Arkadiusz Kozłowski

THE USEFULNESS OF PAST DATA IN SAMPLING DESIGN FOR EXIT POLL SURVEYS

1. Introduction

Exit poll is a survey conducted on the election day in which respondents (voters) leaving the polling station answer, i.a. on who they cast their votes. This survey is so popular mainly thanks to the television stations, for which knowing the election results just after the polling stations have been closed, irrespective of the fact that the result is only approximate, allows them to first comments and live analysis on the election night, which guarantees a very high viewership.

The idea for this type of surveys was born is the US and there it was developed most intensively. As Frankovi (1992) says, the first survey on the election day took place in 1940 in Denver. The first exit poll in the form we know today, i.e. on a large scale and at the request of media, took place in 1967 and was conducted for CBS (Levy, 1983). The creation and development of survey methodology is ascribed to Warren Mitofsky (Moore, 2003). In Poland the first this type of research was conducted by Ośrodek Badania Opinii Publicznej (OBOP) during the first and second round of presidential election in 1990.

Exit poll is one of the few sample surveys, the results of which may be confronted with the complete enumeration and, what is more, in a very short period of time. From the statistical point of view, this gives a possibility of the immediate validation of the applied methodology. For the research centres conducting this type of surveys it is a kind of a challenge because the "malpractice" may cause them to lose their reputation and trust not only to a particular research centre but to the polls in general. In the group of surveys related to election, exit poll has a special place for a few reasons. Firstly, population of survey does not include all people entitled to voting but only people who actually vote. Thanks to that, on contrary to pre-election surveys, the "screening" problem of how to identify likely voters does not exist. Secondly, the questions in exit poll are related to facts and not intentions which may differ

from the actual election decisions. This issue is of particular importance especially in case of changing political preferences a few days before election (so-called late swing). As Hilmer (2008) emphasizes, the exit poll is more clear to respondents, an aim of it is more obvious and not arousing misgivings which result in lower non-response rate compared to other election surveys. Also the size of the sample (for Poland a tens of thousands) is far more higher than in standard surveys. With regard to above-mentioned reasons, the requirements of the survey's recipients concerning its precision are higher than the requirements concerning other election surveys.

However, the aim of exit poll is not only prediction of the election result. This survey delivers a lot of valuable information about votes distribution in different socio-demographic groups, the changes of political preferences in relation to previous election, the motives of choosing a particular party or candidate, the motives of choosing the time of voting etc. This information enables a thorough analysis of the results and will be used until the next election due to the fact that current political surveys, mainly of the above-mentioned reasons, do not provide so detailed data with the necessary precision. In the less stabilized democracies, exit polls indirectly perform a function of legitimacy of election and its results – the official results happen to be questioned if they differ from those obtained from the independent exit poll (the examples of such situations may be found in Andreenkova, 2008). Unintentional effect of exit poll may also be an influence on the potential voters' motivation to go to the polls if the preliminary results are announced before closing the last polling station. This problem concerns mainly the US where there is no legal prohibition on publishing surveys' results before all polling stations have been closed. This issue is widely discussed by, i.a. Seymour (1986), Lensky (2008).

2. Statistical aspects of exit poll

Exit poll is a two-stage survey. Primary stage units are precincts and the secondary stage units are voters. As long as selection of respondents to the sample is concerned there is an agreement between theorists and practitioners that the best choice in this case is a systematic sampling. This approach mainly results from the uneven distribution of particular party voters during the day, which was the object of study i.a. Klorman (1976), Busch and Lieske (1985). The significant influence on the choice of the time of day has an election day, in the US it is usually Tuesday, in the UK Thursday, i.e. working days. In Poland, as in the majority of countries, election takes place on holiday. Respondents chosen to the sample are interviewed by the use of self-administered questionnaire, which is then put in the envelope or deposited in the specially prepared ballot box. Bishop and Fisher (1995) proved that this mode of data

collection, called secret ballot decreases item non-responses and socially desirable responses compared to face-to-face interview, which is reflected in more accurate estimates.

