UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Machine Learning approach to Multidimensional Opinion Polarization

Informatics engineering
Specialization in Computing

Author: David Cárcamo Delgado
Director: Romualdo Pastor Satorras
Co-director: Michele Starnini
GEP tutor: Joan Subirats Soler

January 23, 2023

Contents

List of Figures

List of Tables

Abstract

Due to the globalization we have undergone in recent years, the study of public opinion and how it is influenced has become an increasingly studied field.

To date, many studies have observed behaviors that polarize opinion in the political sphere. Analyzing how small groups of questions have changed over time.

This thesis aims to study the incorporation of machine learning tools for the study of opinion. These will allow us to work with larger volumes of data and find relationships that otherwise might not be found.

We will study the effect of various attributes on people's opinions, focusing mainly on the effect of political leaning on people's ideology. We will also study how different topics encapsulate an ideology and how these ideologies have varied through time.

Resumen

Debido a la globalización que hemos sufrido durante los últimos años, el estudio de la opinión de la población y como esta es influenciada ha resultado en un ámbito cada vez más estudiado.

Hasta la fecha, multitud de estudios han observado comportamientos de polarización de la opinión en el ámbito político. Analizando pequeños grupos de preguntas para ver cómo ha cambiado la opinión a lo largo del tiempo.

Esta tesis pretende estudiar la incorporación de herramientas de aprendizaje automático para el estudio de la opinión. Estas nos permites trabajar con mayores volúmenes de datos y encontrar relaciones que podrían no encontrarse de otra forma.

Estudiaremos el efecto de varios atributos en la opinión de la gente, centrándonos sobre todo en el efecto que tiene la inclinación política en la ideología de la población. También estudiaremos como diferentes tópicos encapsulan una ideología y como estas han evolucionado con el tiempo.

1 Context and scope

1.1 Introduction

In the last century, due to the rapid development of technologies, the interaction between societies and individuals of all kinds has gone from being slow and having less impact to being instantaneous and accessible to all. These advances have contributed to anyone being able to get to know new cultures, and an infinity of opinions, and thus be able to rethink their way of seeing the world. However, this global opening may not have caused greater flexibility, but rather the opposite effect on society.

Contact with the whole world could have degenerated into even more closed groups. The facility offered by the Internet to interact and start new groups encouraged by specific ideas could have promoted increasingly polarized communities.

The objective of this final degree project is to find whether we are suffering a polarization of opinion and analyze how this polarization behaves.

We will try to reach a conclusion by taking advantage of the calculation power that computers offer us. Specifically, we will base this study on finding a solution through the use of machine learning techniques, which will allow us to work with unaffordable volumes of data for one person and find relationships that would otherwise remain hidden. These types of techniques are increasingly useful as we can reach increasingly complex conclusions by increasing the amount of information we process.

In this study we will focus specifically on the case of the United States, using the data collected by ANES [1] (American National Election Studies). ANES surveys the population every two or four years asking all kinds of questions about their political vision and allows us to know the socio-economic framework behind these individuals. The study by ANES has been carried out since 1948 and will allow us to study the aspects that have changed in North American society during the last 70 years.

1.1.1 Implied Actors

This project is aimed at researchers in social disciplines. Seeing the techniques used in this work, they will be able to evaluate the use of machine learning implies in their respective studies. These can be both within the sector of political polls and of any

kind that can benefit from these strategies.

The personnel involved in the realization of this project will be:

- Romualdo Pastor Satorras: Project director.
- Michele Starnini: Project co-director.
- David Cárcamo Delgado: Project author.

1.2 Contextualization

1.2.1 Previous studies

As already discussed, the objective of this project is to study the evolution of the polarization of opinion. For this, we will use the data provided by ANES to carry out a study on the last years.

The use of ANES to carry out studies on political polarization and its evolution is not new, to date, multiple articles have studied different facets of political polarization. Article [2] studied the evolution of affective polarization towards political parties using different political opinion survey questions, see figure 1.

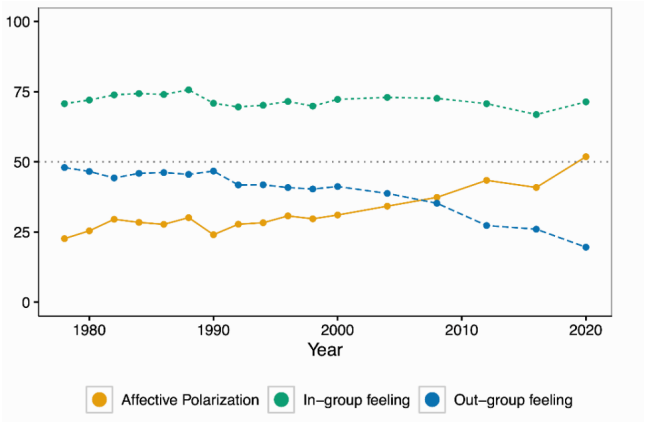

Figure 1: Evolution of political affection [2].

Article [3] studies opinion evolution for democrats against republicans for multiple individual variables. See figure 2.

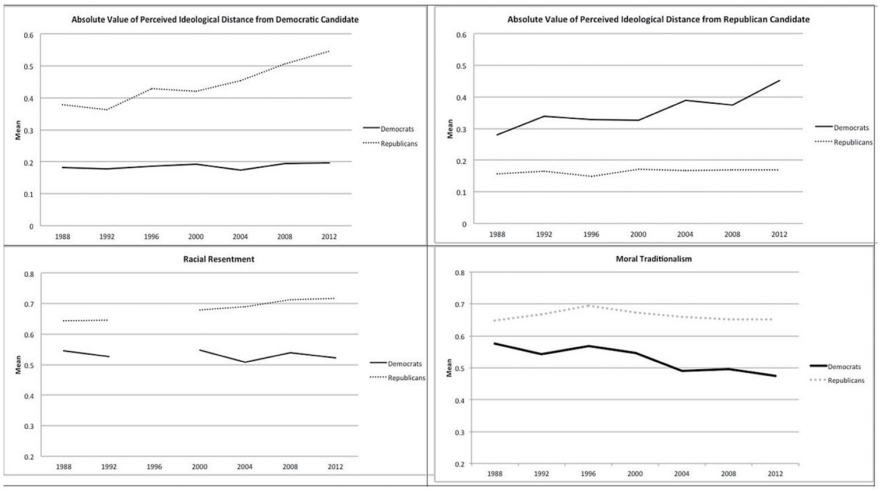

Figure 2: Evolution of opinion between Republicans and Democrats [3].

Multiple other studies have studied group polarization studying its effect on multiple questions of the dataset.

1.2.2 Justification

Many studies have displayed political behavior using ANES but the use of machine learning techniques has remained minimal. Instead, lots of effort has been centered in study the evolution of single variables over some time and the correlation between them.

The purpose of our project is to use machine learning techniques to study the evolution of opinion polarization. This allows and encourages us to work with larger data volumes. In this way, we will be able to find complex relationships between individuals and study the evolution of polarization in the United States. We will also study distinct communities defined by their ideology.

The work does not involve any technical innovation but will study the use of different techniques which haven't been applied before for political analysis.

1.3 Scope

1.3.1 Project objectives and sub-objectives

The main objectives that we have set for the project are:

- Perform an embedding of opinion vectors into low-dimensional spaces to quantify global polarization also with respect to political leaning.

- Reconstruct the network of topics to identify clusters of topics that form an ideology
- Analyze the time evolution of opinion polarization and topic ideology.

To meet the objectives of the project, we must meet a series of sub-objectives that will serve to achieve the main objectives.

- Understand the data to make correct use of it.
- Choose the optimal data set for the study, clean it and prepare it for data processing.
- Identify and study different methods of dimensionality reduction.
- Identify and study different clustering methods.

1.4 Methodology and tools

1.4.1 Methodology

This project is based on research and therefore is not subject to a clear development in which clear tasks can be specified to be carried out at all times. The way of working will be advancing the work week by week and defining the objectives to be met for the following week. That is why an agile work philosophy is the one that best suits our needs. The agile methodology allows us to refine the work of one of the four large groups of tasks that we have defined week by week and allows us to constantly improve considering the new setbacks and results that we find. In this way, we will have the slight flexibility that is needed when doing research work.

To be able to carry out this development, weekly meetings will be held in which the work carried out during that week will be presented through a presentation made by the engineer. Based on the results obtained or difficulties encountered, an action plan and/or objectives to be met for the following meeting will be discussed.

1.4.2 Tools

To be able to process all the information, we will need a program that allows us to manage the data comfortably. With the tools needed to transform the data. For this,

we will program in Python using Visual Studio Code with the help of several of its libraries. The principal ones are:

- Numpy, for array management.
- Sklearn and scipy, for applying different machine learning techniques.
- Matplotlib and seaborn, for data representation.

2 Methods

This section will explain all the methods use in order to obtain the results of the study.

2.1 Data preprocessing

2.1.1 Dataset characteristics

ANES [1] (American National Election Studies) performs surveys every 2 to 4 years in which they ask the American population about their opinions on all sorts of topics before and after elections. ANES has been performing these studies since 1948 and proposes a random sample of opinions scattered throughout the United States.

Every study consists of two sessions in which the questions are asked to the respondent. The first session is conducted previous to the election and the second one afterward.

Every edition of the study presents the respondents with a large set of questions designed for every edition. The citizen is presented with questions that will record a large set of aspects of his opinion.

As the ANES studies questions vary between every edition, they have compiled all the data from all the year's surveys into a single time series. In this dataset, they have joined all the questions performed in different editions which can be compared. To accomplish this they have removed all the questions asked in one or two editions. Questions that depend on previous answers have been joined with their predecessors.

The time series dataset compiles the opinion of 68225 respondents, for the edition covering as far as 2020, for which it stores 1029 variables. Out of these 1029, 7 variables are information relative to the survey, such as the corresponding year or information about the interviewer, 26 collect information about the respondent, such as age or ethnicity and the remaining 996 collect all of the respondent ideologies collected through the questions of the survey.

2.1.2 Selection of useful variables

Our objective is to study the ideology of the American population and so, not all of the time-series questions will fit our needs. Out of the 996 questions in the dataset, we

need to select those which are useful to encapsulate the ideology without introducing significant bias.

The first type of questions that are not desirable is the categorical ones. The dataset has a selection of categorical questions, hence, is not possible to place them on an ordinal scale, some of them ask for:

- Occupation of relatives.
- Specific likes and dislikes about parties and candidates.
- Qualities of a candidate which make the respondent want to vote against him.

Once the categorical topics are discarded the next type of question that we want to avoid is personal opinions on political representation. Being ANES a study interested in capturing all the information as possible of the population the dataset has questions of all sorts. A great chunk of them is focused on obtaining all of the opinions of the respondent about the elections, such as:

- Opinion of each party candidate.
- Positioning of each party candidate on diverse socio-economic topics.
- Grading the parties' policies on different socio-economic topics.
- Rating of different parties.
- Opinion on the congress.
- Number of likes and dislikes about a candidate.
- Opinion of senators.
- Comparison between both parties.

The main problem these questions oppose is that they don't have much information about the ideology of the individual. A party represents a set of ideas, and opposing a party or favoring them does not inform about their opinion on those specific topics, only their opinion on the whole picture. That's why, directly working with the topics the population is concerned with, and not with information around parties, results in a more accurate view of the individual.

Another problem presents when studying the population's opinion through time. The people involved, as the candidates, are not the same every edition and would result in joining opinions of different people as if it was only one.

Once filtered all the unwanted variables the list of questions to use collects the respondent's opinion about topics such as:

- Spending importance on different areas.
- Abortion.
- The church and religion.
- Women rights
- Foreigner rights.
- Black rights and status.
- Concern about war.

These questions inform about how every respondent views the world and what the correct way to proceed is.

Once the questions have been filtered we account for another factor. The study aims to detect any sign of opinion polarization throughout the years. Working with questions that have only two possible responses forces the respondent to choose a side. This way, polarization is synthetically introduced and so we create a bias that will route us to conclude there is a polarization of opinion. These wouldn't be beneficial, so we also ignore this type of question.

After all the data has been filtered, results in a total of 96 questions of interest amongst all years, these questions, along with their identifier can be found in annex A.

2.1.3 Selection of shared variables

ANES studies vary their questions according to the specific situations encountered for every specific edition. This means that some questions are not shared between different surveys. In early editions of the studies, for example, the Vietnam war was still on course and it was asked to the respondents their opinion on the matter. Nowadays

Vietnam is not discussed in the survey. Instead more questions on personal values are presented.

Figure 3 displays the number of questions any pair of surveys share. These are displayed only for editions after 1984, as editions previous to this rarely share more than 20 questions with any other survey, peaking at 25 questions between 1972 and 1976. For the surveys in the figure, we observe in the diagonal the number of questions each survey has that are contained in the 96 questions deemed as relevant for our study. Out of the 15 studies represented 4 of them have significantly fewer shared questions with other years than the rest. These are 1984, 1986, 1998, and 2002. We also observe that 1992, 2000, 2004, 2008, 2012, and 2016 have the largest number of questions in common and will be the surveys that will allow us to keep the maximum number of variables to work with.

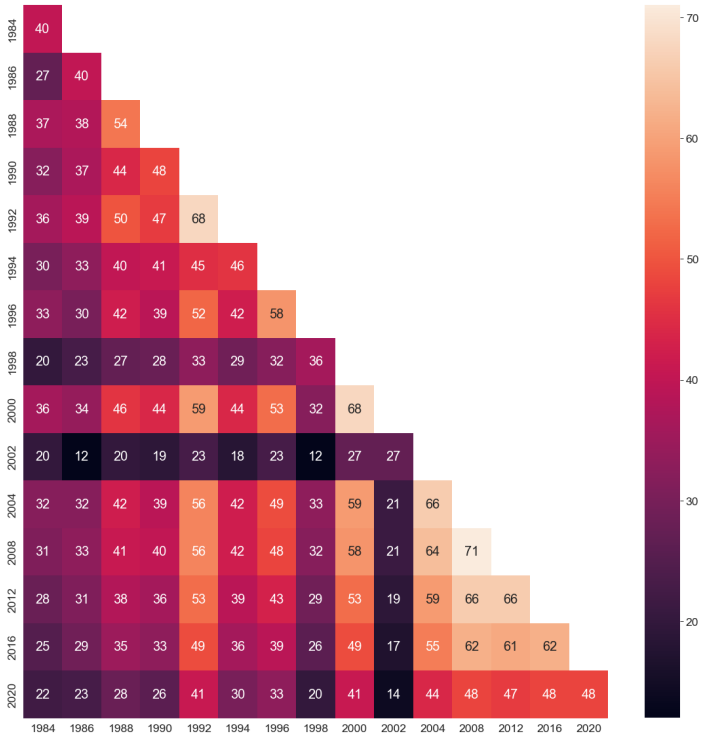

Figure 3: Shared number of questions between every year's study (1984-2020), own elaboration.

