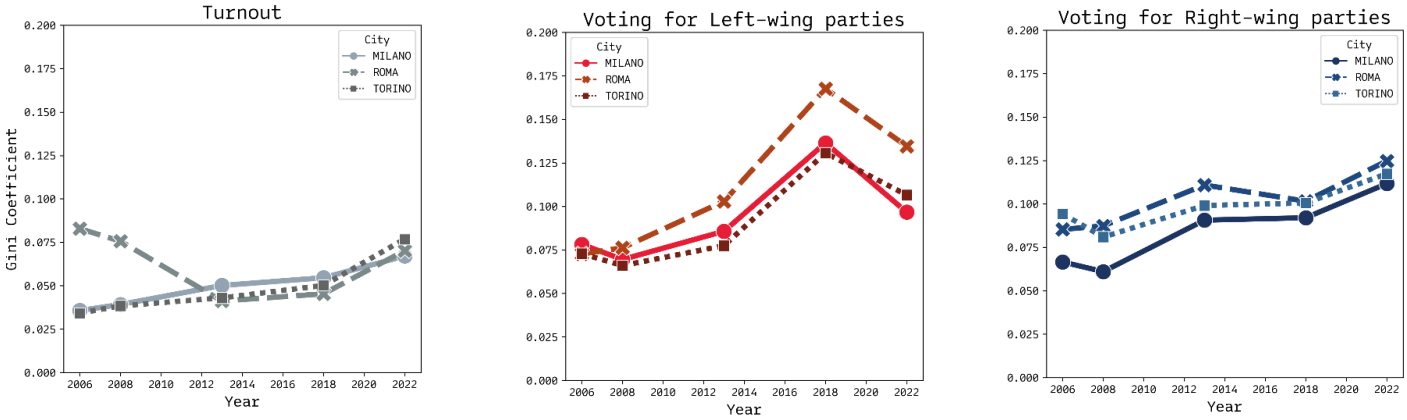


Figure 8: Spatial Inequality in voting behaviour – Moran I index

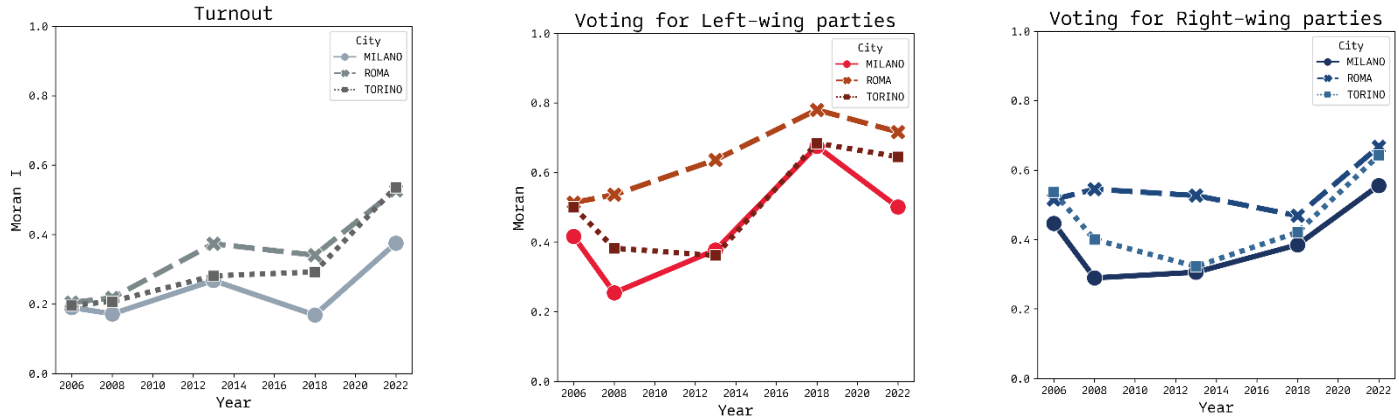
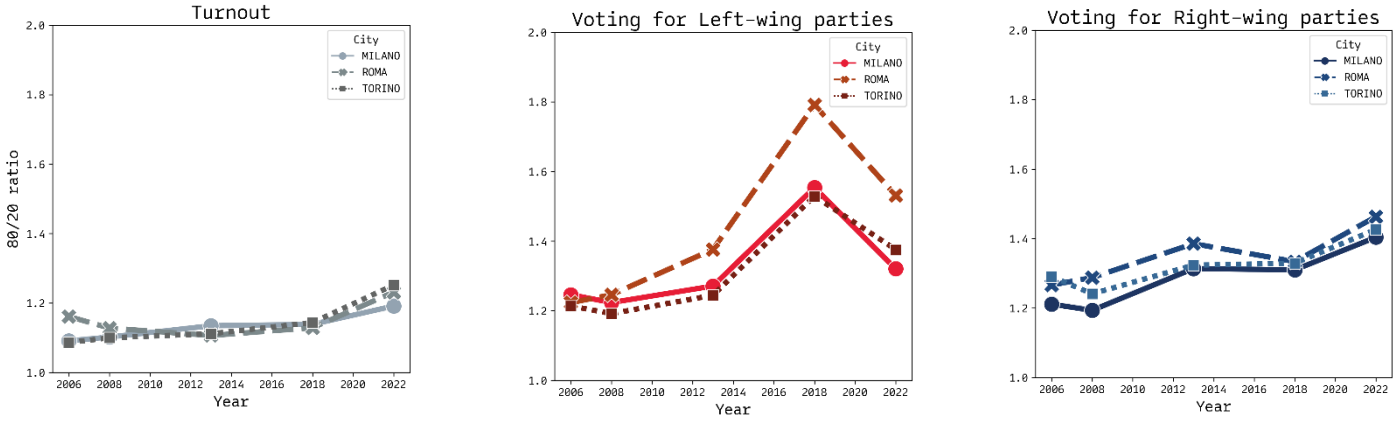


Figure 9: Spatial Inequality in voting behaviour – Polarization index (80/20 percentiles ratio)



One possible explanation for the increase in spatial inequality and polarization in voting behavior that we observe could be the increasing economic and social divide between different geographic areas. In cities, it is common for certain neighborhoods or districts to have higher levels of wealth and education, while others may be disadvantaged and have lower levels of these resources. This can lead to divergent political preferences, with those in more affluent areas tending to support different parties than those in poorer areas. Furthermore, the concentration of certain groups

in specific geographic areas can contribute to political polarization, as these groups may have more influence in shaping the political landscape in their area. Other potential factors that could be contributing to the observed trend include changes in media and communication patterns, changes in political parties and their platforms, or shifts in social and cultural values within different communities (Agnew, 1996; Cini et al., 2021; Crulli, 2022; Dijkstra et al., 2020; Emanuele, 2018; Pratschke et al., 2021; Truglia et al., 2018; Walks, 2005). Further research will be needed to fully understand the underlying causes and consequences of these trends.

In the last part of our exploratory analysis, we put our magnifying glass in the most central areas of the city, also known as ZTL (which stands for Limited Traffic Zones). In doing so, we try to test a definition that has become very popular among commentators in the aftermath of the 2018 elections: "*partito delle ZTL*" (party of the ZTL). Commentators use this definition to refer to left-wing parties (especially the *Democratic Party* and its allies) and the progressive concentration of their consensus in the most central and wealthiest areas of cities at the expense of peripheral areas.[29]

In Figure 11, we plot the mean share of votes for left-wing parties in those precincts inside or outside the ZTL. While we do indeed find that left-wing parties have become more popular in ZTL areas compared to the past, we do not see the huge differences often described in public discussion (D'albergo, 2022; Domani, 2022; Ferrara, 2018; Zurlo, 2022). If we look closer at Figure 11 (above), we see that much of the loss incurred by right-wing parties in those areas have been captured by the Third Pole (pink polygons).

[29] The term "*Partito delle ZTL*" became popular after being mentioned by the Journalist Massimo Giannini in a Political Talk Show (*Di Martedi*) on 13 March 2018 (just after the 4th March Elections).. However, the term had already appeared in some twitter discussions at the time, where it seems that its creation is to be attributed to the journalist Leonardo Panetta (Mediaset): https://twitter.com/marcobreso/status/972047046332035072

Figure 10: First party/coalition in national election at the precinct level

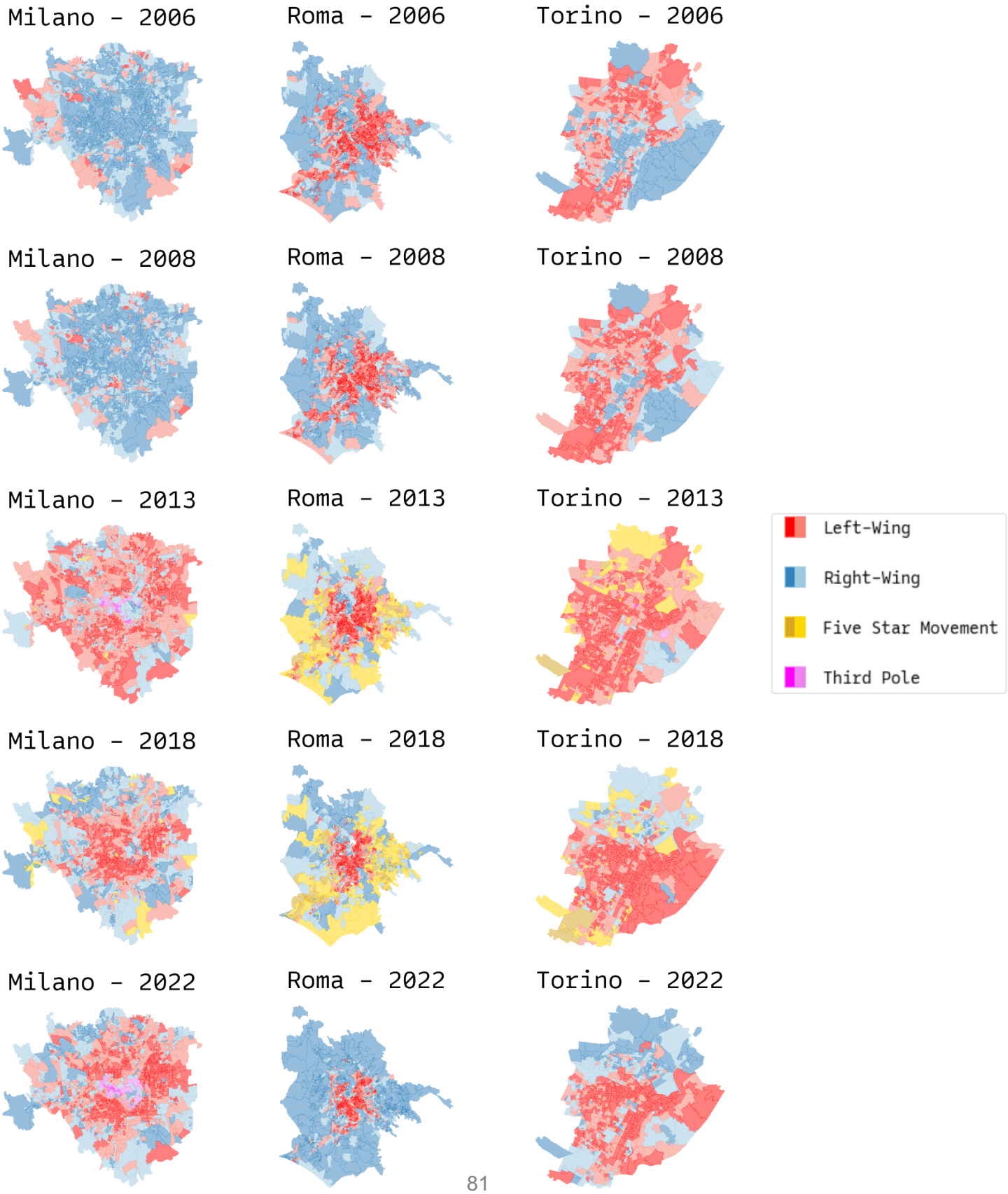
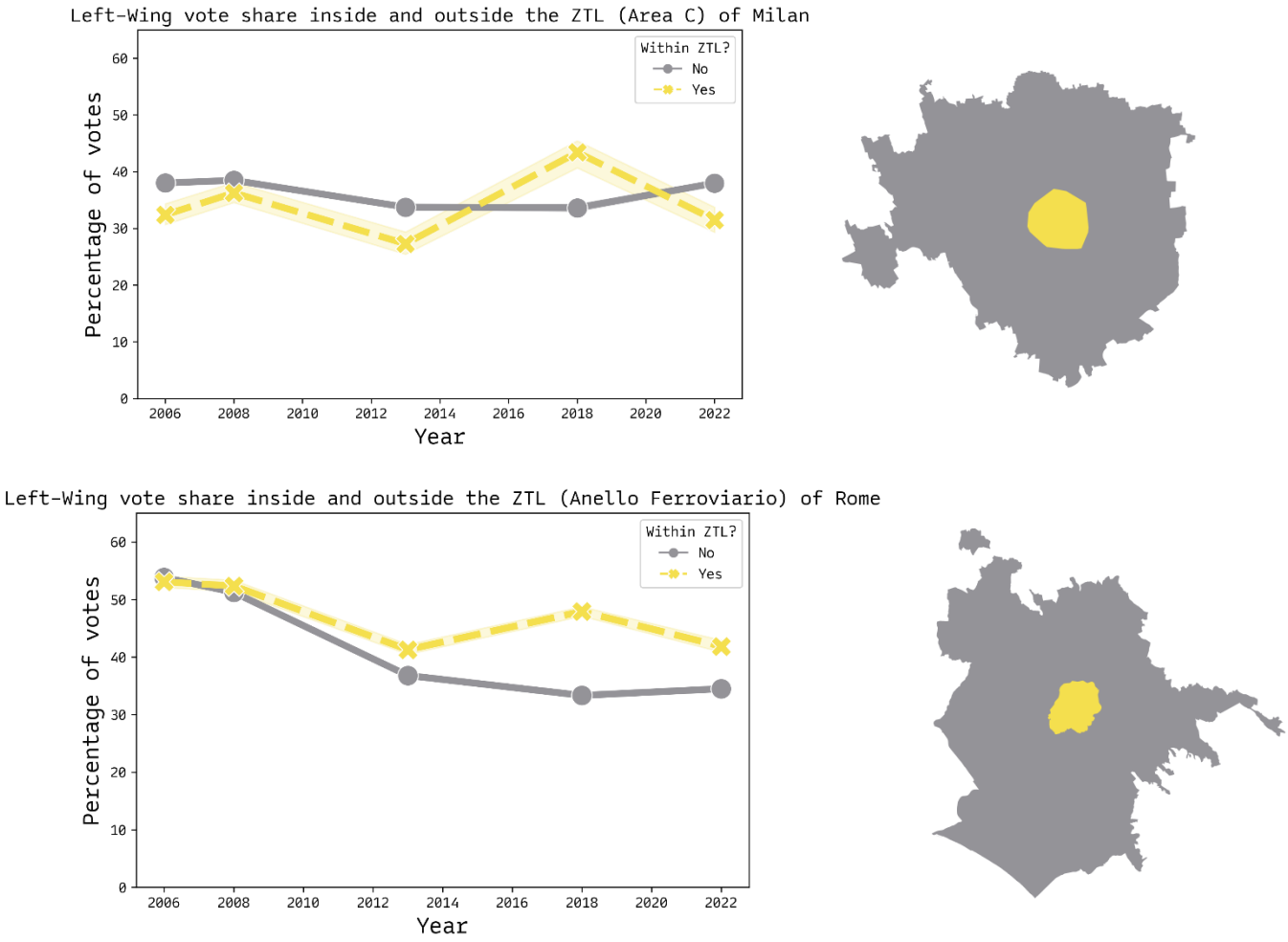


Figure 11: Support for left-wing parties in the Limited Traffic Zones (National Elections)



8. Conclusions

Precinct-level data and GIS files can be incredibly valuable for researchers studying electoral patterns, but they can also be some of the most challenging types of data to collect. In this study, we develop a new methodology and technical guide specifically tailored to overcome these challenges and make these data more accessible in Italy. Our dataset currently includes data from the 10 largest cities in Italy for the past 20 years, but we plan to expand to include more cities and countries in the future.

Our exploratory analysis of this dataset revealed some interesting trends in voting patterns in large Italian cities. Specifically, we found evidence of increasing spatial polarization of voting behavior, with certain areas becoming more politically heterogeneous over time. We also observed a trend towards a concentration of left-wing voters in wealthier, central areas of these cities. These trends may be influenced by a variety of socio-spatial factors, such as income, education, and geographic location.

The granularity of these data allows us to delve more deeply into these trends and explore the underlying causes of spatial polarization and inequality in voting behavior. This information can be useful for policymakers and political analysts looking to better understand

the political landscape in these cities and how it may change over time. In the future, we hope to continue expanding our dataset to include more cities and countries, further advancing our understanding of electoral patterns and the factors that shape them.

References

Agnew, J. (1996). Mapping politics: how context counts in electoral geography. *Political Geography*, *15*(2), 129–146.

