Socioeconomic Classifier based on Ballot Box (Kalpi) Results

A project in data science and data analysis

•••

November 2023

Overview

This project applies machine learning techniques to a key topic in Israeli socio-politics:
The relationship between socioeconomic status and Knesset voting patterns.
It predicts a neighborhood's socioeconomic index (SEI) based on its kalpi results.

The presentation covers the following:

- Data collection and tagging.
- Exploratory statistics: trends, intricacies, exceptions.
- Considerations towards ML classification: approaches, features, optimization.
- Classifier evaluation: results, challenges, solutions and insights.
- Combining classifiers and future work.

Understanding the problem

Socioeconomics and voting in Knesset elections

- **Israeli voting patterns are largely sectarian:**
  - **Jews**: secular / traditional / religious / ultra-orthodox / immigrants.
  - **Arabs**: Palestinian (urban/rural, Muslim/Christian) / Bedouin / Druze-Circassians-Alawites.

- **Different sectors are quite stratified socioeconomically:**
  - Jews: secular > religious > immigrants > traditional > ultra-orthodox
  - Arabs: Christian > Druze etc. > urban Muslim > rural Muslim > Bedouin

- **Research question:**
  - Are these coarse country-wide generalizations manifested in fine numerical patterns?
  - Can the ballots in a given kalpi indicate the socioeconomic index of the neighborhood?

Socioeconomic data (from the Central Bureau of Statistics):

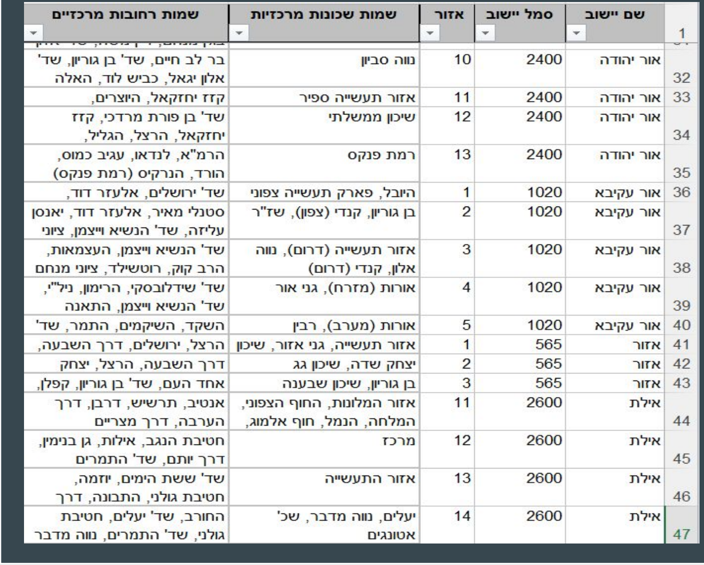
Most recent (2019) socioeconomic index (SEI): 1 - 10 ordinal scale

- One SEI per ‘socioeconomic zone’:
  - Individual small rural locality (~1000 rural zones).
    - E.g. *Bat Shlomo - 8; Amirim - 6; Rumat Heib - 2.*
  - Neighborhood-size section of urban locality (~2000 urban zones)
    - E.g. *Holon 312 ‘Rasko G’ (major streets: Yerushalaim, HaShita, ...) - 7;*
      Rehovot 121 ‘No name’ (major streets: HaShomrim, Sireni, ...) - 5.
    - **No direct mapping** from address to SE zone.
  - For Arab localities (except Nazareth, Rahat): no internal zone indices provided.

Socioeconomic data

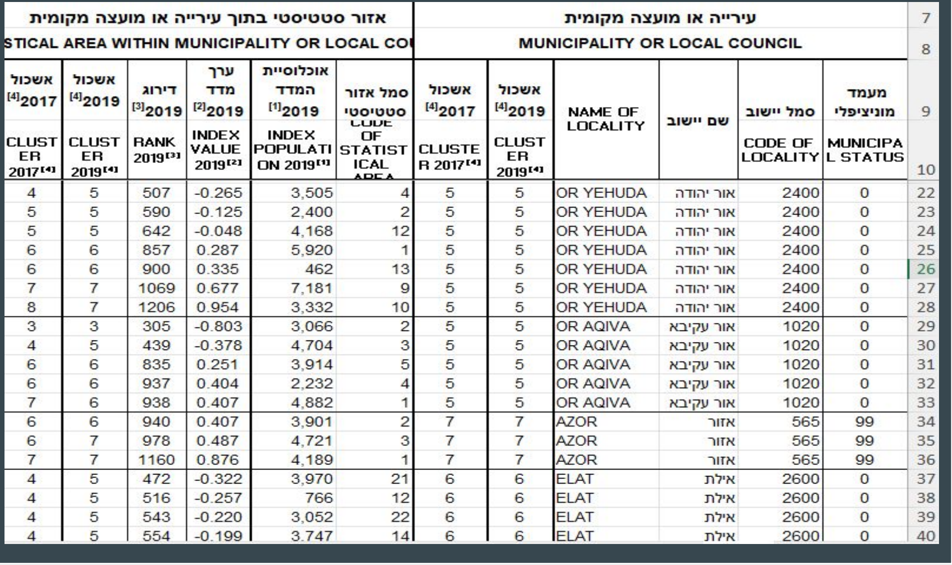
Original CBS Excel sheets

Zoning:



	שם יישוב	סמל יישוב	אזור	שמות שכונות מרכזיות	שמות רחובות מרכזיים
32	אור יהודה	2400	10	נווה סביון	בר לב חיים, שד' בן גוריון, שד' אלון יגאל, כביש לוד, האלה
33	אור יהודה	2400	11	אזור תעשייה ספיר	קזז יחזקאל, היוצרים,
34	אור יהודה	2400	12	שיכון ממשלתי	שד' בן פורת מרדכי, קזז יחזקאל, הרצל, הגליל,
35	אור יהודה	2400	13	רמת פנקס	הרמ"א, לנדאו, עגיב כמוס, הורד, הנרקיס (רמת פנקס)
36	אור עקיבא	1020	1	היובל, פארק תעשייה צפוני	שד' ירושלים, אלעזר דוד,
37	אור עקיבא	1020	2	בן גוריון, קנדי (צפון), שז"ר	סטנלי מאיר, אלעזר דוד, יאנסן עליזה, שד' הנשיא וייצמן, ציוני
38	אור עקיבא	1020	3	אזור תעשייה (דרום), נווה אלון, קנדי (דרום)	שד' הנשיא וייצמן, העצמאות, הרב קוק, רוטשילד, ציוני מנחם
39	אור עקיבא	1020	4	אורות (מזרח), גני אור	שד' שידלובסקי, הרימון, ניל"י, שד' הנשיא וייצמן, התאנה
40	אור עקיבא	1020	5	אורות (מערב), רבין	השקד, השיקמים, התמר, שד'
41	אזור	565	1	אזור תעשייה, גני אזור, שיכון	הרצל, ירושלים, דרך השבעה,
42	אזור	565	2	יצחק שדה, שיכון גג	דרך השבעה, הרצל, יצחק
43	אזור	565	3	בן גוריון, שיכון שבענה	אחד העם, שד' בן גוריון, קפלן,
44	אילת	2600	11	אזור המלונות, החוף הצפוני, המלחה, הנמל, חוף אלמוג,	אנטיב, תרשיש, דרבן, דרך הערבה, דרך מצריים
45	אילת	2600	12	מרכז	חטיבת הנגב, אילות, גן בנימין, דרך יותם, שד' התמרים
46	אילת	2600	13	אזור התעשייה	שד' ששת הימים, יוזמה, חטיבת גולני, התבונה, דרך
47	אילת	2600	14	יעלים, נווה מדבר, שכ' אטונגים	החורב, שד' יעלים, חטיבת גולני, שד' התמרים, נווה מדבר

SE-indexing:



