Final Statistics Lab Report

Balad

By Mutaz Ayesh 206595472 and Mohammad Majadli 206416554

1 + 2 Background, About Balad

The Israeli legislative elections are general elections where Israeli citizens vote for the Israeli parliament, the Knesset. Citizens vote for a party or a list including an alliance of parties, and the number of seats which it receives in the Knesset is proportional to the number of voters who voted for it. Until 2013, the electoral threshold which allowed parties or lists to enter the Knesset was 2%, but before the 2015 elections were held the threshold was raised to 3.25% with the aim to weaken Arab parties that usually ran separately.

To combat the move which was described as 'racist' by several publications, Arab parties formed the Joint List which included the 4 main parties known by their Hebrew acronyms Ra'am, Hadash, Ta'al, and Balad, and became the third biggest party of that election.

Recently in 2020, the list was also the third biggest party with over half a million votes resulting in 15 seats, the most seats ever. The 2021 elections witnessed Ra'am breaking away from the list, with the other 3 parties remaining, and in 2022 Hadash and Ta'al ran together while Balad ran independently.

In our party Balad, we will be taking into account these differences wherever relevant by splitting the Joint List of 2021 into 60% Hadash-Ta'al and 40% Balad. That is an approximate estimate based on the aftermath of 2022 elections, where Hadash-Taal received 170k and 131k votes, respectively. Further analyses in this report showed that a similar proportion of voters of 2021's Joint List immigrated to 2022's Hadash-Ta'al and Balad, further strengthening our statistical decision.

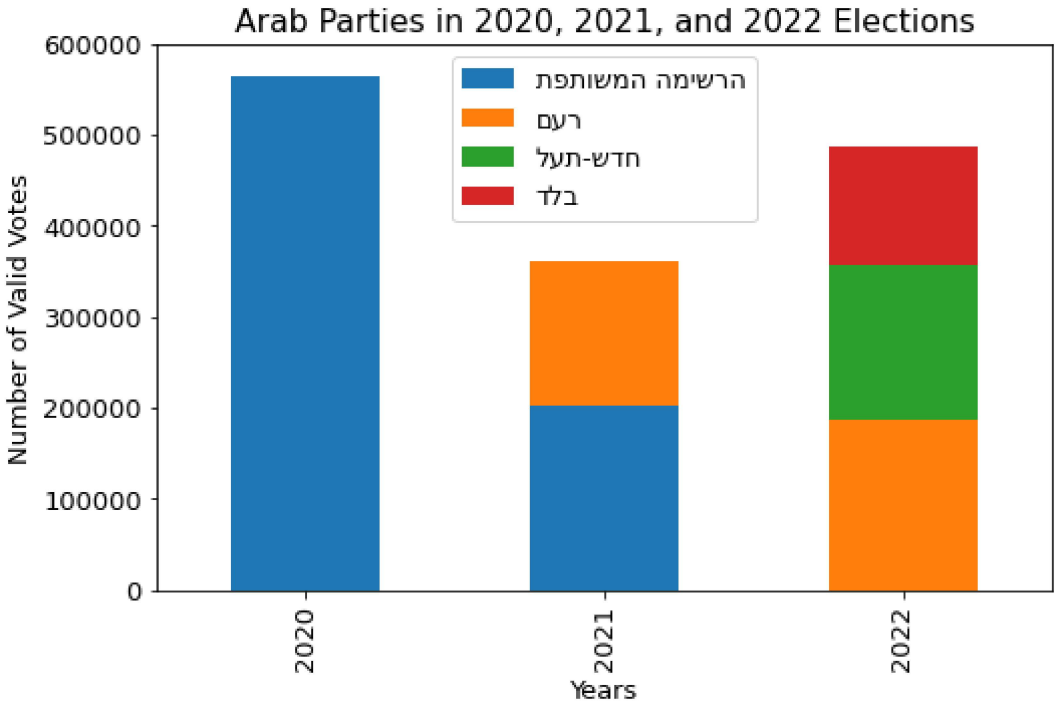
The datasets that we will be using were extracted from the official website of the Israeli elections. Mainly, we will be using the datasets based on polling stations (and there are 11707 polling stations in 2022 and 12127 in 2021) and minimally on those based on cities as a whole. We used the datasets of polling stations as we assume that they represent the population of the neighborhoods in which they are located. As a result, what are known as "double envelopes" were excluded completely from all analyses.

3. Voter turnout

Balad is a secular Arab party whose agenda centers Palestinian citizens of Israel and advocates for their rights. In all the elections between 2003 and 2013 Balad ran independently and consistently maintained 3 seats and increased in popularity with 66k votes cast in 2003 and 97k in 2013.

In recent years, the figure below shows the noticeably low turnout of Arab citizens in 2021 compared to the years before and after it where 2021's Joint List (which included Balad) received 201k votes. The Arab voter turnout in 2022 increased by 35% compared to 2021, and if the list remained together in 2022 it would have had 301k in total, a 50% increase for the alliance from last year.

However, Balad failed to enter the 2022 Knesset despite gaining an all-time record of 131k votes. This can be attributed to both the raising of the electoral threshold which was specifically designed to limit Arab participation in the Knesset as mentioned above and to the division of Arab parties which left its voter base divided as well. This effectively means that Balad is a party that is specifically vulnerable to voter turnout compared to the other two Arab parties.



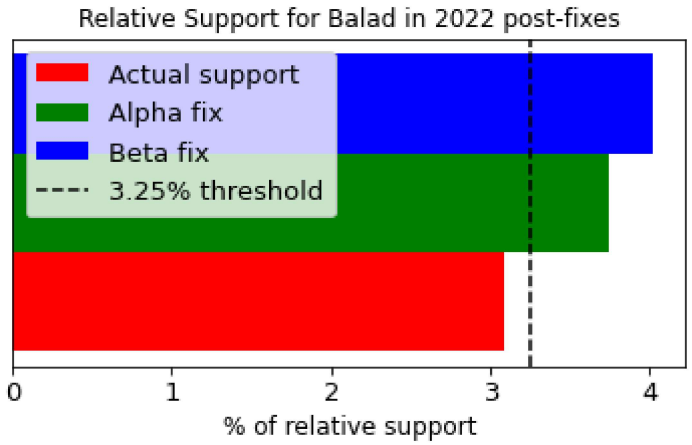
To further strengthen our claims above regarding Balad's vulnerability to Arab voters not voting in the elections, we decided to analyze Balad's performance had all potential voters voted.

First we fixed the elections results by assuming constant turnout per *ballot*. We checked the number of those with the right to vote who did not exercise their right proportionally to those who did and multiplied the result ballot-wise.

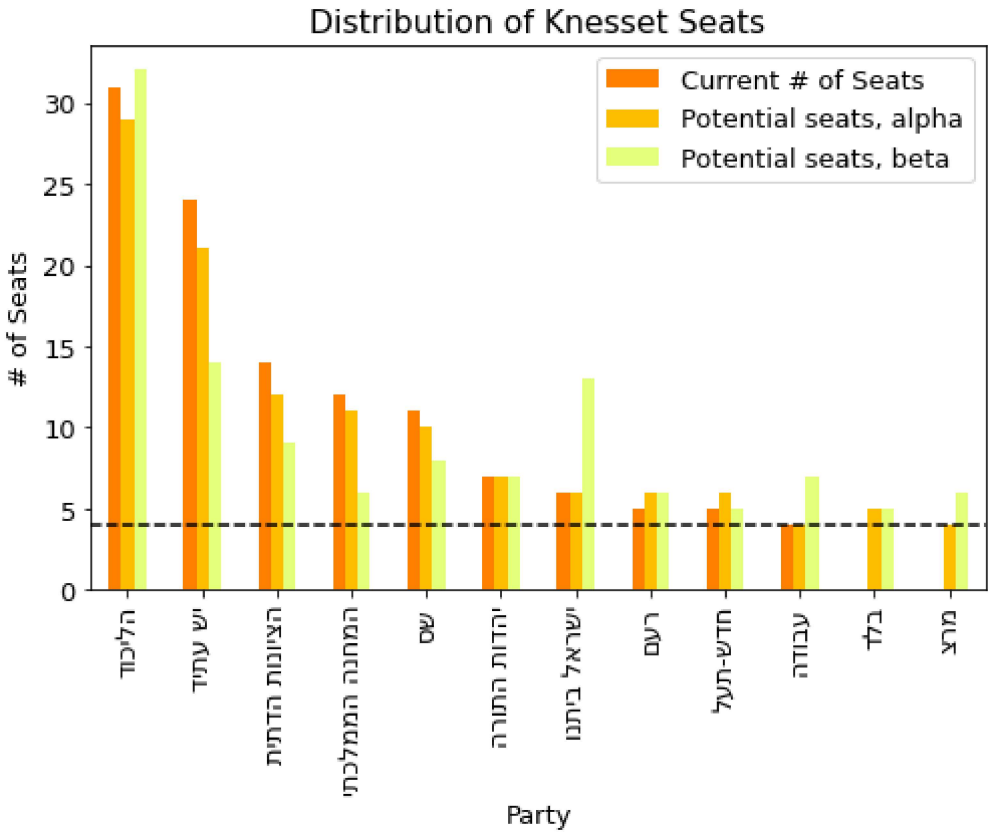
Later, we fixed the results by assuming constant turnout per *party* using the ordinary least squares method of linear regression.

The results of both fixes can be seen in the figure below. The *alpha* fix and the linear regression-based *beta* fix seem to vary in estimating the potential relative support for each party. However, they both agree that Balad passes the electoral threshold and enters the 2022 Knesset. Our

analysis of the *alpha* method even shows that Balad would receive more votes than Labor and Meretz if all potential voters voted.



To further exemplify these fixes we illustrated how the political map would look like by calculating the number of seats that each party would get in comparison to the current distribution of seats in the Knesset.



Several conclusions can be drawn from these two figures above.

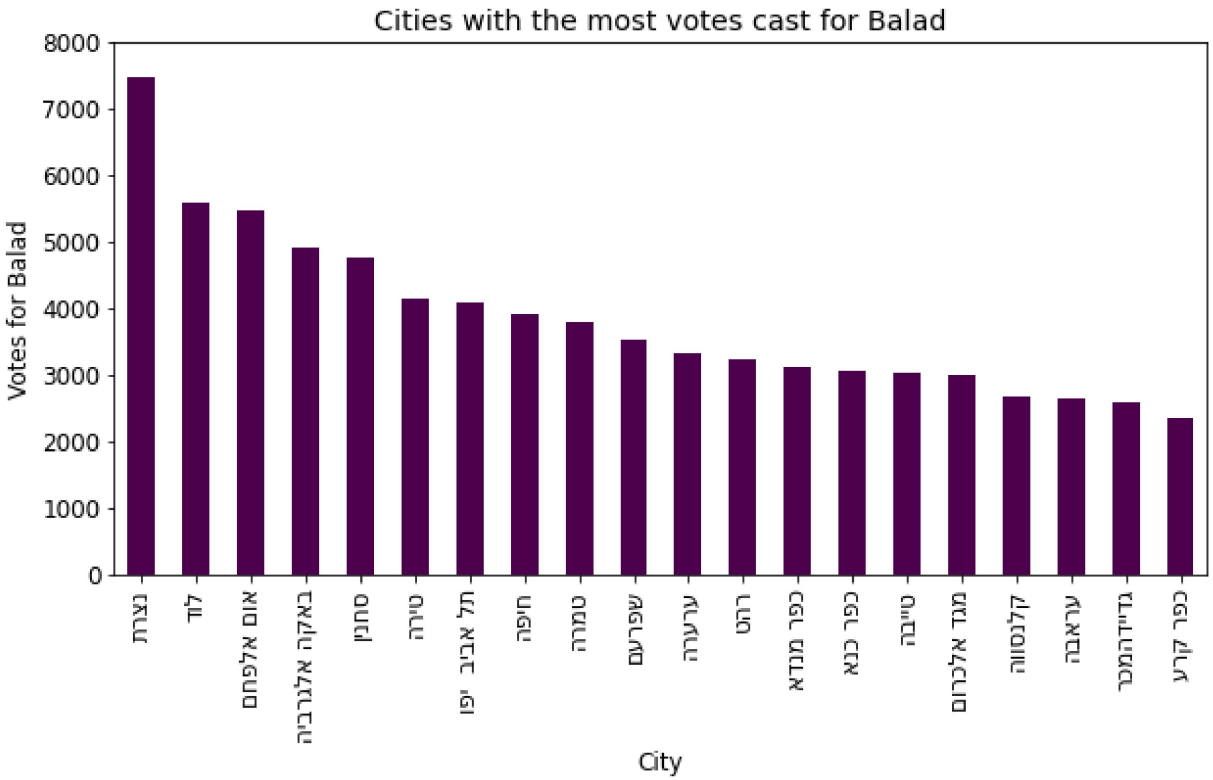
The corrections exemplify the consequences of not voting in Arab neighborhoods and localities on not only Balad, but also Hadash-Taal and Raam which both gained an extra seat. Specifically, it shows how vulnerable Balad is to both the threshold and to voters abstaining from voting.

