Analysis of Israeli Census data in correlation with Electoral Results.

Gaby Yeselson. Ben Gurion University of the Negev.

Github Repository: https://github.com/roflmoqkz/Analysis-of-Israeli-Census-data-in-correlation-with-Electoral-Results

Abstract

This study quantifies how demographic and household characteristics vary with party support by integrating administrative, web, and geospatial data. We do this through analysis of the 2022 Israeli Census together with electoral data from the November 2022 election. Census indicators were joined with election results through a pipeline that scraped polling-station metadata, geocoded curated addresses, and spatially matched stations to statistical areas. Party vote totals were aggregated to (city, area) units and combined with city-level results to compute party-weighted averages of census features. Results, translated to common party labels and ordered for comparability, reveal consistent gradients: ultra-Orthodox and Arab-aligned parties are associated with younger populations, larger households, higher dependency, and denser settings; secular/center-left parties align with older, smaller, higher-asset households; national right and nationalist-religious parties occupy intermediate profiles.

Introduction

Elections reflect underlying social structure, yet administrative reporting often fragments the link between where people live and how they vote. This research addresses that gap by reconciling city-level and area-level election results from the November 2022 election with fine-grained census indicators from the 2022 Israeli Census. It constructs a spatial bridge from polling-station outcomes to statistical areas using curated address lookups, rate-limited geocoding, and polygon-based joins. Vote metrics from multiple sources are normalized, mapped to standard party labels, and aggregated to consistent (city, area) units.

The analytic goal is to estimate, for each party, the distribution of demographic and household characteristics across the places where it performs well. A weighted averaging model leverages city-level shares where statistical-area detail is missing and area-level shares where available, preserving locality structure via population-based multipliers. Visualizations—single-metric and grouped bar charts—summarize contrasts across parties in age structure, household composition, economic proxies, and origin indicators. The approach yields a coherent portrait of socio-demographic correlates of party support and a reproducible framework for merging administrative, web, and spatial data.

Gathering Information

Data was ingested from three primary files. Census indicators were loaded from census.xlsx, a file we got from the Israeli Census website[1]. This file contains a wide table of locality-level and statistical-area features including population approximations and indicators such as sex ratio, age structure, household composition, and demographics.

Vote tallies were gathered from the Israeli election 25 website[2]. From here we gathered 2 files: expc.csv which contains voting results for every city, and expb.csv which contains results for every polling booth. Both of these files were used for analysis.

Our next task was to match the polling booths to the statistical areas. Since the expb.csv file did not contain any geographic information about the location of the booths, we used

Party lists were normalized from Hebrew to English using a fixed translation dictionary (trans_dict), which was later applied to all party-facing outputs.

The first attempt of getting the Polling-station locations was done programmatically. Unique city identifiers were extracted from the census file and used to drive a Selenium browser to the official election site's per-city results pages. For each city, the polling-station selector was parsed and its options were harvested to capture the human-readable location (such as the name of a school) and the station's numeric value. These human-readable locations were geocoded via Nominatim with rate limiting and a simple cache. Progress bars were used for long-running loops.

This however was proven unsuccessful because the geocoder could not identify the location of most polling stations, only about 1000 out of 9758 polling stations (in large cities) were identified.

Instead we used the file 'adresses.xlsx' which we obtained from the Israeli Government Website[3]. adresses.xlsx was opened and a lookup was built from (city code in column C, station value in column E) to the address in column G. Due to the fact that the poll numbers were slightly different, expb.csv was opened to map, we used area codes in column F (in both files) to map the poll booths before the address lookup. The resulting addresses were geocoded via Nominatim just as before. This time we managed to geocode 7773 booths, not all, but we did account for the missing ones in the analysis.

Geospatial processing used polygons read from geo.gdb.zip, which is the file that contains the geography of the different statistical areas. This file was obtained from the Bureau of Statistics website[4]. Layers were scanned to retain geometry and key identifiers, including 2022_STAT (statistical area), SEMEL_YISHUV (locality code), and SHEM_YISHUV_HEB (locality name).

Coordinate reference systems were aligned, and the geocoded polling-station points were spatially joined to polygons to assign each station to a statistical area, producing the polling_in_areas table with city, value (polling-station ID), and area (statistical area).