7	עירייה או מועצה מקומית						אזור סטטיסטי בתוך עירייה או מועצה מקומית					
8	MUNICIPALITY OR LOCAL COUNCIL						STICAL AREA WITHIN MUNICIPALITY OR LOCAL CO					
9	מעמד מוניציפלי	סמל יישוב	שם יישוב	NAME OF LOCALITY	אשכול [4]2019	אשכול [4]2017	סמל אזור סטטיסטי	אוכלוסיית המדד [1]2019	ערך מדד [2]2019	דירוג [3]2019	אשכול [4]2019	אשכול [4]2017
10	MUNICIPAL STATUS	CODE OF LOCALITY			CLUSTER 2019[4]	CLUSTER 2017[4]	CODE OF STATISTICAL AREA	INDEX POPULATION 2019[1]	INDEX VALUE 2019[2]	RANK 2019[3]	CLUSTER 2019[4]	CLUSTER 2017[4]
22	0	2400	אור יהודה	OR YEHUDA	5	5	4	3,505	-0.265	507	5	4
23	0	2400	אור יהודה	OR YEHUDA	5	5	2	2,400	-0.125	590	5	5
24	0	2400	אור יהודה	OR YEHUDA	5	5	12	4,168	-0.048	642	5	5
25	0	2400	אור יהודה	OR YEHUDA	5	5	1	5,920	0.287	857	6	6
26	0	2400	אור יהודה	OR YEHUDA	5	5	13	462	0.335	900	6	6
27	0	2400	אור יהודה	OR YEHUDA	5	5	9	7,181	0.677	1069	7	7
28	0	2400	אור יהודה	OR YEHUDA	5	5	10	3,332	0.954	1206	7	8
29	0	1020	אור עקיבא	OR AQIVA	5	5	2	3,066	-0.803	305	3	3
30	0	1020	אור עקיבא	OR AQIVA	5	5	3	4,704	-0.378	439	5	4
31	0	1020	אור עקיבא	OR AQIVA	5	5	5	3,914	0.251	835	6	6
32	0	1020	אור עקיבא	OR AQIVA	5	5	4	2,232	0.404	937	6	6
33	0	1020	אור עקיבא	OR AQIVA	5	5	1	4,882	0.407	938	6	7
34	99	565	אזור	AZOR	7	7	2	3,901	0.407	940	6	6
35	99	565	אזור	AZOR	7	7	3	4,721	0.487	978	7	6
36	99	565	אזור	AZOR	7	7	1	4,189	0.876	1160	7	7
37	0	2600	אילת	ELAT	6	6	21	3,970	-0.322	472	5	4
38	0	2600	אילת	ELAT	6	6	12	766	-0.257	516	5	4
39	0	2600	אילת	ELAT	6	6	22	3,052	-0.220	543	5	4
40	0	2600	אילת	ELAT	6	6	14	3,747	-0.199	554	5	4

Voting data (from the Central Election Committee):

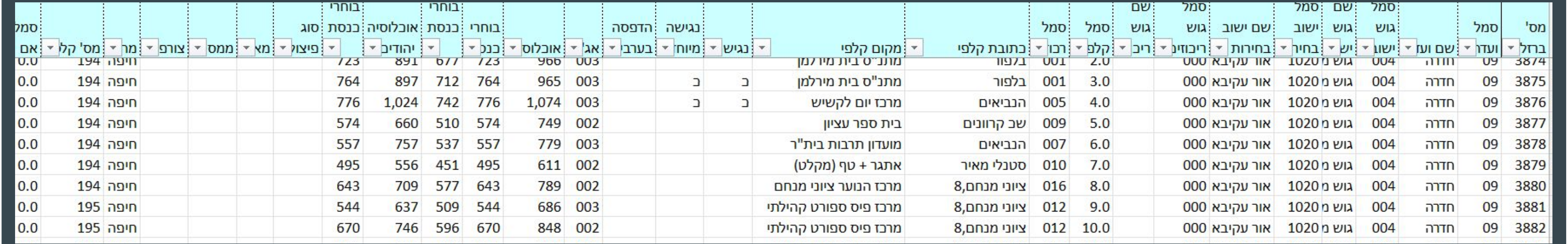
Ballot box (kalpi) data from the 25th Knesset elections (1 Nov. 2022)

- ~11700 kalpiot (+ ~800 double-envelope ‘mobile’ kalpiot), each with:
  - Identification info, including ID and street address / public building, e.g.:
    - *kalpi 1755, Daburiyye, 7 AlYasmin st., Family Health Center*
    - *kalpi 6717, Gan Yavneh, Meiron st., Maccabim Primary School.*
  - Numeric voting data:
    - General: eligible voters - actual voters - legal votes. E.g. *662 - 475 - 471*
    - Ballot: voters per list, e.g. *אמת 13, ב 5, ג 2, ד 0, ום 0, ט 52, כן 83, ל 14, מחל 143*...

Knesset kalpi and voting data

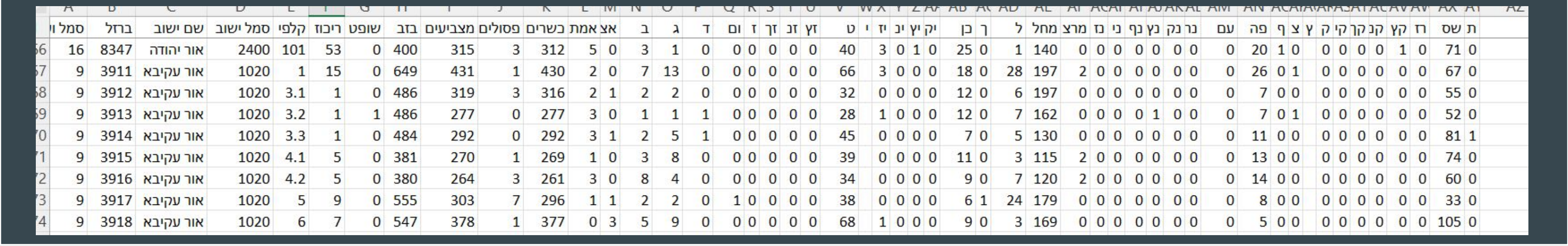
Original CEC Excel sheets

Kalpi details:



סמל אם	מס' קל	מר	צורפ	ממס	מא	סוג פיצו	בוחרי כנסת	אוכלוסיה יהודי	בוחרי כנסת	בוחרי כנכ	אוכלוס	אג'	הדפסה בערבי	נגישה מיוח	נגיש	מקום קלפי	כתובת קלפי	סמל רכו	סמל קלפ	שם גוש ריב	סמל גוש ריכוזי	שם ישוב בחירות	סמל ישוב בחיר	שם גוש יש	סמל גוש ישוב	שם וע	סמל ועדו	מס' ברזל
0.0	194	חיפה					723	891	677	723	966	003				מתנ"ס בית מירלמן	בלפור	001	2.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3874
0.0	194	חיפה					764	897	712	764	965	003		כ	כ	מתנ"ס בית מירלמן	בלפור	001	3.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3875
0.0	194	חיפה					776	1,024	742	776	1,074	003		כ	כ	מרכז יום לקשיש	הנביאים	005	4.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3876
0.0	194	חיפה					574	660	510	574	749	002				בית ספר עציון	שב קרוונים	009	5.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3877
0.0	194	חיפה					557	757	537	557	779	003				מועדון תרבות בית"ר	הנביאים	007	6.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3878
0.0	194	חיפה					495	556	451	495	611	002				אתגר + טף (מקלט)	סטנלי מאיר	010	7.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3879
0.0	194	חיפה					643	709	577	643	789	002				מרכז הנוער ציוני מנחם	ציוני מנחם,8	016	8.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3880
0.0	195	חיפה					544	637	509	544	686	003				מרכז פיס ספורט קהילתי	ציוני מנחם,8	012	9.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3881
0.0	195	חיפה					670	746	596	670	848	002				מרכז פיס ספורט קהילתי	ציוני מנחם,8	012	10.0		000	אור עקיבא	1020	גוש מ	004	חדרה	09	3882

