ENS 491-492 – Graduation Project
Final Report

Project Title: Detecting Most Representative Polling Units in Local Elections

Group Members:

Ege Kaan Özalp

Selim Gül

Cansu Lenger

Supervisor(s): Professor Burak Kocuk

Date: 26.05.2024

Sabancı Üniversitesi

Contents

1. EXECUTIVE SUMMARY

The main aim of this paper is to provide a solution to the issue of the absence of a model that would accurately and affordably find the predetermined number of neighborhoods in a given city of Türkiye such that analyzing solely the results of these selected neighborhoods would provide a highly accurate overview for the outcomes of the future elections. This project addresses this issue by implementing two major methods:

- Data scraping for collecting the past election results (i.e., the results of elections in 2014 and 2019),
- Creating a Mixed-Integer Program to detect the most representative neighborhoods and their weight coefficients for a given city using the scraped data.

In this study, it is shown that creating an optimization model for selecting the most representative neighborhoods provides accurate predictions for the future elections with the weighted absolute error value being less than 10% for each of the test cases which proves the overall high accuracy of the model. To further improve this model, it is discussed to implement statistical methods and create new parameters for socio-political changes and migration options.

2. PROBLEM STATEMENT

2.1. Motivation of the Study

The major motivation of this study is formed by the urge to address three crucial problems in current pre-election polling methods: the excessive cost of polling individuals prior the elections, the lack of accuracy in current methods and the absence of a model that finds the most representative polling units in Türkiye. The need for vast amount of data for creating an estimation model is understandable – especially if the pollster aims to create a machine learning, deep learning, or a neural network model. However, creating a random sample and gathering this vast data by polling the voters is highly costly. One concrete example would be the costs of pre-election polling in the United States of America. According to the firm OpenSecrets (2024) – which focuses on “tracking the flow of money in American politics and providing the data and analysis” – the total amount of money spent on pre-election polling is $1,043,059 which is shared among companies

American Polling, Guidant Polling & Strategy, National Research & Polling Group, Public Policy Polling and SEA Polling & Strategic Design.

In addition to the fact that polling being costly, the accuracy of these polls suffers greatly from the current approach of focusing on polling individuals. Pilnacek et al. (2021, p. 1) emphasizes that current conditions for a high-quality polling consists of 3 points: "representative sampling", "the correct model for turnout probabilities" and "compliance between stated vote intention and the actual vote". The major issue in these conditions is that the third point assumes that the people who participated in the poll will comply to their answer during the actual elections which is not the case since voters are increasingly postponing their decisions to a time closer to the elections (Pilnacek et al., 2021, p. 1-2).

Furthermore, there are significant problems on pre-election polls due to not only the participants but also the pollsters – especially in Türkiye. According to Aydaş et al. (2022, p. 102), many election polls in Türkiye significantly deviate from the scientific standards, particularly in terms of representation, which are considered the gold standard in survey methodology literature. Consequently, their predictions regarding election outcomes tend to be inaccurate.

Considering these issues, it is decided to create a model that accurately provides the most representative polling units in a city of Türkiye by focusing on the historical trend of the neighborhoods rather than solely focusing on the answers of individuals themselves – since analyzing only the individuals yields a highly unpredictable and questionably accurate results – to gather the accurate overview of the future elections in this study. By doing so, it is aimed to decrease the cost by merely focusing on a predetermined amount of neighborhoods that the model suggests and increase the accuracy of pre-election polls by not only focusing on the raw data in polls but rather the weighted values of the polls with weight values provided by the model.

2.2. Objectives/Tasks

The objective of this study is to create a model that finds a predetermined (i.e., user-defined) number of neighborhoods – with their corresponding weights – in a given city of Türkiye such that conducting polls in these neighborhoods would yield a highly accurate overview of the future elections that are conducted in this city. To achieve this, there are three major tasks:

- Scraping of election data of 2014 and 2019 from YSK to feed the data into the model,
- Creating a Mixed-Integer Program that finds the user-defined number of neighborhoods – with their weights – such that it maximizes the accuracy function.
- Testing the outcomes of the model using the results of 2024 metropolitan municipality elections.

Here, it is mentioned that the model maximizes an accuracy function. To further explain this part, the accuracy function itself is in fact the Euclidean distance function in which the data points are represented as a point in m-dimensional space where "m" represents the number of political parties that participated to the elections of the given city. The coordinates of each data point are represented by the vote percentages of each of the political parties. To maximize the accuracy function, it is required to minimize this Euclidean distance function. To make the logic of the accuracy function (i.e., the objective function of the model) clearer, analyze the below example:

Assume that there are only two political parties in the given city. The data point R represents the final cumulative percentage result of both political parties P_1 and P_2 in the given city in year T. The data point X represents the final percentage result of the P_1 and P_2 in the neighborhood X, the data point Y represents the final percentage result of the P_1 and P_2 in the neighborhood Y and similarly, the data point Z represents the result in the neighborhood Z.

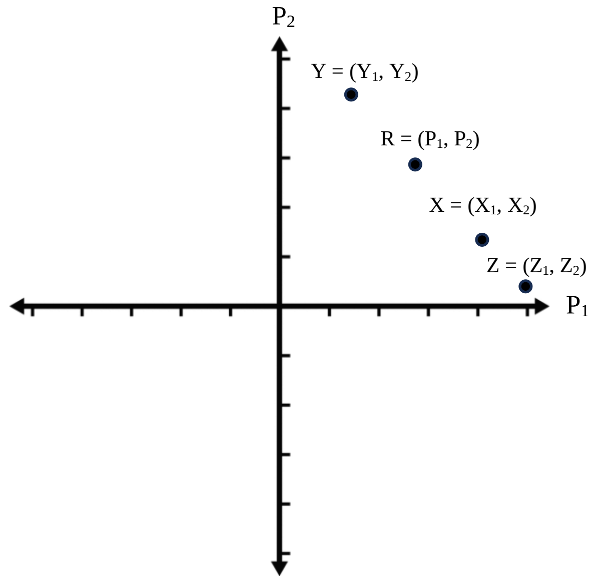

Figure 2.2.1.: The representation of the election results for the objective function.

Given the data points of the results and the user-defined number for neighborhood selection, the model aims to find a linear combination of the results of the selected neighborhoods such that this combination would represent the data point C which is the closest possible combination point to the data point R. The coordinates of the point C are shown as (C_1, C_2) where C_1 equals to "$w_1X_1 + w_2Y_1 + w_3Z_1$" and C_2 equals to "$w_1X_2 + w_2Y_2 + w_3Z_2$".

