STATISTICS IN TRANSITION new series, Winter 2014
Vol. 15, No. 1, pp. 37–58

THE USE OF NON-SAMPLE INFORMATION IN EXIT POLL SURVEYS IN POLAND

Arkadiusz Kozłowski[1]

ABSTRACT

Exit poll is a commonly used tool to predict election outcome in democratic countries. The aims of this survey, however, go beyond the standard prediction which usually loses its value after 1-2 days. Lasting benefits of exit poll result from the possibility of estimating vote distribution in socio-demographic groups, changes of political preferences, the motives for choosing a candidate, etc. No other survey is capable of providing such detailed data with satisfactory precision. Nonetheless, the exit poll accuracy, both in Poland and abroad, often leaves much to be desired. It seems that while conducting the research the non-sample information is not used sufficiently, which could improve the quality and the precision of the survey.

The sources of auxiliary variables, which can be used in exit poll, along with the analysis of technical aspects of their acquisition and combination are outlined in this paper. Statistical methods aiming at incorporating the information about those variables to the survey, both at the stage of selecting the sample of precincts and at the stage of forecasting election results are proposed. Developed solutions were subjected to the simulation testing on the parliamentary election to the Sejm 2011 data. The results confirm the possibility of a significant increase in the effectiveness of estimates by improving 'representativeness' of a sample and by applying complex estimation of parameters.

Key words: exit poll, auxiliary variables, balanced sampling, complex estimation.

1. Introduction

An exit poll, conducted on the election day in which respondents (voters) leaving the selected polling stations answer, i.a. on who they cast their votes, is a commonly used tool to predict the election outcome in democratic countries. Thanks to work of a few thousand pollsters within appropriately organized logistics and IT operation, the TV viewers can know the approximate election results on the same day, right after the last polling station has been closed. These

[1] University of Gdańsk. E-mail: arekozlowski@wzr.ug.edu.pl.

estimates allow first commentaries and live analysis to be presented on the election night (which usually guarantees wide audience).

The aim of exit poll is not only to predict the election result (the very same forecast quickly becomes useless). The survey gives the opportunity for in-depth analysis of voting results in such aspects as vote 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 participating in election, etc. These analyses remain to be the most reliable source of citizens' political behaviours until the next election, due to the fact that current political surveys are unable to provide such detailed data with the necessary precision (Szreder, 2011).

In terms of statistics, exit poll is a unique survey because it is a sample survey research the general result of which is quickly confronted with the result of complete enumeration (the same cannot be said about pre-election surveys as the population of interest is much larger – apart from actual voters, the pre-election survey encompasses also people entitled to vote but not participating in the election). The degree of compatibility between the general forecast and official results announced by the National Electoral Commission (PKW) is the basis for validation of applied methodology, and also has influence on the quality of more detailed data sets.

The key elements of the exit poll methodology are the sampling design and the method of estimation. The sample is chosen in two stages. In the primary stage the precincts are sampled and in the secondary stage the voters leaving the polling station are chosen. As far as the selection of the respondents to the sample is concerned there is an agreement between theorists and practitioners that the best choice in this case is systematic sampling (in Poland it is usually every tenth person leaving polling station). This approach mainly results from the uneven distribution of particular party voters during the day, which was the object of study, for instance by Klorman (1976), Busch and Lieske (1985). It is especially important in countries where the election day falls on working day, like in the US (Tuesday) and the UK (Thursday). In Poland, as in the majority of countries, election takes place on holiday.