Aggregation to the (city, area) level proceeded by aligning vote metrics from expb.csv with areas. A mapping from (city, value) to area was constructed from polling_in_areas; the pipeline also supported reading area directly from expb where available (area from column B and value from column F).The result was an area_results table keyed by (city, area) with party vote totals as columns. In parallel, city-level party metrics were prepared from city_results, selecting the city identifier (column index 2) and treating columns from index 8 onward as party metrics.

Part 1: Expected Voter Analysis

The biggest issue in analyzing the data for this project is the fact that Israel has 13 major parties. At first, we tried to map the various statistical features in the cities and areas to the major party that was selected in that city/area, and produce a map similar to K-means clustering. But then, we came up with a more clever approach.

To relate census features to party support, a weighted averaging model was constructed. For each census feature (columns from index 6 onward) and each party, weights were derived from population approximations and vote-share distributions. The weight was calculated as:

$w = p * v * r$

Where: w is the weight, p is the population, v is the vote share (number of votes for the party divided by the total votes. And r is the 'city multiplier'

For small towns without a statistical area (blank StatArea),The city multiplier is 1 (irrelevant). But for larger cities, a locality-specific multiplier was computed from the total population of each city as follows:

$r = p/c$

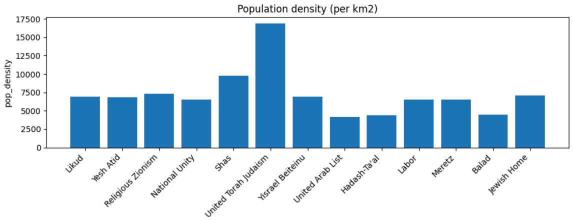
Where p is the population of the statistical area/city and c is the total population of the city.

The rationale behind this approach was to account for the missing polling stations, and to account for statistical areas that had no polling places. This gave more weight to the entire city, which was guaranteed to have proper census and voting data, as well as to the localities, which had more precise data, although less complete. This is especially important as large cities can have different neighborhoods with completely different statistics and voting patterns.

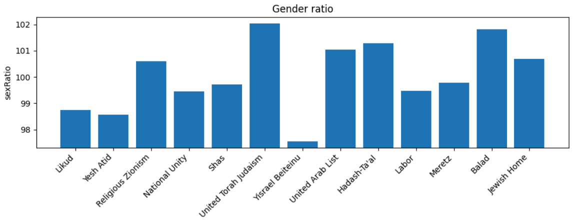
Feature-by-party weighted averages were assembled into a matrix (weighted_avgs). This matrix was then transformed into the final results by renaming party columns using trans_dict, selecting the top 13 parties (those that received at least 1% of the vote) and arranging them from the most chosen party to the least.

For visualization, bar charts were generated for each specified census feature or grouped set of features. Single-feature plots displayed party-wise bars for that feature; grouped plots displayed clustered bars aligned by party, with a consistent color assigned to each statistic across parties. For statistics where the numbers were large and close together (such as gender ratio), y-axes were scaled to the data’s min–max range rather than anchored at zero to emphasize variation in central ranges.

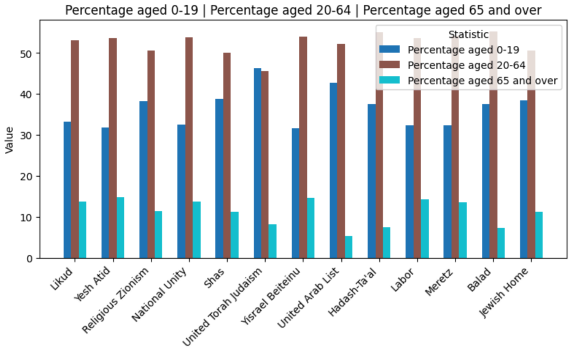
Results

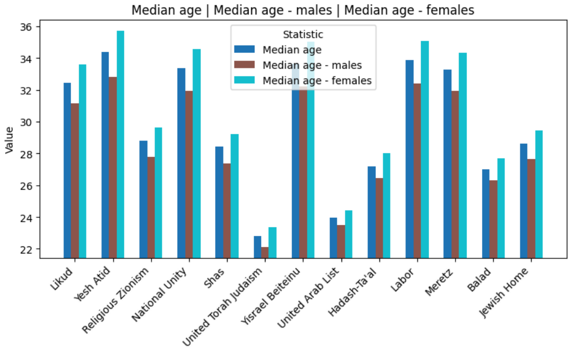


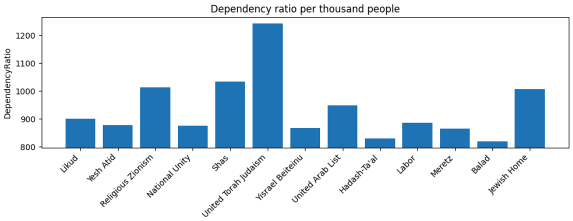
Population density was highest in areas voting for United Torah Judaism and, to a lesser extent, Shas, indicating strong support in very dense, urban localities. Parties such as United Arab List, Balad, and Hadash–Ta’al were associated with lower average densities, while Likud, Yesh Atid, Labor, and Meretz occupied midrange urban contexts.