When grouping more than just two surveys, the number of shared questions will diminish and as more surveys are taken into account the lower it will be. Figure 4 displays the number of shared questions when considering different sizes for the groups of surveys. We observe that considering groups of 2 to 4 surveys yields between 50 and 55 variables in common. For groups with more than 4 surveys, the number of

shared questions decreases significantly with each added survey. For this reason, we will focus on these small groups of surveys.

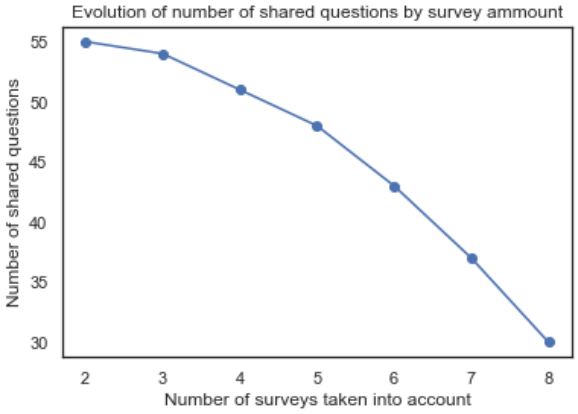

Figure 4: Shared questions with different sizes of surveys, own elaboration.

Table 1 contains the set of surveys that we found more interesting to study and their corresponding number of shared questions.

Survey editions	Number of shared variables
(1992, 2016)	49
(1992, 2012)	52
(1992, 2008)	55
(1992, 2004, 2016)	47
(1992, 2000, 2008)	52
(1992, 2004, 2008)	54
(1992, 2000, 2008, 2016)	47
(1992, 2000, 2004, 2012)	48
(1992, 2004, 2008, 2012)	51
(1992, 2000, 2004, 2008, 2012, 2016)	43

Table 1: Number of shared questions for a different set of surveys, own elaboration.

Out of the sets of years obtained, there are 2 groups containing 3 surveys that maintain an equal distance between their years. (1992, 2004, 2016) and (1992, 2000, 2008), even so, they're not the set of surveys with more shared questions for a subset of three surveys. Similarly, with sets of 4 surveys, there's (1992, 2000, 2008, 2016) which maintains equal distances between years. We also have a set of six surveys with almost full coverage since 1992.

We opted to focus on a set of three studies for simplicity, specifically (1992, 2000, 2008) as it provides three surveys with a similar quantity of respondents, around 2,000 per survey. (1992, 2004, and 2016) in contrast, contains more than half its population in the 2016's survey.

While will focus on a specific set of years, others will be used in order to contrast the information retrieved.

2.1.4 Variable consistency

We find that in the dataset different variables present inconsistencies on their format. In order to have a direct meaning between the question and the answer we have coded the questions such that the code of the answer implies the opinion of the respondent. To accomplish this, we have reordered the codes of the answers so the lowest value represents disagreement with the question and the highest value represents total agreement with the statement.

For example question, VCF0847 asks the individual their guidance from religion. The responses to this are coded in the following manner:

1. Some
2. Quite a bit
3. A great deal
5. Religion not important

In this case, 'Religion not important' should be the lowest coded answer.

Also were found questions with gaps between some answers' codes. Answers coded {1,2,4,5} were present. When normalizing these questions in the [0,1] interval they would produce a large gap in the middle which is not representative of the answers. These questions were re-coded so they would have contiguous codes.

The final change to the codes of the answers made was to the missing codes. The dataset uses different missing codes to differentiate between types of missings. Missings could be generated by not wanting to answer or because no particular opinion is held by the respondent. We treat all these missings the same and remove their code.

With all the data correctly coded, we normalize the data in the [0,1] interval so all the questions have the scale no matter the number of answers they propose.

2.1.5 Missing Data

Missing values in the dataset refer to all the answers to specific questions which are not filled in. The underlying factor to which the data is missing categorizes missing data into one of three categories.

MCAR

Missing completely at random (MCAR) presents when the missing values are randomly distributed, finding no pattern to the missing data. This scenario would be optimal.

MAR

Missing at random (MAR) presents when the missingness of the data is not random and has different value rates depending on some characteristics of the respondent.

MNAR

Missing not at random (MNAR) presents when the missing value depends on the true value. This case is the most difficult to treat, as to know whether the missingness is explained by the value itself would be needed to know this value.

ANES missing pattern

ANES [13] does not assure that all the missing data is MCAR, finding several variables with MAR missing data. No study has found the ANES data to be MNAR, thus, we will consider the data to follow MAR.

2.1.5.1 Missing identification

In figure 5 can be observed the missing rates for each of the 52 questions/topics and for each of the three years of the set. All the data corresponding to 1992 is consistent in their missing rates, always below 20%. But, 2000 and 2008 present spikes of missing rates in some questions which surpass the 45%. Combining both years there are a total of 13 of these questions.

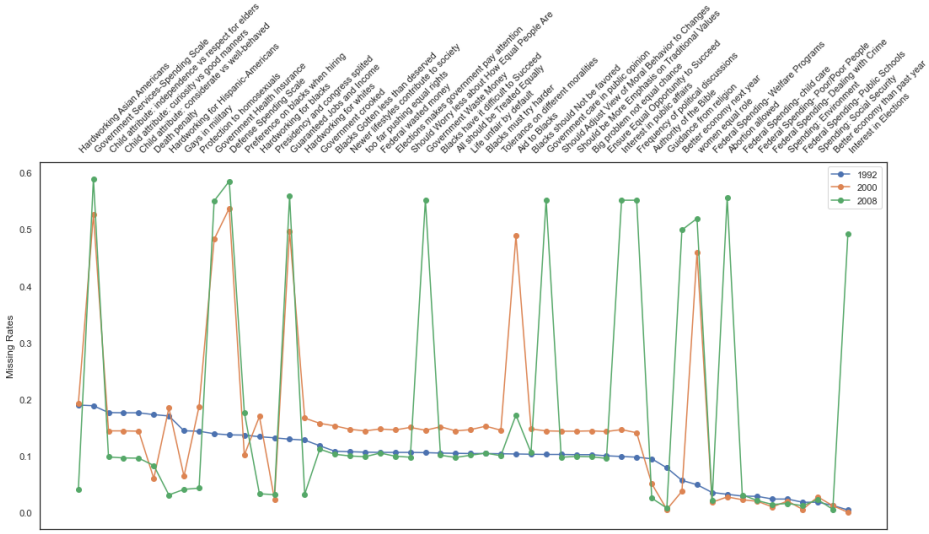

Figure 5: Missing Rates 1992, 2000, and 2008, own elaboration.
X-axis contains every question out of the 52 selected questions shared by the 1992, 2000, and 2008 surveys. Y-axis displays the missing rate for every question.

Multiple algorithms require the data to have no missings, therefore we will need to process the data to fill in or delete the missings encountered.

2.1.5.2 Variable deletion

When a large percentage of the population does not respond to a given question can lead to bias. To avoid this we opted to delete variables with high missing rates. We considered high-missing variables all those that exceed a 25% missing rate.

We have seen that for the set of surveys of 1992, 2000, and 2008, the resulting number of questions after deleting these variables is reduced from 52 to 39.

2.1.5.3 Missing imputation

As the data is assumed to be MAR, the missing data is not related to the value that should fill it, thus we can fill the missing data with new values. There are multiple methods to impute missing values. Some of the most common are:

- Filling with the mean value of the variable.
- Filling missing values with the last non-missing value found.
- Applying a regression model to estimate the specific value.
- Use the nearest neighbors to an individual to interpolate the answer.

Out of these using a regression model or nearest neighbors provide typically the best results. We opted to use nearest neighbors (KNN) for the task. We will specifically use sklearn's KNNImputer.

In [11] the effectiveness of KNN imputation is discussed and is found that increasing the number 'K' of neighbors to impute the data resulted in increasingly distorted space. They found that using 1NN, while not resulting in the best method for applying regression was the only one that preserved the original data structure. For this reason, we will use 1NN to impute all the missing data.

2.2 Low dimensional embedding of the respondents

2.2.1 Dimensionality reduction

The type of data with which we will work relates an individual with a series of fields, in this case with each of their responses to the survey. We could represent any response to a singular question as a value in one axis. Having multiple variables would be treated as a multidimensional space with every question defining a single dimension. This way, if in the survey we were presented with the data of 10,000 individuals and the survey had a total of 100 questions, we could represent the system in a space of 100 dimensions in which there would be 10,000 points, each one representing an individual.

Dealing with such a high number of dimensions poses several problems.

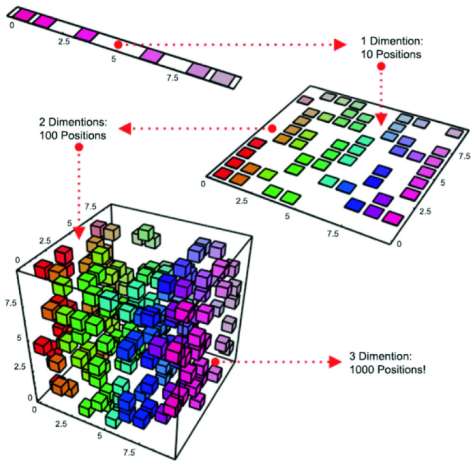

Figure 6: High dimensional sparsity [6].

In a space with high dimensions, the points are isolated from each other. Figure 6 visualizes this problem with lower dimensions. The possible positions in space grow

exponentially with the number of dimensions while the samples are the same. This data sparsity makes it difficult to find relationships between data.

Another challenge high dimensional data opposes is computation cost. Many techniques suffer when working with high dimensional data, resulting in slow execution times.

Finally, we encounter the representation challenge. Understanding high-dimensional data is an exponentially difficult endeavor for us, as we cannot visualize it. The bigger the space is, the harder it gets to get a grasp on the data.

To fix these problems is common to use some sort of dimensionality reduction technique. Which will embed our data into a lower dimensional space which we will refer to as embedding. We will embed our data into two dimensions.

2.2.2 Dimensionality reduction techniques explored

There are multiple dimensionality reduction techniques. Out of them, we tried some of the most widely used.

The techniques studied are: UMAP [21], TSNE [22], PCA [23], metric MDS [24], and ISOMAP [25].

From all of these, we opted for the use of ISOMAP. ISOMAP is an extension of the MDS algorithm consisting of three phases:

1. **Nearest neighbor search.** To relate the data ISOMAP uses a graph. This graph will have as vertices all the users of the surveys. The algorithm used to generate this graph will be the K-nearest neighbors. An edge will join every user to the K-nearest users in the original space. Each edge will have a weight corresponding to the euclidean distance between every pair of connected vertices.

2. **Shortest-path graph search** From the graph, all the distances between any two points are computed using a Dijkstra or a Floyd-Warshall algorithm.

3. **Partial eigenvalue decomposition.** This step performs a metric MDS. Metric MDS finds a set of points in lower-dimensional space such that the distances between the points match the distances between the data as much as possible.

We have chosen ISOMAP mainly due to the application of distances in the embedding.

ISOMAP preserves as much as possible the structure of the data by preserving the distances found between any two points in space. This yields a space from which we can analyze points based on their position in the embedding.

Methods such TSNE or UMAP retain the local or global structure of the data but do so compromising the other, being necessary a fine-tuning of the hyperparameters to reach a correct result. This compromise makes that the distance between two points is not guaranteed to have a meaning.

PCA and MDS are also a good candidate for measuring the data in the embedding, but ISOMAP tipically yields a more accurate representation of the data.

2.2.3 Creation of the embedding

Creating different embeddings for data from different surveys would more accurately resemble the opinion for that specific year. But, different embeddings, generated with different data wouldn't be directly comparable as the spaces would be different. So we create a single embedding containing all the surveys to study. The data fed to ISOMAP to create this new embedding is only the individual responses to the set of questions. Information about the year of the survey or personal information about the respondent such as age or gender is not introduced.

ISOMAP hyperparameters

The hyperparameters used in sklearn's ISOMAP function were:

- Number of neighbors (n_neighbors): 10.
- Number of dimensions (n_components): 2.
- Path method (path_method): Dijkstra.
- Eigen solver (eigen_solver): Dense.

2.2.4 Comparison of political and socio-demographic attributes

With the 2-dimensional embedding representing individual ideologies, we can classify every respondent based on different attributes. Each of these attributes divides the population into different communities. Some of the attributes divide the population

into more than two communities. In these cases, we have adjusted the communities to be a pair, opposite between them. The attributes used to classify the population are:

- Political leaning, comparing democrats and republicans.
- Social class, comparing the working to the middle class.
- Ethnicity, comparing the white and the black population.
- Gender, comparing males to females.
- Age, comparing the population younger than 35 years old and older than 55 years old.
- Income, comparing the lower 33% (0-33%) to the higher 33% (67-100%).
- Education, comparing those with some college attendance and those without it.

The belonging to one community is given by the response of each respondent. All respondents that do not fit into a community, whether there was not an answer by their side or because it represents a community outside of the ones studied, are not taken into account for the study of that specific attribute. There is no imputation of missing values for personal attributes.

Once the population is divided into two communities according to an attribute, we can locate the center of the opinion of every community. We compute this center of opinion as the center of the system. The mean of all individual ideologies on the embedding belonging to a community will be the center of the opinion of that community.

With these centers of opinion, we can track how different communities are related between them and how they evolve throughout time.

2.2.5 Statistical tests

We will make use of two statistical tests to study the distribution of the communities.

Null model

We make use of a Null model to prove if the center of opinion obtained for every community has significance. Our null hypothesis then is that the center of opinion obtained can be explained by random data distribution. This would mean that a random set of opinions could have yielded the center of opinion, thus, the center of the opinion of that community does not provide new information. The alternative hypothesis is that the center of the opinion of a given community has significance, and will be proved if the null hypothesis is disproved.

To test the null hypothesis we will run a simulation. This simulation will run for 10.000 iterations, computing the location of centers of opinion that results from assigning randomly every attribute. If the real center of a community is found within the 5% most uncommon positions of the test, then we can reject the null hypothesis. This 5% is from the assumption of a p-value of 0.05.

The computation of the experiment to test the null hypothesis follows the following steps:

1. Take a single attribute, such as sex.
2. Consider only the individuals belonging to the two studied communities, male and female in this case. Assign every individual to one of the two communities randomly, in such a way that the proportion between males and females remains equal to the one found in the data.
3. Repeat the random assignment of community 10.000 times, saving the center of opinion for all iterations.
4. Find the center of all random iterations for both communities and compute the distance to every random center. Additionally, compute the distance between the center of the random distributions and the real center.
5. Sort the distances to the center.
6. Take the distance which is larger than 95% of them all.

Running this algorithm for every attribute yields for every community the distance between the random distribution center of opinion and two locations. The first is the real center found on the embedding for the given community. The other is the distance from which only 5% of the random centers will be further apart, let's call this distance p-distance as it represents the p-value.

