Decision Memo:
Considerations Before Implementing E-Voting For EU Countries

Anna Pauxberger & Sherington Anton Amarapala
Minerva Schools at KGI
Class of 2020

Code for this assignment:
Matching on Salta Data:
https://gist.github.com/anonymous/a2fbfe3b75f85055a473e8c7235344af
Genetic Matching on Salta Data:
https://gist.github.com/anonymous/3b0d757dc757560d24370144e0d087c9
Match Balance Comparioson Tables:
https://gist.github.com/anonymous/31c35814bdc63b27a494fa7cf388474c

Decision Memo: Considerations Before Implementing E-Voting For EU Countries

TO: Jean-Claude Juncker, President of the European Commission
FROM: Anna Pauxberger, Statistician at Eurostat & Sherington Anton Amarapala, Statistician at Eurostat
DATE: December 15, 2017
RE: Considerations Before Implementing E-Voting For EU Countries

---

Executive Summary

The European Commission is considering whether or not to require electronic voting for member countries of the European Union. Before making this decision, many factors have to be taken into account, such as feasibility, cost, usefulness, etc. Us at Eurostat were asked by the European Commission to look into how e-voting affects citizens perception of the voting process. The paper "Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions" by Alvarez R. et al. looks into citizens perception of e-voting in Salta, Argentina. We replicated their results and extended their analysis using genetic matching and came to the conclusion that e-voting is perceived positively by citizens, but there are some concerns about the secrecy of their votes. ***We, therefore, recommend the European Commission to look further into how ballot secrecy can be ensured when implementing e-voting and how to ensure that citizens feel that their vote is kept secret.***

Introduction

Electronic Voting was first experimented with in the 1960s but was increasingly implemented in the recent years in Europe and the Americas. E-voting, as opposed to i-voting, refers to local electronic polling stations, while i-voting is possible remotely via the internet. Either one facilitates counting votes for governmental units, but the effect on voters themselves is debated.

The paper "Voting Made Safe and Easy: The Impact of e-voting on Citizens Perceptions" written by Alvarez, M. R., et al. was published in *The European Political Science Association* journal in June 2013. The paper uses data from the partial implementation of e-voting in Salta, Argentina, during the 2003 election to assess the impact of e-voting on how citizen perceive the voting process. The implementation of e-voting was not randomized at the voter level but at the voting-station level. Voting stations covering some districts implemented e-voting, whereas other districts used the traditional paper ballot method. After voting, voters were asked to fill out a survey where they answered questions about their demographics and voting experience.

Alvarez et al. used propensity-score matching to evaluate the effect of introducing e-voting on the voting experience. We replicate their results by first running their code and confirming their results. Alvarez et al. use the *MatchIt* package in R (Ho, D. et al., 2011) but the *Matching* package (Sekhon, J., 2011) can also be used. Therefore, we replicate their results using the *Matching* package in R as well, to see if that improves the match-balance and leads to more accurate results. Because of the fundamental flaws of propensity score matching we extended

Alvarez et al.'s paper by using multivariate genetic matching, as outlined by Alexis Diamond and Jasjeet Sekhon (2013), instead of propensity score matching

Part 1 - Verifying the original methods by Alvarez et al.

We were able to replicate their code in R using the data set and code provided by Harvard Dataverse (n.d.). Our replication gives us the same results as Alvarez et al. present in the paper. (See Appendix 1) Their code runs properly and the data set provided is complete. Their main finding is that e-voting is perceived as easier to use than the traditional paper ballot and most e-voters support substituting traditional voting with e-voting. However, people assigned to e-voting were also more concerned about the secrecy of their vote. (See Table 1)

	Mean	Standard Deviation	Posterior Interval (5% — 95%)	
Select candidates electronically	29.58	4.99	21.32	37.46
Evaluation voting experience	27.05	4.35	19.95	34.16
Agree substitute TV by EV	21.54	4.15	14.92	28.36
Difficulty voting experience	21.16	6.34	11.62	32.32
Elections in Salta are clean	17.44	3.74	11.20	23.84
Qualification of poll workers	13.22	5.54	4.38	22.52
Sure vote was counted	8.38	3.52	3.25	14.56
Speed of voting process	7.03	6.97	-3.88	19.32
Confident ballot secret	-9.15	4.05	-16.00	-3.21