Before pollsters begin interviews it is crucial to establish next to which polling stations the survey will be conducted. In Poland over twenty five thousands of precincts are created during the election. The sample reflecting most faithfully nationwide results needs to be chosen from this population. Barreto et al. (2006) state: "In fact, this is the most important step in exit polling". Unofficially, according to the one of the research centres, those past errors in Polish exit poll mainly result from the unrepresentative sample of polling stations. The conventional approach towards this issue is a random selection of precincts, however, this approach does not give the enough guarantee of representativeness of sample. Moreover, as Szreder (2007) emphasizes, relying only on the random sampling means in fact that the pollster admits that he/she lacks the valuable a priori knowledge about the surveyed population. Since such knowledge exists we should ask not "if" but "how" it should be used? One of the often utilized and widely accepted methods is the division of a population to strata. The choice of stratifying variables and establishing the number of strata is not an obvious thing. Additionally, optimal parameters may differ between countries and may change with time. For example Levy (1983), while characterizing American practices, mentions 2 to 6 strata created based on past voting behaviours, geographic regions, urban vs. rural counties, percent foreign stock, type of voting equipment, or poll closing times. The analysis of the distribution of results in particular strata from past election will surely make it easier to design strata properly. Another technique which also uses the data about past election results is tied sampling procedure. Tied sample means that the basis for creating election forecast is a sample of precincts which turned out to be the most representative one during the past election (Hofrichter, 1999). This consists in sampling a certain amount of samples (from a statistical point of view each of them is of the same value) and choosing the sample which reflected the particular past election results chosen as a reference point. Of course, this technique requires the complete data about election results on the level of precincts. If this data is unavailable, for example in the UK, the same precincts may be surveyed in the following elections and the own data collected in previous years may be used to correct the results. The successful appliance of this method in 2005 is described by Curitce & Firth (2008).

As far as the number of sampled polling stations is concerned, it is the result of a compromise between budget restrictions and the statistical theory. Fewer polling stations in a sample and more respondents from one precinct increase sampling error, whereas more polling stations in a sample requires more

pollsters, which increases the costs. Number of polling stations in a sample in Polish exit polls oscillates usually from five hundred to one thousand.

The aim of this paper is the empirical verification of the usefulness of a priori data, mainly information about past election results, in order to increase the quality, i.e. representativeness of polling stations sample in exit poll.

3. Data

Data about election results since presidential election 2000 on the level of precincts is widely available on the Państwowa Komisja Wyborcza (PKW) web sites. This data is great for simulative analysis of the process of sampling to exit poll. However, comparing the election results for particular precincts between two elections may cause some formal and substantive difficulties. According to voting system (Act from July, 16th 1998) the precincts are created by authority of municipality in the way that they include from five hundred to three thousand citizens. Between the successive elections the division of municipality to precincts may change, and in fact this happens very often, due to the change of municipality's borders, number of citizens in municipality or precinct, the change of the number of councillors in town council or the change in the division of municipality to electoral districts. The lack of the central supervision over creating precincts and the above-mentioned changes result in the lack of an unequivocal key to identify precincts between elections. Substantive difficulties result from the natural demographic changes (reaching voting age, deaths, migrations), voting outside voter's district (this phenomenon was very significant during the second round of presidential election 2010 due to untypical election day and the widespread information about such possibility) and changes on the political scene.

Databases shared by PKW include the following information:

- Territorial identification of unit (names and codes of voivodeships, counties and municipalities);
- Precinct number (numeration applied within municipality);
- Precinct address (location of board of elections);
- Type of territorial unit (city, urban area on the urban-rural area, rural area on the urban-rural area, village, districts of capital city Warsaw);
- Number of people entitled to vote;
- Number of ballots distributed (turnout);
- Valid votes;
- Number of votes cast on a particular committees/candidates.