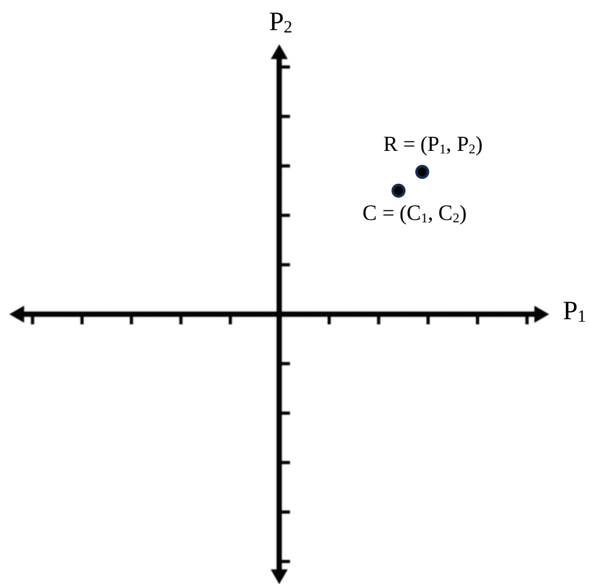

Figure 2.2.2.: An example representation of the optimal linear combination point for the neighborhoods.

Hence, by selecting the user-defined number of neighborhoods and assigning each one of them a weight, the model aims to minimize the Euclidean distance function. It is crucial to note that this example focuses on a single year. For multiple years, the model aims to find neighborhoods and weights that minimizes the sum of the result of Euclidean distance function of each year.

2.3. Realistic Constraints

Now that the accuracy function of the model is defined, there are also several crucial factors that the model should obey:

- There is a need for binary variables to keep track of the selected neighborhoods and continuous variables to keep track of the assigned weights for each of the neighborhoods. Weight values are decided to be between 0 and 1 (inclusive) for standardization.
- The total number of binary variables that gets the result “1” cannot exceed the number of neighborhoods that the user would prefer to select.
- The weight of a given neighborhood should be less than or equal to the value of the binary variable that is assigned to it since if it is not selected, it should not possess a weight value.
- The sum of weights of each neighborhood should be equal to 1.

3. METHODOLOGY

3.1. Data Scraping

To gather the necessary data for this study, we developed a couple of scripts using Python and mainly the Selenium library, along with other Python libraries such as *json, pandas, numpy* and so forth. The scripts listed below automate the processes of navigating the Supreme Election Council of Turkey's (YSK) official website to download election data from the 2014 and 2019 local elections, cleaning the data, and formatting them for the needs of our project.

3.1.1. Data Downloading

Below is an outline of the data scraping process:

- **Setup and Configuration:**
    - The script named *scraper.py* sets up download preferences to automatically save files to a specified directory without prompting the user.
    - Initializes the Chrome WebDriver with the necessary options to handle file downloads and browser configurations.

- **Navigation and Selection:**
    - The script navigates to the YSK election results page.
    - It prompts the user to select the election year and type of election. The options include various local elections such as mayoral and municipal council elections. The users give input “1” for 2014 elections, “2” for 2019 elections on the initial run.
    - The script uses Selenium's WebDriverWait to ensure all elements are loaded and clickable before interaction.

- **Data Download:**
    - For each city, the script iterates through all districts (ilçe) of each city and selects each one to download the corresponding election results.
    - After selecting a district, the script triggers the download of the election results in an Excel format.
    - It waits for the download to complete, then renames the file based on the city and district to maintain organization.
    - The renamed file is moved to a designated directory structured by year and type of election.

- **File Handling:**
    - The script includes functions to sanitize filenames to ensure they are valid for the file system.
    - It waits for downloads to complete and verifies the presence of the downloaded file before proceeding to the next district.

- **Error Handling and Reattempts:**
    - The script includes mechanisms to handle invalid inputs and retry failed operations, ensuring robust data scraping.
    - If user inputs a wrong number, it asks the user to enter a correct number, displaying the options.
    - If the website does not fully load before clicking on an element, the script waits for the elements to be clickable and resumes from the city or district where it last downloaded the data of.

This automated data scraping method efficiently collects the historical election data required for the study, minimizing manual effort and errors.

3.1.2. Data Cleansing and Aggregation

Following the data scraping process, we developed a script to aggregate and cleanse the election data. This script processes the downloaded Excel files, consolidates the results by city, and performs necessary data cleansing. Below is an outline of the data processing workflow:

- **Setup and Configuration:**
    - The script named *sum_up_by_city.py* identifies all Excel files in the specified directory containing the election results.
    - Extracts unique city names from the filenames to categorize the data accordingly.

- **Data Aggregation:**
    - For each city, the script filters the relevant files and initializes a DataFrame to store the aggregated results.
    - Each file is read into a DataFrame, assuming the header row is specified (e.g., row 10). The script dynamically identifies the columns representing political parties.
    - For all the excel files, the actual data that we are interested in, starts at row 10. The data above it is not relevant with our project, therefore they can be dismissed.

- **Grouping and Summing:**
    - The data is grouped by district (ilçe) and neighborhood (mahalle/köy) to sum the vote counts for each political party.
    - The intermediate results from each file are concatenated into a single DataFrame for the city.

- **Final Aggregation:**
    - The script performs a final grouping and summing operation to consolidate the vote counts across all files for each city.
    - The aggregated results are saved into a new Excel file prefixed with an asterisk (*) to denote the processed output (e.g., *Summed_Up_Results_CityName.xlsx).

- **Output and Verification:**

- The output files are saved in the specified directory, maintaining the structure, and ensuring the data is ready for further analysis.
- The script prints a confirmation message upon successful processing of each city.

3.1.3. Data Filtering and Cleaning

Following the aggregation of election data, it was crucial to ensure the consistency and accuracy of the dataset by removing neighborhoods that were not present in both election years (2014 and 2019) and excluding data related to prisons. This step was essential for maintaining the integrity of the dataset used for modeling, because the prisons may not actually represent the choice of people who live there, and by removing the neighborhoods that were not present in both election years, our model was able to provide more consistent results. Below is an outline of the data filtering and cleaning process:

- **Setup and Configuration:**
  - The script named *remove_prisons_and_neighborhoods.py* is configured to process directories containing election results for various election types (e.g., "Belediye Başkanlığı", "Büyükşehir Belediye Başkanlığı", "İl Genel Meclis Üyeliği", “Belediye Meclisi Üyeliği”).
  - Initializes lists to track information on removed rows and those removed due to the presence of "cezaevi" (prison) related keywords.