Table 1 - Effect of e-voting obtained using propensity score matching

Part 2 - Replicating the results using the *Match()* function

Many of the balance statistics provided by the original paper have *MatchBalance()* p-values well below 1, indicating a large difference between the control and the treatment group, which leaves

large space for confounding variables potentially biasing the result. On the contrary, a more desirable p-value of 1 for each covariate would indicate identical treatment and control groups, and thus perfect balance. While the paper applied all tools correctly, we decided to verify the procedure by matching the dataset with a different package, called *Matching*. To account for the same settings as used in the paper, we set the caliper to 0.05, indicating a small standard deviation distance for each covariate and lower acceptance of observations. This leaves 23 observations but should result in a higher p-value as the matches are of higher quality. Furthermore, we used propensity score matching as used in the paper, to fairly compare the two methods.

Using the *Match()* function from the *Matching* package for propensity score matching did not outperform *MatchIt*, as the balance statistics performed poorer, with lower p-values (see Table 2). The higher the p-value, the lower the statistical significance for the difference between the two groups, which indicates better matching for that covariate. The p-value was generated using a T-test for binary variables, and the Kolmogorov-Smirnov test for ordinal variables, since it is a more precise measure of differences in two distributions. The *Match()* function from the *Matching* package alone was not able to outperform *MatchIt*, as the balance statistics performed poorer, with lower p-values (see Table 2). For example, *MatchIt* was able to balance age group with a p-value of 1.00, while *Match()* decreased the initial p-value of 0.53 to 0.04.[1] Nonetheless, the real advantage of *Matching* lies in its extensions of *MatchBalance*, which allows for

[1] #evidencebased - We provide an example from the original paper as well as our own code in comparison as evidence for the argument that Match does not outperform MatchIt, but GenMatch is able to outperform both with regards to balancing two groups. We explain the results in detail.

evaluation of the matched sets in detail, as well as *GenMatch*, which runs genetic matching generating more reliable matching results.

	MatchIt BM	MatchIt AM	Gen/Match BM	Match AM	GenMatch AM
Age group	0.55	1.00	0.53	0.04	1.00
Education	0.00	0.72	0.00	0.06	1.00
White collar	0.29	0.80	0.60	0.59	1.00
Not full time worker	0.02	0.80	0.05	0.29	1.00
Male	0.87	1.00	0.75	0.33	1.00
Technology count	0.00	0.59	0.01	0.05	1.00
Political information	0.00	0.55	0.00	0.62	1.00

Table 2 - MatchBalance p-value comparison
BM and AM stand for before matching and after matching respectively. Values for the MatchIt function were taken from the original paper Table 2, which used the Kolmogorov-Smirnov test for ordinal variables and the difference in proportions test for binary variables. Match and GenMatch values were calculated using their respective functions, and extracting the Kolmogorov-Smirnov Bootstrap p-value for ordinal variables, and the T-test p-value for binary variables.

Propensity score matching estimates the effect of e-voting on the voting experience by matching based on the probability of receiving the treatment, which is defined by the covariates. The goal is to reduce the bias induced by confounding variables that could affect the treatment effect based on whether or not a unit received treatment instead of control due to their characteristics.

Gary King and Richard Nielson argue that propensity score matching can increase imbalance, inefficiency, model dependence, and bias (2016). They recommend using other matching methods instead because propensity score matching "attempts to approximate a completely randomized experiment, rather than, as with other matching methods, a more efficient fully blocked randomized experiment" (King, G. & Nielson, R., 2016). Propensity score matching makes the assumption that the variables affecting the data generation process are known. This is usually not the case, and you should, therefore, not use that assumption as the basis for matching.

Based on the criticism of propensity score matching we decided to extend their analysis using multivariate genetic matching.

Part 3 - Extending with multivariate genetic matching

Genetic matching is "a multivariate matching method, that uses an evolutionary search algorithm to determine the determine the weight each covariate is given" (Diamond, A. et al., 2012). Quickly explained, an evolutionary algorithm uses a cost function, which is the balance achieved, to grade the population and decides which units to keep and which ones to discard. It uses random mutations to change parts of the "gene" and crossover to mix the "genes" of the already existing population. The gene, in this case, is the set of weights for each of the covariates. After multiple iterations, the balance improves and if the algorithm runs for enough iterations, you will achieve an optimal balance. If the matching improves the covariate balance, we can conclude that the method is better than the propensity-score matching used by Alvarez et al.

