GRADUATE SCHOOL OF ENGINEERING AND BASIC SCIENCES

Master in Big Data Analytics, Thesis

PROCEDURE TO ESTIMATE THE INTENTION TO VOTE IN SPAIN THROUGH NEURAL NETWORKS

MAURINE LÉONTINE BESSIÈRE
100439778

SUPERVISOR: ANDRÉS M. ALONSO FERNÁNDEZ

JUNE 2021

Declaration

I hereby declare that this Master Thesis is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university. I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year.

Abstract

In this work, we study the use of Artificial Neural Networks (ANN) for the imputations of missing values. More specifically, we designed and carried out a procedure of imputation using Multilayer Perceptrons with the final aim of estimating voting intentions in Spanish general elections. The work is based on the February 2021 barometer survey of the Centro de Investigaciones Sociológicas (CIS).

Acknowledgements

I would first like to thank my thesis supervisor Andrés M. Alonso Fernández. Prof. Alonso Fernández was always of a great help whenever I had a question about my research or writing. He suggested many times very useful and ingenious workarounds when I encountered situations I could not see how to solve. He/She consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

I would also like to acknowledge Prof. Ricardo Aler Mur for the high quality of his lectures on methodology in machine learning, to which I resorted many times during my work.

Finally, I must express my gratitude to my professors of my previous master in quantitative sociology, which endowed me with the capacities to carry out the preliminary analysis of this thesis.

Contents

List of Figures

List of Tables

Nomenclature

Adam	Adaptative Moment Estimation
AUC	Area Under the ROC Curve
ANN	Artificial Neural Network
BNG	Bloque Nacionalista Galego
CCa-NC	Coalición Canaria - Nationalista Canario
CIS	Centro de Investigaciones Sociológicas
EAJ-PNV	Euzko Alderdi Jeltzalea-Partido Nacionalista Vasco
ERC	Esquerra Republicana de Catalunya
JxCat	Junts per Catalunya
MAR	Missing At Random
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
MI	Multiple Imputation
MICE	Multiple Imputation with Chained Equations
MLP	MultiLayer Perceptron
NA	Not Available
NA+	Navarra Suma
PACMA	Partido Animalista Contra el Maltrato Animal
PP	Partido Popular
PRC	Partido Regionalista de Cantabria
PSOE	Partido Socialista Obrero Español
UPN	Unión del Pueblo Navarro

1 Introduction

Failures in election polls are recurrent. The history will remember the United States presidential election of 1936 that opposed Alfred Landon to the incumbent President, Franklin D. Roosevelt. The *Literary Digest*, a magazine that had been predicting correctly the winners of presidential elections for the last 20 years, conducted its most expensive poll on 2.4 million voters. They predicted Landon would get 57% of votes. The shock was high when Roosevelt was elected with 62%. At the same time, George Gallup promised he would predict the winner of the 1936 presidential election by interviewing only 50,000 people and forecasted a win for Roosevelt at 54%. After this event, the importance of correct sampling and non-response correction came to light, leading to the modern opinion polls.

This historic example demonstrates the difficulty of estimating correctly voting intentions and the necessity to correct the many biases that can occur on a survey. Especially, the *Literary Digest* did not take into account non-responses during their surveys, leading to an overestimation of Landon's partisans. Nowadays, statisticians are much aware of the many difficulties arising from such studies and the need to account for them.

In Spain, different organisms carry out voting intentions pools. In this study, we choose to focus on the surveys carried out by the Centre for Sociological Research (CIS, Centro de Investigaciones Sociológicas), previously known as the Spanish Institute of Public Opinion (IOP, Instituto Español de la Opinión Pública) founded in 1963. It is an autonomous public organism charged of scientific research on the Spanish society, specifically through periodic surveys. More pre-

cisely, we will study the data of its monthly Barometers, designed to measure Spanish public opinion throughout the time. Phone interviews permit gathering extensive social, demographic and opinions of around 4,000 people of 18 years old or more.

In the public sphere, the CIS is particularly known for its voting intentions estimated from the barometers, which are very commented by media after each publication. Their results often provoke a lively argument around the "cooking" of the Centre for reconstructing such estimations. On the one hand, such randomized voluntary surveys bring many difficulties around the representativeness of interviewees and their opinions, which are accounted for by specific weighting. On the other hand, the uniqueness of electoral opinion polls bring many obstacles, such as non-responses and indecisiveness or biases in responses, which need to be overcome to obtain results. It is precisely this methodology of the CIS that is sharply criticized as a "cooking" (Castro, 2018), that is far from actual election outcome and may influence public opinion. As a result of these controversies, the institution had to review their methods to account for the new diversity of the political sphere in Spain and explain publicly their models for estimating voting intentions (Llaneras and Domínguez, 2018). A detailed document on the model (CIS V108) used by the CIS for the 2019 elections was published (CIS, 2019). It explains which variables are used, how the filters, weighting and imputations are realized. It also recall that the use of such estimations is not to predict the vote but is only a photography of the current state of the opinion. This model is still the one used on more recent barometers.

The main difficulty of such estimations remain in the allocation of non-responses or indecisive electors, that is the imputation of the missing information. The steps leading to final estimations of the CIS mainly fall under the "classical" approach to imputations. In this study, we will explore the use of neural networks for estimating voting intentions and the possible increase in performance they can bring as compared to "traditional" methods. More precisely, we will

study multilayer perceptrons (MLP), a type of feed-forward artificial neural networks (ANN) for this task. We will focus on the last available barometer at the time we started the analysis, February 2021.

The structure of our work will be as follow. First, we will proceed with a descriptive analysis of our dataset, in order to understand more clearly the variables of interest for our problem and the characteristics of the sample. This preliminary analysis will motivate the need to impute missing values. Given this necessity, we will draft in a second part the theoretical background behind missing values imputations, and introduce the use of neural networks in that matter. Finally, we will carry out our case study, in order to estimate Spanish voting intentions of the February 2021 barometer and evaluate the performance of our methodology.

2 | Barometer of the CIS

In this chapter, we will carry out the preliminary exploratory analysis of our dataset. We chose to focus on the last available barometer data published by the CIS at the moment we started this research, February 2021. This chapter is important to get a first understanding of our data and the challenges arising to estimate voting intentions. First, the selection of variables of interest for our case study will be motivated. Then, we will present a univariate descriptive analysis of those variables to get a better view of the characteristics of our population. Given those independent analyses, we will then explore the relationships between the variables and our target, the voting intention. Finally, we will present the challenges arising from the missing values present on our dataset, that are at the chore of the work we will pursue subsequently.

The February 2021 barometer was conducted between the 3 and the 11 of that month. It targeted 4000 randomly chosen people of 18 years old or more from phone number (fixed and cellular). In the end, 3.869 interviews were conducted by phone, in 1.240 cities and 50 provinces. As stated by the the official technical documentation, individuals were randomly chosen according to ratios of sex and age using stratified sampling. Weights are available for each observation to obtain a representative sample at the national or autonomous community level.

The total dataset therefore contains 3.869 observations of 360 variables covering extensive socio-demographic information and political opinions.

2.1 Variables of interest for estimating voting intentions

The CIS barometers are not uniquely designed towards estimating voting intentions. Therefore, the 360 variables are not all relevant for our study. In the dataset, our main variable of interest is the question labeled *INTENCIONG*: "*Suponiendo que mañana se celebrasen nuevamente elecciones generales, es decir, al Parlamento español, ¿a que partido votaria Ud.? (RESPUESTA ESPONTANEA)*". Before any further analysis, we need to carefully select our variables of interest that could help us impute missing values in *INTENCIONG*.

To do so, we can rely on the existing literature as researchers have been studying explanatory factors of voting preferences for decades. Specifically, political sociology has seen different paradigms arise around this topic. In total, three explanatory models can be distinguished (Mayer and Perrineau, 1996).

The oldest one was developed by Lazarsfeld and his colleagues, by studying voters' behavior across the duration of American presidential campaigns (1940 and 1944) (Lazarsfeld et al., 1948). They concluded that mass media do not influence much voting preferences. Interpersonal interactions are more likely to determine ones's preferences, especially throughout influential individuals. As a result, according to them, group belongingness is the most important explanatory factor of voting behavior. Socio-demographic characteristics are therefore the discriminatory variables able to predict one's voting intentions.

A decade later, researchers of the *Survey Research Center* at the University of Michigan called into question this socio-demographic determinism (Campbell et al., 1960). They argue that voting intentions can be best predicted using the "funnel model", based on partisanship. According to the authors, individuals inherit a party identification from their parents. This partisanship then shapes their understanding of social issues, leading to partisan feelings in favor of this original party. This last model has also been widely criticized due to the blindness

and lack of evolution it supposes.

An alternative is found in the rational theory (Downs, 1957). It argues for a utilitarian paradigm of the voter, able to counterbalance his benefits and costs at a given time in order to determine his vote. This paradigm leaves more room to change in voting intentions between elections and throughout campaigns according to past and current events.

This theoretical background gives us a first understanding of the variables of interest we should consider for subsetting our dataset. A combination of different paradigms can most likely lead to better accuracy. Following Lazarsfeld and al., socio-demographic information about respondents can be chosen; variables such as the age, sex, religion, educational background, professional and social category... Then, following the partisanship model, we should take into account past voting participation and vote, feelings about political parties and leaders. Finally, according to the rational theory, we need to consider respondent's opinion about the nation's and their current socio-economic situation.

Furthermore, the CIS has years of experience in estimating voting intentions given pre-electoral polls. In an existing work, researchers drew back on the processes employed by the institution in the last decade (Mercado et al., 2014). The authors distinguish four types of questions usually included in the CIS barometers: electoral, ideological, evaluation and socio-demographic. Among those questions, they select a list of relevant ones for their influence on voting intentions. Those four types of questions go well with the scheme drawn above: sociodemographic questions are useful for the first paradigm; electoral and ideological questions for the second; and evaluation questions (valuation of current socio-economic situations) for the last paradigm. The surveys have changed since this publication, but remain an interesting basis to verify the relevance of our variable selection.

All in all, we decided as a first step to select the following 49 variables in the

February barometer [1]:

- CCAA: Comunidad autónoma (important for regional parties);
- Basic socio-demographic variables: SEXO, EDAD, P0 (nationality), ESCUELA (previous schooling), ESTUDIOS (aggregated studies level achieved), RELIGION (religiosity), ECIVIL (civil status: single, married, divorced...);
- Economic variables: SITLAB (professionnal status) and CNO11 (occupation);
- CLASESUB: subjective identification of social class (aggregated levels);
- Questions linked to the COVID-19 crisis:
  - P6: who would you like to take charge of the combat against pandemia;
  - P7: How do you judge the behavior of Spanish people in their way of dealing with the measures against COVID-19?
- P12: Valuation of the general economic situation in Spain;
- P13: Valuation of your personal economic situation at the moment;
- REGFREFE: preferred political regime;
- PESPANNA1, PESPANNA2, PESPANNA3: first, second and third problem existing in Spain at the moment;
- PPERSONAL1, PPERSONAL2, PPERSONAL3: first, second and third social problems that personally affect you at the moment;
- LIDERESCORONA_1, LIDERESCORONA_2, LIDERESCORONA_3, LIDERESCORONA_4, LIDERESCORONA_5: scale of valuation of national leaders in link with the current COVID-19 situation (Pedro Sánchez, Pablo Casado, Pablo Iglesias, Santiago Abascal, Inés Arrimadas);

[1] We must underline that this preliminary selection was carried out without analyzing the responses, and more importantly the non-response rates, which could yield to subsequent elimination of some of them.

- CONFIANZAPTE: degree of trust in the current central government president Pedro Sánchez;
- CONFIANZAOPOSIC: degree of trust in the leader of the principal party of opposition (PP) Pablo Casado;
- PREFPTE: personal preference as president of central government;
- INTENCIONGR: voting intention in hypothetical general elections (our target variable);
- INTENCIONGALTERR: alternative voting intention in hypothetical general elections;
- VOTOSIMG: political party that you feel the most sympathy towards in general elections (taken from SIMPATIA and INTENCIONGR);
- AUTOPTCA1, AUTOPTCA2: self-definition of your political ideology (first and second option);
- ESCIDEOL: scale of ideological self-position (1-10 from left to right);
- ESCAIDEOLLIDERES_1, ESCAIDEOLLIDERES_2, ESCAIDEOLLIDERES_3, ESCAIDEOLLIDERES_4, ESCAIDEOLLIDERES_5: scale of ideological position of national political leaders (Pedro Sánchez, Pablo Casado, Pablo Iglesias, Santiago Abascal, Inés Arrimadas);
- VALORALIDERES_1, VALORALIDERES_2, VALORALIDERES_3, VALORALIDERES_4, VALORALIDERES_5: scale of valuation of national political leaders (1-10) (Pedro Sánchez, Pablo Casado, Pablo Iglesias, Santiago Abascal, Inés Arrimadas);
- FIDEVOTO: vote loyalty in elections;
- PARTICIPACIONG: electoral participation in general elections of November 2019;

- RECUERDO: vote memory in general elections of November 2019 for voters;
- CERCANIA: political party that you consider the closest to your ideas;

2.2 Summary of missing values profile

After this first reduction of our dataset, we have in total 193 450 observations of which 3 311 (1.7%) are missing. However, this percentage is greatly underestimated, as most variables include categories that need to be recoded as NAs. This is what we do next, replacing all concerned values (99 or 9 depending on the variables, which corresponds to "no contesta"). We also decide to recode as missing the N.S. ("No sabe"). Although they are not exactly missing values (not knowing the answer to one question can be an interesting indicator), for the purpose of our study we judged better to consider them as such. Similarly, in some questions regarding political leaders (LIDERESCORONA, ESCAIDEOLLIDERES, VALORALIDERES), some people answered they did not know the person. We also recode them as NA.

The variable concerning the memory of vote in last general elections (RECUERDO) has categories of not having the right to vote at that time, with few respondents. We recode them as NA.

The variables related to vote display specific categories such as "Voto nulo", "No votaría"/"No votó". Those cannot be considered missing values, but rather both linked to abstention. We therefore recode them in the same category.

Some variables display a high rate of missing values in the original dataset: it is the case of the INTENCIONGALTERR and SIMPATIA, because they are conditional questions. INTENCIONGALTERR was asked only to people who declare an intention to vote for a particular political party in INTENTIONGR (that is, do not answer they would vote blank, null, would not vote, do not know or not answer). And SIMPATIA was asked on the contrary only to the persons

that did not declare an intention to vote for a particular political party. It is easily possible to impute the missing values for those variables, using the answer of INTENTIONGR. This strategy had already been performed in the dataset by the CIS, under the variable VOTOSIMG, which can be understood as the voting intention if declared or the sympathy towards one political party in the context of general elections. For INTENCIONGALTERR, we would use a similar approach, by taking the answers (blank, null, would not vote, NC or NA) of INTENCIONGR. Using such a strategy of imputing the missing values with the information present in INTENCIONGR however presents a major issue. Our final aim is to estimate the voting intention, that is the NAs of INTENCIONGR. When training a model on the complete set for this variable, we would include VOTOSIMG and the completed INTENCIONGALTERR as predictors. But those variables were created using answers of INTENCIONGR when the question was not asked. Therefore, in the complete set of INTENCIONGR, VOTOSIMG would just be INTENCIONGR. The models would put a major weight on that variable to have the largest accuracy. But in the incomplete set of INTENCIONGR, the variable is just SIMPATIA, and not the real voting intention. INTENCIONGALTERR would probably also be an important feature in the model, as the answers of INTENCIONGR and INTENCIONGALTERR are often close on the ideological spectrum. But on the incomplete set of INTENCIONGR, INTENCIONGALTERR would be missing for most variables. We would have imputed it beforehand as we will follow an iterative approach, but is does not make much sense to predict an alternative voting intention before the primary one. To overcome this issue, we decided to create a new variable that combines the answers of SIMPATIA and INTENCIONGALTERR. It is interesting as it acts as a secondary voting intention for both the respondents that answered INTENCIONGR and those that did not. This variable "EXTRA" takes as value SIMPATIA when it is available. Else it takes the value of INTENCIONGALTERR except in some specific cases. When INTENCIONGALTERR is "Same party" (31 respondents), it means that the respondent

would not vote for any other party in general elections. We therefore use instead the value of INTENCIONGR. When INTENCIONGALTERR is blank vote (116 persons) or abstention (348 persons), the assignment is more complicated as it could refer to loyalty or abstention. We make the hypothesis that answering this as an alternative voting intention is more the sign of loyalty and therefore use the value of INTENCIONGR for our new variable. Therefore, the variable has no category "Blank vote" as it is not a particular party. The sympathy towards no particular party suggest abstention such that we group the INTENCIONGR value of abstention and "None" of SIMPATIA. With this new variable, we drop SIMPATIA and INTENCIONGALTERR from our final subset of predictors.

After these recodings, our dataset now presents 21 887 (11.5%) missing values. Figure 2.1 allows us to see visually which variables display the most missing values. This first diagnostic demonstrates that our main variable of interest (INTENCIONGR) is not the one with the most important proportion of missing values, but only the 8th. Thereafter, we make a brief analysis of the other variables.

Firstly, the variable with most missing values is AUTOPTCA2, with around 66% missing values (2544). We must remember that this variable is a second option that can be chosen by respondents, after answering AUTOPTCA1 ("Autodefinición de su ideología política"). Most likely, it has that many NAs because people only identify with the first option. Nevertheless, the amount of missing values in AUTOPTCA1 is also very high. And if we compare the answers to INTENTIONGR and AUTOPTCA1 for the few respondents answering both, the counts are not really helpful to separate people. It is not a surprise, as it is quite difficult for anyone to identify themselves on a given range of political ideologies. As the imputations for this variable would be hard and quite hazardous, we also decide to drop it from our analysis. Looking at the Goodman and Kruskal's tau [2] between the non missing values of this variable and INTENCIONGR, we obtain a

[2]Goodman and Kruskal's tau is a measure of assymetric association for categorical factors. Assymetric because it is based on the fraction of variability in the categorical variable y that can be explained by the categorical variable x.