In addition, Balad's passing the threshold would have caused a loss of seats in the biggest 5 parties and drastically changed the results of the elections; Balad making it into the Knesset would have weakened the pro-Netanyahu bloc and made it harder for them to form a government as they in total have 58 seats based on the *alpha* fix and 56 based on the *beta* fix.

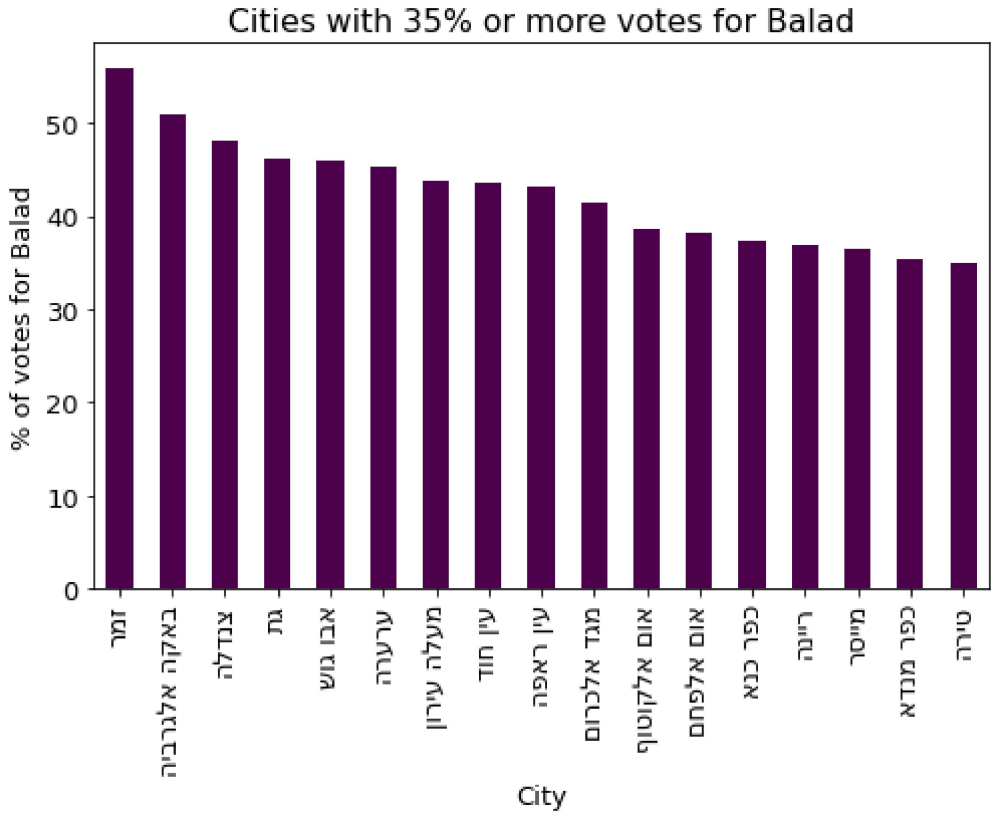
4. Social aspects

1. The variation in the frequency of Balad's voters in different localities and polling stations. Where are the party's strongholds? Where doesn't it have enough voters? (An analysis by socioeconomic ranking)
2. Additional parameters can be shown such as settlement size, Jewish/Arab, secular/religious, peripherality index, per capita income, inequality, etc.

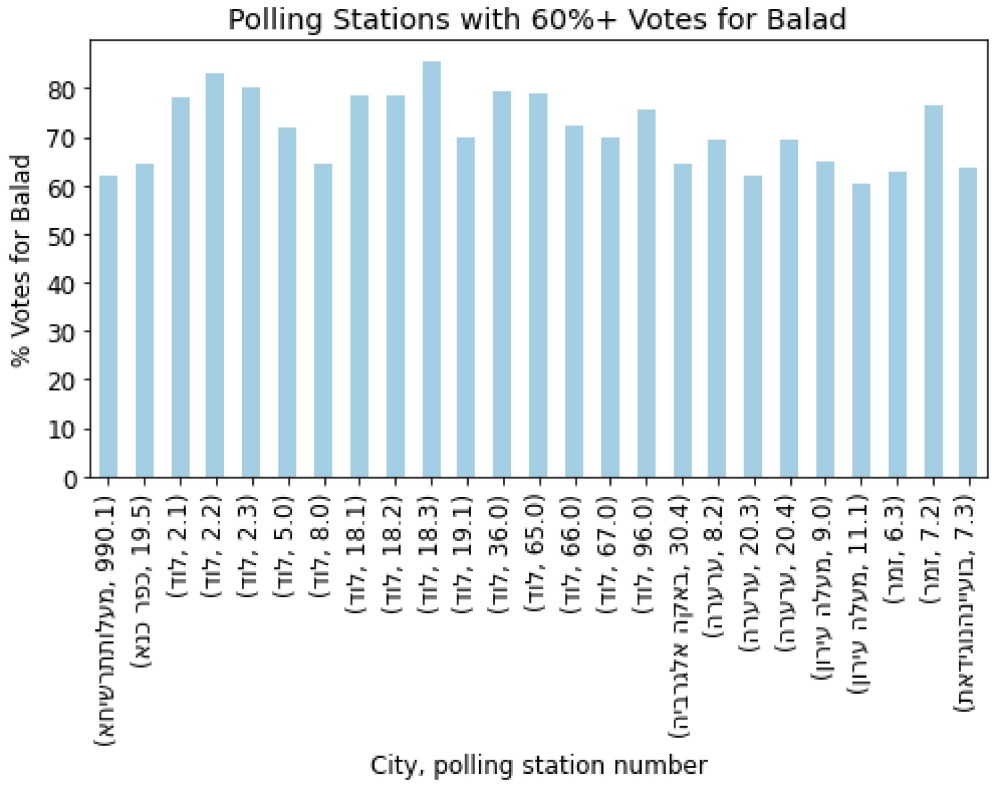
To check what Balad's stronghold cities are, we first checked which cities casted the most votes for Balad. The figure below shows Nazareth leading with almost 7.5k votes cast to Balad, followed by Lod and Umm Al-Fahem with 5.5k each.



Later we checked the cities in which Balad received 35% or more of the total votes. The figure below shows that Balad is in the lead in almost all of these cities, such as Zemer with almost 55% of the votes going to Balad, except a few where they're very close to Hadash-Taal such as Kufr Manda and Tira. Nazareth and Lod cannot be found below as Nazareth has a strong Hadash base and Lod is a mixed Jewish/Arab city. These cities can definitely be considered Balad's strongholds.



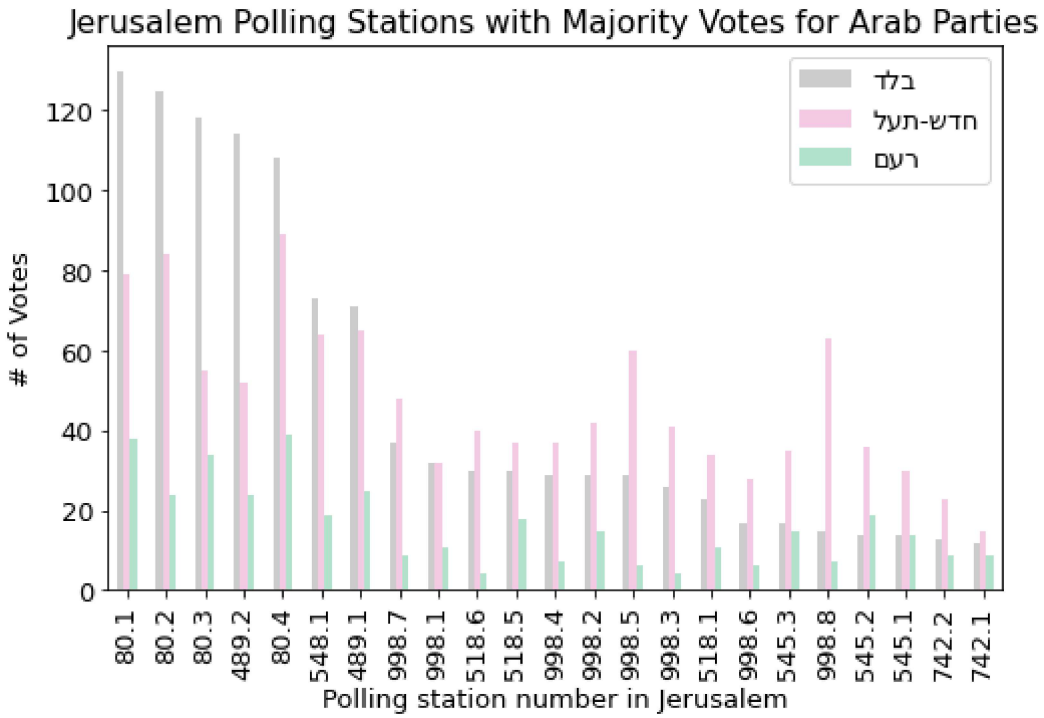
When we checked polling stations, what we found in the first graph regarding Lod was even more distinct as 14 polling stations out of the 25 with 60%+ votes for Balad are from Lod. The polling stations on the x-axis were sorted to further illustrate these polling stations. We deduce that they are polling stations in Arab neighborhoods of Lod, and such an outstanding turnout for Balad in Lod explains its becoming the second biggest party in the city behind Likud.



An interesting case that arose in our analysis of polling stations is Jerusalem. In Jerusalem almost all Palestinian residents who constitute a third of the city's population do not have the right to

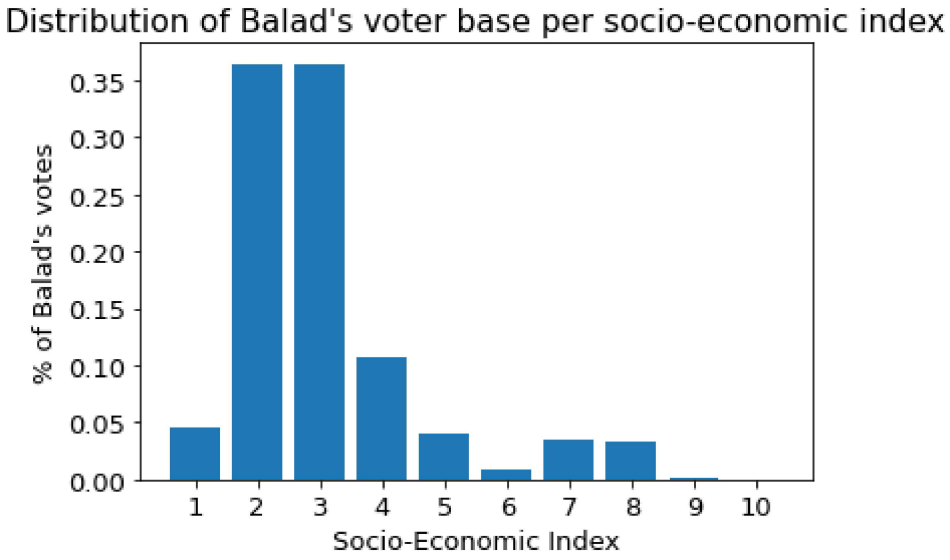
vote. Within those who do, though, 38% of them voted for Balad, 37% for Hadash-Ta'al and the remainder for Ra'am.

Below are the polling stations which we strongly assume are polling stations in Jerusalem's Arab neighborhoods such as Beit Safafa in which Arab parties received the majority of the votes. The distribution of votes in Jerusalem specifically would look much different if all residents of Jerusalem could vote, and if more polling stations existed in the eastern part of the city, which many consider a form of voter suppression. We claim that Jerusalem could potentially be another stronghold city of Balad's.