If the distance between the random and real centers is larger than the p-distance, then the probability of the real center being explained by random data is below 5% and we reject the null hypothesis.

2.2.5.1 Kolmogorov-Smirnov test

When segmenting the population by some attributes we can find that an attribute has no impact on the final center of the opinion of the community. To prove if this applies to any attribute we will perform a Kolmogorov-Smirnov test.

Specifically, we will use a 2D 2-sample Kolmogorov-Smirnov test [27]. This determines for any two communities if they are significantly different. We will test only the communities defined by the same attribute, as democrats to republicans or males to females. If the p-value obtained is lower the 0.05 both communities would follow different distributions, therefore, the actual placement related to that attribute is significant of the ideology in the embedding.

2.3 Cluster of topics forming an opinion

Another objective of the project is to identify clusters of topics that form an ideology and analyze its time evolution. To get these we need to relate the answers of all the individuals to the different questions. To compute how different questions relate to one another we will make use of the set of shared questions between 1992, 2000, and 2008 surveys reached in section 2.1.

2.3.1 Creation of clusters

To create clusters of topics we first need a measure of the similarity of any two variables. For this, we will use the Pearson correlation coefficient. Pearson correlation measures the linear correlation between two variables, resulting in a value in the range [-1,1]. Where a positive number indicates a positive correlation between both variables. And a negative indicates a negative correlation. The higher the absolute value of the correlation the greater this one is.

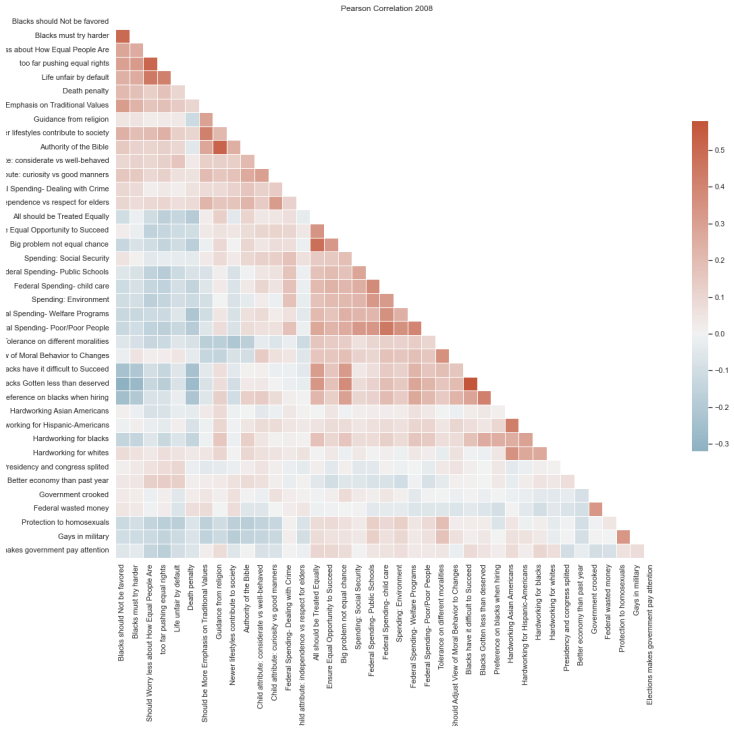

Figure 7: Pearson correlation matrix for data in the year 2008, own elaboration.

Figure 7 displays the Pearson correlation for all variables for the 2008's respondents. We see that the Pearson correlation for topics reaches higher values for directly related topics. We also see that, although some questions seem very similar, the highest correlation peaks lower than 0.6, far from having two redundant variables.

To create the clusters of topics we have taken the Pearson correlation for every survey in the set. Then, for every matrix, the dissimilarity between any two variables is defined by $1 - |Correlation|$. Using this method we lose the information on the sign of the correlation. To cluster the different variables, we perform agglomerative clustering using Scipy [31].

Agglomerative clustering starts by treating every individual variable as a single cluster. Then, the dissimilarity matrix is used to compute distances between clusters. The two nearest clusters are joined into one and the distances are updated. By repeating this method we can obtain the exact quantity of clusters we desire.

We will use all the different measures of distance provided by Scipy to test how they behave.

2.3.2 Comparison of clusters

As we do not have attributes to use to compare different sets of clusters we will create our own to have as a baseline.

From all the questions we deemed as relevant to the individual ideology, we have created clusters. Variables have been grouped when they asked for opinions of the same nature. For example, economic questions have been grouped and questions about ethnicity and foreign opinion have been also joined. This procedure has been performed for 4,5, and 6 clusters, annex D contains the specific clusters used.

To compare the generated clusters to our own, we have used normalized mutual information score [29], which provides a score between zero and one for the similarity of two communities.

3 Results

3.1 A low dimensional representation of individuals in the space of topics/questions

The embedding of the respondents in the space of the topics was performed as explained in section 2.2

To create the 2-dimensional representation of the respondents' ideology the individuals of all three surveys were used. None of the data preserved either the year they conducted the study or any individual attribute such as age or gender.

This data have been fed to ISOMAP to create the 2-D embedding of the individuals. This results in a cloud of points representing every single individual. Figure 8 shows the result of the embedding in a density map to see how the respondent ideology has been embodied. We find in the center a region with a high density of opinions. Around it, we have a bigger pocket with less density of opinions expanding to the left and down.

The area represented on top and to the right is less densely populated and we can expect them to be the less shared opinions.

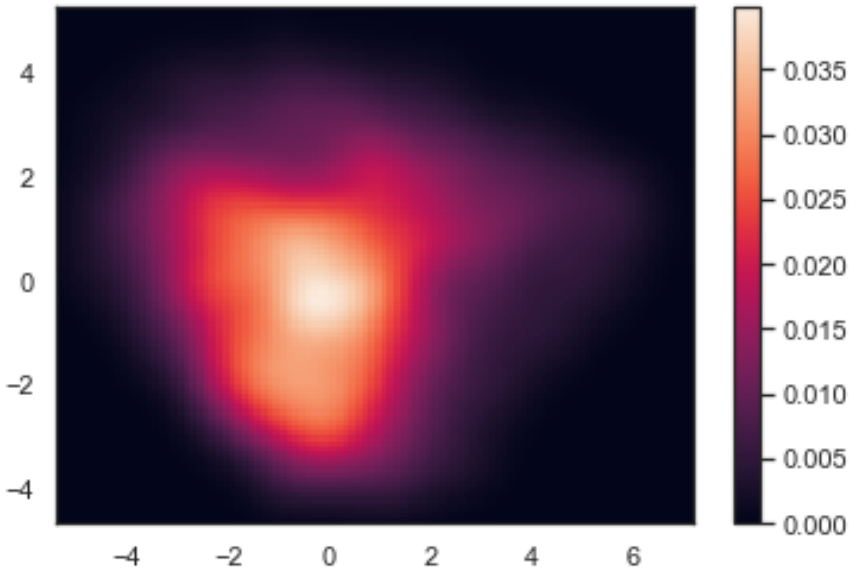

Figure 8: Density map of the 2-dimensional embedding of the respondents, own elaboration.

3.1.1 Effect of political leaning on ideology

When trying to differentiate between people's ideologies the most common feature to look up is their political leaning. Democrats and Republicans are defined by their political opinion and as such is a candidate factor to have significance on the space by itself.

Figure 9 displays the center of opinion in the embedding for democrats and republicans. This center of opinion is displayed for the population of every survey of 1992, 2000, and 2008.

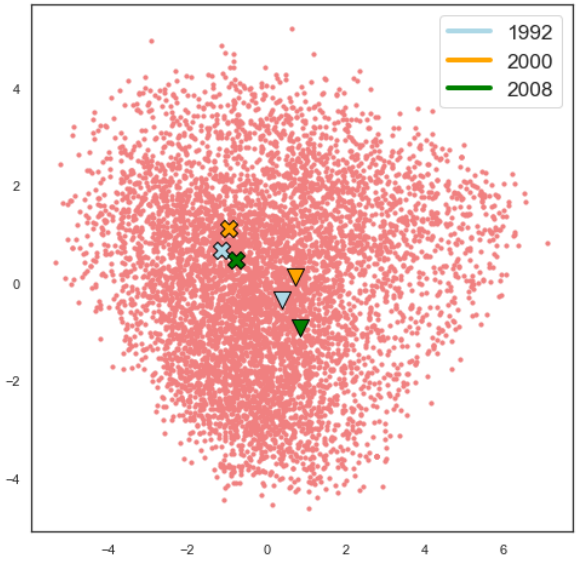

Figure 9: Center of the opinion of Democrat and Republican communities, own elaboration.
The figure displays in red the positions of all opinions in the embedding. On top of them displays the center of opinions for each community. Crosses represent the center of opinion for republicans while democrats are represented by triangles. Blue, orange, and green correspond to the survey considered for finding each center, from 1992, 2000, and 2008 respectively.

We can see that the placement for democrats and republicans is different enough their centers of opinion don't overlap. The republican population seems to be more oriented toward the top left area while the democrats are oriented to the bottom right. We observe that the location found for both parties throughout the years has changed. This shift in opinion seems to be global to the whole population as both democrats and republican centers of opinion have shifted similarly. This shift doesn't present any tendency as 2000 deviates upwards from 1992 while 2008 is situated below. Even so, the democrat opinion has deviated more, as the republican centers are found closer between them.

Both Republican and Democrat centers are situated around the orbit of the most densely populated area of the embedding. This was expected as both populations are constituted by a similar amount of individuals.

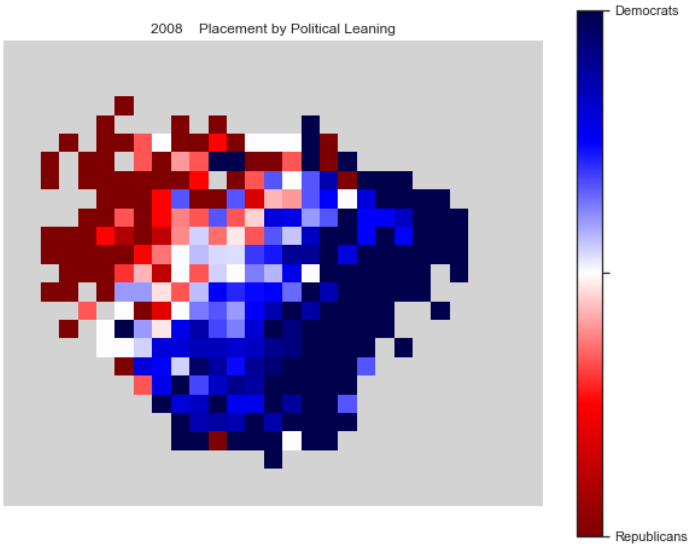

Figure 10: Presence of republican and democrat population in the embedding for 2008's respondents, own elaboration.
Displays for every section in the embedding the relative population for both political communities for the 2008's population. The color in each bin displays how dominant a community is in that location. A deep blue color means 100% of the population in that section is democrat, while dark red means 100% of that section's population is republican. White represents an equal amount of both communities. Gray is reserved for all locations where no individual is placed.

Figure 10 exemplifies better how democrat and republican communities are distributed in the embedding by showing how the space is filled with both communities. We observe that Democrats constitute almost all the population in the bottom and right part of the embedding. Republicans are dominant only in the top left. The center part, where most respondents are situated, is filled with a similar amount of republicans and democrats. This stripe where both leanings meet spans across all the embedding in a diagonal.

With these figures, we see how relevant is political leaning to the outcome of the answers to the questions of the surveys. ISOMAP has embedded the opinion of every respondent without knowledge of the party each individual supported. And yet, republicans and democrats were placed in opposite sections of the embedding.

3.1.2 Comparing other socio-demographic attributes

Figure 11 displays the centroids in the embedding for all the additional attributes. Social class, ethnicity, gender, age, income, and education.

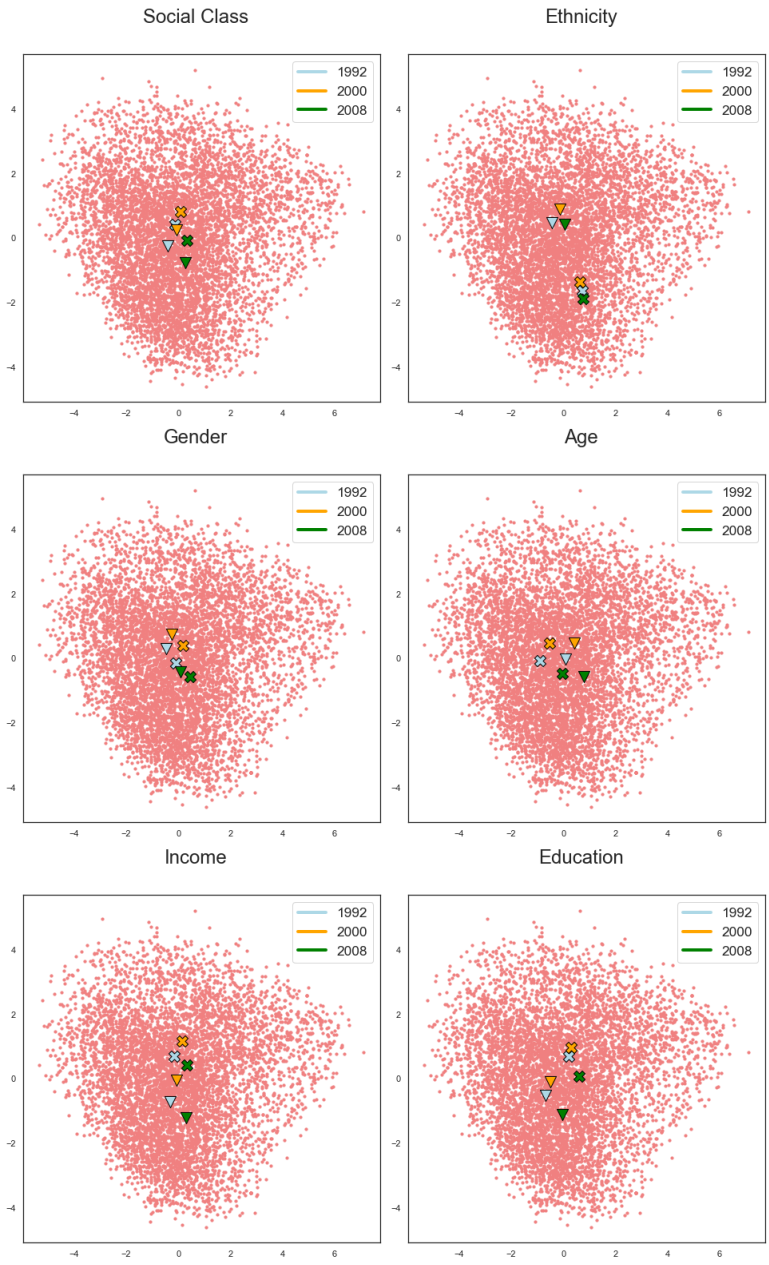

Figure 11: Community centers of other socio-demographic attributes, own elaboration. For every socio-demographic attribute the center of the opinion of their communities is displayed for every one of the three surveys, 1992, 2000, and 2008. The centroids are represented with the following symbols: For social class; middle with crosses, working with triangles. For ethnicity; black with crosses, white with triangles. For gender; females with crosses, males with triangles. For age; older than 55 with crosses, younger than 35 with triangles.For annual income; the top 33% with crosses, bottom 33% with triangles. For education; college with crosses, no college with triangles.

