Software solutions for identifying outliers

Nicolae-Marius Jula*

Nicolae Titulescu University and Romanian Academy

Abstract

An outlier is an observation that appears to deviate evidently from other observations in the sample. It is important to identify an outlier because it may suggest erroneous data or, in some cases, outliers may be due to random variation or may indicate something scientifically interesting. However, if the data contains significant outliers, the analyst should consider the use of robust statistical techniques.

We demonstrate how to identify outliers in electoral data using informatics methods. An outlier in these datasets may suggest a not necessarily an erroneous data, but an untypical situation – more votes from special lists that the regular registered in that area.

Keywords: outlier, electoral data, electoral outcome.

1. Introduction

An outlier is an observation that seems to deviate significantly from the other sample observations. Identification of potential atypical points is important for several reasons.

An outlier may indicate incorrect data. For example, data can be encoded incorrectly, or an experiment was not performed correctly. If it can be determined that a point is actually atypical, then that value should be removed from the analysis (or corrected if possible).

In some cases, it may not be possible to determine whether or not an outlier negatively affects the analysis. Atypical may be due to random variation or indicate something interesting from a scientific perspective. In any case, it is not recommended to discard that value. However, if the data contains significant outliers, it is recommended to use statistical techniques.

Identification of atypical observations depends on the data's distribution. Usually, the tests start from the hypothesis of normal distribution. Under these circumstances, if the normality assumption for the data being tested is not valid, then a determination that there is an outlier may in fact be due to the non-normality of the data rather than the presence of an outlier.

Some tests are designed to determine the presence of a single point, others can identify multiple points. It is not recommended swapping their use (you can obtain mixed results if you use a single-point identification test for multiple points).

The most common tests for identifying the outliers are:

- Z-Scores and Modified Z-Scores:

$$Z_i = \frac{Y_i - \bar{Y}}{s}$$

Where $\bar{Y}$ is the average value and *s* is the standard deviation.

The Modified Z-Score formula is:

$$M_i = \frac{0.6745(Y_i - \tilde{Y})}{MAD}$$

Where $\tilde{Y}$ is the median values and MAD denotes the median absolute deviation:

$$MAD = median(|Y_i - \tilde{Y}|)$$

*Corresponding author: mariusjula@univnt.ro.

Iglewicz and Hoaglin[1] recommend that modified Z-scores with an absolute value of greater than 3.5 be labeled as potential outliers.

Other formal outlier tests (for normally distributed data) are:

- Grubbs' Test - recommended when testing for a single outlier.
- Tietjen-Moore Test - a generalization of the Grubbs' test to the case of more than one outlier. It has the limitation that the number of outliers must be specified exactly.
- Generalized Extreme Studentized Deviate (ESD) Test - this test requires only an upper bound on the suspected number of outliers and is the recommended test when the exact number of outliers is not known.

Using E-Views to identify outliers in election data

We use the election outcomes from November 2014. The tested hypothesis is that there are some circumscriptions where there is a visible gap between registered voters and the total number of recorded votes. The difference is represented by the voters registered on special lists.

There are situations where the difference can be explained by the usual extra traffic, like airports, railway stations and touristic destinations. It is interesting to identify the cases when the big difference cannot be explained by the previous arguments.

We use econometric methods to identify the outliers, represented by the situations when the difference is not in the statistical limits.

In EViews 8, we use Influence Statistics to identify the outliers. According to EViews help, "influence statistics are a method of discovering influential observations, or outliers. They are a measure of the difference that a single observation makes to the regression results, or how different an observation is from the other observations in an equation's sample. EViews provides a selection of six different influence statistics: RStudent, DRResid, DFFITS, CovRatio, HatMatrix and DFBETAS". For our analyses, we use RStudent, DFFITS and CovRatio.

We create a regression using total recorded votes from list lists (PLS) and the mean values of PLS, using 18533 values from the presidential elections from Romania, November 2014, first round[2]. We calculate the ratio of total votes from special lists (PLS) on total votes (P), as a PLS_P variable.

[1] Iglewicz B., Hoaglin D., "Volume 16: How to Detect and Handle Outliers", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor, 1993.