Out[ ]: (0.38, 0.37)

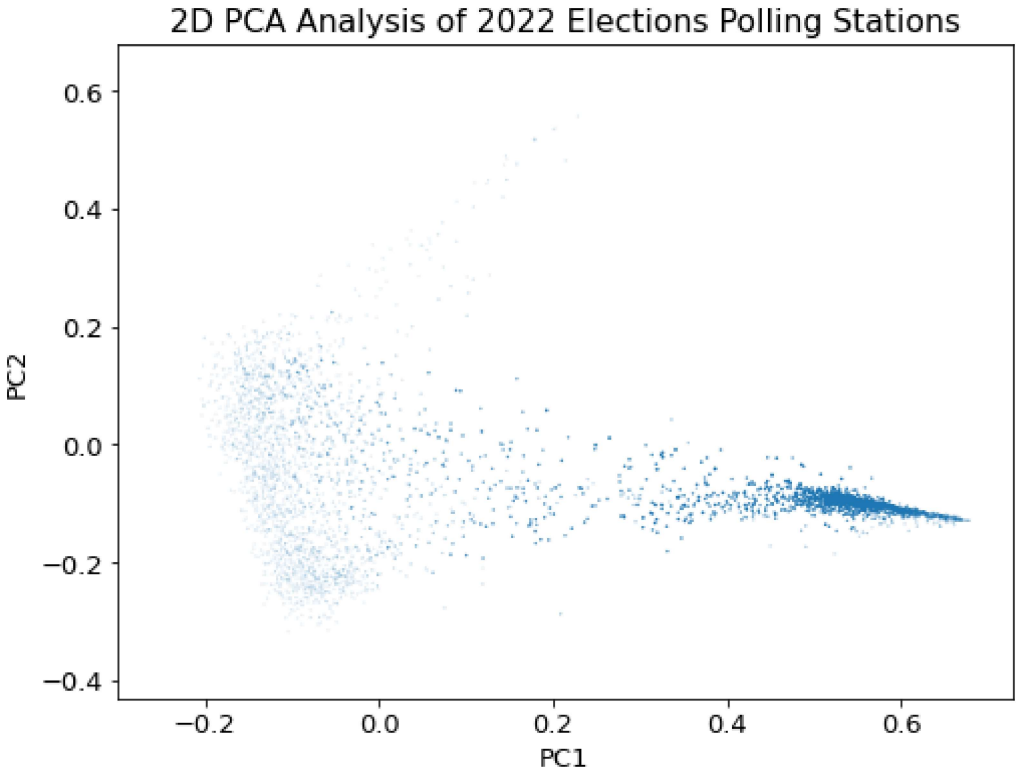
The socio-economic indices of cities and localities in Israel seem to be an important social aspect of Balad's voter base. The figure below shows that more than 70% of Balad's voters live in cities and localities that were given low scores of 2 and 3. This effectively legitimizes Balad's political agenda which aim to improve the living conditions of Palestinian citizens of Israel.



5. PCA

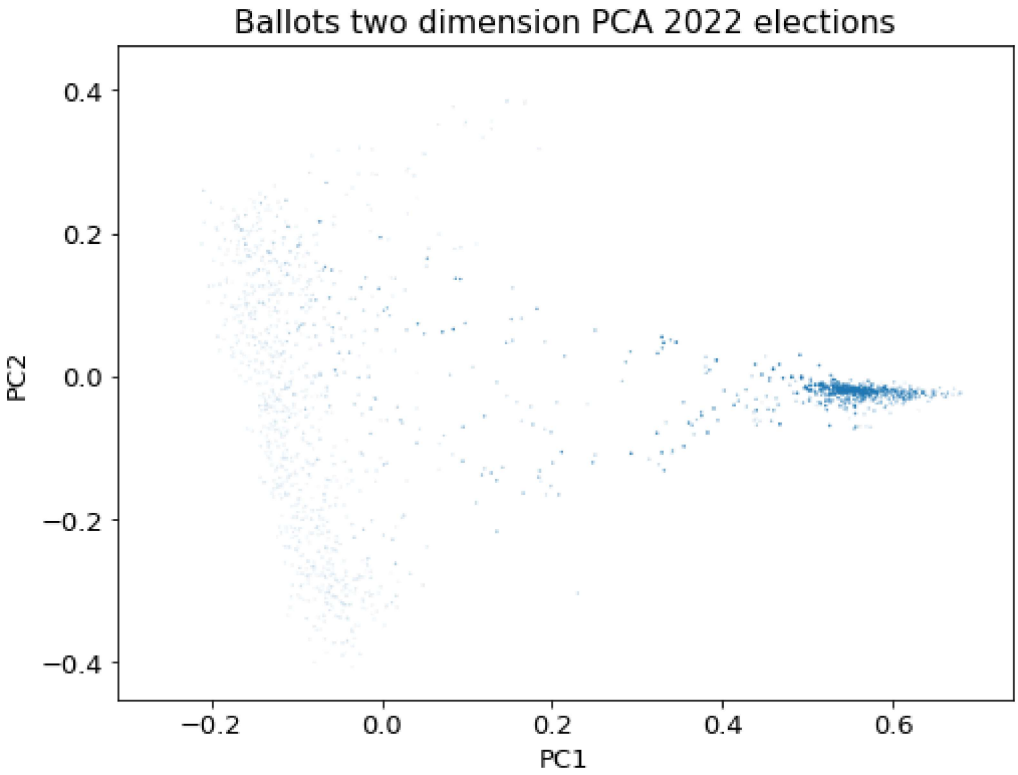
A principal component analysis, an unsupervised method for dimensionality reduction in data, was computed onto the polling stations dataset. The dataset was first scaled so that all features are at the same scale to avoid irregularities due to PCA being a variance maximizing exercise. The PCA figure below has the size of every point (which represents a polling station) decided by the percentage of votes that Balad received.

The graph makes it easy to see the clustering of polling stations that have a high percentage of votes for Balad on the the right side. They are spread horizontally right below 0 which indicates that they are relevant to the first principal component the most. The first principal component may be analyzed as the national background of the ballot station (Jewish/Arab).

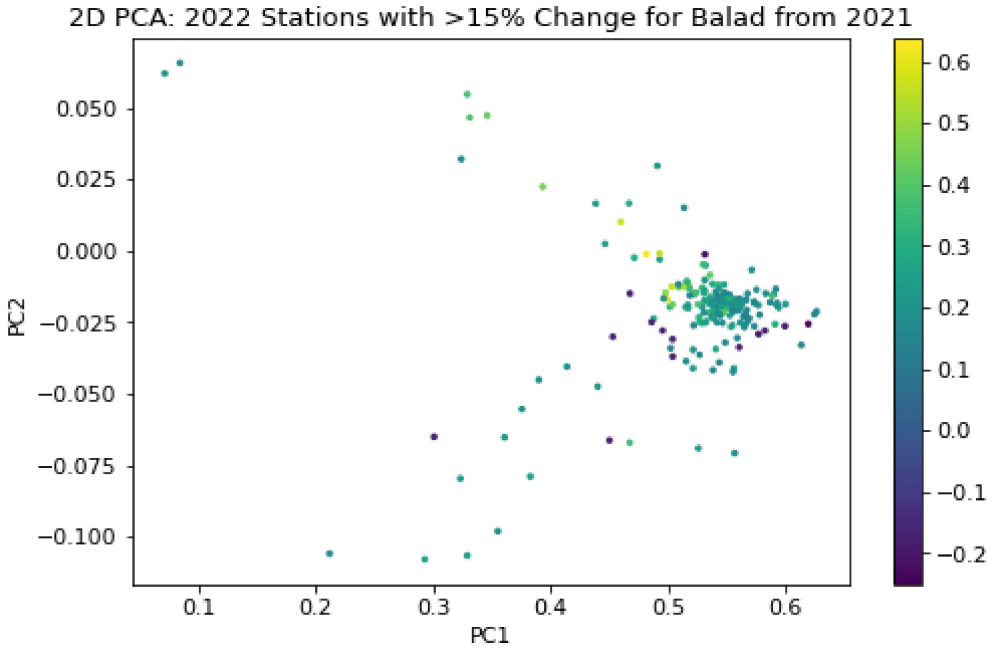


To calculate the differnece in the percentage of votes for Balad in polling stations between 2021 and 2022 we first split the Joint List of 2021 into 60% Hadash-Ta'al and 40% Balad and then

found the shared polling stations between both rounds of elections. There were almost 7k such stations. We then performed a PCA analysis onto this data set and result was the figure below. We got a very similar graph to the one before, where those with the highest difference were clustered together to the right as well. The scale of the second principal component was different (-0.4:0.6 > -0.4:0.4) which makes sense since both graphs do not have the exact same polling stations, but the horizontal scale and clustering about the zero x-axis to the right is remained unchanged, further strengthening our previous analysis.



We also extracted all the polling stations with 15+% increase/decrease compared to 2021 elections. We can see that those clustered together are colored green indicating a ~30% increase, and some are even colored yellow which indicates a ~50+% increase, while those with a decrease in percentage are negligible. The dimensions are even clearer in that the clustering is on the right and is about the zero x-axis.



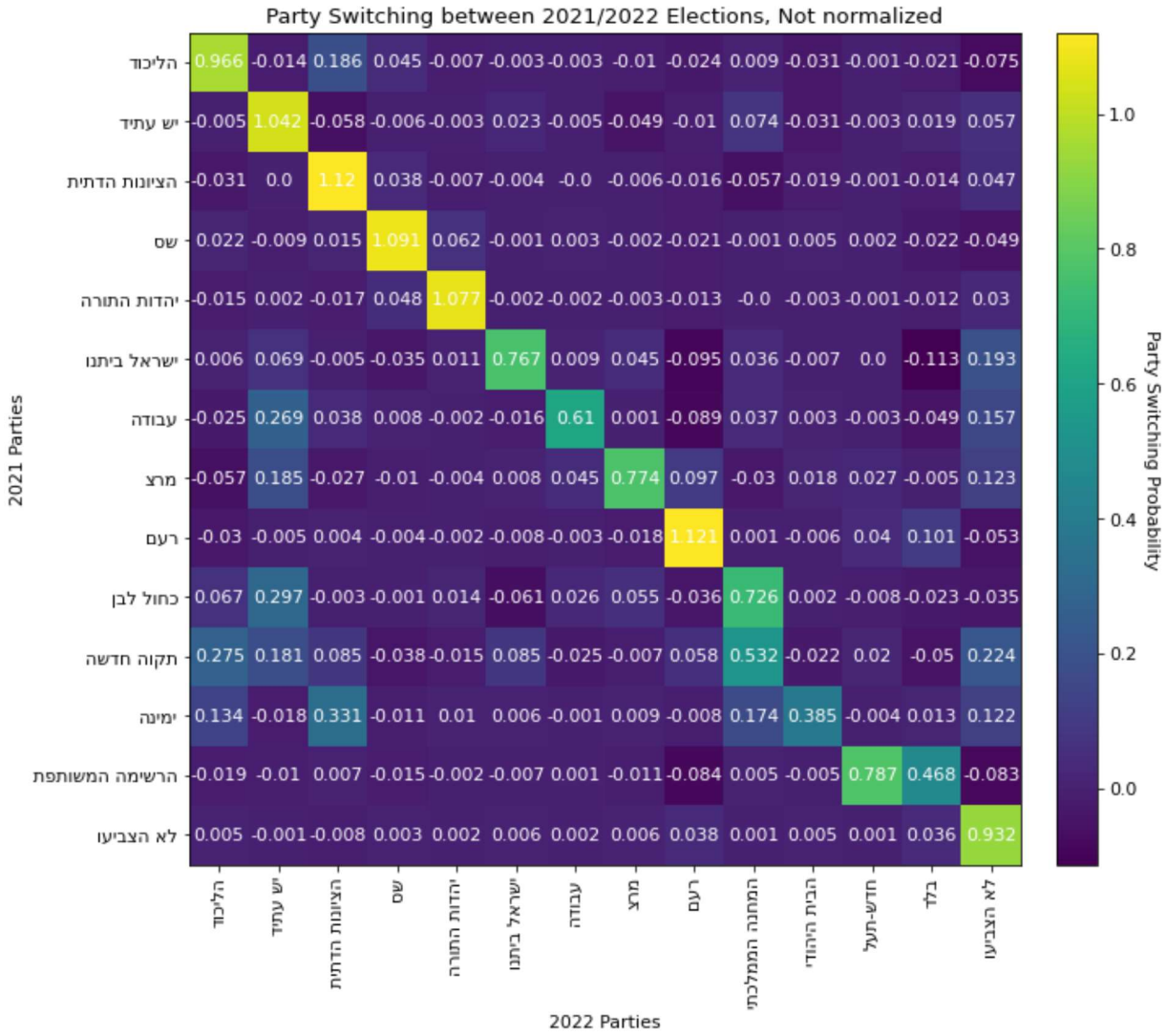
6 + 7 + 8 Party Switching

In this section we computed the M matrix which minimizes the MSE loss of the following equation:

$M^* = argmin_M||N^{(a)}M - N^{(b)}||_F^2$, wherein $N^{(a)}$ and $N^{(b)}$ are the values of the datasets of each elections cycle which include the shared polling stations between them.