Based on the comparison of the precinct number, address and the number of people entitled to vote, the precincts which have not changed during successive elections may be identified. From technical point of view, this

requires a scrupulous and hard work because due to the fact that there is incoherence in noting some of the variables (mainly address) it is impossible to apply one algorithm that would pair precincts from two elections. Taking that into consideration, decision was made to identify precincts only between two elections, i.e. presidential election 2010 (first voting) (WP10) and parliamentary election to the Sejm 2007 (WS07). The object of analysis is setting such a sampling plan that will maximize probability of choosing the best sample of precincts for estimation of results of WP10 by using detailed results from WS07.

In this analysis only the regular precincts were taken into consideration, excluding the precincts created in hospitals, prisons, detention centres, on ships, social welfare centres and abroad. In WS07, the committees which did not have candidates in all precincts were excluded. Additional information about combined districts is presented in table 1.

Table 1

Number of precincts and people entitled to vote in all and combined precincts

	Number of precincts		Number of people entitled to vote	
	total	combined	total	combined
WP10	25 774	22 964 (89,1%)	30 813 005	28 602 904 (92,8%)
WS07	25 476	22 964 (90,1%)	30 615 471	28 558 000 (93,3%)

4. Analysis

The object of analysis is the first stage of exit poll, i.e. sampling of the precinct. Due to this fact, the variability of the results occurring on the second stage, i.e. resulting from random sampling of respondents, is not taken into account. While calculating the result, the actual results in the sampled precincts were taken into consideration. As a measure evaluating the similarity of the sample to the whole population, average Manhattan distance has been used:

$$AMD_i = \frac{1}{n}\sum_{j=1}^{n}\left|p_{ij} - P_j\right| \cdot 100\% \tag{1}$$

where:
AMD_i – metrics for i^{th} sample,
p_{ij} – relative result of j^{th} committee/candidate in i^{th} sample,
P_j – relative result of j^{th} committee/candidate in the whole country,
n – size of sample (one hundred).

The same measure has also been used in calculation of the difference between the nationwide result and particular precincts' results. In order to make the results comparable, for every tested technique the size of sample is one hundred precincts.

Firstly, it was decided to experimentally check how tied sample procedure affects the effectiveness of sampling technique compared to simple random sampling. For this purpose, a following simulation was designed:

1) ***m*** independent samples were drawn (sampling without replacement),
2) from ***m*** samples the one which had the least AMD value in the parliamentary election to the Sejm 2007 was chosen,
3) for the chosen sample the AMD was calculated for the presidential election 2010,
4) points 1-3 were repeated one thousand times for five different ***m*** values.

In table 2 the results of the above-mentioned simulation are presented, i.e. basic descriptive statistics for the distribution of one thousand AMDs in relation to WP10, for different values of ***m*** parameter. The first case, in which number of generated samples equals one, is de facto simple random sampling.

Table 2

AMDs for tied sample procedure and with different ***m*** values

m	*Min.*	*1st Q*	*Median*	*Mean*	*3rd Q*	*Max.*
1	0,043	0,195	0,291	0,336	0,438	1,255
5	0,033	0,129	0,182	0,194	0,244	0,556
10	0,035	0,118	0,162	0,171	0,210	0,440
100	0,027	0,102	0,137	0,143	0,175	0,389
1000	0,029	0,094	0,128	0,136	0,170	0,318

It easily to notice that with increasing number of generated samples (***m***), out of which the best sample in terms of WS07 is chosen, the better samples are obtained in terms of WP10. Both average levels of AMDs and the dispersion of distribution are reduced. Tied sample procedure is thus an effective technique increasing the quality of the sample of precincts. Moreover, the great improvement of the results in relation to SRS is obtained just after 5 generated samples and by increasing ***m*** value to the level of 1000 the improvement is still noticeable but is not so significant.

The number of committees which had candidates in the whole country in WS07 is seven while the number of candidates in WP10 is ten. However, the vast majority of votes was obtained by three parties/candidates (apart from the fourth in the order in WS07 Polskim Stronnictwem Ludowym (PSL) with the result of 8,91%, committees/candidates outside top three obtained less than 3%

of the votes). Taking that into consideration, in the further analysis the estimation of the results only of the three most popular candidates is under focus. In table 3 the results of the above described simulation are presented, having regard only to the three highest results.