The more problematical stage is the choice of accurate sample of precincts which would be the most representative of the population. Barreto et al. (2006, p. 479) state, "In fact, this is *the* most important step in exit polling", suggesting that selecting the inaccurate sample of precincts was the reason of unsatisfactory exit poll results during U.S. presidential election 2004 (the survey conducted by Edison-Miofsky Research forecasted the victory of John Kerry over Georg W. Bush). In Poland around 25-26 thousand of precincts are created during the election (around 24 thousand are the so-called regular precincts). The conventional approach towards the issue of sampling the precincts is stratified sampling, in which strata are created based on geographical regions and the type of territorial unit (city/village). Such a solution increases the representativeness of a sample compared to unrestricted sampling, however, by increasing the degree of

the use of a prior knowledge about the population of interest, the representativeness can be further increased. The sources of knowledge about population can be, on the one hand, the official statistics of the Central Statistical Office (GUS), on the other hand, the official results of the past elections shared by PKW. PKW databases are particularly valuable because they contain detailed voting results for each of over 25 thousand precincts.

The same data sources which help selecting the right sample (supporting sampling process) can also support the process of estimating the election result. It is commonly known that electorates of particular political groups vary between themselves in respect of many demographic, economic and social variables. If this type of characteristics was known for every precinct, it would be, along with information about the previous elections vote structures, a rich source of auxiliary variables for complex estimations.

The aim of this paper is the empirical verification of the assumption that incorporating additional auxiliary variables to the exit poll strategy increases effectiveness of estimating the election result. The additional variables are non-sample and are not directly connected with the unit of research, i.e. the precinct, therefore, statistical analysis is preceded by the presentation of technical aspects of data acquisition and combination along with pointing out the advantages and limitations of a given source. At the stage of selecting the sample, an innovative method of balanced sampling, the so-called cube method, is applied. The object of the analysis is estimation of relative result, i.e. the fraction of votes cast on particular committees across the country. Proposed strategies are subjected to simulation testing on the parliamentary election to the Sejm 2011 data.

2. Characteristics of exit poll

The first exit poll was conducted in the United States in the 60s of the 20th century at the request of CBS (Levy, 1983). The creation and development of the 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 differs from other political preferences surveys in many aspects. First of all, the population of interest is different. Apart from actual voters, the political surveys encompasses also people entitled to vote but not participating in the election, whereas the exit poll surveys only people taking part in the election (this is one of the main arguments of the opinion research centres refuting accusations of the discrepancy between pre-election surveys and the actual election results). Secondly, the exit poll questions refer to facts (the actual votes), not to the intentions which often are different from the actual voting decisions. The survey is conducted directly after leaving the polling station which minimizes the errors connected with ‘gaps in memory’ and the ‘bandwagon effect’ due to the

fact that the final result is still unknown. Another distinguishing feature is much higher percentage of the conducted interviews. In standard surveys conducted by applying the CATI method, the percentage goes up to several percents, in face-to-face interviews it goes up to 50-60% whereas in exit polls in Poland it remains at the level of 85-95% (Domański et al., 2010). It is worth emphasizing that this is a particularly high level, practically unparalleled in the Western countries, where the general trend for the decrease of the sample response rates has also affected exit poll (the examples of the response rate: Germany: 70-72% (Hofrichter, 1999), USA: 45-55% (Lenski, 2008), Great Britain: 86% (Moon 2008)). Furthermore, the survey scale is also noteworthy – the sample size measured by the number of individual respondents is usually several times higher than the sample size of a standard pre-election survey. For example, during the parliamentary election 2011, TNS OBOP conducted research on the sample of 900 polling stations, conducting around 100 000 interviews in total.

The above-mentioned characteristics raise the value of the exit poll information compared to other preferences and political behaviours surveys. The challenge to maintain this advantage is that in a few countries there is a possibility of voting indirectly, not in a polling station (the so-called *absentee ballot*), i.e. via mail, Internet or attorney. Additionally, people voting through mail can do this within a certain period of time before the official election day. This complicates the survey and forces the organizer to apply different techniques, e.g. telephone surveys, in order to supplement the interviews conducted in front of polling stations. However, for the time being, this is not a problem of Polish researchers.