M^* essentially contains the data on party switching, or the difference in each identical ballot between 2021 and 2022. Since not much time passed between the two elections, we assume that the same people who voted in that specific station in 2021 are the same who voted in 2022. Code-wise, we used a function *linalg.pinv* which computes the optimal parameters $\hat{\theta}$ that minimize the expression above.

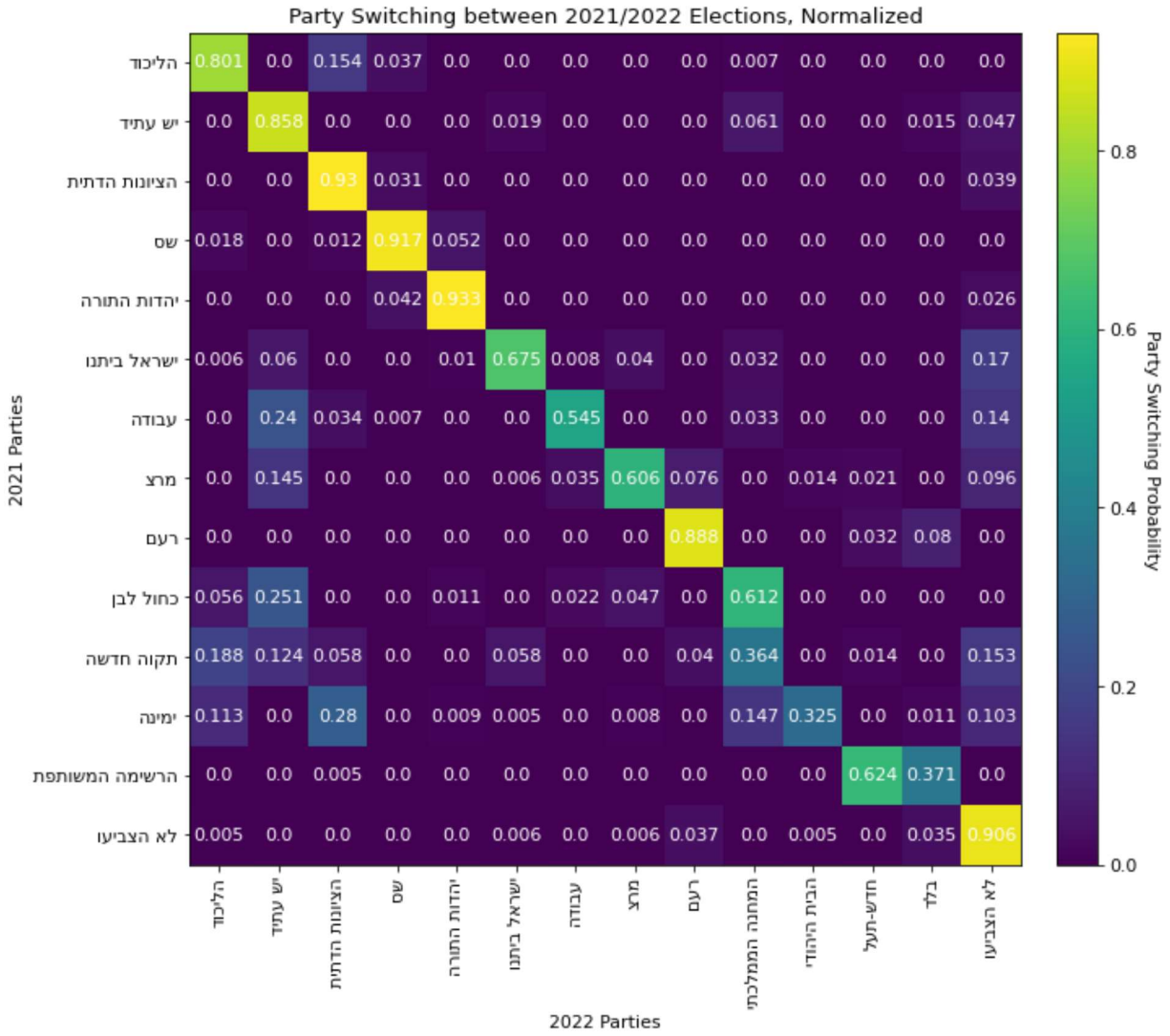
The heatmap below presents the M^* values, without normalization. It shows the party-switching values between each pair of parties. The first five values on the diagonal in addition to Ra'am are all colored in shades of yellow, with most of them exceeding 1. This is sensible as these parties all gained seats compared to 2021, meaning they received more votes.



With respect to Balad, we standardized M^* and zeroed all party-switching values below 0.5% as they are considered negligible in order to infer about party-switching.

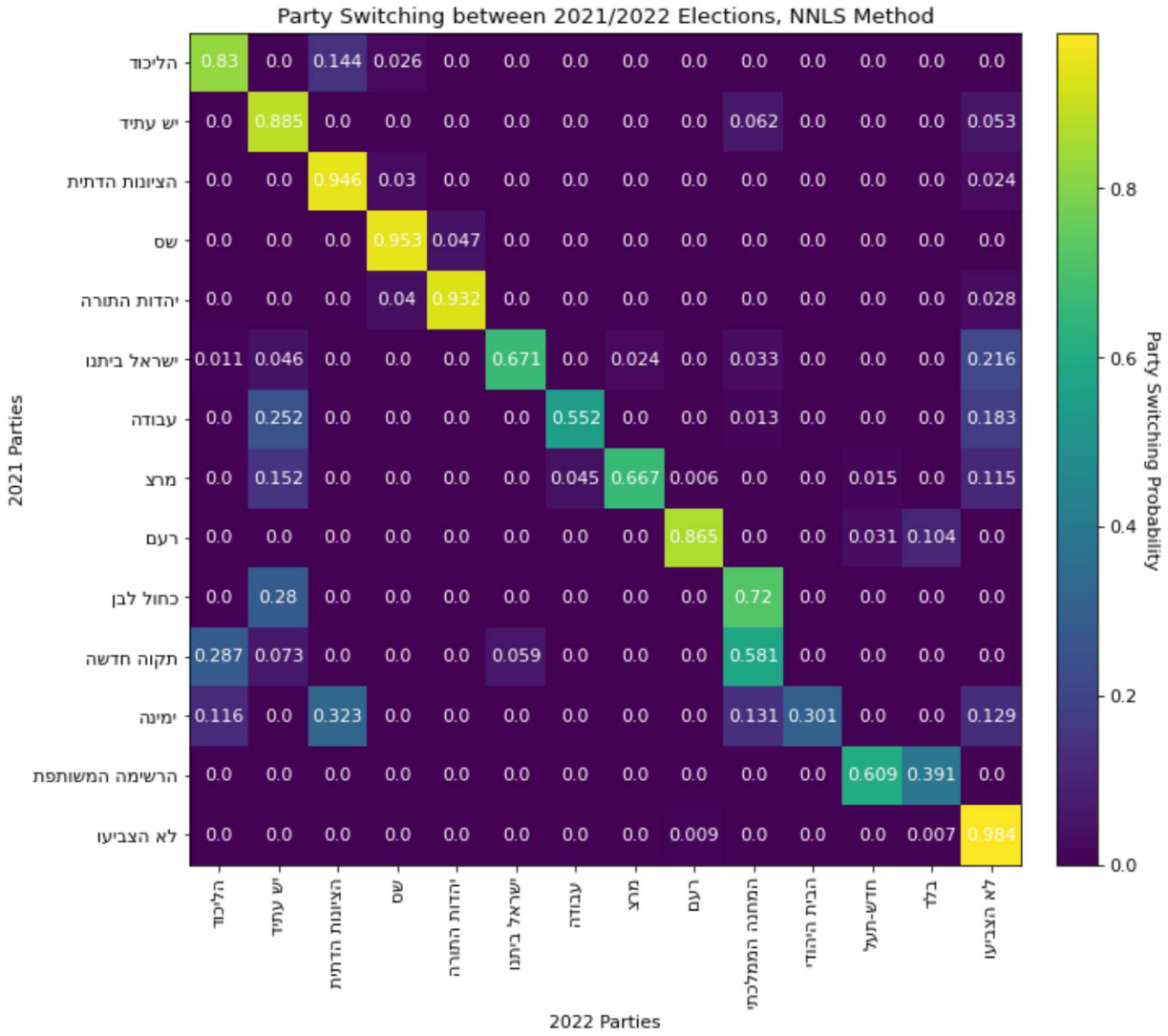
Arab parties showed interesting patterns themselves: the voters of the leftist/secular Joint List were split in a ratio of 0.63-to-0.37 between Hadash-Ta'al and Balad respectively, with no remainder going to the Islamist UAL, while UAL showed little party-switching to Balad mostly and then Hadash-Taal.

And as mentioned before, Hadash-Ta'al and Balad experienced a 50% increase of votes from 2021, which explains the zero of any JL voters abstaining from voting in 2022. The value is even negative as can be seen in the heatmap above.

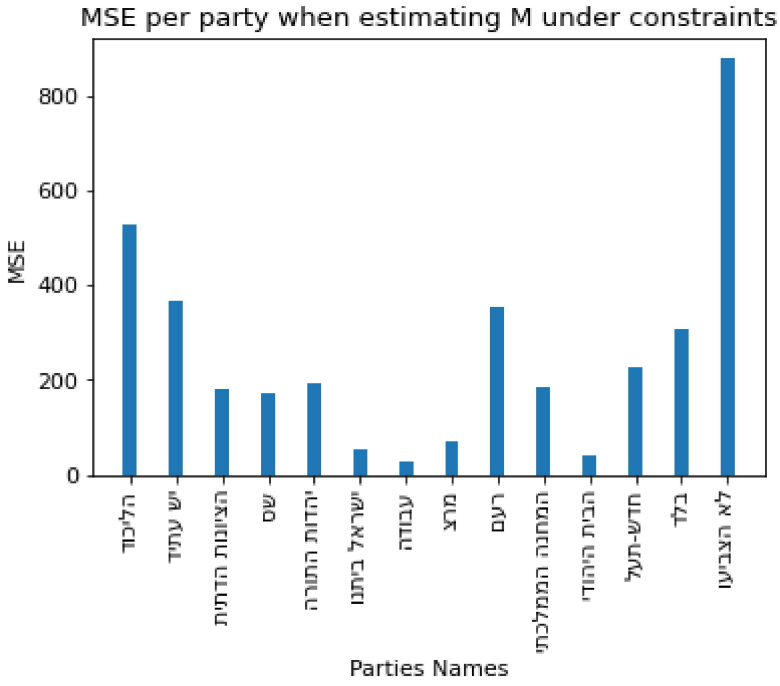


An additional calculation of the coefficients matrix using the non-negative least squares (NNLS) can be seen below. The NNLS method introduces a nonnegativity constraint which constrains our optimization problem defined above. With respect to Balad, the NNLS seems to split the JL voter base into 0.61-0.39 for Hadash-Ta'al and Balad respectively, and raised the party-switching ratio from Ra'am to Balad to 10% compared to the 8% of the OLS method. We conclude that Balad's voter base mainly came from the Joint List and secondarily from Ra'am.