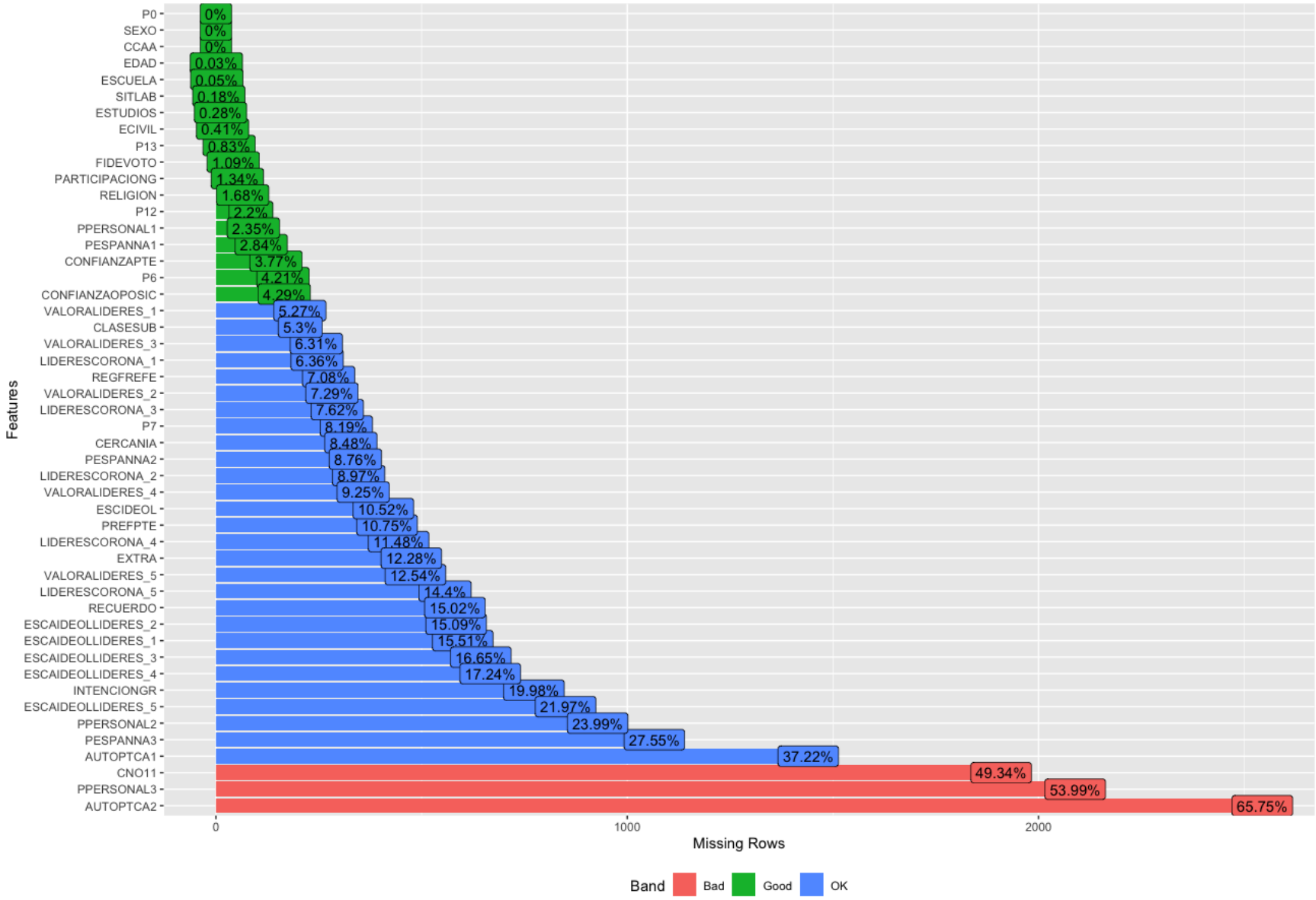

Figure 2.1: Percentage of missing values per variable

small 0.057 for AUTOPTCA2 (taking into account only the observations without ay missing values), and a higher but still very low association with AUTOPTCA1 (0.099). Therefore, for those reasons, we judge reasonable to eliminate it.

Following a similar reasoning, we decide to drop the variable PPERSONAL3 (Problemas sociales que personalmente afectan más: tercero problema). We suppose it has 54% NAs (2089) because respondents cannot think of a third problem. PPERSONAL2 has also a lot of NAs (24%, almost one of four respondents). The Goodman and Kruskal's tau with INTENCIONGR is also small (0.041 for PPERSONAL3, 0.03 for PPERSONAL2, 0.024 for PPERSONAL2). The problem is similar with the variables PESPANNA (Problemas principales que existen actualmente en España). PESPANNA3 has 28% NA. PESPANNA2 and PESPANNA1 have lowest shares of missing values and seemed more important to predict the voting. But when looking at the Goodman and Kruskal's tau with INTENCIONGR the association is very small. It is probably because there is a high number of possible answers but most are linked to the COVID-19 crisis and economy, such that it does not allow to differentiate between electors. We therefore decide to drop those variables for easier imputation processes and because they do not appear as important predictors.

Finally, the last variable with a higher proportion of missing values is CNO11, the professional category of the respondent. It displays 49% missing values (1909). In our selected variables, we have other information related to professional occupation, even if much less precise than this variable (SITLAB). But CNO11 was only asked to people answering SITLAB as current workers. Given the small number of observations in some categories, we judge reasonable to also eliminate CNO11.

In the end, after this preliminary diagnostic of missing values, we decided to drop nine variables of our subset (AUTOPTCA2, AUTOPTCA2, PPERSONAL1, PPERSONAL2, PPERSONAL3, PESPANNA1, PESPANNA2, PESPANNA3, CNO11). We now have 40 variables, and INTENCIONGR is the second one in terms of

missing values. In Table A1.1 (Appendix A1), we present the count and relative frequency of missing values per variable in our subsetted dataset.

On another hand, those analyses of missing values in some variables have underlined an interesting phenomenon: there seems to be some associations between non responses in some variables and our variable of interest INTENCIONGR. Some respondents that did not answer their voting intention may also be more likely to not have answered other specific variables. To further explore this possible "missing value patterns", we need to cross the presence of absence of answers for each respondent among all variables. This can be visualized in Figure 2.2.

In this plot, the horizontal axis corresponds to observations whereas the vertical axis corresponds to variables. Each respondent's answers are therefore represented vertically. In red are displayed missing values whereas in blue are non missing. Finally, variables on the y axis are ordered by rate of missing values: those who contain the most important proportion of NAs are at the top of the graph.

As we can observe, there is less than half respondents that answered all questions. The number of rows with no missing value is 1 595, over 3869, that is only 41.2%. However, as we can notice on Table 2.1, most persons have only between 1 and 4 missing responses. 70.3% of our sample has 2 or less missing values.

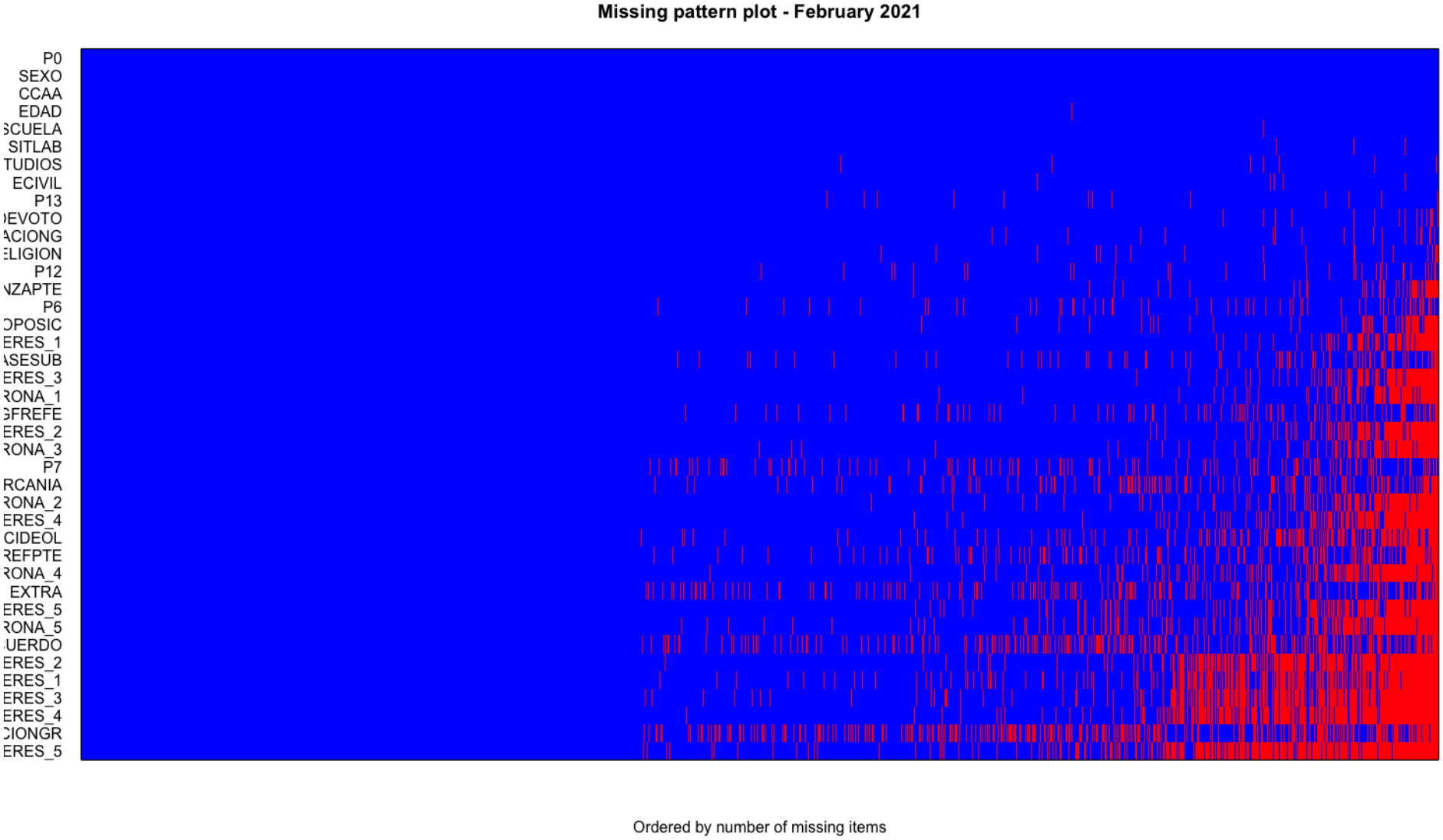

Figure 2.2: Missing values profiles

Number of missing values	Count	Percentage
0	1595	41.2
1	765	19.8
2	359	9.3
3	201	5.2
4	147	3.8
5	111	2.9
6	109	2.8
7	81	2.1
8	78	2.0
9	66	1.7
10	37	1.0
11	39	1.0
12	33	0.9
13	32	0.8
14	27	0.7
15 or more	189	4.9
Total	3869	100.0

Table 2.1: Repartition of counts of missing values per respondent

On the contrary, basic socio-demographic variables have very few missing values, such that we should be able to have this information for most of the sample. Some respondents have nevertheless a very high share of NAs (right part of the plot),for the majority of variables. We can expect that those observations would be the hardest one to impute and may bias our final voting intention estimations.

If we look closely at the missing values profiles of respondents that did not answer INTENCIONGR, we can remark that most of them have not either answered the three variables with a highest NA rate. Another variable that has a high proportion of missing values and that we judge important to comment is RECUERDO, the recall of vote in the last general elections. It presents 15% NAs (581 respondents)[3]. Theoretically, it is one of the most important variables to predict INTENCIONGR. What is interesting to note is that most of the NAs in INTENCIONGR are also NAs in RECUERDO, which is probably due to people not wanting to answer the question.

Other individuals display a more seldom profile, not answering our target variable paired with other NAs in variables where it is not frequent to have miss-

[3]We must keep in mind that are also included in NAs people that did not have the right to vote at the time (45 respondents or 8% of them).

ing values. For them, it may be easier to estimate their voting intentions by using the answers of respondents with similar characteristics in the complete variables.

Overall, with this first analysis of the missing values in our dataset, we could understand that the obstacles do not only reside in non-responses for the voting intention. In order to obtain the best estimations for this variable, we intend to use the answers of respondents to the selected variables of interest. Therefore, we must also design strategies to fill missing values of those variables to reach better accuracy.

Before diving deeper into the existing processes that can be used to impute our missing values, we must beforehand analyze more thoroughly our dataset. In the following section, we will draft a brief univariate analysis of our variables. Then, we will explore the relationships between our candidate predictors and our target variable. This will allow us to better understand which explanatory variables should be the most valued.

2.3 Univariate descriptive analysis

This section is designed to outline a general profile of our sample characteristics. Using simple frequency tables and barplots, we can get a quick understanding of our observations. This preliminary analysis is also useful for further recoding purposes by grouping some categories. As before, those analyses are presented on the original sample, that is without applying weights. Unless specified, results are presented without taking into account NAs.

2.3.1 Socio-demographic variables

The socio-demographic variables are the ones which display the lowest rate of missing values. In our sample, we have a balanced proportion of males and females (Figure 2.3); respectively 48.9% (1891) and 51.1% (1978). Similarly, the age repartition of the sample is quite balanced thanks to the sampling method

behind the Barometer (Figure 2.4.

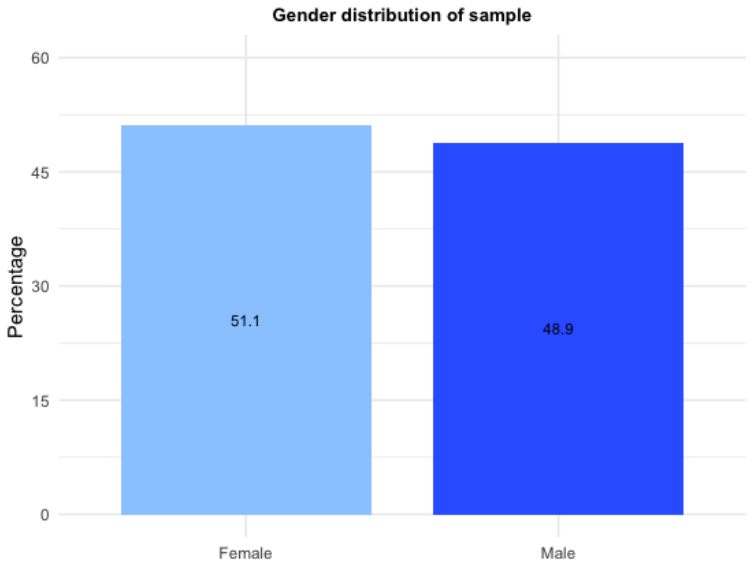

Figure 2.3: Gender distribution of sample

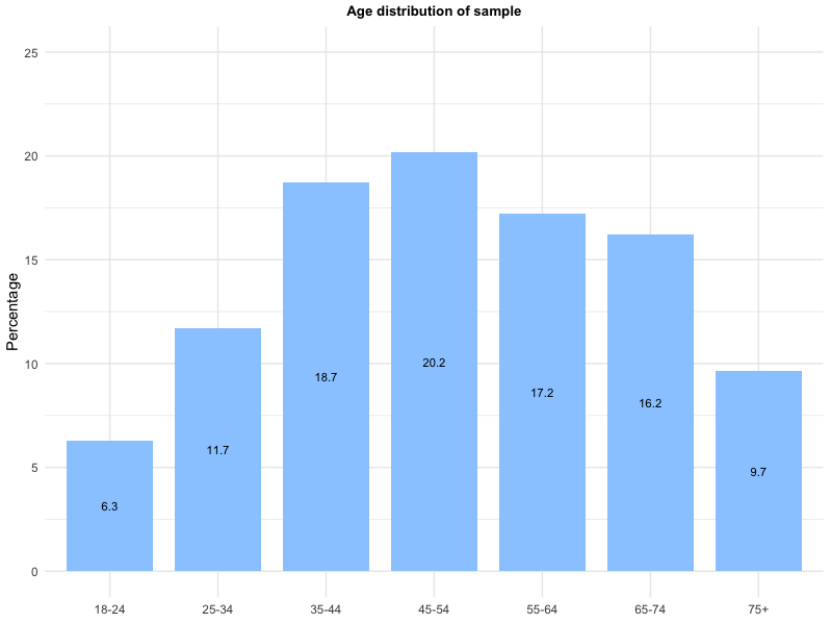

Figure 2.4: Age distribution of sample

Then, all respondents have the Spanish nationality (2.7% also have another one). More than half of them are married (Table 2.2).

Concerning education, almost all sample has been to school in the past (98.2%). Regarding the level of studies (Figure 2.5), it is interesting to notice that approximatively 40% have attended higher education.

Nationality	Count	Percentage
Spanish	3766	97.3
Spanish and another	103	2.7
Total	3869	100.0

(a) Nationality

Civil Status	Count	Percentage
Married	2155	55.9
Single	1120	29.1
Widow	285	7.4
Separated	65	1.7
Divorced	228	5.9
Total	3853	100.0

(b) Civil Status

Table 2.2: Nationality and Civil Status frequency distribution of sample

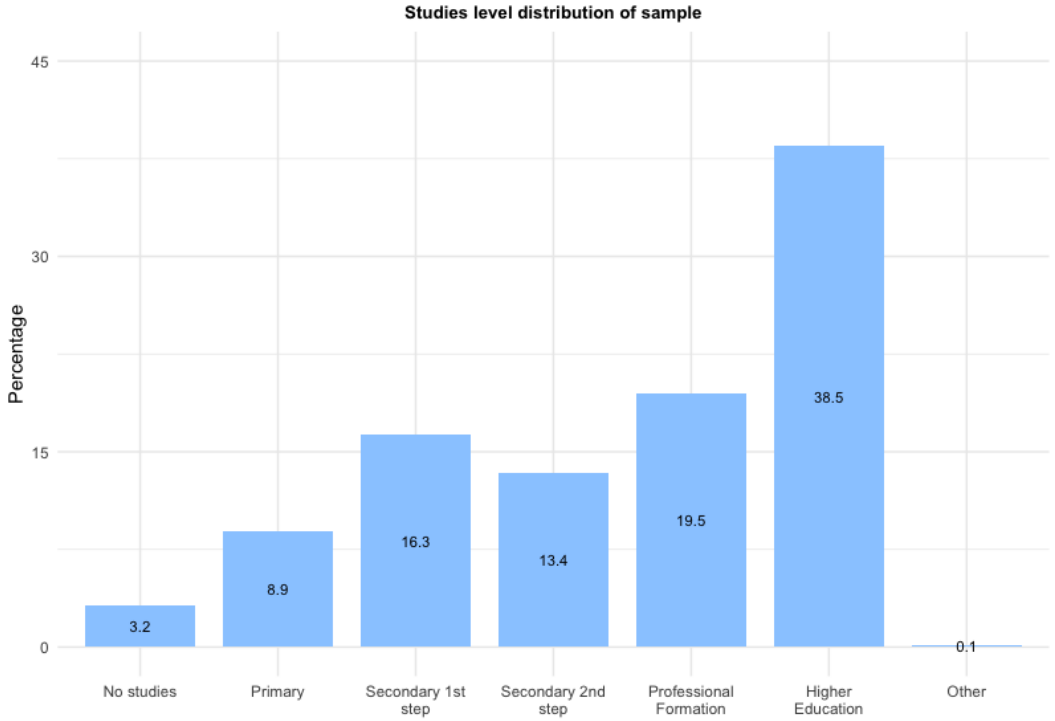

Figure 2.5: Level of studies distribution of sample

65 persons did not answer the question regarding religion. As for the others, the most frequent situation is Catholic believer, even is the majority is not practicing (Figure 2.6).

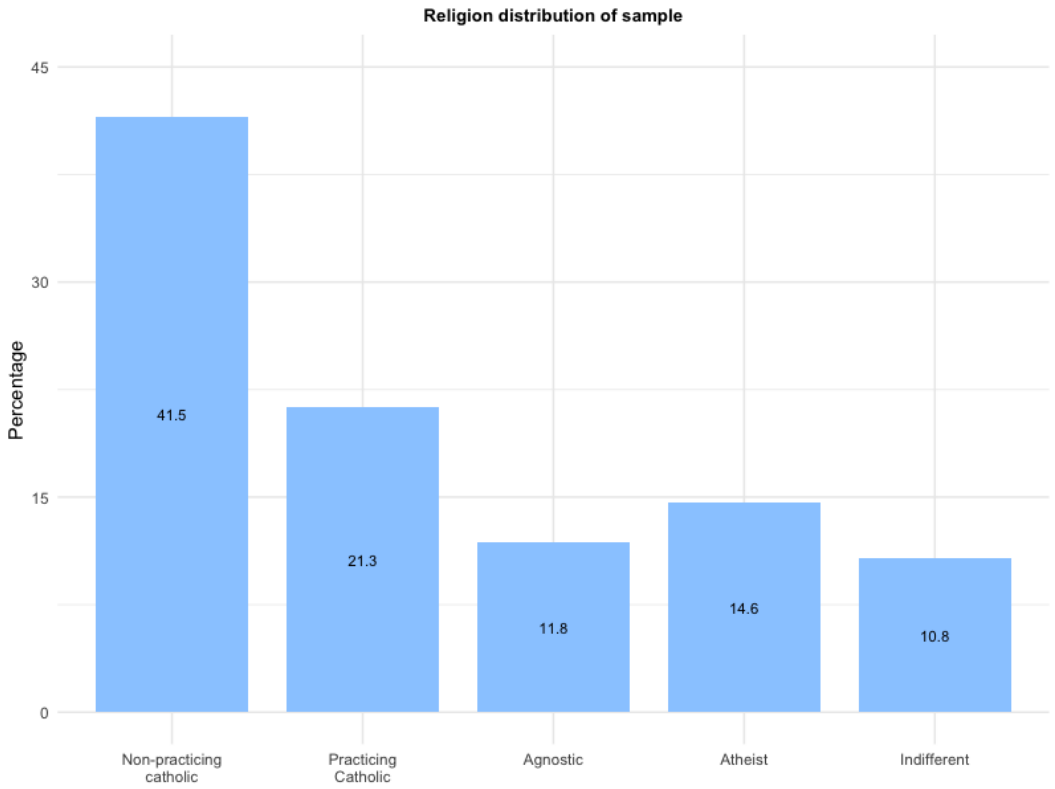

Figure 2.6: Religion distribution of sample

Then, the professional situation variable (SITLAB) has eight levels (after our initial recoding), some of them referring to similar situations (retired and has worked before/retired and has not worked before; unemployed and has worked before/unemployed and has not worked before). We decide to group those similar levels, reducing the number of levels to six (and NA). As we can see on Figure 2.7, more than half of the sample is working, around 27% are retired and the rest are either unemployed, students, homemaker or in another situation. The professional situation is NA for only seven persons.