[2] Central Electoral Bureau, http://www.bec2014.ro/rezultate-finale-2-noiembrie-2014/

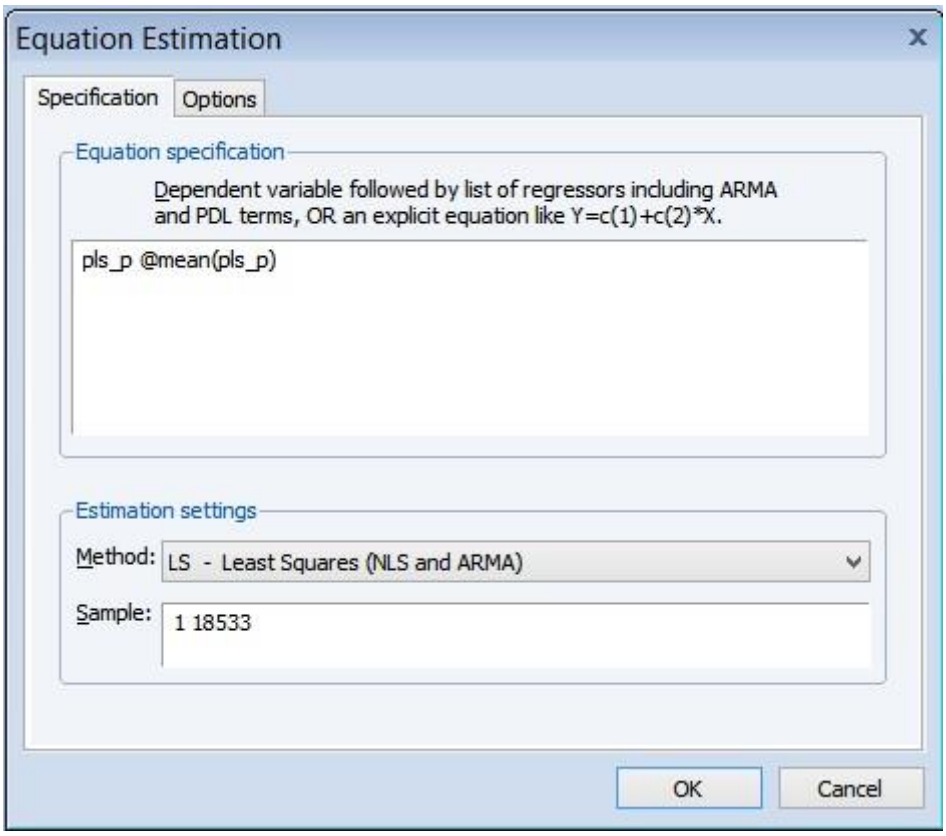

Fig. 1. Equation Estimation

We expect that some values to differ significant from the average, suggesting that in some electoral circumscriptions there are a large number of voters on special lists. It is important to analyze this aspect in regard to the possibility of electoral fraud (multiple voting).