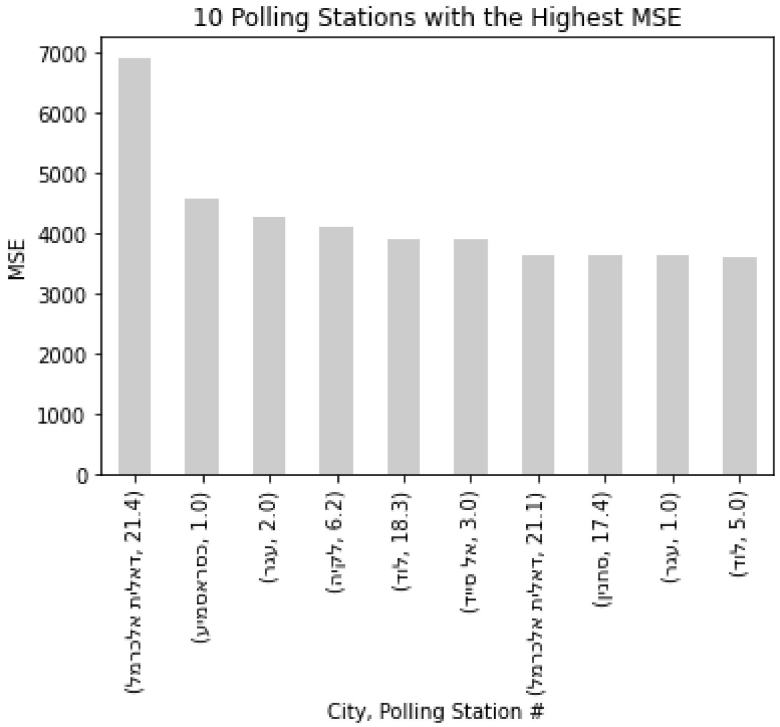
Since Balad re-emerged only this year, nothing exactly can be said about party-switching from it to other parties.



To check how well our model performed, we looked into the residuals and and the mean squared error (MSE) by squaring and summing the residuals. The MSE seems to be the highest in bigger parties; it could not predict accurately "non-voters", Likud, and Yesh Atid. The model seems to be weak in predicting Arab parties as well, specifically Ra'am and Balad, which is sensible as Ra'am was the biggest winner out of the 3 parties and Balad reached an all-time record in terms of # of votes.

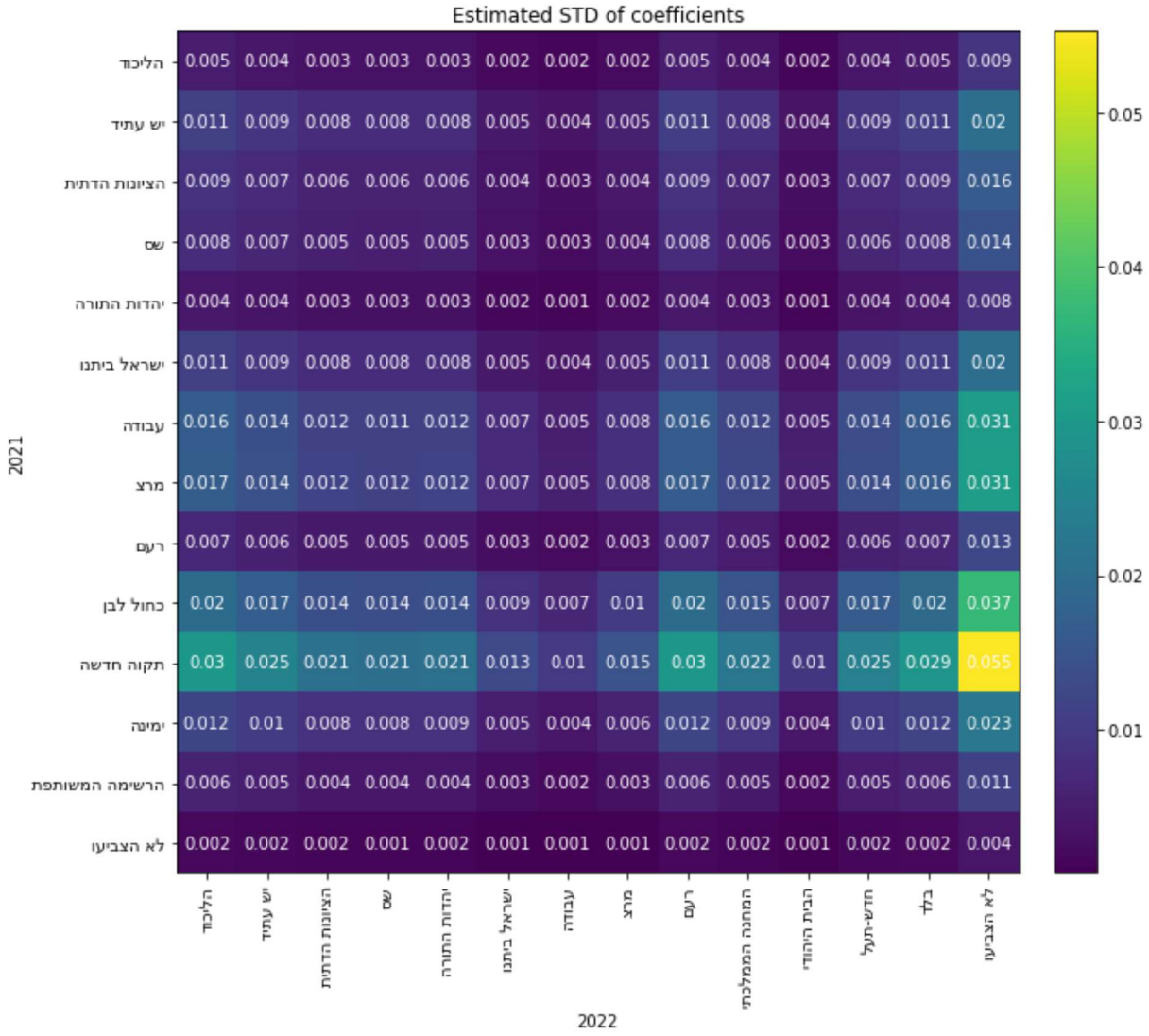


We looked into polling stations with the highest MSE, in the figure below. To explain these results, we noticed that A-Sayed, Daliyat al-Karmel, Kisra-Smei', and Lod all experienced significantly less numbers of non-voters, while the Sakhnin, Likya and Ghajar had an opposite shift of greater non-voters in 2022. In addition, UAL got stronger in Bedouin cities like Likya and A-Sayyed and weaker in Sakhnin in favor of Balad and Hadash-Taal. The Druze cities of Kisra Smei' and Daliyat al-Karmel casted significantly greater votes to the National Camp coupled with a significant increase in their turnout. All these changes in voting patterns might have strongly affected the MSE.



To determine which vote transfers from/to the party are significant, we calculated the standard deviation of our non-normalized coefficient, followed by the t-statistic and the p-values.

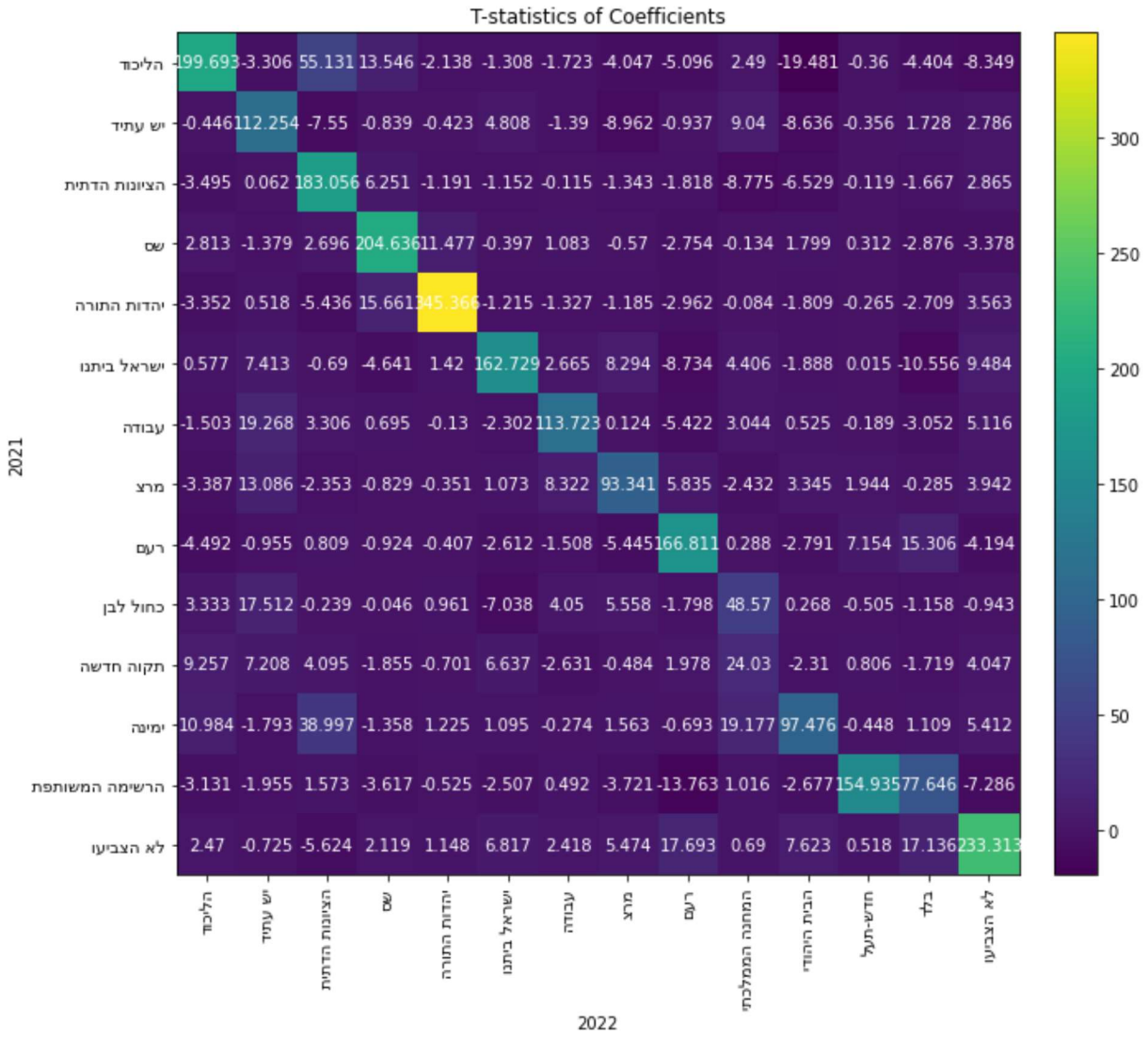
The standard deviation of the coefficients in the matrix implies that the correlation or association between the main sources of Balad's voters, namely the Joint List, non-voters, and transitioned voters from Ra'am, is consistent with minimal variation at these percentages. This is evident from the small deviation of the percentages of these sources, which are (0.6, 0.2, 0.7)% respectively.