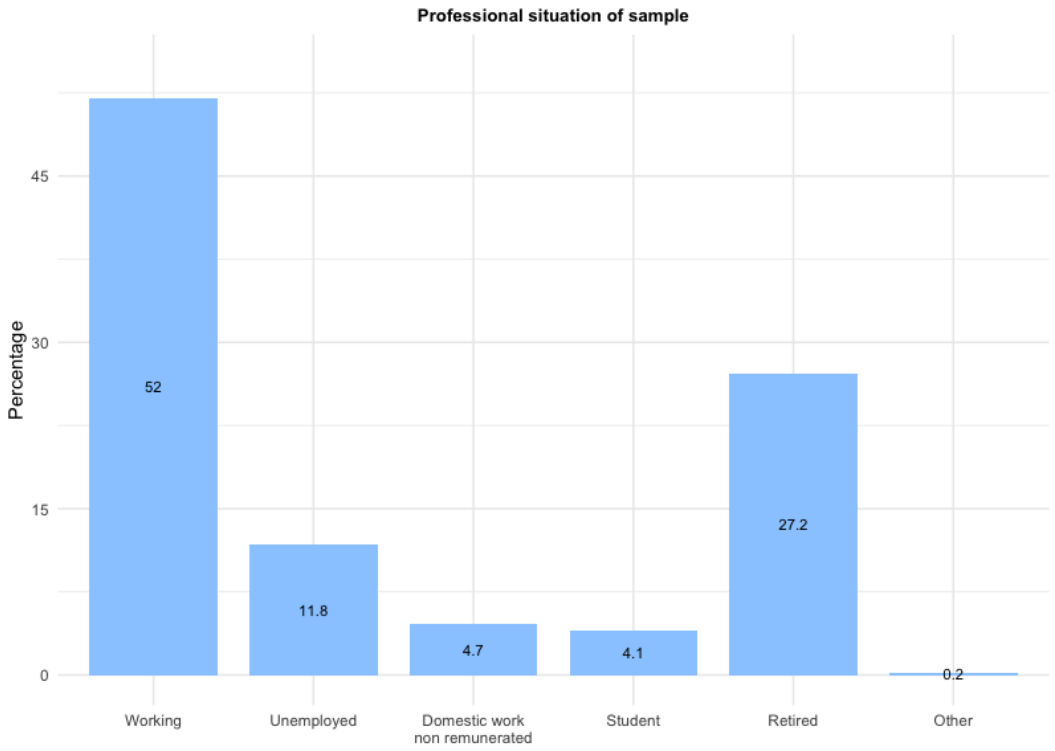

Figure 2.7: Professional situation distribution of sample

Finally, regarding the subjective social class identification, there are 205 NA (5% of sample). Without considering the missing values, more than half individuals consider themselves as middle class and 17% as lower middle class. Only 6% as upper or upper-middle class and the rest consider themselves as working, lower-class or other (Figure 2.8).

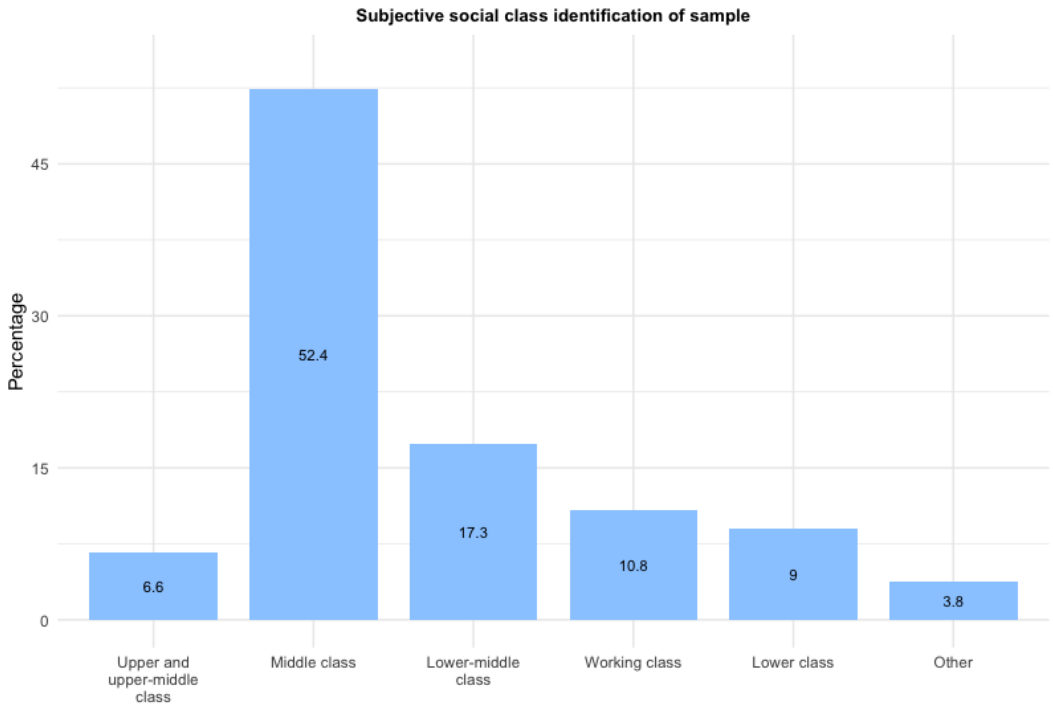

Figure 2.8: Distribution of the subjective social class identification of sample

2.3.2 Questions related to the current situation

Afterward, a group of questions are related to the current situation in Spain. Interviewees were asked who they wanted to take charge of the fight against the pandemic (P6). This variable has 163 missing values. For those who answered, the majority (68%) would prefer the Spanish government and the autonomous communities to work together (Table 2.3). Respondents are more divided concerning the behavior of Spanish people in regard to the measures adopted (317 NA).

	Count	Percentage
Spanish government	709	19.1
Autonomous communities	245	6.6
Both together	2512	67.8
Doctors, scientifics and independent experts	117	3.2
Organizations or Institutions (WHO, EU)	23	0.6
Other	77	2.1
None of them	23	0.6
Total	3706	100.0

(a) P6

	Count	Percentage
The majority is reacting with civicism and solidarity	2181	61.4
The majority is reacting with little civicism and indiscipline	1371	38.6
Total	3552	100.0

(b) P7

Table 2.3: Opinion regarding COVID-19 situation

Interviewees also gave their opinion on the current economic situation, both for Spain (P12) and for their personal situation (P13). This could affect their voting choice depending on their valuation of current difficulties. There are few missing values for these variables (85 and 32 respectively). As we can see on Figure 2.9, respondents have in general a pessimistic opinion on the general economic situation in Spain as compared to their personal situation. More than half of the sample judge their personal economic situation as good whereas for the country, 51% judge it very bad, which is probably even accentuated by the crisis.

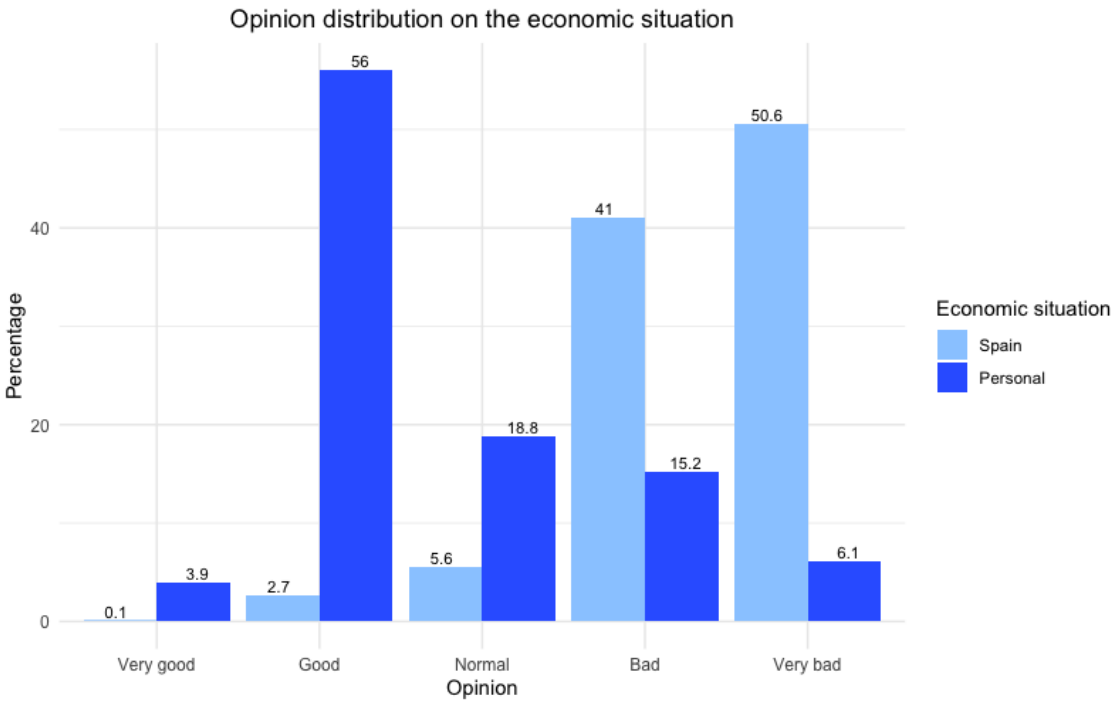

Figure 2.9: Valuation of the country and personal economic situation

Finally, some questions related to the current situations are more related to political parties and politicians. It is the case for the LIDERESCORONA five variables. They question interviewees on their valuation of the main political leaders regarding the COVID-19 situation. They are interesting variables as they are indicators on the opinion of the respondents on the possible candidates for general elections. If they value positively their actions, we can expect they would be more likely to vote for them. After looking closely at those variables, we noticed there was a category (97) that indicates that respondents do not know the politician. We decide to recode those values as NA. For Pedro Sánchez, Pablo Casado and Pablo Iglesias there are few people (3, 35 and 15 respectively) in this case, but for Santiago Abascal and Inés Arrimadas it is more frequent (91 and 179 respondents). Overall, the percentage of NAs values for all of them remain under 15%.

As we can note on the boxplots below, the current president of the government (Pedro Sánchez) is the most positively valued. On the contrary, Santiago Abascal (from the extreme right party VOX) is the less positively valued, more than half of respondents assigning him the lowest "grade". Overall for all leaders

except Pedro Sánchez, only around 25% of respondents give a valuation superior to 5 regarding their behavior in the health crisis.

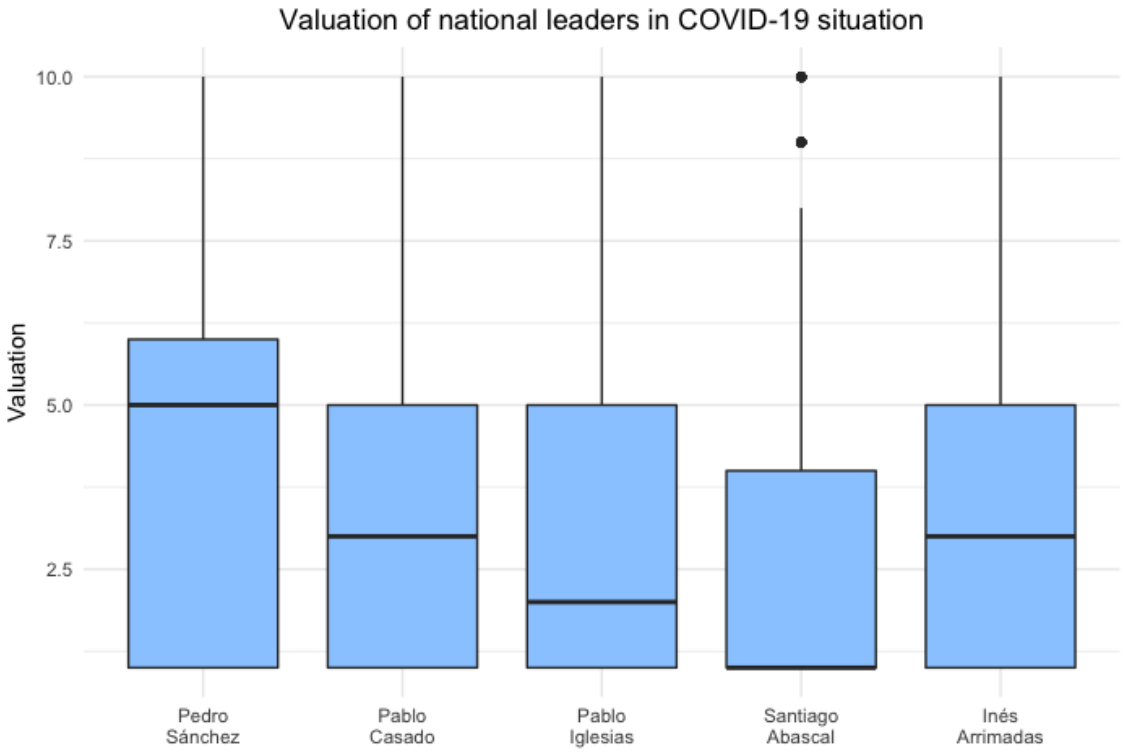

Figure 2.10: Valuation of the national leaders regarding COVID-19

For the two leaders of the two principal political parties (PSOE and PP), interviewees were also asked about their level of trust. This information gives an additional insight on how they value their actions in the current situation. 146 and 166 persons did not answer those questions for Pedro Sánchez and Pablo Casado respectively. As we can see on Figure 2.11, the results are much more negative than when interviewees were asked specifically on the COVID-19 situation. Indeed, more than 70% of respondents have low or very low trust in Pedro Sánchez. Pablo Casado is still less trusted that the president of the government.

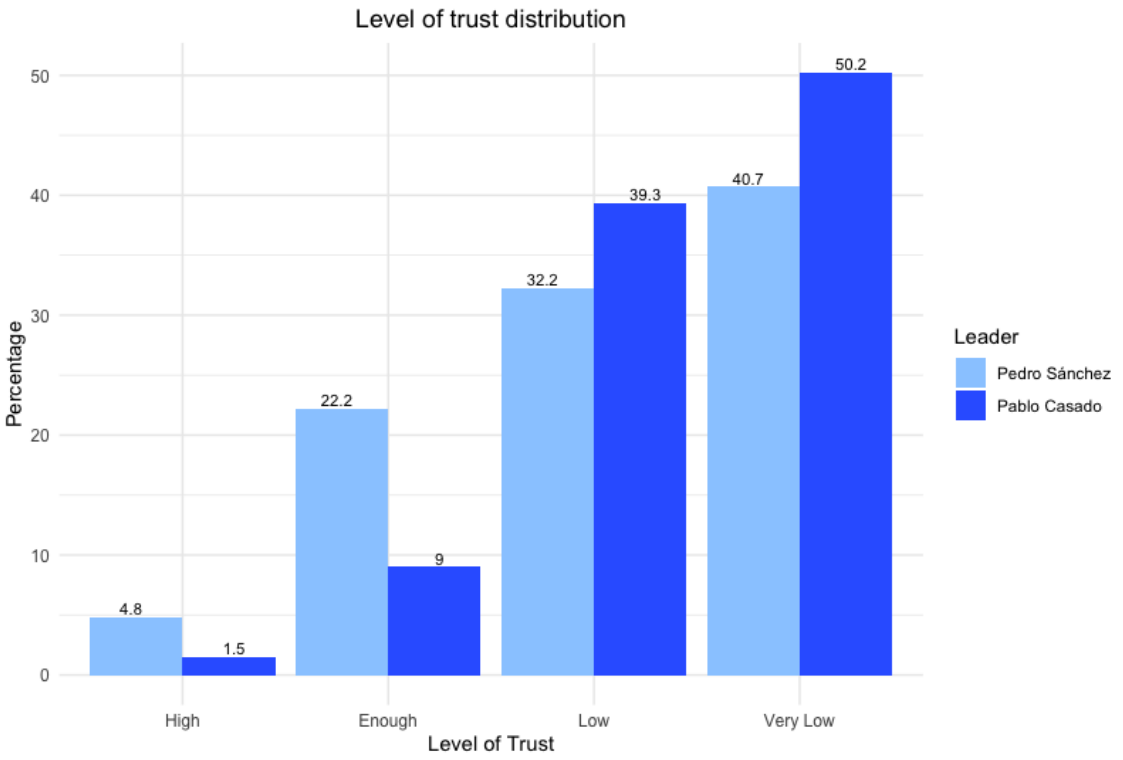

Figure 2.11: Level of trust in the president of government and leader of opposition

2.3.3 Political preferences

Then, a series of questions are precisely linked to preferences of the respondents in political matters.

When asked about their preferred political regime, 274 persons did not answer or did not know (NAs). Among the others, as we can see on Table 2.4, democracy is in great majority preferred. The other answers may be interesting when imputing the voting intention, as for instance someone for which democracy or authoritarian does not matter may be more likely of abstention.

	Count	**Percentage**
Democracy always	3045	84.7
Authoritarian can be preferable	244	6.8
Not matter	306	8.5
Total	3595	100.0

Table 2.4: Distribution of the preferred political regime

More observations are missing concerning the PREFPTE variable (personal

preference as president of government), an important indicator of voting intentions. This variable has nine categories, including seven named political leaders, "Other" and "None of them". The first seven categories are interesting when linked to the party to which the leaders belong. However, as we can notice on Table 2.5, the "None of them" category is quite important, accounting for more than 27% of responses.

	Count	**Percentage**
Pedro Sánchez	1013	29.3
Pablo Casado	351	10.2
Santiago Abascal	238	6.9
Pablo Iglesias	176	5.1
Alberto Garzón	73	2.1
Inés Arrimadas	323	9.4
Iñigo Errejón	111	3.2
Other	203	5.9
None of them	965	27.9
Total	3453	100.0

Table 2.5: Distribution of the preferred person as president of government

A variable that has the advantage of having few missing values (407) is the one asking to self-position on a Left (1) to Right (10) ideological scale. We dropped AUTOPTCA1 because the high rate of NA was linked to the difficulty people encounter to define their ideology, but this self-positioning on scale seems easier. On Figure 2.12, we can see a tendency to position on the center (towards left) for the majority of the sample.