The gender ratio is the number of men per 100 women. There seems to be little variance here, but secular parties tend to skew a bit towards females and religious parties towards men.

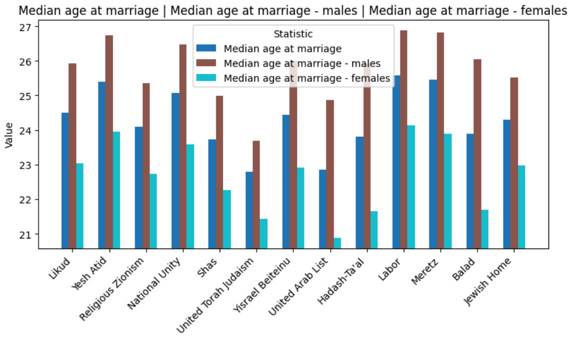


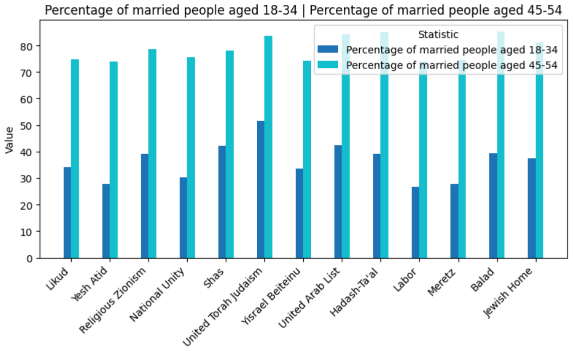


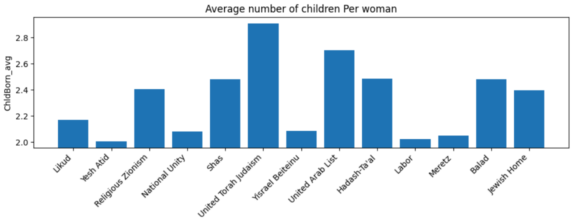


The dependency ratio is calculated by summing the number of dependents (ages 0-14 and 65+) and dividing by the working-age population (ages 15-64),

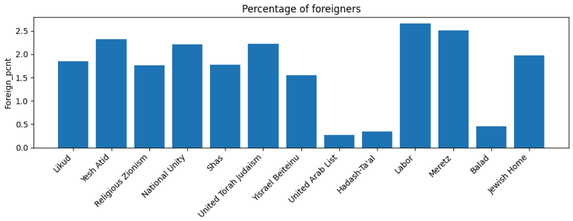
Age structure and dependency patterns showed pronounced contrasts. Ultra‑Orthodox (United Torah Judaism, Shas) and Arab‑aligned (United Arab List, Balad) parties exhibited younger profiles: higher shares of ages 0–19, lower 65+, low median ages, and elevated dependency ratios. Secular/center‑left parties (Yesh Atid, Labor, Meretz) tended toward older age structures, higher 20–64 shares, higher median ages, and lower dependency. Religious Zionism and Jewish Home sat between these poles, skewing younger than the secular blocs but older than the ultra‑Orthodox and Arab‑aligned groups.