The main focus of this paper is on reducing errors relating to the selection of precincts and estimation, nonetheless, it is worth mentioning other potential sources of errors. They mainly occur during the selection of voters and interaction between the respondent and interviewer. One of them is the faulty implementation of systematic sampling scheme. The threat is that the selection discipline of taking every n^{th} person will break down and interviewers will approach individuals who they think will respond. That would naturally introduce selection bias. Another problem is that of *co-location* of precincts, when two or more precincts are in the same physical facility. For someone who is leaving such a facility it is usually not clear which precinct he/she voted in, which makes it difficult to maintain the desired selection probabilities (Scheuren and Alvey, 2008, p.12).

Another potential source of errors are non-respondents. As stated above, the response rate in Polish exit polls is still quite high, but it will probably decline in the near future. There are two types of non-respondents in exit polls: refusals and misses. A refusal is when a sampled voter refuse to participate in the survey and a miss is when a sampled voter cannot be asked to fill out the questionnaire because the interviewer is too busy or the voter does not pass the interviewer. Merkle and Edelman (2002) estimate that about three-fourths of non-response in exit polls is attributable to refusals and about a quarter to misses. Refusals pose a greater threat to the survey outcome than misses because the voter's reluctance to participate in exit polls can result from specific political attitudes and

consequently cause certain bias. Merkle and Edelman (2002) studied factors correlated with non-response on the basis of various exit polls conducted from 1992 through 1998 in the USA. Conclusions from their investigation are as follows: of the voter's characteristics, age is the most strongly related factor to the response rate (older voters have lower response rates), of the Election Day factors, interviewing position at the polling place (closeness to the polling place) is the main factor that can have a dramatic effect on response rates (response rates decline as the interviewer moves further away from the voting room), and of the interviewer characteristics, again age is the most significant factor (the age of the interviewer is positively correlated with response rates). What is surprising is the authors found very little or no correlation between response rates and exit poll error measures. Neither refusal rates nor miss rates were significant predictors of errors.

Apart from unit non-response, when information is missing on all questionnaire variables, researchers conducting exit polls experience item non-response (when only some answers are missing) and false answers. The crucial factor for the scale of these occurrences seems to be the mode of data collection. One of the most popular solution here is the so-called secret ballot. In this mode 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 experimentally that this mode of data collection decreases item non-responses and gives socially desirable responses, compared to the face-to-face interview. Using secret ballot one assumes that voters can read and understand questions well enough to give a reasonable answer. This can not be the case in countries with low literacy level. Bautista et al. (2006) give the example of Mexico, where due to low educational level mixed-mode of data collection were used (face-to-face with secret ballot).

3. Sources of additional information

Exit poll does not always meet recipients' expectations as far as the compatibility between the forecast and the actual result is concerned, irrespective of the fact that the survey is conducted on a large sample size with a low rate of refusals and potentially low measurement error. As a result, the need to strengthen the survey with non-sample information arises. Two main sources of non-sample information are specified: data referring to past election results and GUS data not directly connected with the elections but strongly related to voters' decisions.

As far as the past general election data is concerned, the election results starting from the presidential election 2000 are available on each aggregation level (from voivodeship to precincts) on the PKW website. The key issue for using this data is the possibility of confronting the results for two or more elections between the corresponding precincts. This process, however, causes some problems. The main issue is that according to the Election code (2011) the

division into the election precincts is made by authority of municipality, however, the division is not permanent. Before each election a new division is made and both the number and the boarders of precincts within municipality can change. Further difficulties arise from typical demographic changes (reaching voting age, deaths, migrations), voting outside the voter's district. The above-mentioned reasons lead to a situation in which the voters from *i*-th precinct in the *X* municipality are not exactly the same voters who participated in the elections a few years ago. However, the differences are insignificant, thus the informative value of the past election results should remain high.