We see that the centers considering age, social class, gender, and income have centers of opposing communities near each other. For the education level and ethnicity of the respondents, the centers are more apart from each other. For ethnicity, the center of opinions is significantly more distant than for the other attributes.

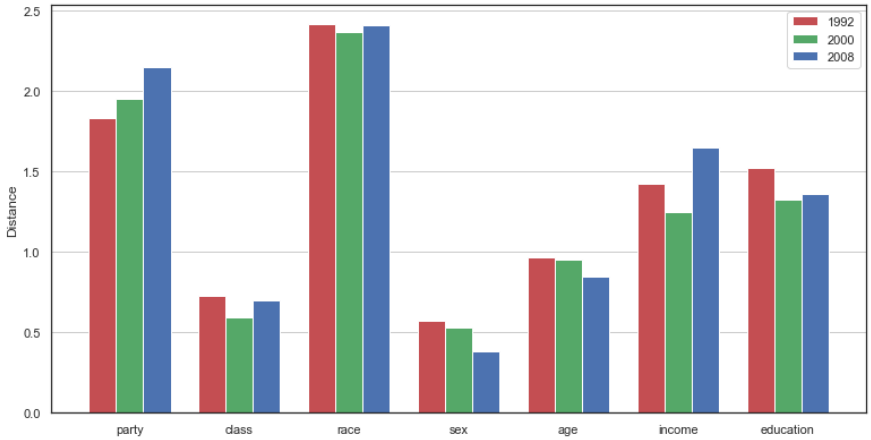

Figure 12: Distance between centroids of each political or socio-demographic feature, own elaboration.

Figure 12 compiles the distances between both communities within each attribute and tracks its evolution through time. The two attributes which present a larger gap between their communities are ethnicity and political leaning. These are followed by income and education which have similar distances between their communities. Last we find age, social class, and gender.

We find some tendencies throughout the surveys. The effect of political leaning on opinion seems to increment with time, this is the only attribute that seems to have this tendency. Other attributes such as sex and age behave the opposite way, with each survey having smaller distances than the last. Ethnicity seems to remain constant. And social class, income, and education don't present any tendency in their distances.

Based on these distances, we observe that the opinion extracted from the questions and reflected in the new embedding always presents a certain difference between communities. We see that the attributes which the embedding has differentiated best are political leaning, ethnicity, income, and education.

3.1.2.1 Effect of ethnicity in the location in the embedding

Since ethnicity resulted in the single factor which segmented the communities the most is worth checking how the black and white populations compare and are placed in the

embedding.

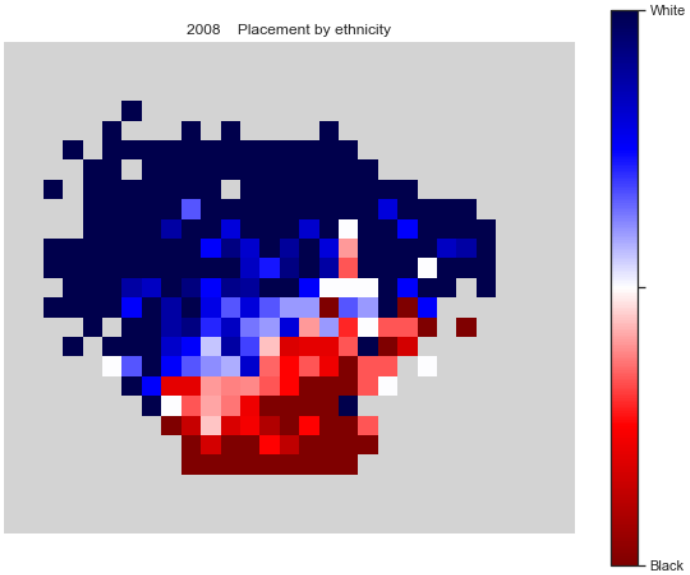

Figure 13: Presence of white and black population in the embedding for 2008's respondents, own elaboration.

Figure 13 shows the dominance in the embedding for the white and black communities for the 2008 survey. In 2008 the black population of the survey is half as large as the white. Even so, the opinion of the black community is concentrated in a small region at the bottom of the embedding. Leaving all the top areas to be occupied by the white population.

Comparing these proportions to the ones found when dealing with political leaning we observe how the democrats are differentiable by race. Black democrats constitute the majority of the bottom area while white democrats are more present towards the right sector. The section with a similar amount of black and white has democrat predominance. And so, the republicans are located in general away from the black population. This distance between black and republicans is supported by the fact that for the 2008 survey only 4.6% of the black population identified themselves as republicans.

3.1.2.2 Comparison of the most relevant attributes

Figure 14 displays the location of the center of opinion for the top 4 most differentiable attributes according to the distances found between their centers: political leaning, ethnicity, income, and education. These centers correspond to the population found

in the 2008's survey.

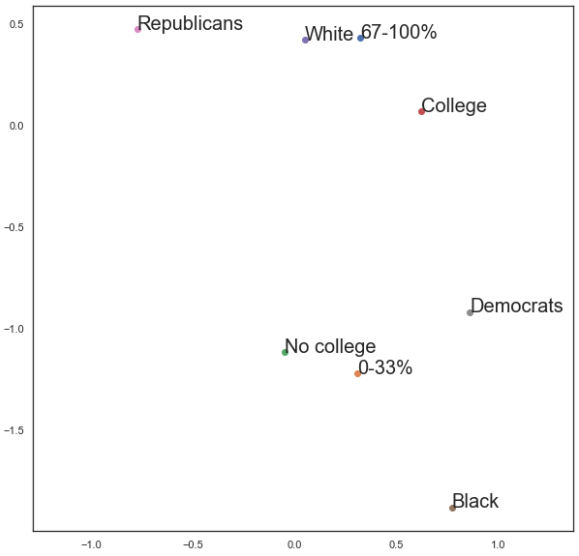

Figure 14: Location of most significant community centers in the embedding, own elaboration.

We observe that there is a close distance between income and having attended college with the highest earners center being close to the college population and the low-income population having their nearest neighbor in the population without college. The high-income population also finds themselves close to the white and republican populations, while the low-income population is close to the democrats and the black.

In general, the community centers seem to be divided into two groups:

- Republicans, white, and the top third in annual earnings.
- Democrats, Black, and the bottom third in annual earnings.

But this case is not the case for education and political leaning. As can be seen in table 2 the center for democrats is similarly close to both people with and without college attendance. For the republicans this distance with both levels of education is not as similar, being nearer to the college population by some margin. In 2008's survey 50% of democrats had attended college and the other half didn't. Republicans didn't behave the same way, two-thirds of republicans had attended college. This way the level of education doesn't seem to propose a great difference in the political leaning.

	College	No college
Republicans	1.45	1.74
Democrats	1.02	0.93

Table 2: Distances between political leaning and education centers of opinion, own elaboration.

3.1.3 Low dimensional representation of different sets of years

Once seen the results obtained from the surveys of the years 1992, 2000, and 2008, other sets of surveys may be explored to get more insight into the evolution of ideology. These can help us to discern if the conclusions are consistent throughout different groups of surveys. Annex B contains the results obtained on different sets of surveys.

After studying different sets of surveys, results of different populations present to be robust. Despite different surveys having different questions in common and having different populations the conclusions drawn from them are very similar to the ones retrieved for the first set of surveys studied.

All the embeddings follow the same hierarchy of distances between centers of opinion. Being political leaning and ethnicity always the most distant factors. The rest of the attributes behaved similarly to the first set of surveys studied, with gender being the attribute it varied less.

While no pattern seems to be preserved for the rest of the attributes, political leaning presents always increasing distances between the center of its communities. Being 2012 the year where it suffers the greatest increase. These results can be interpreted as an increase in political polarization throughout time.

3.1.4 Statistical tests

We tested both statistical tests discussed in section 2.2.5, to see the relevance of attributes in the embedding.

3.1.4.1

*Null model

Table 3 contains the results of the execution testing the null hypothesis. The table contains the results for both communities in every attribute and every survey

separately. We observe that for every community the actual distance is further than the distance of the 0.05 values. We can reject the null hypothesis, thus, we prove the alternative hypothesis that the data does not follow a random distribution.

	1992		2000		2008	
	distance	p-distance	distance	p-distance	distance	p-distance
Democrats	0.98	0.12	1.09	0.14	0.86	0.1
Republicans	1.3	0.16	1.42	0.19	1.81	0.21
Middle	0.51	0.14	0.35	0.14	0.52	0.15
Working	0.47	0.13	0.41	0.16	0.4	0.11
White	0.43	0.05	0.4	0.06	0.98	0.1
Black	2.64	0.33	2.7	0.41	1.99	0.21
Female	0.34	0.12	0.28	0.13	0.21	0.11
Male	0.39	0.14	0.36	0.17	0.28	0.15
17-34	0.6	0.15	0.68	0.22	0.54	0.17
55+	0.67	0.17	0.53	0.17	0.44	0.14
0-33%	0.95	0.17	0.69	0.18	0.76	0.13
67-100%	0.88	0.15	0.88	0.23	1.32	0.23
No college	0.96	0.13	1.06	0.19	0.94	0.13
College	1.07	0.14	0.67	0.12	0.86	0.12

Table 3: Null model distances, own elaboration.
The distance column represents the distance between the real center and the center of all the random iterations performed. The p-distance column represents the distance from the random iterations center which comprehends 95% of the results.

Kolmogorov-Smirnov test

Table 4 contains the p-values for the 2-sample Kolmogorov-Smirnov test. All p-values are smaller than 0.05. Thus all communities are drawn following different distributions compared to their counterpart.

Year	Party	Class	Race	Gender	Age	Income	Education
1992	1.01e-33	7.06e-14	2.58e-43	3.7e-06	5.61e-12	2.26e-26	4.14e-39
2000	3.84e-32	1.42e-05	2.25e-37	1.13e-05	8.21e-09	1.72e-14	5.54e-24
2008	1.93e-48	5.57e-11	1.7e-70	0.000301	2.05e-07	5.91e-26	5.11e-38

Table 4: Kolmogorov-Smirnov test for every attribute, own elaboration.

3.1.5 Effect of ISOMAP hyperparameters

As it's studied in annex C, the effect of ISOMAP's K hyperparameter, referring to the number of neighbors considered when generating the graph, has no significant effect on our results. The difference we find when considering other values of K is the magnitude of the distances between centers of opinion. With lower K's creating bigger embeddings, the distances between centers increased while using larger K's decreases the span of the embedding. This has no effect on the importance found for different attributes, political leaning, and ethnicity are the most polarized attributes. The rest of the attributes follow the same hierarchy of importance we found initially no matter the embedding. Being sex less significant attribute for all cases.

Statistical tests performed for different values of K yielded the same results. All communities have significance for the final ideology and communities within an attribute have distinguishable populations.

The results obtained are not defined by ISOMAP's K as different values yield the same conclusions.

3.2 Topic clusters representing an ideology

Using the joined data of 1992, 2000, and 2008 surveys we want to create and analyze topics which form an ideology.

3.2.1 Study of obtained clusters

As explained in section 2.3 we have created clusters of topics forming an ideology. We have compared these clusters with some we handpicked in order to have some to compare the solutions with.

Figure 15 displays the mutual information score between the generated clusters and our selection. The clusters are generated for every survey and for every linkage method. We see that 'weighted' and 'ward' are the two methods that yield the most similar results to our clusters. Despite 'weighted' linkage having the greatest similarity in case, we will use ward, as it generally yields better scores.

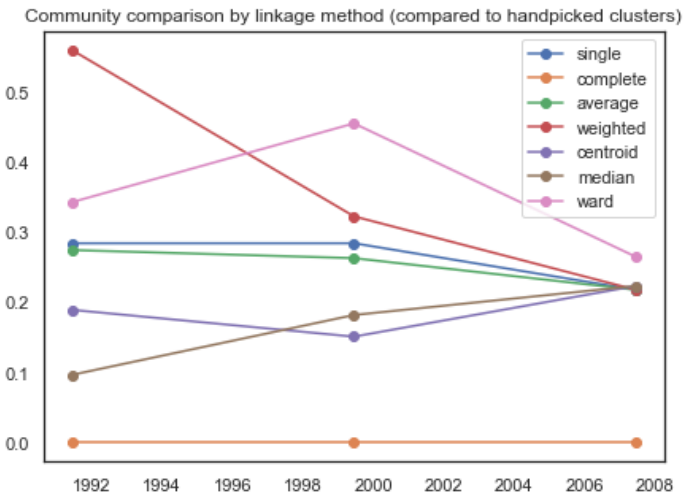

Figure 15: 4 clusters mutual information score, own elaboration.
Mutual information score (Y-axis) between the handpicked clusters and the ones generated using multiple of scipy's linkage methods for every survey (X-axis).

The clusters found for the 2008's survey using 'ward' as the linkage method and generating 4 clusters yield the following groups.

Group 1:

- All should be Treated Equally.
- Blacks Gotten less than deserved.
- Big problem not an equal chance.

- Blacks have it difficult to Succeed.
- Ensure Equal Opportunity to Succeed.
- Federal Spending- Poor/Poor People.
- Federal Spending- Public Schools.
- Federal Spending- Welfare Programs.
- Federal Spending- child care.
- Preference on blacks when hiring.
- Should Adjust View of Moral Behavior to Changes.
- Spending: Environment.
- Spending: Social Security.
- Tolerance on different moralities.
- Death penalty.

Group 2:

- Elections makes government pay attention.
- Better economy than past year.
- Presidency and congress splited.
- Government crooked.
- Federal wasted money.
- Hardworking Asian Americans.
- Hardworking for Hispanic-Americans.
- Hardworking for blacks.
- Hardworking for whites.

Group 3:

- Blacks must try harder.
- Blacks should Not be favored.
- Life unfair by default.
- Too far pushing equal rights.
- Should Worry less about How Equal People Are.

Group 4:

- Federal Spending- Dealing with Crime.
- Child attribute: curiosity vs good manners.
- Child attribute: considerate vs well-behaved.
- Child attribute: independence vs respect for elders.
- Gays in military.
- Guidance from religion.
- Newer lifestyles contribute to society.
- Protection to homosexuals.
- Should be More Emphasis on Traditional Values.
- Authority of the Bible.