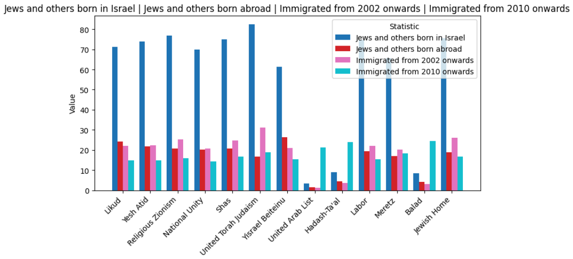


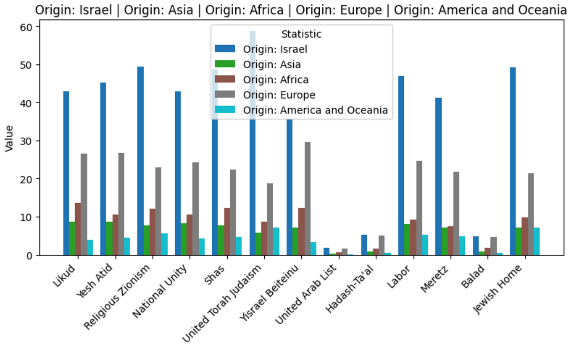




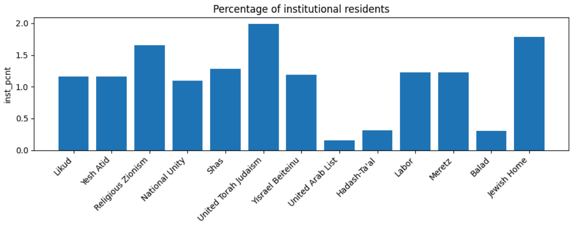
Marital patterns and fertility aligned with the age structure. Median marriage ages were lowest in United Torah Judaism and United Arab List and highest for Yesh Atid, Labor, and Meretz. Average children per woman was highest for United Torah Judaism and Shas, elevated for Arab-aligned parties, and lowest for Yesh Atid and Meretz.

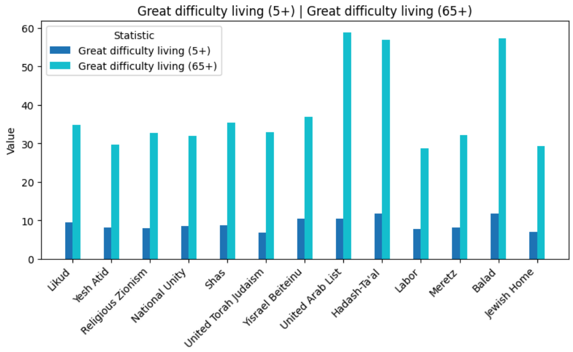






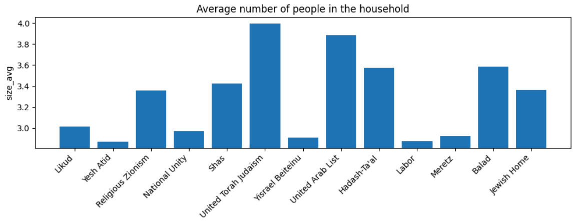
Origin and migration indicators suggested distinct compositional profiles. “Jews and others born in Israel” and related origin fields were highest in ultra-Orthodox concentrations and lower among Arab-aligned parties by construction; center-left parties (Labor, Meretz) and National Unity showed relatively higher shares of foreign-born/immigrant cohorts and recent immigration proportions, consistent with their urban, higher-education geographies.

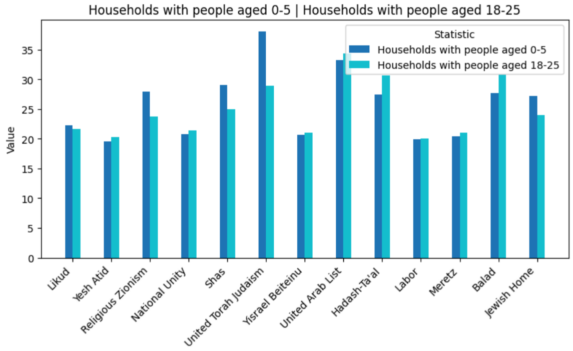




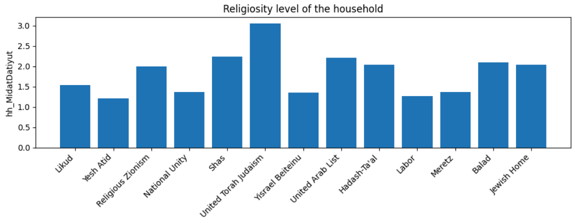
Difficulty-of-living indicators were elevated among Arab-aligned parties and generally lower among Yesh Atid, Labor, and Meretz; ultra-Orthodox parties were mixed—relatively low difficulty among working-age but higher among older cohorts

.