Alabrese, E., Becker, S. O., Fetzer, T., & Novy, D. (2019). Who voted for Brexit? Individual and regional data combined. *European Journal of Political Economy*, *56*, 132–150.

Amos, B., McDonald, M. P., & Watkins, R. (2017). When boundaries collide: Constructing a national database of demographic and voting statistics. *Public Opinion Quarterly*, *81*(S1), 385–400.

Amos, B., Smith, D. A., & Claire, C. S. (2017). Reprecincting and voting behavior. *Political Behavior*, *39*(1), 133–156.

Baltz, S., Agadjanian, A., Chin, D., Curiel, J., DeLuca, K., Dunham, J., Miranda, J., Phillips, C. H., Uhlman, A., Wimpy, C., & others. (2022). American election results at the precinct level. *Scientific Data*, *9*(1), 1–12.

Beck, N., Gleditsch, K. S., & Beardsley, K. (2006). Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly*, *50*(1), 27–44.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly*, *65*(1), 22–44.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717). Springer.

Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2015). *Principles of geographical information systems*. Oxford university press.

Cini, L., Colloca, P., Maggini, N., Tomassi, F., & Valbruzzi, M. (2021). Inchiesta su periferie urbane, disagio socio-economico e voto. I casi di Bologna, Firenze e Roma. *Quaderni Di Scienza Politica*, *28*(2), 137–177.

Commissione Collegi Elettorali. (2020). *Proposta dei collegi uninominali e plurinominali per la Camera dei deputati e il Senato della Repubblica*. http://www.riformeistituzionali.gov.it/media/1364/commissione-collegi-elettorali_relazione_13-novembre-2020-con-errata-corrige.pdf

Cooper, C. A., Haspel, M., & Knotts, H. G. (2009). The value of Voterfiles for US state politics research. *State Politics \& Policy Quarterly*, 102–121.

Corbetta, P., & Piretti, M. S. (2009). *Atlante storico-elettorale d'Italia: 1861-2008*. Zanichelli.

Crulli, M. (2022). Vote metropolitanization after the transnational cleavage and the suburbanization of radical right populism: the cases of London and Rome. *Quaderni Dell'Osservatorio Elettorale QOE-IJES*.

D'albergo, L. (2022, September 26). Il Pd va sotto anche nel centro storico di Roma. Non era mai successo: "Non siamo più il partito della Ztl." *La Repubblica*. https://roma.repubblica.it/cronaca/2022/09/26/news/pd_roma_centro_storico_ztl_sconfitta-367334149/

Darmofal, D. (2015). *Spatial analysis for the social sciences*. Cambridge University Press.

Dijkstra, L., Poelman, H., & Rodr\'\iguez-Pose, A. (2020). The geography of EU discontent. *Regional Studies*, *54*(6), 737–753.

Domani, E. (2022). Il Pd resta chiuso in una Ztl sempre più stretta. *Domani*. https://www.editorialedomani.it/politica/italia/il-pd-resta-chiuso-in-una-ztl-sempre-piu-stretta-uvfs78pu

Emanuele, V. (2018). *Cleavages, institutions, and competition: understanding vote nationalization in Western Europe (1965-2015)*. Rowman \& Littlefield.

Ferrara, E. (2018). *PD, OVVERO IL PARTITO DELLE ZTL*. https://ricerca.repubblica.it/repubblica/archivio/repubblica/2018/06/27/pd-ovvero-il-partito-delle-ztlFirenze01.html

Kollman, K., Hicken, A., Caramani, D., Backer, D., & Lublin, D. (2019). Constituency-Level Elections Archive (CLEA)(Data file and codebook). *Ann Arbor, MI: Center for Political Studies, University of Michigan (Producer and Distributor). Available at: Http://Www. Electiondataarchive. Org*.

Lelo, K., Monni, S., & Tomassi, F. (2021). *Le mappe della disuguaglianza: una geografia sociale metropolitana*. Donzelli Editore.

Ministero Dell'Interno. (2021). *Eligendo*. https://elezioni.interno.gov.it/

Mobasheri, A., Mitasova, H., Neteler, M., Singleton, A., Ledoux, H., & Brovelli, M. A. (2020). Highlighting recent trends in open source geospatial science and software. In *Transactions in GIS* (Vol. 24, Issue 5, pp. 1141–1146). Wiley Online Library.

Pratschke, J., Vitale, T., Morelli, N., Cousin, B., Piolatto, M., & Del Fabbro, M. (2021). Electoral support for the 5 Star Movement in Milan: An ecological analysis of social and spatial factors. *Journal of Urban Affairs*, 1–24.

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, *38*(2), 337–341.

Russo, L. (2017). The use of aggregate data in the study of voting behavior: ecological inference, ecological fallacy and other applications. In *The Routledge Handbook of Elections, Voting Behaviorand Public Opinion* (pp. 484–495). Routledge.

Schwartz, S. (1994). The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *American Journal of Public Health*, *84*(5), 819–824.

Selb, P., & Munzert, S. (2013). Voter overrepresentation, vote misreporting, and turnout bias in postelection surveys. *Electoral Studies*, *32*(1), 186–196.

Sinton, D. S., & Lund, J. J. (2007). *Understanding place: GIS and mapping across the curriculum*. ESRI, Inc.

Truglia, F. G., Fruncillo, D., & Addeo, F. (2018). Territorializzazione del consenso elettorale: Determinati spaziali, socio-economiche e culturali. *Fruncillo e Addeo (a Cura Di) Le Elezioni Del*, 83–111.

Voting and Election Science Team. (2020). *2020 Precinct-Level Election Results*. Harvard Dataverse. https://doi.org/10.7910/DVN/K7760H

Walks, R. A. (2005). City--suburban electoral polarization in Great Britain, 1950--2001. *Transactions of the Institute of British Geographers*, *30*(4), 500–517.

Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. *Journal of Peace Research*, *54*(2), 125–140.

Willis, D., Merivaki, T., & Ziogas, I. (2021). Election Data Transparency: Obtaining Precinct-Level Election Returns. *Public Integrity*, 1–16.

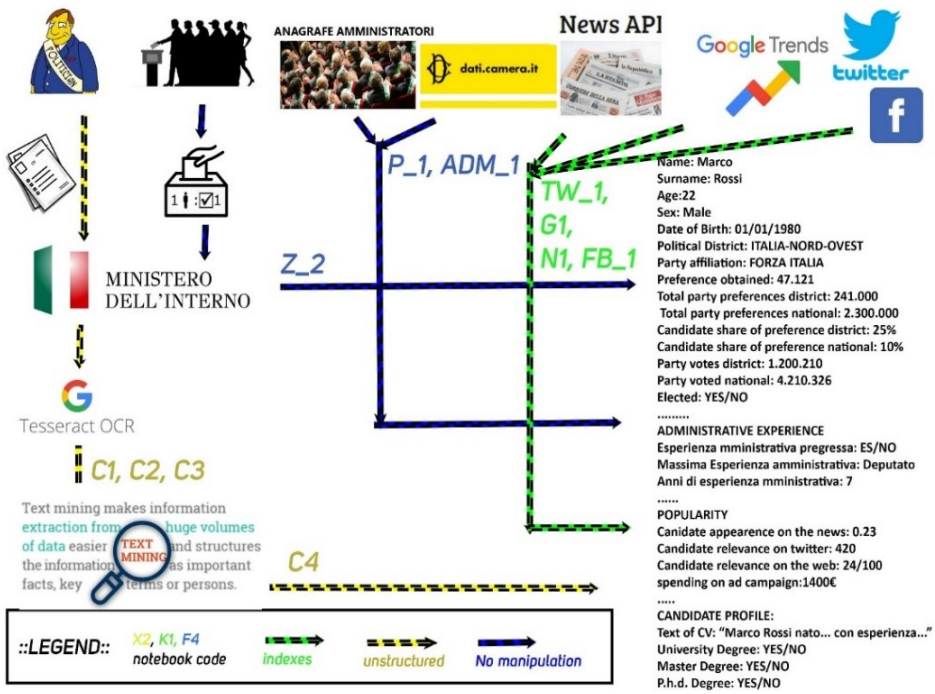
Zurlo, S. (2022). "il pd è il partito della ztl, Ormai Fuori Dalla realtà." In *ilGiornale.it*. il Giornale. https://www.ilgiornale.it/news/politica/pd-partito-ztl-ormai-fuori-realt-2070351.html

Appendix for : "*What characteristics of politicians affect electoral success? Estimates from a new dataset of European candidates' CV*"

1.1 Dataset construction

To guarantee replicability and transparency of our work, we made available all the code and descriptions of all the steps made to build our dataset. Those codes come in the form of Jupyter/IPython notebooks that are accessible on a GitHub repository together with the dataset stored in various formats[1]. Figure 1 below shows a diagram of the input sources used and schematically recaps the manipulation made on the original structure of the data to produce as an output a structured and "well-shaped" dataset. The colors of the lines in Figure A -4 represent the class of manipulation to which the data have been subjecting, where the blue lines refer to data that has not been transformed (data that are already in a structured form), the yellow lines are the data that come from the scanned CVs (unstructured text form) and the green lines represent the data manipulated to build popularity indexes (Google, Twitter[2], etc.). The codes placed near

Figure A.4: Map of sources and manipulation



[1] The notebooks can be found at this link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9PJDL9

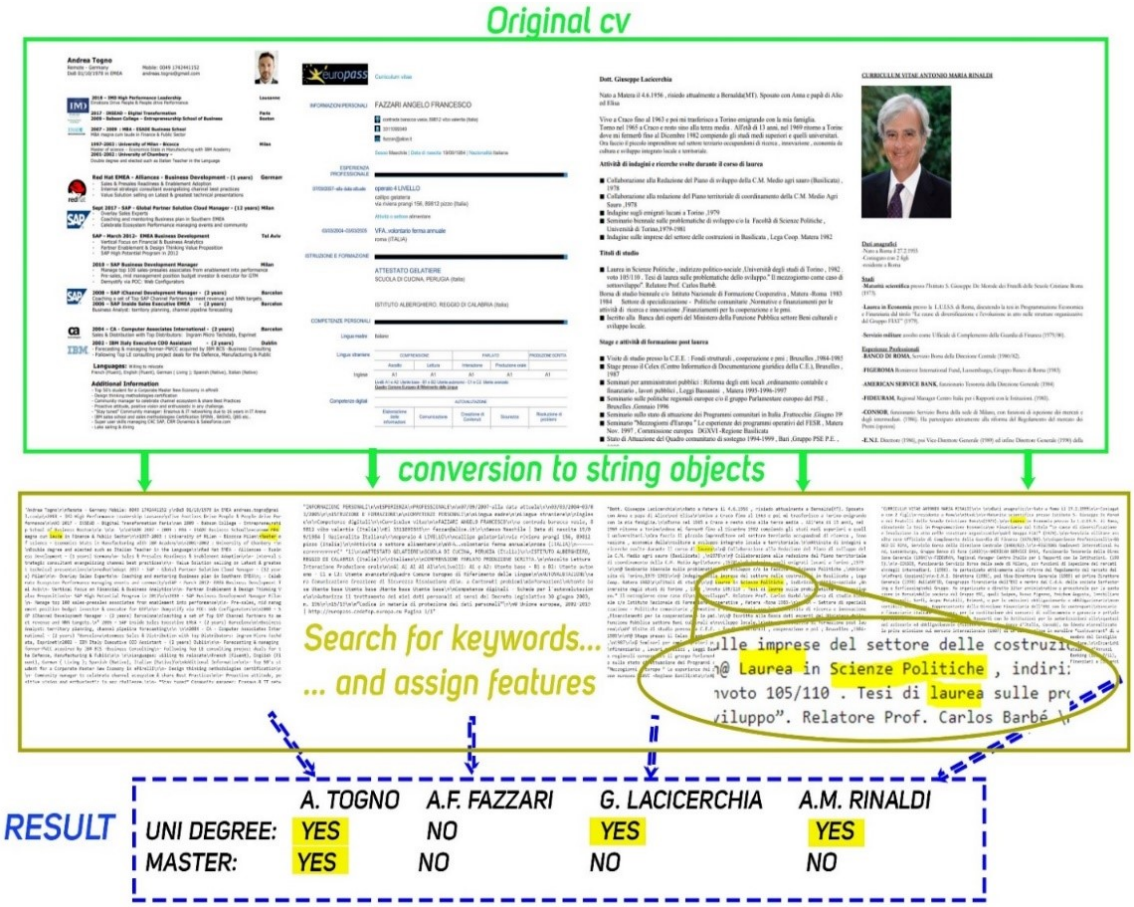
[2] Please note that the Twitter API offers free access only for the data of the week before. Since our data have been collected in the week before the election, in order to reproduce the code the user needs a paid subscription. The code used to make the call to the Twitter API is available on the github repository. Apart from the language detection used to ensure relevance of the tweets with respect to the candidate, the text of the tweet is not used at all (except for clear and obvious irrelevant tweets). The indicators for popularity are build using the counter for "favorite" and "retweet" for all the tweets associated with the name and surname of the candidate. In order to be compliant with the developer policy of twitter (Twitter, 2019), the raw results of the API call are not published but are available on request to the author.

the colored lines (C1, C2, P_1.etc) are used to reference the name of the corresponding jupyter/IPython notebook with the related code.

Figure A-5 shows the pipeline used to extract the information from the CVs. In the top green rectangle, we show some examples of the different formats the CVs are given on the Ministry of Interior website[3]. Since formats differ, we run a "tesseract" algorithm on each CV to extract string objects that contain all the characters of the CVs. We then apply a text analysis to those string objects (yellow rectangle in Figure A5). For example, we look for the presence of any of the following words in the CVs: *'Degree','Degreeto','Degreeta','Degreetosi','facoltà','facolta','giurisprudenza','scienze politiche','bachelor','university','avvocato','laude','chirurgo','medico','dottore'*.

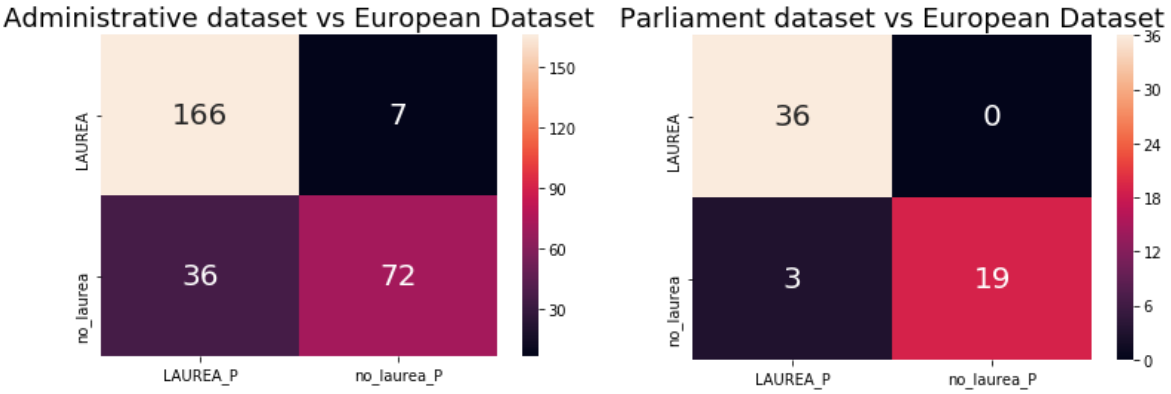
If a candidate's CV contains any of the above words, it will be labeled as a candidate with a university degree (Degree). Similarly, if a candidate has a "PhD" degree, we look up the keyword "*PhD, Dottorato, etc..."* in the CV, and so on. The choice of the keywords is based on prior knowledge and careful analysis of some CV examples, and a detailed

Figure A.5: Pipeline of CV analysis



[3] https://dait.interno.gov.it/elezioni/trasparenza/europee2019

Figure A.6: Cross-Validation of the dataset.
Column sums are the "prediction" of our dataset. Row sums are the actual values contained in the Administrative and Parliament dataset



description of all the keywords we used can be found in the notebook stored in the GitHub repository.[4] We are aware that this text mining approach could produce errors; an example of this error is a candidate that reports having attended a university course but did not fully complete the degree. There might also be candidates who do not report their university degrees in their CV. To tackle this, we construct a validation procedure of our approach by making use of an "official" dataset of all the Italian administrators (at local, municipal, and regional level) and another dataset of the Members of Parliament (MP). Those "official datasets" are based on self-reporting information, and they have been used for several empirical studies[5].

Since some of the candidates might have previous experience running in public offices, we can match those observations and compare the information we obtained with our approach and the information contained in those "official datasets." Below we show the confusion matrix (Figure) that compares the information we obtained through our procedure (the "European Dataset") with those contained in the official datasets ("Administrative" and "Parliament" datasets). When compared with the administrative dataset (left matrix in the figure), we have contradictory results in 16 percent of the cases ($\frac{36+7}{166+72}$), while in the parliament dataset we have contradictory results in 5 percent of the cases ($\frac{0+3}{36+19}$). These percentages might seem high, but a careful "manual" examination of true and false positive observations (bottom-left and top-right square of the confusion matrix) reveals that the contradictions are due in most part to errors in the administrative dataset. In Section 6.2 of the Appendix, we show observations taken randomly from the two groups to explain the comparison between the information given in the CV and those contained in the administrative dataset[6].