Votes:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
	סמל ו	ברזל	שם ישוב	סמל ישוב	קלפי	ריכוז	שופט	בזב	מצביעים	פסולים	כשרים	אמת	אצ	ב	ג	ד	ום	ז	זך	זנ	זץ	ט	י	יד	ינ	יץ	יק	כן	ך	ל	מחל	מרצ	נז	ני	נף	נץ	נק	נר	עם	פה	ף	צ	ץ	ק	קי	קך	קנ	קץ	רז	שס	ת
66	16	8347	אור יהודה	2400	101	53	0	400	315	3	312	5	0	3	1	0	0	0	0	0	0	40		3	0	1	0	25	0	1	140	0	0	0	0	0	0	0	0	20	1	0		0	0	0	0	1	0	71	0
67	9	3911	אור עקיבא	1020	1	15	0	649	431	1	430	2	0	7	13	0	0	0	0	0	0	66		3	0	0	0	18	0	28	197	2	0	0	0	0	0	0	0	26	0	1		0	0	0	0	0	0	67	0
68	9	3912	אור עקיבא	1020	3.1	1	0	486	319	3	316	2	1	2	2	0	0	0	0	0	0	32		0	0	0	0	12	0	6	197	0	0	0	0	0	0	0	0	7	0	0		0	0	0	0	0	0	55	0
69	9	3913	אור עקיבא	1020	3.2	1	1	486	277	0	277	3	0	1	1	1	0	0	0	0	0	28		1	0	0	0	12	0	7	162	0	0	0	0	1	0	0	0	7	0	1		0	0	0	0	0	0	52	0
70	9	3914	אור עקיבא	1020	3.3	1	0	484	292	0	292	3	1	2	5	1	0	0	0	0	0	45		0	0	0	0	7	0	5	130	0	0	0	0	0	0	0	0	11	0	0		0	0	0	0	0	0	81	1
71	9	3915	אור עקיבא	1020	4.1	5	0	381	270	1	269	1	0	3	8	0	0	0	0	0	0	39		0	0	0	0	11	0	3	115	2	0	0	0	0	0	0	0	13	0	0		0	0	0	0	0	0	74	0
72	9	3916	אור עקיבא	1020	4.2	5	0	380	264	3	261	3	0	8	4	0	0	0	0	0	0	34		0	0	0	0	9	0	7	120	2	0	0	0	0	0	0	0	14	0	0		0	0	0	0	0	0	60	0
73	9	3917	אור עקיבא	1020	5	9	0	555	303	7	296	1	1	2	2	0	1	0	0	0	0	38		0	0	0	0	6	1	24	179	0	0	0	0	0	0	0	0	8	0	0		0	0	0	0	0	0	33	0
74	9	3918	אור עקיבא	1020	6	7	0	547	378	1	377	0	3	5	9	0	0	0	0	0	0	68		1	0	0	0	9	0	3	169	0	0	0	0	0	0	0	0	5	0	0		0	0	0	0	0	0	105	0

Data preparation and tagging:

Combining CBS and CEC data:

- Matching locality zoning with per-zone index (from different CBS Excel sheets)
- Matching kalpi street address with kalpi results (from different CEC Excel sheets)
- Obtaining SEIs for ~20% of kalpiot (respectively from CBS and CEC data):
    - **~1500** kalpiot in small rural localities:
        - Straightforward mapping of kalpi to SEI
        - Problem: **overrepresentation** of certain sectors (Jews, middle class).
    - **~860** urban kalpiot from ~30 urban/larger localities to counter the imbalance.
        - Problem: **extensive manual tagging**, as obtaining urban SE zone from kalpi address requires manual search on a map.

Data preparation and tagging (cont.):

Towards meaningful kalpi data (still in Excel):

- Excluded ‘double envelope’ kalpiot (they don’t represent residential areas).
- Manually added available locality-based sector affinity, in the form of two binary ‘national’ variables - *Jewish* and *Palestinian* - as follows:

Locality is ...		*Palestinian*	
		yes	no
Jewish	yes	Mixed (*Ramle*): 891 kalpiot	Jewish (*Eilat*): 9165 kalpiot
	no	Arab (*Taibe*): 1507 kalpiot	Druze/Circassians/Alwaites (*Julis*): 144 kalpiot

- Converted absolute numbers to proportions (0<p<1):
  - Legal votes out of eligible voters.
  - Party list ballots out of legal votes: the 13 lists with >1% of the country-wide vote distribution, plus ‘others’.

Exploratory statistics of the 2363 SEI-tagged kalpiot:

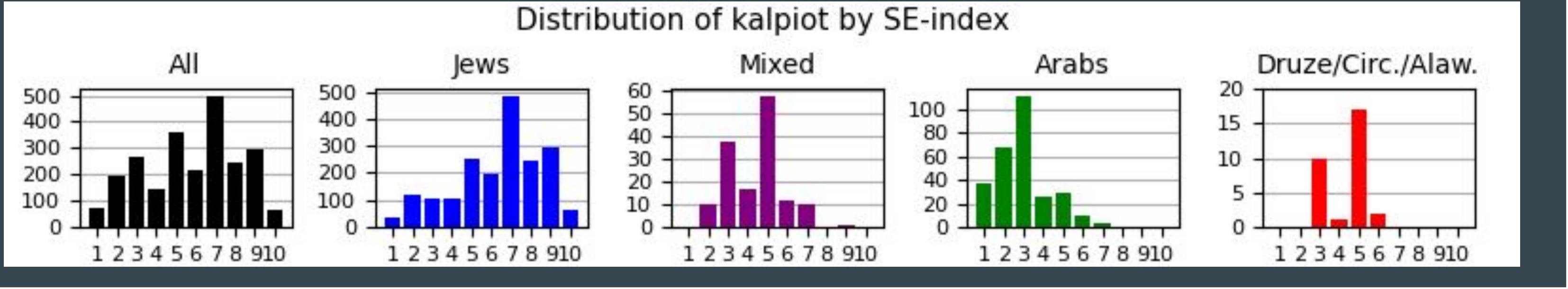
Analyses using Python (*numpy, pandas, matplotlib*):

- SEI-tagged kalpiot by sector (percentage of kalpiot in sector):

Jews: 1903 (21%) *Mixed*: 146 (16%) *Arabs*: 284 (19%) *Druze etc.*: 30 (21%)

- Kalpiot distribution by SEI:

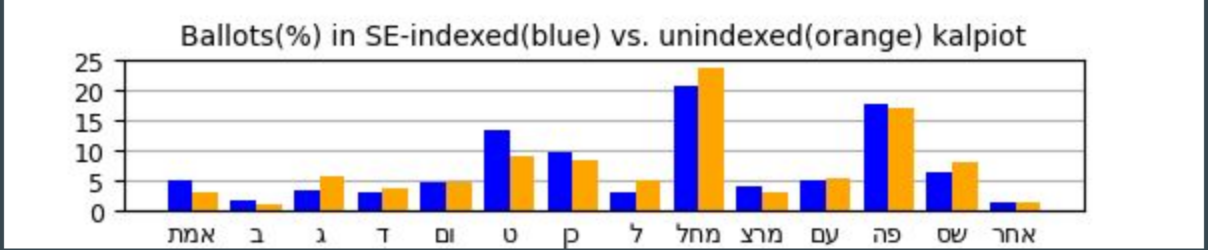
SEI	1	2	3	4	5	6	7	8	9	10
kalpiot	73	196	266	147	359	219	499	242	297	65