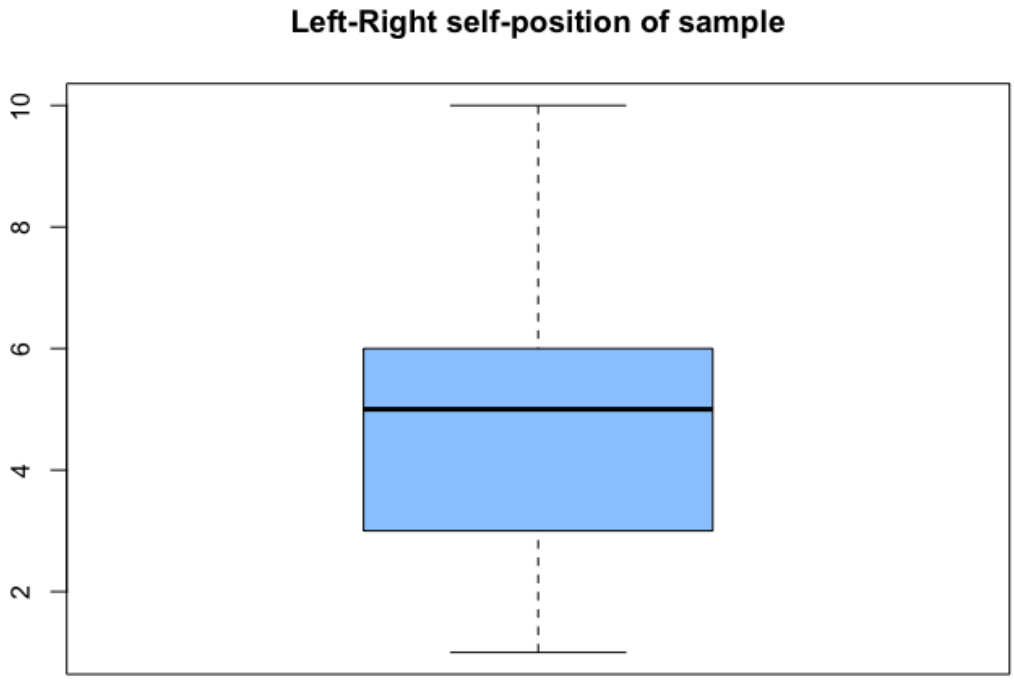

Figure 2.12: Distribution of the self-position on Left-Right scale

This information is particularly interesting to cross with how the respondents position the main political leaders on the same scale (Figure 2.13). For creating this graph, we had to remove the missing values, which are quite important, all the more after adding as missing the category "Does not know him/her" (more than 600 NAs for each leader).

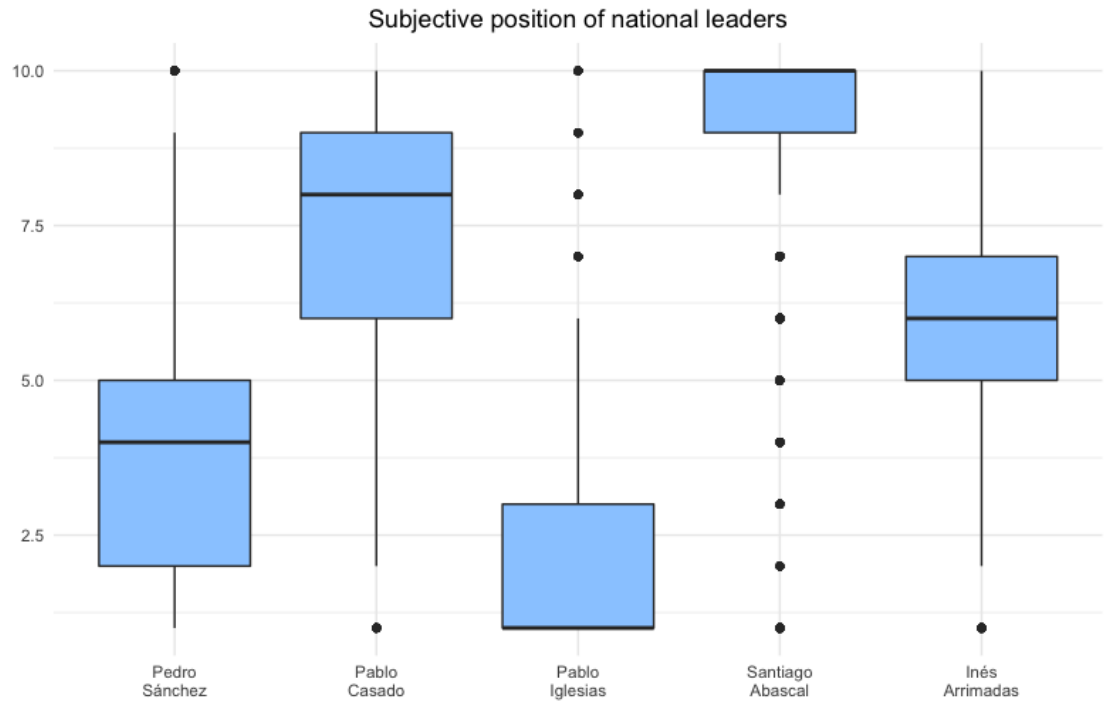

Figure 2.13: Distribution of the subjective position of main leaders on Left-Right scale

On the same scheme of a 1-10 scale, respondents were asked to position each leader according to how they value them. Again, we recoded as NAs the "Does not know him/her", which leads to around 200-300 NAs for each leader. We could expect that a more positive valuation is likely to increase the chances to see the interviewee vote for a particular political party. As compared to Figure 2.10, Figure 2.14 shows that the positive valuation given to Pedro Sánchez is overall less important when we are not talking about the COVID-19 situation.

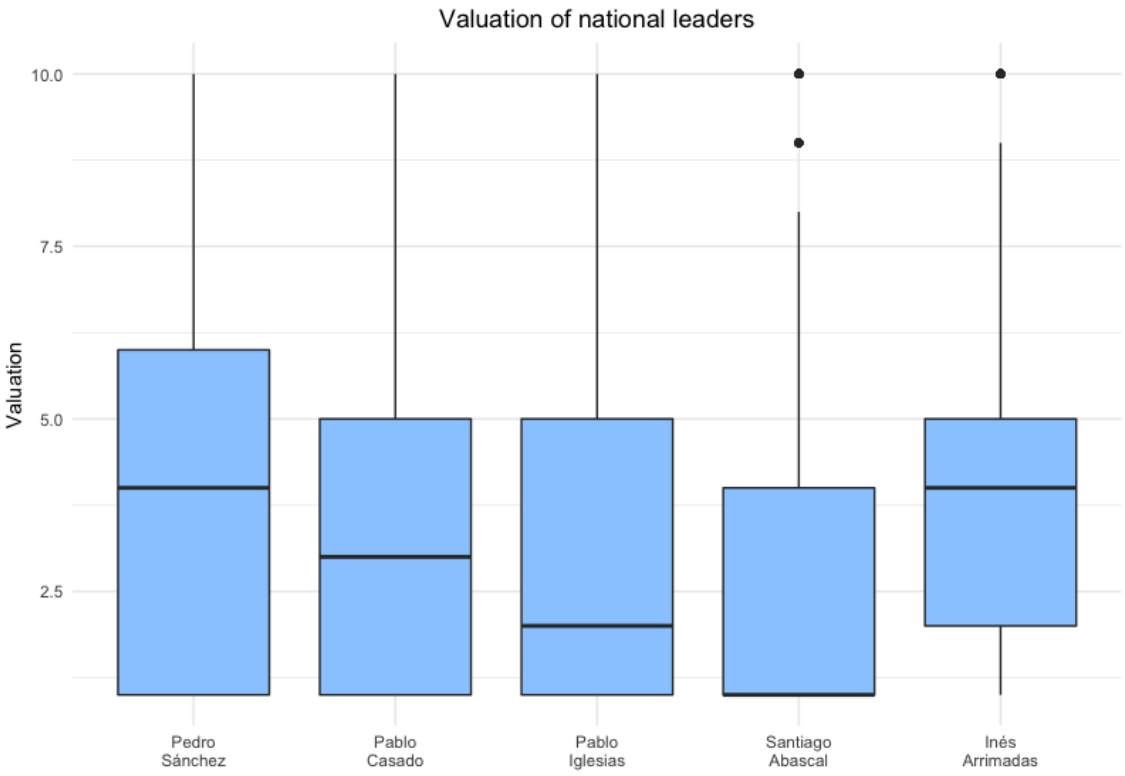

Figure 2.14: Valuation of the main leaders on a 1-10 scale

Finally, the CERCANIA variable (political party that you consider the closest to your ideas) has only 328 missing values. This variable, has a large range of possible categories that had not been grouped like in other variables related to voting intentions (INTENCIONGR, INTENCIONGALTERR, RECUERDO). We decided to group under "Others" the categories not appearing in the latter ones. Moreover, following the results published by the CIS for their voting intentions estimations of February 2021 (CIS, 2021), we grouped together Compromís and Más País under Més Compromís (electoral coalition). The same by grouping Unidas Podemos,Podemos, EQUO, IU and En Comú Podem, and En Común-Unidas Podemos under Unidas Podemos. The frequency table for this recoded variable is available in Appendix A1 (Table A1.2). There we can remark that parties specific to some Spanish communities hold a very low percentage. It would be difficult to estimate the voting intentions of the missing values for such variable. It is the same with all variables related to political parties. To tackle this problem, we decided to group such parties under two broad categories: left re-

gionalist/nationalist party (BNG, CUP, EH Bildu, PRC, Teruel Existe, Més Compromís) and right regionalist/nationalist party (PNV, CCa-NC, JxCat, NA+). ERC is a left nationalist party of Cataluña but as it has a higher number of respondents in most variables, it was kept as a category. Those recodings will also be applied to the other variables regarding political parties. The strategy for imputations of those parties will be explained in the following chapter. Table A1.3 in Appendix A1 presents the frequency table for this second level of recodings.

2.3.4 Variables related to elections

To finish, the last variables of interest are the ones linked directly to elections.

First, our "EXTRA" variable, combining SIMPATIA and INTENCIONGALTERR, is very linked to INTENCIONGR as it is a combination of the answers to SIMPATIA (for NAs in INTENCIONGR), INTENCIONGALTERR (for the 'complete' answers to INTENCIONGR) and INTENCIONGR itself in some cases. We display in Appendix A1 the frequency tables for this variable, with two levels of recoding (Table A1.4 and Table A1.5).

Then, we have the direct variables linked to the voting: RECUERDO on the last elections and the voting intention for future general elections (INTENCIONGR).

As we can notice on Table 2.6 and Table 2.7, the distribution is quite similar to that observed in the CERCANIA variable if we adjust without taking into account the "None" category. However, we can note some discrepancies, for instance for the extreme-right party VOX. 6.7% of the respondents declared it as the closest to their ideas, when 9.2% would vote for it in general elections, maybe because they judge smaller parties closer to their ideas have less chances to win and represent them.

	Count	Percentage
PP	442	14.3
PSOE	836	27.0
Ciudadanos	249	8.0
Més Compromís	38	1.2
ERC	80	2.6
JxCat	22	0.7
EAJ-PNV	31	1.0
EH Bildu	15	0.5
CCa-Nc	5	0.2
NA+	16	0.5
PACMA	33	1.1
VOX	285	9.2
CUP	14	0.5
Unidas Podemos	296	9.6
BNG	21	0.7
PRC	4	0.1
Teruel Existe	3	0.1
Other	27	0.9
Blank vote	160	5.2
Abstention	519	16.8
Total	3096	100.0

Table 2.6: Voting intentions distribution

	Count	Percentage
PP	442	14.3
PSOE	836	27.0
Ciudadanos	249	8.0
ERC	80	2.6
PACMA	33	1.1
VOX	285	9.2
Unidas Podemos	296	9.6
Left regionalist/nationalist	95	3.1
Right regionalist/nationalist	74	2.4
Other	27	0.9
Blank vote	160	5.2
Abstention	519	16.8
Total	3096	100.0

Table 2.7: Voting intentions distribution, second level of recoding

Similarly, we proceeded to the same recodings for the variable of the memory of vote in the last elections (2019), which frequency table is presented on Table A1.6 and Table A1.7. As compared to Table 2.6, as for the non-missing values,

we can see an evolution with less votes estimated for PP, Unidas Podemos and Ciudadanos; whereas respondents declare they would vote more for the VOX or PSOE. We can also notice there were only around 2% of blank votes and 13.7% of abstention in 2019 but, for respondents, nowadays 5.2% declare the intention to vote blank and 16.8% to not vote. But it is not possible to know if they would actually do so the day of general elections.

Related to this comparison, we can also take a look at the FIDEVOTO variable, which questions interviewees about their "fidelity" in voting between elections. There are only 42 missing values for this variable. In Figure 2.15, we can see that more than half of respondents declare to change of vote depending on their current opinions. However, 20% declare always voting for the same party, which could be useful for imputing voting intentions if the RECUERDO variable is non-missing.

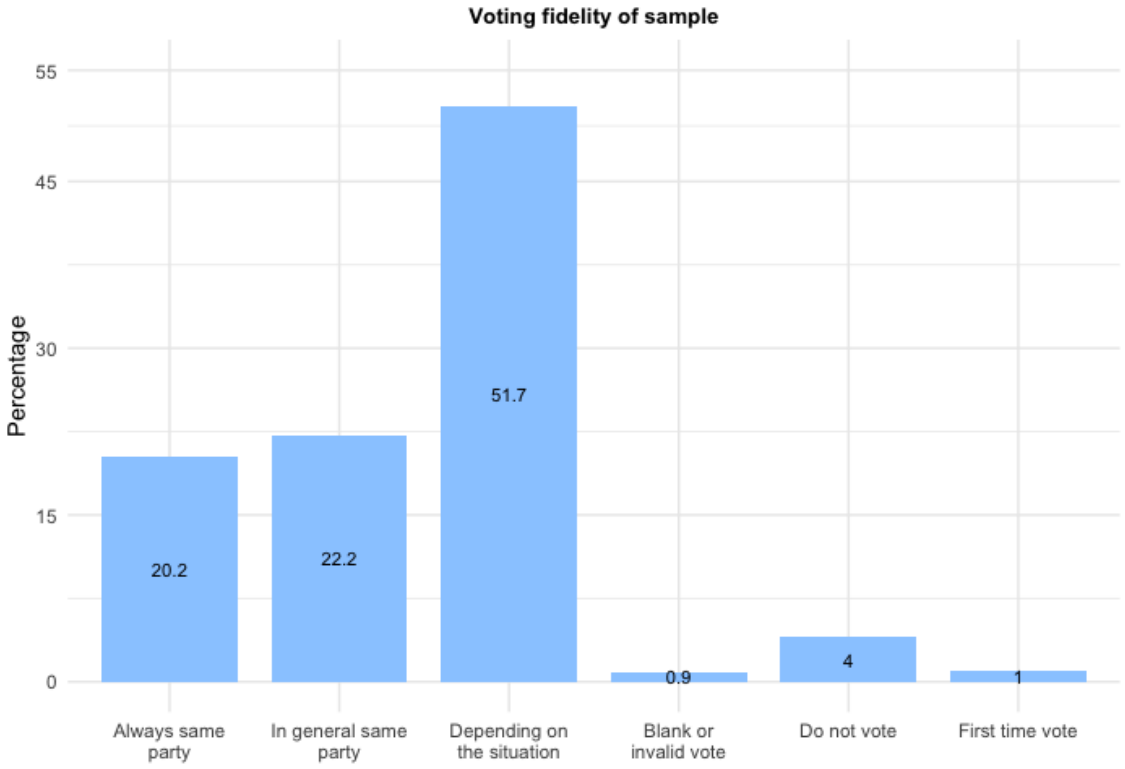

Figure 2.15: Distribution of fidelity of vote between elections

To finish, the last variable of interest we selected is PARTICIPACIONG, that indicates if the person participated in the November 2019 general elections. This

variable is interesting as it is an indicator of the investment of respondents in the electoral domain, even if their behavior can change throughout time. 52 persons did not answer this question. 8 declare they went but could not vote, 33 were underage, 127 could vote and 11 did not have the right. We decide to group them as "Could not vote". We also grouped together the category "Went to vote and voted" and "Voted by mail" together under "Voted". As we can read on Table 2.8, most interviewees (87.9%) did vote in the last general elections.

	Count	**Percentage**
Voted	3356	87.9
Could not vote	171	4.5
Did not want to vote	290	7.6
Total	3817	100.0

Table 2.8: Distribution of the participation in the last general elections

To conclude, this preliminary univariate descriptive analysis of our variables of interest has permitted us to get a first view of the characteristics of the sample. We could also proceed to different recodings that will be necessary for further analyses. Thereafter, we proceed to a brief bivariate analysis of those variables, in relation with our target variable. It will help us understand which candidate variables would be the best predictors to impute the voting intention.

2.4 Bivariate descriptive analysis

There are different ways to measure the relationships between independent features (predictors) and a dependent feature (target or response). Simply computing contingency tables or representing them visually is not enough, as it does not indicate the statistical significance of the relationships we may observe.

A difficulty for such analysis on our dataset is that some some categories very few observations (for instance 5 or less for CCa-NC, PRC and Teruel Existe). When computing contingency tables, the observed and expected number of observations in each case can be very small and lead to an overestimation of the χ^2

statistic for example. For that reason and for easier visualization, we decided to momentarily group under the "Other" category the political parties with less than 100 observations for our testing in order not to bias the results. On Table 2.9 is the frequency table of this recoded INTENCIONGR variable; as we can see, we now have only 5 named political parties, the blank vote and abstention, other answers being categorized as "Other" (the previously recoded NAs are not included in the following analyses).

	Count	**Percentage**
PSOE	836	27.0
PP	442	14.3
Ciudadanos	249	8.0
Unidas Podemos	296	9.6
VOX	285	9.2
Blank	160	5.2
Other	309	10.0
Abstention	519	16.8
Total	3096	100.0

Table 2.9: Distribution of the recoded voting intentions

We will also need to group together some categories for our independent variables where there are very low absolute frequencies. Finally, we do not perform those tests exhaustively for each candidate predictor, but only those for which analyzing the relationship appears relevant and the results significant.

We will now present the methodology behind the statistical tests we will perform before commenting our results.

2.4.1 Methodology

Our response variable is categorical. Our predictors are of two types: most are categorical, some are numeric.

To analyze a relationship between two categorical variables, a common procedure is to perform a chi-square test. For the quantitative variables we have, we decide to recode them into groups to be able to perform also chi-square tests. This test allows us to examine if two variables X and Y are dependent, that is test the

hypotheses:

$$H0 : F(x,y) = F(x)F(y) \qquad \text{(variables are independent)}$$
$$H1 : F(x,y) \neq F(x)F(y) \qquad \text{(variables are dependent)}$$

If X and Y were independent (H_0 true) then, the expected value for the number of observations in $A_i \cdot B_j$ (category i of variable X and category j of variable Y) would be:

$$E_{ij} = \frac{n_i \cdot n_j}{N}$$

where n_i is the number of observations of A_i, n_j the number of observations of B_j and N the total number of observations. We also denote O_{ij} the number of observations in the joint class $A_i \cdot B_j$ (obtained from the contingency table, that is absolute joint frequency table). Then, we may use the chi-square statistic:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

with $(k-1) \cdot (r-1)$ degrees of freedom (k is the total number of categories of X, and r the total for Y).

A small Chi-square value would mean that the observed count is close to the expected count, and that the two features are independent. Therefore, we are more looking for high values that indicate that H_0 is incorrect and that the independent variable should be interesting for our models. We choose a significant level $\alpha = 0.05$. After computing our chi-square values, we will verify if they fall in the error region ($p - value < 0.05$).

2.4.2 Qualitative variables

To begin, we can distinguish some differences according to the sex, as Figure 2.16 demonstrates. The chi-square test points out to these apparent differences (X-squared = 58.308, df = 7, p-value = 3.283e-10).