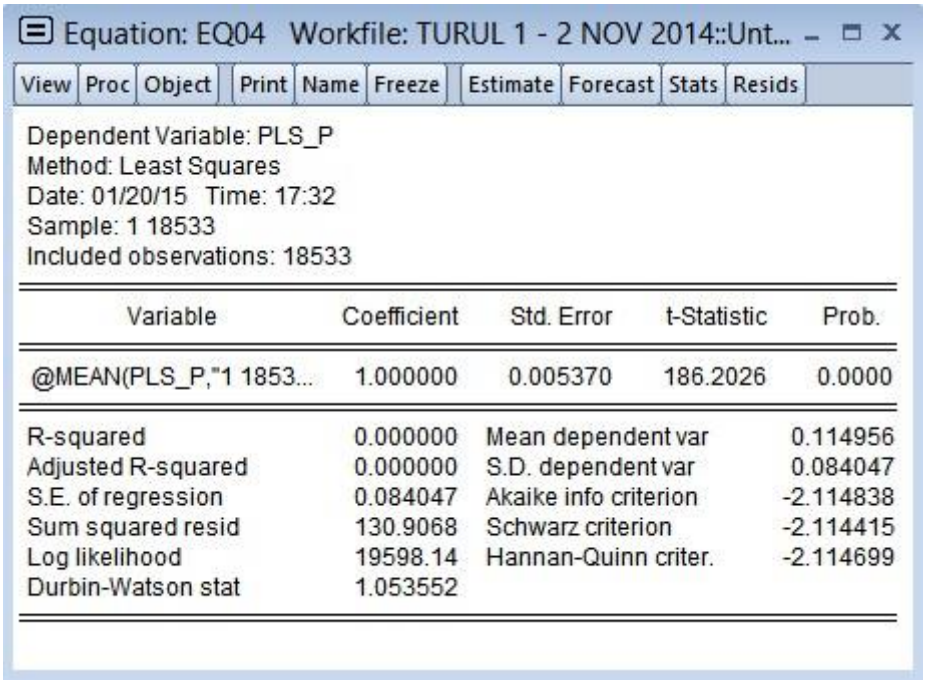

Fig. 2. Estimation results

Theoretically, there should be an interesting situation only when the value is very large compared to its mean. There is no interest in values above mean (meaning that in those circumscriptions the number of voters on special lists is low).

Using influence statistics, we obtain the following results:

Table 1. Top 10 possible outliers

Influence Statistics

Date: 01/20/15 Time: 17:41

Sample: 1 18533

Included observations: 18533

Obs.	Resid.	RStudent	DFFITS	COVRATIO
5169	0.852257	10.16854	-0.074696	0.994558
14056	0.841834	10.04350	-0.073778	0.994693
5887	0.807493	9.631701	-0.070753	0.995126
13786	0.760044	9.063146	-0.066576	0.995694
15492	0.729588	8.698461	-0.063897	0.996041
4930	0.717105	8.549041	-0.062800	0.996179
2580	0.693396	8.265330	-0.060715	0.996435
13817	0.682145	8.130741	-0.059727	0.996553
11145	0.681044	8.117565	-0.059630	0.996564
6013	0.673279	8.024692	-0.058948	0.996645

High values on the three tests suggest that these observations are highly likely to be outliers.

These tests, according to EViews implementation, are:

- The RStudent is numerically identical to the t-statistic that would result from using a dummy variable in the original equation, variable equal to 1 on that particular observation and zero elsewhere. Thus it can be interpreted as a test for the significance of that observation.
- DFFITS is the scaled difference in fitted values for that observation between the original equation and an equation estimated without that observation, where the scaling is done by dividing the difference by an estimate of the standard deviation of the fit.
- COVRATIO is the ratio of the determinant of the covariance matrix of the coefficients from the original equation to the determinant of the covariance matrix from an equation without that observation.

We use VLOOKUP function in Microsoft Excel to identify the corresponding circumscription for the top values from out test.

Table 2. Voters on permanent lists vs. voters on special lists

			Total number of voters:			
			Recorded on permanent lists	**Total votes**	**Total recorded voters from permanent lists**	**Votes on special lists**
			TLP	**P**	**PLP**	**PLS**
No.	**ID.**	**Identification**	**1**	**2**	**3**	**4**
1	5169	Judetul CLUJ, Sectia de votare CJ_1, MUNICIPIUL CLUJ-NAPOCA, Adresa: CLUJ-NAPOCA / Colegiul Economic "Iulian Pop"	1065	366	12	354
2	14056	Judetul SIBIU, Sectia de votare SB_105, MUNICIPIUL SIBIU, Adresa: PĂLTINIŞ / Păltiniş	27	162	7	155
3	5887	Judetul CONSTANŢA, Sectia de votare CT_65, MUNICIPIUL CONSTANŢA, Adresa: CONSTANŢA / ŞCOALA GIMNAZIALA NR. 37	1253	735	57	678
4	13786	Judetul SĂLAJ, Sectia de votare SJ_147, DRAGU, Adresa: UGRUŢIU / Cămin cultural, nr.20	24	72	9	63
5	15492	Judetul TIMIŞ, Sectia de votare TM_284, ORAŞ JIMBOLIA, Adresa: JIMBOLIA / Casa de cultură	1534	669	56	565
6	4930	Judetul CARAŞ-SEVERIN, Sectia de votare CS_362, ZĂVOI, Adresa: POIANA MĂRULUI / OCOLUL SILVIC	53	131	22	109
7	2580	Judetul BIHOR, Sectia de votare BH_552, SÎNMARTIN, Adresa: BĂILE FELIX / Gradinita Baile Felix	493	1341	257	1084
8	13817	Judetul SĂLAJ, Sectia de votare SJ_178, HIDA, Adresa: MILUANI / CAMIN CULTURAL	30	69	14	55
9	11145	Judetul MEHEDINŢI, Sectia de votare MH_180, GOGOŞU, Adresa: OSTROVU MARE / Grădiniţa P.T. II	92	250	51	199
10	6013	Judetul CONSTANŢA, Sectia de votare CT_191, MUNICIPIUL CONSTANŢA, Adresa: CONSTANŢA / UNIVERSITATEA "OVIDIUS"	54	85	18	67