Another aspect of the use of information about past results is choosing the elections which will serve as a reference point to strengthen the estimates of the current survey. The most reasonable option seems to be choosing the chronologically nearest election as in such case, the changes on the political scene along with demographic and organizational changes are not so significant. In the case of the parliamentary election 2011 such a reference point can be the presidential election 2010 (some of the main candidates can be linked to political parties). However, the parliamentary election's character is different in respect of the division into electoral districts and different set of candidates in each district, hence the parliamentary election 2007 can be considered as a better reference point. The question arises as how much the informative value of the data has deteriorated due to the changes in precincts during 4 years. Another possibility is choosing the European Parliament election 2009, however, irrespective of being nearest in time, this choice has some drawbacks, i.e. low election turnout (24,53%), different division into electoral districts and generally speaking different attitude towards the European election among both the voters and the politicians than towards the national elections. In the conducted simulation analysis the use of the presidential election 2010 results (first round) and the parliamentary election to the Sejm 2007 results was studied.

Only the technical issue of matching the corresponding precincts in the mentioned elections needs to be resolved. Due to the fact that the division into precincts lies within the competence of municipality, there is no main key identifying the precincts between elections. By comparison of the precinct address, precinct number (numeration applied within municipality) and the number of registered voters, the corresponding precincts can be identified with the high credibility. The probably correctly linked precincts in which the difference in the number of people entitled to vote exceeded 200 were excluded (this operation also eliminates the precincts in tourist resorts, in which the vast majority of voters are out-of-towners). In further analysis only the regular precincts were taken into account, as only this type of precinct encompasses relatively unchanged voter groups. The number of linked precincts and the number of registered voters in comparison with the whole population are presented in Table 1.

Table 1. Data on populations subjected to simulation test

	Number of precincts (Number of registered voters)			
	all	regular	linked S11 – P10	linked S11 – P10 – S07
Population:	U	U1	U2	U3
Sejm 2011 (S11)	25 993 (30 762 931)	24 217 (30 387 730)	23 553 (29 382 340)	22 209 (27 305 152)
Presidential 2010 (P10)	25 774 (30 813 005)	24 144 (30 382 814)	23 553 (29 388 485)	22 209 (27 349 352)
Sejm 2007 (S07)	25 476 (30 615 471)	23 903 (30 188 868)		22 209 (27 332 149)

Source: Own calculation based on PKW data.

Another source of information that can increase the quality of exit poll are GUS official statistics for the units of territorial division of the country referring to social and economic characteristics. In this case, the main factors limiting the possible uses are: the level of data aggregation, the range of described population and the timeliness of data. As far as the aggregation level is concerned, the most helpful would be the data at the level of precincts, which of course does not exist. The lowest available level of aggregation is municipality and only for a limited range of variables. The second limitation is a problem due to the fact that GUS data refers naturally to the whole population and not only to the active voter groups which are analyzed in exit poll. As far as the timeliness of data is concerned, it depends on the type of variables, however, usually a few years' delay in relation to the election date has to be taken into account. In that case, the most reasonable approach is to use the features which do not change significantly in time.

Despite these limitations, it is believed that incorporating certain variables can improve both selecting a sample and the result estimations. After the analysis of available socio-economic data and their relation to the past election results, it was decided to incorporate to the study two variables at the level of municipality:

- economic entities registered in REGON per every 10 000 population (2010, *podm_gm*),
- the area of agricultural land (in ha) per every 1 000 population (2005, *uzyt_gm),*

and two variables at the level of powiat (the second-level unit of local government and administration in Poland):

- registered unemployment rate (2011, *bezr_pow),*
- the average monthly gross salary in comparison to national average salary (2010, *wyn_pow).*

The very valuable source of the supportive information can be the distribution of voters in view of the features like sex, age and education. The character of election does not enable the official collection of such data, however, the opinion research centre conducting such survey in the past has own estimates at its disposal and can use them to correct the estimates at the level of a single unit. In view of the fact that presented analysis encompasses only official data in which the most detailed information is the general election result in the precinct, this possibility is not taken into consideration.