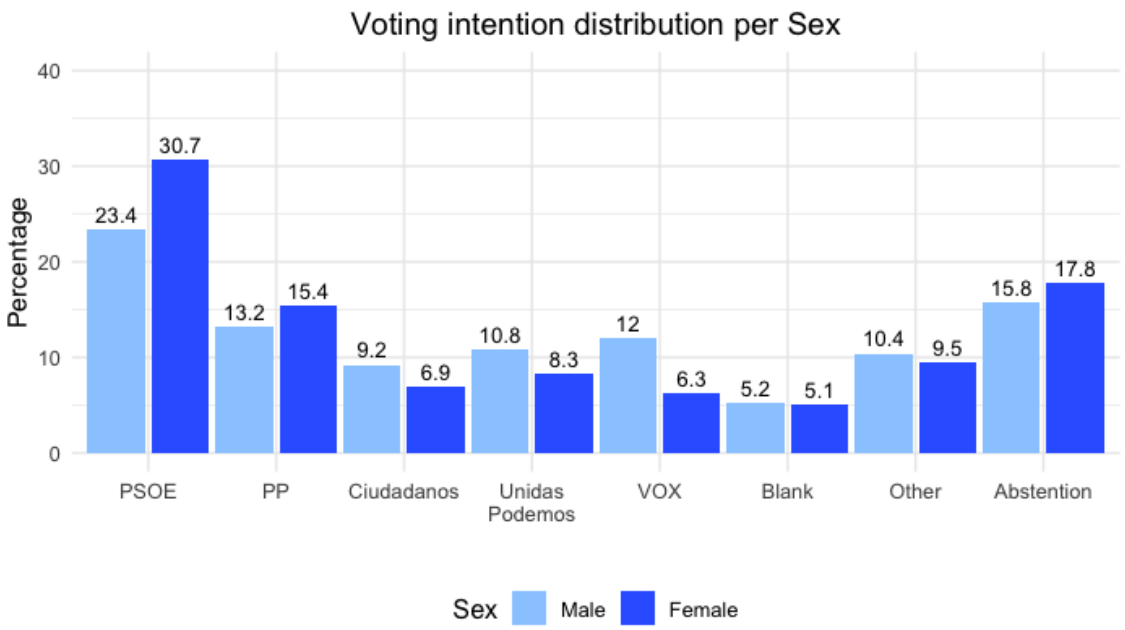

Figure 2.16: Distribution of the voting intentions per Sex

Then, we looked at the relationships between the level of studies and the declared voting intention. For that purpose, we grouped together some categories (No studies and Primary; Secondary 1st step, Secondary 2nd Step and Professional Studies) and recoded as NA the two observations "Other". Figure 2.17 clearly shows the existence of important differences between the level of studies in terms of voting intentions, as well as the chi-square test (X-squared = 92.453, df = 14, p-value = 1.303e-13). We can note that persons who did not go to school or only followed primary studies tend to vote more for the "leader" parties PP and PSOE. Higher educated respondents are more to vote for Ciudadanos or for other parties (denoting maybe a further interest into specific political spheres). Secondary or professional educated persons are more than higher educated ones to vote for the PSOE but also the most numerous in term of voting intentions for VOX. Abstention is less common for higher educated respondents.

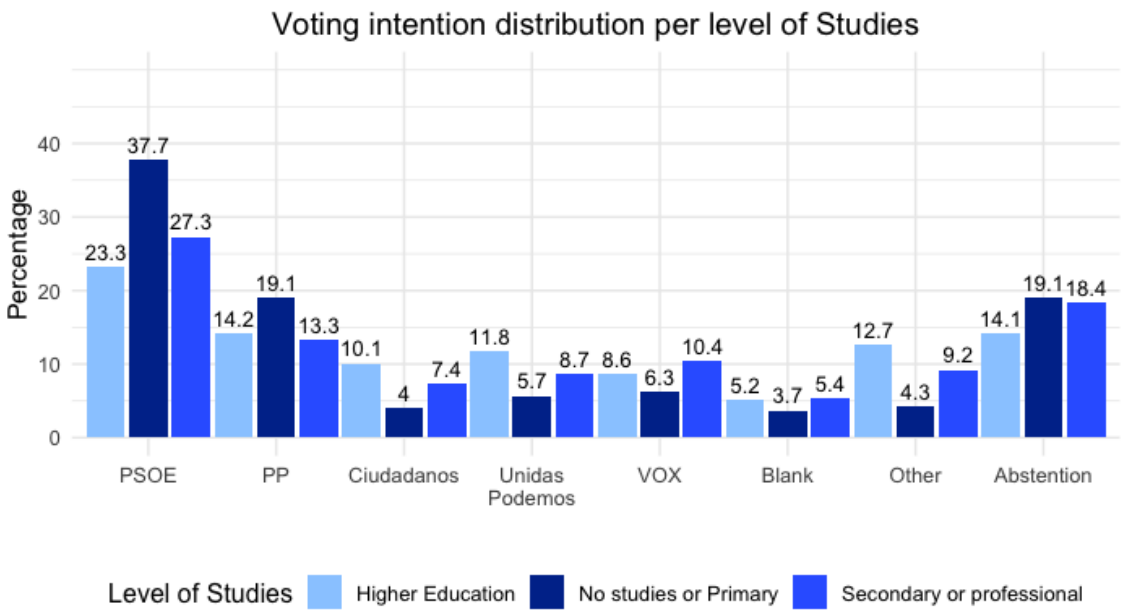

Figure 2.17: Distribution of the voting intentions per level of Studies

As for the professional situation, we again found significant differences between groups in terms of voting intentions. We recoded as NA the 9 "Other" professional situations. The chi-square test also confirms our visual analysis (X-squared = 140.23, df = 28, p-value < 2.2e-16). We can notice how retirees and non-remunerated domestic workers are numerous to vote for the two main political parties in Spain. Persons in the working market (unemployed or working) have more similar voting intention proportions, and specifically appear to be more to intend voting for VOX. Finally, students appear to be the less homogeneous voters, with higher proportions in almost all categories, including blank voting intentions.

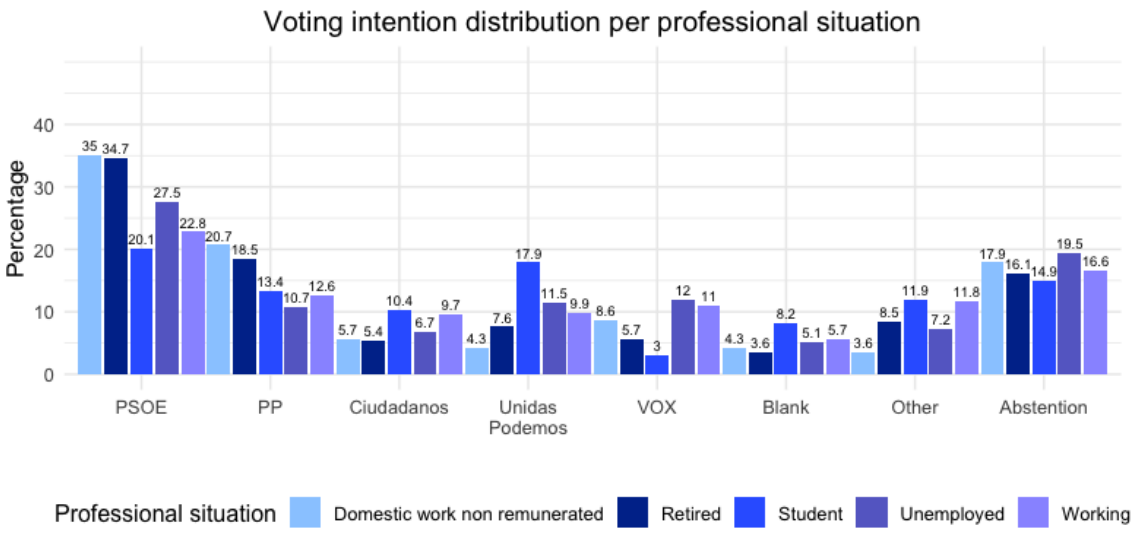

Figure 2.18: Distribution of the voting intentions per professional situation

To continue, we also looked at the link between the memory of vote to last general elections (RECUERDO) and the declared voting intention. We therefore recoded RECUERDO as we did for INTENCIONGR. We moreover recoded as "Other" the blank vote as it was too little represented in the RECUERDO variable and may bias the chi-square test results. Unsurprisingly, the results are clear: there is a strong dependence between the last vote to the general elections and the intention declared for the next ones (X-squared = 7459.9, df = 36, p-value < 2.2e-16). As Figure 2.19 shows, voters appear to be very loyal to their preferred political party along elections, at least when the data is not missing. In the horizontal axis are the voting intentions whereas in colors are represented the memory of the last vote. This "loyalty" is less the case for Ciudadanos, for which electors seem to be redirecting to the PP or VOX. People who do not vote also seem to change of behavior more frequently. Voting for VOX appears as the most "loyal" voter career. This variable should therefore be the one with the most discriminatory power to help us estimate voting intentions.

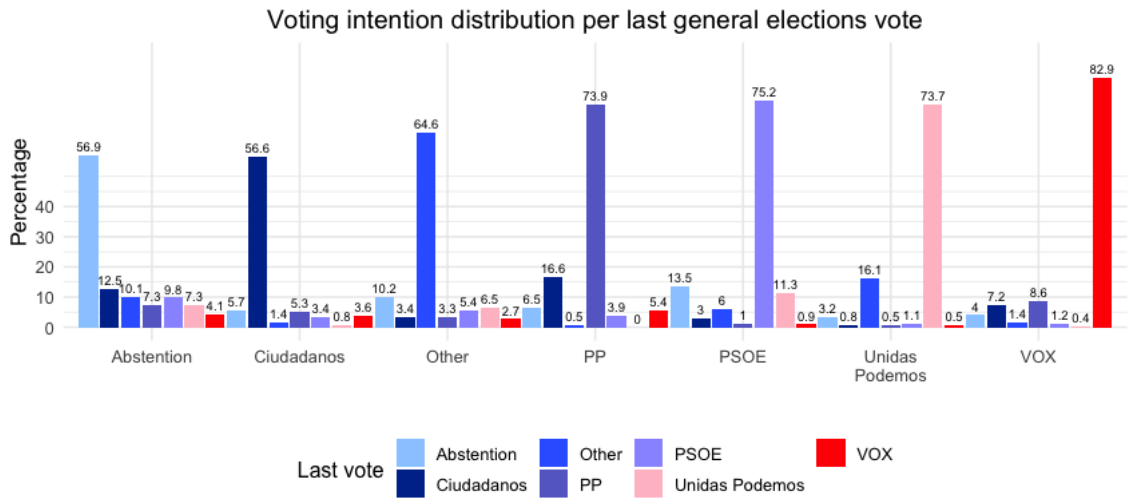

Figure 2.19: Distribution of the voting intentions per last vote recall

Finally, we also crossed the voting intentions with the political party interviewees declare as closest to their ideas (CERCANIA). The result of the chi-square test demonstrates the dependence of the two variables (X-squared = 8591.7, df = 42, p-value < 2.2e-16), as we could have expected. Figure 2.20 shows similar relationships in the distributions as Figure 2.19. Particularly, we find here a variable helping to predict blank vote and abstention intentions, as declaring no political party is close to one's ideas is the most represented affiliation of people intending to vote blank.

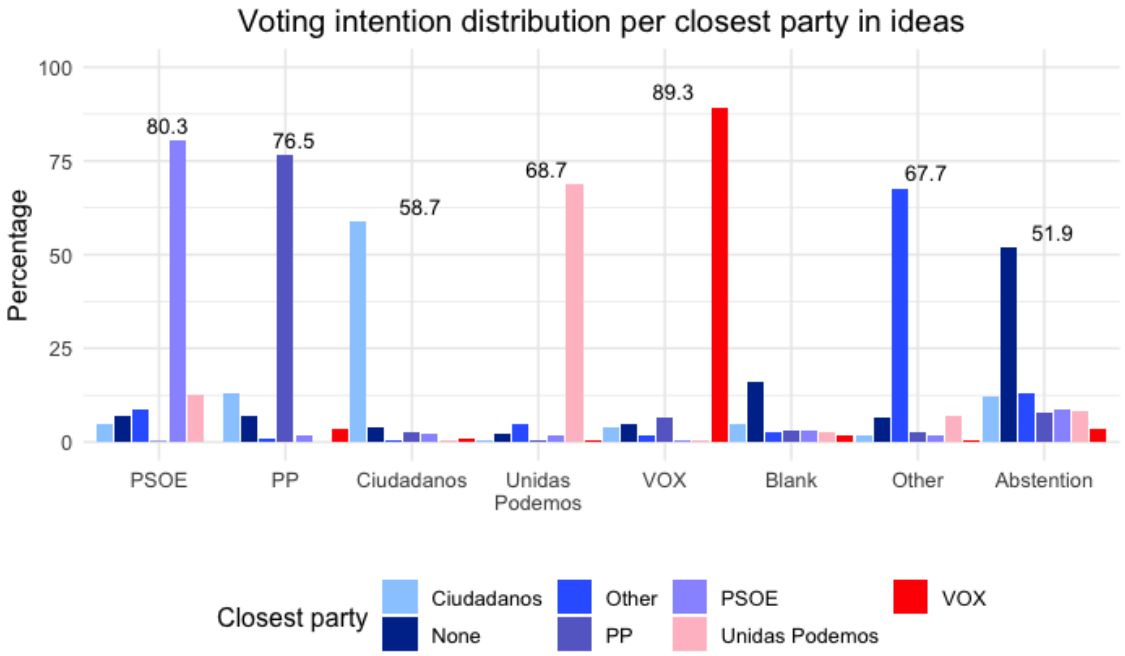

Figure 2.20: Distribution of the voting intentions per closest party in ideas

2.4.3 Quantitative variables

Then, we also looked at the relationship between the voting intentions and some quantitative (continuous or discrete) variables.

First, the age [4] seems to play into the voting intentions. The chi-square test attests it (X-squared = 201.95, df = 42, p-value < 2.2e-16). As Figure 2.21 shows, young voters are more to declare they would vote for Unidas Podemos, VOX or not vote at all. More traditional parties such as PP or PSOE are on the contrary more popular among older adults. However, overall, we can see that all parties have more or less potential voters for all age ranges.

[4] It was recoded in bins such as in Figure 2.4

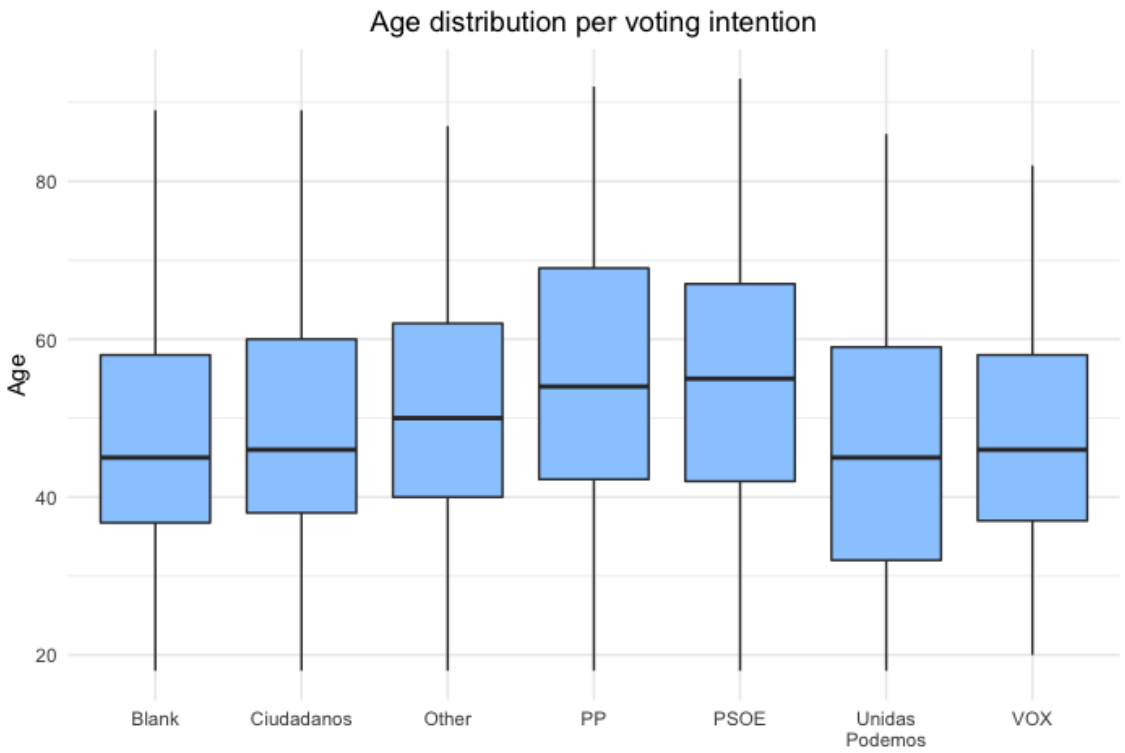

Figure 2.21: Distribution of the age per voting intention

Finally, we studied the relationship between the voting intentions and the Left-Right scale self-positioning of the respondents. Figure 2.22 confirms what was expected; parties associated to the left wing tend to have more potential voters that self-identify to the left, and the contrary for the right wing. But it is not always the case, as we can see some outliers, and for example VOX, which is categorized as extreme-right, does not actually have potential voters that identify to the extreme. To test this association, we grouped the values that had less respondents together (1,2,3 together and 8,9,10 together) and converted the variable to a factor. The p-value of the chi^2-test is very small (<2.2e-16) and the chi-squared statistic of 1938.2, confirming this association (df=35).

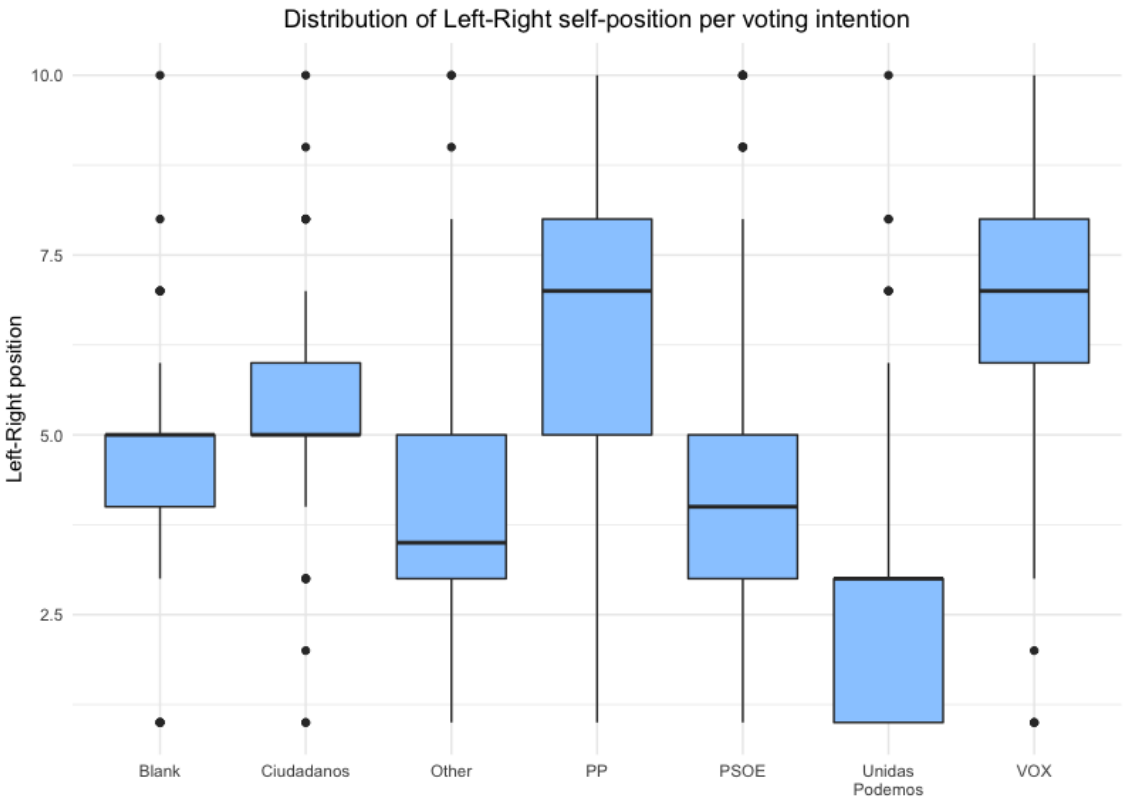

Figure 2.22: Distribution of the Left-Right self-position per voting intention

Therefore, this multivariate analysis has permitted us to get a quick view at the existing relationships between our target variable and some variables of interest. We could discern some predictors that have a high dependency with the voting intention and that should help us impute our missing values.

Now that we have a better understanding of our variables of interest and of the main characteristics of the respondents in regards to their voting intentions, we can tackle our main object of study: estimating the voting intention. To do so, as we have proposed, we are going to impute missing values. In the next chapter, we will cover the methodology behind the method we intend to carry out. We will briefly review the existing research in missing values imputations and present the advantages and the steps of the process we will implement for our case study.