These results suggest that there are some values that should be analyzed. As one can observe, in some voting circumscription where the votes on special lists exceed the votes from regular lists with more than 2000%, with a maximum of 2950% in the observation 5169.

Conclusions

Before using any data for an analysis, one should test that data. When using big data, some data points will be further away from the sample mean than what is deemed reasonable. It is sometimes difficult to say that a particular point is statistically correct or not. Sometimes, outliers are identified as minimum or maximum. One must decide, based on statistic test that a point should be considered as outlier and analyzed in a particular way and/or extracted from the dataset.

When dealing with sensitive data, like electoral results, observation like the ones presented above should raise some questions and there are some facts that should cleared before continuing the analyses.

„This work was financially supported through the project "Routes of academic excellence in doctoral and post-doctoral research - READ" co-financed through the European Social Fund, by Sectoral Operational Programme Human Resources Development 2007-2013, contract no POSDRU/159/1.5/S/137926."

References

[1] Iglewicz B., Hoaglin D., "*Volume 16: How to Detect and Handle Outliers*", The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor, 1993.
[2] Alesina A., Roubini N., & Cohen G.D. *Political Cycles and the Macroeconomy*, Cambridge – Massachusetts: MIT Press, 1997
[3] Beber B., Scacco A., *What the Numbers Say: A Digit-Based Test for Election Fraud*, Political Analysis (2012) 20, 211–234
[4] Fair C. R., *Predicting Presidential Elections and Other Things*, Second Edition, Stanford University Press, Stanford, California, USA, 2012
[5] Gaertner M., *Democracy, Elections and Macroeconomic Policy; Two Decades of Progress*, European Journal of Political Economy, 10: 85-110., 1994
[6] Lemann L., Bochsler D., *A systematic approach to study electoral fraud*, Electoral Studies, 35, 33-47, 2014
[7] Nannestad P., & Paldam M. , *The VP-Function: A Survey of Literature on Vote and Popularity Functions after 25 Years*, in Public Choice 79, 213-245, 1994
[8] Barnett, Lewis, *Outliers in Statistical Data*, 3rd. Ed., John Wiley and Sons, 1994

Annexes

Table 3. Top 50 possible outliers

Influence Statistics
Date: 01/20/15 Time: 18:32
Sample: 1 18533
Included observations: 18533

Obs.	Resid.	RStudent	DFFITS	COVRATIO
5169	0.852257	10.16854	-0.074696	0.994558
14056	0.841834	10.04350	-0.073778	0.994693
5887	0.807493	9.631701	-0.070753	0.995126
13786	0.760044	9.063146	-0.066576	0.995694
15492	0.729588	8.698461	-0.063897	0.996041
4930	0.717105	8.549041	-0.062800	0.996179
2580	0.693396	8.265330	-0.060715	0.996435
13817	0.682145	8.130741	-0.059727	0.996553
11145	0.681044	8.117565	-0.059630	0.996564
6013	0.673279	8.024692	-0.058948	0.996645
9736	0.660924	7.876939	-0.057862	0.996771
18513	0.659388	7.858573	-0.057727	0.996786
3894	0.656215	7.820628	-0.057449	0.996818
15319	0.653004	7.782237	-0.057167	0.996850
3455	0.629961	7.506763	-0.055143	0.997076
17919	0.629230	7.498028	-0.055079	0.997083
16665	0.625231	7.450230	-0.054728	0.997121
4713	0.620338	7.391755	-0.054298	0.997168
11489	0.615813	7.337681	-0.053901	0.997211