Table 3

AMDs for tied sample procedure and with different ***m*** values, only the three highest results

m	*Min.*	*1st Q*	*Median*	*Mean*	*3rd Q*	*Max.*
1	0,053	0,514	0,835	0,951	1,265	3,375
5	0,028	0,272	0,425	0,471	0,635	1,502
10	0,021	0,254	0,389	0,417	0,537	1,316
100	0,026	0,200	0,315	0,335	0,438	1,015
1000	0,012	0,195	0,302	0,329	0,431	1,067

In case of the 3 most popular candidates, the similar dependencies exist as in the case of all candidates, i.e. the smaller AMD values for the higher number of generated samples and the decreasing improvement of effectiveness. The further increase of ***m*** in the above simulation procedure would significantly increase the calculation needs and simulation execution time, so, in order to check if further increase of ***m*** value leads to the improvement of the results, the following procedure was proposed:

1) ***N*** independent samples were drawn (sampling without replacement);
2) for every sample the AMD was calculated in WS07;
3) 100 samples with the smallest AMD were chosen and for all of the them the AMD in WP10 was calculated (table 4).

Table 4

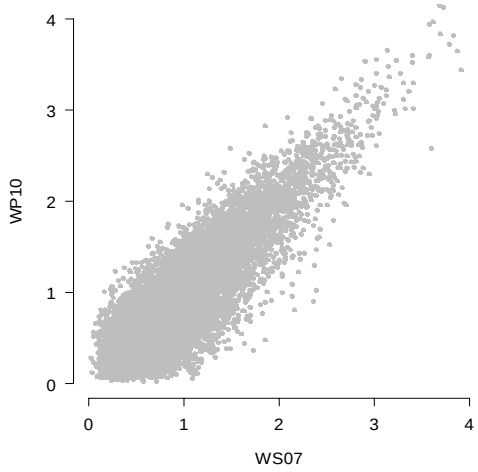
AMDs for the best 100 samples in WS07

N	*Min.*	*1st Q*	*Median*	*Mean*	*3rd Q*	*Max.*	*for min(WS07)*
10^3	0,040	0,223	0,344	0,371	0,523	0,830	0,375
10^4	0,061	0,209	0,324	0,350	0,466	0,806	0,278
10^5	0,060	0,195	0,310	0,334	0,434	0,846	0,434
10^6	0,060	0,193	0,307	0,341	0,432	0,927	0,060

Taking into consideration average values and quartiles, the results are getting better (only the average for ***N*** = 1 000 000 is worse than previous one), however the differences are relatively small. The computational capabilities of modern computers enables generating millions of samples without any problem, however, it seems that generating more than ten thousand samples does not make much sense. This follows from the fact that very good samples may be

obtained already with a several thousand draws, however, these samples are not necessarily the most representative ones in the previous election.

In the last column of table 4 the AMDs for the best samples in WS07 are presented. In two cases out of four the values are higher than the average. Based on that, the conclusion may be drawn that the choice of the best sample out of $\boldsymbol{N}$ generated not always is the best solution. This situation is illustrated by graph 1 and 2.