3 | Methodology: state of the art of missing values imputation

Data quality is a crucial determinant of the correctness and utility of statistical analyses. In particular, completeness of the data can affect the results and lead to incorrect interpretations. If the complete subset of the data remains representative of the studied population, it is not important. However, in social surveys, it is well known that respondents may voluntarily not provide some information, for a diverse range of reasons (misunderstanding, refusal, desirability bias...). This type of missing values is specific and cannot be treated using simple techniques if one wishes to obtain confident results.

In this chapter, we will first define missing values and the different mechanisms that can generate them. Then, we will present the classical imputation procedures and their shortcomings. We will finally review existing research supporting the promises of neural networks for missing values imputation.

3.1 Missing values definition

Missing values are usually categorized in three different categories (Rubin, 1976):

- Missing Completely at Random (MCAR): the probability that a value of variable X_i is missing is independent of the values of the others variables in the dataset;

- Missing at Random (MAR): the probability that a value of variable X_i is missing depends on the values of other variables, but not on the values of X_i itself;

- Missing Not at Random (MNAR): the probability that a value of variable X_i is missing depends on the the value (unknown) of X_i itself.

A simple way to deal with missing data is by using deletion techniques. *Listwise* and *pairwise* techniques, focus on the complete data by omitting the records with missing variables (either only the variable or the entire observation/row). However, this can only work in the presence of MCAR mechanism. Apart from deletion techniques, other techniques can be categorized under imputation methods (Buhi, 2008). Imputation means computing appropriate values for replacing the missing data, and it is necessary when data are not MAR.

The important difference between MCAR and other patterns is the likelihood of missing value to occur. If we can detect a relationship between the observed variables and missing values, then it is a hint that the data may not be missing completely at random. It is difficult to test for MAR or MNAR, as we do not have information on the missing values.

We can get an intuition in our case that our values are not MCAR, just by looking at Figure 2.2 which shows the missing value profiles. We underlined that the variables with the highest NA rates are most related to political opinions. As we are in a social survey situation, it is very unlikely that our missing values are MCAR, as most variables are intrinsically intertwined. We have missing values coming from two sources if we simplify: the actual NAs that were missing in the original dataset; and the missing values coming from recoding categories related to lack of knowledge or unfamiliarity. When someone refuses to answer a question, we may believe is it motivated by a particular reason, that could be linked either to the response one refuses to disclose or linked to other personal determinants. When the underlying problem is the lack of knowledge, it is also correlated to personal determinants. Therefore, the mechanisms behind the miss-

ing values in social surveys, especially regarding the delicate political opinions, are not random at all. Many factors can influence one's ability or willingness to answer opinion questions. The age, gender and class positions distributions of non response to political questions are linked with the belief of being entitled to a political opinion (Bourdieu, 1979). But the relationships are not simple, as Bourdieu (1984, p.409) underlines in his works:

> "To understand the relationship between educational capital and the propensity to answer political questions, it is not sufficient to consider the capacity to understand, reproduce, and even produce political discourse, which is guaranteed by educational qualifications; one also has to consider the (socially authorized and encouraged) sense of being entitled to be concerned with politics, authorized to talk politics, by applying a specific political culture, i.e., explicitly political principles of classification and analysis, instead of replying ad hoc on the basis of ethical principles."

Therefore, when trying to compute appropriate values for imputation, the complex relationships that may prevail behind non-response must be taken into account. Relying on simple methods, or giving too much power to a variable that could appear as determinant (like education), is forgetting the complex underlying principles leading to this situation.

Finally, another distinction can be useful between missing values (Silva-Ramírez et al., 2015). In monotone missing data patterns, the missing values are observed for the same observations and variables. In this case, variables can be ordered such that if variable X_i is missing, then for all variables X_j with $j > i$ are also missing. In non-monotone patterns, any observation and any variable can be affected. In our dataset, the missing values are non-monotone as we could observe during the preliminary analysis.

To summarize, the issue of missing values in our case study is complicated. First, because we are interested in estimating one variable with a high rate of

missing values that are not MAR, which requires imputations. Second, because this variable is very correlated with other variables of our dataset, which also have (sometimes many) missing values. Finally, because the mechanisms behind non-response in political opinion questions are complex. Their overcoming require imputation methods able to take into account the many hidden relationships that prevail.

Imputation techniques have been thoroughly studied in the last decades. They can be categorized into two categories (García-Laencina et al., 2010): statistical techniques and machine learning-based techniques. In the following sections, we will briefly present and review the advantages and shortcomings of both groups and justify the approach we will follow.

3.2 Classical imputation methods

Classical imputation methods can be classified according to their degree of complexity[1].

3.2.1 Single imputation methods

First, single imputation methods only compute one value for each missing record.

The easiest and simplest imputation techniques are mean or mode imputations. They consist in replacing the missing values by the mean or mode (especially for categorical variables) of the complete (observed) values of that variable. The computation is easy and the results consistent, but presents several disadvantages: the variance is underestimated (and therefore information is lost), all records are imputed with the same value and correlation with other variables is ignored. This type of imputation is more useful in cases of MCAR mechanism.

Then, regression (conditional mean imputation) goes a step further by fit-

[1]Only the most classical or well known methods are presented here. The extent of the research is high in that field and the purpose of our work is not an exhaustive state of the art.

ting a regression model to predict the missing value. The variable containing missing value is used as the dependent variable, and the other variables as predictors. The process is repeated by changing the target variable until all variables with missing values are covered. As it is an iterative process, one can use the predictions for incomplete predictors done beforehand, which requires cautiously selecting the order of the imputations. Logistic regression can be used for categorical variables. This method is more useful under MAR situations. However, it still underestimates the variance and in the case of linear regression (mostly used), it ignores non-linear relationships.

Finally, hot-deck imputation is using similar observations of the dataset, that are complete, to impute the missing values. It works by using auxiliary variables to determine similarity between observations (Schafer, 1997).

We can also mention deductive imputation, that works using logical rules with auxiliary variables in order to obtain imputations. These methods are useful when the true value is easy to recover through such deductions, which is not often the case.

The main shortcoming of simple imputation methods and that is common to all of them is that it does not take into account the uncertainty of the imputations. Indeed, as only one value is computed, there is no information on the standard error of such estimation. To overcome this, multiple imputation (combining several simple imputations) was proposed (Little and Rubin, 1987).

3.2.2 Multiple imputation methods

Multiple imputation (MI) methods require three phases. First, the imputation: m different datasets are created using values coming from a specific distribution. Secondly, the analysis: the m complete datasets are analyzed. Thirdly, the combination: the m datasets are pooled together to obtain a single final imputation (for example using the mean of the imputed values in each dataset). The first phase allows to gain back uncertainty in the estimations, which is later

introduced in the imputed dataset with the pooling method. The underlying assumption behind most MI methods is that data are MAR.

We can distinguish two approaches when confronted with multiple incomplete variables (Huque et al., 2018). Fully conditional specification (known as sequential regression multiple imputation) imputations use univariate conditional distributions for each variables and performs iteratively. On the other hand, joint modelling uses a multivariate normal distribution and uses the joint posterior to impute values.

There exists different imputation sampling methods, which should be chosen following the missing value mechanism (MCAR, MAR or MNAR) and pattern (monotone or non-monotone). For instance, for monotone patterns, parametric regression methods appear proper; whereas for non-monotone patterns, Markov chain Monte Carlo (MCMC) is more appropriate (Song and Shepperd, 2007).

One of the most generalized imputation methods thanks to its easy implementation in statistical packages is the Multiple Imputation using Chained Equations method (MICE) (Azur et al., 2011). It is a fully conditional specification method based on MCMC sampling. The efficiency of MICE has been extensively underlined, all the more as compared to simple single imputation methods. However, this efficiency can be severely reduced when the dimensionality of the data or the rate of missing value are high. In these cases, the number of iterations needs to be increased (Graham et al., 2007), rising the computational cost. Moreover, the MICE algorithm is based on the linear regression, and therefore does not take into account more complex relationships that could exist between our variables. In this matter, several improvements or changes in the MICE algorithm have been proposed, based on random forests (Stekhoven and Buhlmann, 2011) or hybrid methods combining MICE with machine learning algorithms (Ratolojanaharya et al., 2019).

3.3 Machine Learning for missing values imputation

As we precised, imputation techniques can be classified into two categories: statistical methods and machine-learning based methods (García-Laencina et al., 2010). We described above the first category, and underlined its shortcomings, particularly in dealing with high rates of missing values. Imputation methods based on machine learning could be an interesting alternative. They consist in creating models to predict values that will replace the missing ones. To do so, they rely on the complete cases. The research in that domain has been prolific in the last few years and many options appear promising.

Among them, we can cite using K-nearest neighbors (KNN) to improve hot-deck imputations by considering more than one neighbor and is remarkably outperforming even when missing rates are high (Troyanskaya et al., 2001). Self organizing maps (SOM) has also been proved to work well for data imputation (Fessand and Midenet, 2002), by first computing distances only with the complete variables when choosing image-nodes. Then, weights of the neighbors of the image-node are used to compute missing values in the corresponding dimensions.

Another group of machine learning-based imputation methods rely on artificial neural networks (ANN). Existing literature has explored their performance using different architectures.

The one we chose to rely on are Multilayer Perceptrons (MLP). This may be the most used ANN method for imputation (Gupta and Lam, 1996; Sharpe and Solly, 1995), and has demonstrated its performance in the case of survey datasets (Nordbotten, 1996). The steps for MLP imputation is the following. First, the variables containing missing values are listed. Secondly, for each incomplete variable X_i, a MLP is constructed, using as target this variable. For the inputs, one can either only consider the complete variables or records, or keep the full dataset

by considering NAs as a category[2]. This model is trained on the dataset without missing values in X_i. Thirdly, the MLP model is used for predictions on the missing records for variable X_i.

Other types of ANN have appeared as good candidates for missing value imputation. For instance, the feedback in Recurrent Neural Networks (RNN) can be leveraged to update an initial initialization of missing values (Bengio and Gingras, 1996). Other researchers have explored the use of fully connected neurons in Auto-Associative Neural Networks (AANN) (Narayanan et al., 2002). We can also cite the use of auto-encoders, unsupervised networks that consist of an encoder that ends in a bottleneck layer and a decoder to reconstruct input data. For imputation tasks, the network is able to train on incomplete data by replacing missing values with a simple estimation (average of known values) and only using complete records for computing the error (Abiri et al., 2019). Then, after training, the same "masking" approach can be used to compute the output of the decoder that will give the imputations. Results are promising, as imputations appear better using auto-encoders than statistical methods for imputations, even in the case of high rates of missing data. Finally, hybrid approaches make advantage of the performance of different imputation methods. For instance, combining unsupervised K-means algorithm for computing initial imputations using cluster centers that are then refined using a MLP (Narravula and Vadlamani, 2011). Similarly, it is possible to initialize an auto-encoder using KNN rule to then reconstruct the output instead of a simple mean estimation (Choudhury and Pal, 2010). Another process would be to first train a MLP on complete records to first impute the missing values. Then, with a weighted sum of similarities on complete and imputed cases, KNN is used for final imputatins (Silva-Ramírez et al., 2015). However, as compared to a simple imputation using MLP, this last approach seems better only on quantitative variables.

Finally, one has to note that machine learning methods can be also used not

[2]This is only possible if it is possible to consider all variables as categorical. It is particularly useful in case of high rates of missing values.

for imputation purposes, but for handling classification problems with incomplete inputs. In our case study, as our final aim is to estimate the voting intentions, our problem could be tackled as such. The variable INTENCIONGR would be the target to be predicted, training performed on the complete records for this variable (regardless the missing values in other variables) and predictions made on the uncomplete records. Several machine learning methods are able to deal with missing data to solve a classification task. Decision trees can handle missing values in train and test sets, with algorithms such as ID3 (Quinlan, 1986), C4.5 (Quinlan, 1989) or CN2 (Clark and Niblett, 1989). By extension, ensemble models such as random forests or gradient boosting can also do so. Neural networks ensembles have also been studied to perform classification with incomplete inputs (Sharpe and Solly, 1995).

Therefore, the extent of research in the field of missing values imputation is wide. From the pioneer work of Rubin to the latest hybrid machine learning models, there are many possibilities to handle incomplete data. Simple deletion when missing values are not MCAR is not anymore satisfactory. It has also been demonstrated in the Spanish context in the specific case of people not answering voting intentions questions (Urquizu-Sancho, 2006): this population is very heterogeneous and its vote very split among different parties. Simply deleting the incomplete cases or simple imputation methods would not result in accurate results. In this study, we decide to focus on the use of Multilayer Perceptron for imputations. With many incomplete variables and high rates of missing values, this approach appears as promising. In the next chapter, we will carry out this method on our case study, the CIS Barometer of February 2021, in order to obtain final estimations on the voting intentions. We will follow a similar procedure as the one exposed above, which we will describe in detail, and will require several added steps to adapt to our particular dataset.

4 Case Study

As previously mentioned, our case study in using neural networks for missing values imputation is the barometer of February of the CIS. More particularly, we are interested in estimating a particular type of missing values, the voting intentions. We underlined how the missing information in such a variable is not a simple problem, and hides many complex relationships with other variables. By using neural networks and the methodology explained above, we hope for a more robust estimation of the voting intentions.

4.1 Methodology and preprocessing

In the first chapter, we carried out an analysis of the missing values in our dataset. We could see that there are a lot of NAs, and not only in our variable of interest. Yet, the voting intentions are very related to other variables, as we saw in the multivariate analysis. In order to estimate the most accurately possible the voting intentions, we decided we needed to also impute the missing values of the other variables selected in our dataset, following an iterative approach of imputations. Therefore, we will repeat the procedure for each variable with missing values, following the increasing order in the share of NAs we could observe in Table A1.1 (Appendix A1). At each step, when estimating the missing values for the new variable, we will use the imputations we have done in the previous step.

Our main variable of interest, voting intentions, is not the variable with the largest share of NAs (second one in order). In order to obtain the best results possible for the imputations, we will perform a double pass on this variable, first

imputing it in the "usual" order from less to more NAs and, then imputing it again at the end. A final step of imputations will also be carried out for the regionalist/nationalist parties we grouped together. If one has been imputed to the Left or Right regionalist/nationalist, we will look at his/her community of residence (CCAA) and assign his/her voting intention to the corresponding party.

4.1.1 Training, test and estimation sets

As our methodology is based on the use of machine learning models to impute missing values, one crucial step is to divide our dataset. Here, we are not aiming to predict future values but to estimate missing values in our available dataset. We therefore have to separate our records between the ones that require a prediction, and the ones that are not missing and can be used for training. This is simply done by isolating the complete set from the incomplete set in the variable we wish to impute. As we are going to impute many variables, those sets will differ for each of them. For example, if we aim to impute REGFREFE (preferred political regime), the complete (train) set will be constituted of records that are not NA in REGFREFE. The incomplete (estimate) set will be constituted of records that are on the contrary missing for this variable. If we aim to impute another variable, like INTENCIONGR, it is likely that the incomplete set will be different, as one person can have answered REGFREFE (is in the complete set) but not INTENCIONGR (is in the incomplete set). In regards to which variables are included in the sets next to the variable to impute, two solutions are possible. First, it is possible to use only the complete variables, that do not have any NAs. This approach has the inconvenience of severely reducing the number of variables included in the models which can affect the reliability of our imputations. However, as we are using an iterative scheme (impute values in the increasing order of NAs values), the most crucial variables at the end will benefit from a wider sets of predictors because they have already been imputed. Another approach is to use all variables, regardless if they contain NAs or not. This approach has the

inconvenience that it requires to add an "NA" category to the levels, and therefore recode the quantitative variables as categorical (especially the ones based on a 1-10 scale). The choice between those two approaches will affect the number of columns (variables) in our complete and incomplete matrices, but the number of rows (observations) remain the same, as they were divided depending on the target variable.

As our method is based on the use of neural networks, which are able to deal with large amounts of data and infer complex relationships between variables, we believe that the second option is preferable. It will increase the dimensionality of our datasets but will also most likely lead to better imputations, which is our ultimate goal.

Another consideration to have is that neural networks are stochastic models with a high variance. Every time they are run, they generate a different model, even if the training set is the same. This is because the initial weights are random. To obtain better imputations, the best strategy would be to perform multiple imputations, that is train several models and aggregate the predictions of each model on our estimate set.

After dividing our dataset into training and estimation sets, it is also common to subdivide the training sets for several reasons. First, we can be interested in evaluating our trained models in order to have an idea on how well they would perform on unseen observations (estimation sets). This is important to evaluate the capacity of our models to produce reliable imputations. For each variable, we can compare our performance to that of a stratified dummy classifier (generates random predictions by respecting the training set's class distribution). For the voting intention, we could also compare our final results to the voting estimations published by the CIS, but those must not the considered as the "true" voting intentions since they are unknown. This outer evaluation is done by separating our complete set into a training set, used to train the weights of the network, and a test set to evaluate the accuracy on unseen values. In the test set, we create

artificial NAs that need to be imputed, and compare the results of imputations to our known values on a particular metric. After this outer-evaluation, we need to train the model again on the whole complete set to do our final estimations on the incomplete set, as the more data we have, the better [1].

Furthermore, it is also useful to divide the training set into a training and validation set. The validation set is used like a test set during training for inner-evaluation. This is useful for two reasons. First, it allows to avoid overfitting, that is updating the weights of our network until it learns perfectly the training set (learning noise). Figure 4.1 presents the evolution of the error/loss (y) on the number of epochs (x). As the number of epochs increases, the training error (blue line) decreases. But this is not a good thing, as the model will not be able to generalize to new observations that are not in the training set. To avoid this, we can divide the training set in two subsets: a train and a validation set. The validation set will be used to compute the classification error every time the weights are updated. As we can see on Figure 4.1, at the point marked with a yellow triangle, the validation error increases (red line), whereas the blue one keeps decreasing: this is when overfitting happens. The minimum of the red curve should be the stopping point of the training as it is the time when both validation and training errors are the lowest (early stopping).

[1] And therefore, we should be aware that the outer-evaluation will be pessimistic.

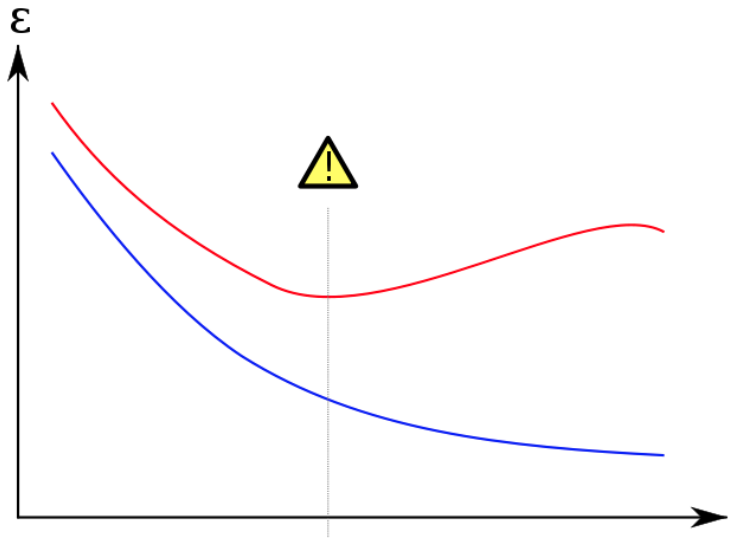

Figure 4.1: Training and validation error

A validation set is also useful for another reason. In a MLP, there are diverse parameters that can be adjusted to improve the performance of the model. Choosing the right combination of parameters is called hyper-parameter tuning. Such parameters affect the complexity of the model (and therefore help prevent overfitting) and can improve the efficiency of training. In a MLP, such parameters are for example the activation function, the solver, the learning rate (value and type of updating). The architecture of the network (number of hidden layers and number of neurons in the hidden layers) can also be treated as an hyper-parameter. In this situation, the training set would be used to train the model on different combinations of hyper-parameters. The validation set will be used for inner-evaluation, that is to evaluate which model with which parameters leads to the better accuracy.