8586	0.604145	7.198258	-0.052877	0.997319
11449	0.604066	7.197321	-0.052870	0.997320
5505	0.602435	7.177829	-0.052727	0.997335
12484	0.599329	7.140725	-0.052454	0.997364
8811	0.594721	7.085670	-0.052050	0.997406
13569	0.579074	6.898755	-0.050677	0.997546
5761	0.578592	6.893005	-0.050635	0.997550
17940	0.577351	6.878186	-0.050526	0.997561
9052	0.542845	6.466144	-0.047499	0.997856
4754	0.541978	6.455792	-0.047423	0.997864
13440	0.541760	6.453190	-0.047404	0.997865
13823	0.530877	6.323274	-0.046449	0.997955
5645	0.518377	6.174076	-0.045354	0.998055
9213	0.510044	6.074624	-0.044623	0.998120
14298	0.496155	5.908891	-0.043406	0.998227
15521	0.492887	5.869899	-0.043119	0.998252
1327	0.489882	5.834053	-0.042856	0.998274
5687	0.489695	5.831815	-0.042839	0.998276
13785	0.488817	5.821345	-0.042762	0.998282
12434	0.488647	5.819317	-0.042748	0.998284
18226	0.488369	5.815998	-0.042723	0.998286
18228	0.487520	5.805863	-0.042649	0.998292
8481	0.485044	5.776324	-0.042432	0.998310
5482	0.482266	5.743184	-0.042188	0.998331
15776	0.479103	5.705453	-0.041911	0.998354
5656	0.478794	5.701762	-0.041884	0.998356
17256	0.473549	5.639202	-0.041424	0.998395
12301	0.467825	5.570920	-0.040923	0.998436
5651	0.462821	5.511238	-0.040484	0.998471
5643	0.461967	5.501043	-0.040410	0.998477
5473	0.456472	5.435511	-0.039928	0.998516

Table 4. Voters on permanent lists vs. voters on special lists

			Total number of voters:			
			Recorded on permanent lists	**Total votes**	**Total recorded voters from permanent lists**	**Votes on special lists**
			TLP	**P**	**PLP**	**PLS**
No.	**ID.**	**Identification**	**1**	**2**	**3**	**4**
1	5169	Judetul CLUJ, Sectia de votare CJ_1, MUNICIPIUL CLUJ-NAPOCA, Adresa: CLUJ-NAPOCA / Colegiul Economic "Iulian Pop"	1065	366	12	354
2	14056	Judetul SIBIU, Sectia de votare SB_105, MUNICIPIUL SIBIU, Adresa: PĂLTINIŞ / Păltiniş	27	162	7	155
3	5887	Judetul CONSTANŢA, Sectia de votare CT_65, MUNICIPIUL CONSTANŢA, Adresa: CONSTANŢA / ŞCOALA GIMNAZIALA NR. 37	1253	735	57	678
4	13786	Judetul SĂLAJ, Sectia de votare SJ_147, DRAGU, Adresa: UGRUŢIU / Cămin cultural, nr.20	24	72	9	63
5	15492	Judetul TIMIŞ, Sectia de votare TM_284, ORAŞ JIMBOLIA, Adresa: JIMBOLIA / Casa de cultură	1534	669	56	565
6	4930	Judetul CARAŞ-SEVERIN, Sectia de votare CS_362, ZĂVOI, Adresa: POIANA MĂRULUI / OCOLUL SILVIC	53	131	22	109
7	2580	Judetul BIHOR, Sectia de votare	493	1341	257	1084