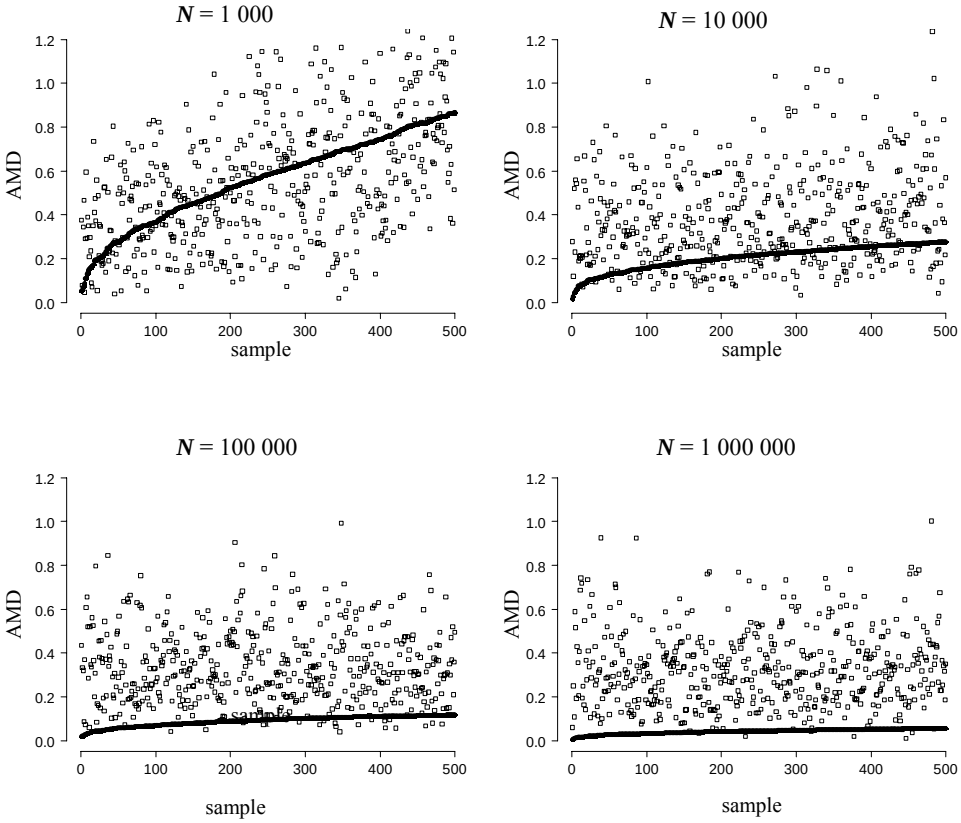


Graph 1. AMDs for N = 10 000 samples

As it is clearly shown in graph 1, there is a dependence between AMD in WS07 and AMD in WP10. However, the closer to the coordinate system's origin, the weaker the dependence. Therefore, the samples which are the closest to X-axis (AMD in WS07) are not always the closest to Y-axis (AMD in WP10). The similar thing may be noticed in graph 2, where the best five hundred samples in WS07, sorted in the non-decreasing order, and their AMD in WP10 are presented. With the increase of the number of generated samples ($\boldsymbol{N}$) the better samples in terms of the similarity to the general results in WP07 are obtained, however, the similarity of those samples to the general results in WP10 remains more or less the same and the values are pretty much dispersed.

Therefore, it seems justified to introduce modification to the tied sample procedure which would mean that not the best sample in terms of representativeness in the past election would be chosen to the survey but one out of one hundred to five hundred best samples would be selected. The more samples are generated ($\boldsymbol{N}$) the more justified the modification seems to be. The issue that needs to be further analyzed is how this one sample should be chosen. Perhaps, there are attributes which will allow to separate samples which will remain representative in the subsequent elections from those which

representativeness will significantly deteriorate. The author compared turnout, the number of people entitled to vote and the variability of the results of particular committees in the best samples, however, no significant differences were identified.



Graph 2. AMDs for the best five hundred samples (sorted) out of N generated (□ - WP10, ● - WS07)

Another method to increase the representativeness of sample is applying stratified sampling. Based on the analysis of the differentiation of results in WS07, eight strata were identified using following features: territorial division of the country (two variants: north-western area including nine voivodeships and south-eastern area including seven voivodeships) and the type of subdivision (four variants: cities – municipalities with the number of people entitled to vote over eighty thousand, towns, other urban area, rural area). Allocation proportional to the average of relative share of two features: number of people entitled to vote and number of precincts in a stratum was applied (table 5).