Combining both approaches is possible: we can use the same validation set to do early stopping and inner-evaluate our model on the hyper-parameter values. Another solution is to keep a small stratified part of the train set as another validation set to perform the early stopping (it is a different validation set than the one used to do the inner-evaluation).

Such a division between the sets to perform evaluation is what is called and holdout method. However, it is possible that the test partition does not repre-

sent well the problem (by chance), mainly if the dataset is small. This is even more plausible in our case where some categories are underrepresented in the data. A possible solution is to repeat the train and test (or train and evaluation) by shuffling the data randomly and then dividing. However, this induces an overlap in partitions. A similar approach using non-overlapping partitions is the k-fold cross-validation method. The procedure starts by shuffling and dividing the whole complete set into k folds of equal size. Then, for i in $[1, k]$:

1. Choose i as the test fold.
2. Choose one of the remaining $k - 1$ folds as the validation fold.
3. Choose all remaining folds the training fold.
4. Tune hyper-parameters. Use the the training fold to train the model. Use the validation loss on each trained model to do early stopping. Choose the combination of hyper-parameters that has the minimum loss (inner-evaluation).
5. Train the model with the selected hyper-parameters on the training and validation folds together. Compute the accuracy on the test set (outer-evaluation).

Aggregate results across all the folds (i in $[1, k]$) to obtain the final outer-evaluation. Finally, for doing the final estimations on the incomplete set, we need to repeat partly this procedure in order to train with all available data. First, hyper-parameter tuning is carried out using only a training and validation set. Once the best combination of parameters is found, training of the model is done on the whole complete set in order to predict the incomplete set.

In the above procedure, k-fold cross-validation is done for the outer evaluation, and holdout for the inner-evaluation (hyper-parameter tuning). Another possible approach for more accurate results would be to perform both cross-validation for the inner and outer evaluations. This is the approach we intend to follow, by using a three-folds cross-validation for hyper-parameter tuning (and a

supplementary validation set for early stopping) and a three-folds cross-validation for outer-evaluation. Also, we decide to repeat the process three times (three nested cross-validations) to ensure more accurate results. Nested cross validation does not use the same data to tune the hyper-parameters and perform the outer-evaluation. We therefore avoid data leakage and overfitting (Cawley and Talbot, 2010). In the inner loop we fit a model to each training set and choose the set of hyper-parameters that maximizes the balanced accuracy on the validation set. In the outer loop, we estimate our model performance by averaging the test set scores using again cross-validation (the test sets have not been used for the inner-evaluation).

4.1.2 Preprocessing steps

Before being able to train the models, we need to perform some preprocessing transformations to obtain the best estimations and the most efficient training.

First, specific preprocessing transformations must be considered when training neural networks. Such models are not able to deal with categorical variables. As our dataset contains mostly such types of variables, they need to be transformed to numerical type. Since we do not have many, we can use one-hot encoding for that purpose. It works by replacing each observation by a list of boolean values with 1 in the present category index and 0 in the others. For example, if the observations in the INTENCIONGR are ['PP', 'PSOE', 'VOX, PP'], the 'PP' category would be represented by [1,0,0] and the transformed column would be [[1,0,0],[0,1,0],[0,0,1],[1,0,0]]. This is better than simply encoding the categories as indexes (PP=1, PSOE=2, VOX=3), which would introduce an ordinal order that is not present in the original variable. It has the inconvenience of increasing data dimensionality as one new variable (column) will be created for each category of the original variable.

Then, for the only numerical variable (age, as NA will be introduced as a category for the 1-10 scale variables), another transformation must be performed.

Each variable in the training set may present a different scale. The one-hot encoded variables will be on a 0-1 scale whereas the age will have a larger scale. When this happens, the model will give more importance when updating the weights to the variables that have larger values (here the age). The solution is to scale the variable, such that they all live within the same scale. We can either use normalization (subtract the minimum value from each variable and divide it with the range max-min) or standardization (subtract the mean and divide by the standard deviation). Here we choose to perform normalization, to make sure we obtain a [0,1] range that is the same as the range of our encoded categorical variables.

Finally, as we could see in our descriptive analysis, most of our variables are very imbalanced, that is the class distribution is not even. Some categories hold a high number of respondents, whereas some others very few. It is especially the case for the voting intention variable, that even recoded is not evenly distributed (Table 2.7). Knowing that our dataset is imbalanced, when we are evaluating our models, we should not focus on the simple accuracy measure but consider other metrics such as balanced accuracy or AUC (Area Under the ROC Curve) that are more suitable in this situation. When dividing the sets for training, validation and test, we must also ensure to perform a stratified split on the variable to impute to have the same representation of classes in the partitions (as we are dealing with imbalanced dataset).

Once all those pre-processing steps are performed, it is possible to train our models.

4.2 Results

In Table A1.1 (Appendix A1), we presented the count and relative frequency of missing values per variable in our subset dataset. As we studied before, some variables present a high share of missing values whereas some others only have few. For example, the variable age only has one missing value, which is probably

MAR. The building and tuning of neural networks is a lengthy process, such that carrying out the whole procedure for 40 variables would be very long. Moreover, in the iterative process, those values would be imputed with a high number of missing values in the other variables such that our imputations may be of bad quality. We decide that for the variables with smaller amounts of missing values, we can perform a Multiple Imputation using Chained Equations method (MICE) and not lose too much performance. We do that for the variables with a number of missing values of 1% or less of the total missing values in the dataset (that is in Table A1.1 the last 8 variables that have missing values). Among the 37 variables that have missing values, 30 were imputed using neural networks. The MICE imputation for eight variables was done using package "mice" in R, by imputing the whole dataset but only keeping the imputations done on the eight variables concerned. We realized four imputations with a maximum number of 30 iterations to reach convergence. The computation of the four imputations was done in parallel to reduce the execution time.

We then followed the methodology outlined in Section 4.1 to train models for each variable in increasing order of missing values. We trained our model using sklearn implementation of the MLP classifier (MLPClassifier), as well as diverse libraries useful for preprocessing, hyper-parameter tuning and cross-validation.

4.2.1 Hyper-parameters

As stated, we are using Multilayer Perceptrons (MLP), that is a type of artificial neural network with more than a single layer. The input layer takes our input features (pre-processed). The output layer is the layer that produces the output variables. In our case of multi-class classification, the output layer uses a softmax activation function to assign probabilities to each class of the multi-class problem and then chooses the class with maximum probability. In between, one can choose a given number of hidden layers. For the input and hidden layers, we can choose different non-linear activation functions as well as the num-

ber of neurons (only for the hidden layers). Those are the parameters related to the architecture of the network. By default in the class used (MLPClassifier from scikit-learn), the loss function that is minimized each time weights are updated is the log-loss (cross-entropy loss). It is suitable for classification models with probabilities as outputs. The loss increases as the predicted probability diverges from the true label, as we can see in Figure 4.2. In multiclass classification, a separate loss is computed and summed for each class label per observation: $logloss = -\sum_{c=1}^{M} y_{i,c} \log(p_{i,c})$ where M is the total number of classes, $y_{i,c}$ the binary label of observation i for class i, $p_{i,c}$ the predicted probability for i and class c. Finally, one can choose the optimization algorithm in charge of updating the weights for minimizing the loss function.

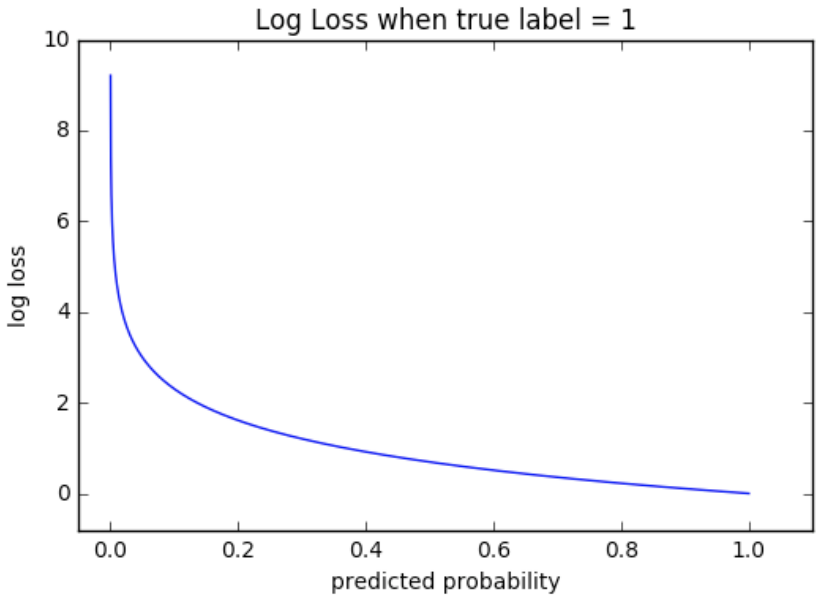

Figure 4.2: Log loss

The computational cost is large for training and evaluation models with our methodology. After several trials, we could narrow the hyper-parameter search space on the parameters that worked the best (in terms of balanced accuracy on the three rounds of nested cross-validations). First, we decided to use two hidden layers, that had better results than a unique one. More layers did not improve accuracy, as two hidden layers are enough to represent any kind of function. As for the number of neurons in each hidden layer, too few would result in underfitting

our training set, and too many would lead to overfitting. For the first hidden layer, we fell upon a fixed number of neurons, 100, that is below the average number of neurons of the input layers (around 300 after pre-processing of the 39 predictors). For the size of the second hidden layer, the performance oscillated depending on the target variable because they differ in number of classes. We kept it as an hyper-parameter to tune between 25 and 50. In Figure 4.3, we can see a representation of this type of architecture for a MLP with two hidden layers and multiple outputs.

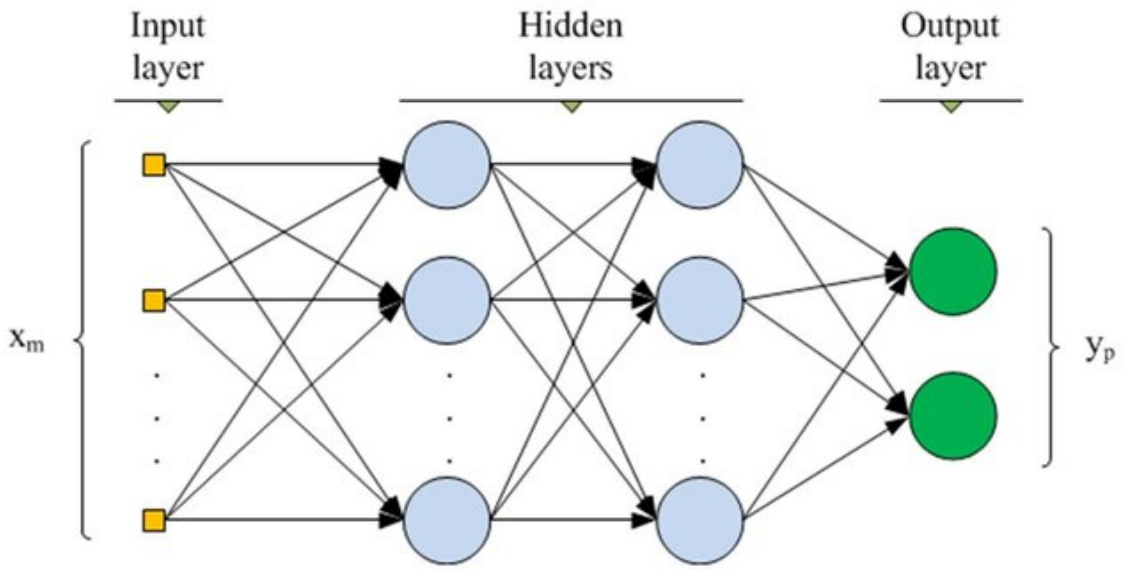

Figure 4.3: Typical architecture of a MLP with two hidden layers

The optimization algorithm used is "Adam", Adaptative Moment Estimation (Kingma and Ba, 2017), a stochastic gradient-based optimizer that computes adaptive learning rates for each parameter which allows to speed up convergence. By using batches and shuffling at each iteration, it is also able to generalize well. Finally, it requires few hyper-parameters to tune, which is convenient as we have many MLPs to train. Finally, we chose as activation function for the input and hidden layers the hyperbolic tangent (Figure 4.4). It proved better on our trials as compared to other functions (logistic or rectified linear unit), and is also known for faster convergence (Lecun et al., 1998). Finally, we set apart 10% of the training set to perform early stopping and terminate training when the validation

score is not improving for ten consecutive epochs.

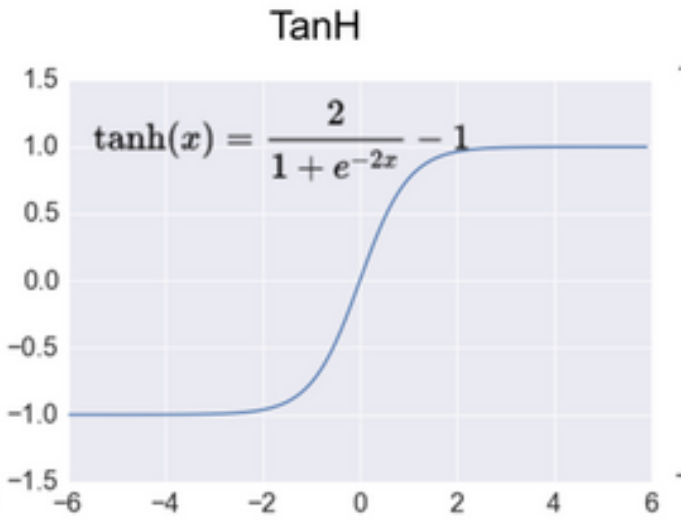

Figure 4.4: Tanh activation function

4.2.2 Outer-evaluation

In the end, we carried out three full rounds of iterative training and predictions to obtain more robust estimations. As explained before, in each round, we did three loops of nested cross-validations. In each loop, we first use three-folds cross-validation for inner evaluation on a training set to select of our final subset of hyper-parameters (number of neurons of second hidden layer). Then we evaluate the model trained on the best parameters using a three-folds cross-validation on a test set. After the outer evaluation, we then repeat the process without any test set to select the hyper-parameters using three-folds cross-validation. The best parameters are taken and the full complete set is used to train the final MLP to predict our incomplete set. Once we have estimations for one variable, we use them as input for the next variable. INTENCIONGR was imputed twice as it is the second variable in terms of missing values (first estimations were used to impute ESCAIDEOLLIDERES_5 then dropped to be estimated again using the imputations of ESCAIDEOLLIDERES_5). Such lengthy process was repeated three times to obtain three final estimations of INTENCIONGR.

In Table 4.1, we present the average balanced accuracy of our three rounds

of estimations for each classification task. We also present the average balanced accuracy of three estimations from a stratified dummy classifier for comparison. As we can remark, we always obtain a better balanced accuracy using MLPs classifiers than a dummy one. More specifically, our results are considerably better for the variables the most associated to the political opinions: CONFIANZAPTE, CONFIANZAOPOSIC, VALORALIDERES, LIDERESCORONA, CERCANIA, PREFPTE, EXTRA, RECUERDO and INTENCIONGR. The results are not perfect, but as we are using balanced accuracy, it is expected as in each variable there are classes that are very little represented. If we simply had used accuracy, the dummy classifier would also have achieved very good results.

	MLP	**Dummy Classifier**
RELIGION	0.28	0.17
P12	0.26	0.20
CONFIANZAPTE	0.62	0.26
P6	0.15	0.14
CONFIANZAOPOSIC	0.48	0.26
VALORALIDERES_1	0.43	0.10
CLASESUB	0.20	0.17
VALORALIDERES_3	0.33	0.10
LIDERESCORONA_1	0.43	0.10
REGFREFE	0.34	0.33
VALORALIDERES_2	0.28	0.10
LIDERESCORONA_3	0.36	0.10
P7	0.57	0.51
CERCANIA	0.56	0.09
LIDERESCORONA_2	0.28	0.10
VALORALIDERES_4	0.32	0.10
ESCIDEOL	0.26	0.10
PREFPTE	0.47	0.11
LIDERESCORONA_4	0.29	0.10
EXTRA	0.35	0.10
VALORALIDERES_5	0.30	0.10
LIDERESCORONA_5	0.27	0.10
RECUERDO	0.57	0.08
ESCAIDEOLLIDERES_2	0.23	0.10
ESCAIDEOLLIDERES_1	0.23	0.10
ESCAIDEOLLIDERES_3	0.17	0.10
ESCAIDEOLLIDERES_4	0.18	0.10
INTENCIONGR	0.61	0.08
ESCAIDEOLLIDERES_5	0.20	0.10
INTENCIONGR	0.61	0.08

Table 4.1: Average outer evaluation of MLP and dummy stratified classifier

Such average outer evaluations, especially regarding CERCANIA, RECUERDO and INTENCIONGR are interesting to give us an idea on our ability to classify correctly the missing values of INTENCIONG. Nevertheless, we must take into account that such evaluations are biased, as the respondents having NA values in voting intentions hold specific characteristics that may not be observed in the complete set.

4.2.3 Estimations

Using the best subset of hyper-parameters for each variable to impute and training on the whole complete set, we could obtain final voting estimations. The left and right regionalist/nationalist categories were reassigned using the community (CCAA) of residence of the respondent according to the political party. In our first imputations, five people remained unassigned in the left regionalist/nationalist party after this first reassignment. Two people from Navarra, that were imputed as EH Bildu (as this party presents candidacies for the elections in the Basque Country and Navarra). Then, two persons were from Madrid and one from Asturias. Looking at the original complete dataset, we saw that eight people from Madrid and three from Asturias had declared an intention to vote for Més Compromís, and therefore assigned the four imputated respondents to this party. In the second and third imputations, this was the case only for left regionalist/nationalist imputed values from Navarra, that were recoded in EH Bildu. In Appendix A1 Table A1.8, we present the detailed counts and frequencies for each round of estimations, which are very similar. In Table 4.2, we present the average relative frequency of each category. At first sight, there does not seem to be any implausible values. However, as those are voting intentions, we cannot compare them to any "true" value. To comment them, we can compare them to estimations obtained from other imputation methods and to the estimations published by the CIS.

	Average percentage
PP	14.1
PSOE	26.2
Unidas Podemos	9.4
VOX	9.0
Ciudadanos	7.9
ERC	2.5
Més Compromís	1.2
EAJ-PNV	1.2
PACMA	1.0
BNG	0.6
JxCat	0.7
EH Bildu	0.5
NA+	0.5
CUP	0.4
CCa-Nc	0.2
PRC	0.1
Teruel Existe	0.1
Other	0.9
Blank vote	5.1
Abstention/Null	18.4
Total	100.0

Table 4.2: Average relative frequency estimations of voting intentions

We also performed alternative imputations by taking another approach for estimating the smallest parties. Instead of creating two categories for regionalist/nationalist parties, we decided to create a unique one. We only grouped into this category the political parties that had very few respondents in the complete dataset of INTENCIONGR: CCa-Nc (five respondents), PRC (four respondents) and Teruel Existe (three respondents). Other regionalist/nationalist parties were kept as separate categories. Again, we ran three full rounds of nested cross-validations and estimations, that are presented in detailed in Appendix A1 Table A1.9 (outer-evaluations), Table A1.10 (detailed estimations) and Table A1.11 (average relative frequencies). The results are almost exactly the same that the last estimations we presented, with very small changes.

In Table 4.3 below, we compare our results with those of Multiple Imputation using Chained Equations method (MICE) with 30 maximum iterations. The results are extremely similar. We can note a difference in the estimations for blank

votes: we estimate 5.1% of blank votes, which is a bit smaller than the 5.9% estimated by MICE imputations. The percentages not allocated to blank vote mostly go to PSOE votes in our models, which are estimated to be 26.2% as compared to 25.2% in MICE. Our approach of estimating voting intentions using an iterative imputation with Multilayer Perceptrons therefore leads to very similar results than an iterative imputation using linear regression models. Nevertheless, we believe that in case of having a much larger sample to impute, our approach would be computationally faster.