		BH_552, SÎNMARTIN, Adresa: BĂILE FELIX / Gradinita Baile Felix				
8	13817	Judetul SĂLAJ, Sectia de votare SJ_178, HIDA, Adresa: MILUANI / CAMIN CULTURAL	30	69	14	55
9	11145	Judetul MEHEDINŢI, Sectia de votare MH_180, GOGOŞU, Adresa: OSTROVU MARE / Grădiniţa P.T. II	92	250	51	199
10	6013	Judetul CONSTANŢA, Sectia de votare CT_191, MUNICIPIUL CONSTANŢA, Adresa: CONSTANŢA / UNIVERSITATEA "OVIDIUS"	54	85	18	67
11	9736	Judetul IAŞI, Sectia de votare IS_149, MUNICIPIUL IAŞI, Adresa: IAŞI / UNIVERSITATEA TEHNICĂ "GH. ASACHI"	1672	1874	420	1454
12	18513	Judetul SECTOR 6, Sectia de votare B6_1205, MUNICIPIUL BUCUREŞTI, Adresa: BUCUREŞTI SECTORUL 6 / COLEGIUL UCECOM "SPIRU HARET"	876	1068	241	827
13	3894	Judetul BRĂILA, Sectia de votare BR_34, MUNICIPIUL BRĂILA, Adresa: BRĂILA / Căminul de Bătrâni "Lacu Sărat"	389	555	127	428
14	15319	Judetul TIMIŞ, Sectia de votare TM_111, MUNICIPIUL TIMIŞOARA, Adresa: TIMIŞOARA / Colegiul National de Arta "Ion Vidu"	605	1211	281	930
15	3455	Judetul BRAŞOV, Sectia de votare BV_43, MUNICIPIUL BRAŞOV, Adresa: BRAŞOV / Vila Orizont	330	541	138	403
16	17919	Judetul SECTOR 3, Sectia de votare B3_611, MUNICIPIUL BUCUREŞTI, Adresa: BUCUREŞTI SECTORUL 3 / ŞCOALA GIMNAZIALĂ Nr.55;	1965	43	11	32
17	16665	Judetul VÂLCEA, Sectia de votare VL_141, ORAŞ CĂLIMĂNEŞTI, Adresa: CĂLIMĂNEŞTI / Hotel Traian (Cofetăria)	543	1070	278	792
18	4713	Judetul CARAŞ-SEVERIN, Sectia de votare CS_145, ORAŞ ORAVIŢA, Adresa: MARILA / Centrul Administrativ	35	68	14	50
19	11489	Judetul MUREŞ, Sectia de votare MS_238, AŢINTIŞ, Adresa: SÎNIACOB / Casa de locuit Siniacob	22	26	7	19
20	8586	Judetul HARGHITA, Sectia de votare HR_33, MUNICIPIUL GHEORGHENI, Adresa: LACU ROŞU / Corp de clădire cu 20 de apartamente	76	89	25	64
21	11449	Judetul MUREŞ, Sectia de votare MS_198, ORAŞ SOVATA, Adresa: SOVATA / Biblioteca Orăşenească Sovata Băi	484	573	161	412
22	5505	Judetul CLUJ, Sectia de votare CJ_337, BELIŞ, Adresa: SMIDA / Scoala Generala Smida	35	46	13	33
23	12484	Judetul OLT, Sectia de votare OT_179, DEVESELU, Adresa: DEVESELU / Asociaţia locatari Nr.5	213	196	56	140
24	8811	Judetul HARGHITA, Sectia de votare HR_258, SUSENI, Adresa: LIBAN / Scoala Generala Liban	58	31	9	22