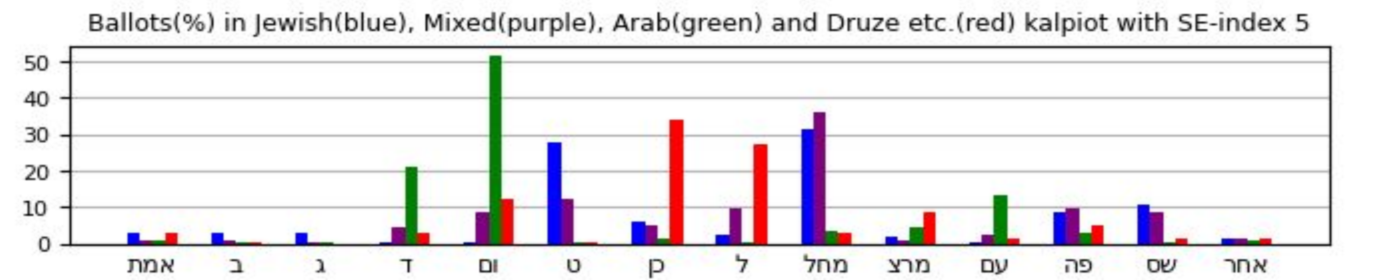
Exploratory statistics of the SEI-tagged kalpiot results:

Analyses using Python:

- Slightly distorted vote distribution, i.e suboptimal sample, mostly due to SE biases:

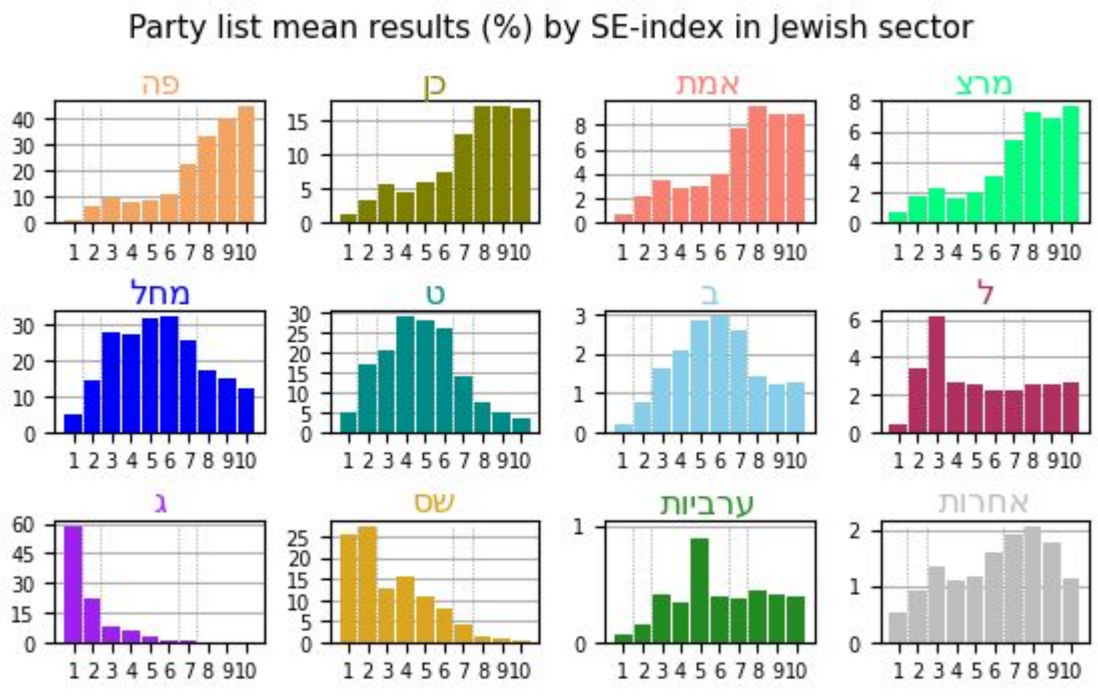


- Sector effect for kalpiot with the same SEI:



Exploratory statistics

Analyses of voting data in the Jewish sector



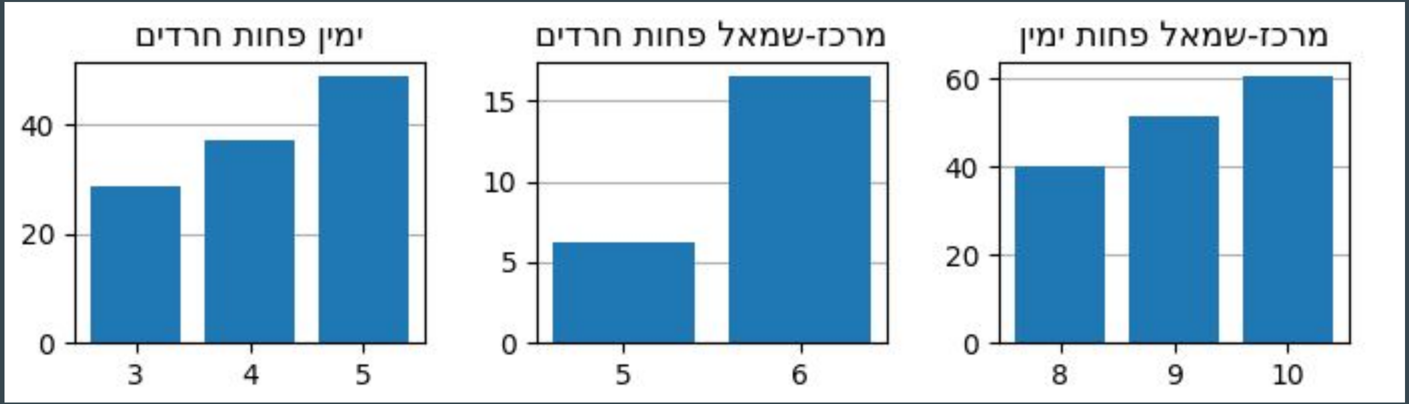
SEI	Trends
1	Mainly Haredi with ג dominance. Little Right, Center-Left negligible.
2	Dominance of שס as ג drops. Pronounced rise of Right. Rise of Ctr-Lf to low plateau. Emergence of ל to plateau.
3	Rise of Right to dominance Drop of שס and ג. Local peak of ל.
4-6	Right dominance. Gradual decline of שס as ג fades
7	Mixed Rt - Ctr-Lf dominance. Decline of Right + שס. Steep rise of Center-Left.
8-10	Center-Left dominance. Drop + decline of Rt, שס fading.

Exploratory statistics

Analyses of voting data in the Jewish sector (cont.)

- Potential additional between-index distinctions for apparently similar SEIs

SEI 3 vs. 4 vs. 5 SEI 5 vs. 6 SEI 8 vs. 9 vs. 10

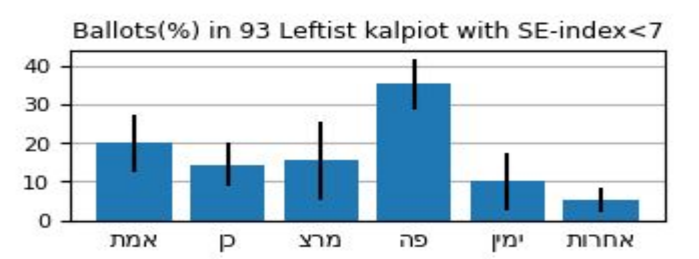


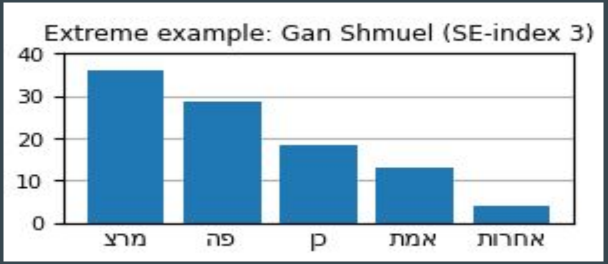
- Hopefully ML classifiers would identify such features.