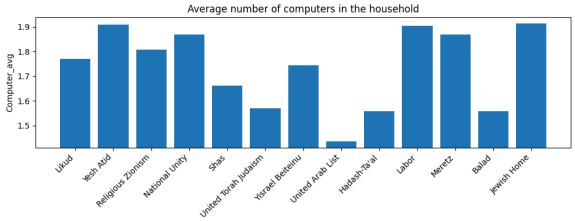
Household composition mirrored these demographics. Average household size and the prevalence of households with young children (0–5) were highest in United Torah Judaism and Shas, and elevated among Arab-aligned parties; they were lowest for Yesh Atid, Labor, and Meretz. Households with members aged 18–24 were most common in Arab-aligned parties and relatively high for ultra-Orthodox parties, consistent with the younger age distributions.

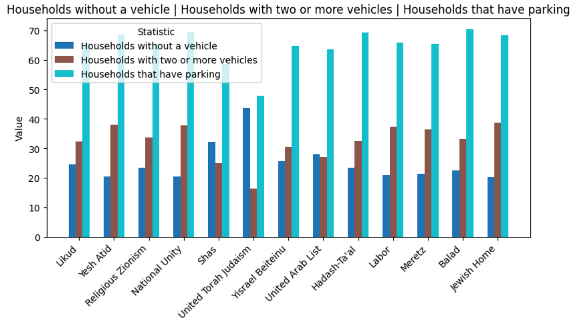


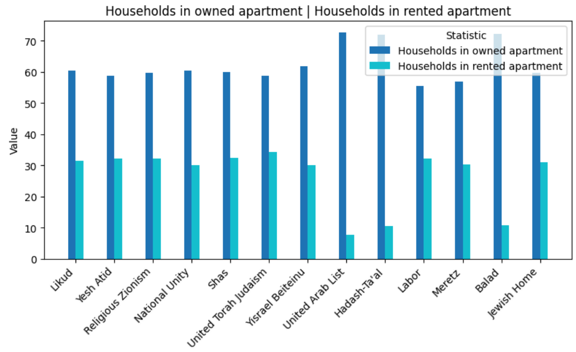
For this statistic, each household was given a score as follows, depending on the majority in the city/area:

Self Identification	Score
Secular	1
Traditional	2
Religious/Very Religious	3
Ultra Orthodox	4

Unsurprisingly, the more religious districts lean towards the religious parties while the secular districts lean towards the left-leaning parties.







Socioeconomic proxies differentiated blocs. The prevalence of households without vehicles was highest where United Torah Judaism and Shas performed well and lowest in higher-income, car-rich strongholds of Yesh Atid, National Unity, Labor, Meretz, and Jewish Home. Two-plus-vehicle ownership, parking availability, and average computers per household were highest in these latter parties and lowest for United Torah Judaism and United Arab List.

Part 2: ML Prediction Model

The target party for each census row was derived from election summaries. Sample weights equaled the approximate population to bias learning toward more populous observations.

A KMeans model was fit with sample_weight. The number of clusters K was selected by scanning values around the number of unique parties using silhouette score; K was then constrained to be at least the number of parties. To guarantee coverage (at least one cluster per party), clusters were mapped to parties via a weighted assignment: first by solving a maximum-weight (Hungarian) assignment on the cluster×party weight matrix, then assigning any remaining clusters by weighted majority. This mapping yields a predictable party label for every cluster and ensures every major party has at least one cluster. Model quality was summarized with weighted purity (population-weighted majority share within clusters) and homogeneity, and predictions were obtained by applying the cluster→party mapping to cluster labels.

To visualize structure, principal component analysis (PCA) was applied to the standardized feature matrix (same features as the clustering). A two-component PCA captured the dominant axes of variation; coordinates were plotted as a scatter colored by the predicted party (or cluster), with cluster centroids overlaid in PC space. Explained-variance ratios summarized how much variation PC1 and PC2 captured, and loadings identified the features contributing most to each component. This view highlighted separations consistent with the clusters: gradients in age structure, household size, and socioeconomic proxies aligned with PC directions, and parties clustered along these axes.

Population weighting prioritizes high-population areas; patterns in small localities may be underemphasized. KMeans assumes convex, spherical clusters; alternative models (e.g., Gaussian mixtures) could capture anisotropy. The cluster→party assignment enforces coverage but may split large parties across multiple clusters; this is by design to preserve heterogeneity while maintaining predictability.