	MLP	**MICE**
PP	14.1	14.3
PSOE	26.2	25.2
Unidas Podemos	9.4	9.4
Ciudadanos	7.9	7.9
VOX	9.0	8.9
ERC	2.5	2.6
Més Compromís	1.2	1.2
EAJ-PNV	1.2	1.0
PACMA	1.0	1.1
JxCat	0.7	0.8
BNG	0.6	0.7
EH Bildu	0.5	0.5
NA+	0.5	0.5
CUP	0.4	0.4
CCa-Nc	0.2	0.2
PRC	0.1	0.1
Teruel Existe	0.1	0.1
Other	0.9	0.8
Blank vote	5.1	5.9
Abstention/Null	18.4	18.5
Total	100.0	100.0

Table 4.3: Comparison of relative frequencies of voting intentions imputations using MLPs and MICE

Finally, we can compare our estimations with those published by the CIS for the February 2021 Barometer (CIS, 2021). The CIS publishes relative frequency estimations only taking into account valid votes, that is excluding the abstention/null vote category. We therefore also exclude that category of our estimations and are able to compare the exact same other categories as we designed our

estimations for this purpose [2].

	MLP estimations	**CIS estimations**
PSOE	32.1	30.7
PP	17.3	18.8
Unidas Podemos	11.6	11.2
VOX	11.0	13.6
Ciudadanos	9.7	9.3
ERC	3.0	3.5
Més Compromís	1.5	1.7
EAJ-PNV	1.5	1.5
PACMA	1.3	1.4
JxCat	0.9	1.3
BNG	0.7	0.7
EH Bildu	0.6	0.8
NA+	0.6	0.4
CUP	0.5	0.9
CCa-Nc	0.2	0.4
PRC	0.2	0.2
Teruel Existe	0.1	0.1
Other	1.0	1.7
Blank vote	6.3	1.8

Table 4.4: Comparison of estimations using MLPs and estimations published by the CIS

In Table 4.4, we note that most categories hold similar relative frequencies in both estimations. We can notice one significant difference for blank vote, which is much more prominent in our estimations (6.3%) than in the CIS estimations. Then, we estimated more votes allocated to PSOE (32.1% against 30.7%). We have a 14.8 point difference between the voting intentions for PSOE and PP, whereas the difference is slightly smaller for the CIS estimations (12.9 points). Their estimations are also more important for VOX, 2.6 points more. For the other parties with smaller intention shares, our results are quite similar.

In the end, our results are a little different from the estimations of the CIS. We believe that the major explanation, also explaining the similar results obtained using MICE, is that our methodology uses the complete sets to train the models. Indeed, in Table 4.5, we compare the relative frequencies of our estimations and

[2]The presented percentages are the average relative frequencies of our three final estimations.

of the original dataset among valid answers (not NA, nor abstention or null vote). We can indeed notice that our results are almost exactly the same.

	MLP estimations	**Original complete values**
PSOE	32.1	32.4
PP	17.3	17.2
Unidas Podemos	11.6	11.5
VOX	11.0	11.1
Ciudadanos	9.7	9.7
ERC	3.0	3.1
Més Compromís	1.5	1.5
EAJ-PNV	1.5	1.2
PACMA	1.3	1.3
JxCat	0.9	0.9
BNG	0.7	0.8
EH Bildu	0.6	0.6
NA+	0.6	0.6
CUP	0.5	0.5
CCa-Nc	0.2	0.2
PRC	0.2	0.2
Teruel Existe	0.1	0.1
Other	1.0	1.0
Blank vote	6.3	6.2

Table 4.5: Comparison of estimations using MLPs and original complete values

5 | Conclusion

The barometers of the CIS are mostly known by the Spanish population for their estimation of voting intentions, often commented in the media. One matter that is rarely underlined is that such surveys do not intend to predict the results of general elections, but rather give a picture of the current political opinions of the voters. To provide this snapshot of the political state, one major difficulty arises: the missing values of the question on voting intentions.

In this work, we have reviewed different methodologies for tackling imputation of missing data. In the non trivial context of imputing voting intentions, we have demonstrated the necessity to resort to more complex methods than simple imputations. After a descriptive analysis of our dataset, we saw that political opinions are very related to other variables present in the survey. Based on this analysis, we decided to impute all variables with missing values to obtain better estimations. To do so, we implemented an iterative procedure based on Multilayer Perceptrons, a dense neural network. The main strategy consisted in imputing variables in increasing order of their number of missing values in order to obtain a complete set for our final imputation of the voting intentions.

This procedure was carried out on the barometer of the CIS of February 2021. We took particular care of imputing correctly the under-represented classes for all variables, and especially the small regional political parties for the voting intentions. As the real voting intentions of non-respondents are unknown, we cannot conclude on the accuracy of our result. One interesting observation was the important similarity of our estimations with the imputations carried our us-

ing a Chained Equations method (MICE). Comparing with the results published by the CIS for the barometer of February 2021, the main difference was in the proportion of estimated blank votes, higher in our results. The main explanation for such result is that our models were trained using the complete set of values, which displayed a similar rate of blank votes. In the actual general elections, such blank votes may be lower, as an intention is not a commitment to vote blank. For instance, respondents may be inclined to declare an intention to vote blank instead of abstention for reputation matters. To overcome such biases, one interesting solution would be to include expert knowledge in the models.

Finally, one major disadvantage of this procedure of imputation using artificial neural networks it the "black box" effect behind them. The estimations published by the CIS are often criticized for their opaque "cooking". The use of neural networks does not escape this issue, as they exploit complex underlying methods that would be difficult to explain and justify for an uninformed public.

Bibliography

Abiri, N., Linse, B., Edén, P., and Ohlsson, M. (2019). Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. *Neurocomputing*, 365:137–146.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49.

Bengio, Y. and Gingras, F. (1996). Recurrent neural networks for missing or asynchronous data. *In: Touretzky and al. (eds) Advances in Neural Information Processing Systems 8*, (MIT Press):395–401.

Bourdieu, P. (1972/1979). Public Opinion Does Not Exist. *Communication and Class Struggle*, 1:124–130.

Buhi, E. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior*, 32(1):83–92.

Campbell, A., Converse, P. E., Miller, W. E., and Stokes, D. E. (1960). *The American voter.* John Wiley Sons, Inc., New York.

Castro, C. (2018). El CIS cierra la cocina. La Vanguardia. `https://www.lavanguardia.com/politica/20181008/452207324031/cis-sondeo-polemica-conversion-estimacion-voto.html`. Accessed: 01.05.2021.

Cawley, G. C. and Talbot, N. L. C. (2010). On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.*, 11:2079–2107.

Choudhury, S. J. and Pal, N. R. (2010). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, 182.

CIS (2019). Estudio nº 3242. Macrobarómetro de Marzo 2019. Preelectoral elecciones generales 2019. 1. Nota Metodológica. Modelo CIS v108.

CIS (2021). Estudio nº 3309. Barómetro de febrero 2021. Estimación de voto.

Clark, P. and Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283.

Downs, A. (1957). An Economic Theory of Political Action in a Democracy. *Journal of Political Economy*, 65(2):135–150.

Fessand, F. and Midenet, S. (2002). Self-organizing map for data imputation and correction in surveys. *Neural Computing and Applications*, 10(4):300–310.

García-Laencina, P., Sancho-Gómez, J.-L., and Figueiras-Vidal, A. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2):263–282.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213.

Gupta, A. and Lam, M. (1996). Estimating missing values using neural networks. *Journal of the Operational Research Society*, 47(2):385–401.

Huque, M. H., Carlin, J. B., Simpson, J. A., and Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, 18(168).

Kingma, D. P. and Ba, J. L. (2017). Adam: A Method for Stochastic Optimization.

Lazarsfeld, P., Berelson, B., and Gaudet, H. (1948). *The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign*. Columbia University Press, New York.

Lecun, Y., Bottou, L., Orr, G., and Müller, K. (1998). *Neural Networks: Tricks of the Trade*. Springer.

Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*. John Wiley Sons, New York.

Llaneras, K. and Domínguez, I. (2018). El CIS se ve obligado a explicar la 'cocina' de su encuesta. El País. `https://elpais.com/politica/2018/09/27/actualidad/1538071871_763849.html`. (Accessed: 01.05.2021).

Mayer, N. and Perrineau, P. (1996). Les modèles d'analyse des comportements électoraux. *Cahiers français*, 276:41–50.

Mercado, M. E., Gómez, J. R., and Conde, R. C. (2014). *Los pronósticos electorales con encuestas : Elecciones generales en España (1979-2011)*. Centro de Investigaciones Sociológicas, Madrid.

Narayanan, S., Marks, R. J., Vian, J. L., Choi, J. J., El-Sharkawi, M. A., and Thompson, B. B. (2002). Set constraint discovery: missing sensor data restoration using auto-associative regression machines. *Proceedings of the 2002 International Joint Conference on Neural Networks*, 3:2872–2877.

Narravula, A. and Vadlamani, R. (2011). A novel soft computing hybrid for data imputation. *Proceedings of the 7th international conference on data mining*.

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics*, 12(4):385–401.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Quinlan, J. R. (1989). Unknown attribute values in induction. *Proceedings of Machine Learning Research*, page 164–168.

Ratolojanaharya, R., Ngouna, R. H., Medjaher, K., Junca-Bourié, J., Dauriac, F., and Sebilo, M. (2019). Model selection to improve multiple imputation for handling high rate missingness in a water qualit dataset. *Expert Systems with Applications*, 131:299–307.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman Hall, Florida.

Sharpe, P. K. and Solly, R. J. (1995). Dealing with missing values in neural network-based diagnostic systems. *Journal of the Operational Research Society*, 3(2):73–77.

Silva-Ramírez, E. L., Pino-Mejías, R., and López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, 29:65–74.

Song, Q. and Shepperd, M. (2007). Missing Data Imputation Techniques. *International Journal of Business Intelligence and Data Mining*, 2(3):261–291.

Stekhoven, D. J. and Buhlmann, P. (2011). Missforest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Troyanskaya, O., Cantor, M., Alter, O., Sherlock, G., and al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

Urquizu-Sancho, I. (2006). The non-declared vote in the surveys: The Spanish case in the 1980s. *Electoral Studies*, 25(1):103–128.

A1 | Appendix: Frequency tables

	Count	Percentage
PP	470	13.3
PSOE	988	27.9
Ciudadanos	378	10.7
Més Compromís	51	1.4
ERC	85	2.4
JxCat	34	1.0
EAJ-PNV	41	1.2
EH Bildu	21	0.6
CCa-Nc	7	0.2
NA+	8	0.2
PACMA	29	0.8
VOX	236	6.7
CUP	24	0.7
Unidas Podemos	393	11.1
BNG	18	0.5
PRC	3	0.1
Teruel Existe	1	0.0
Other	55	1.6
None	699	19.7
Total	3541	100.0

Table A1.2: Distribution of the political party closest to one's ideas

Variable	Count of missing values	Percentage
ESCAIDEOLLIDERES_5	850	22
INTENCIONGR	773	20
ESCAIDEOLLIDERES_4	667	17
ESCAIDEOLLIDERES_3	644	17
ESCAIDEOLLIDERES_1	600	16
ESCAIDEOLLIDERES_2	584	15
RECUERDO	581	15
LIDERESCORONA_5	557	14
VALORALIDERES_5	485	13
EXTRA	475	12
LIDERESCORONA_4	444	11
PREFPTE	416	11
ESCIDEOL	407	11
VALORALIDERES_4	358	9
LIDERESCORONA_2	347	9
CERCANIA	328	8
P7	317	8
LIDERESCORONA_3	295	8
VALORALIDERES_2	282	7
REGFREFE	274	7
LIDERESCORONA_1	246	6
VALORALIDERES_3	244	6
CLASESUB	205	5
VALORALIDERES_1	204	5
CONFIANZAOPOSIC	166	4
P6	163	4
CONFIANZAPTE	146	4
P12	85	2
RELIGION	65	2
PARTICIPACIONG	52	1
FIDEVOTO	42	1
P13	32	1
ECIVIL	16	0
ESTUDIOS	11	0
SITLAB	7	0
ESCUELA	2	0
EDAD	1	0
CCAA	0	0
SEXO	0	0
P0	0	0

Table A1.1: Repartition and count of missing values per variable in the subsetted dataset

	Count	Percentage
PP	470	13.3
PSOE	988	27.9
Ciudadanos	378	10.7
ERC	85	2.4
PACMA	29	0.8
VOX	236	6.7
Unidas Podemos	393	11.1
Left regionalist/nationalist	118	3.3
Right regionalist/nationalist	90	2.5
Other	55	1.6
None	699	19.7
Total	3541	100.0

Table A1.3: Distribution of the political party closest to one's ideas, second level of recoding

	Count	Percentage
PP	413	12.2
PSOE	633	18.7
Ciudadanos	479	14.1
Més Compromís	65	1.9
ERC	57	1.7
JxCat	36	1.1
EAJ-PNV	28	0.8
EH Bildu	15	0.4
CCa-Nc	10	0.3
NA+	20	0.6
PACMA	61	1.8
VOX	259	7.6
CUP	24	0.7
Unidas Podemos	447	13.2
BNG	15	0.4
PRC	9	0.3
Teruel Existe	2	0.1
Other	113	3.3
No particular party/Abstention	708	20.9
Total	3394	100.0

Table A1.4: Distribution of the political party towards one feels the most sympathy in general elections (for respondents that did not mention a particular party in INTENCIONG) or for alternative vote in general elections

	Count	Percentage
PP	413	12.2
PSOE	633	18.7
Ciudadanos	479	14.1
ERC	57	1.7
PACMA	61	1.8
VOX	259	7.6
Unidas Podemos	447	13.2
Left Nationalist	130	3.8
Right Nationalist	94	2.8
Other	113	3.3
No particular party/Abstention	708	20.9
Total	3394	100.0

Table A1.5: Distribution of the political party towards one feels the most sympathy in general elections (for respondents that did not mention a particular party in INTENCIONG) or for alternative vote in general elections. Second level of recodings.

	Count	Percentage
PP	467	14.2
PSOE	1035	31.5
Ciudadanos	318	9.7
Més Compromís	35	1.1
ERC	100	3.0
JxCat	27	0.8
EAJ-PNV	39	1.2
EH Bildu	23	0.7
CCa-Nc	12	0.4
NA+	23	0.7
PACMA	23	0.7
VOX	237	7.2
CUP	11	0.3
Unidas Podemos	375	11.4
BNG	20	0.6
PRC	4	0.1
Teruel Existe	1	0.0
Other	22	0.7
Blank vote	66	2.0
Abstention	450	13.7
Total	3288	100.0

Table A1.6: Distribution of the memory of vote in last general elections

	Count	Percentage
PP	467	14.2
PSOE	1035	31.5
Ciudadanos	318	9.7
ERC	100	3.0
PACMA	23	0.7
VOX	237	7.2
Unidas Podemos	375	11.4
Left regionalist/nationalist	94	2.9
Right regionalist/nationalist	101	3.1
Other	22	0.7
Blank vote	66	2.0
Abstention	450	13.7
Total	3288	100.0

Table A1.7: Distribution of the memory of vote in last general elections, second level of recoding

	First imputations		Second imputations		Third imputations	
	Count	Percentage	Count	Percentage	Count	Percentage
PP	547	14.1	547	14.1	547	14.1
PSOE	1014	26.2	1015	26.2	1015	26.2
Ciudadanos	304	7.9	305	7.9	305	7.9
Més Compromís	47	1.2	47	1.2	47	1.2
ERC	94	2.4	95	2.5	96	2.5
JxCat	27	0.7	27	0.7	27	0.7
EAJ-PNV	46	1.2	46	1.2	46	1.2
EH Bildu	20	0.5	20	0.5	20	0.5
CCa-Nc	6	0.2	6	0.2	6	0.2
NA+	20	0.5	20	0.5	20	0.5
PACMA	40	1.0	40	1.0	40	1.0
VOX	346	8.9	347	9.0	347	9.0
CUP	16	0.4	16	0.4	16	0.4
Unidas Podemos	369	9.5	364	9.4	364	9.4
BNG	23	0.6	23	0.6	23	0.6
PRC	5	0.1	5	0.1	5	0.1
Teruel Existe	3	0.1	3	0.1	3	0.1
Other	33	0.9	33	0.9	33	0.9
Blank vote	192	5.0	199	5.1	196	5.1
Abstention/Null	717	18.5	711	18.4	713	18.4
Total	3869	100.0	3869	100.0	3869	100.0

Table A1.8: Detailed estimations of voting intentions for each round of imputations

	MLP	**Dummy Classifier**
RELIGION	0.28	0.17
P12	0.26	0.20
CONFIANZAPTE	0.63	0.26
P6	0.14	0.14
CONFIANZAOPOSIC	0.49	0.25
VALORALIDERES_1	0.42	0.10
CLASESUB	0.17	0.17
VALORALIDERES_3	0.33	0.10
LIDERESCORONA_1	0.43	0.10
REGFREFE	0.34	0.33
VALORALIDERES_2	0.34	0.10
LIDERESCORONA_3	0.35	0.10
P7	0.57	0.51
CERCANIA	0.39	0.06
LIDERESCORONA_2	0.29	0.10
VALORALIDERES_4	0.33	0.10
ESCIDEOL	0.26	0.10
PREFPTE	0.50	0.11
LIDERESCORONA_4	0.30	0.10
EXTRA	0.28	0.05
VALORALIDERES_5	0.30	0.10
LIDERESCORONA_5	0.30	0.10
RECUERDO	0.49	0.05
ESCAIDEOLLIDERES_2	0.22	0.10
ESCAIDEOLLIDERES_1	0.24	0.10
ESCAIDEOLLIDERES_3	0.17	0.10
ESCAIDEOLLIDERES_4	0.12	0.10
INTENCIONGR	0.54	0.05
ESCAIDEOLLIDERES_5	0.23	0.10
INTENCIONGR	0.53	0.05

Table A1.9: Average outer evaluation of MLP and dummy stratified classifier, alternative imputations

	First imputations		Second imputations		Third imputations	
	Count	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**
PP	547	14.1	547	14.1	547	14.1
PSOE	1014	26.2	1014	26.2	1014	26.2
Ciudadanos	305	7.9	305	7.9	305	7.9
Més Compromís	47	1.2	47	1.2	47	1.2
ERC	95	2.5	95	2.5	95	2.5
JxCat	26	0.7	26	0.7	26	0.7
EAJ-PNV	47	1.2	47	1.2	47	1.2
EH Bildu	19	0.5	19	0.5	19	0.5
CCa-Nc	5	0.1	5	0.1	5	0.1
NA+	20	0.5	20	0.5	20	0.5
PACMA	40	1.0	40	1.0	40	1.0
VOX	347	9.0	347	9.0	347	9.0
CUP	16	0.4	16	0.4	16	0.4
Unidas Podemos	366	9.5	366	9.5	366	9.5
BNG	22	0.6	22	0.6	22	0.6
PRC	5	0.1	5	0.1	5	0.1
Teruel Existe	3	0.1	3	0.1	3	0.1
Other	33	0.9	33	0.9	33	0.9
Blank vote	202	5.2	202	5.2	202	5.2
Abstention/Null	710	18.4	710	18.4	710	18.4
Total	3869	100.0	3869	100.0	3869	100.0

Table A1.10: Detailed estimations of voting intentions for each round of alternative imputations

	Average percentage
PP	14.1
PSOE	26.2
Unidas Podemos	9.5
VOX	9.0
Ciudadanos	7.9
ERC	2.5
Més Compromís	1.2
EAJ-PNV	1.2
PACMA	1.0
BNG	0.6
JxCat	0.7
EH Bildu	0.5
NA+	0.5
CUP	0.4
CCa-Nc	0.1
PRC	0.1
Teruel Existe	0.1
Other	0.9
Blank vote	5.2
Abstention/Null	18.4
Total	100.0

Table A1.11: Average relative frequency estimations of voting intentions, alternative imputations