Table 5

Allocation of the sample

	Cities	Towns	Other urban area	Rural area
North-west	16	11	7	18
South-east	12	8	4	24

Subsequently, the stratified sampling was compared by simulation with unrestricted sampling using also the relation to the past results. For this purpose ten thousand samples were drawn in accordance with every scheme and descriptive statistics with AMDs were calculated in the way it was done in the previous simulations (table 6). On average, thanks to stratified sampling more representative samples were obtained compared to unrestricted sampling. The same advantage of stratified sampling occurs when the elements of tied sample procedure are introduced, i.e. out of the previously generated ten thousand samples, the one hundred most representative in WS07 are chosen. The samples obtained by applying this method, with given $\boldsymbol{N}$, turned out to be, on average, the most representative ones.

Table 6

Comparison of the results of simple random sampling simulation and stratified sampling simulation for N=10 000 samples

	Min.	*1st Q*	*Median*	*Mean*	*3rd Q*	*Max.*
	All samples					
SRS	0,023	0,483	0,811	0,940	1,277	4,576
Stratified	0,036	0,471	0,783	0,859	1,180	3,108
	The best 100 samples in terms of WS07					
SRS	0,061	0,209	0,324	0,350	0,466	0,806
Stratified	0,073	0,207	0,292	0,320	0,401	0,860

The last aspect of exit poll is an intentional omission of the smallest precincts in sampling procedure. Such behaviour from the research institution's point of view is acceptable due to financial benefits because while surveying the smaller number of large precincts, the same sample of respondents may be obtained with lower costs. Obviously, intentional exclusion of some of the units out of the sampling population affects the estimates. This fault, however, can be eliminated if tied sample procedure is applied, because the sample is chosen in such way that it reflects the total result, which also includes the precincts omitted in sampling. In table 7 the characteristics of the best one hundred (in terms of accuracy in WS07) out of ten thousand generated samples are presented, the number of sampling population was reduced in the subsequent rows by the precincts with the number of people entitled to vote smaller than $\boldsymbol{k}$.

Table 7

AMDs for the best 100 samples out of 10 000, with the omission of precincts smaller than k

k	*Min.*	*1st Q*	*Median*	*Mean*	*3rd Q*	*Max.*
500	0,048	0,216	0,342	0,348	0,465	0,872
600	0,024	0,213	0,319	0,338	0,434	0,821
700	0,025	0,197	0,277	0,316	0,409	0,802
800	0,061	0,229	0,321	0,345	0,437	0,932
900	0,028	0,228	0,345	0,366	0,491	0,822
1000	0,029	0,186	0,307	0,322	0,405	0,871
1100	0,059	0,217	0,349	0,368	0,465	0,862
1200	0,068	0,217	0,344	0,365	0,477	1,039
1300	0,073	0,231	0,401	0,423	0,560	1,162
1400	0,098	0,319	0,454	0,473	0,584	1,094
1500	0,109	0,336	0,474	0,534	0,696	1,222
2000	0,070	1,090	1,286	1,269	1,487	1,834

The above results show that in case of WP10 the omission of precincts smaller than six hundred to eight hundred people entitled to vote would not reduce the representativeness of sample but, in fact, it would increase it. Even limiting the sampling population only to precincts larger than one thousand people entitled to vote would not involve the loss of quality of the sample if the tied sample procedure is applied.

5. Conclusions

The detailed data about the past elections results is a valuable source of additional information enabling the improvement of sampling in exit poll. Based on the full results of the presidential election 2010 and the parliamentary election to the Sejm 2007 it was proven by means of simulation experiments that applying tied sample procedure significantly improves the representativeness of the sample. This benefit increases along with the growth of the number of generated samples. However, the more samples are generated the more advisable it is to modify the procedure in a way that instead of choosing the best sample from the past election, the choice is made from the first several hundred samples. The method of this choice requires further analysis. The improvement was obtained by applying stratified sampling in which the population was divided to eight strata based on two variables: geographic division and the type of territorial unit. It was also proven that by applying tied sample procedure the smallest precincts may be omitted in the survey, which is beneficial from financial and organizational point of view and does not further affect the results.

Acknowledgements

The author would like to thank Prof. Mirosław Szreder for the inspiration to conducting the research and valuable comments when writing the article and Mirosław Bogdanowicz from Krajowe Biuro Wyborcze for sharing the data in the useful format.