4. Sampling plan

The proposed sampling technique, which expands the conception of restricted sampling compared to typical stratified sampling, is balanced sampling. The sampling design is called balanced in relation to certain additional characteristics (*auxiliary variables)* if it generates samples from which the estimates of additional variables sums (by Horvitz–Thompson estimator, HT) match the known actual sums (Deville, 2004). In other words, in balanced sampling the auxiliary variables are estimated without an error. The above definition can be generalised for any samples, not necessarily chosen in random sampling.

The idea of balanced sampling is not new. It appeared along with the representative method and is connected with the very same term of representativeness. The first use of this conception in practice refers to famous sampling of precincts during the Italy census (Gini, Galvani, 1929, after Langel, Tille, 2011). 29 precincts were selected in such way that the averages from the sample for a few auxiliary variables would match the average from population. Both Nayman and Yates (Langel, Tille, 2011) condemned such behaviour as the sample was selected purposive. It was later observed that the balanced sample can be selected in a probabilistic way. A special example is stratified sampling, in which the sample is random and at the same time balanced on the specified indicator variables of the strata (such variables take on the value 1 for the units belonging to stratum and, otherwise, 0; the number of variables corresponds with the number of strata).

From a technical side, the probabilistic way of selecting a balanced sample is not evident. There is a number of methods enabling this choice, the majority of which is based on elimination process (the so-called *rejective sampling*), i.e. rejecting some of the sampled units (or the whole sample) if the condition of the balance is not satisfied (one of the variants of the method is the so-called tied sampling (Kozłowski, 2012). This requires conducting a number of interactions, which, depending on complexity of limiting conditions, are more or less time-consuming. The majority of methods also have some constraints resulting from the possible applicability only in a chosen sampling schemes, the lack of possibility of differentiating inclusion probabilities as well as limited number and type of auxiliary variables. The method which overcomes these difficulties and, in

this context, is the most general is the so-called cube method proposed by Deville (2004).

The starting point in the cube method is geometrical conceptualisation of all possible samples (in sampling without replacement) of N-element population as vertices of N-cube C, i.e. $C=[0,1]^N$. Any sample s is defined as a vector $(s_1,\ldots,s_k,\ldots s_N)'$, where s_k takes on the value 1 if the k^{th} unit is in the sample and, otherwise, 0. The number of all possible samples (of any size) equals to the number of the vertices of the cube C, i.e. 2^N. In the instance of 3-element population (N=3) the sample space can be presented as vertices of a cube (Figure 1). Starting from the point defined by the vector of the first order inclusion probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_k,\ldots,\pi_N)$, the selection of the sample can be illustrated as random 'reaching' to the one of vertices.

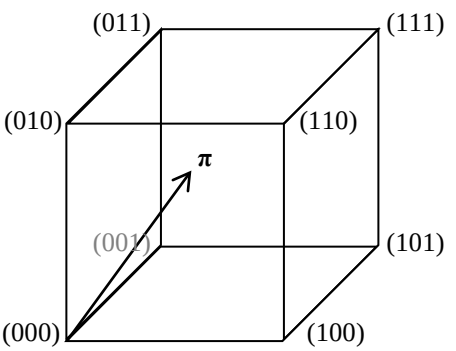

Figure 1. Geometrical representation of sample space for 3-element population

Source: Deville, 2004, p. 896.

The design is balanced on auxiliary variables only if the data is at the unit level, i.e. for every unit of the population the vector $\boldsymbol{x_k}=(x_{k1},\ldots,x_{kj},\ldots,x_{kp})'$ should be known, where p – the number of auxiliary variables. The totals of auxiliary variables $\boldsymbol{X}=\sum_{k\in U} x_k$ are estimated by $\widehat{\boldsymbol{X}}_{HT}=\sum_{k\in U}\frac{x_k S_k}{\pi_k}$, where U – population. The balanced sampling, as per definition, reassures:

$$\widehat{\boldsymbol{X}}_{HT}=\boldsymbol{X} \tag{1}$$

for every possible sample, in other words $\widehat{\boldsymbol{X}}_{HT}$ variation equals 0. In practice, this condition is usually fulfilled only approximately. The equation (1) is a set of limiting conditions, which defines an affine subspace (hyperplane) Q in $\mathbb{R}^N$ in dimension *N-p*. The idea of balanced sampling is to randomly 'reach' to such a vertex of cube C, which at the same time belongs to hyperplane Q (exactly balanced) or is located as close as possible (approximately balanced).