25	13569	Judetul SATU MARE, Sectia de votare SM_264, POMI, Adresa: ACIUA / CAMIN CULTURAL	75	134	41	93
26	5761	Judetul CLUJ, Sectia de votare CJ_593, SĂVĂDISLA, Adresa: VĂLIŞOARA / Căminul Cultural	55	62	19	43
27	17940	Judetul SECTOR 3, Sectia de votare B3_632, MUNICIPIUL BUCUREŞTI, Adresa: BUCUREŞTI SECTORUL 3 / ŞCOALA GIMNAZIALĂ Nr.78;	1650	26	8	18
28	9052	Judetul HUNEDOARA, Sectia de votare HD_209, ORAŞ GEOAGIU, Adresa: GEOAGIU-BĂI / Scoala Primara Geoagiu Bai	388	564	193	371
29	4754	Judetul CARAŞ-SEVERIN, Sectia de votare CS_186, BREBU NOU, Adresa: GĂRÎNA / Căminul Cultural	278	137	47	90
30	13440	Judetul SATU MARE, Sectia de votare SM_135, BELTIUG, Adresa: BOLDA / CĂMINUL CULTURAL BOLDA	51	67	23	44
31	13823	Judetul SĂLAJ, Sectia de votare SJ_184, HIDA, Adresa: PĂDURIŞ / SCOALA GENERALA	34	48	17	31
32	5645	Judetul CLUJ, Sectia de votare CJ_477, IARA, Adresa: BORZEŞTI / CASA	56	60	22	38
33	9213	Judetul HUNEDOARA, Sectia de votare HD_370, GHELARI, Adresa: PLOP / MAGAZIN ABC	25	32	12	20
34	14298	Judetul SIBIU, Sectia de votare SB_347, ŞEICA MARE, Adresa: PETIŞ / Imobil Petiş	48	36	14	22
35	15521	Judetul TIMIŞ, Sectia de votare TM_313, BARA, Adresa: LĂPUŞNIC / Căminul cultural Lăpuşnic	33	51	20	31
36	1327	Judetul ARGEŞ, Sectia de votare AG_449, RUCĂR, Adresa: SĂTIC / Şcoala Sătic	104	124	49	75
37	5687	Judetul CLUJ, Sectia de votare CJ_519, MĂRIŞEL, Adresa: MĂRIŞEL / Şcoala Generală	92	129	51	78
38	13785	Judetul SĂLAJ, Sectia de votare SJ_146, DRAGU, Adresa: ADALIN / Şcoala cu cls.I-IV,nr.100	128	106	42	64
39	12434	Judetul OLT, Sectia de votare OT_129, BĂRĂŞTI, Adresa: MOŢOEŞTI / Sediu Fostă Brigadă Agricolă	66	111	43	67
40	18226	Judetul SECTOR 5, Sectia de votare B5_918, MUNICIPIUL BUCUREŞTI, Adresa: BUCUREŞTI SECTORUL 5 / FACULTATEA DE DREPT	695	842	334	508
41	18228	Judetul SECTOR 5, Sectia de votare B5_920, MUNICIPIUL BUCUREŞTI, Adresa: BUCUREŞTI SECTORUL 5 / FACULTATEA DE DREPT	663	727	289	438
42	8481	Judetul GORJ, Sectia de votare GJ_255, PADEŞ, Adresa: CERNA-SAT / SC GEN CERNA SAT	59	50	20	30
43	5482	Judetul CLUJ, Sectia de votare CJ_314, APAHIDA, Adresa: CÎMPENEŞTI / Sediu sector Piscicola	54	72	29	43
44	15776	Judetul TIMIŞ, Sectia de votare TM_568, TOPOLOVĂŢU MARE, Adresa: CRALOVĂŢ / Căminul Cultural Cralovăţ	101	101	41	60

45	5656	Judetul CLUJ, Sectia de votare CJ_488, IARA, Adresa: VALEA VADULUI / SCOALA	42	32	13	19
46	17256	Judetul VRANCEA, Sectia de votare VR_306, TULNICI, Adresa: LEPŞA / Şcoala Lepşa	360	435	179	256
47	12301	Judetul NEAMŢ, Sectia de votare NT_481, VÎNĂTORI-NEAMŢ, Adresa: VÎNĂTORI-NEAMŢ / MANASTIREA SIHASTRIA	164	151	63	88
48	5651	Judetul CLUJ, Sectia de votare CJ_483, IARA, Adresa: LUNGEŞTI / CASA	41	45	19	26
49	5643	Judetul CLUJ, Sectia de votare CJ_475, GÎRBĂU, Adresa: CORNEŞTI / CAMINUL CULTURAL CORNESTI	106	104	44	60
50	5473	Judetul CLUJ, Sectia de votare CJ_305, ALUNIŞ, Adresa: VALE / CAMIN CULTURAL VALE	81	77	33	44