Exploratory statistics

Analyses of voting data in the Jewish sector (major exceptions)

- **Left-leaning exceptional trend:**
  - 93 of 806 kalpiot with SEIs 1-6 are Center-Left dominant (60%>, many 80%>)

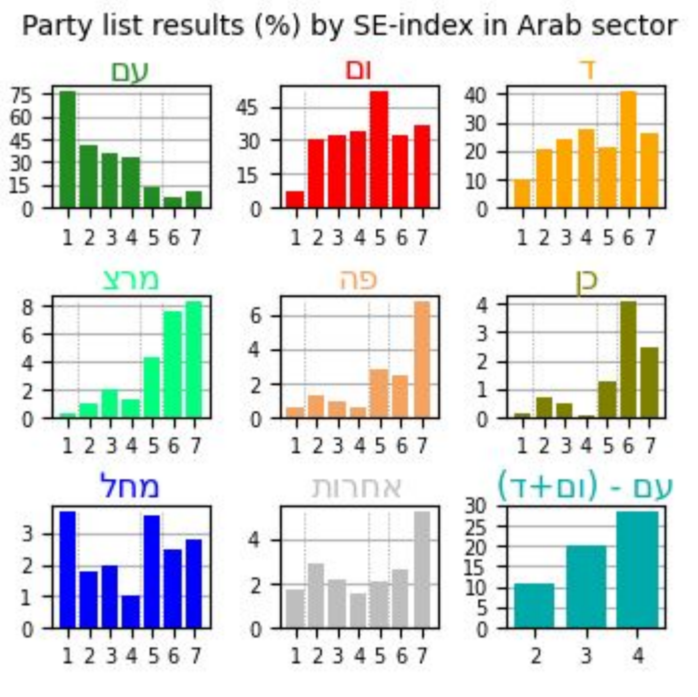




  - 56% of these are in cooperative kibbutzim (*Gan Shmuel, Be'eri, Yotvata* etc.)
  - Coop-economy thwarts CBS's SEIs for otherwise SE-strong localities.
  - These kibbutzim are known and their kalpiot can be marked with a binary variable.
- **No opposite trend:** only 14 of 604 kalpiot SEI-tagged 8-10 with 60%> Right-Haredi votes.

Exploratory statistics:

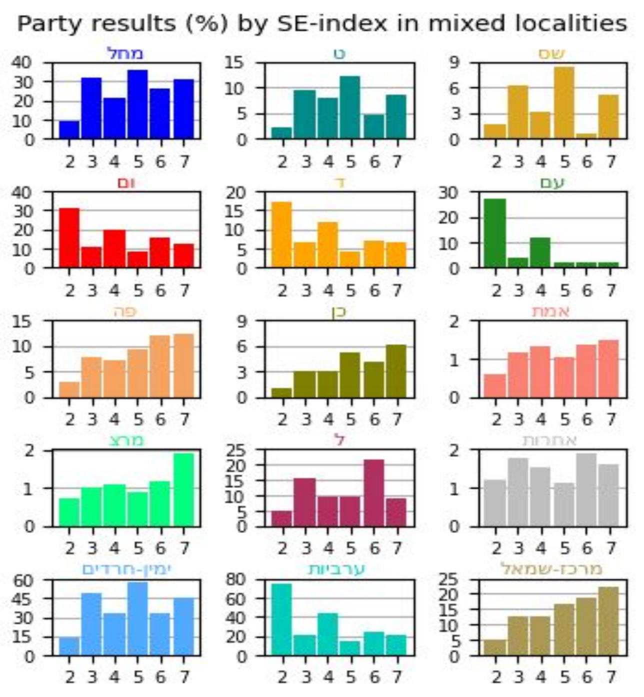
Analyses of voting data in the Arab sector:


Party list results (%) by SE-index in Arab sector

SEI	Dominant sector	Trends
1	Bedouin	Clear dominance of עם (religious). Limited presence of ום/ד (secular). Zionist parties combined: <8%
2-4	Rural/Urban Muslim	Drop of עם and rise of ום/ד, to mixed three-party dominance. Then gradual rise of ום/ד at the expense of עם as SE-index ('secular minus religious' as potential internal cut-off). Zionist parties combined: still <8%
5	Muslim-Christian	Further drop of עם. Dominance of ום. Rise of Zionist Parties: ~13%
6-7	Christian	Mixed ום/ד dominance Further rise of Zionist Parties: >18% Too few kalpiot to generalize further.

Exploratory statistics:

Analyses of voting data in mixed cities:



General trends:

- Irregularities due to small sample: data from 3 out of 8 mixed cities, each with its own SE pattern of Jewish-Arab ‘mixture’.
- Block-specific rather than party-specific index-related trends: שס same as ט/מחל and עם same as ד/ום (cf. Jewish/Arab localities).
- Popularity of ל, further exaggerated by sample.
- Weak but consistently rising Center-Left correlated with SEI.

Ind.	Trends
2	Dominance of Arab parties.
3	Major Arab drop + Right rise (artifact?). Peak of ל as in Jewish locs.
4	Right drop and Arab rise to mixed dominance, plausibly more representative than SEI 3 given general by-sector SE trends.
5	Right rise to dominance as Arab parties drop to low plateau.
6	Major ל peak (artifact?) as Right drops only to rise again in SEI 7.

From statistics to classifiers

General considerations

- The data are characterized by clear trends:
  - Often with apparent cut-offs between adjacent SEIs.
  - Mostly manifesting known socio-political tendencies, e.g. vis a vis religion and nationalism.
  - These suggest a *white-box* approach, e.g. **decision tree** or **logistic regression**.

- However...
  - Distributions in individual kalpiot are more diverse than the averages suggest (wide SDs).
  - Many local and complex discernable patterns are not apparent in averaged observation.
  - These suggest a *black-box* approach, e.g. **random forest** or **neural network**.