In the cube method the vector $\boldsymbol{\pi}$ is randomly transformed into a vector containing only values 0 and 1 (i.e. vector $\boldsymbol{s}$ defining a sample) in such a way that inclusion probabilities are exactly satisfied and balance condition (1) for every variable is satisfied to the furthest extent possible (Tillé, 2011). The method is divided into two phases: flight phase and landing phase. Flight phase is a random walk on in the intersection of the cube C and the constraint subspace, which starts from the point defined by the vector $\boldsymbol{\pi}$ and ends on the vertex of intersection ($\boldsymbol{\pi}^*$). If the reached point is at the same time the vertex of the cube C (i.e. all elements of $\boldsymbol{\pi}^*$ equals 0 or 1), then the balance is exactly satisfied and the process of sampling is finished. Otherwise, the landing phase begins in which (by applying linear programming) the vertex of cube C located as close as possible to the point reached in the flight phase and at the same time satisfying the inclusion probabilities is set.

In most cases the perfect balance cannot be achieved due to the so-called rounding problem. Nevertheless, it is proved that (Tillé 2006, p. 165):

$$\left|\widehat{X}_{j\,HT} - X_j\right| \leq p * \max_{k\in U}\left|\frac{x_{kj}}{\pi_k}\right| \tag{2}$$

The accuracy of the balance is decreased along with the increase in the amount of auxiliary variables, and is improved along with the increase in a sample size if it is set before the sampling.

5. Methods of estimation

The estimated parameter is the fraction of votes cast on *J* committee, which can be presented as a quotient of two sums:

$$P_J = \frac{Y_J}{Y} = \frac{\sum_{k\in U} y_{Jk}}{\sum_{k\in U} y_k} \tag{3}$$

where:

Y_J – the sum of valid votes cast on the *J* committee across the country,

Y – the sum of valid votes in total across the country,

y_{Jk} – the number of valid votes cast on *J* committee in the k^{th} precinct,

y_k – the number of valid votes in total in k^{th} precinct.

The problem of estimation is to estimate the total number of valid votes and the total number of valid votes cast on *J* committee based on *n*-element sample of precincts. In the case of both sums, it was decided to test three types of estimators: Horvitz-Thompson estimator (HT), ratio estimator (q) and estimator using the log-linear model (P). The model is the so-called Poisson regression – the type of a generalized regression model, in which it is assumed that the response variable Y has a Poisson distribution. The function linking linear combination of explanatory variables with the response variable is a natural logarithm. This model was chosen because it is particularly useful in the analysis of count variables (taking on integer nonnegative values).

The Horvitz-Thompson estimators used for estimating the sum of votes in total and the sum of votes cast on *J* committee respectively are presented with the following formulas:

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{4}$$

$$\hat{Y}_{J,HT} = \sum_{i \in s} \frac{y_{J,i}}{\pi_i} \tag{5}$$

where:

y_i, $y_{J,i}$ – the number of valid votes, in total and on *J* committee respectively, cast in i^{th} precinct selected to the sample,

π_i - the first order inclusion probability.

The Horvitz-Thompson estimator does not use the auxiliary variables directly, however, it can use them indirectly if in the sampling design the additional variable is used to establish the inclusion probabilities.