We want to stress that text mining methods could be more prone to errors than other approaches. For instance, given the relatively limited number of observations (1076), we could have opted for a manual inspection of the CVs (e.g., by "humanly" reading each CV and manually coding the information on a spreadsheet). This procedure might be much less prone to errors caused by a misunderstanding of the information contained in

[4] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/9PJDL9
[5] for example Baltrunaite et al. (2014) use the Administrator dataset. Galasso & Nannicini (2011) use the MP dataset.
[6] The notebook C5_validation in the github repository also makes available the full list of links to the cv.

the CVs and might be relatively not too much time expensive for our case[7]. But there are also some significant pitfalls: first, "human" transcription is likely to be prone to errors too ("spreadsheet errors"). Moreover, it is difficult to guarantee transparency of the process[8]. Hand-coding might work if we need to translate only 1000 CVs. Still, the task becomes much more expensive when this number increases and the methodology used here can be easily applied to other cases in which similar data sources are available.

1.2 Validation of the dataset

In this section, we manually inspect the errors we get in our cross-validation procedure (Figure A.6) by comparing the information we obtain by mining the text of the CVs (and constructing our dataset) with the information contained in the Administrator dataset ("Anagrafe degli amministratori") and the Parliament dataset (dati.camera.it). Both datasets are based on self-reported measures.

FALSE POSITIVE
Observations labeled with a university degree by us but not in the administrative and Parliament dataset. The figures below report two screenshots, on the left, there is a screenshot from the "browsable" version of the administrative and Parliament dataset, on the right, there is a screenshot from the CV mined to obtain the label of our dataset. The color of the rectangle (yellow, green, or red) is used to tell if the error is due to misinterpretation of the text or error in the administrative/parliament dataset.

TRUE POSITIVE
(observations that have NOT been labeled with a degree by us but not in the administrative dataset).

Those observations are mainly related to the CV because the university degree of the candidate cannot be easily inferred from the reading of the text.

Example 1
The candidate has declared to have obtained a degree in his CV (right) but not in the official administrative dataset (left).

[7] If we estimate 2 minutes per CV the task would be done within 30 hours of work
[8] There is not easy replicability, there are difficulties in identifying sources of potential errors, and transparency concerns in the rules used for contested cases.

Example 2

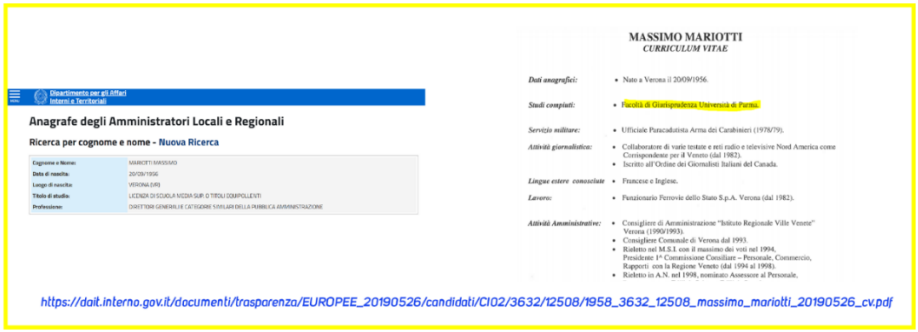
The candidate declares to be a "Medicine Student" in her CV (right) but not in the administrative dataset (left). However, from her CV it is not clear if she has obtained at least a bachelor's degree or if she only attended the university.

Example 3

The candidate clearly declares to have obtained a degree in his CV (right), that is not the case in the official administrative dataset (left).

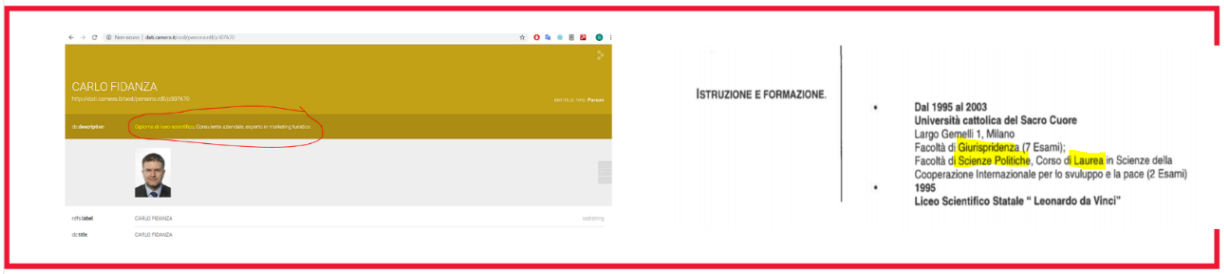
Example 4

The candidate declares to have obtained a degree in his CV (right), that is not the case in the official administrative dataset (left). However, the candidate only reports the name of the faculty and university in his CV. Unclear if he has obtained the degree.



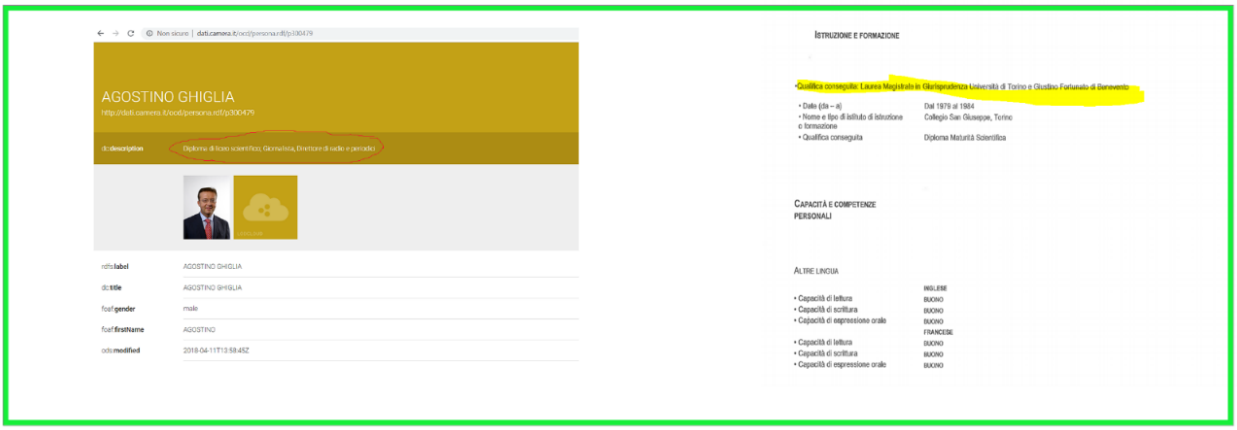
Example 5

The candidate reports the name of the faculty and university and the passed exams (degree not obtained). In the official administrative dataset he correctly reports not having obtained the degree (left). This can be considered a clear error of the procedure.



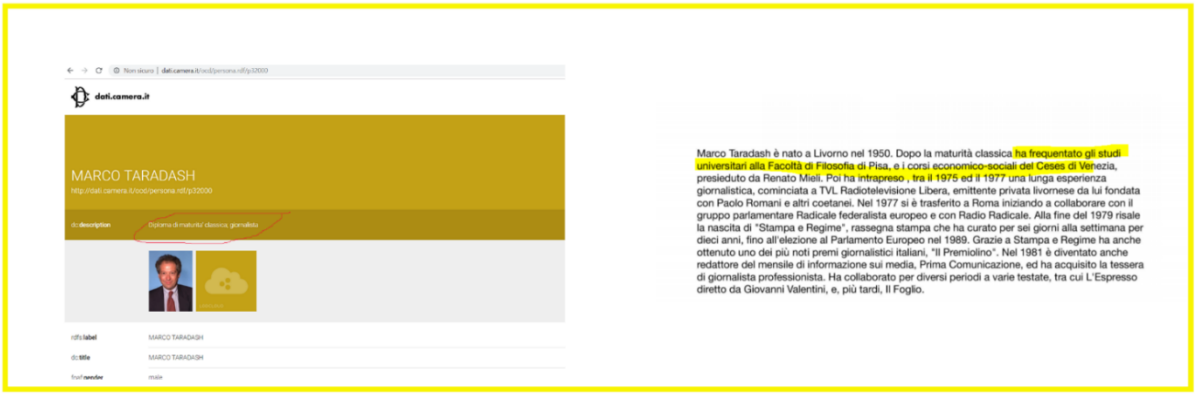
Example 6

The candidate declares to have obtained a degree in his CV (right), that is not the case in the official administrative dataset (left). Note that he is the same Candidate of Example 1.

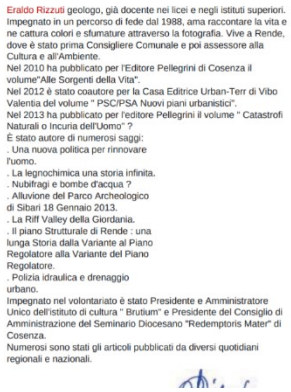


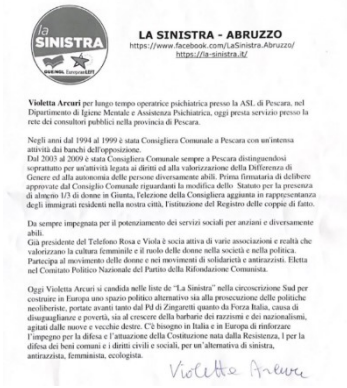
Example 7

The candidate declares to have attended the university in his CV (right). He does not report to have obtained the degree in the official administrative dataset (left). Unclear if the candidate has obtained the degree and/if this is an error of the procedure.



Example 8 and Example 9


Eraldo Rizzuti geologo, già docente nei licei e negli istituti superiori. Impegnato in un percorso di fede dal 1988, ama raccontare la vita e ne cattura colori e sfumature attraverso la fotografia. Vive a Rende, dove è stato prima Consigliere Comunale e poi assessore alla Cultura e all'Ambiente.
Nel 2010 ha pubblicato per l'Editore Pellegrini di Cosenza il volume"Alle Sorgenti della Vita".
Nel 2012 è stato coautore per la Casa Editrice Urban-Terr di Vibo Valentia del volume " PSC/PSA Nuovi piani urbanistici".
Nel 2013 ha pubblicato per l'editore Pellegrini il volume " Catastrofi Naturali o Incuria dell'Uomo" ?
È stato autore di numerosi saggi:
. Una nuova politica per rinnovare l'uomo.
. La legnochimica una storia infinita.
. Nubifragi e bombe d'acqua ?
. Alluvione del Parco Archeologico di Sibari 18 Gennaio 2013.
. La Riff Valley della Giordania.
. Il piano Strutturale di Rende : una lunga Storia dalla Variante al Piano Regolatore alla Variante del Piano Regolatore.
. Polizia idraulica e drenaggio urbano.
Impegnato nel volontariato è stato Presidente e Amministratore Unico dell'Istituto di cultura " Brutium" e Presidente del Consiglio di Amministrazione del Seminario Diocesano "Redemptoris Mater" di Cosenza.
Numerosi sono stati gli articoli pubblicati da diversi quotidiani regionali e nazionali.



1.3 Twitter, Google Trend and Facebook Ads

Twitter
For each candidate, we retrieve tweets found by the Search Twitter API by sending a query with the name and the surname of the candidate for an interval . We exclude retweet and tweet not in Italian Language.

Google Trend Index
For each candidate, we retrieve the daily google trend index for the 2 months before the day of the election (23/03/2019 and 23/05/2019).

Facebook Ads
Retrieved from facebook public data about political advertising.
https://www.facebook.com/ads/library/report/.

1.4 Robustness tests

Table A-1: EXCLUDING TOP AND BOTTOM 95th percentile

This is the same as the results reported in Table 3. But we exclude the bottom 5th and top 95th quantiles.

Dep.Var. Preference Votes	I	II	III	IV	V
Degree	1.15**	2.53***	1.23**	2.37***	1.91***
	(0.58)	(0.63)	(0.54)	(0.59)	(0.55)
PhD	-0.69	-0.61	0.00	0.23	0.21
	(0.94)	(0.95)	(0.88)	(0.90)	(0.79)
Political Experience			0.50	2.54***	2.05***
			(0.63)	(0.71)	(0.66)
Female			-2.42***	-1.98***	-1.59***
			(0.53)	(0.47)	(0.41)
Age (log)			0.35	0.20	-0.07
			(1.02)	(0.98)	(0.93)
Position List			-0.82***	-0.69***	-0.58***
			(0.08)	(0.07)	(0.06)
Google Index (log)					0.90***
					(0.11)
Twitter (log)					1.50***
					(0.54)
Facebok Ads (log)					0.21**
					(0.10)
N	*21604*	*21604*	*21604*	*21604*	*21604*
CONTROLS: Province and Party (in column 2,4 and 5)					

*,**,***; p-value at 1,5 and 10 per cent; standard errors are clustered at candidate level. Dependent variable is expressed in percentage (min=1, max=100). Magnitude interpretation: A coefficient of -2.42 for a dummy variable means an decrease of 2.42 percentage points for the relative share of preferences. The dependent variable has mean 6.9, with 12.8 standard deviation

Table A-2: Alternative measure of Candidates' success

This is the same as the results reported in Table 1. But we use an alternative measure of candidate success. That is the change in the rank of a candidate between initial ballot position and rank on the election outcome. This is the same measure used in (Mechtel, 2014).There are several changes worthy to discuss. Firstly the sign of the PhD and Female coefficients in some of the specifications turns sign (despite not always statistically significant), this might be explained by the fact that those factors might help in "differentiating" positively within the rank of the list, but not in terms of absolute of relative success (note that the distribution of the votes is highly skewed to the bottom part).

Dep.Var. Rank Change	I	II	III	IV	V
Degree	-0.38	-0.10	-0.15	0.24	0.21
	(0.27)	(0.29)	(0.16)	(0.16)	(0.15)
PhD	0.23	0.44	-0.05	0.36	0.25
	(0.52)	(0.54)	(0.32)	(0.31)	(0.29)
Political Experience			0.01	0.77***	0.56***
			(0.17)	(0.19)	(0.18)
Female			-0.21	-0.08	0.06
			(0.15)	(0.14)	(0.13)
Age (log)			-0.23	-0.00	0.02
			(0.30)	(0.28)	(0.27)
Position List			0.70***	0.75***	0.79***
			(0.02)	(0.02)	(0.02)
Google Index (log)					0.18***
					(0.03)
Twitter (log)					0.51***
					(0.11)
Facebok Ads (log)					0.09***
					(0.03)
N	*24052*	*24052*	*24052*	*24052*	*24052*
CONTROLS: Province and Party (in column 2 4 and 5)					

*,**,***; p-value at 1,5 and 10 per cent; standard errors are clustered at candidate level. Dependent variable is expressed in absolute value and is the change of the candidate rank between the initial position and the rank on the election outcome (min=-14, max=14). Magnitude interpretation: A coefficient of +0.36 for a dummy variable means an increase of 0.36 positions on the change of rank . The mean value for the dependent variable is 0.29, with 4.77 standard deviation.

Table A-3 Excluding Giorgia Meloni

This is the same as the results reported in Table 3. But we exclude the bottom 5th and top 95th quantiles.

Dep.Var. Preference Votes (%)	**I (Full-Sample**	**II (excluding G. Meloni)**
LA SINISTRA	2.20	2.34
	(1.51)	(1.50)
PARTITO DEMOCRATICO	2.21	2.10
	(1.64)	(1.58)
MOVIMENTO 5 STELLE	-0.58	-0.36
	(1.25)	(1.20)
FORZA ITALIA	-1.68	-1.98
	(2.13)	(2.13)
LEGA SALVINI PREMIER	-0.30	-0.67
	(2.03)	(2.02)
FRATELLI D'ITALIA	6.69***	3.14**
	(2.07)	(1.38)
N	*24052*	*23945*
CONTROLS: Twitter, Facebook, Party, Province, Age and Political Experience		

*,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. Dependent variable is expressed in percentage (min=1, max=100). The table reports the value of the interaction coefficients between the Party and the dimension listed in the first row. (e.g. in the example of Degree, the delta δ coefficient of the following regression: $P_{c,p} = \alpha * Degree_{c,p} + \gamma * Party_{c,p} + \delta * Party_{c,p} * Degree_{c,p} + Controls_{c,p} + u_{c,p}$)

Table A-4: Different Dependent Variable

Dep.Var. Number of Preferences/ Total Votes	**I**	**II**	**III**	**IV**	**V**
Degree	-0.12	-0.32**	-0.20	-0.33**	-0.31**
	(0.15)	(0.16)	(0.16)	(0.16)	(0.13)
PhD	0.12	-0.06	0.20**	0.02	-0.11
	(0.11)	(0.11)	(0.09)	(0.10)	(0.12)
Political Experience			0.82***	0.38***	0.23***
			(0.14)	(0.11)	(0.09)
Female			-0.17**	-0.24***	-0.13**
			(0.09)	(0.08)	(0.06)
Age (log)			-0.19	-0.10	0.06
			(0.15)	(0.13)	(0.11)
Position List			-0.07***	-0.07***	-0.04***
			(0.01)	(0.02)	(0.01)
Google Index (log)					-0.00
					(0.01)
Twitter (log)					0.76***
					(0.25)
Facebok Ads (log)					0.12***
					(0.03)
N	*24052*	*24052*	*24052*	*24052*	*24052*

CONTROLS: Age, Education, Political Experience, Province and Party

*,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. **Dependent variable** is expressed in percentage (min=1, max=100) and is equal to the number of preferences got by a candidate over total number of preferences in that Province

Table A-5 - Quadratic Age

Dep.Var. Preference Votes	I	II	III	IV	V
Degree	0.61	1.11	0.18	0.85	0.67
	(0.72)	(0.82)	(0.66)	(0.76)	(0.64)
PhD	-0.76	-0.73	0.03	0.28	-0.56
	(0.90)	(0.96)	(0.80)	(0.88)	(0.84)
Political Experience			2.66***	3.76***	2.35***
			(0.74)	(0.74)	(0.64)
Female			-2.61***	-2.42***	-1.59***
			(0.57)	(0.54)	(0.46)
Age			-0.16	-0.12	-0.07
			(0.20)	(0.21)	(0.18)
AGE_squared			0.00	0.00	0.00
			(0.00)	(0.00)	(0.00)
Position List			-1.02***	-0.94***	-0.71***
			(0.08)	(0.08)	(0.06)
Google Index (log)					0.92***
					(0.12)
Twitter (log)					2.85***
					(0.63)
Facebok Ads (log)					0.77***
					(0.15)
N	*24052*	*24052*	*24052*	*24052*	*24052*

CONTROLS: Province and Party Fixed Effect

*,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. **Dependent variable** is expressed in percentage (min=1, max=100).