For this instance, we observe that most of the topics are placed into 2 of the 4 groups. Group 1 contains topics of all sorts of nature, economics, the black situation, and opinion on the death penalty. In group 2 we find two different types of questions which we would expect together, opinions on different ethnicities and critiques of the government. Group 3 congregates topics about equality. Finally, group 4 joins religion topics with homosexuality and education which could be expected to be together.

We find that some topics such as the black opinion are scattered throughout multiple groups which intuitively we would expect together.

When looking at the mutual information score for different clusters created, see figure 16, we find that using fewer amount of clusters results in having low similarity while increasing the clusters yields increasingly better similarities.

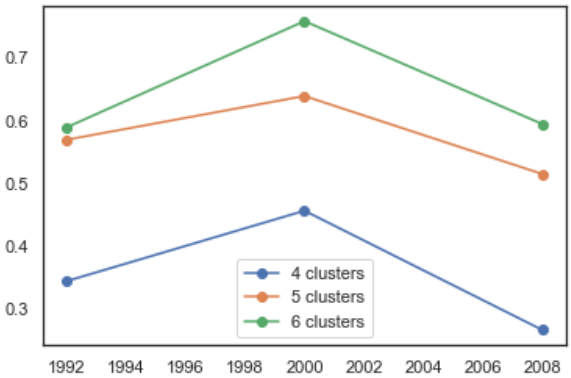

Figure 16: Mutual information score for different quantity of clusters, own elaboration. Mutual information score Y-axis) between the handpicked clusters of different sizes and the ones generated using 'ward' as linkage method for every survey (X-axis).

In fact, when comparing the generated clusters for a survey with the one generated for another survey we find the same behavior. Figures 17 and 18 display the similarity of the groups generated for 4 and 10 clusters respectively for all pairs of surveys. We see that with a greater amount of clusters the similarity is larger.

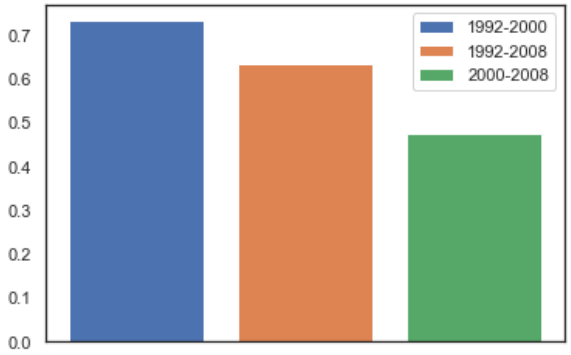

Figure 17: Similarity between generated clusters, 4 groups, own elaboration.
Mutual information score between the 4 clusters generated for the different surveys.

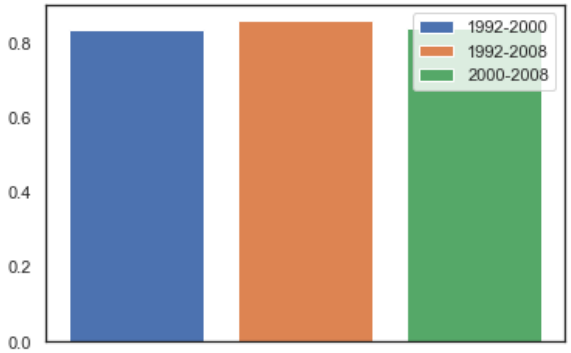

Figure 18: Similarity between generated clusters, 10 groups, own elaboration.
Mutual information score between the 10 clusters generated for the different surveys.

When trying to find a low amount of ideologies defined by topics these are defined loosely and we do not observe any pattern in how the ideologies evolve. When increasing the amount of topic-related ideologies the similarity between these ideologies remains similar throughout all surveys. These inform us that segmenting into groups with fewer topics, these have a close similarity. Remaining small groups of questions related to themselves.

4 Conclusions

We have seen that no matter which set of surveys is used, with their corresponding questions, the spaces generated by ISOMAP have all similar trends. We have observed that for all embeddings, ethnicity and political leaning are the single most relevant attributes in explaining the ideology of an individual. We have also seen how education and annual income remain the third and fourth most significant attributes. Social class on the other side has not presented such significance despite being highly related to income. Finally, age and gender joined the social class as the lesser attributes for defining ideology.

We have proved that all attributes are relevant to the final ideology of an individual. Being gender the one it approaches the most to have the population behave the same.

We have observed the tendency of opinion based on the population's political leaning, finding for all embeddings that the republican and democrat opinions are every time more divided. This leads us to conclude that the American population presents a polarization of opinion based on political orientation during the last 30 years.

The embeddings created for different surveys proved to be consistent in the results obtained and these results were not given by a specific configuration of ISOMAP as the same conclusion could be reached for different hyperparameters for the method.

We have created a network identifying clusters of topics forming an ideology. The obtained clusters resulted significantly different from each other and we have not been able to retrieve a pattern on the evolution of such ideologies. We have observed that small groups of questions remain related to one another throughout time.

Studying the clusters from the absolute value of the Pearson correlation have not yielded consistent ideologies nor has shown any evolution. Thus, other methods to construct the ideologies must be explored so negative correlations are not confused with positive ones. Then the study of their relationship might present more consistent results.

5 Temporal planning

The duration of the final degree project is stipulated in 18 ects equivalent to 25 hours per ects. Therefore, the duration of the project that we have foreseen is 450 hours. Starting the same day as the start of GEP, September 19, and ending a week before the start of the defenses of the TFG the project is to be able to prepare it correctly. Therefore, the duration of this project will reach 18 weeks. Therefore, the objective will be to work 25 hours per week during the indicated period.

5.1 Description of tasks

5.1.1 Project management (GP)

Under this label are grouped all the tasks specified during the GEP, as well as the meetings that will have to be carried out during the project. These tasks define the project and must be completed at the beginning, however, they will be carried out at the same time as the initial steps of the project itself.

The tasks included within GP are:

- GP.1: Scope of the project: Contextualize the framework in which the project is located, define its scope and justify its reasons and innovations.
- GP.2: Temporary planning: Define the objectives to be met during the project and plan the execution of all the tasks involved with the time available.
- GP.3: Budget and sustainability: Plan the material and human costs of the project, as well as analyze sustainability at various levels.
- GP.4 Documentation: Writing the project memory.
- GP.5 Meetings: Time devoted to discussion among group members about the work done and planning of the work to be done.
- GP.6 Defense. Planning and preparation of the defense of the project.

5.1.2 Data preparation (PD)

The objective of data preparation is to choose which of all the information we start with will be useful to us and prepare it so that it can be used correctly in the following stages of the project.

- PD.1 Study of the data set: Human study of the data set, study the questions that exist, what they are like, what format they have, and how they are distributed over time.
- PD.2 Data selection: Selection of the data that will be used for the study.
- PD.3 Data cleaning: Preparation of the data so that they can be used and the rest of the techniques can be applied to them.

5.1.3 Dimensionality reduction (RD)

For the reduction of dimensionality, a study of the techniques to be used, a technique selection exercise and its implementation will be necessary.

- RD.1 Study of the techniques: Study of the different dimensionality reduction techniques available, their requirements, and advantages.
- RD.2 Test and selection of the technique: Assessment and selection of the technique or techniques to be used.
- RD.3 Implementation of the technique: Implementation of the chosen techniques.
- RD.4 Study of the result: Study of results and relationships obtained on polarization.

5.1.4 Clustering (C)

In the same way as with the reduction of dimensionality, a previous study will be necessary for clustering in order to detect the possible ways of doing it. In this case we can perform clustering in various ways in the project and all of them will have to be studied.

- C.1 Study of the techniques: Study of the different grouping methods.

- C.2 Selection of techniques: Assessment of and selection of the techniques that will be used.
- C.3 Implementation of the techniques: Implementation of the chosen techniques
- C.4 Study of the results: Study of the groups obtained through clustering.

5.1.5 Results study (ER)

The objective of the study of results is to verify if the initial premise of the project is true, that society is polarizing. This does not contribute to a yes or no answer but requires a study by ideologies that can give varied results. Under this label, all the information obtained will be studied to reach a conclusion of the study.

- ER.1 Study of patterns: Study of the relationships obtained during the project.
- ER.2 Conclusion: Elaboration of a conclusion considering all the data.

5.2 Necessary resources

5.2.1 Human resources

The project has two differentiated roles that are separated into organization and control and its implementation:

- Director: Including both the director and co-director, they organize the planning of the project and direct the meetings for the correct development of the work.
- Engineer: Develop the project. In charge of implementing all the objectives.

5.2.2 Material resources

For the development of this project, three computers will be necessary. Two of the computers will be used to monitor the project by both directors, since they do not require excessive computing power they will be mid-range. On the part of the engineer, a medium-high-range computer with a processor capable of executing the written programs will be required. On the part of the graph, the choice is a high-end one to be able to explore deep learning options. The resulting computer has the following

specifications. AMD Ryzen 7 5800X 3.8GHz processor, GeForce RTX 3070 TI GPU, and 32GB of RAM.

5.2.3 Software resources

Visual Studio Code and all the libraries to be used are free.

5.3 Time estimate

Table 5 shows the table of tasks to be carried out, the estimated time for the resolution of each of them, the dependencies they have, and an assessment of the risk of lengthening that they have. A numerical estimate of the risk of the tasks will be provided in the following sections.

ID	Task	Time	Dependency	Risk
GP	Project management	155h		
GP.1	Scope of the project	25h		Low
GP.2	Temporary planning	10h	GP.1	Low
GP.3	Budget and sustainability	10h	GP.2	Low
GP.4	Documentation	70h		Low
GP.5	Meetings	8pm		Low
GP.6	Defense	20h	GP.4, ER.2	Low
PD	Data preparation	90h		
PD.1	Study of the data set	25h		Low
PD.2	Data selection	40h	PD.1	Low
PD.3	Data clearing	25h	PD.2	Low
RD	Dimensionality Reduction	80h		
RD.1	Study of techniques	20h	PD.3	Low
RD.2	Test and selection of the technique	25h	RD.1	Medium
RD.3	Implementation of the technique	15h	RD.2	Low
RD.4	Study of the result	20h	RD.3	Low
C	Clustering	75h		
C.1	Study of techniques	20h	RD.4	Low
C.2	Selection of techniques	15h	C.1	Low
C.3	Implementation of techniques	25h	C.2	Medium
C.4	Study of results	15h	C.3	Low
ER	Study Results	50h		
ER.1	Study of patterns	35h	C.4	Low
ER.2	Conclusion	15h	ER.1	Low
-	Total	450h		

Table 5: Time estimates and task dependency, own elaboration.

5.4 Gantt

Figure 19 shows the Gantt chart of work planning.

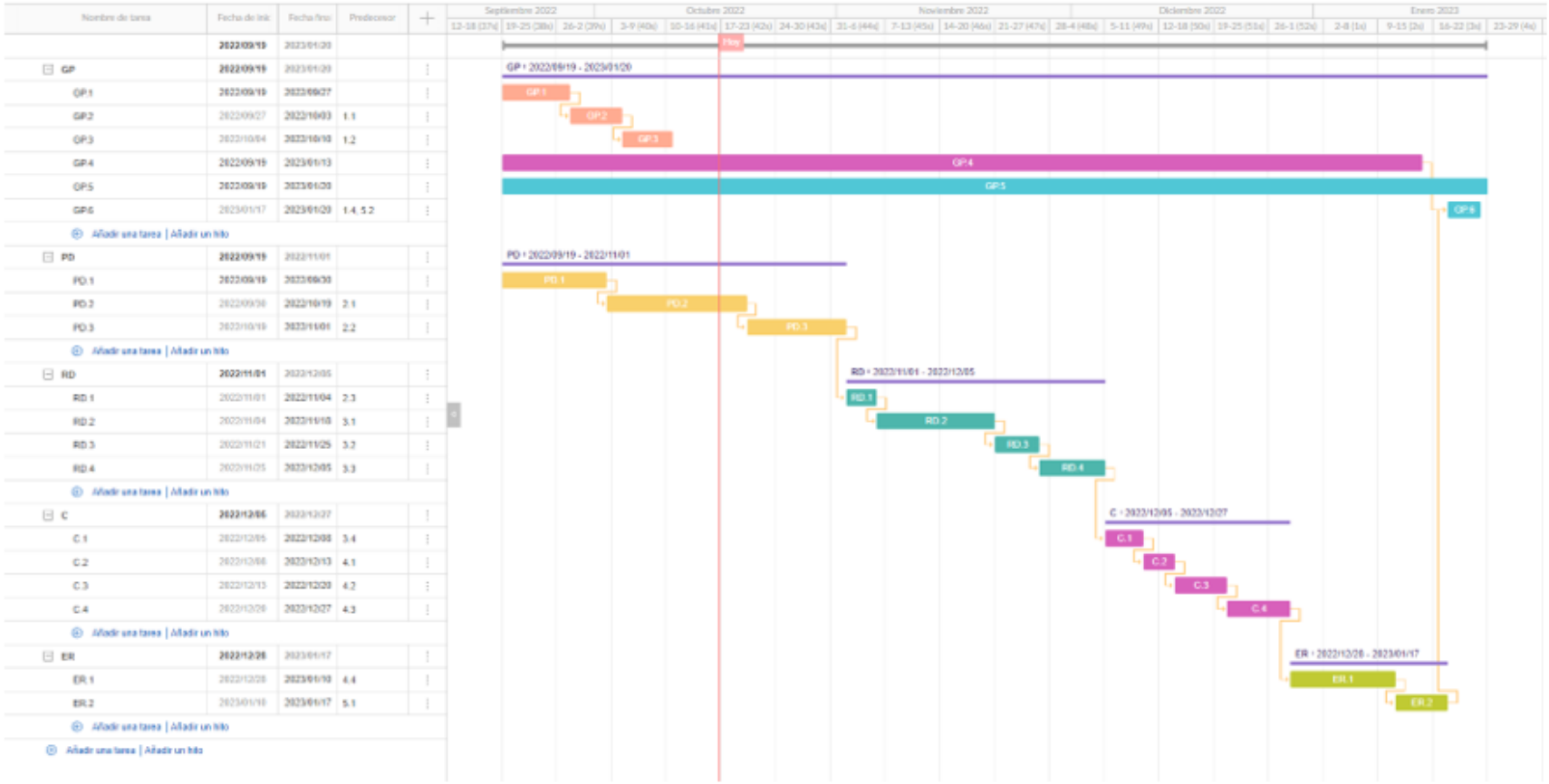

Figure 19: Gantt diagram, own elaboration.

5.5 Risk management

This section is intended to clarify the course of action if a task poses a setback.

In the project management section, the first three tasks (GP.1, GP.2, GP.3) have a deadline to be prepared, and therefore, in case of a setback, the course of action should be to dedicate more hours to it. The same happens for the defense (GP.6) by having a deadline that has not yet been closed but fixed.