The analogous set of ratio estimators is as follows:

$$\hat{Y}_q = \frac{\hat{Y}_{HT}}{\hat{X}_{HT}} X \tag{6}$$

$$\hat{Y}_{J,q} = \frac{\hat{Y}_{J,HT}}{\hat{X}_{J,HT}} X_J \tag{7}$$

where:

$\hat{X}_{HT}$, $\hat{X}_{J,HT}$ - HT estimators of the sums for auxiliary variables,

X ,X_J – the known sums of auxiliary variables.

Ratio estimators use the information about one auxiliary variable for which the values from the sample and the actual sum in population are known. In the conducted analysis the auxiliary variable is usually the same variable but in the past (e.g. the result of the same committee in the previous election).

In the third case, the numbers of votes were modelled by the Poisson regression, by using the following formula (the same for both parameters):

$$\hat{y}_k = e^{x_k \hat{\beta}} \cdot X_k^* \tag{8}$$

where:

$\hat{y}_k$ - the theoretical number of votes in k^{th} precinct,

$\boldsymbol{x_k}$ – the vector of explanatory variables (independent variables),

$\widehat{\boldsymbol{\beta}}$ – the vector of regression factor estimated based on the sample *s*,

X_k^* - offset variable.

The aim of the offset variable, which is added to the basic model, is to eliminate the differences between number of votes resulting only from the precinct's size. In the case of modelling the number of votes in total, the offset variable was the number of registered voters in S11, whereas in the case of modelling the number of votes cast on committee *J*, the offset variable was a previously estimated total number of votes. Estimation of the sum of votes across the country is the simple prediction aggregation for all precincts in the population.

The estimators are presented as follows:

$$\widehat{Y}_P = \sum_{k \in U} \hat{y}_k \tag{9}$$

$$\widehat{Y}_{J,P} = \sum_{k \in U} \hat{y}_{Jk} \tag{10}$$

where:

$\hat{y}_k$, $\hat{y}_{Jk}$ - the theoretical number of votes in k^{th} precinct according to the model (8), in total and on committee *J*, respectively.

The general rule was adopted (with the exception of formula (12)) that the final estimator of fraction of votes cast on committee *J* would be the quotient of two sums estimated with the same type of estimator. Complex estimators can use supportive data from other sources and to different extents, which results in a high number of possible variants. Finally, it was decided to separate seven estimators, the effectiveness of which will be subjected to simulation testing later in this paper:

$$\widehat{P}_J^{(HT)} = \frac{\widehat{Y}_{J,HT}}{\widehat{Y}_{HT}} \tag{11}$$

$$\widehat{P}_J^{(Q-S11)} = \frac{\widehat{Y}_{J,HT}}{\widehat{Y}_q^{(S11)}} \tag{12}$$

where:

$\hat{Y}_q^{(S11)}$ - the ratio estimator according to formula (6) in which the auxiliary variable is the number of registered voters during the parliamentary election to the Sejm 2011;

$$\widehat{P}_J^{(Q-P10)} = \frac{\widehat{Y}_{J,q}^{(P10)}}{\widehat{Y}_q^{(P10)}} \tag{13}$$

where:

$\widehat{Y}_{J,q}^{(P10)}$ - the ratio estimator according to formula (7), in which the auxiliary variable is the number of votes cast on the candidate linked to the committee J during the presidential election 2010 (in the case of the Ruch Palikota (RuchP) committee, the auxiliary variable was the result of Bronisław Komorowski committee),

$\widehat{Y}_q^{(P10)}$ - ratio estimator according to formula (6), in which the auxiliary variable is the number of valid votes in total during the presidential election 2010;

$$\widehat{P}_J^{(Q-S07)} = \frac{\widehat{Y}_{J,q}^{(S07)}}{\widehat{Y}_q^{(S07)}} \tag{14}$$

where:

$\widehat{Y}_{J,q}^{(S07)}$ - the ratio estimator according to formula (7), in which the auxiliary variable is the number of votes cast on the same political party during the parliamentary election to the Sejm 2007 (in the case of Ruch Palikota committee, the auxiliary variable was the result of Platforma Obywatelska RP committee),