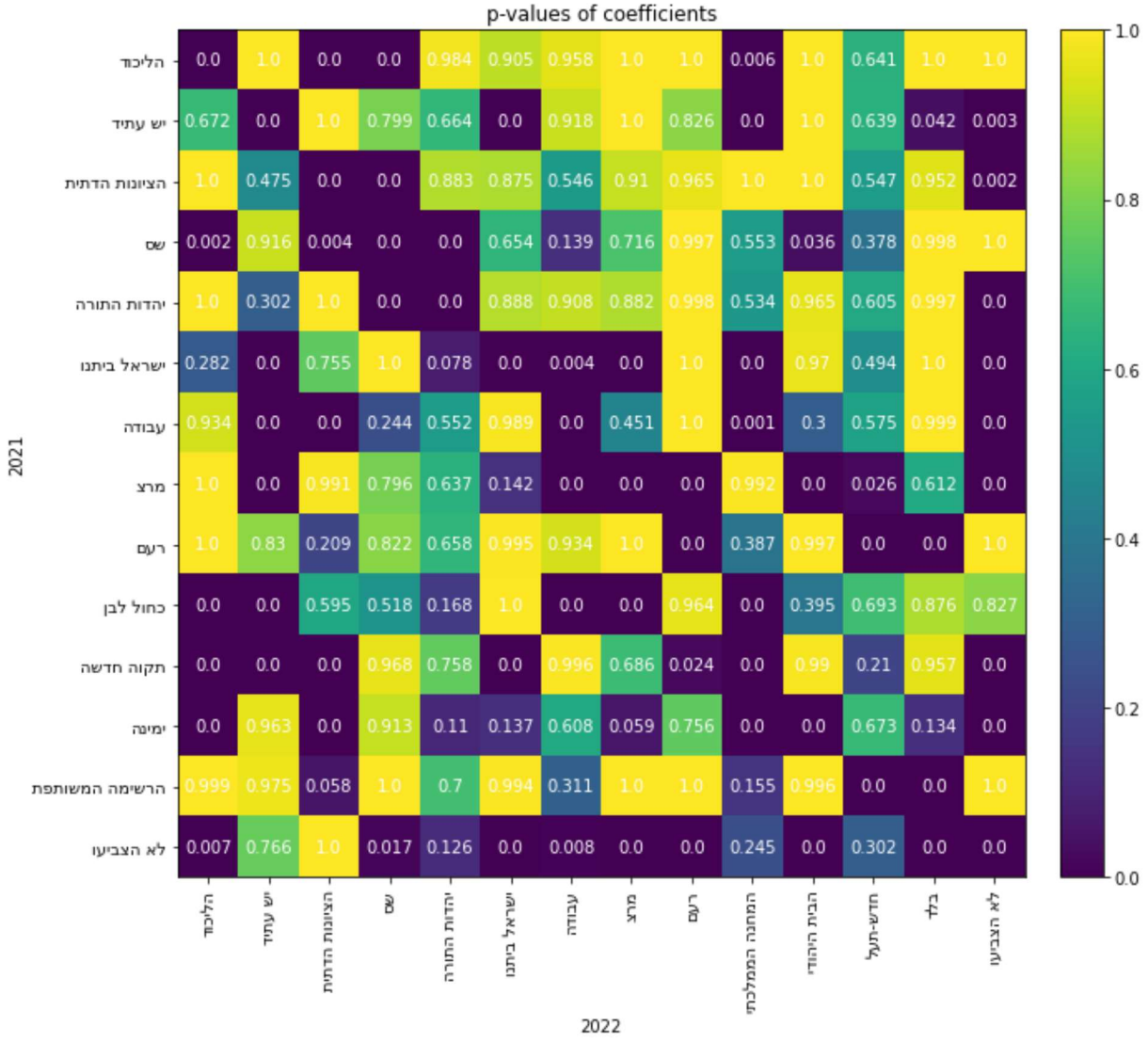
Given the assumptions about voter behavior, we are conducting a one-tailed t-statistic test to determine if there is a correlation between each party's voters and any other party from previous elections. Specifically, the null hypothesis states that there is no correlation, and the coefficient is equal to zero. The alternative hypothesis is that the coefficient is not equal to zero, and is significant.

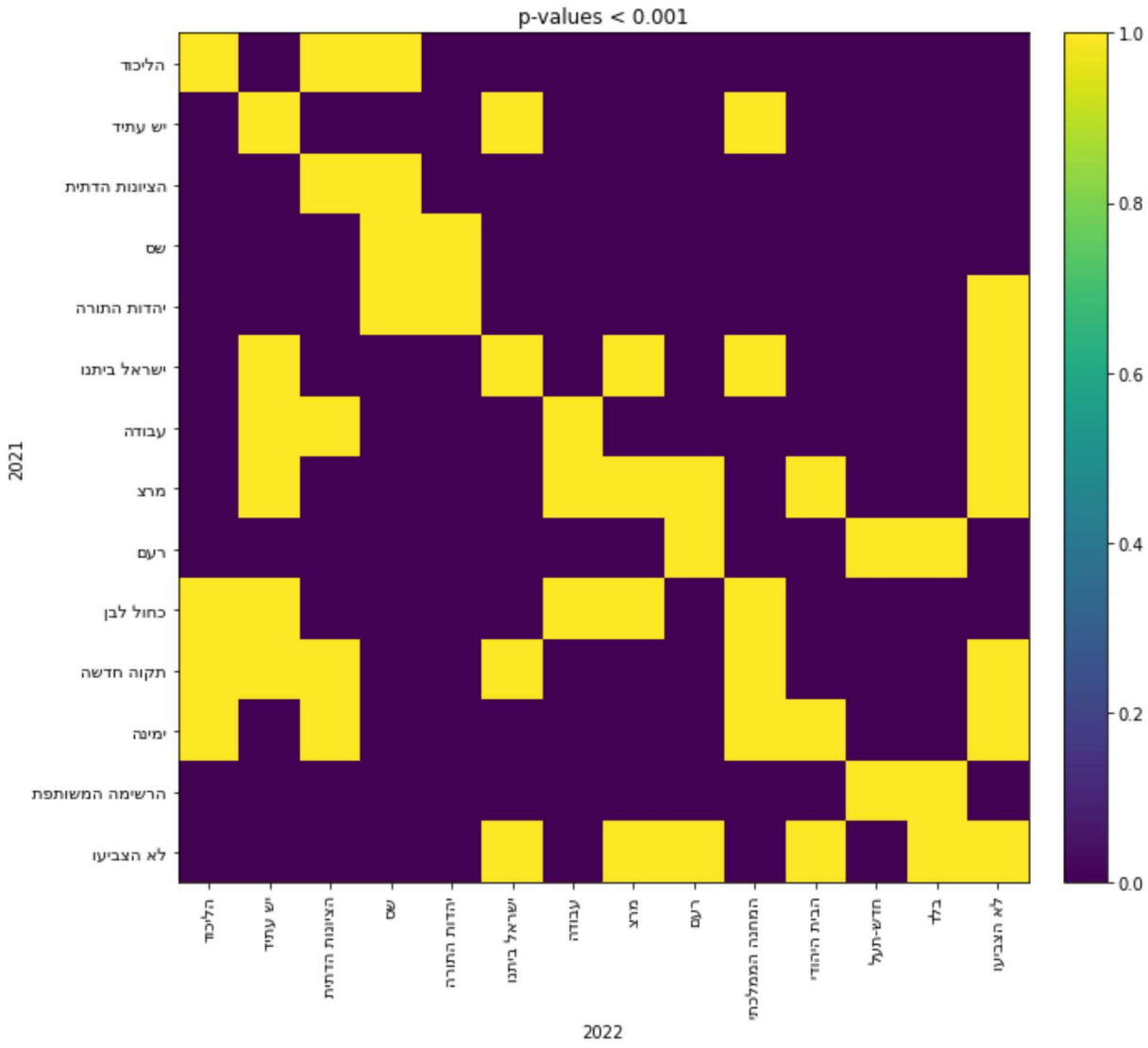
We have a 13 x 13 matrix representing real-world values, not based on estimations. This means we have 169 degrees of freedom. According to this test, any values above 3.345 falls within the 99.9% confidence level.

By observing the matrix, this test allows us to confirm that Balad's votes came from the sources we previously identified. It's important to note that the values and the confidence level mentioned in the above explanation are arbitrary and is dependent on the actual dataset and the specific test being used.



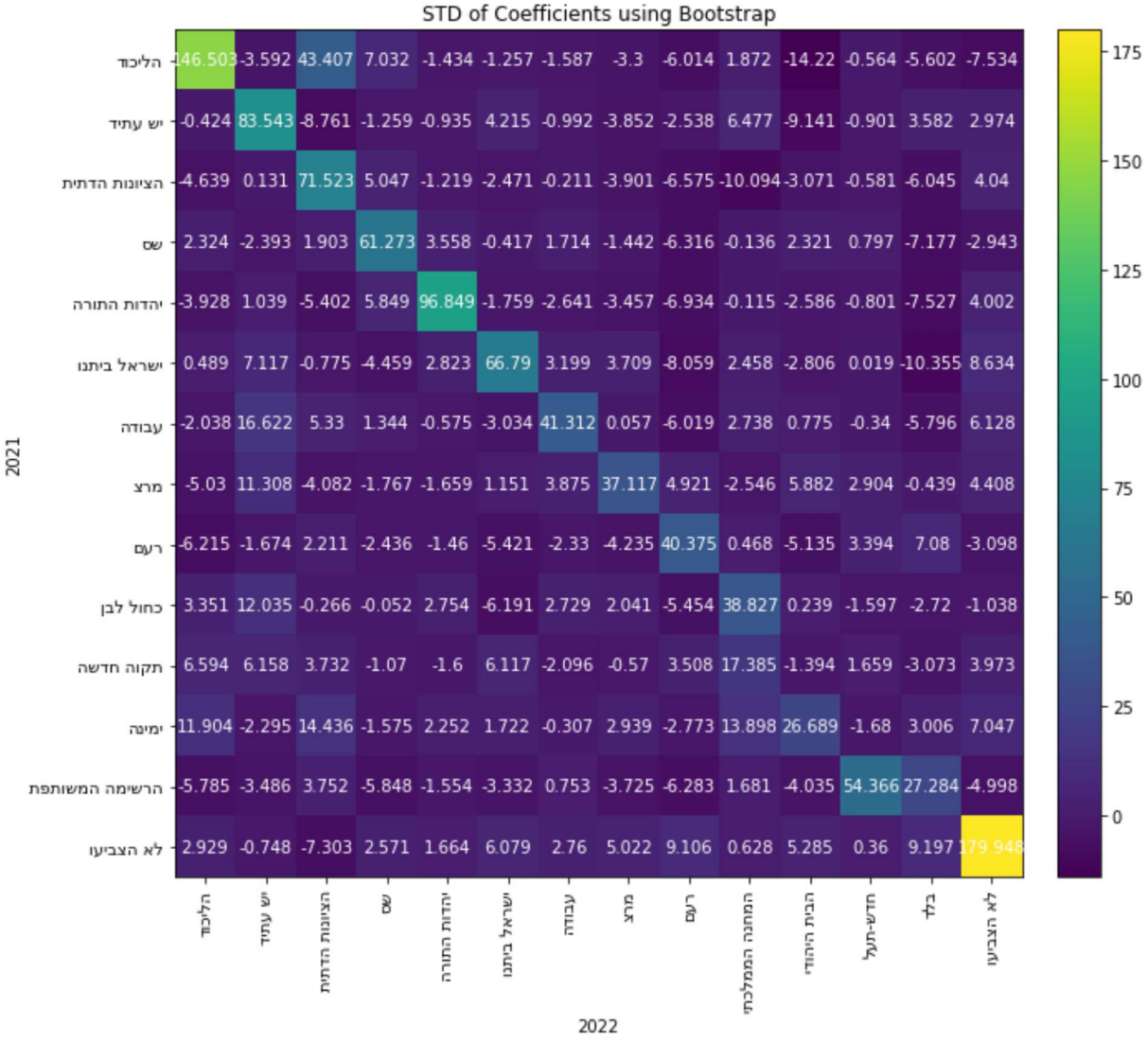
Analyzing the p-value matrices gives us additional assurance that the coefficients we have calculated are statistically significant. With a 99.9% confidence level, we can confidently reject the null hypothesis that the coefficients are zero and conclude that there is a meaningful association between the parties' voters and those from previous elections. This suggests that Balad's voter base has a clear correlation with the voters of other parties (JL, Ra'am , Non-voters) in the previous election.