The GP.5 task must be executed constantly and in case it seems that it is lagging, it is necessary to dedicate more time to bring it up to date.

The GP.6 meetings have been scheduled every week, an increase in the hours dedicated to them translates into making longer meetings in a specific case or an extra meeting under some circumstances, which will hardly affect the execution of the project.

Data preparation is extremely important and a delay in time will result in an increase in the weekly hours dedicated to the project or a delay in the following tasks. As it is such an important point, the time required for its completion has been extended in the initial estimate to ensure its completion on time.

The reduction of the dimensionality supposes a medium risk in terms of the test and selection of the technique (RD.2), again in a problem more time could be spent. However, if it is a significant setback, the most viable option would be to choose a good enough method, even if it is not ideal, and continue the project with that choice.

The study of clustering should not be a setback, you can try many methods and analyze the results for all of them. The possible hiccup here would be at the time of the implementation itself (C.3) but a reduction in the techniques used could mitigate the problem.

The study of results has no shortcuts, in case of any problem the only solution is to increase the time dedicated and use some of the time allocated for the preparation of the defense.

6 Budget

In this section, the economic factors concerning the project will be analyzed and its costs will be specified.

6.1 Personnel costs

As can be seen in Table 6, the prices per hour are instantiated depending on each role, based on the salaries observed in the GlassDoor platform [8] for the project manager and engineer.

Role	Price/ hour	Social security	Remuneration
Director (D)	30€	9€	39€
Co-director (CD)	30€	9€	39€
Engineer (I)	10€	3€	13€

Table 6: Personal cost, own elaboration.

ID	Task	Time	Actors	Gross cost	Cost
GP	Project management	155h	-	9200€	11.960€
GP.1	Scope of the project	25h	D,CD	1500€	1.950€
GP.2	Temporary planning	10am	D,CD	600€	780€
GP.3	Budget and sustainability	10h	D,CD	600€	780€
GP.4	Documentation	70h	D,CD,I	4900€	6370€
GP.5	Meetings	20h	D,CD,I	1400€	1.820€
GP.6	Defense in court	20h	I	200€	260€
PD	Data preparation	90h	-	900€	1.170€
PD.1	Study of the data set	25h	I	250€	325€
PD.2	Data selection	40h	I	400€	520€
PD.3	Data cleaning	25h	I	250€	325€
RD	Dimensionality Reduction	80h	-	800€	1.040€
RD.1	Study of techniques	20h	I	200€	260€
RD.2	Test and selection of the technique	25h	I	250€	325€
RD.3	Implementation of the technique	15h	I	150€	195€
RD.4	Study of the result	20h	I	200€	260€
C	Clustering	75h	-	750€	975€
C.1	Study of techniques	20h	I	200€	260€
C.2	Selection of techniques	15h	I	150€	195€
C.3	Implementation of techniques	25h	I	250€	325€
C.4	Study of results	15h	I	150€	195€
ER	Study Results	50h	-	500€	650€
ER.1	Study of patterns	35h	I	350€	455€
ER.2	Conclusion	15h	I	150€	195€
-	Total	450h	-	12150€	15.795€

Table 7: Time estimates and task dependency, own elaboration.

Table 7 shows the cost of carrying out each task. The gross cost has been calculated taking into account a deduction of 30%

6.2 Generic costs

6.2.1 Amortizations

Hardware

Taking 4 years for the amortization of the computers, as allowed by the Treasury, a dedication per day of 5 hours over the total of 450 hours of work, we obtain an amortization as shown in table 8.

Item	Quantity	Price	Amortization
Mid-end computer	2	750€	153.4€
High-end computer	1	2000€	204.5€
Total	3	3500€	357.9€

Table 8: Hardware amortizations, own elaboration.

Software

All the software used in the project is free and therefore its amortization is 0€.

6.2.2 Energy consumption

The energy cost has been considered taking into account that the energy cost of a computer is around 0.1 kWh. Considering the estimated hours of work for each role, we can estimate the number of computer hours used and with them its cost. This representation is shown in Table 9. The price has been considered considering the average price of the kWh at this time, 0.31€/kWh according to the Spanish electricity network [9].

Role	Hours of use	Price
Director (D)	135h	41.85€
Co-director (CD)	135h	41.85€
Programmer (I)	405h	125.55€
Total	675h	209.25€

Table 9: Price of electricity consumption in desktop computers, own elaboration.

6.2.3 Internet consumption

The internet rate is about 20€ per month, being 3 people for 5 months who will use it in different places. Internet spending will amount to 60€ per month, 300€ in the total project.

6.2.4 Total generic costs

Table 10 shows the total generic costs separated by category.

Concept	Cost
Amortizations	357.9€
Electricity consumption	209.25€
Internet consumption	300€
Total	766.15€

Table 10: Generic costs differentiated by category, own elaboration.

6.3 Contingencies

Concept	Cost	Contingency
Personal	15.795€	3.159€
Energy consumption	209.25€	41.85€
Hardware	357.9€	71.6€
Total	16,124.25€	3,272.45€

Table 11: Contingency margin, own elaboration.

To have a wide margin, a contingency cost of 20% has been chosen since, due to the nature of the investigation, we do not have a clear idea of the results that we will be obtaining. This cost will also apply to spending on hardware and electricity consumption, which are volatile on the date of completion of the project. Internet consumption is a fixed rate and therefore will not have variations on the previously mentioned price. The contingency costs can be seen in Table 11.

6.4 Unforseen

Finally, it is necessary to evaluate the cost that the different unforeseen events could cause. As mentioned in section 5 on time planning, an unforeseen event in a task will mean an increase in hours dedicated to it. Table 12 presents for each task the cost that it will entail and the result of the cost of the unforeseen event considering a probability that the task will be extended (Risk) and the additional burden of time that its failure would entail.

Unforeseen	Cost	Additional Workload (%)	Risk	Total
Project management	9200€	10%	10%	92€
Data preparation	900€	20%	10%	18€
Dimensionality reduction	800€	20%	30%	48€
Clustering	750€	10%	30%	22.5€
Study Results	500€	10%	15%	7.5€
Total	-	-	-	188€

Table 12: Cost of contingencies according to the tasks, own elaboration.

An unforeseen event related to hardware failure will mean the purchase of a similar computer. Unforeseen events due to hardware are represented in table 13.

Unforeseen	Cost	Risk	Total
Medium end computer 1	750€	5%	37.5€
Medium end computer 2	750€	5%	37.5€
Medium-high end computer	2000€	5%	100€
Total	3500€	-	175€

Table 13: Cost of hardware contingencies, own elaboration.

Table 14 shows the total costs due to unforeseen events.

Unforeseen	Total
Tasks	188€
Hardware	175€
Total	363€

Table 14: Total costs due to unforeseen events, own elaboration.

6.5 Total cost

Table 15 collects all the costs to generate a total.

Concept	Cost
Personnel	15.795€
Energy consumption	209.25€
Hardware	357.9€
Internet consumption	300€
Contingencies	3,272.45€
Contingencies	363€
Total	20297.6€

Table 15: Total cost of the project, own elaboration.

7 Sustainability

7.1 Environmental dimension

Regarding the environmental weight of the project, we have tried to reduce it as much as we could. First, taking advantage of the virtual resources at our disposal to hold meetings remotely and not contribute to the carbon footprint through transportation. Also, the use of the same computer to do the calculations instead of using remote servers with more power means a reduction in energy costs by not having to use both systems at the same time.

The study of polarization is currently done with a lot of human interaction, requiring the person to relate the topics individually. This project aims to automate this work which will result in less time in which the computer must be turned on and therefore less energy consumption.

7.2 Economic dimension

I believe that both material and human costs have been analyzed correctly, this information is detailed in section 6.

I find that this project, although it will result in a higher material cost than similar studies, will reduce all the necessary human resources by leaving all the drudgery to the machine. Economically, it means a decrease in spending, since humans are a more expensive resource for computing, as well as being slow.

7.3 Social dimension

On a social scale, it is perhaps where this project could be found most weak. Being a research work, it will not have a direct impact on society and aims to contribute to a change of mentality and an increase in self-awareness. This work is not essential for life, but its conclusions portray the modern mentality and I believe that a correct attitude is always important.

On a personal level, the project implies an awareness of the polarization of opinion, a subsequent introspection, and knowledge about observing my attitude toward seeing the entire spectrum.

This work will use methods rarely used and in a simpler way in similar studies. With the completion of this project, other articles may appear expanding on the conclusions of this and giving a more complete vision of the current vision of Western society.

8 References

[1] American National Election Studies. (2023, January 05). Retrieved September 19, 2022, from https://electionstudies.org/.

[2] Affective polarization in the American public - Northwestern University. (n.d.). Retrieved January 16, 2023, from https://www.ipr.northwestern.edu/documents/working-papers/2021/wp-21-27.pdf

[3] Marc J. Hetherington, Meri T. Long, Thomas J. Rudolph, Revisiting the Myth: New Evidence of a Polarized Electorate, Public Opinion Quarterly, Volume 80, Issue S1, 2016, Pages 321–350, https://doi.org/10.1093/poq/nfw003.

[4] Altman, N., Krzywinski, M. The curse(s) of dimensionality. Nat Methods 15, 399–400 (2018) https://rdcu.be/c1t02.

[5] Arce, C. and Gärling, T., 1989. Multidimensional scaling. Anuario de psicología/The UB Journal of psychology, pp.63-80.

[6] StuartReid. (2015, June 22). Dimensionality Reduction Techniques. Retrieved October 17, 2022, from http://www.turingfinance.com/artificial-intelligence-and-statistics-principal-component-analysis-and-self-organizing-maps/.

[7] What is hierarchical clustering? (n.d.). Retrieved October 17, 2022, from https://www.kdnuggets.com/2019/09/hierarchical-clustering.html .

[8] Company salaries | glassdoor. (n.d.). Retrieved September 17, 2022, from https://www.glassdoor.com/Salaries/index.htm

[9] Precio del KWH de Luz por Horas. (n.d.). Retrieved October 10, 2022, from https://tarifaluzhora.es/?tarifa=pcb&fecha=10%2F10%2F2022

[10] A. Rogier T. Donders, Geert J.M.G. van der Heijden, Theo Stijnen, Karel G.M. Moons. Review: A gentle introduction to imputation of missing values (2006). https://doi.org/10.1016/j.jclinepi.2006.01.014.

[11] Beretta, L., Santaniello, A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 16 (Suppl 3), 74 (2016). https://doi.org/10.1186/s12911-016-0318-z

[12] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. https://doi.org/10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.

[13] How to analyze anes data - american national election studies. (n.d.). Retrieved October 19, 2022, from https://electionstudies.org/wp-content/uploads/2018/05/HowToAnalyzeANESData.pdf

[14] Jcf2d, W. (n.d.). University of Virginia Library Research Data Services + Sciences. Retrieved January 14, 2023, from https://data.library.virginia.edu/getting-started-with-multiple-imputation-in-r/

[15] Pietryka, M., & MacIntosh, R. (2017). An Analysis of ANES Items and Their Use in the Construction of Political Knowledge Scales. Political Analysis, 21(4), 407-429. doi:10.1093/pan/mpt009

[16] U.S. Census Bureau quickfacts: United States. (n.d.). Retrieved January 2, 2023, from https://www.census.gov/quickfacts/fact/table/US/PST045221

[17] King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." American Political Science Review 95 (1): 49–69. https://gking.harvard.edu/files/gking/files/evil.pdf

[18] Hanson, M., & Checked, F. (2023, January 01). Average cost of college [2023]: Yearly tuition + expenses. Retrieved January 8, 2023, from https://educationdata.org/average-cost-of-college

[19] Crespo Márquez, A. (2022). The Curse of Dimensionality. In: Digital Maintenance Management. Springer Series in Reliability Engineering. Springer, Cham. https://doi.org/10.1007/978-3-030-97660-6_7

[20] Elmore, K. L., & Richman, M. B. (2001). Euclidean Distance as a Similarity Metric for Principal Component Analysis, Monthly Weather Review, 129(3), 540-549. Retrieved Jan 14, 2023, from https://doi.org/10.1175/1520-0493(2001)129<0540:EDAASM>2.0.CO;2

[21] Uniform manifold approximation and projection for dimension reduction. (n.d.). Retrieved November 2, 2022, from https://umap-learn.readthedocs.io/en/latest/

[22] SKLEARN.MANIFOLD.TSNE. (n.d.). Retrieved November 2, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[23] SKLEARN.DECOMPOSITION.PCA. (n.d.). Retrieved November 2, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[24] SKLEARN.MANIFOLD.MDS. (n.d.). Retrieved November 2, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html

[25] SKLEARN.MANIFOLD.ISOMAP. (n.d.). Retrieved November 2, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.Isomap.html

[26] Dong Y, Peng CY. Principled missing data methods for researchers. Springerplus. 2013 May 14;2(1):222. https:/doi.org/10.1186/2193-1801-2-222. PMID: 23853744; PMCID: PMC3701793.

[27] Syrte. (n.d.). Syrte/NDTEST: Multi-dimensional statistical test with python. Retrieved November 23, 2022, from https://github.com/syrte/ndtest/

[28] Yegelwel, W. (n.d.). Wil Yegelwel. Retrieved December 6, 2022, from https://wil.yegelwel.com/cluster-correlation-matrix/

[29] Leon Danon et al J. Stat. Mech. (2005) P09008 https://doi.org/10.48550/arXiv.cond-mat/0505245

[30] Scipy.cluster.hierarchy.fcluster. (n.d.). Retrieved December 10, 2022, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.fcluster.html

[31] Scipy.cluster.hierarchy.linkage. (n.d.). Retrieved December 10, 2022, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

9 Annex A: Chosen variables

This annex contains a list of all the questions we deemed as useful for the study. The variables are presented with their code and a short description of the question.

- 'VCF0310': "Interest in Elections".
- 'VCF0313': "Interest in public affairs".
- 'VCF0601': "Approve participation in protests".
- 'VCF0602': "Approve social disobedience".
- 'VCF0603': "Approve demonstrations".
- 'VCF0604': "Government waste Money".
- 'VCF0606': "Federal government waste money".
- 'VCF0608': "Government crooked".
- 'VCF0609': "Government care in public opinion".
- 'VCF0622': "Government pays attention to public opinion".
- 'VCF0623': "Parties makes government pay attention".
- 'VCF0624': "Elections makes government pay attention".
- 'VCF0625': "Congressmen attention to constituents".
- 'VCF0733': "Frequency of political discussions".
- 'VCF0675': "Trust in media".
- 'VCF0806': "Government Health Insurance".
- 'VCF0809': "Guaranteed Jobs and Income".
- 'VCF0811': "Urban unrest force-solve".
- 'VCF0814': "Civil rights pushed too fast".
- 'VCF0815': "Segregation vs desegregation".
- 'VCF0817': "Use school bus to integrate communities".
- 'VCF0818': "Ensure jobs/house for blacks".
- 'VCF0820': "Whites favor segregation".
- 'VCF0821': "Black favor desegregation".
- 'VCF0830': "Aid to Blacks".