Table A-6 Probit Model

Dep.Var. Elected (1=Yes, 0=No)	**I**	**II**	**III**	**IV**	**V**
Degree	0.22	-0.03	0.14	0.12	0.08
	(0.14)	(0.19)	(0.15)	(0.16)	(0.22)
PhD	0.19	0.07	0.29	0.15	0.21
	(0.22)	(0.27)	(0.22)	(0.28)	(0.30)
Political Experience			0.80***	0.38**	0.32*
			(0.13)	(0.17)	(0.19)
Female			-0.20	-0.47***	-0.51***
			(0.13)	(0.13)	(0.17)
Age (log)			-0.20	0.15	0.13
			(0.25)	(0.31)	(0.39)
Position List			-0.07***	-0.11***	-0.11***
			(0.02)	(0.02)	(0.02)
Google Index (log)					0.11***
					(0.04)
Twitter (log)					-0.13
					(0.11)
Facebok Ads (log)					-0.00
					(0.03)
N	*24052*	*24052*	*24052*	*24052*	*24052*

CONTROLS: Province and Party Fixed Effect

*,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. **Dependent variable** is expressed as a dummy variable equal to 1 if the candidate has been elected, 0 otherwise

Table A-7 Interaction between Web presence and Sex

Dep.Var. Preference Votes	I	II	III	IV	V	VI
Degree	0.61	1.11	0.26	0.93	0.73	0.71
	(0.72)	(0.82)	(0.67)	(0.76)	(0.63)	(0.61)
PhD	-0.76	-0.73	-0.06	0.20	-0.61	-0.56
	(0.90)	(0.96)	(0.80)	(0.86)	(0.84)	(0.88)
Political Experience			2.62***	3.73***	2.33***	2.43***
			(0.73)	(0.74)	(0.65)	(0.64)
Female			-2.68***	-2.48***	-1.64***	1.19**
			(0.56)	(0.53)	(0.45)	(0.51)
Age (log)			0.46	0.65	1.04	0.78
			(1.15)	(1.18)	(1.08)	(1.05)
Position List			-1.02***	-0.94***	-0.71***	-0.69***
			(0.08)	(0.08)	(0.06)	(0.06)
Google Index (log)					0.92***	1.23***
					(0.12)	(0.18)
Twitter (log)					2.85***	4.07***
					(0.63)	(0.87)
Facebok Ads (log)					0.77***	0.68***
					(0.15)	(0.19)
Female*Twitter (log)						-3.40***
						(1.09)
Female*Google Index (log)						-0.68***
						(0.19)
Female*Facebok Ads (log)						0.23
						(0.27)
N	*24052*	*24052*	*24052*	*24052*	*24052*	*24052*

CONTROLS: Province and Party Fixed Effect

*,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. **Dependent variable** is expressed in percentage (min=1, max=100).

Table A-8 Different type of Degree

Dep.Var. Preference Votes	I	II	III	IV	V
PhD	-0.73	-0.84	0.04	0.25	-0.20
	(0.90)	(0.97)	(0.81)	(0.89)	(0.79)
Political Experience			2.14***	3.72***	2.81***
			(0.77)	(0.84)	(0.76)
Female			-2.89***	-2.51***	-1.83***
			(0.64)	(0.58)	(0.52)
Age (log)			2.68*	3.03**	2.50**
			(1.37)	(1.32)	(1.24)
Position List			-0.98***	-0.82***	-0.68***
			(0.09)	(0.08)	(0.07)
Google Index (log)					0.93***
					(0.13)
Twitter (log)					1.75***
					(0.61)
Facebok Ads (log)					0.45***
					(0.14)
Degree Law	0.24	0.69	-0.14	0.57	0.17
	(1.01)	(0.81)	(0.89)	(0.74)	(0.67)
Degree Economics	-0.90	-0.57	-0.61	-0.32	-0.91
	(0.92)	(0.77)	(0.82)	(0.71)	(0.67)
Degree medical	0.67	1.66	-0.55	0.39	-0.42
	(1.24)	(1.08)	(1.14)	(0.98)	(0.95)
N	*16703*	*16703*	*16703*	*16703*	*16703*

CONTROLS: Province and Party Fixed Effect

*,**,***; p-value at 1,5 and 10 percent; standard errors clustered at candidate level. **Dependent variable** is expressed in percentage (min=1, max=100). Reference Degree is "other type of degree"

1.5 Description of the variables

Table A-9: Description of the Variables (for full and detailed information on all the variables of the dataset see the github repository)

dataset name	paper name	Description
retweet_counter_mean	***Twitter***	mean of retweet related tweets
REG_EXPER	***Regional experience***	number of years of experience in local govt (REGION)
PROV_EXPER	***Provincial experience***	number of years of experience in local govt (PROVINCE)
PROVINCE	***Province***	Province where the number of preferences is counted
PREF_RELATIVE_100	***Preferences (%)***	PREF_RELATIVE *100
PREF_RELATIVE	***Preferences***	number of preferences obtained by the candidacy relative to the total number of preferences obtained by the Party in a certain province
POS_LIST	***Position List***	position of the candidate on the ballot list
POL_EXPER	***Political experience***	a dummy of having previous administrative or political experience
PHD_nlp_code	***PhD***	dummy for whether the candidate has a P.h.D. degree_code
PARL_SEN_EXP	***Parliament experience (Senate)***	number of legislative mandate in Parliament (SENATO_DELLA_REPUBBLICA)
PARL_CAM_EXP	***Parliament experience (Camera)***	number of legislative mandate in Parliament (CAMERA_DEPUTATI)
COM_EXPER	***Municipality experience***	number of years of experience in local govt (MUNICIPALITY)
GOOGLE_SUM	***Google Index***	the sum of the daily GOOGLE TREND index over the previous 2 months
SEX_code_female	***Female***	sex of the candidate (male or female)
fb_ads_budget	***Facebook ads***	amount spent on ads by the Facebook page between March and June 2019
EP_EXP	***European experience***	a dummy of incumbent candidate (experience in EUROPEAN PARLIAMENT)
LAUREA_nlp_code	***Degree***	dummy for whether the candidate has a university degree (any level)_code
AGE	***Age***	age of the candidate expressed in years
CIRCOSC		political district
CIRCOSC_code		political district_code
CAND_FULL_NAME		name and surname of the candidate
CAND_NOME		name of the candidate
CAND_COGNOME		the surname of the candidate
DATA_NASCITA		Date of Birth
SEX		sex of the candidate (male or female)
ID		candidate id
LISTE_DESCR_LISTA		party of the candidate
LISTE_DESCR_LISTA_code		party of the candidate_categorical
Casellario_Penale		text of the criminal record certificate
CP_NULLE		dummy for whether the criminal record certificate is empty (no crime record)

num_pag_cp	number of page of the criminal record certificate (might include the cv)
lenght_cp	length in characters of the criminal record certificate (might include the cv)
Curriculum_vitae	text of the curriculum vitae
num_pag_cv	number of pages of the cv
lenght_cv	length in characters of the curriculum vitae (sometimes include the crime record certificate)
ID_CANDIDATURE	a unique identifier for a candidacy (associated with a candidate and a district)
LAUREA_nlp	dummy for whether the candidate has a university degree (any level)
MASTER_nlp	dummy for whether the candidate has a master degree (not including Laurea Specialistica or Laurea Magistrale)
PHD_nlp	dummy for whether the candidate has a P.h.D. degree
MASTER_nlp_code	dummy for whether the candidate has a master degree (not including Laurea Specialistica or Laurea Magistrale)_code
LINK_CP	URL link to the criminal record certificate
NOTE_CP	flag notes on criminal record certificate
ANOMAL_CP	flag note on criminal record certificate
LINK_CV	URL link to the curriculum vitae pdf
NOTE_CV	flag notes on curriculum vitae
ANOMAL_CV	flag notes on curriculum vitae
NA_TWITTER	a dummy on whether the Twitter API gave no response to the call with the name and surname of the candidate
NA_GOOGLE	a dummy on whether the google trend API gave no response to the call with the name and surname of the candidate
NA_CASELLARIO_PENALE	a dummy on whether the criminal records were not present
NA_FACEBOOK	a dummy on whether the Facebook API gave no response to the call with the name and surname of the candidate
favorite_sum	the sum of favorite related tweets
retweet_counter_sum	the sum of retweet related tweets
favorite_max	max of favorite related tweets
retweet_counter_max	max of retweet related tweets
favorite_min	min of favorite related tweets
favorite_mean	mean of favorite related tweets
GOOGLE_MEAN	mean of the daily GOOGLE TREND index over the previous 2 months
GOOGLE_STD	the standard deviation of the daily GOOGLE TREND index over the previous 2 months
fb_page_name	name of the Facebook associated page
PREF_LISTA_CIRCOSC	number of preferences obtained by the affiliated Party in that district (circoscrizione)
retweet_counter_min	min of retweet related tweets
PREFERENZE	number of preferences obtained by the candidacy in a certain province
PREF_LISTA	number of preferences obtained by the affiliated Party in that province

1.6 Other materials

Figure A-7: Example of a Ballot list were information about the candidate is given.source (Mechtel, 2014)

Amtlicher Stimmzettel für die Wahl des Gemeinderats in Ulm am 07. Juni 2009

Sie haben insgesamt 40 Stimmen.
Bitte beachten Sie:
- Kein Bewerber/keine Bewerberin darf mehr als drei Stimmen erhalten.
- Auch wenn Sie mehrere Stimmzettel verwenden, dürfen Sie insgesamt nicht mehr als 40 Stimmen abgeben.
- Wenn Sie mehr als insgesamt 40 Stimmen abgeben, sind alle von Ihnen verwendeten Stimmzettel ungültig!

▶ Bitte lesen Sie vor der Stimmabgabe unbedingt das Merkblatt „Wichtige Hinweise für die Stimmabgabe"! ◀

Wahlvorschlag

Sozialdemokratische Partei Deutschlands **SPD**

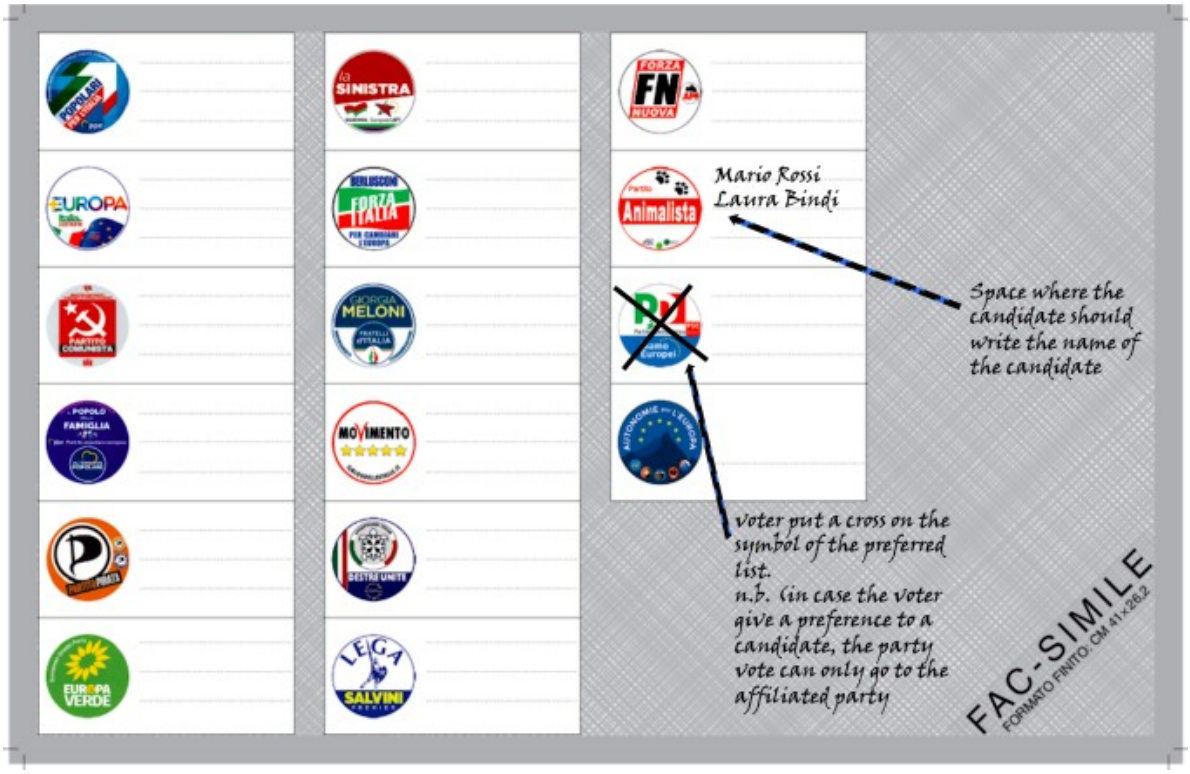
0201	**Kühne** Dorothee, Fraktionsgeschäftsführerin, Ochsensteige 90	0201	
0202	**Rivoir** Martin, Dipl.-Ingenieur u. Landtagsabgeordneter, Sulzbachweg 31	0202	
0203	**Dr. Grimm** Susanne, Ärztin, Gartenhalde 24, Mähringen	0203	
0204	**Bergmann** Ingo, Politikwissenschaftler, Brunnenstr. 39, Unterweiler	0204	
0205	**Schanz** Lisa-Barbara, Medizinisch-Technische Assistentin, Rechbergweg 24	0205	
0206	**Pflüger** Hartmut, Freier Architekt, Griesbadgasse 7	0206	
0207	**Adler** Katja, Rechtsanwältin, Zeppelinstr. 7	0207	
0208	**Kriechbaum** Jürgen, Kriminalbeamter, Dullisbrunnen 2, Lehr	0208	
0209	**Dr. Süslü** Haydar, Allgemeinarzt, Blauäcker 2/1	0209	
0210	**Doll** Dieter, Internet-Sicherheitsbeauftragter, Gartenweg 7	0210	
0211	**Ansbacher** Martin, Jurist, Schlossergasse 22	0211	
0212	**Hofmann** Nick, Buchhändler, Deinselsgasse 12	0212	
0213	**Schulze** Bettina, Verwaltungsangestellte, Trollingerweg 66	0213	
0214	**Giannopoulos** Georgios, Informatiker, Königstr. 26	0214	
0215	**Wimmer** Therese, Verwaltungsleiterin, Böfinger Weg 30/3	0215	
0216	**Dr. Dahlbender** Brigitte, Biologin u. Geographin, Schlehdornweg 1, Einsingen	0216	
0217	**Turath-Schröder** Angelika, Dipl.-Verwaltungswirtin, Thymianweg 2, Einsingen	0217	
0218	**Herrmann** Marc, Moderator, Seßlerweg 11	0218	
0219	**Ullemeyer** Renate, Sachbearbeiterin, König-Wilhelm-Str. 15/6	0219	
0220	**Obermeier** Peter, Architekt, Beim Haslachkreuz 26, Donaustetten	0220	
0221	**Takac** Sabine, Hausfrau, Biberacher Str. 51	0221	
0222	**Lud** Susann, Sozialfachwirtin u. Krankenschwester, Am Sudhaus 10	0222	
0223	**Lambacher** Michael, Industriemechaniker, Wörthstr. 74	0223	
0224	**Bender** Claudia, Kaufmännische Angestellte, Höhenblick 37, Einsingen	0224	
0225	**Schanz** Otmar, Rentner, Soldatenstr. 23	0225	
0226	**Yelen** Fadime, Zahnarzthelferin, Silcherweg 27	0226	
0227	**Stiffel** Günter, Kaufmännischer Angestellter, Königstr. 12	0227	
0228	**Filius** Christine, Tagesstättenleiterin, Riedleinweg 29	0228	
0229	**Schleicher-Rövenstrunck** Hermann, Sozialarbeiter, Im Grünen Winkel 49	0229	
0230	**Bogdashkin** Eugen, Sozialpädagoge, Thomas-Dehler-Weg 23	0230	
0231	**Dr. Klump** Elisabeth, Hausfrau, Zwischen den Wegen 48, Jungingen	0231	
0232	**Dr. Flämig** Günther, Medizincontroller, Margarethe-von-Wrangell-Weg 12/1	0232	
0233	**Schiele** Andrea, Verwaltungsangestellte, Veitsbrunnenweg 1	0233	
0234	**Wecker** Gert, Lehrer, Stuifenweg 29	0234	
0235	**Gabsi** Fathi, Arbeiter, Karlstr. 67	0235	
0236	**Braun-Vogt** Leonore, Dipl.-Ökonomin i.R., Ruländerweg 52	0236	
0237	**Syburra** Christel, Sekretärin, St.-Gallener-Str. 27	0237	
0238	**Haciok** Michaela, Hausfrau, Klosterstr. 25	0238	
0239	**Sprandel** Horst, Rektor a.D., Heidenheimer Str. 99	0239	
0240	**Gnahm** Bernhard, Oberstudienrat, Jörg-Syrlin-Str. 141	0240	

Fig. A.1. Ballot paper: Social Democratic Party, Ulm, local elections 2009.