We use the *Matching* and *rgenound* packages to run genetic matching on the balance matrix they use to check their match balance. We included demographic variables for our genetic matching such as age group, education, use of technology, where they get political info from, job data, and gender. We run the *GenMatch* function with e-voting as the treatment to get the covariate weights. We used a population size of 500 to ensure that we get the optimal balance. The caliper is set to 0.05, the same as in the original paper, to ensures that the matching does not happen if units are further than 0.05 standard units away to prevent a decrease in matchbalance. We then

run nine different matching functions, one for each of the nine outcomes[2], using the weight matrix we got from running the genetic algorithm.

The estimated effect of e-voting obtained by our genetic matching model is similar to the results in the original paper (See Table 3). Our standard deviation for all the outcomes is lower than in the original paper which gives us smaller confidence intervals. All the p-values are under 0.05 which means that all our estimates pass our 0.05 significance test. The lowest p-value is for "speed of voting process" at 0.026. The genetic matching function only found matches for 487 of the original 1054 observations, however a lower quantity of data does not necessarily mean lower quality of the results if the data can provide more insight. Despite the lower number of observations, we can assume our results are valid since our matches are very high with p-values of 1 and the estimates prove statistically significant with p-values of <0.05. We do a balance test to see how the genetic matching function performs in improves the covariate balance. Before the matching nine of the 16 variables in the balance are statistically different between the treatment and control group. After the genetic matching, none of the variables have a statistically significant difference between the treatment and control group. Actually, there is no difference between the treatment and control group after the genetic matching in implemented (See Table 2). All the p-values are 1.00 which indicates a perfect match on the covariates between the two groups. This indicates that by running genetic matching, we can remove all possible confounding

[2] The 9 outcomes from the survey used in the analysis are; evaluation of the voting experience, qualification of the poll workers, whether or not they agree that traditional voting should be substituted by e-voting, speed of the voting process, easet of the voting procedure, their preferred method for selecting candidates from different political parties, confidence in the vote being counted, confidence on ballot secrecy, and how clean they believe the Salta elections are.

cause by the observed covariates. It is important to note that we do not know anything about the unobserved covariates and whether or not they influence the treatment effect.

	Mean	Standard Deviation	Conf.int 5%	Conf.int 95%
Select candidates electronically	26.37	1.88	22.68	30.06
Evaluation voting experience	24.69	2.10	20.57	28.81
Agree substitute TV by EV	19.16	1.89	15.46	22.86
Difficulty voting experience	19.13	1.87	15.46	22.80
Elections in Salta are clean	23.52	2.14	19.32	27.72
Qualification of poll workers	8.69	1.60	5.56	11.83
Sure vote was counted	16.27	1.64	13.06	19.49
Speed of voting process	3.76	1.68	0.46	7.06
Confident ballot secret	-7.33	1.70	-10.65	-4.00

Table 3: Effect of e-voting obtained using multivariate genetic matching

Conclusion

Given the decreased confidence in ballot secrecy for electronic voting, the European Commission should with priority investigate how to ensure ballot secrecy, as well as its perception by citizens. The other indicators suggest that electronic voting positively influenced the perception of voting, making it less difficult and made people feel more confident their vote was counted. These results align with the original paper's conclusions. However, the data these two papers are based on stems from a survey held in 2003. Since then, technologies and people's perception of it changed greatly, therefore we recommend to supplement these findings with newer methods and analyses if the budget for allows it.

Word Count: 1849

Code for this assignment:

Matching on Salta Data:
https://gist.github.com/anonymous/a2fbfe3b75f85055a473e8c7235344af

Genetic Matching on Salta Data:
https://gist.github.com/anonymous/3b0d757dc757560d24370144e0d087c9

Match Balance Comparioson Tables:
https://gist.github.com/anonymous/31c35814bdc63b27a494fa7cf388474c