- 'VCF0832': "Protect rights of accused".
- 'VCF0834': "Women should have an equal role".
- 'VCF0837': "Abortion allowed".
- 'VCF0839': "Government Services-Spending Scale".
- 'VCF0842': "Harder environmental regulations".
- 'VCF0843': "Defense Spending Scale".
- 'VCF0844': "Use of military force".
- 'VCF0847': "Guidance from religion".
- 'VCF0848': "Concern conventional war".
- 'VCF0848a': "Concern nuclear war".
- 'VCF0850': "Authority of the Bible".
- 'VCF0851': "Newer lifestyles contribute to society".
- 'VCF0852': "Should Adjust View of Moral Behavior to Changes".
- 'VCF0853': "Should be More Emphasis on Traditional Values".
- 'VCF0854': "Tolerance on different moralities".
- 'VCF0867a':"Preference on blacks when hiring".
- 'VCF0871': "Better economy than past year".
- 'VCF0872': "Better economy next year".
- 'VCF0876a':"Protection to homosexuals".
- 'VCF0877a':"Gays in military".
- 'VCF0879': "Allow more immigrants".
- 'VCF0886': "Federal Spending- Poor/Poor People".
- 'VCF0887': "Federal Spending- child care".
- 'VCF0888': "Federal Spending- Dealing with Crime".
- 'VCF0889': "Federal Spending- AIDS".
- 'VCF0890': "Federal Spending- Public Schools".
- 'VCF0891': "Federal Spending- Students".
- 'VCF0892': "Federal Spending- foreigners".
- 'VCF0893': "Federal Spending- homeless".

- 'VCF0894': "Federal Spending- Welfare Programs".
- 'VCF9003': "Opinion on evangelical groups".
- 'VCF9004': "Opinion on elderly".
- 'VCF9005': "Opinion on supreme court".
- 'VCF9013': "Ensure Equal Opportunity to Succeed".
- 'VCF9014': "too far pushing equal rights".
- 'VCF9015': "Big problem not equal chance".
- 'VCF9016': "Life unfair by default".
- 'VCF9017': "Should Worry less about How Equal People Are".
- 'VCF9018': "All should be Treated Equally".
- 'VCF9038': "Guaranteed equal Opportunity by govt".
- 'VCF9039': "Blacks have it difficult to Succeed".
- 'VCF9040': "Blacks should Not be favored".
- 'VCF9041': "Blacks must try harder".
- 'VCF9042': "Blacks Gotten less than deserved".
- 'VCF9043': "Opinion on school prayer".
- 'VCF9046': "Federal Spending- food stamps".
- 'VCF9047': "Spending: Environment".
- 'VCF9048': "Federal Spending- science".
- 'VCF9049': "Spending: Social Security".
- 'VCF9050': "Federal Spending- assist blacks".
- 'VCF9206': "Presidency and congress split control".
- 'VCF9223': "immigrants take jobs away".
- 'VCF9226': "unemployment increase".
- 'VCF9233': "Position on torture".
- 'VCF9237': "Position on the death penalty".
- 'VCF9238': "Should be harder to buy a Gun ".
- 'VCF9244': "Trust other people".
- 'VCF9245': "Satisfied with life".

- 'VCF9246': "Child attribute importance: curiosity vs good manners".
- 'VCF9248': "Child attribute importance: considerate vs well-behaved".
- 'VCF9249': "Child attribute importance: independence vs respect for elders".
- 'VCF9251': "Knowledge political issues".
- 'VCF9254': "Satisfied with democracy in U.S.".
- 'VCF9267': "Opinion on muslims".
- 'VCF9268': "Opinion on rich".
- 'VCF9269': "Opinion on christians".
- 'VCF9270': "Hardworking for whites".
- 'VCF9271': "Hardworking for blacks".
- 'VCF9272': "Hardworking for Hispanic-Americans".
- 'VCF9273': "Hardworking Asian Americans".
- 'VCF9274': "Black influence".

10 Annex B: Low dimensional representation of different sets of years

Out of the different sets of surveys to use we decided on the following.

Surveys 1992, 2004, 2016:

Figure 20 displays the distance between centers of each socio-demographic and political attribute for the set of surveys. We observe that despite the new respondents and questions the distances found between communities behave similarly. Ethnicity and political leaning are still the most far-between attributes. In this embedding, the tendency for political leaning presents polarization once again.

For the rest of the attributes, we observe that the apparent tendencies for the first embedding don't follow through into this.

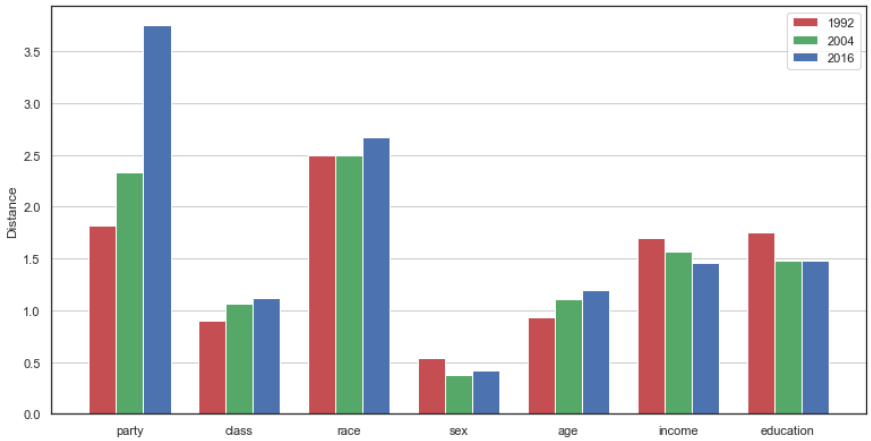

Figure 20: Distance between centroids of each political or socio-demographic feature, for data comprehended in years 1992, 2004, 2016, own elaboration.

Figure 21 displays the location of the opinion centers for the most relevant attributes of the 2016 survey. We find relations between centers similar to those found initially, where the higher education, white and high-income communities are closer to each other. Their counterparts are also close together. For political leaning and education, we once again find them to be nearly orthogonal between them.

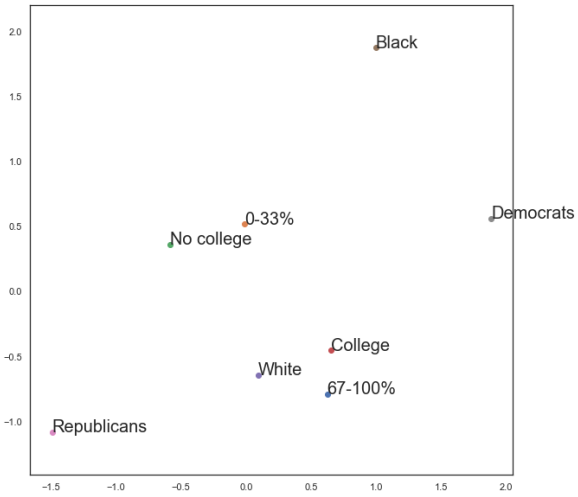

Figure 21: Location of most significant centers for 2016 survey in the set 1992, 2004, 2016, own elaboration.

Surveys 1992, 2000, 2008, 2016:

Figure 22 displays the distances found for the embedding with the new set of surveys.

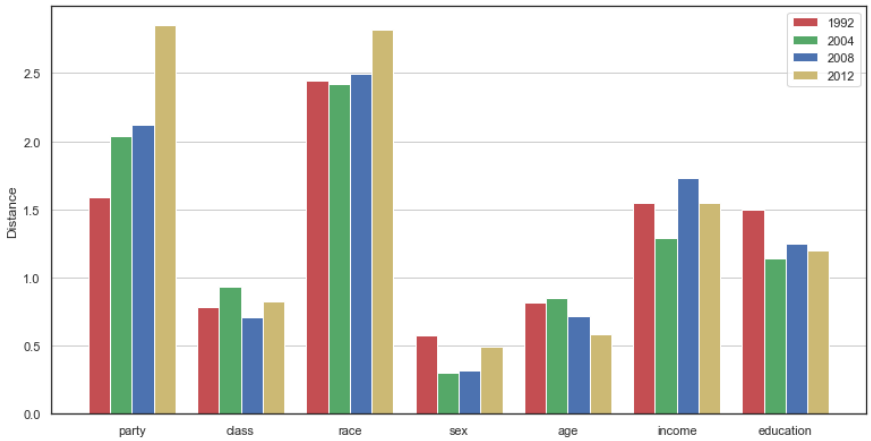

Figure 22: Distance between centroids of each political or socio-demographic feature for data comprehended in years 1992, 2000, 2008, 2016, own elaboration.

Figure 23 displays the location of the opinion centers for the most relevant attributes.

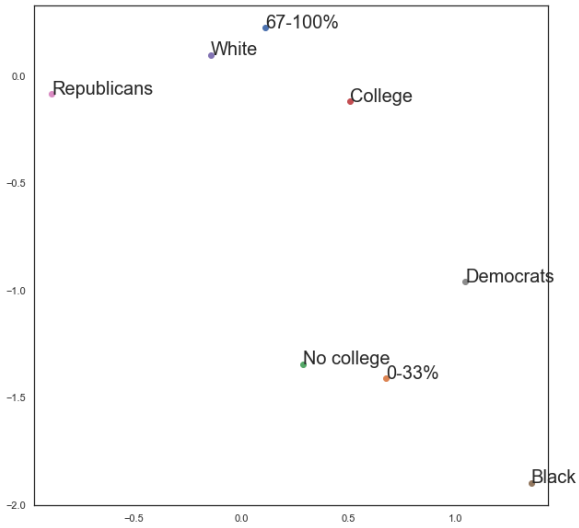

Figure 23: Location of most significant centers for 2016 survey in the set 1992, 2000, 2008, 2016, own elaboration.

Surveys 1992, 2000, 2004, 2008, 2012, 2016

Figure 24 displays the distances found for the embedding with the new set of surveys.

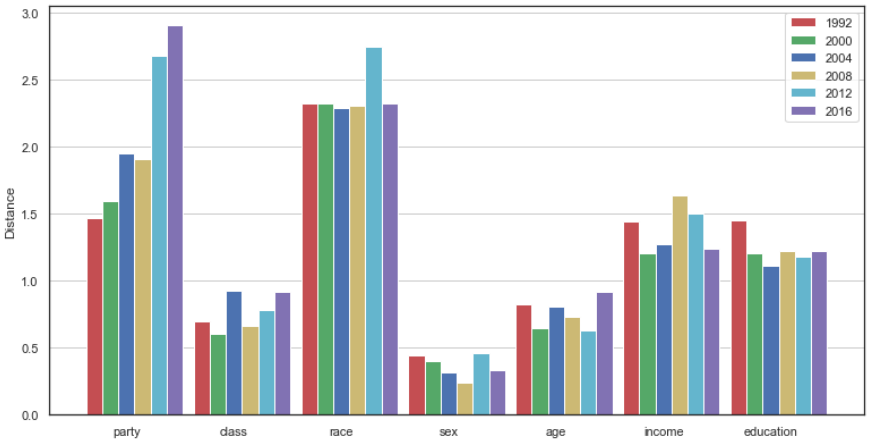

Figure 24: Distance between centroids of each political or socio-demographic feature, for data comprehended in years 1992, 2000, 2004, 2008, 2012, 2016, own elaboration.

Figure 25 displays the location of the opinion centers for the most relevant attributes.

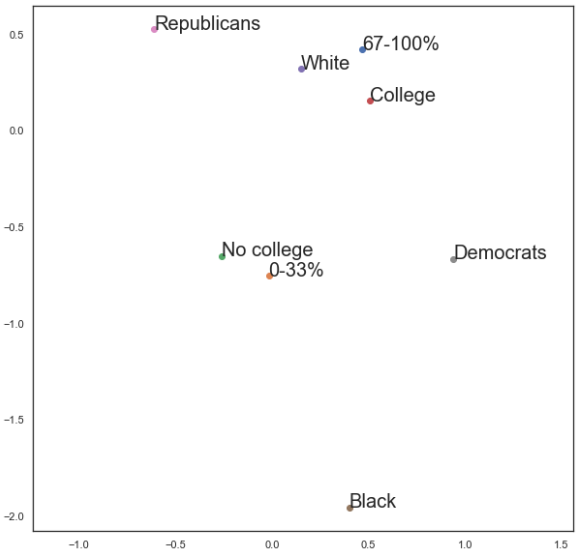

Figure 25: Location of most significant centers for 2016 survey in the set 1992, 2000, 2004, 2008, 2012, 2016, own elaboration.

Conclusion from figures

All different embeddings behave in similar manners. Ethnicity and political leaning have proved to be the most significant attributes for defining the individual's ideology. Gender remained the attribute with less difference between its communities.

Political leaning and the level of education positions in the embedding present to be orthogonal in the space.

From all the attributes only political leaning presents an increasing distance between its centers. This increase in polarization resulted minimal between 2004 and 2008. Then in 2012 increased greatly. From what we can observe an increase in polarization through time.

11 Annex C: Effect of ISOMAP hyperparameters

This annex will study the effect of the K hyperparameter on the results obtained to see if the conclusions are robust for different values. To do these we will see if the results obtained for K 5 and 20 are consistent with those found for K = 10.

K=5

Figure 26 displays the distances between centers of opinions for the set of surveys 1992, 2000, and 2008.

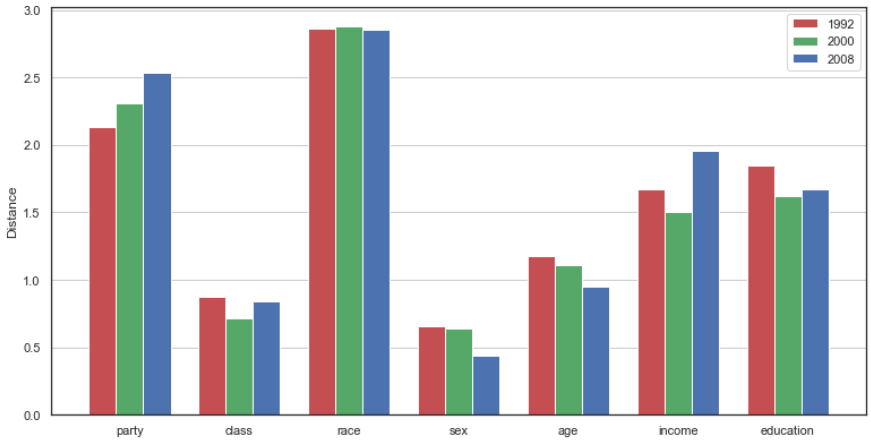

Figure 26: Distances between centers of opinion for K = 5, own elaboration.