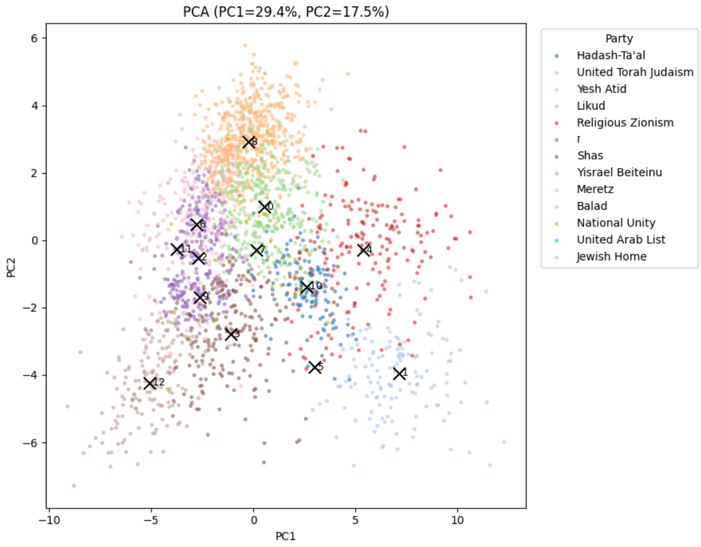
Results

The sizes of the clusters are as follows:

Party	Size weighted based on population (per 1000)	Size weighted based on voters
Likud	885	0.7932138835
Yesh Atid	1785	2.106851853
Religious Zionism	561	1.085910121
National Unity	243	0.562913601
Shas	1295	3.296383384
United Torah Judaism	1293	4.614588464
Yisrael Beiteinu	608	2.847482533
United Arab List	1006	5.183898746
Hadash-Ta'al	1172	6.557361457
Labor	0	0
Meretz	563	3.735584543
Balad	2264	16.33385515
Jewish Home	493	8.681109643

*Labor did not get a majority in any district or city.

The sizes of these clusters generally indicate how diverse the voters for that party are. Parties that are more diverse receive a lower per voter size. This is because there are less districts where they are a distinct majority. Therefore: Labor,National Unity and Likud are the least diverse, as those are the ones that tend to appeal to broader demographics, where as Arab and Ultra Orthodox parties tend to be very concentrated.



This is the PCA component analysis for the various parties. The components are:

PC1	29.40%	PC2	17.50 %
Variable	Value	Variable	Value
Percentage aged 0-19	0.276	Households with two or more vehicles	0.322
Average number of people in the household	0.275	Households without a vehicle	0.308
Median age	-0.267	Median age at marriage	0.295
Median age - females	-0.264	Median age at marriage - females	0.295
Median age - males	-0.260	Average number of computers in the household	0.284
Households with people aged 0-5	0.255	Median age at marriage - males	0.256
Average number of children Per	0.227	Households that have parking	0.247

woman			
Percentage aged 65 and over	-0.224	Great difficulty living (5+)	0.246
Origin: Israel	0.219	Jews and others born abroad	0.227
Religiosity level of the household	0.217	Jews and others born in Israel	0.227
Percentage of married people aged 45-54	0.201	Great difficulty living (65+)	0.194
Jews and others born in Israel	0.195	Percentage of married people aged 18-34	0.187
Jews and others born abroad	-0.195	Origin: Israel	0.185
Origin: Europe	-0.190	Religiosity level of the household	0.162
Dependency ratio per thousand people	0.181	Origin: Europe	0.138

We see from the table that PC1 mainly refers to Demographic Parameters, especially to age, whereas PC2 mainly refers to socioeconomic factors.

Both PC1 and PC2 have very high contribution factors for all of the top parameters. In addition, the difference between the dominance of PC1 and PC2 is quite big, and is also dominant compared to the other Principal Components.

This signifies that many of the parameters are linked: Demographic age parameters, and Socioeconomic parameters. It also suggests that Age and Wealth are the dominant factors contributing to the chosen party, rather than religious or ethnic factors.

Part 3: Mapping the Voters

Finally, we wanted to investigate whether geography has any effect on voting patterns.

The map visualizes the locally winning party and its strength using administrative polygons and aggregated vote data. Polygon geometry was sourced from the official layer with locality code. In the Israeli census, large cities are divided into wards, which are divided into subwards, which are divided into statistical areas. For larger cities, statistical areas were grouped by hundreds (e.g., 300–399) using integer division of the statistical area. This is because 100-group represents the ward for which it belongs.