- **Data Processing:**
  - Reads Excel files from the specified directories for the years 2014 and 2019.
  - Extracts and stores data from each file into a DataFrame, categorized by city names.

- **Filtering Rows:**
  - The script filters out rows containing specific keywords related to prisons (e.g., 'ceza', 'ceza evi', 'cezaevi') to exclude any data associated with prisons.
  - The number of removed rows is tracked and recorded.

- **Matching Data:**
  - Performs an inner join on the DataFrames for 2014 and 2019 to retain only the rows (neighborhoods) present in both years.
  - The script identifies and records the rows that are removed because they do not exist in both datasets.

- **Saving Cleaned Data:**
  - The cleaned data for each city is saved into new Excel files, prefixed to indicate that they are cleaned results (e.g., CityName_2014_Cleaned_Results.xlsx and CityName_2019_Cleaned_Results.xlsx).

- **Logging Removed Data:**
  - Information about the removed rows is saved into JSON files for transparency and further analysis.
  - Separate JSON files are created to log the details of rows removed due to prison- related keywords and those removed because they were not present in both years.

3.1.4. Conversion to Percentage Results

After cleaning the election data, the next step is to convert the raw vote counts into percentage results. This transformation is essential for normalizing the data and facilitating meaningful comparisons and analyses. Below is an outline of the process for converting the cleaned data to percentage results:

- **Setup and Configuration:**
  - The script named *convert_to_percentage.py* is set to process Excel files from the directory containing cleaned election results outputted by the script above.
  - Configures the directory paths to locate input files and save the output files.
- **Reading and Processing Files:**
  - The script iterates over all cleaned result files in the specified directory, identifying those that need to be processed.
  - Each file is read into a DataFrame using the pandas library.
- **Conversion to Percentage:**
  - The script identifies the columns corresponding to political parties' vote counts.
  - For each political party column, the script calculates the percentage of total valid votes (Toplam Geçerli Oy) that the party received in each neighborhood (mahalle/köy).
  - The vote percentages are rounded to five decimal places for precision.
- **Saving the Results:**
  - The transformed DataFrame, now containing percentage results, is saved into a new Excel file with a modified filename to indicate the conversion (e.g., CityYear_Percentage_Results.xlsx).

- The output files are saved in a designated directory for percentage results, ensuring organized storage.

This conversion process standardizes the election data by expressing the vote counts as percentages, which is critical for the subsequent steps of analysis and modeling.

3.2. Format of the Scraped Data

3.2.1. Summed_Up_Results_X_T.xlsx

These files contain the number of votes for each of the political parties in each neighborhood of a given city X in year T. The columns of the excel sheet consist of "District Name", "Neighborhood Name", "Total Number of Electors", "Electors Who Vote", "Number of Valid Votes Without Objection", "Number of Valid Votes with Objections", "Total Valid Votes", "Total Invalid Votes" and the names of the political parties.

İlçe Adı	Mahalle/Köy	Kayıtlı Seçmen Sayısı	Oy Kullanan Seçmen Sayısı	İtirazsız Geçerli Oy Sayısı	İtirazlı Geçerli Oy Sayısı	Toplam Geçerli Oy	Toplam Geçersiz Oy	DSP	HKP	BBP	AK PARTİ	MİLLET	SAADET	HAK-PAR	LDP	BTP	İP	HDP	CHP	MHP
ALADAĞ	AKPINAR MAH.	541	524	503	0	503	21	1	1	0	118	0	2	1	1	0	1	0	48	330
ALADAĞ	AKÖREN MAH.	878	844	794	0	794	50	1	4	1	284	1	6	0	0	0	0	1	92	400
ALADAĞ	BAŞPINAR MAH.	643	610	589	0	589	21	1	0	1	100	4	1	0	0	0	0	0	93	389
ALADAĞ	BOZTAHTA MAH.	215	197	197	0	197	0	0	0	0	71	0	2	0	0	0	0	0	1	123
ALADAĞ	BÜYÜKSOFULU MAH.	1155	1098	1067	0	1067	31	0	0	3	679	1	14	0	0	0	1	0	123	245
ALADAĞ	CERİTLER MAH.	717	698	631	0	631	67	2	1	0	214	0	8	0	0	0	0	0	114	292
ALADAĞ	DARILIK MAH.	120	120	113	0	113	7	0	0	2	48	2	1	0	0	0	0	0	31	29
ALADAĞ	DAİLER MAH.	111	105	102	0	102	3	0	0	0	42	0	0	0	1	0	0	0	8	50
ALADAĞ	DÖLEKLİ MAH.	470	453	383	0	383	70	0	1	3	94	1	9	0	0	0	3	0	53	219

Figure 3.2.1.1.: Example rows of the Excel sheet for X = Adana and T = 2014.

These files provide electoral information for both the neighborhoods and the prisons of X, ensuring that the resulting data aligns consistently with the actual election results in X.

3.2.2. X_T_Percentage_Results.xlsx

These files contain the percentage value for each of the political parties in each neighborhood of a given city X in year T. The columns are the same, but this time the values under each of the political party columns represents the percentage of the number of received votes for this party in the corresponding neighborhood.

İlçe Adı	Mahalle/Köy	Kayıtlı Seçmen Sayısı	Oy Kullanan Seçmen Sayısı	İtirazsız Geçerli Oy Sayısı	İtirazlı Geçerli Oy Sayısı	Toplam Geçerli Oy	Toplam Geçersiz Oy	DSP	HKP	BBP	AK PARTİ	MİLLET	SAADET	HAK-PAR	LDP	BTP	İP	HDP	CHP	MHP
ALADAĞ	AKPINAR MAH.	541	524	503	0	503	21	0.00199	0.00199	0	0.23459	0	0.00398	0.00199	0.00199	0	0.00199	0	0.09543	0.65606
ALADAĞ	AKÖREN MAH.	878	844	794	0	794	50	0.00126	0.00504	0.00126	0.35768	0.00126	0.00756	0	0	0	0	0.00126	0.11587	0.50378
ALADAĞ	BAŞPINAR MAH.	643	610	589	0	589	21	0.0017	0	0.0017	0.16978	0.00679	0.0017	0	0	0	0	0	0.15789	0.66044
ALADAĞ	BOZTAHTA MAH.	215	197	197	0	197	0	0	0	0	0.36041	0	0.01015	0	0	0	0	0	0.00508	0.62437
ALADAĞ	BÜYÜKSOFULU MAH.	1155	1098	1067	0	1067	31	0	0	0.00281	0.63636	0.00094	0.01312	0	0	0	0.00094	0	0.11528	0.22962
ALADAĞ	CERİTLER MAH.	717	698	631	0	631	67	0.00317	0.00158	0	0.33914	0	0.01268	0	0	0	0	0	0.18067	0.46276
ALADAĞ	DARILIK MAH.	120	120	113	0	113	7	0	0	0.0177	0.42478	0.0177	0.00885	0	0	0	0	0	0.27434	0.25664
ALADAĞ	DAİLER MAH.	111	105	102	0	102	3	0	0	0	0.41176	0	0	0	0.0098	0	0	0	0.07843	0.4902
ALADAĞ	DÖLEKLİ MAH.	470	453	383	0	383	70	0	0.00261	0.00783	0.24543	0.00261	0.0235	0	0	0	0.00783	0	0.13838	0.5718

Figure 3.2.2.1.: Example rows of the Excel sheet for X = Adana and T = 2014.