Figure A-8: Example of Ballot list poster. (European Elections 2019)

11) la SINISTRA – GUE/NGL European LEFT	12) BERLUSCONI FORZA ITALIA PER CAMBIARE L'EUROPA	13) European Green Party EUROPA VERDE	14) +EUROPA Italia in comune	15) Partito Democratico PSE Siamo Europei
Marilena GRASSADONIA nata a PALERMO (PA) il 06/10/1970	**Antonio TAJANI** nato a ROMA (RM) il 04/08/1953	**Annalisa CORRADO** nata a CIVITAVECCHIA (RM) il 08/09/1973	**Emma BONINO** nata a BRA (CN) il 09/03/1948	**Simona BONAFE'** nata a VARESE (VA) il 12/07/1973
Marco BENEDETTELLI nato a ANCONA (AN) il 11/03/1978	**Alessandra MUSSOLINI** nata a ROMA (RM) il 30/12/1962	**Ferdinando BONESSIO detto Nando** nato a FIRENZE (FI) il 15/02/1958	**Marco TARADASH** nato a LIVORNO (LI) il 19/05/1950	**David Maria SASSOLI** nato a FIRENZE (FI) il 30/05/1956
Ginevra Roberta BOMPIANI nata a MILANO (MI) il 05/08/1939	**Raffaella BONSANGUE** nata a SIRACURA (SR) il 25/09/1963	**Beatrice BRIGNONE** nata a SENIGALLIA (AN) il 31/01/1978	**Costanza HERMANIN DE REICHENFELD detta HERMANIN** nata a FIRENZE (FI) il 02/06/1981	**Roberto GUALTIERI** nato a ROMA (RM) il 19/07/1966
Stefano CICCONE nato a ROMA (RM) il 07/02/1964	**Giovanni Paolo BERNINI** nato a PARMA (PR) il 06/04/1963	**Adriano CARDOGNA** nato a ANCONA (AN) il 25/09/1952	**Riccardo TRAVAGLINI** nato a ROMA (RM) il 26/09/1979	**Camilla LAURETI** nata a ROMA (RM) il 20/05/1975
Maurella CARBONE nata a ROMA (RM) il 30/11/1953	**Salvatore DE MEO** nato a FONDI (LT) il 27/10/1971	**Elena PULCINI** nata a PRATOLA PELIGNA (AQ) il 10/03/1950	**Niccolo' RINALDI** nato a FIRENZE (FI) il 03/12/1962	**Pietro BARTOLO** nato a LAMPEDUSA E LINOSA (AG) il 10/02/1956
Giovanni DE LUCIA detto Vanni nato a CIVIDALE DEL FRIULI (UD) il 12/08/1953	**Maria DIMASI** nata a POMEZIA (RM) il 11/02/1974	**Bengasi BATTISTI** nato a CORCHIANO (VT) il 15/02/1959	**Carlotta CAPONI** nata a FOLIGNO (PG) il 31/03/1982	**Beatrice COVASSI** nata a FIRENZE (FI) il 01/12/1968
Ivanilde CARVALHO nata a GUAJARA MIRIM (BRASILE) il 20/07/1974	**Alessandra FEDUZI** nata a URBINO (PU) il 16/10/1987	**Carmela LUONGO detta Carmen** nata a ALTAVILLA IRPINA (AV) il 10/06/1960	**Francesco MINGIARDI** nato a AUGUSTA (SR) il 14/02/1976	**Nicola DANTI** nato a PELAGO (FI) il 06/09/1966
Nicola FRATOIANNI nato a PISA (PI)	**Jacopo Maria FERRI** nato a FIRENZE (FI)	**Mario CANINO** nato a ROMA (RM)	**Laura SERVA** nata a RIETI (RI)	**Bianca VERRILLO** nata a SANTA MARIA LA FOSSA (CE)

Figure A-9: voting paper fac-simile



1.7 Machine learning models

Table A-10: Variables used in the ML models

Variables
Degree
PhD
Political Experience
Female
Age (log)
Position List
Google Index (log)
Twitter (log)
Facebok Ads (log)
District
Party List

Figure A-10 – Feature importance of Random Forest model

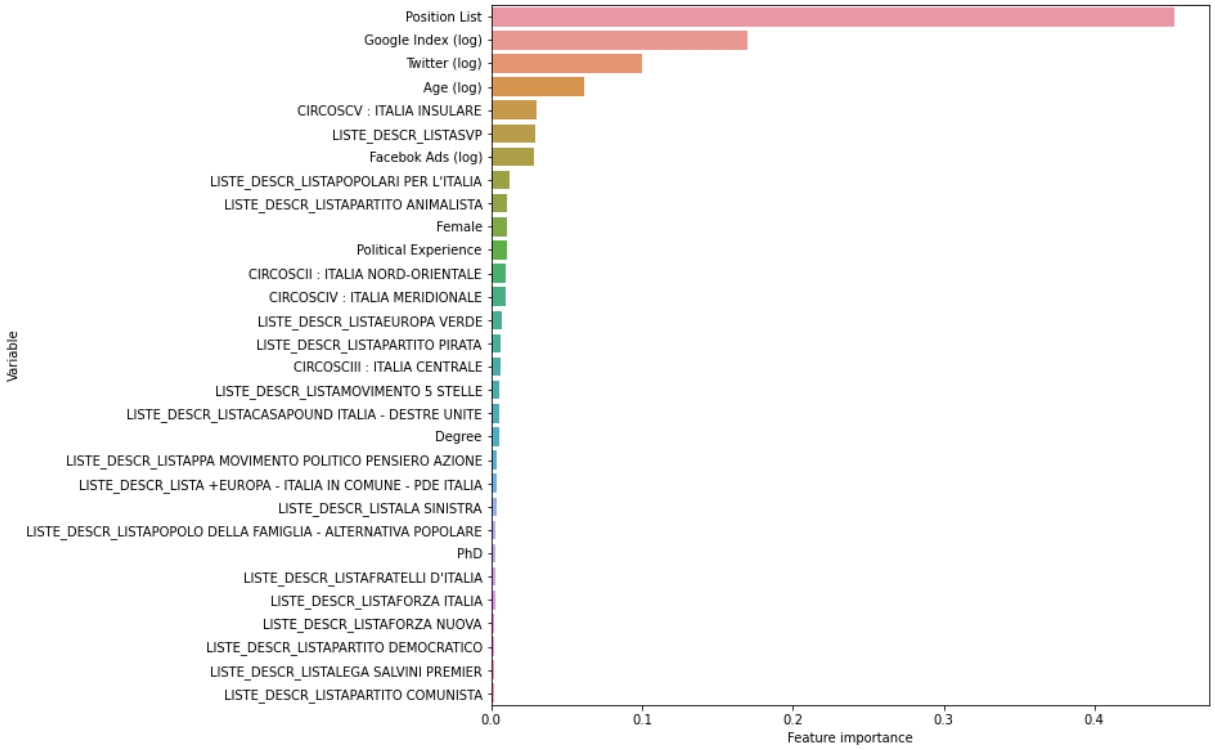
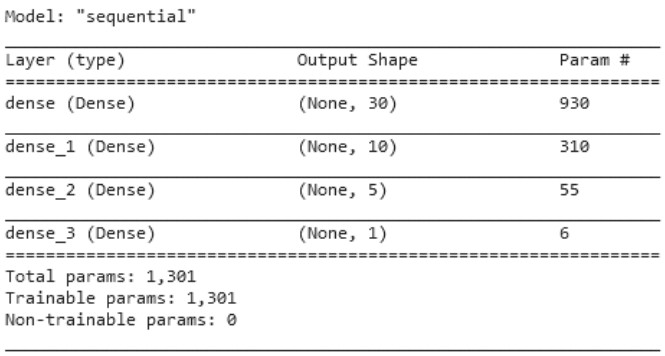