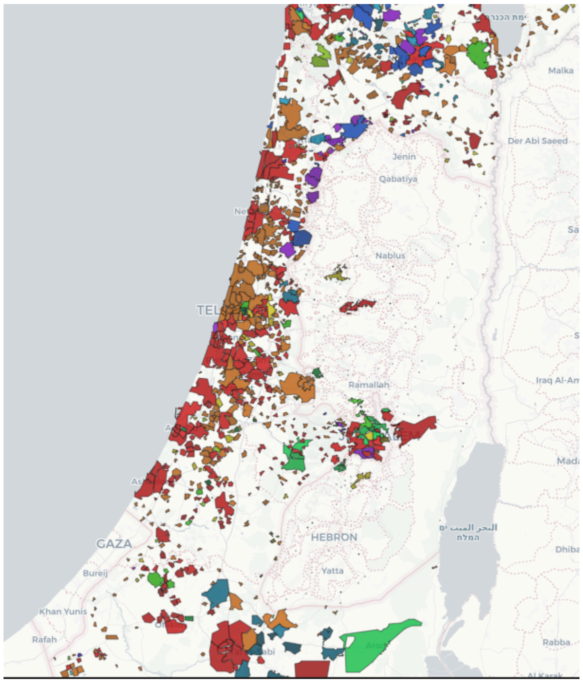
For each 100-group, party totals were summed from area_results across all member areas, the winning party and its share were computed, and the corresponding polygons were dissolved to a single geometry per group.

Party color was encoded on the HSV circle, assigning evenly spaced hues by the predefined party order to ensure consistent, interpretable coloring across views. Brightness scaled with the winner's vote share (brighter at low shares, darker at high shares), conveying margin of victory.

The resulting records (city name, hundred-group index, winner, percent, geometry) were assembled into a

GeoDataFrame, reprojected to WGS84, and rendered as an interactive web map with Folium on a light CartoDB basemap. Polygons were styled by the computed color, and tooltips display the city name, ward number, winning party, and vote share. The workflow uses unary dissolves for group geometries, normalizes keys across sources to avoid join mismatches, applies progress tracking for city iteration, and gracefully skips incomplete cases.

Results



Note: A larger and scrollable map can be found in the ipynb file.
From the map, we can see that there is a small but significant difference between voting patterns in cities, even for the diverse parties. Cities like Tel Aviv have a higher share of Yesh Atid voters across all districts, whereas cities like Beer Sheva, and Rishon le Zion have majority Likud voters. And this applies consistently throughout the wards in the city, with at most one ward having a different party than the majority. The exceptions to this are the Ultra Orthodox areas, as well as Haifa, which is the only major city that is split between Likud and Yesh Atid.

Another conclusion we can infer is about the various regions of Israel. According to the map, there is no strong party preference in different districts of Israel, rather the differences are more between cities than between regions, and several regions can have different cities next to each other, with completely different voting patterns.

Conclusion

Taken together, the grouped charts,k-means,PCA, and maps, depict a consistent structure: parties associated with younger, larger, denser households (United Torah Judaism, Shas, and Arab-aligned lists) contrast with parties associated with older, smaller, higher-asset households (Yesh Atid, Labor, Meretz, National Unity), with right-of-center national parties (Likud, Yisrael Beiteinu) occupying intermediate profiles and Religious Zionism/Jewish Home bridging to the younger, higher-dependency end. These gradients are stable across population density, age/household composition, economic proxies (vehicles, parking, computers), difficulty-of-living measures, and marital/fertility statistics.

Further Research

Another way we tried to analyse the data was by taking a combination of 2 different statistics, and making a scatter plot with the majority party based on them. This approach should be expanded in a way so that we can not only see the winning party, but also the percentage that they got, that’s the reason we chose our approach instead.

Another way to expand the research is to study past elections and census data, as this paper solely focuses on the data from 2022.

Acknowledgment

We used Github Copilot with GPT-5 while writing this paper.

Sources

[1]:“נתונים נבחרים, לפי יישובים ואזורים
סטטיסטיים - מפקד האוכלוסין 2022.”
https://census.cbs.gov.il/he/extra-info.
[2]:”תוצאות לפי קלפיות, תוצאות לפי ישובים
בחירות לכנסת ה-25”
https://votes25.bechirot.gov.il/
[3]:“דו"ח קלפיות טופס א, נכון ליום 10.7.22”
https://www.gov.il/he/pages/kalpi_place
[4]: “שכבת אזורים סטטיסטיים 2022”
https://census.cbs.gov.il/he/extra-info.