These files provide electoral information for the neighborhoods that existed during both 2014 and 2019 elections. For all practical reasons and consistency, the results of neighborhoods that did not exist in both elections and prisons are excluded since these percentage files are crucial for the decision part of the model.

3.3. The Mixed-Integer Program

Parameters:

Parameter Name	**Explanation**	**Range**
n	Number of neighborhoods in the given city.	$n \geq 0$
m	Number of political parties in the given city.	$m \geq 0$
k	The number of neighborhoods that the user of this program aims to conduct polls in.	$n \geq k \geq 0$
M^T	Represents the election result matrix in year T. It is a "n x m" matrix. Each of the entries of the matrix represents the percentage of votes that the corresponding party got from the given neighborhood. To access the indices of this parameter, we used M^T_{ij} representation in the objective function.	$i \in \{1, 2, \ldots, n\}$ $j \in \{1, 2, ..., m\}$ $T \in \{2014, 2019\}$
G_i	Represents the ratio of received votes for ith political party. It is calculated by using the following equation: (the total number of votes the political party received) / (the total number of voters).	$i \in \{1, 2, \ldots, m\}$

Decision Variables:

Variable Name	Explanation	Range
x_i (Binary)	$x_i = 1$ if the given neighborhood is selected to the most representative neighborhood group, $x_i = 0$ otherwise.	$x_i \in \{0, 1\}$ $i \in \{1, 2, \ldots, n\}$
c_i (Continuous)	c_i represents the weight coefficient for the selected neighborhood.	$x_i \geq c_i \geq 0$ $i \in \{1, 2, \ldots, n\}$

Objective Function:

(min) $\left[\sum_{j=1}^{m}((G_j - \sum_{i=1}^{n} c_i M_{ij}^{2014}))^2\right] + \left[\sum_{j=1}^{m}((G_j - \sum_{i=1}^{n} c_i M_{ij}^{2019}))^2\right]$. Note that since Euclidean distance function is not a quadratic function, we did not directly use this equation in the Gurobi Optimizer but rather we minimized the terms that are inside of the square root, which is practically the same as minimizing the Euclidean distance.

Constraints:

Constraint	Explanation
$\sum_{i=1}^{n} x_i = k$	The model should select exactly user-defined number of neighborhoods.
$c_i \leq x_i$ for $i \in \{1, 2, ..., n\}$	The weight coefficient of a neighborhood should be zero if it is not selected. If it is selected, it should be between [0, 1].
$\sum_{i=1}^{n} c_i = 1$	The sum of all assigned weights should be equal to 1.
$x_i \in \{0,1\}$ and $0 \leq c_i \leq x_i$ for $i \in \{1, 2, \ldots, n\}$.	Binary variable and continuous variable constraints. Included once again for formality.

Crucial Points:

- While calculating the ratios of actual election results for the political parties; we decided to include prisons, and every neighborhood to get the exact result of the given year.
- While implementing the logic of selection of the most representative neighborhoods, we preferred the model to select only the neighborhoods that existed both in 2014 and 2019. In addition to that, we excluded prisons for the selection phase. This step was necessary to provide consistency.
- The *MIPGapAbs* value of the model is set to 10^{-3} to avoid unnecessarily long execution time (i.e., to implement pruning).

3.4. Outcomes of the Model

The model is used for the selection of the most representative neighborhoods in 4 different elections of 2014 and 2019:

- Mayor elections of metropolitan municipalities,
- Mayor elections of municipalities (for towns and districts),
- Membership elections of provincial councils,
- Membership elections of municipal councils.

The selected neighborhoods and their corresponding weights for each of the cities (and districts depending on the election type) are stored inside text files for the testing phase.

3.5. Testing Phase

The testing phase of this study aims to compare the results of the model to the results of mayor elections for metropolitan municipalities of Türkiye in 2024. It is decided to make the model select the most representative 3 neighborhoods by analyzing the top 5 most voted political parties.

3.5.1. Istanbul

The mayor election of metropolitan municipality of Istanbul differs from other cities since 2019 elections were repeated. To handle this issue, it is decided to analyze the elections in Istanbul as if the model analyzes three different elections in three different years. Based on the results of 2014, 2019 and 2019 (repeated) mayor elections of metropolitan municipalities, the model suggests the following neighborhoods of Istanbul with given weights:

Name of the Neighborhood	**Weight**
Süleymaniye (Fatih)	0.4779
İskenderpaşa (Fatih)	0.0193
Meclis (Sancaktepe)	0.5027

Based on the data provided by YSK – i.e., The Supreme Election Council of Turkey – the top 5 most voted parties and their corresponding percentages are as follows in Istanbul in 2024:

Name of the Political Party	**Percentage**
CHP	51.14%
AK PARTİ	39.59%
YENİDEN REFAH	2.61%
ZAFER PARTİSİ	2.12%
DEM	2.12%

In addition to that, the combined version of the results in the selected neighborhoods are as follows:

Name of the Neighborhood	**CHP**	**AKP**	**YRP**	**ZP**	**DEM**
Süleymaniye (Fatih)	38%	50.6%	3.6%	2.3%	2.3%
İskenderpaşa (Fatih)	41%	48.8%	2.8%	1.8%	2.2%

Meclis (Sancaktepe)	61%	30.9%	1.7%	1.8%	1.7%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Ankara is as follows:

Name of the Political Party	Prediction Calculation	Absolute Error
CHP	(0.38 x 0.4779) + (0.41 x 0.0193) + (0.61 x 0.5027) = 0.4961	|0.5114 – 0.4961| = 0.0153
AK PARTİ	(0.506 x 0.4779) + (0.488 x 0.0193) + (0.309 x 0.5027) = 0.4065	|0.3959 – 0.4065| = 0.0106
YENİDEN REFAH	(0.036 x 0.4779) + (0.028 x 0.0193) + (0.017 x 0.5027) = 0.0262	|0.0261 – 0.0262| = 0.0001
ZAFER PARTİSİ	(0.023 x 0.4779) + (0.018 x 0.0193) + (0.018 x 0.5027) = 0.0203	|0.0212 – 0.0203| = 0.0009
DEM	(0.023 x 0.4779) + (0.022 x 0.0193) + (0.017 x 0.5027) = 0.0199	|0.0212 – 0.0199| = 0.0013

3.5.2. Ankara

Based on the results of 2014 and 2019 mayor elections of metropolitan municipalities, the model suggests the following neighborhoods of Ankara with given weights:

Name of the Neighborhood	Weight
Ahmetadil (Akyurt)	0.2666
Atatürk (Akyurt)	0.3131
Demirci (Çubuk)	0.4201

Based on the data provided by YSK – i.e., The Supreme Election Council of Turkey – the top 5 most voted parties and their corresponding percentages are as follows in Ankara in 2024:

Name of the Political Party	Percentage
CHP	60.5%
AK PARTİ	31.6%
YENİDEN REFAH	3.1%
ZAFER PARTİSİ	1.5%
İYİ PARTİ	0.9%

In addition to that, the combined version of the results in the selected neighborhoods are as follows:

Name of the Neighborhood	CHP	AKP	YRP	ZP	İYİ
Ahmetadil (Akyurt)	23.6%	66.3%	6.1%	0%	1.8%
Atatürk (Akyurt)	40.4%	48.3%	7.2%	0.9%	0.6%
Demirci (Çubuk)	92%	0.4%	0.7%	1.4%	0%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Ankara is as follows:

Name of the Political Party	Prediction Calculation	Absolute Error
CHP	(0.236 x 0.2666) + (0.404 x 0.3131) + (0.92 x 0.4201) = 0.575	|0.605 – 0.575| = 0.03
AK PARTİ	(0.663 x 0.2666) + (0.483 x 0.3131) + (0.04 x 0.4201) = 0.344	|0.316 – 0.344| = 0.028
YENİDEN REFAH	(0.061 x 0.2666) + (0.072 x 0.3131) + (0.007 x 0.4201) = 0.0417	|0.031 – 0.0417| = 0.0107
ZAFER PARTİSİ	(0 x 0.2666) + (0.009 x 0.3131) + (0.014 x 0.4201) = 0.0087	|0.015 - 0.0087| = 0.0063
İYİ PARTİ	(0.018 x 0.2666) + (0.006 x 0.3131) + (0 x 0.4201) = 0.0066	|0.009 – 0.0066| = 0.0024

3.5.3. Izmir

Based on the results of 2014 and 2019 mayor elections of metropolitan municipalities, the model suggests the following neighborhoods of Izmir with given weights:

Name of the Neighborhood	Weight
Balaban (Bergama)	0.2197
Gültepe (Bergama)	0.0388
Konak (Konak)	0.7414

Based on the data provided by YSK – i.e., The Supreme Election Council of Turkey – the top 5 most voted parties and their corresponding percentages are as follows in Izmir in 2024:

Name of the Political Party	Percentage
CHP	48.97%
AK PARTİ	37.06%
DEM	4.19 %
İYİ PARTİ	3.64 %
ZAFER PARTİSİ	2.52 %

In addition to that, the combined version of the results in the selected neighborhoods are as follows:

Name of the Neighborhood	CHP	AKP	DEM	İYİ	ZP
Balaban (Bergama)	53.8%	42.3%	0%	3.8%	0%

Gültepe (Bergama)	46.1%	42.3%	0%	7.6%	3.8%
Konak (Konak)	50.7%	37.3%	4.5%	1.4%	1.5%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Izmir is as follows:

Name of the Political Party	**Prediction Calculation**	**Absolute Error**
CHP	(0.538 x 0.2197) + (0.461 x 0.0388) + (0.507 x 0.7414) = 0.5119	|0.4897 - 0.5119| = 0.0222
AK PARTİ	(0.423 x 0.2197) + (0.423 x 0.0388) + (0.373 x 0.7414) = 0.3858	|0.3706 - 0.3858| = 0.0152
DEM	(0 x 0.2197) + (0 x 0.0388) + (0.045 x 0.7414) = 0.0333	|0.0419 - 0.0333| = 0.0086
İYİ PARTİ	(0.038 x 0.2197) + (0.076 x 0.0388) + (0.014 x 0.7414) = 0.0216	|0.0364 - 0.0216| = 0.0148
ZAFER PARTİSİ	(0 x 0.2197) + (0.038 x 0.0388) + (0.015 x 0.7414) = 0.0125	|0.0252 - 0.0125| = 0.0127

3.5.4. Erzurum

Based on the results of 2014 and 2019 mayor elections of metropolitan municipalities, the model suggests the following neighborhoods of Erzurum with given weights:

Name of the Neighborhood	**Weight**
Karasu (Aşkale)	0.0721
Geventepe (Karayazı)	0.0652
Şükrüpaşa (Yakutiye)	0.8626

Based on the data provided by YSK – i.e., The Supreme Election Council of Turkey – the top 5 most voted parties and their corresponding percentages are as follows in Erzurum in 2024:

Name of the Political Party	**Percentage**
AK PARTİ	50.42%
İYİ PARTİ	22.18%
DEM	7.80%
CHP	6.52%
ZAFER PARTİSİ	5.11%

In addition to that, the combined version of the results in the selected neighborhoods are as follows:

Name of the Neighborhood	**AKP**	**İYİ**	**DEM**	**CHP**	**ZP**
Karasu (Aşkale)	35.3%	17.6%	11.7%	23.5%	5.9%
Geventepe (Karayazı)	18.7%	6.2%	56.2%	18.7%	0%
Şükrüpaşa (Yakutiye)	53.7%	28.2%	0.8%	2.7%	6.5%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Erzurum is as follows:

Name of the Political Party	**Prediction Calculation**	**Absolute Error**
AK PARTİ	(0.353 x 0.0721) + (0.187 x 0.0652) + (0.537 x 0.8626) = 0.5	|0.5042 – 0.5| = 0.0042
İYİ PARTİ	(0.176 x 0.0721) + (0.062 x 0.0652) + (0.282 x 0.8626) = 0.2599	|0.2218 – 0.2599| = 0.0381