Figure A-7 – Neural Network model



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 30)                930
_________________________________________________________________
dense_1 (Dense)              (None, 10)                310
_________________________________________________________________
dense_2 (Dense)              (None, 5)                 55
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 6
=================================================================
Total params: 1,301
Trainable params: 1,301
Non-trainable params: 0
_________________________________________________________________
```

Table A-11 – Statistical significance of ML models

Model 1	**Model 2**	**T-stat**	**P-value**
mean	OLS	5.526885899	0.000
mean	Lasso	6.408692575	0.000
mean	Ridge	2.334439706	0.011
mean	Random Forest	6.839643303	0.000
mean	Neural Network	5.298255478	0.000
OLS	Lasso	-1.526973581	0.065
OLS	Ridge	-1.579959612	0.059
OLS	Random Forest	6.312362343	0.000
OLS	Neural Network	3.256366975	0.001
Lasso	Ridge	-0.952832067	0.171
Lasso	Random Forest	5.570822375	0.000
Lasso	Neural Network	3.373104866	0.001
Ridge	Random Forest	4.989219928	0.000
Ridge	Neural Network	3.64060874	0.000
Random Forest	Neural Network	-0.370206836	0.356

Computed following the SKlearn code at: https://scikit-learn.org/0.24/auto_examples/model_selection/plot_grid_search_stats.html

1.8 Electoral rules and Results of the European Elections

While each European country follows its electoral system, the common legal basis for the European Parliament elections is ordered by articles 14, 20, 22, and 223 of the Treaty on the Functioning of the European Union (TFEU). These rules norm the apportionment of seats to member states in proportion to the number of citizens[9]. In contrast, seat allocation within member states needs to follow a proportional system under plurinominal or uninominal preferences (devolved to member states' choice)[10].

Italy divided the apportionment of seats into 5 districts (North-West, North-East, Center, South, Islands) resulting from the aggregation of existing regions (See Figure 1).

Figure A. 1

Candidates' seats are allocated first on a proportional party share (computed at the national level) within the districts; second, the candidate's relative share of votes within a list. Each Party can present a maximum of 15 candidates in each district (with multiple candidacies allowed in a different district). Voters of each district can vote for one single Party and optionally give a preference for a minimum of 1 to a maximum of 3 candidates within the voted list. To express a preference for a candidate, the voters need to write the name or surname of the candidate on the voting paper (a list of the names of all the candidates is generally available outside the ballot box; an example of this ballot list and the voting paper are shown in the Appendix). When a voter expresses a vote for more than one candidate within a list, at least one has to be of the opposite sex of the first preference. The 2019 European Elections took place on different dates between 23rd and 26th May; in Italy, citizens have voted on Sunday May 26th between 7.00 a.m. and 11.00 p.m. Voter turnout at the National level reached 56 percent, which is lower compared to the result of the previous

[9] The ratio of MEP to citizen varies from 838.000 in Germany to 70.000 in Malta.
[10] Article 1 ,2002/772/EC

European Election (57 percent), and also compared to the last National Parliament Elections (72 percent), this lower turnout rate is well explained by proximity voting.

The turnout rate varies significantly between regions, from 36 percent in Sardegna to 67 percent in Umbria (central plot in Figure A-3). The ratio of the number of preferences over total valid votes at the national level is 48 percent[11] meaning each voter expressed an average of 0.48 preferences (out of a maximum of 3). That is, on average, at least half of the voters did not opt to express a preference. This behavior has quite some variance between regions and districts, ranging from 1.01 in Sicily to 0.21 in Piemonte (right plot in Figure A-3). These differences in the ratio of preferences/votes can have different interpretations. The "political resource" model (Andrè et al. 2012, Wauters et al. 2012) suggests that citizens that are less involved in politics, less educated, or simply less informed, are less likely to cast a preference vote[12]. A preference vote is thus interpreted as a more "*sophisticated*" vote compared to a "*simpler"* party vote. In light of this prediction, what we observe in Italy is evidence that is counter-intuitive to the prediction of the "political resource" model (right plot of Figure); in areas where "political resources" are notably lower -the south-, we observe a significantly higher average number of preferences cast[13]. A possible explanation to this puzzle would refer to licit and illicit party strategy that might differ by region or can be used as an indicator of the prevalence of exchange of votes (*voto di scambio*)[14]. The strong correlation between the turnout rate and the average number of preferences (left plot in Figure A-3) might serve as an additional accounting of the phenomena; it might be argued that the "turnout rate" acts as a selection filter for "*sophisticated*" voters, leading to a higher number of preferences in an area where turnout is lower. We leave the discussion of this relationship to future research.

[11] In average each voter has expressed 0.48 preference (note that this value can range between 0 and 3 according to the electoral law).

[12] "...Less affluent –working class and unemployed – citizens are less involved in social networks within which such information is exchanged and because they participate less in many areas of public life they are more difficult to reach in election campaigns. As such, they will be less likely to evaluate individual candidates and more likely to rely on simple decision rules such as party identification…". (Wauters, Verlet, & Ackaert, 2012)

[13] The southern regions of Italy see the coexistence of low "political resources" (low education, low consumption of newspaper, etc..) and high use of "sophisticated votes" (expressed in terms of number of casted preferences). The coexistence of this two phenomena is in contradiction with the model of "political resource" that would suggests the opposite relationship.

[14] This is the predominant view expressed in a rich literature focused on Italy by predominantly political scientists as well some leading opinionists that defined preferential voting the "flour of the devil" (*"farina del diavolo"* because of its link to illegal mafia-practices of the "*voto di scambio"* (Panebianco, 2014). There is in fact an historical trend in the rate of "preferences" expressed by southern regions, which is well summarized by De Luca (De Luca, 1997, 2011) This is mainly attributed to different party strategy due to regional contexts (such as promoting specific candidates instead of party), or voter behaviours that are specific to southern regions. The hypothesis of the "*voto di scambio"*, is supported by Cartocci (Cartocci, 1990), that use as a supporting argument the difference between the turnout rate in political elections and referendum. An empirical verification of the factor influencing the cast of a preference is given by Katz (Katz, 1985), that put under exam the three categories of votes proposed by Parisi et Pasquino (Parisi & Pasquino, 1985) (*voto di scambio, voto di opinion e voto di appartenenza*) which results in a partial contradiction of the hypothesis of the *"voto di scambio"* , while finding supporting evidence on the thesis of preference votes as a manifestation of a "sophisticated" vote (in line with the idea of "political resource" model).

Figure A.2

Vote share of the three main parties

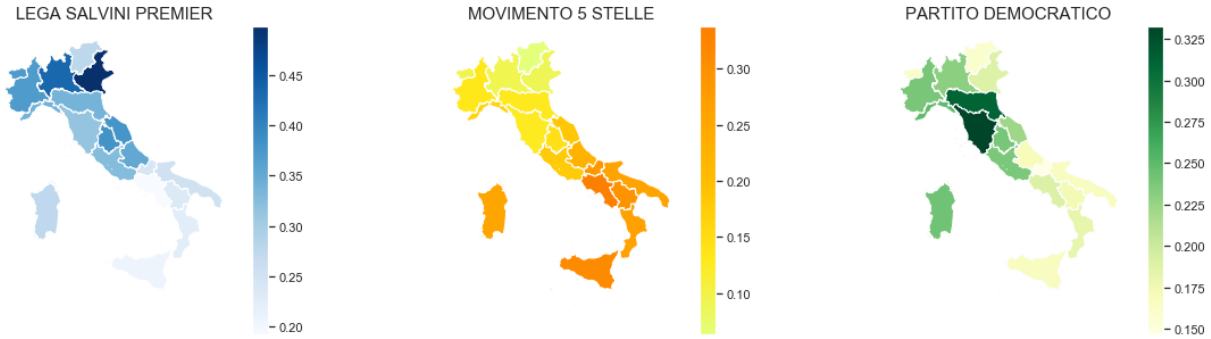
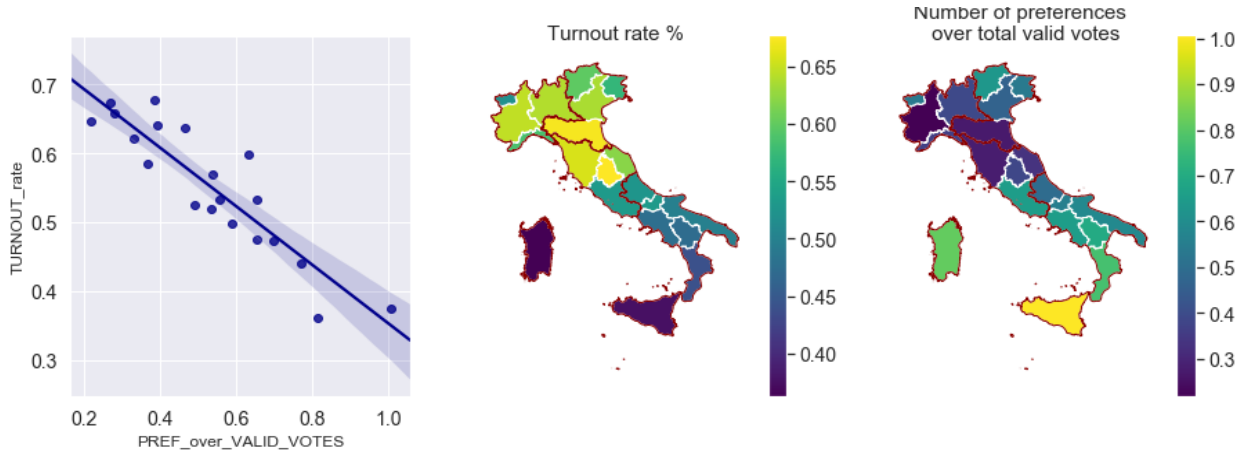


Figure A.3



The electoral results confirm support of the two parties leading the incumbent Government, LEGA and M5S (Movimento 5 Stelle), which together took the majority share of votes, respectively 34.26 and 17.06 percent, while the third Party (PARTITO DEMOCRATICO) took 22.74 percent of the votes at the national level. Overall, there are significant spatial differences in the three main parties, with LEGA more successful in Northern Regions[15], M5S in the south. At the same time, PARTITO DEMOCRATICO has been more successful in Central Regions[16].

[15] Historically LEGA (previously LEGA NORD and LEGA PADANIA), for more than one decade had a strong territorial orientation towards the objective of reaching autonomy for the northern regions of Italy.

[16] Tuscany and Emilia-Romagna are historically called “Le Regioni Rosse” (The Red Regions).

Appendix for : "*What characteristics of politicians affect electoral success? Estimates from a new dataset of European candidates' CV*"

Figure A- 2: Distribution of votes in the Survey Sample

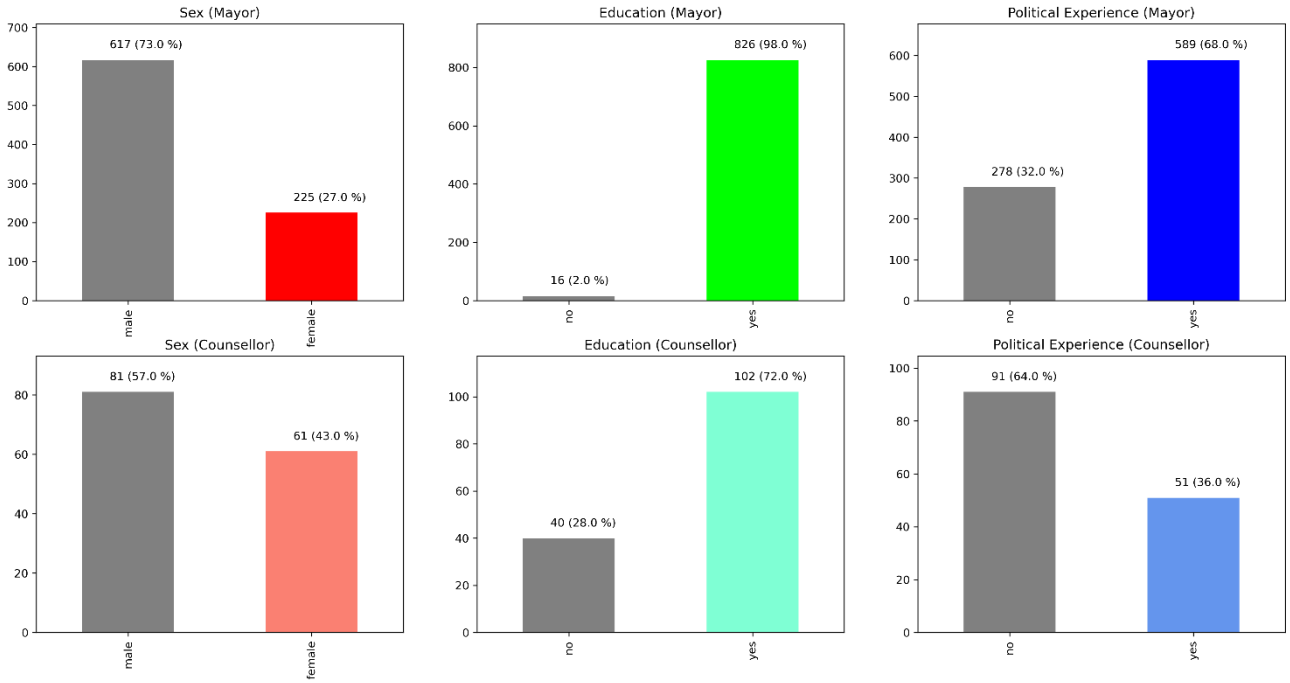


Figure A- 1: Coverage of the literature on the effect of Transparency on voting behavior (based on table A-1)

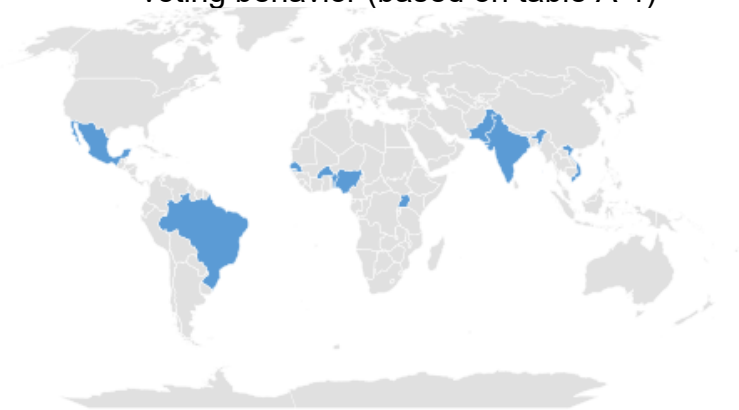


Table A- 1:Coverage of the literature on the effect of Transparency on voting behaviour

Study	Country
Information's Effect on Voter-Driven Accountability in Senegal. *American Journal of Political Science*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/ajps.12591	Senegal
Chong, A., De La O, A. L., Karlan, D., & Wantchekon, L. (2015). Does corruption information inspire the fight or quash the hope? A field experiment in Mexico on voter turnout, choice, and party identification. *The Journal of Politics*, *77*(1), 55–71.	Mexico
Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., Nellis, G., Adida, C. L., Arias, E., Bicalho, C., Boas, T. C., & others. (2019). Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*, *5*(7), eaaw2612.	
Malesky, E., & Schuler, P. (2020). Single-party incumbency advantage in Vietnam: A conjoint survey analysis of public electoral support. *Journal of East Asian Studies*, *20*(1), 25–52.	Vietnam
Pande, R., Banerjee, A., Green, D., & McManus, J. (2014). Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India. *Political Communications*, *31*, 391–407.	India
Pande, R., Banerjee, A., & Walton, M. (2012). *Delhi's Slum-Dwellers: Deprivation, Preferences, and Political Engagement among the Urban Poor*. International Growth Center.	India
Bhavnani RR. 2009. Do electoral quotas work after they are withdrawn? Evidence from a natural experiment in India. Am. Polit. Sci. Rev. 103(1):23–35	India
Beaman L, Chattopadhyay R, Duflo E, Pande R, Topalova P. 2009. Powerful women: Does exposure reduce bias? Q. J. Econ. 124(4):1497–540	India
Collier P, Vicente P. 2008. Votes and violence: evidence from a field experiment in Nigeria.Work. Pap., Cent. Study Afr. Econ.	Nigeria
Gine X, Mansuri G. 2010. Together we will: evidence from a field experiment on female voter turnout in Pakistan. Work. Pap., World Bank, Washington, DC	Pakistan
Ferraz C, Finan F. 2008. Exposing corrupt politicians: the effect of Brazil's publicly released audits on electoral outcomes. Q. J. Econ. 123(2):703–45	Brazil
Wantchekon L. 2009. Can informed public deliberation overcome clientelism? Experimental evidence from Benin. Work. Pap., New York Univ	Benin
Wantchekon L. 2003. Clientelism and voting behavior: evidence from a field experiment in Benin. World Polit. 55(3):399–422	Benin

When Does Information Influence Voters? The Joint Importance of Salience and Coordination Claire Adida, Jessica Gottlieb, Eric Kramon, and Gwyneth McClendon, Comparative Political Studies (2019).	Benin
Accountability Backlash: Negative Electoral Responses to Public Service Provision in Brazil Taylor C. Boas, F. Daniel Hidalgo, and Guillermo Toral	Brazil
Citizens at the Council: Comparing the Impact of Mediated Information and First-Hand Experience on Voter Turnout in Municipal Elections Marcus Holmlund, Malte Lierl	Burkina Faso
Using Local Networks to Increase Accountability Simon Chauchard, Neelanjan Sircar	India
Common Knowledge, Relative Performance, and Political Accountability Eric Arias, Horacio Larreguy, John Marshall, Pablo Querubin	Mexico
Meet the Candidates: Information and Accountability in Primary and General Elections Melina Platas, Pia Raffler	Uganda
Repairing Information Underload Mark Buntaine, Sarah Bush, Ryan Jablonski, Daniel Nielson, Paula Pickering	Uganda

QUESTIONNAIRE

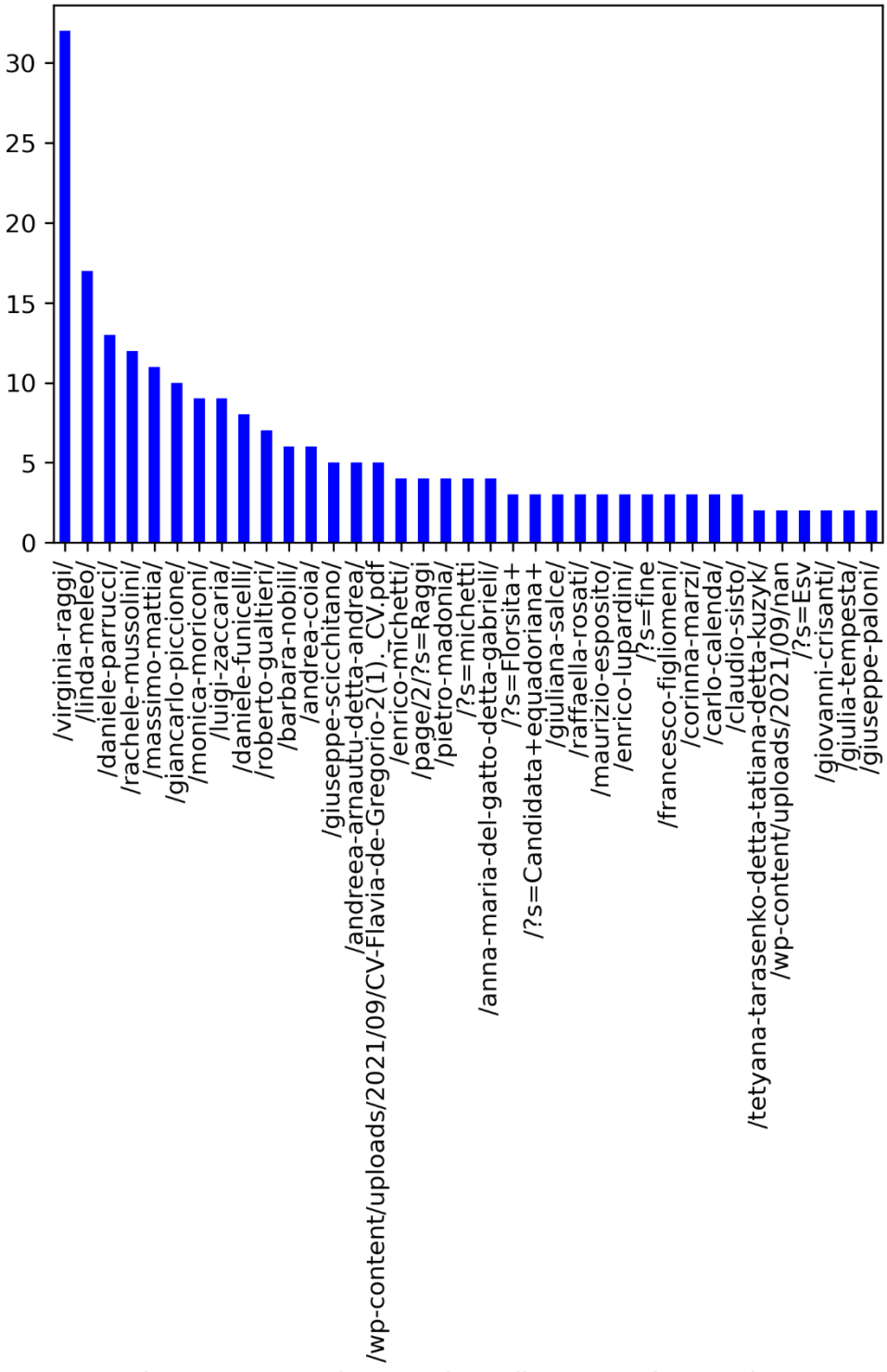
Questionario round 1

Codice domanda istat (Aspetti della vita	Codice Domanda	Tema	Obbligatoria	Testo domanda	tipologia risposta (U=unica, M=multipla)	1	2	3	4	5	6	7	8
		demografia											
	1	D1_1	NO	**Sesso**	U	maschio	femmina	preferisce non rispondere	preferisce descrivere				
	2	D1_2	NO	**età**	U	anno							
	3	D1_3	NO	**Livello d'istruzione**	U	laurea e post-laurea	diploma	licenza di scuola media	licenza di scuola elementare o nessun titolo di studio				
	4	D1_4	NO	**Municipio di residenza**	S								
		discute di politica ?											
20,1	5	D1_5	SI	**con quale frequenza le capita di parlare di politica**	U	tutti i giorni	qualche volta alla settimana	Una volta alla settimana	Qualche volta al mese (meno di quattro volte)	Qualche volta l'anno	Mai		
20,2	6	D1_6	NO	**è iscritto a:**	U	partiti politici	Organizzazioni sindacali (sindacati	comitati di quartiere	associazioni o gruppi di volontaria	associazioni ecologiste, per i diritti civili, per la	associazioni culturali, ricreative o di altro	Associazioni professionali o di categoria	
		partecipa alla politica ?											
20,4	7	D1_7	NO	**Negli ultimi 12 mesi:**	M	ha partecipato a un comizio	ha partecipato a un corteo	ha sentito un dibattito politico	ha partecipato ad una cena/aperitivo elettorale	ha dato soldi a un partito (per sottoscrizione, iscrizione, sostegno)	Ha svolto attività gratuita per associazioni o gruppi di volontariato	ha svolta attività gratuita per una associazione non di volontariato	ha svolto attività gratuita per un partito