Table 16 contains the results of the null model. The table contains the results for both communities in every attribute and every survey separately. We observe that for every community the actual distance is further than the distance of the 0.05 values. We can reject the null hypothesis thus we prove the alternative hypothesis that the data does not follow a random distribution.

	1992		2000		2008	
	distance	p-distance	distance	p-distance	distance	p-distance
Democrats	0.92	0.11	1.0	0.13	0.82	0.09
Republicans	1.21	0.15	1.3	0.17	1.71	0.18
Middle	0.45	0.13	0.32	0.13	0.47	0.14
Working	0.42	0.12	0.39	0.16	0.37	0.11
White	0.4	0.05	0.37	0.06	0.94	0.1
Black	2.46	0.32	2.51	0.38	1.91	0.21
Female	0.3	0.11	0.28	0.13	0.19	0.1
Male	0.35	0.13	0.36	0.16	0.25	0.14
17-34	0.56	0.14	0.63	0.2	0.52	0.16
55+	0.62	0.15	0.49	0.16	0.43	0.13
0-33%	0.87	0.15	0.66	0.16	0.71	0.12
67-100%	0.81	0.15	0.85	0.2	1.25	0.21
No college	0.87	0.11	0.99	0.17	0.87	0.12
College	0.98	0.12	0.63	0.11	0.8	0.11

Table 16: Null model distances for K=5, own elaboration.
The distance column represents the distance between the real center and the center of all the random iterations performed. the p-distance column represents the distance from the random iterations center which comprehends 95% of the results.

Table 17 contains the p-values for the 2-sample Kolmogorov-Smirnov test. All p-values are smaller than 0.05. Thus all communities are drawn following different distributions compared to their counterpart.

Year	Party	Class	Race	Gender	Age	Income	Education
1992	2.39e-34	3.19e-11	2.87e-47	1.65e-06	1.76e-12	9.22e-26	8.83e-40
2000	8.57e-35	3.61e-06	5e-36	2.64e-06	2.86e-07	3.73e-16	8.67e-23
2008	7.15e-49	3.95e-11	2.03e-73	3.83e-05	6.82e-08	2.34e-27	8.69e-38

Table 17: Kolmogorov-Smirnov test for every attribute for K=5, own elaboration.

K=20

Figure 27 displays the distances between centers of opinions for the set of surveys 1992, 2000, and 2008.

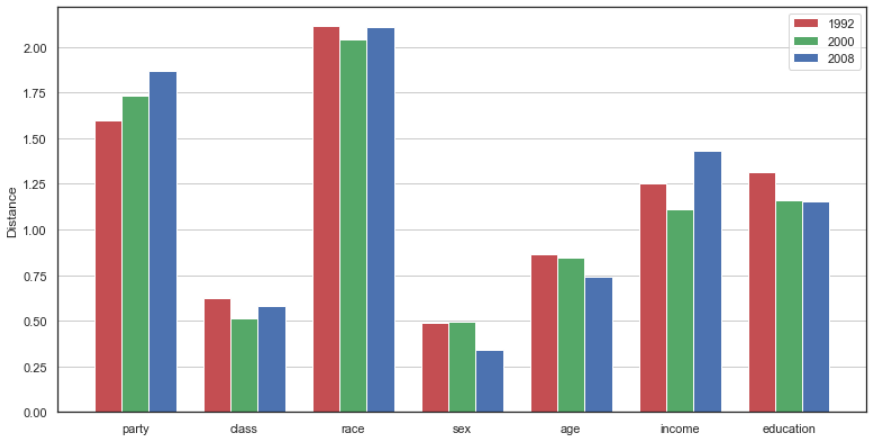

Figure 27: Distances between centers of opinion for K = 20, own elaboration.

Table 18 contains the results of the null model. The table contains the results for both communities in every attribute and every survey separately. We observe that for every community the actual distance is further than the distance of the 0.05 values. We can reject the null hypothesis thus we prove the alternative hypothesis that the data does not follow a random distribution.

	1992		2000		2008	
	distance	p-distance	distance	p-distance	distance	p-distance
Democrats	0.69	0.08	0.75	0.09	0.6	0.06
Republicans	0.91	0.1	0.99	0.12	1.26	0.13
Middle	0.32	0.09	0.23	0.1	0.33	0.1
Working	0.3	0.08	0.28	0.11	0.25	0.08
White	0.3	0.04	0.26	0.04	0.7	0.07
Black	1.81	0.22	1.78	0.27	1.41	0.15
Female	0.23	0.08	0.22	0.08	0.15	0.07
Male	0.26	0.09	0.28	0.11	0.19	0.09
17-34	0.41	0.1	0.47	0.14	0.41	0.12
55+	0.45	0.11	0.37	0.11	0.33	0.1
0-33%	0.64	0.11	0.48	0.11	0.52	0.08
67-100%	0.6	0.1	0.62	0.14	0.91	0.15
No college	0.62	0.08	0.71	0.12	0.6	0.09
College	0.69	0.09	0.45	0.08	0.55	0.08

Table 18: Null model distances for K=20, own elaboration.
The distance column represents the distance between the real center and the center of all the random iterations performed. the p-distance column represents the distance from the random iterations center which comprehends 95% of the results.

Table 19 contains the p-values for the 2-sample Kolmogorov-Smirnov test. All p-values are smaller than 0.05. Thus all communities are drawn following different distributions compared to their counterpart.

Year	Party	Class	Race	Gender	Age	Income	Education
1992	1.85e-37	6.25e-13	1.04e-45	5.69e-08	8.29e-14	2.02e-33	9.19e-43
2000	4.44e-34	3.99e-08	3.24e-35	5.07e-06	3.97e-09	2.42e-18	2.05e-25
2008	2.19e-52	6.27e-10	2.89e-81	2.2e-05	4.09e-11	1.42e-29	5.63e-37

Table 19: Kolmogorov-Smirnov test for every attribute for K = 20, own elaboration.

Conclusions from the results

ISOMAP embedding results consistent with the results obtained when changing its K. Both values of 5 and 20 had very similar hierarchy of distances between different attributes. The distances between partisanships follow the same increasing tendency ecnountered for K = 10.

The null model and Kolmogorov-Smirnov tests yielded the same results. Being all distributions non-random and all attributes being significant in the final ideology of an individual.

12 Annex D: Handpicked clusters

This annex displays the clusters of topics we selected as similar. They contain all 96 questions we deemed as useful and for each set of 96 questions these were grouped into 4, 5, and 6 clusters.

4 Clusters

Category 1:

- Federal Spending Poor/Poor People
- Federal Spending child care
- Federal Spending Dealing with Crime
- Federal Spending AIDS
- Federal Spending Public Schools
- Federal Spending Students
- Federal Spending foreigners
- Federal Spending homeless
- Federal Spending Welfare Programs
- Federal Spending food stamps
- Federal Spending science
- Federal Spending assist blacks
- Government Services-Spending Scale
- Defense Spending Scale
- Spending: Environment
- Spending: Social Security
- Federal wasted money
- Government Health Insurance

Category 2:

- Blacks have it difficult to Succeed
- Blacks should Not be favored

- Blacks must try harder
- Blacks Gotten less than deserved
- Black influence
- Segregation vs desegregation
- Whites favor segregation
- Black favor desegregation
- Ensure jobs/house for blacks
- Aid to Blacks
- Preference on blacks when hiring
- Opinion on evangelical groups
- Opinion on elderly
- Opinion on supreme court
- Opinion on muslims
- Opinion on rich
- Opinion on christians
- Hardworking for whites
- Hardworking for blacks
- Hardworking for Hispanic-Americans
- Hardworking Asian Americans
- Child attribute: curiosity vs good manners
- Child attribute: considerate vs well-behaved
- Child attribute: independence vs respect for elders
- Guidance from religion
- Authority of the Bible
- School prayer
- Protection to homosexuals
- Gays in military
- Allow more immigrants
- Immigrants take jobs away

- Trust other people
- Satisfied with life

Category 3:

- Ensure Equal Opportunity to Succeed
- Too far pushing equal rights
- Big problem not equal chance
- Life unfair by default
- Should Worry less about How Equal People Are
- All should be Treated Equally
- Guaranteed equal Opportunity by govt
- Use school bus to integrate
- Protect rights of accused
- Women equal role
- Abortion allowed
- Civil rights pushed too fast
- Tolerance on different moralities
- Guaranteed Jobs and Income
- Should Adjust View of Moral Behavior to Changes
- Should be More Emphasis on Traditional Values
- Torture
- Death penalty
- Should be harder to buy a Gun
- Approve participation in protests
- Approve demonstrations
- Approve civil disobedience

Category 4:

- Government Waste Money
- Government crooked
- Government care in public opinion

- Government pays attention to public opinion
- Parties makes government pay attention
- Elections makes government pay attention
- Congressmen attention to constituents
- Better economy than past year
- Better economy next year
- Knowledge political issues
- Satisfied with democracy in U.S.
- Interest in Elections
- Interest in public affairs
- Frequency of political discussions
- Trust in media
- Presidency and congress splited
- Use of military force
- Concern conventional war
- Concern nuclear war
- Urban unrest force-solve
- Harder environmental regulations
- Newer lifestyles contribute to society
- Unemployment increase

5 Clusters

Category 1:

- Federal Spending Poor/Poor People
- Federal Spending child care
- Federal Spending Dealing with Crime
- Federal Spending AIDS
- Federal Spending Public Schools
- Federal Spending Students
- Federal Spending foreigners
- Federal Spending homeless
- Federal Spending Welfare Programs
- Federal Spending food stamps
- Federal Spending science
- Federal Spending assist blacks
- Government Services-Spending Scale
- Defense Spending Scale
- Spending: Environment
- Spending: Social Security
- Federal wasted money
- Government Health Insurance

Category 2:

- Blacks have it difficult to Succeed
- Blacks should Not be favored
- Blacks must try harder
- Blacks Gotten less than deserved
- Black influence
- Black influence
- Segregation vs desegregation
- Whites favor segregation

- Black favor desegregation
- Ensure jobs/house for blacks
- Aid to Blacks
- Preference on blacks when hiring
- Opinion on muslims
- Hardworking for whites
- Hardworking for blacks
- Hardworking for Hispanic-Americans
- Hardworking Asian Americans

Category 3:

- Opinion on Evangelical group
- Opinion on elderly
- Opinion on supreme court
- Opinion on rich
- Opinion on christian
- Child attribute: curiosity vs good manner
- Child attribute: considerate vs well-behaved
- Child attribute: independence vs respect for elder
- Guidance from religion
- Authority of the Bible
- School prayer
- Protection to homosexual
- Gays in military
- Allow more immigrant
- Immigrants take jobs away
- Trust other people

Category 4:

- Ensure Equal Opportunity to Succeed
- Too far pushing equal rights

- Big problem not equal chance
- Life unfair by default
- Should Worry less about How Equal People Are
- All should be Treated Equally
- Guaranteed equal Opportunity by govt
- Use school bus to integrate
- Protect rights of accused
- Women equal role
- Abortion allowed
- Civil rights pushed too fast
- Tolerance on different moralities
- Guaranteed Jobs and Income
- Should Adjust View of Moral Behavior to Changes
- Should be More Emphasis on Traditional Values
- Torture
- Death penalty
- Should be harder to buy a Gun
- Approve participation in protests
- Approve demonstrations
- Approve civil disobedience
- Government Waste Money

Category 5:

- Government crooked
- Government care in public opinion
- Government pays attention to public opinion
- Parties makes government pay attention
- Elections makes government pay attention
- Congressmen attention to constituents
- Better economy than past year

- Better economy next year
- knowledge political issues
- Satisfied with democracy in U.S.
- Interest in Elections
- Interest in public affairs
- Frequency of political discussions
- Trust in media
- Presidency and congress splited
- Use of military force
- Concern conventional war
- Concern nuclear war
- Urban unrest force-solve
- Harder environmental regulations
- Newer lifestyles contribute to society
- Unemployment increase
- Satisfied with life

6 Clusters

Category 1:

- Federal Spending- Poor/Poor People
- Federal Spending- child care
- Federal Spending- Dealing with Crime
- Federal Spending- AIDS
- Federal Spending- Public Schools
- Federal Spending- Students
- Federal Spending- foreigners
- Federal Spending- homeless
- Federal Spending- Welfare Programs
- Federal Spending- food stamps
- Federal Spending- science
- Federal Spending- assist blacks
- Government Services-Spending Scale
- Defense Spending Scale
- Spending: Environment
- Spending: Social Security
- Federal wasted money
- Government Health Insurance

Category 2:

- Blacks have it difficult to Succeed
- Blacks should Not be favored
- Blacks must try harder
- Blacks Gotten less than deserved
- Black influence
- Segregation vs desegregation
- Whites favor segregation
- Black favor desegregation

- Ensure jobs/house for blacks
- Aid to Blacks
- Preference on blacks when hiring
- Opinion on muslims
- Hardworking for whites
- Hardworking for blacks
- Hardworking for Hispanic-Americans
- Hardworking Asian Americans

Category 3:

- Opinion on Evangelical groups
- Opinion on elderly
- Opinion on supreme court
- Opinion on rich
- Opinion on christians
- Child attribute: curiosity vs good manners
- Child attribute: considerate vs well-behaved
- Child attribute: independence vs respect for elders
- Guidance from religion
- Authority of the Bible
- School prayer
- Protection to homosexuals
- Gays in military
- Allow more immigrants
- Immigrants take jobs away
- Trust other people

Category 4:

- Ensure Equal Opportunity to Succeed
- Too far pushing equal rights
- Big problem not equal chance

- Life unfair by default
- Should Worry less about How Equal People Are
- All should be Treated Equally
- Guaranteed equal Opportunity by govt
- Use school bus to integrate
- Protect rights of accused
- Women equal role
- Abortion allowed
- Civil rights pushed too fast
- Tolerance on different moralities
- Guaranteed Jobs and Income
- Should Adjust View of Moral Behavior to Changes
- Should be More Emphasis on Traditional Values
- Approve participation in protests
- Approve demonstrations
- Approve civil disobedience

Category 5:

- Government Waste Money
- Government crooked
- Government care in public opinion
- Government pays attention to public opinion
- Parties makes government pay attention
- Elections makes government pay attention
- Congressmen attention to constituents
- Better economy than past year
- Better economy next year
- knowledge political issues
- Satisfied with democracy in U.S.
- Interest in Elections

- Interest in public affairs
- Frequency of political discussions
- Trust in media
- Presidency and congress splited
- Harder environmental regulations
- Newer lifestyles contribute to society
- Unemployment increase
- Satisfied with life

Category 6:

- Urban unrest force-solve
- Use of military force
- Concern conventional war
- Concern nuclear war
- Torture
- Death penalty
- Should be harder to buy a Gun