From statistics to classifiers

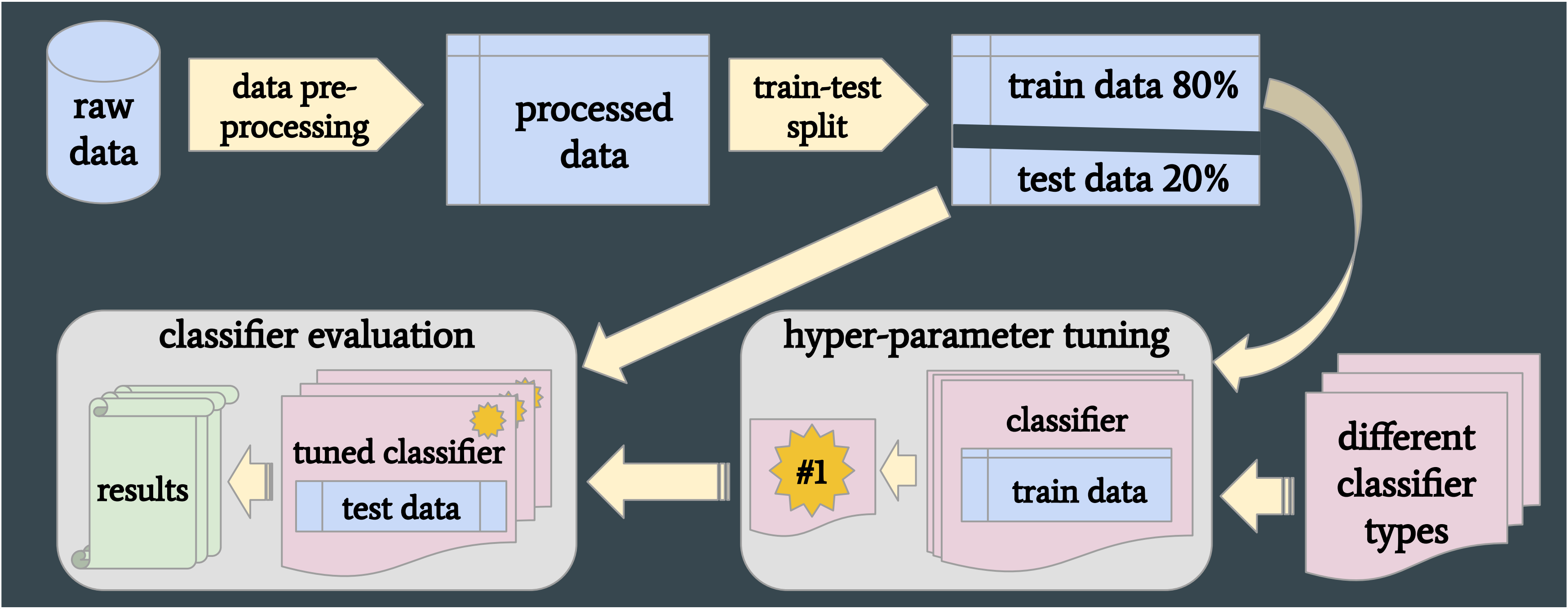
Features

- **Basic set** - Only numerical features of the kalpiot **(15 in total)**:
  - Proportion of valid ballots out of eligible voters.
  - Proportions of each of the party lists (>1%):
    מחל, פה, ט, כן, שס, ג, ל, עם, ום, אמת, מרצ, ד, ב
  - Proportion of all remaining party lists taken together

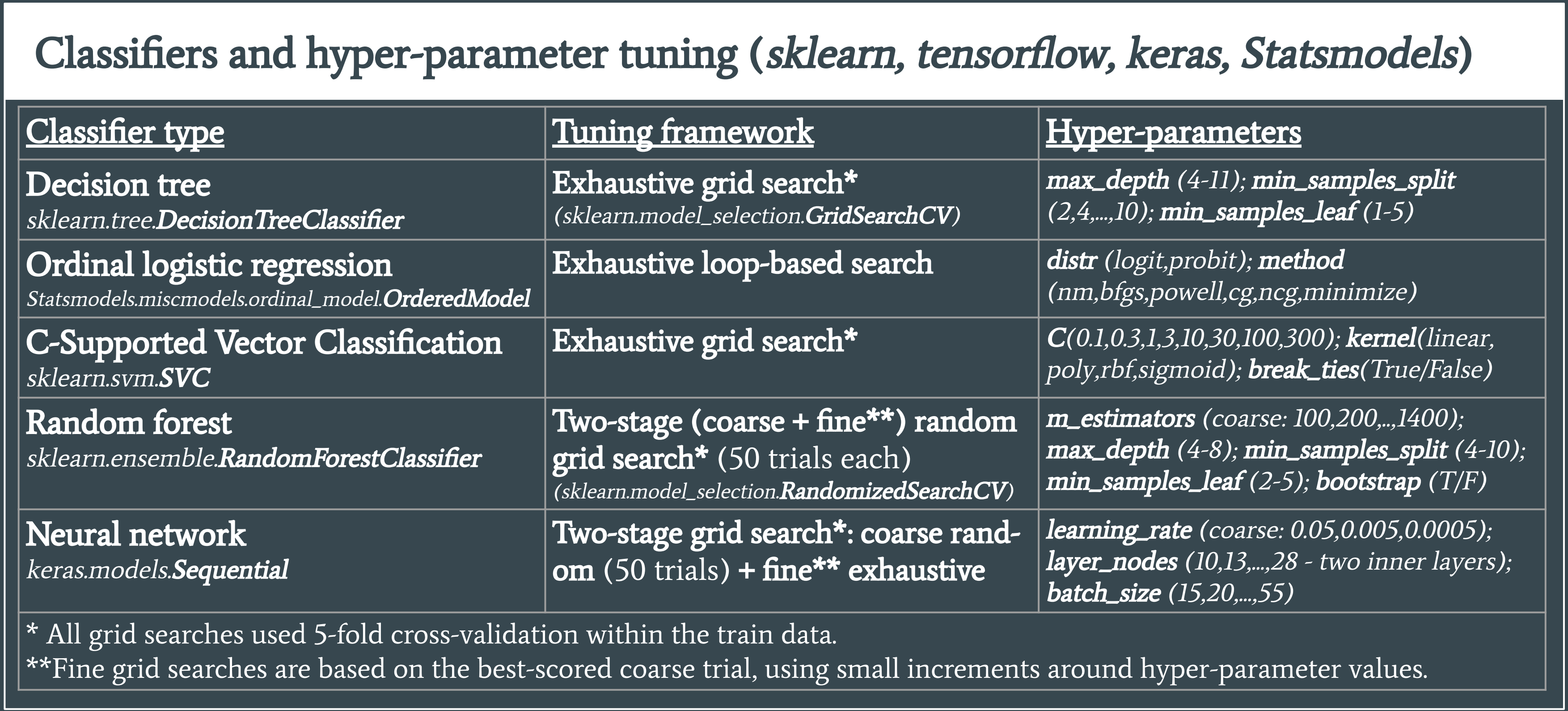
- **Extended set** - Basic set plus 3 additional known per-locality binary features **(18 in total)**:
  - is_Jewish: TRUE for Jewish and mixed localities, FALSE for Arab and Druze etc.
  - is_Palestinian: TRUE for Arab and mixed localities, FALSE for Jewish and Druze etc.
  - is_coop: TRUE for cooperative kibbutzim, FALSE elsewhere.

Socioeconomic classifiers

System pipeline



Socioeconomic classifiers



Classifiers and hyper-parameter tuning (*sklearn, tensorflow, keras, Statsmodels*)

Classifier type	Tuning framework	Hyper-parameters
Decision tree *sklearn.tree.**DecisionTreeClassifier***	**Exhaustive grid search*** *(sklearn.model_selection.**GridSearchCV**)*	***max_depth*** *(4-11);* ***min_samples_split*** *(2,4,...,10);* ***min_samples_leaf*** *(1-5)*
Ordinal logistic regression *Statsmodels.miscmodels.ordinal_model.**OrderedModel***	**Exhaustive loop-based search**	***distr*** *(logit,probit);* ***method*** *(nm,bfgs,powell,cg,ncg,minimize)*
C-Supported Vector Classification *sklearn.svm.**SVC***	**Exhaustive grid search***	***C****(0.1,0.3,1,3,10,30,100,300);* ***kernel****(linear, poly,rbf,sigmoid);* ***break_ties****(True/False)*
Random forest *sklearn.ensemble.**RandomForestClassifier***	**Two-stage (coarse + fine**) random grid search*** (50 trials each) *(sklearn.model_selection.**RandomizedSearchCV**)*	***m_estimators*** *(coarse: 100,200,..,1400);* ***max_depth*** *(4-8);* ***min_samples_split*** *(4-10);* ***min_samples_leaf*** *(2-5);* ***bootstrap*** *(T/F)*
Neural network *keras.models.**Sequential***	**Two-stage grid search*: coarse random** (50 trials) **+ fine** exhaustive**	***learning_rate*** *(coarse: 0.05,0.005,0.0005);* ***layer_nodes*** *(10,13,...,28 - two inner layers);* ***batch_size*** *(15,20,...,55)*

* All grid searches used 5-fold cross-validation within the train data.

**Fine grid searches are based on the best-scored coarse trial, using small increments around hyper-parameter values.

Socioeconomic classifiers

Classifiers and hyper-parameter tuning - additional details

- Before eventual grid search, preliminary tuning for each classifier:
    - Certain hyper-parameters were discarded if values proved costly and ineffective (e.g. ‘entropy’ criterion in Decision tree and Random forest classifiers - used only ‘gini’).
    - Various score metrics were tried: *accuracy, F1-score, macro F1-score, balanced accuracy*. **Accuracy** was chosen (more on this later).

- Each classifier was tried on the data using both ‘basic’ and ‘extended’ feature settings:
    - Best variant was found for each setting.
    - The best of the two was chosen to ‘represent’ the classifier.