20.6	8a	D1_8a	SI	**Con quale frequenza si informa dei fatti della politica italiana**	S	tutti i giorni	qualche volta alla settimana	Una volta alla settimana	Qualche volta al mese (meno di quattro volte)	Qualche volta l'anno	Mai		
20.8	8b	D1_8b	SI (solo se NON ha rispost "mai" alla 8a)	**Lei si informa dei fatti della politica anche attraverso internet**	U	No	Si						
20.9	8c	D1_8c	SI (solo se NON ha rispost "mai" alla 8a)	**In che modo si informa di politica attraverso internet**	M	Leggo o scarico giornali, news, riviste online	Ascolto radio e tv online	Tramite siti di partiti politici	Tramite i social network (es: Facebook, Twitter, Youtube)	Tramite blog, newsgroup, o forum di discussione online	Altro		
20,10	9	D1_9	SI (solo se ha risposto "Mai" alla domanda 8a	**Quali sono i motivi prevalenti per cui non si informa mai di politica**	M	non mi interessa	non ho tempo	è un argomento troppo complicato	sono sfiduciato dalla politica italiana	altro			
		Candidato ideale											
	I0	D1_I0	SI	**Alcuni studi, ritengono che la competenza dei rappresentanti politici, in termini di livello d'istruzione, esperienza professionale/amministrativa ed esperienza politica, possa avere un impatto benefico sull'amministrazione delle città. Quanto è d'accordo con questa affermazione ?**	U	molto	abbastanza	poco	per niente				
	I1	D1_I1	SI	**Quale dei seguenti fattori secondo lei un candidato ideale dovrebbe avere:**	M	un alto livello d'istruzione	un alto livello di esperienza politica o esperienza in associazioni , comitati, etc..	un ideologia ed una visione politica forte	un alta capacità di comunicare	un etica incorruttibile	una fedina penale pulita	altro: (specificare)	
		Intenzioni di voto											
	I2	D1_I2	NO	**Ha già deciso per quale candidato sindaco votare, se si quale ?**	U	candidato 1	candidato 2	candidato 3	ancora non ho deciso	non andrà a votare			
	I3	D1_I3	NO	**Ha già deciso se esprimere anche una preferenza per i candidati al consiglio comunale ? Se si quale**	M	candidato 1	candidato 2	candidato 3	non esprimerò nessuna preferenza				
	I4	D1_I4	NO	**Conosce almeno il nome di un candidato al consiglio comunale ?**	M	candidato 1	candidato 2	candidato 3	non conosco nessun nome				
		Issues											
	I5	D1_I5	NO	**Secondo lei, quali sono le priorità che i candidati dovrebbero risolvere ?**	M	-Traffico e parcheggi (auto e scooter)	-Sicurezza (polizia, movida, e strade)	-Trasporto pubblico (bus, tram e metro)	-Decoro urbano (parchi, pulizia strade, strutture abbandonate)	-Corruzione e legalità	-Economia e Lavoro	-Emergenza abitativa	-Inquinamento atmosferico e ambientale
		Trasparenza politica											
	I6	D1_I6	SI	**Quanto riterrebbe utile avere delle norme che obblighino i partiti ad adottare maggiore trasparenza ? Ad esempio pubblicare i curriculum dei candidati, conflitti di interesse, dichiarazione dei redditi e patrimoni, etc…**	U	Molto utile	Abbastanza utile	poco utile	inutile				

Questionario round 2

	Obbligatoria		SINGOLA (S) / MULTIPLA (M)	SCELTA DI VOTO					
D2_1	SI	è andato a votare per le elezioni del sindaco ?	S	si	no				
D2_2	solo se domanda 1 risponde si	Se è andato a votare, Per quale candidato sindaco ha votato ?	S	candidato 1	candidato 2	candidato 3	etc..	voto nullo/scheda bianca	
D2_3	solo se domanda 1 risponde si	Ha espresso anche una o due preferenze per un consigliere?	S	si	no				
D2_4	solo se domanda 3 risponde si	Per quali candidati consiglieri ha espresso la preferenza	M	candidato 1	candidato 2	…			
raccolta di informazioni									
D2_5	SI	Ha visitato il sito elezionitrasparentiroma.it ?	S	Si	No				
D2_6	solo se domanda 5 risponde si	Ha ritenuto utili tali informazioni ?	S	Si	No				
D2_7	solo se domanda 6 risponde si	Quali informazioni ha ritenuto più utili	M	Il livello di istruzione del candidato	L'esperienza Professionale del candidato	L'esperienza Politica del candidato	L'età	altro	
D2_8	SI	Se ha espresso un voto di preferenza per un candidato consigliere, può gentilmente dirmi come è venuto a conoscenza della sua candidatura ?	M	Conoscevo personalmente il candidato	Conoscevo indirettamente il candidato tramite amici e/o famigliari	Ho partecipato ad un comizio/incontro con il candidato	Ho letto il nome del candidato su di un manifesto e/o volantino	Ho raccolto informazioni sul candidato su internet	
D2_9	solo se domanda 1 risponde si	è soddisfatto del suo voto (soddisfazione) ?	M	Molto	abbastanza	poco	per niente		

Figure A- 3: most visited pages by respondents



1

Figure A- 4: Representativness of the Sample (internet users)

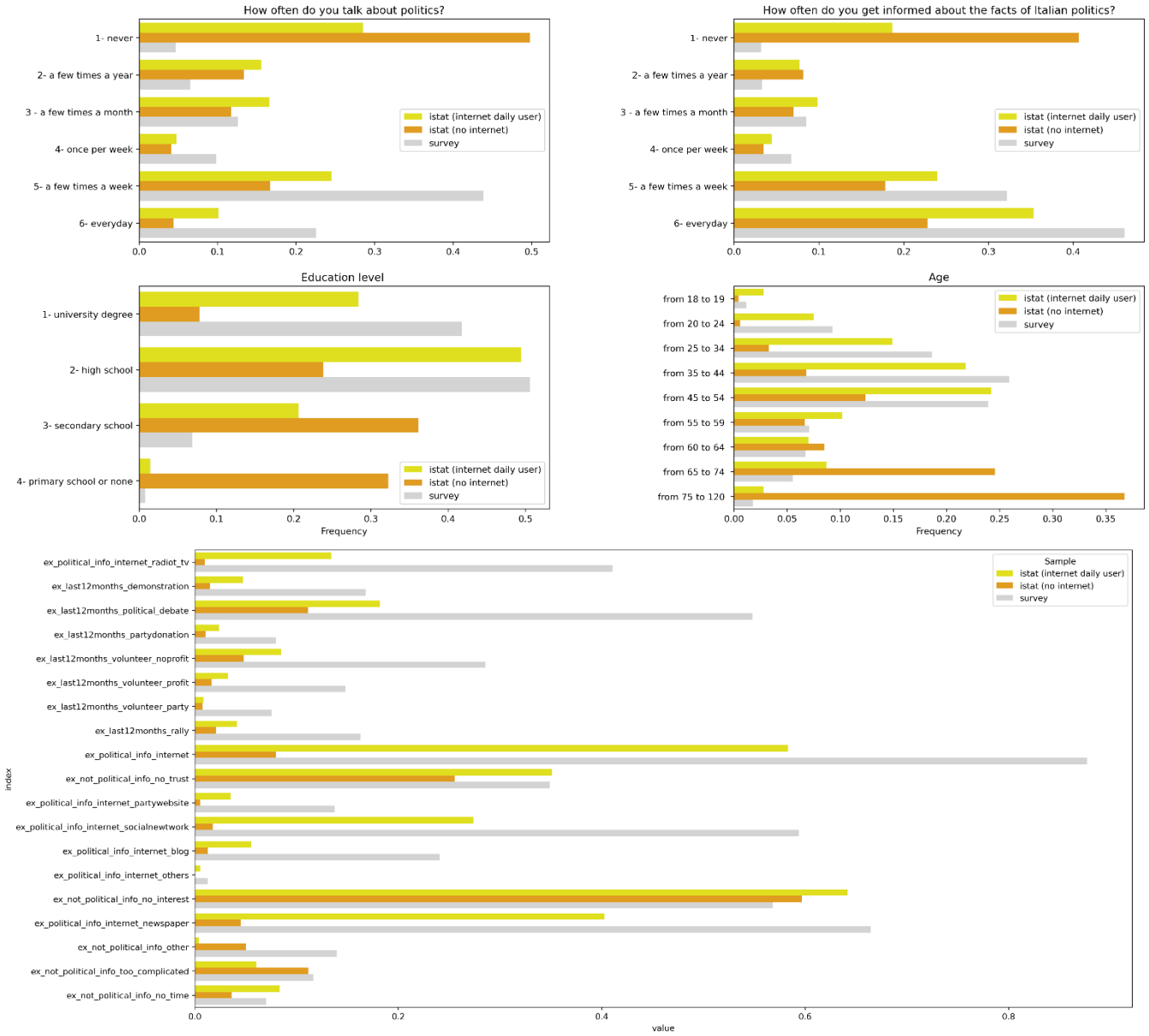


Figure A- 5: Interaction with Populist (intention to vote for Raggi or Michetti)

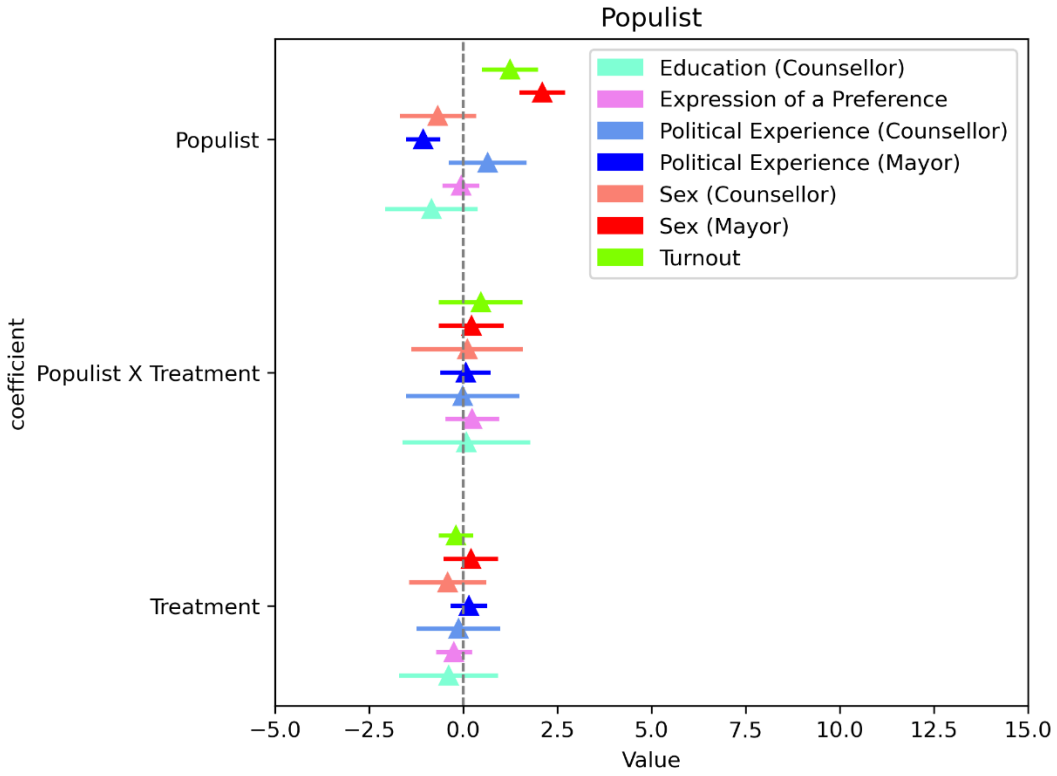


Figure A- 6: Time spent on the website

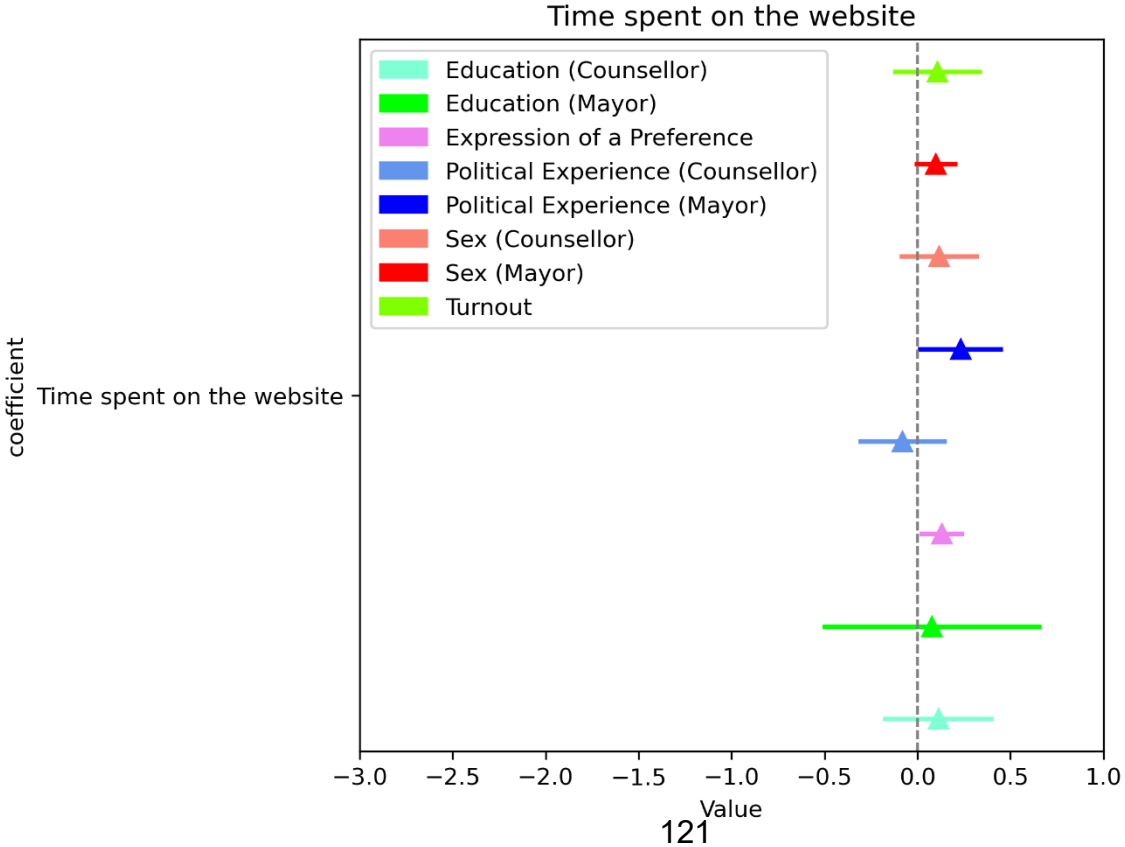


Table A- 2: Turnout among samples

Sample	Treatment	Turnout
2nd questionnaire only	no	0.866
1st and 2nd questionnaire	no	0.867
	yes	0.860

Table A- 5: robustness tests (expression of a preference)

	post_counsellor_vote I	t_counsellor_vote_alternati	post_counsellor_vote II	post_counsellor_vote III	post_counsellor_vote IIII
Treatment	-0.0731	0.1671	-0.0236	0.1036	0.1065
	(0.1654)	(0.1873)	(0.1930)	(0.2038)	(0.1997)
constant	-1.1309***	-1.6939***	-1.5871***	-2.0009***	-1.9052***
	(0.1001)	(0.1187)	(0.1476)	(0.1726)	(0.1688)
›w at least one name of a counsellor				1.5857***	2.0407***
				(0.2250)	(0.2008)
Will vote a counsellor			1.8982***	1.0930***	
			(0.2130)	(0.2461)	
N	866	866	680	673	675

Table A- 5: robustness tests (turnout)

	post_turnout I	post_turnout II	post_turnout III	post_turnout IIII	post_turnout IIIII
Treatment	-0.0397	-0.0544	-0.0470	-0.0850	-0.1097
	(0.1885)	(0.2092)	(0.2353)	(0.2440)	(0.2406)
constant	1.8563***	2.1493***	2.2593***	2.9901***	1.5796***
	(0.1166)	(0.1671)	(0.1690)	(0.2295)	(0.2254)
Will not vote			-3.6507***	-4.3637***	
			(0.3918)	(0.4193)	
Not yet decided who to vote		-0.9665***		-1.7930***	
		(0.2171)		(0.2599)	
Made a donation to a party					1.1499
					(0.7405)
political info on internet newspaper					0.4896**
					(0.2459)
Member of a local committee					0.7966*
					(0.4158)
N	1008	792	792	792	684

Table A- 5: robustness tests (characteristics of candidates)

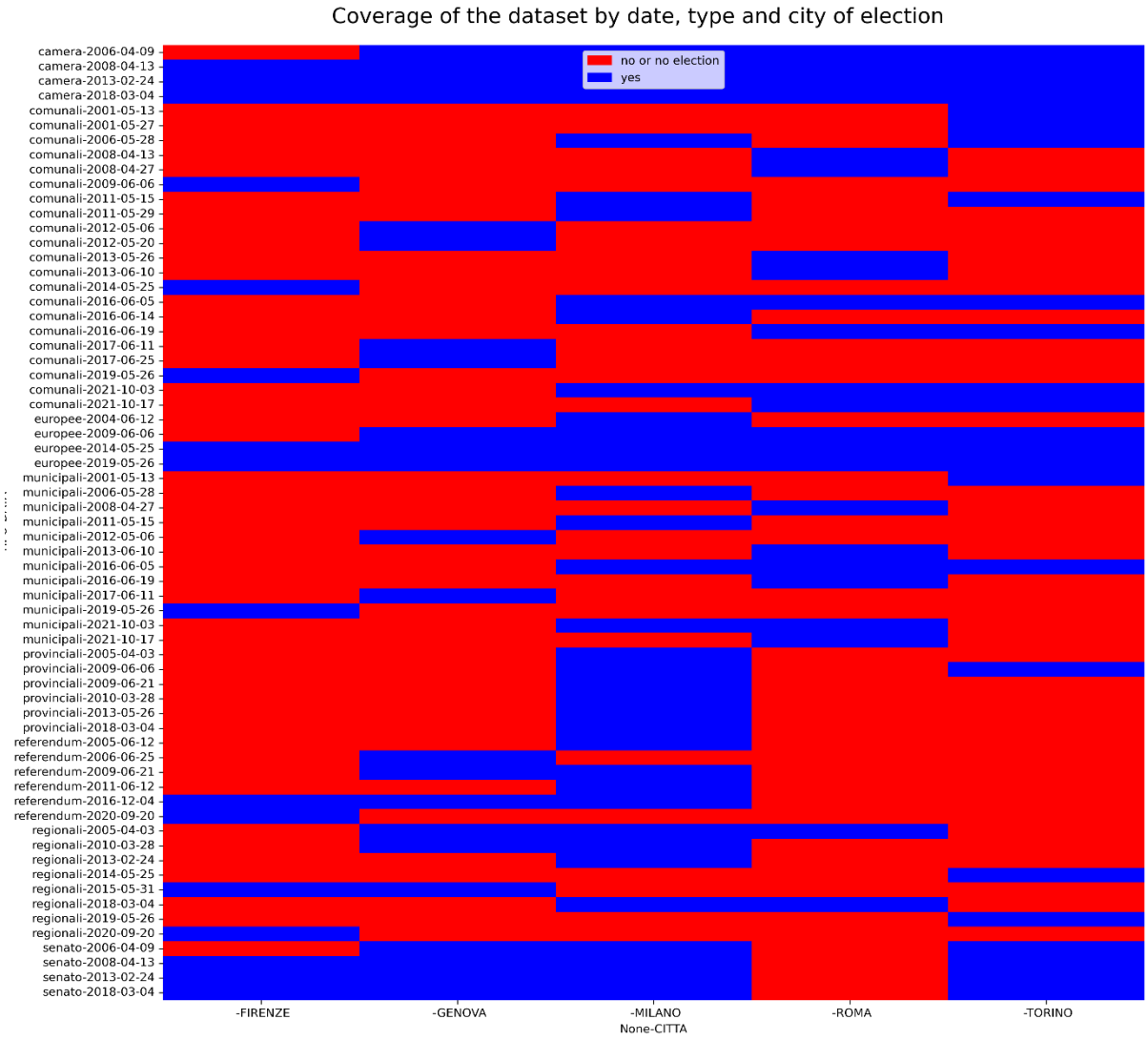
	sex_mayor I	sex_mayor II	istruzione_mayor I	istruzione_mayor II	esperienza politica_mayor I	esperienza politica_mayor II	sex_counsellor I	sex_counsellor II	istruzione_counsellor I	istruzione_counsellor II	esperienza politica_counsellor I	esperienza politica_counsellor II
Treatment	0.2443	0.3698*	-0.5199	-0.7526	0.1027	0.1327	-0.4711	-0.0625	-0.0950	-0.4516	0.0214	-0.3733
	(0.1591)	(0.1950)	(0.5052)	(0.6277)	(0.1510)	(0.1817)	(0.3498)	(0.4389)	(0.3783)	(0.4749)	(0.3557)	(0.4292)
constant	-1.1037***	-1.0935***	4.1705***	3.7063***	0.7126***	0.6709***	-0.0953	0.0498	0.9754***	0.8287	-0.5878***	-1.2087**
	(0.1008)	(0.2156)	(0.3563)	(0.6455)	(0.0915)	(0.1981)	(0.2185)	(0.4942)	(0.2447)	(0.5333)	(0.2277)	(0.5393)
Not yet decided who to vote		-0.8536***		-0.4161		-0.7719***		0.1289		-0.1259		0.2717
		(0.3002)		(0.6834)		(0.2139)		(0.6400)		(0.6708)		(0.6094)
Know at least one name of a counsellor		0.3249		1.3043		0.0212		-0.8120*		0.4824		0.6837
		(0.2138)		(1.0533)		(0.2056)		(0.4510)		(0.4776)		(0.4615)
Made a donation to a party		-0.7751**		25.6930		0.6379*		0.2234		-0.8415		0.4549
		(0.3792)		(360972.5451)		(0.3625)		(0.6872)		(0.7269)		(0.6704)
Get political info on internet newspaper		-0.1596		0.6211		0.2960		-0.3896		0.3754		0.5923
		(0.2079)		(0.6012)		(0.1930)		(0.4651)		(0.4985)		(0.4788)
Member of a local committee		0.3591		-0.1737		-0.1920		0.6848		0.0777		0.1210
		(0.2397)		(0.7939)		(0.2329)		(0.5818)		(0.6596)		(0.5707)
N	842	581	842	581	867	599	142	104	142	104	142	104

Table A- 6: robustness test (mayor votes)

	vote_enrico michetti I	vote_enrico michetti II	vote_enrico michetti III	vote_carlo calenda I	vote_carlo calenda II	vote_carlo calenda III	vote_virginia raggi I	vote_virginia raggi II	vote_virginia raggi III	vote_roberto gualtieri I	vote_roberto gualtieri II	vote_roberto gualtieri III
Treatment	-0.0321	-0.2579	-0.1979	-0.1827	0.0150	-0.0015	0.3216**	0.7673**	0.8056**	-0.1209	-0.2910	-0.3234
	(0.1624)	(0.2668)	(0.2692)	(0.1951)	(0.2856)	(0.2878)	(0.1608)	(0.3147)	(0.3221)	(0.1605)	(0.2389)	(0.2412)
Intention to vote for Virginia Raggi		-1.5270***	-1.5292***		-2.2372***	-2.2729***		3.8421***	3.9533***		-1.7028***	-1.7157***
		(0.3908)	(0.3924)		(0.7543)	(0.7545)		(0.3374)	(0.3496)		(0.3700)	(0.3713)
Intention to vote for Carlo Calenda		-1.4448***	-1.4472***		3.5136***	3.4732***		-2.3581**	-2.2582**		-0.9723**	-0.9669**
		(0.4984)	(0.5004)		(0.3747)	(0.3763)		(1.0354)	(1.0384)		(0.3803)	(0.3818)
Intention to vote for Enrico Michetti		3.3769***	3.3645***		-0.4071	-0.4301		-2.2291***	-2.1722***		-2.8065***	-2.8079***
		(0.3326)	(0.3342)		(0.4125)	(0.4133)		(0.7526)	(0.7575)		(0.6091)	(0.6099)
Intention to vote for Roberto Gualtieri		-2.9440***	-2.9445***		-0.4147	-0.4399		-1.8036***	-1.7175***		2.7618***	2.7480***
		(0.7364)	(0.7373)		(0.4124)	(0.4128)		(0.6321)	(0.6372)		(0.2985)	(0.3004)
constant	-1.0912***	-1.1575***	-0.7645*	-1.5984***	-2.1339***	-2.3635***	-1.2551***	-2.4182***	-2.0506***	-0.9859***	-0.8541***	-1.0608**
	(0.0991)	(0.2181)	(0.4412)	(0.1149)	(0.2825)	(0.5413)	(0.1035)	(0.3066)	(0.5856)	(0.0966)	(0.2012)	(0.4413)
Get political info on internet			-0.4605			0.2935			-0.5243			0.2510
			(0.4341)			(0.5102)			(0.5614)			(0.4309)
N	867	681	674	867	681	674	867	681	674	867	681	674

Appendix for : "*Sezioni Elettorali Italiane (SEI): A new database of Italian electoral results geocoded at the precinct level.*"

Figure A 1



The figure shows the coverage of the dataset by type, date and city of elections. Note that municipality and regional elections do not take place always at the same time in each city. Thus, some elections might be missing simply because of this. For national elections (European, Camera and Senato) we have an almost full coverage.

Data Cleanup

The geocoder results might often provide wrong coordinates of addresses. How can we deal with these errors? Here, we will provide a detailed explanation that primarily relies on rule-of-thumb rules derived from manual inspection of these errors.

Before explaining our cleaning procedure, we want to make some crucial premises. First, different geocoders and contexts (city or country) might require different approaches. Second, while the procedure might be prone to errors, the final results might not be affected, especially when we employ the census procedure to clean out residual error points through the “majority” rule (step 3 of the census procedure). Third, we can evaluate the quality of the results by comparing the final precinct polygons against an externally hand-drawn dataset (such as we do in the main corpus of the paper for the city of Rimini). Of course, if we had an external dataset of geocoded addresses, we could evaluate the errors of the geocoder, but, in that case, we would not need a geocoder at all! If the project *Archivio nazionale degli stradari e dei numeri civici (ANSC)* is delivered in the future we would no longer need anymore a geocoder. In this example, we will look at the city of Turin, but the same reasoning applies for other cities.

Cleaning points outside a city or a municipality (Figure A-2)

The first thing we do is check whether the coordinates are in the municipality of Turin. In the example below (Figure A-2), we can see that some points (those highlighted in yellow) fall well outside the municipality border (retrieved from ISTAT). These points are errors in the geocoder. At the same time, some points fall just after or over the borders (see the green markers). Those points are less likely errors. They might be points that fall precisely or very close to the boundaries. To avoid deleting these points, we slightly buffer city borders. After buffering the city borders, we keep all points within the city. By doing this, we have eliminated all points that are surely geocoder mistakes because they fall into the wrong municipality.

Cleaning errors within the municipality

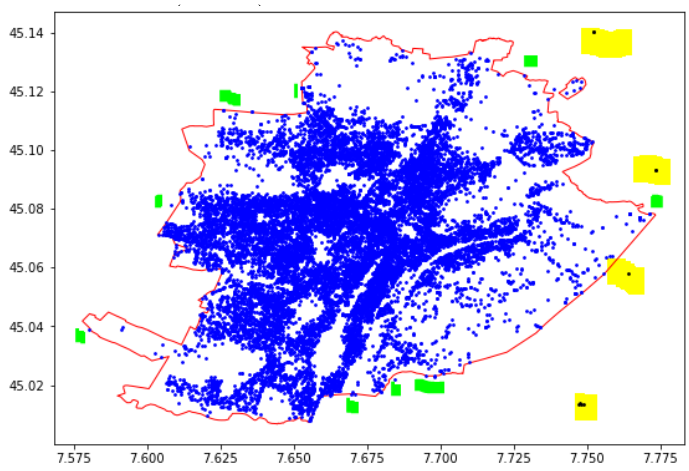
Identifying geocoder errors within the municipality is much more difficult. We distinguish two fundamental cases.

The first case is when the geocoder cannot find the address. In this case, most of the geocoder will always return to the centroid of the city. How do we identify those errors? The fastest way is to sort all geocoded points by their frequency. If some addresses have abnormal frequencies, those are most likely errors. Table 1 shows an example of Turin and Palermo, highlighting in yellow the mistakes.

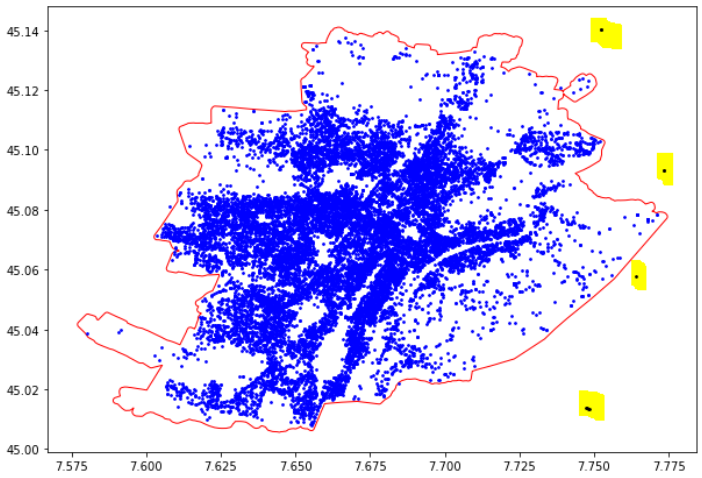
The second case builds on a simple intuition. A point should be close to all other points of the same precinct. If a point is far from all other points of the same precinct, it is likely an error. To identify these errors, we compute an index of how much a point is an outlier in each precinct. Figure A-4 shows the calculation with a graphical example. By manual inspection of these cases, we set a threshold as a rule of thumb where points with an index larger than 3 will be labeled as mistakes and dropped. For precincts with only two points, this procedure is not applicable. In these cases, we simply compute distances between points of the same precinct and check which have the highest value. When this value is higher than plausibly expected (3km), we inspect these points are not wrong manually.

Figure A 2: delete points outside city borders

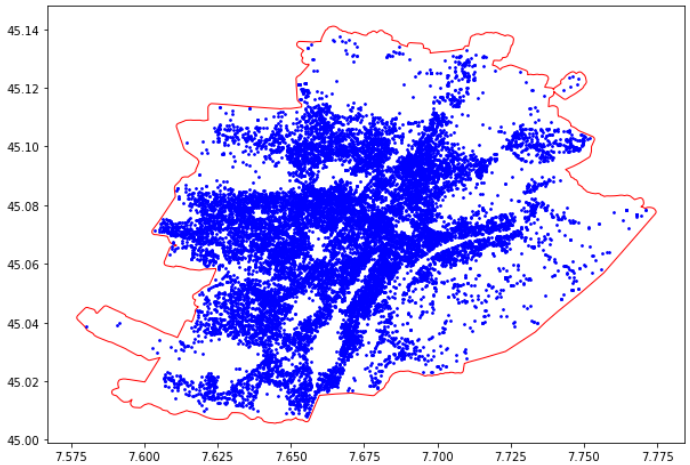
Check points that fall outside city borders



Slightly buffer the red



Drop all points outside red



a

Figure A 3: geocoder errors in Palermo and Turin

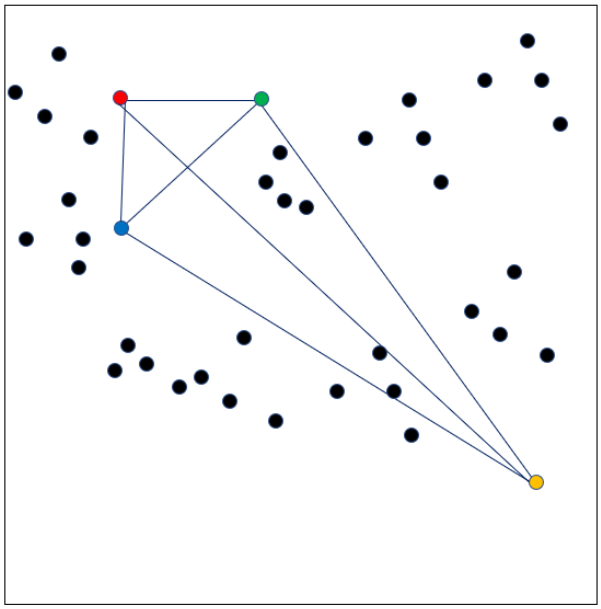
```
df_new.address.value_counts().head(10)
```