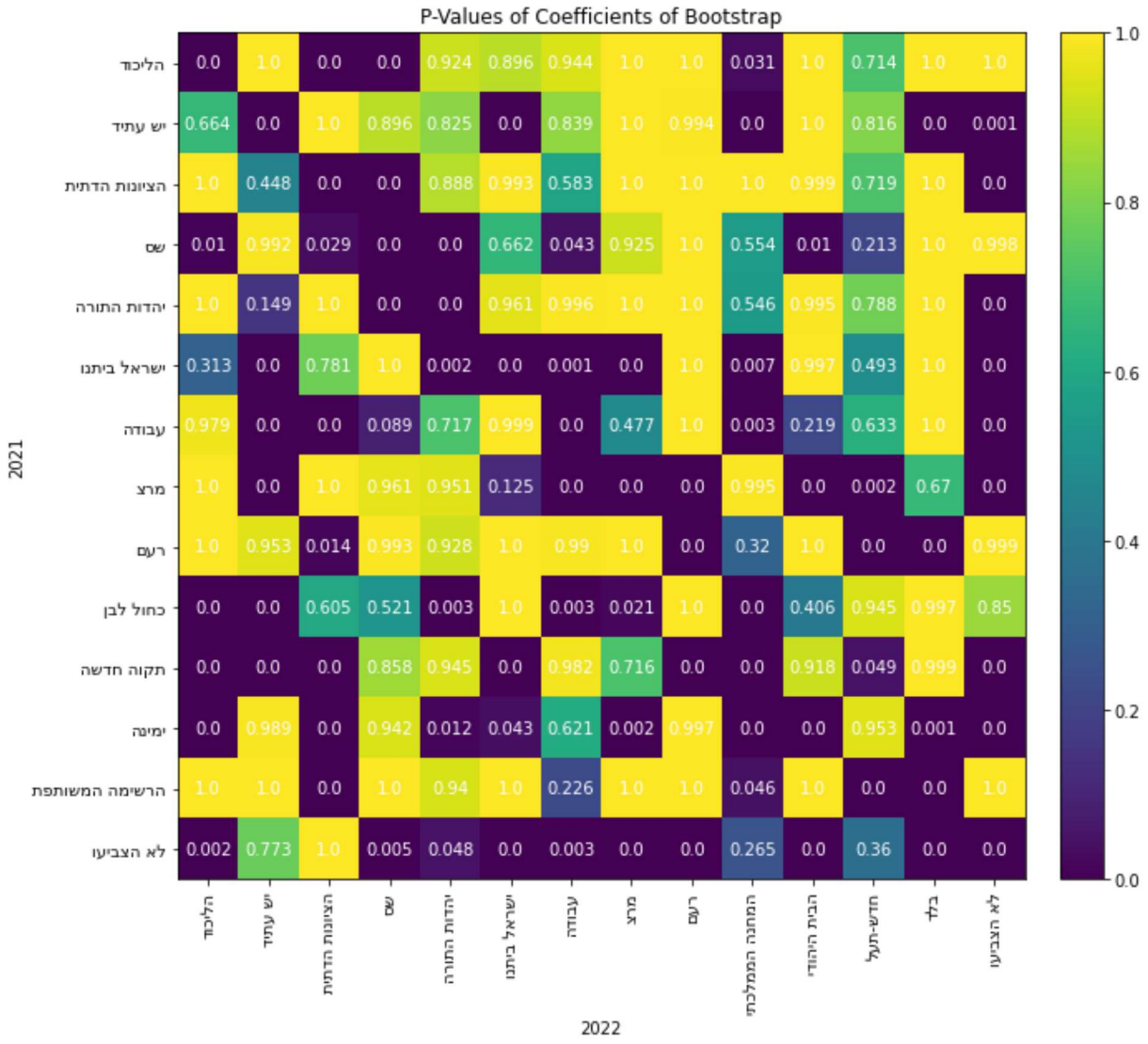

p-values of coefficients
2021
2022
הליכוד
יש עתיד
הציונות הדתית
שס
יהדות התורה
ישראל ביתנו
עבודה
רעם
כחול לבן
תקוה חדשה
ימינה
הרשימה המשותפת
לא הצביעו
מרצ
המחנה הממלכתי
הבית היהודי
חדש-תעל
בלד
0.0 1.0 0.0 0.0 0.984 0.905 0.958 1.0 1.0 0.006 1.0 0.641 1.0 1.0
0.672 0.0 1.0 0.799 0.664 0.0 0.918 1.0 0.826 0.0 1.0 0.639 0.042 0.003
1.0 0.475 0.0 0.0 0.883 0.875 0.546 0.91 0.965 1.0 1.0 0.547 0.952 0.002
0.002 0.916 0.004 0.0 0.0 0.654 0.139 0.716 0.997 0.553 0.036 0.378 0.998 1.0
1.0 0.302 1.0 0.0 0.0 0.888 0.908 0.882 0.998 0.534 0.965 0.605 0.997 0.0
0.282 0.0 0.755 1.0 0.078 0.0 0.004 0.0 1.0 0.0 0.97 0.494 1.0 0.0
0.934 0.0 0.0 0.244 0.552 0.989 0.0 0.451 1.0 0.001 0.3 0.575 0.999 0.0
1.0 0.0 0.991 0.796 0.637 0.142 0.0 0.0 0.0 0.992 0.0 0.026 0.612 0.0
1.0 0.83 0.209 0.822 0.658 0.995 0.934 1.0 0.0 0.387 0.997 0.0 0.0 1.0
0.0 0.0 0.595 0.518 0.168 1.0 0.0 0.0 0.964 0.0 0.395 0.693 0.876 0.827
0.0 0.0 0.0 0.968 0.758 0.0 0.996 0.686 0.024 0.0 0.99 0.21 0.957 0.0
0.0 0.963 0.0 0.913 0.11 0.137 0.608 0.059 0.756 0.0 0.0 0.673 0.134 0.0
0.999 0.975 0.058 1.0 0.7 0.994 0.311 1.0 1.0 0.155 0.996 0.0 0.0 1.0
0.007 0.766 1.0 0.017 0.126 0.0 0.008 0.0 0.0 0.245 0.0 0.302 0.0 0.0
1.0
0.8
0.6
0.4
0.2
0.0

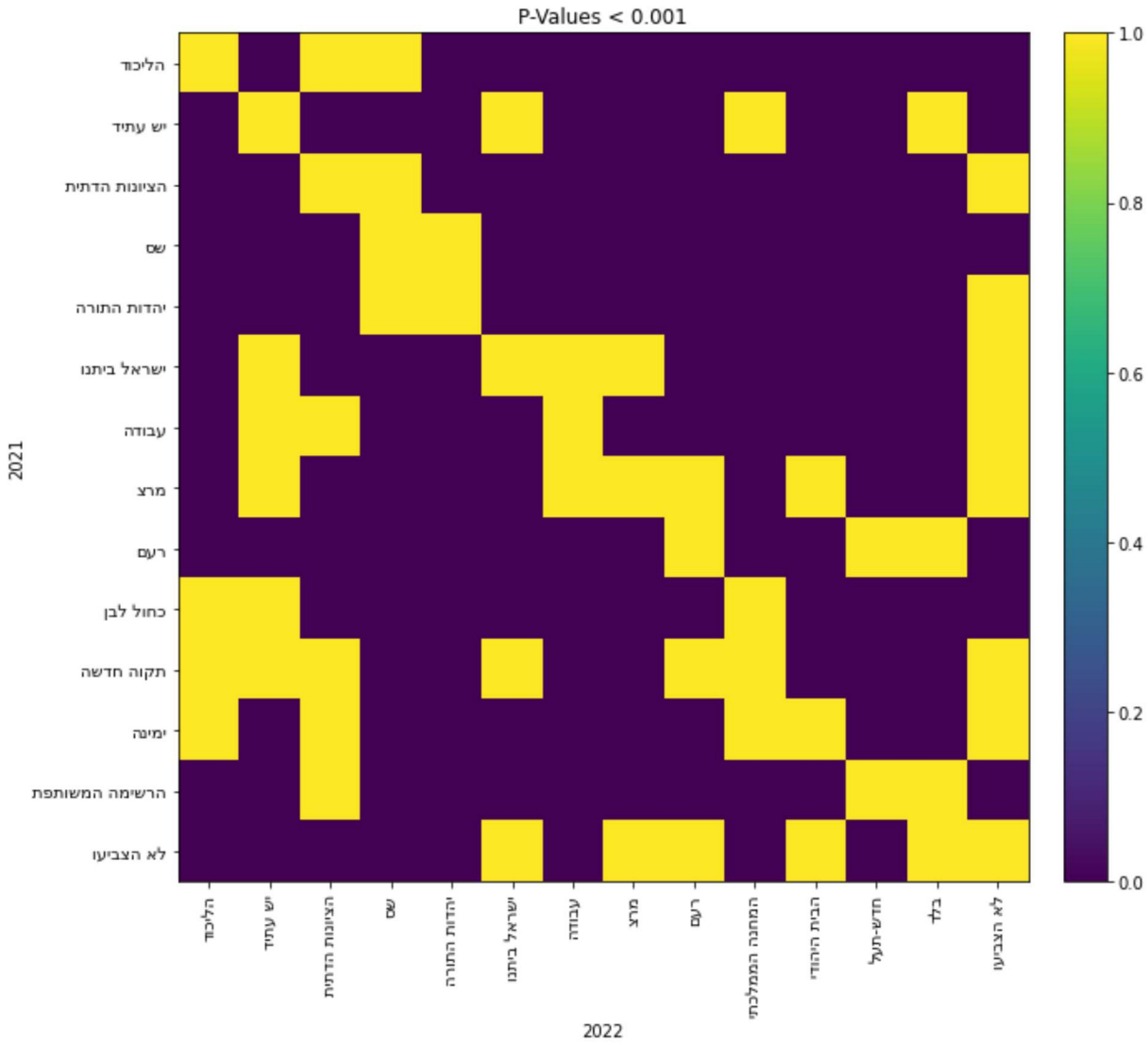


To test how good the model predictions of vote transfers are for the party relative to other parties, we applied the bootstrap method in addition to dividing into train/test sets.

Even after using the bootstrap method, a technique that involves repeatedly drawing random samples with replacement from the original dataset and calculating the statistic of interest, we still obtained the same results. This further confirms our previous conclusions and adds an extra layer of robustness to our analysis. The figures of STD, t-statitics, and p-values following bootstrapping can be found below.

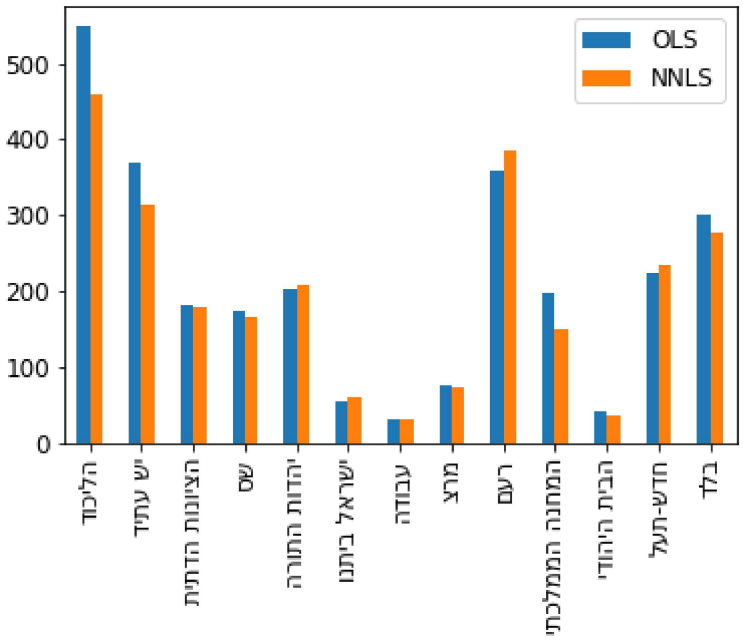

STD of Coefficients using Bootstrap
2021
2022
הליכוד
יש עתיד
הציונות הדתית
שס
יהדות התורה
ישראל ביתנו
עבודה
מרצ
רעם
כחול לבן
תקוה חדשה
ימינה
הרשימה המשותפת
לא הצביעו
146.503 -3.592 43.407 7.032 -1.434 -1.257 -1.587 -3.3 -6.014 1.872 -14.22 -0.564 -5.602 -7.534
-0.424 83.543 -8.761 -1.259 -0.935 4.215 -0.992 -3.852 -2.538 6.477 -9.141 -0.901 3.582 2.974
-4.639 0.131 71.523 5.047 -1.219 -2.471 -0.211 -3.901 -6.575 -10.094 -3.071 -0.581 -6.045 4.04
2.324 -2.393 1.903 61.273 3.558 -0.417 1.714 -1.442 -6.316 -0.136 2.321 0.797 -7.177 -2.943
-3.928 1.039 -5.402 5.849 96.849 -1.759 -2.641 -3.457 -6.934 -0.115 -2.586 -0.801 -7.527 4.002
0.489 7.117 -0.775 -4.459 2.823 66.79 3.199 3.709 -8.059 2.458 -2.806 0.019 -10.355 8.634
-2.038 16.622 5.33 1.344 -0.575 -3.034 41.312 0.057 -6.019 2.738 0.775 -0.34 -5.796 6.128
-5.03 11.308 -4.082 -1.767 -1.659 1.151 3.875 37.117 4.921 -2.546 5.882 2.904 -0.439 4.408
-6.215 -1.674 2.211 -2.436 -1.46 -5.421 -2.33 -4.235 40.375 0.468 -5.135 3.394 7.08 -3.098
3.351 12.035 -0.266 -0.052 2.754 -6.191 2.729 2.041 -5.454 38.827 0.239 -1.597 -2.72 -1.038
6.594 6.158 3.732 -1.07 -1.6 6.117 -2.096 -0.57 3.508 17.385 -1.394 1.659 -3.073 3.973
11.904 -2.295 14.436 -1.575 2.252 1.722 -0.307 2.939 -2.773 13.898 26.689 -1.68 3.006 7.047
-5.785 -3.486 3.752 -5.848 -1.554 -3.332 0.753 -3.725 -6.283 1.681 -4.035 54.366 27.284 -4.998
2.929 -0.748 -7.303 2.571 1.664 6.079 2.76 5.022 9.106 0.628 5.285 0.36 9.197 179.948
הליכוד
יש עתיד
הציונות הדתית
שס
יהדות התורה
ישראל ביתנו
עבודה
מרצ
רעם
המחנה הממלכתי
הבית היהודי
חדש-תעל
בלד
לא הצביעו
175
150
125
100
75
50
25
0