Socioeconomic classifiers

NN-specifics (*scikeras.wrappers, keras.losses/callbacks, sklearn.preprocessing*)

- **Scaling:** Proportions scaled to train-based z-scores (surprisingly outperformed min-max).
- **Wrapper:** *Sequential* wrapped by ***KerasClassifier*** for compatibility with grid search:
  - variable input layer: basic vs. extended feature set.
  - variable number of neurons in internal layers.
- **Loss function:** No standard for ordinal variables.*
  - Non-standard functions failed (e.g. *coral_ordinal.**OrdinalCrossEngropy***).
  - ***SparseCategoricalCrossentropy*** was chosen as it substantially outperformed *MSE, MAE.*
- **Callbacks for optimal setting:** Used ***ModelCheckpoint*** and ***EarlyStopping*** to save and later reload the setting with minimal *validation loss.*

* See e.g.: Lazaro, M. & Figueiral-Vidal, A. (2023), “Neural network for ordinal classification of imbalanced data by minimizing a Bayesian cost”, *Pattern Recognition* 137
Elbe, F. & Hall, M. (2001), “A Simple Approach to Ordinal Classification”. *Lecture notes in Computer Science* .

Classification results

First attempt - no sample weights

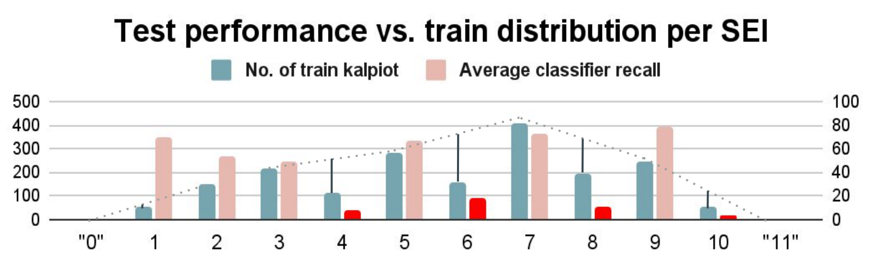
Classifier (top settings)	Features	Acc.	Diff≤1	Macro Recall	Recall per SEI									
					1	2	3	4	5	6	7	8	9	10
DT *(max_depth=9; min_samples_leaf=1; min_samples_split=2)*	*extended*	.463	.789	.433	.65	.47	.41	.27	.63	.30	.42	.36	.74	.09
OLR *(distr=probit; method=powell)*	*basic*	.478	.833	.394	.60	.49	.45	0	.76	0	.76	0	.88	0
SVC *(C=300; kernel=poly; break_ties=False)*	*extended*	.514	.839	.458	.80	.36	.53	.29	.57	.26	.78	.15	.84	0
RF *(n_estimators=570; bootstrap=False; max_depth=9; min_samples_split=4; min_samples_leaf=2)*	*extended*	.522	.829	.456	.80	.49	.61	.09	.68	.14	.76	.17	.82	0
NN *(layer_nodes=28; batch_size=20; learing_rate=0.01)*	*extended*	.507	.820	.439	.80	.56	.29	.03	.74	.30	.74	.13	.80	0

- Promising results, compared to a ‘per-sector mode’ baseline accuracy of 0.252.
- Feature importance (DT,RF): **All** features contribute. Also reflected in OLR coefficients.
- **But:** strong distribution bias, poor recall for certain SE-indices.

Classification results and analysis

Distribution bias across classifiers

		1	2	3	4	5	6	7	8	9	10
4	DT	.03	.09	.26	**.26**	.24	.06	.06			
	OLR		.06	.24	**0**	.59		.09		.03	
	SVC		.06	.21	**.29**	.32	.06	.03	.03		
	RF		.09	.15	**.09**	.62	.03	.03			
	NN	.03	.21	.09	**.03**	.62			.03		
6	DT		.02	.04	.02	.30	**.30**	.30	.02		
	OLR			.02		.42	**0**	.56			
	SVC			.02		.20	**.26**	.52			
	RF			.04		.34	**.14**	.48			
	NN			.02		.38	**.30**	.30			
8	DT		.02	.02			.04	.27	**.36**	.27	
	OLR							.58	**0**	.42	.02
	SVC			.02		.04		.40	**.15**	.40	
	RF					.02		.42	**.17**	.40	
	NN					.06		.50	**.13**	.31	
10	DT								.09	.82	**.09**
	OLR									1.0	**0**
	SVC									1.0	**0**
	RF									1.0	**0**
	NN									1.0	**0**



- SEIs 4,6,8,10 ‘linger’ samples to flanking attractor classes 3,5,7,9.
- Not quantity problem per-se, e.g. SEI 1 vs. 8.
- Rather, **local ‘troughs’**: Poor when quantity is less than average of flanking indices.

Classification attempt, Round #2

Countering distribution bias

- **Counter-measures:**
  - **Selectively add data?** *The dataset already manifests this relative to earlier attempts.*
  - **Duplicate kalpiot for poorly-recalled indices?** *Risk of overfitting to specific kalpiot.*
  - **Change loss function?** *No standard function favoring balanced cross-class recall.*
  - **Add class-based sample weights?** ***YES****... but only if principled and not tweaked!*

- Weighting for samples of SEI *i* with n_i SEI-tagged kalpiot:
  - *if* $n_i < ((n_{i-1}+n_{i+1})/2$ *, then:* $W_i = (n_{i-1} + n_{i+1}) / 2n_i$ *, else:* $W_i = 1$ $(n_0 = n_{11} = 0)$
  - Yielding: $W_1 = 1.34$ $W_4 = 2.13$ $W_6 = 1.96$ $W_8 = 1.64$ $W_{10} = 2.28$

- Desired effect:
  - Boosted recall for SEIs 4,6,8,10 without Class 1 becoming attractor.
  - No significant degradation in total accuracy.

Classification results and analysis

Second attempt - with sample weights* **

Classifier (top settings)	Features	Acc.	Diff≤1	Macro Recall	Recall per SEI									
					1	2	3	4	5	6	7	8	9	10
DT *(max_depth=10; min_samples_leaf=1; min_samples_split=4)*	*extended*	.438 .463	.763 .789	.425 .433	.75 .65	.40 .47	.43 .41	.21 .27	.45 .63	.38 .30	.43 .42	.38 .35	.66 .74	.27 .09
SVC *(C=300; kernel=poly; break_ties=False)*	*extended*	.512 .514	.846 .839	.528 .458	.80 .80	.33 .36	.43 .53	.50 .29	.45 .57	.56 .26	.58 .78	.46 .15	.62 .84	.55 0
RF *(n_estimators=1310; bootstrap=False; max_depth=9; min_samples_split=4; min_samples_leaf=1)*	*extended*	.505 .522	.846 .829	.476 .456	.80 .80	.36 .49	.61 .61	.29 .09	.49 .68	.54 .14	.60 .76	.42 .17	.56 .82	.09 0
NN *(layer_nodes=28; batch_size=55; learing_rate=0.001)*	*extended*	.516 .507	.854 .820	.494 .439	.75 .80	.31 .56	.49 .29	.29 .03	.54 .74	.50 .30	.59 .74	.54 .13	.66 .80	.27 0

* Results of the first attempt are repeated in small fonts.

** The OLR model is experimental and has no sample weight option.

Classification results and analysis

Accuracy and recall

- Decision tree - overall degradation (weakest to begin with).
- Support vector machine - overall improvement:
  - No degradation in accuracy.
  - Model with most balanced recall - greatest improvements and least degradations.
- Random forest - expected (and welcome) tradeoff:
  - Slightly degraded accuracy (but improved near-accuracy).
  - More balanced recall per SEI.
- Neural network - overall improvement:
  - No degradation in accuracy (and improved near-accuracy).
  - Much more balanced recall per SEI.