```
Palermo, PA, Italy                                              364
Via Principe di Palagonia, 2, 90145 Palermo PA, Italy           103
Cortile, 90135 Palermo PA, Italy                                 91
Cortile Regina, 90134 scalilla PA, Italy                         33
Corso Camillo Finocchiaro Aprile, Palermo PA, Italy              14
Via del Fondo, 90135 Palermo PA, Italy                           12
Via dei Nebrodi, Palermo PA, Italy                               12
Cortile Scimeca, 2, 90100 Palermo PA, Italy                      12
Passaggio Giuseppe De Felice Giuffrida, 90121 Palermo PA, Italy  12
Baglio Polito, 90146 Palermo PA, Italy                           11
Name: address, dtype: int64
```

```
df_new["address"].value_counts()
```

```
Via Conte di Roccavione, 10147 Torino TO, Italy          25
Corso Lucio Quinzio Cincinnato, 10151 Torino TO, Italy   20
Superga TO, Italy                                        15
S.da Lanzo, 10148 Torino TO, Italy                       15
10148 Turin, Metropolitan City of Turin, Italy           14
                                                         ..
Via Revello, 34, 10139 Torino TO, Italy                   1
Via Revello, 17, 10139 Torino TO, Italy                   1
Via Revello, 47a, 10139 Torino TO, Italy                  1
Via Revello, 39, 10139 Torino TO, Italy                   1
Viale Virgilio, 107, 10126 Torino TO, Italy               1
Name: address, Length: 16439, dtype: int64
```

Figure A 4: label wrongly geocoded points



- The coloured points (red, green, blue and orange) are labelled as belonging to the same precinct.
- The lines measure the distance between those points.
- All black points belong to other precincts.

It is clear from the figure that the orange point has been wrongly located by the geocoder as it is far away all other points. In other words it is an outlier in that precinct. To identify such issues, we construct a measure C (ratio) that will tell us how much a point is an outlier in each precinct.

We compute the matrix of distances for the four coloured points and then we compute the average distance for each point and an average of the whole group. The ratio will give a measure of how much a point is an outlier. In this case we can clearly see that the orange point as an higher ratio compared to the others.

distance matrix

point	red	green	blue	orange
red		3.3	2.78	9
green	3.3		2.92	8.69
blue	2.78	2.92		5.82
orange	9	8.69	5.82	

A	**average of each points**	5.03	4.97	3.84	7.84
B	**average distance of all points**	*5.42*	*5.42*	*5.42*	*5.42*
C	**ratio (A/B)**	**0.93**	**0.92**	**0.71**	**1.45**