DEM	(0.117 x 0.0721) + (0.562 x 0.0652) + (0.008 x 0.8626) = 0.0519	|0.078 – 0.0519| = 0.0261
CHP	(0.235 x 0.0721) + (0.187 x 0.0652) + (0.027 x 0.8626) = 0.0524	|0.0652 – 0.0524| = 0.0128
ZAFER PARTİSİ	(0.059 x 0.0721) + (0 x 0.0652) + (0.065 x 0.8626) = 0.06	|0.0511 - 0.06| = 0.0089

3.5.5. Trabzon

Based on the results of 2014 and 2019 mayor elections of metropolitan municipalities, the model suggests the following neighborhoods of Trabzon with given weights:

Name of the Neighborhood	Weight
Onur (Hayrat)	0.3405
Tarlacık (Vakfıkebir)	0.4370
Yokuşlu (Yomra)	0.2223

Based on the data provided by YSK – i.e., The Supreme Election Council of Turkey – the top 5 most voted parties and their corresponding percentages are as follows in Trabzon in 2024:

Name of the Political Party	Percentage
AK PARTİ	51.48%
CHP	28.46%
YENİDEN REFAH	9.14%
İYİ PARTİ	4.23%
SAADET	2.38%

In addition to that, the combined version of the results in the selected neighborhoods are as follows:

Name of the Neighborhood	AKP	CHP	YRP	İYİ	SP
Onur (Hayrat)	68%	8%	4%	0%	0%
Tarlacık (Vakfıkebir)	54.9%	27%	6%	3%	1.4%
Yokuşlu (Yomra)	53.7%	24.2%	5.6%	10.2%	1%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Trabzon is as follows:

Name of the Political Party	Prediction Calculation	Absolute Error
AK PARTİ	(0.68 x 0.3405) + (0.549 x 0.4370) + (0.2223 x 0.537) = 0.5908	|0.5148 – 0.5908| = 0.076
CHP	(0.08 x 0.3405) + (0.27 x 0.4370) + (0.242 x 0.537) = 0.275	|0.2846 – 0.275| = 0.0096
YENİDEN REFAH	(0.04 x 0.3405) + (0.06 x 0.4370) + (0.056 x 0.537) = 0.07	|0.0914 – 0.07| = 0.0214
İYİ PARTİ	(0 x 0.3405) + (0.03 x 0.4370) + (0.102 x 0.537) = 0.067	|0.0423 – 0.067| = 0.0247
SAADET	(0 x 0.3405) + (0.014 x 0.4370) + (0.01 x 0.537) = 0.011	|0.0238 – 0.011| = 0.0128

4. RESULTS & DISCUSSION

To discuss the results, performance, and accuracy of the model, it is decided to create several benchmarks in this study: accurate ordering of the political parties, weighted average of the absolute errors of the model and execution time of the model.

4.1. Benchmarks

4.1.1. Accurate Ordering of the Political Parties

One of the most crucial metrics for the accuracy of the model is the ordering of the political parties using only the selected neighborhoods and their corresponding weights. Based on the test cases of 5 mayor elections of the metropolitan municipalities, the model succeeds to correctly predict which party will be 1st, 2nd, 3rd, 4th, and 5th in all test cases.

4.1.2. Weighted Average of the Absolute Errors

The model succeeds to keep an absolute error value that is less than 0.05 for the overwhelming majority of the test cases. However, to further understand the overall accuracy performance of the model, it is decided to focus on weighted absolute error value. To calculate the weighted absolute error value of a city, below formula is used:

$\frac{\sum_{i=1}^{5} AE_i \ x \ (Percentage_i \ x \ 100)}{\sum_{i=1}^{5}(Percentage_i \ x \ 100)}$ where AE_i and $Percentage_i$ represents corresponding values of the i^{th} political party.

The weighted average of the absolute errors are as follows:

Name of the City	Weighted Absolute Error Value
Istanbul	[(0.0153 x 51.14) + (0.0106 x 39.59) + (0.0001 x 2.61) + (0.0009 x 2.12) + (0.0013 x 2.12)] / (51.14 + 39.59 + 2.61 + 2.12 + 2.12) = 0.0123
Ankara	[(0.02222 x 60.5) + (0.0152 x 31) + (0.0086 x 3.1) + (0.0148 x 1.5) + (0.0127 x 0.9)] / (60.5 + 31.6 + 3.1 + 1.5 + 0.9) = 0.0279
Izmir	[(0.04 x 48.97) + (0.04 x 37.06) + (0.20 x 4.19) + (0.4 x 3.64) + (0.5 x 2.52)] / (48.97 + 37.06 + 4.19 + 3.64 + 2.52) = 0.0192
Erzurum	[(0.0042 x 50.42) + (0.0381 x 22.18) + (0.0261 x 7.8) + (0.0128 x 6.52) + (0.0089 x 5.11)] / (50.42 + 22.18 + 7.8 + 6.52 + 5.11) = 0.0150
Trabzon	[(0.076 x 51.48) + (0.0096 x 28.46) + (0.0214 x 9.14) + (0.0247 x 4.23) + (0.0128 x 2.38)] / (51.48 + 28.46 + 9.14 + 4.23 + 2.38) = 0.0471

As shown in the above table, the weighted absolute error of the model is considerably low (i.e., less than 0.05). It explains the reason of the prediction results being significantly close to the real-life results.

4.1.3. Execution Time of the Model

- Processor: AMD Ryzen 5 5600H with Radeon Graphics,
- Clock Frequency: 3.30 GHz,
- CPU Time: 192.06 seconds (for 30 metropolitan municipality elections),
- Elapsed Time: 215,03 seconds (for 30 metropolitan municipality elections),
- Average Elapsed Time (1 metropolitan municipality election): 7 seconds.

4.2. Further Discussions

4.2.1. The Impact of Data

Analyzing the weighted absolute error amount of Istanbul, it is lower compared to the other cities. It is concluded that this finding signals the significance the existence of more data, which help the model to yield better predictions, since Istanbul was the only city in which the model analyzed 3 elections rather than 2.

4.2.2. The Curse of Relative Absolute Error

Name of the Political Party	Prediction Calculation	Relative Absolute Error
CHP	(0.538 x 0.2197) + (0.461 x 0.0388) + (0.507 x 0.7414) = 0.5119	|0.4897 - 0.5119| / 0.4897 = 0.04
AK PARTİ	(0.423 x 0.2197) + (0.423 x 0.0388) + (0.373 x 0.7414) = 0.3858	|0.3706 - 0.3858| / 0.3706 = 0.04
DEM	(0 x 0.2197) + (0 x 0.0388) + (0.045 x 0.7414) = 0.0333	|0.0419 - 0.0333| / 0.0419 = 0.20
İYİ PARTİ	(0.038 x 0.2197) + (0.076 x 0.0388) + (0.014 x 0.7414) = 0.0216	|0.0364 - 0.0216| / 0.0364 = 0.40
ZAFER PARTİSİ	(0 x 0.2197) + (0.038 x 0.0388) + (0.015 x 0.7414) = 0.0125	|0.0252 - 0.0125| / 0.0252 = 0.50

Figure 4.2.2.1.: The prediction calculations and RAE for Izmir.