Classification results and analysis

Confusion matrices for best models (SVC and NN)

No major lingerers/attractors (except Class 9 for SEI 10 in NN).

SVC	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	**.80**		.05	.15						
2	.16	**.33**	.27	.16	.02	.02		.02	.02	
3		.12	**.43**	.20	.18	.02	.02		.02	
4		.03	.15	**.50**	.18	.09		.06		
5		.03	.09	.05	**.45**	.24	.11	.04		
6			.02		.04	**.56**	.36	.02		
7				.01	.07	.11	**.58**	.21	.02	
8			.02		.02		.21	**.46**	.21	.08
9							.06	.20	**.62**	.12
10									.46	**.55**

NN	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	**.75**	.15	.10							
2	.20	**.31**	.24	.18		.02	.02	.02		
3		.16	**.49**	.14	.14	.04		.02		
4		.12	.21	**.29**	.24	.09	.03	.03		
5			.12	.04	**.54**	.18	.09	.03		
6					.16	**.50**	.28	.06		
7			.02		.07	.09	**.59**	.21	.02	
8				.02			.23	**.54**	.17	.04
9							.06	.20	**.66**	.08
10									.73	**.27**

Classification results and analysis

Accuracy by sector

	Jewish	Arab	Mixed	Druze etc.
SVC	.512	.525	.450	.600
RF	.494	.590	.450	.600
NN	.499	.574	.650	.600

- Socio-economic voting patterns are mostly discussed in the context of the Jewish sector, yet:
  - Performance of the models are just as good (or better) for all other sectors.
  - Granted, the Jewish sector is inherently more confusable as it spans the entire SE ladder.

Classification results and analysis

Correlation and ‘collaboration’ between classifiers

- 80% of test samples were classified correctly by **some** classifier, yet:
  - No classifier exceeded 53%.
  - All classifiers together:
    - (Rounded) average: 48%
    - “Majority”: 53.5%
  - Classifier-pair:
    - 42% < agreement < 69%
    - 46% < average is correct < 50%

	DT	OLR	SVC	RF	NN
		⇓ classifier1 vs. classifier2 ⇓			
DT		Agr: .569 MAD: .727 MSD: 1.467	Agr: .429 MAD: .848 MSD: 1.600	Agr: .584 MAD: .628 MSD: 1.178	Agr: .476 MAD: .774 MSD: 1.450
OLR	Cor: .479 MAE: .738 MSE: 1.352		Agr: .465 MAD: .662 MSD: .953	Agr: .535 MAD: .611 MSD: .958	Agr: .545 MAD: .545 MSD: .757
SVC	Cor: .463 MAE: .754 MSE: 1.320	Cor: .481 MAE: .718 MSE: 1.226		Agr: .681 MAD: .448 MSD: .757	Agr: .679 MAD: .412 MSD: .619
RF	Cor: .480 MAE: .756 MSE: 1.413	Cor: .475 MAE: .716 MSE: 1.212	Cor: .495 MAE: .700 MSE: 1.263		Agr: .683 MAD: .425 MSD: .691
NN	Cor: .463 MAE: .740 MSE: 1.265	Cor: .488 MAE: .702 MSE: 1.183	Cor: .495 MAE: .692 MSE: 1.212	Cor: .493 MAE: .688 MSE: 1.188	
	⇑ (classifier1+classifier2)/2 vs. truth ⇑				

Towards a better classifier

Future additions and improvements

- Combining classifiers together (in progress):
  - Tried DT and SVC models trained on the train predictions of the five classifiers.
  - Results nearly identical to those of Random Forest.
- Custom loss function (in progress):
  - Manifest ordinality (greater penalties for greater-distance errors)
  - Mitigate against class imbalance bias by favoring similar recall across classes.
  - Transcended ‘differentiability’ obstacle, but still far from satisfactory.
- Using data from multiple campaigns:
  - 5 Election campaigns in 3 years - further consolidate class models, identify local trends.
  - Irrelevant for Arab sector (Joint List, ד+עם “technical bloc”).

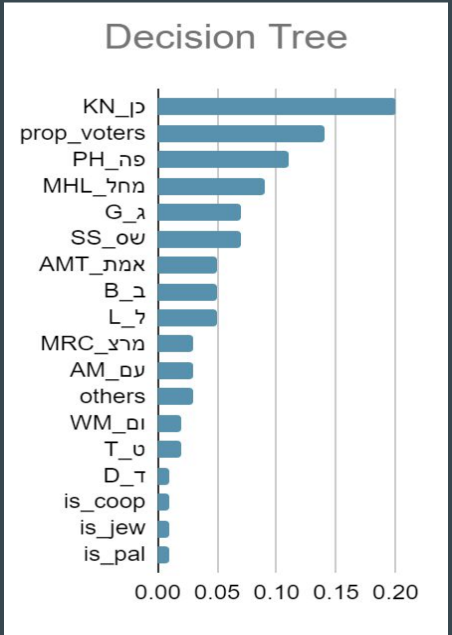
Thank you!!!

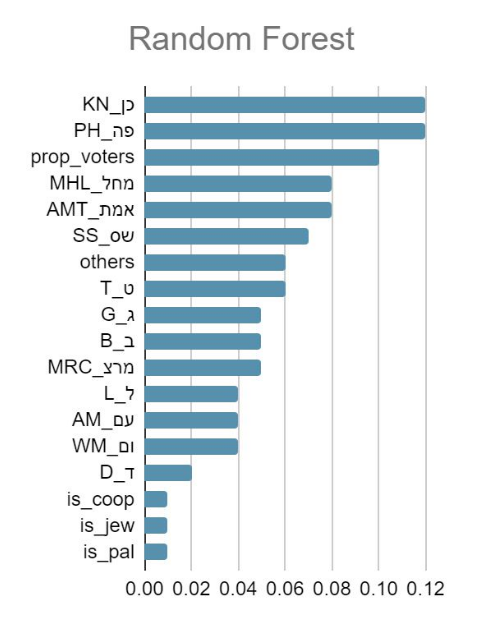
Feel free to ask me anything

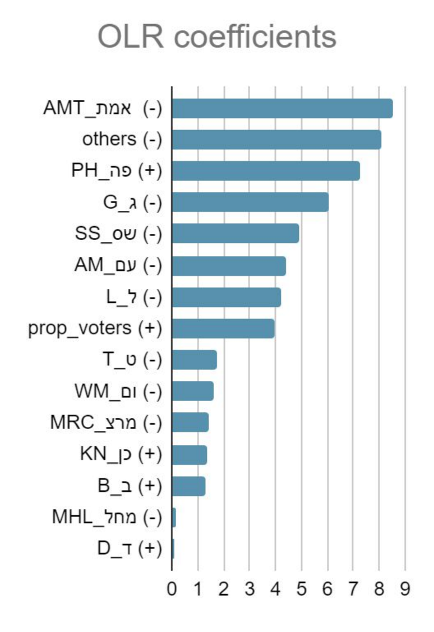
•••

Classification results

Feature importance


Decision Tree
KN_כן
prop_voters
PH_פה
MHL_מחל
G_ג
SS_שס
AMT_אמת
B_ב
L_ל
MRC_מרצ
AM_עם
others
WM_ום
T_ט
D_ד
is_coop
is_jew
is_pal


Random Forest
KN_כן
PH_פה
prop_voters
MHL_מחל
AMT_אמת
SS_שס
others
T_ט
G_ג
B_ב
MRC_מרצ
L_ל
AM_עם
WM_ום
D_ד
is_coop
is_jew
is_pal


OLR coefficients
AMT_אמת (-)
others (-)
PH_פה (+)
G_ג (-)
SS_שס (-)
AM_עם (-)
L_ל (-)
prop_voters (+)
T_ט (-)
WM_ום (-)
MRC_מרצ (-)
KN_כן (+)
B_ב (+)
MHL_מחל (-)
D_ד (+)