Appendix

Appendix 1 - Replication

Replication Table 1

	N	prop.all	prop.ev	prop.tv	diff	pvalue
eselect.cand	1405	71.53025	83.99519	53.48432	30.510866	2.695479e-35
eval.voting	1486	36.27187	46.64391	21.25206	25.391854	2.457725e-23
easy.voting	1495	24.74916	34.01361	11.41925	22.594356	4.349730e-23
agree.evoting	1430	75.24476	84.08551	62.58503	21.500477	3.348404e-20
how.clean	1303	50.65234	57.69231	40.98361	16.708701	3.620959e-09
sure.counted	1444	82.34072	86.19883	76.74024	9.458593	5.066539e-06
capable.auth	1441	81.33241	84.99400	76.31579	8.678208	4.009402e-05
speed	1468	82.90191	84.30699	80.84034	3.466651	9.662635e-02
conf.secret	1455	80.06873	77.14959	84.15842	-7.008828	1.226635e-03

Replication Table 2

	ev	tv	diff	pvalue	ev	tv	diff	pvalue
age.group	2.475751	2.443350	0.03240082	0.58400000	2.451890	2.451890	0.00000000	1.0000000
educ	4.771363	4.142857	0.62850544	0.00000000	4.219931	4.206186	0.01374570	0.6860000
white.collar	30.254042	27.586207	2.66783467	0.29287524	29.209622	28.350515	0.85910653	0.7956642
not.full.time	27.713626	33.497537	-5.78391108	0.01998267	30.756014	31.958763	-1.20274914	0.7046433
male	49.653580	49.096880	0.55669955	0.87472467	48.969072	48.969072	0.00000000	1.0000000
tech	4.183603	3.909688	0.27391476	0.00000000	4.008591	3.941581	0.06701031	0.4940000
pol.info	1.474596	1.310345	0.16425102	0.00000000	1.360825	1.323024	0.03780069	0.5780000

Replication Table 3

	N	prop.ev	prop.tv	diff	pvalue	N	prop.ev	prop.tv	diff	pvalue
eselect.cand	1388	83.84333	53.41506	30.428268	1.237404e-34	1101	82.73381	54.12844	28.605373	3.059922e-24
eval.voting	1460	46.33295	21.29784	25.035108	1.832936e-22	1151	45.58059	20.90592	24.674666	1.141483e-18
easy.voting	1469	33.64269	11.53213	22.110566	5.420419e-22	1159	32.46978	11.89655	20.573224	6.309668e-17
agree.evoting	1409	84.14044	62.43568	21.704758	2.864300e-20	1114	82.43728	63.30935	19.127923	1.130750e-12
how.clean	1284	57.97297	40.99265	16.980326	2.561411e-09	1022	57.56972	41.53846	16.031260	4.149376e-07
sure.counted	1418	86.34731	77.01544	9.331868	7.444164e-06	1117	85.73975	76.97842	8.761333	2.294892e-04
capable.auth	1416	85.13514	76.24585	8.889288	2.953841e-05	1123	84.48905	76.00000	8.489051	4.835931e-04
speed	1443	84.05627	80.84746	3.208814	1.297603e-01	1137	83.21678	80.70796	2.508819	3.062633e-01
conf.secret	1431	77.10843	84.52579	-7.417357	6.506368e-04	1133	76.92308	84.32056	-7.397481	2.088613e-03

References:

Alvarez, R. M., Levin, I., Pomares, J., & Leiras, M. (2013). Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions. *Political Science Research and Methods*, 1(01), 117-137. doi:10.1017/psrm.2013.2

Alvarez, R. M., Levin, I., Pomares, J., & Leiras, M. (2015). Replication data for: Voting Made Safe and Easy: The Impact of e-voting on Citizen Perceptions. *Harvard Dataverse,* doi:10.7910/DVN/24896. Retrieved December 3, 2017, from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/24896

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3), 932-945. Retrieved March 10, 2016 from: Retrieved from https://service.sipx.com/service/php/inspect_document.php?id=perma-x-c66c955a-5f28-11e6-a73e-22000b61898b

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software,* 42(8):1–28. Retrieved December 13, 2017, from https://imai.princeton.edu/research/files/matchit.pdf

King, G., & Nielson, R. (2016). Why propensity scores should not be used for matching. [Working Paper]. Retrieved December 13, 2017, from https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching

Mebane, W. R., Sekhon, J. (2011). Genetic Optimization Using Derivatives: The rgenoud Package for R. *Journal of Statistical Software,* 42(11):1–26. Retrieved December 13, 2017, from https://www.jstatsoft.org/article/view/v042i11

Sekhon, J. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. *Journal of Statistical Software,* 42(7):1–52. Retrieved December 13, 2017, from http://sekhon.berkeley.edu/papers/MatchingJSS.pdf