Although the model succeeds to provide absolute error values that are less than 0.05 for the test cases, there is one more aspect to consider: the relative absolute error. Observing the RAE values found in Figure 4.2.2.1., it is shown that as the percentage value of a political party decreases (the percentages of the parties are provided at section 3.5.3. for Izmir), the RAE of the model for this political party significantly increases. The RAE value becomes more significant for parties with low vote percentages which would like to retrieve precise predictions with very low deviations.

4.2.3. The Impact of the Number of Neighborhoods to be Selected

The number of neighborhoods to be selected is a crucial factor for the accuracy and the weighted absolute error value of the model. In this study, the number of neighborhoods to be selected is set to 3 for all practical reasons – to create an intuitive model and test its performance. It is expected that higher number of neighborhoods to be selected would yield better accuracy for the model.

To provide an example to the impact of the number of neighborhoods that a user selects, it is decided to utilize the model with 5 political parties once again, but this time the number of neighborhoods to be selected is set to two. The model provides the following result for Ankara:

Name of the Neighborhood	**Weight**
Ahmetadil (Akyurt)	0.57
Demirci (Çubuk)	0.4299

The combined version of the results in the selected neighborhoods for the metropolitan municipality elections of 2024 are as follows:

Name of the Neighborhood	**CHP**	**AKP**	**YRP**	**ZP**	**İYİ**
Ahmetadil (Akyurt)	23.6%	66.3%	6.1%	0%	1.8%

Demirci (Çubuk)	92%	0.4%	0.7%	1.4%	0%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Ankara is as follows:

Name of the Political Party	**Prediction Calculation**	**Absolute Error**
CHP	(0.236 x 0.57) + (0.92 x 0.4299) = 0.530	|0.605 – 0.530| = 0.075
AK PARTİ	(0.663 x 0.57) + (0.04 x 0.4299) = 0.395	|0.316 – 0.395| = 0.079
YENİDEN REFAH	(0.061 x 0.57) + (0.007 x 0.4299) = 0.037	|0.031 – 0.037| = 0.006
ZAFER PARTİSİ	(0 x 0.57) + (0.014 x 0.4299) = 0.006	|0.015 - 0.006| = 0.009
İYİ PARTİ	(0.018 x 0.57) + (0 x 0.4299) = 0.010	|0.009 – 0.010| = 0.001

It is shown that the absolute error tends to slightly increase as the number of neighborhoods value gets decreased.

4.2.4. The Impact of the Number of Political Parties the Model Analyzes

Adding more political parties for the model to analyze yields higher relative absolute error values due to the fact that after a certain number of political parties (it was around 5-6 in this study), the percentage of the additional parties are less than or equal to 1%. Since the objective function of the model utilizes a Euclidean distance function, the model focuses more on minimizing the distance – recall that each political party has its own coordinate axis (Section 2.2.) – which is mostly dominated by the parties with higher percentages. Therefore, adding more and more parties with very low percentages compared to previously added political parties would cause high relative absolute values for these parties.

On the other hand, the test cases in section 3.5. shows that the absolute error values of the predictions are not affected by the low percentage values of corresponding political parties. To further analyze this with an example, Ankara is analyzed once again with 8 political parties rather than 5 and the number of neighborhoods to be selected is kept the same:

Name of the Neighborhood	Weight
Teberik (Akyurt)	0.4337
Hancılı (Kalecik)	0.416
Elören (Çamlıdere)	0.1501

Name of the Political Party	Percentage
CHP	60.5%
AK PARTİ	31.6%
YENİDEN REFAH	3.1%
ZAFER PARTİSİ	1.5%
İYİ PARTİ	0.9%
DEM PARTİ	0.8%
SAADET	0.4%
MEMLEKET	0.1%

The combined version of the results in the selected neighborhoods for the metropolitan municipality elections of 2024 are as follows:

Name of the Neighborhood	CHP	AKP	YRP	ZP	İYİ	DEM	SP	MP
Teberik (Akyurt)	37.8%	50.9%	8.0%	1.8%	0.6%	0.6%	0.0%	0.0%
Hancılı (Kalecik)	94.4%	3.7%	1.8%	0.0%	0.0%	0.0%	0.0%	0.0%

Elören (Çamlıdere)	55.1%	42.3%	2.5%	0.0%	0.0%	0.0%	0.0%	0.0%

Focusing on the given values, the analysis of the accuracy of the model for mayor election of metropolitan municipality in Ankara is as follows:

Name of the Political Party	**Prediction Calculation**	**Absolute Error**
CHP	(0.378×0.4337) + (0.944×0.416) + (0.551×0.1501) = 0.6392	|0.605 – 0.639| = 0.034
AK PARTİ	(0.509×0.4337) + (0.037×0.416) + (0.423×0.1501) = 0.2996	|0.316 – 0.299| = 0.017
YENİDEN REFAH	(0.080×0.4337) + (0.018×0.416) + (0.025×0.1501) = 0.0460	|0.031 – 0.046| = 0.015
ZAFER PARTİSİ	(0.018×0.4337) + (0×0.416) + (0×0.1501) = 0.0078	|0.015 - 0.007| = 0.008
İYİ PARTİ	(0.006×0.4337) + (0×0.416) + (0×0.1501) = 0.0026	|0.009 – 0.002| = 0.007
DEM PARTİ	(0.006×0.4337) + (0×0.416) + (0×0.1501) = 0.0026	|0.008 – 0.002| = 0.006
SAADET	(0×0.4337) + (0×0.416) + (0×0.1501) = 0	|0.004 – 0| = 0.004
MEMLEKET	(0×0.4337) + (0×0.416) + (0×0.1501) = 0	|0.001 – 0| = 0.001

It is shown that the initial assumption holds: the absolute error value does not increase as the number of political parties the model analyzes increases. However, it is crucial to note that the relative absolute error (RAE) value significantly increases as the user adds more and more political parties with very low percentages:

Relative Absolute Error
|0.605 – 0.639| / 0.605 = 0.056
|0.316 – 0.299| / 0.316 = 0.051

|0.031 – 0.046| / 0.031 = 0.483
|0.015 - 0.007| / 0.015 = 0.480
|0.009 – 0.002| / 0.009 = 0.711
|0.008 – 0.002| / 0.008 = 0.675
|0.004 – 0| / 0.004 = 1.0
|0.001 – 0| / 0.001 = 1.0

5. IMPACT

The major impact of this study is the creation of a model which would provide a new perspective to statistic-heavy estimation models such as machine learning, deep learning, and neural network models by utilizing an optimization model. In addition to that, by providing the users of the model the option to select the amount of neighborhoods they would like to select, it is aimed to decrease the cost of random sampling and polling since now it is possible to narrow down the places the pollster prefers to visit.