P-Values of Coefficients of Bootstrap
2021
2022
הליכוד
יש עתיד
הציונות הדתית
שס
יהדות התורה
ישראל ביתנו
עבודה
מרצ
רעם
כחול לבן
תקוה חדשה
ימינה
הרשימה המשותפת
לא הצביעו
0.0 1.0 0.0 0.0 0.924 0.896 0.944 1.0 1.0 0.031 1.0 0.714 1.0 1.0
0.664 0.0 1.0 0.896 0.825 0.0 0.839 1.0 0.994 0.0 1.0 0.816 0.0 0.001
1.0 0.448 0.0 0.0 0.888 0.993 0.583 1.0 1.0 1.0 0.999 0.719 1.0 0.0
0.01 0.992 0.029 0.0 0.0 0.662 0.043 0.925 1.0 0.554 0.01 0.213 1.0 0.998
1.0 0.149 1.0 0.0 0.0 0.961 0.996 1.0 1.0 0.546 0.995 0.788 1.0 0.0
0.313 0.0 0.781 1.0 0.002 0.0 0.001 0.0 1.0 0.007 0.997 0.493 1.0 0.0
0.979 0.0 0.0 0.089 0.717 0.999 0.0 0.477 1.0 0.003 0.219 0.633 1.0 0.0
1.0 0.0 1.0 0.961 0.951 0.125 0.0 0.0 0.0 0.995 0.0 0.002 0.67 0.0
1.0 0.953 0.014 0.993 0.928 1.0 0.99 1.0 0.0 0.32 1.0 0.0 0.0 0.999
0.0 0.0 0.605 0.521 0.003 1.0 0.003 0.021 1.0 0.0 0.406 0.945 0.997 0.85
0.0 0.0 0.0 0.858 0.945 0.0 0.982 0.716 0.0 0.0 0.918 0.049 0.999 0.0
0.0 0.989 0.0 0.942 0.012 0.043 0.621 0.002 0.997 0.0 0.0 0.953 0.001 0.0
1.0 1.0 0.0 1.0 0.94 1.0 0.226 1.0 1.0 0.046 1.0 0.0 0.0 1.0
0.002 0.773 1.0 0.005 0.048 0.0 0.003 0.0 0.0 0.265 0.0 0.36 0.0 0.0
1.0
0.8
0.6
0.4
0.2
0.0



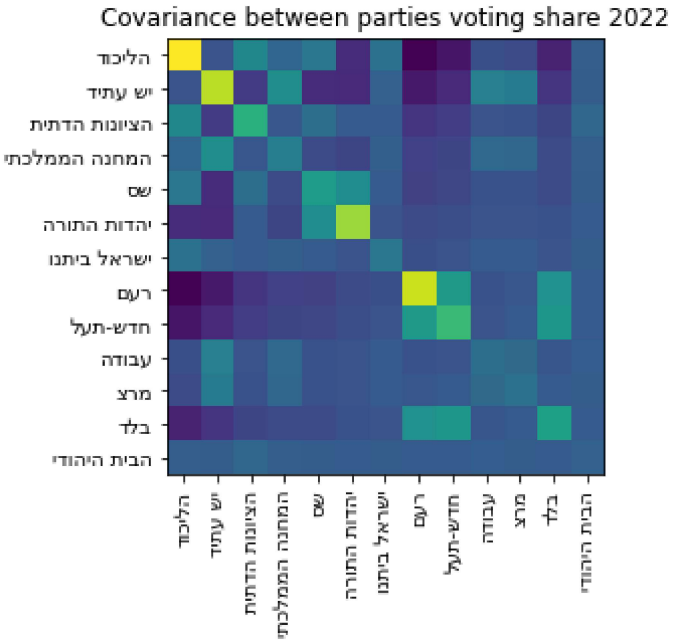
In the following section, we aim to evaluate the performance and accuracy of a estimated results by training it using two separate training sets from dataframes representing the 2021 and 2022 elections. The model will be trained to predict the results of these elections. To gauge the model's performance, we will calculate the mean squared error of the predicted results, applying both the OLS and NNLS methods.

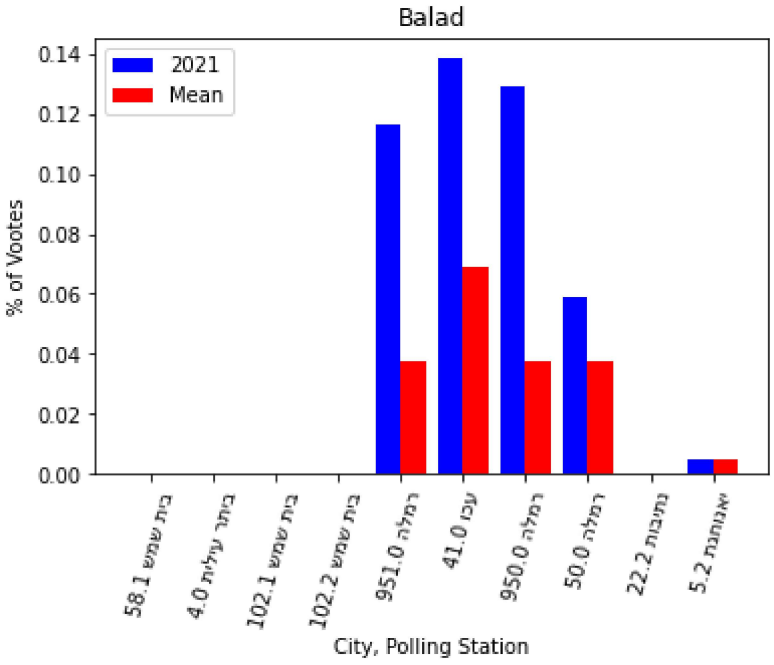
By analyzing the bar plot below, it is evident that the mean squared error (MSE) for the party in question is quite high for both the OLS and NNLS methods. The results obtained from the NNLS method are slightly better, however, the MSE of the party remains high. This is primarily due to the limited information available for the party, as it has only recently re-emerged in the political arena. This lack of data makes it challenging for the model to generalize well to unseen data, resulting in large errors and poor performance. It should also be noted that other parties may also experience high MSE, but for different reasons such as overfitting. In contrast, some parties have relatively low MSE, indicating good overall performance.

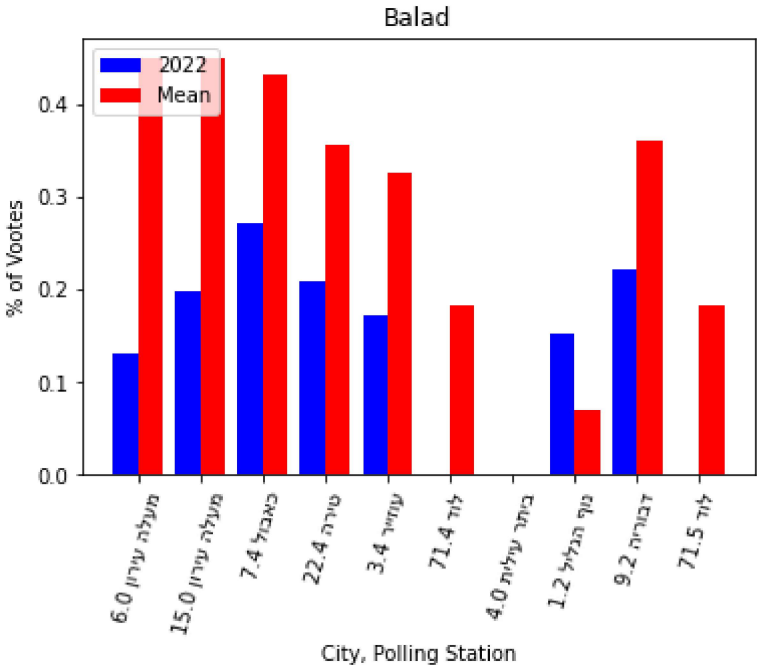


9. Suspicious polling stations

In this section, we aim to identify suspicious ballots by utilizing outlier detection methods, specifically by assessing the Euclidean and Mahalanobis distance between the results of two elections. It is noteworthy that the Mahalanobis distance metric was primarily employed, as it is more robust to outliers, given that it takes into account the covariance structure of the data. This is particularly relevant in this context, because as far as we know, and even though we have a large dataset, the data is not normally distributed. The use of this method allows for a more robust analysis and identification of potential outliers, thereby enhancing the accuracy of identifying suspicious ballots.

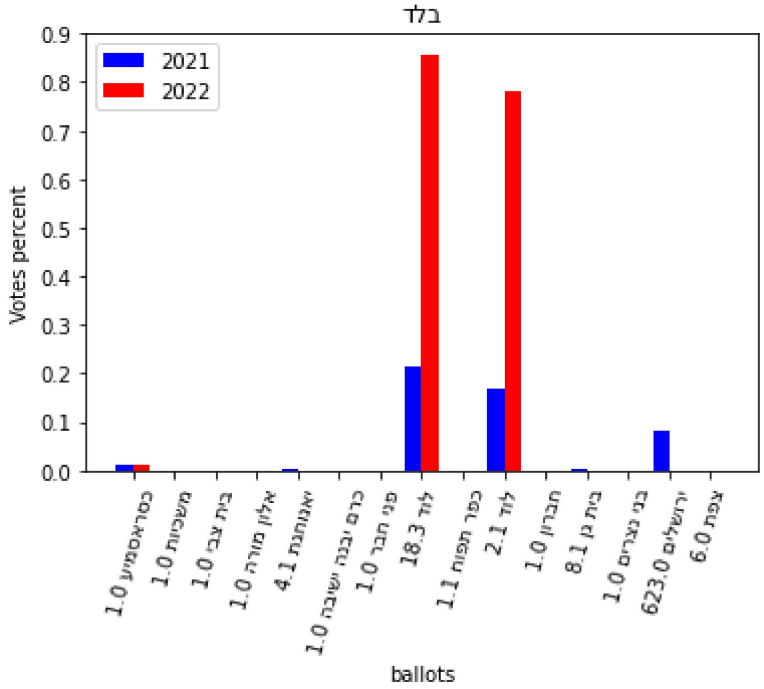






Upon examination of the ballots with the most suspicious behavior across all parties, it was found that a majority of these ballots were from Arab majority or mixed Jewish/Arab cities, and since a significant proportion of non-voters were identified as Arabs in 2021.

Balad is exclusive to recent elections and its mean vote share is calculated based on the prediction of its performance if it were to exist in previous elections. Based on the analysis of suspicious ballots, there is no indication of problematic behavior in the results of Balad. The difference in vote share received by Balad in the suspicious ballots in 2021 is minor and the max deviation -which occured only once- from the projected percentage of 45% (calculated based on the mean) in the 2022 elections is limited to 30%. These observations do not raise suspicions regarding "Balad".



An examination of the euclidean distance was conducted to further assess the potential for suspicious behavior. The results of this analysis do not reveal any indication of problematic behavior. The suspicious ballots in question are primarily located in Arab majority cities, which may suggest that any observed difference in vote share may be attributed to the distribution of support among different political parties within those cities rather than any potential malfeasance on the part of Balad. It is likely that the observed difference in vote share is a result of support for other parties within the Arab majority cities, rather than any suspicious behavior on the part of the "Balad" party.

Finally, we calculated the 10 biggest sum of squared distance between each ballot's results from the two elections. It can be deduced from the analysis of the top 150 suspicious ballots that the mean squared distance (Mahalanobis) of Balad's change is approximately 0.007 ~ 0.7% . This serves as supplementary proof that there is no indication of any irregular behavior present in these polling stations with respect to Balad.