References

Andreenkova, A., Moreno, A. (2008) *Using exit polls to do more than project outcomes: the role and functions of exit polls in advanced and new democracies*. 3MC Conference Proceedings, Berlin.

Barreto, M.A., et al. (2006) *Controversies in exit poll*. "Political Science and Politics" Vol. 39, No. 3, 477-483.

Bishop, G.F., Fisher, B.S. (1995) *"Secret ballots" and self-reports in an exit-poll experiment*. "Public Opinion Quarterly" Vol. 59, No. 4, 568-588.

Busch, R.J., Lieske, J.A. (1985) *Does time of voting affect exit poll results?* "Public Opinion Quarterly" Vol. 49, No. 1, 94-104.

Curtice, J., Firth, D. (2008) *Exit Polling in a Cold Climate: The BBC-ITV Experience in Britain in 2005* [with Discussion]. "Journal of the Royal Statistical Society. Series A (Statistics in Society)" Vol. 171, No. 3, 509--539.

Frankovic, K.A. (1992) *Technology and the Changing Landscape of Media Polls*, in: T.E. Mann, G.R. Orren (eds) "Media Polls in American Politics". Brookings Institution, Washington, DC.

Hilmer, R. (2008) *Exit polls – a lot more than just a tool for election forecasts*, in: M. Carballo, U. Hjelmar (eds.) "Public opinion polling in a globalized World". Springer, Berlin.

Hofrichter, J. (1999) *Exit polls and elections campaigns*, in: B.I. Newman, (ed.) "Handbook of political marketing". Thousand Oaks, Sage Publications.

Klorman, R. (1976) *What Time Do People Vote?* "The Public Opinion Quarterly" Vol. 40, No. 2, 182-193.

Lenski, J. (2008) *New methodological Issues in conducting exit polls*. 3MC Conference Proceedings, Berlin.

Levy, M.R. (1983) *The methodology and performance of election day polls*. "Public Opinion Quarterly" Vol. 47, No. 1, 54-67.

Moore, D.W. (2003) *New Exit Poll Consortium Vindication for Exit Poll Inventor* Inside the polls, Gallup, http://www.gallup.com/poll/9472/new-exit-poll-consortium-vindication-exit-poll-inventor.aspx (30.09.11).

Seymour, S. (1986) *Do exit polls influence voting behavior*. "Public Opinion Quarterly" Vol. 50, No. 3, 331-339.

Szreder, M. (2007) *O roli informacji spoza próby w badaniach sondażowych*. "Przegląd Socjologiczny" LVI/1, 97-107.

Ustawa z dnia 16 lipca 1998 r., *Ordynacja wyborcza do rad gmin, rad powiatów i sejmików województw*. Dz. U. 1998, nr 95, poz. 602, art. 30.

UŻYTECZNOŚĆ DANYCH Z PRZESZŁOŚCI DLA PLANU LOSOWANIA W BADANIACH TYPU *EXIT POLL*

Streszczenie

Głównym zadaniem *exit poll* jest predykcja wyniku wyborczego tuż po zamknięciu lokali wyborczych. Nie mniej ważnym celem badania jest oszacowanie rozkładów głosów w różnych przekrojach społeczno-demograficznych. Kluczową kwestią dla jakości tych oszacowań jest wybór odpowiedniej próby obwodów głosowania. W artykule poddane zostały analizie alternatywne do losowania prostego metody doboru próby obwodów. Główny nacisk położono na wykorzystanie powszechnie dostępnych baz danych ze szczegółowymi wynikami przeszłych wyborów. Za pomocą eksperymentów symulacyjnych oceniono efektywność techniki powiązania wyboru nowej próby z przeszłymi wynikami (*tied sample procedure*) oraz wskazano optymalne dla niej parametry, a także zaproponowano pewną modyfikację procedury. Najlepsze wyniki uzyskano dla losowania warstwowego z zastosowaniem elementów procedury *tied sample*. Wskazano również możliwość redukcji kosztów badania bez straty na efektywności poprzez odpowiedni dobór wyłącznie dużych obwodów.