Based on the research on the website Espacenet – which offers access to millions of patent documents from over 100 countries – there is no Freedom-to-Use issue found.

6. ETHICAL ISSUES

There are no ethical issues in this project.

7. PROJECT MANAGEMENT

To analyze the changes in initial and final project plans, it is decided to focus on the capstones of the project:

- **Conceptual Design:** The main issue that was identified in this step is the absence of a system that provides insights to the user about the neighborhoods to conduct polls such that the user (i.e., pollster) would be able to get the accurate overview of the potential results of the election by solely specifying the name of the city and the amount of neighborhoods the user wants to conduct polls in. In this step, we focused on conducting research, gathering basic requirements, defining objectives, and exploring potential solutions. These basic design ideas included the identification of the need for automatizing the data extraction from the official website of election results, creating an optimization model that utilizes that two-dimensional election data and using this optimization model in a program that takes city name and the number of neighborhoods to be selected information from the user.

- **Preliminary Design:** In the preliminary design phase, the focus was on refining the conceptual ideas into more concrete designs. These preliminary design ideas of the optimization model included assigning both a binary and a continuous decision variable for each of the neighborhoods in a city in the model, identifying the need for an accuracy-based objective function and determining the constraints of the optimization model. The initial version of the accuracy function was to find the pre-determined number of neighborhoods – each of which are interpreted as vectors in m-dimensional space where m is the number of political parties – such that once they are summed up, they maximize the cosine similarity between this newly created vector and the final election result vector.
  In addition to that, we decided on using the Selenium library of Python to automatize the data extraction part of the project.

- **Design Decision:** In this step, we reviewed and discussed our initial design ideas with the supervisor of the project. After this evaluation, we decided to interpret each of the neighborhood datum as a point in m-dimensional space rather than a vector in space and we decided to focus on minimizing the Euclidean distance between the point that is created by summing up the points that are selected by the optimization model and the point that represents the final election result. The major factor that caused this change was the fact that cosine similarity function is not a quadratic function and since we decided to use Gurobi Optimizer, we had to create an objective function which is quadratic.

- **Detailed Design:** In this phase, we extracted the data of elections and created the final mathematical model that optimizes the adjusted objective function. The details of the final model can be accessible at section 3 and 4.

8. CONCLUSION AND FUTURE WORK

8.1. Conclusion

The project succeeds to fulfill its initial aim of providing accurate and more affordable pre-election polling method/strategy by creating and utilizing an optimization model in which the user is able to specify the number of neighborhoods to avoid redundant traveling/polling costs. The findings in section 4.1.2. shows that the weighted absolute error rate of the model is excessively low, which proves the low error rate of the provided model. By just conducting polls in the selected neighborhoods and multiplying the percentage values of each party in each neighborhood with the corresponding weight value, it is possible for users to create their own predictions on future elections.

8.2. Future Work

Although the model is proven to provide accurate predictions, there are still some considerations and improvement opportunities for the future.

8.2.1. The Addition of Statistical Methods to the Current Model

Currently, the model assumes (hence, it requires) the full participation of voters in a certain neighborhood for the pre-election poll to yield accurate results. To further develop this issue and further decrease the required number of participants, it is recommended to create random samples in a neighborhood and predict the poll result of the given neighborhood by using the values of the random sample after processing it with statistical methods. Then the user may proceed to use the weight value assigned by the model to estimate the results of the future election(s).

8.2.2. The Impact of Migration

The intense migration of the residents of a neighborhood may severely change the weight value of a neighborhood – it may even cause the model to not select the certain neighborhood anymore. Hence, the impact of migration must be further analyzed.

8.2.3. The Impact of Politics and Changes in Political Profiles of the Neighborhoods

It is possible that in a worst-case scenario, a politician works so poorly that the stable distribution of political ideas of the residents may change in an extreme manner. In this case, similar to migration example, the model may need to reassign weight to this neighborhood or deselect it to reduce weighted absolute error value. This concept is also remained open for future work.

8.2.4. Objective Function Adjustments for Political Parties with Low Percentages

It is mentioned in section 4.2.2. that the RAE value significantly increases for the predictions of the model on political parties with lower percentages since the minimization model is dominated by the political parties with higher percentages – which is expected since the model uses pure Euclidean distance function. To avoid this, further adjustments to the objective function may be implemented in the future.

9. APPENDIX

All of the findings which consist of the selected neighborhoods in a given city/district with their corresponding weights values for:

- Mayor elections of metropolitan municipalities,
- Mayor elections of municipalities (for towns and districts),
- Membership elections of provincial councils,
- Membership elections of municipal councils,

is accessible via this URL:
https://drive.google.com/drive/folders/1Mc-x1oQaAXg2DtfzfjzMnqsETwVDgj0U?usp=sharing

10. REFERENCES

Aydaş, İ., Moral, M. ve Tosun, Y. (2022). "Türkiye'de Seçim Anketlerinin Toplam Anket Hatası Perspektifinden Bir İncelemesi". *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *51*(1), 87-110. https://doi.org/10.30794/pausbed.1119635

Espacenet. (2024). *Results*. https://worldwide.espacenet.com/patent/search?q=finding%20the%20most%20representative%20polling%20units%20for%20pre-elections&queryLang=en%3Ade%3Afr

OpenSecrets. (2024). *Expenditures Search*. https://www.opensecrets.org/campaign-expenditures/search?type=v&query=polling

Pilnacek, M., Tabery, P., Prokop, D., & Kunc, M. (2021). Apportioning Uncertain Voters in Pre-Election Polls in a Multi-Party System. *International Journal of Public Opinion Research*, *33*(4), 1-13. https://doi.org/10.1093/ijpor/edaa027

Yüksek Seçim Kurulu. (2024). *Sorgu*. https://sonuc.ysk.gov.tr/sorgu