$\hat{Y}_q^{(S07)}$ - the ratio estimator according to formula (6), in which the auxiliary variable is the number of valid votes in total during parliamentary election to the Sejm 2007.

$$\hat{P}_J^{(Poiss-S11)} = \frac{\hat{Y}_{J,P}^{(S11)}}{\hat{Y}_P^{(S11)}} \tag{15}$$

where:

$\hat{Y}_{J,P}^{(S11)}$ - the estimator according to formula (10) based on the model (8) with the explanatory variables:

teren – the type of the area where the precinct is based (*large city* – above 80 thousand registered voters, *town*, *village*),

region – the group of voivodeships (*first group:* Małopolskie, Podkarpackie, Świętokrzyskie, Lubelskie, Łódzkie, Mazowieckie, Podlaskie; *second group:* the remaining voivodeships),

podm_gm, uzyt_gm.

$\hat{Y}_P^{(S11)}$ - the estimator according to formula (9) based on the model (8) with the explanatory variables:

podm_gm, uzyt_gm, bezr_pow, wyn_pow, terrain.

$$\hat{P}_J^{(Poiss-P10)} = \frac{\hat{Y}_{J,P}^{(P10)}}{\hat{Y}_P^{(P10)}} \tag{16}$$

where:

$\hat{Y}_{J,P}^{(P10)}$ - the estimator according to formula (10) based on the model (8) with explanatory variables:

P10_Komorowski – the number of votes cast on Bronisław Komorowski during the presidential election 2010,

P10_KaczynskiJ – the number of votes cast on Jarosław Kaczyński during the presidential election 2010,

P10_Napieralski – the number of votes cast on Grzegorz Napieralski during the presidential election 2010,

podm_gm, uzyt_gm, terrain, region.

$$\hat{P}_J^{(Poiss-S07)} = \frac{\hat{Y}_{J,P}^{(S07)}}{\hat{Y}_P^{(S07)}} \tag{17}$$

where:

$\hat{Y}_P^{(P10)}$ – the estimator according to formula (10), based on model (8) with the explanatory variables:

S07_PO – the number of votes cast on Platforma Obywatelska RP committee during the parliamentary to the Sejm 2007,

S07_PiS – the number of votes cast on Prawo i Sprawiedliwość (PiS) (Law and Justice) during the parliamentary election to the Sejm 2007,

S07_LiD – the number of votes cast on Lewica i Demokraci (LiD) (Left and Democrats) during the parliamentary election to the Sejm 2007,

terrain, region, podm_gm, uzyt_gm.

$\hat{Y}_P^{(S07)}$ - the estimator according to formula (9) based on model (8) with explanatory variables:

S07_votes – the number of valid votes in total during the parliamentary election to the Sejm 2007,

podm_gm, uzyt_gm, bezr_pow, wyn_pow, S07_PO, terrain.

6. Description of simulation

In the conducted simulation test the single-stage cluster sampling was applied instead of two-stage sampling typical for exit poll, which results from the character of available data. No sampling at the second stage was simulated as no unit information being capable to support estimation process was available, therefore, the result in the sampled precinct was taken as given without errors. The sampled units are precincts, i.e. the groups of voters participating in voting.

For reference purposes along with the balanced sampling, the simple random sampling without replacement (SRS) and stratified sampling were tested (STRAT). The division into strata was made based on the variation of the past election results in the section of following variables: *teren* and *region.* 6 strata were created as combination of 3 variants of a *teren* variable and 2 variants of a *region* variable. With regard to the large disproportion between the number of precincts and the number of votes cast in a stratum, the location was chosen proportionally to the number of valid votes cast in the parliamentary election to the Sejm 2007.

The balanced sampling was conducted in 3 variants depending on the type of auxiliary variables that were used:

- balance in reference to GUS variables (podm_gm, uzyt-gm, bezr_pow, wyn_pow) and the number of registered voters during S11 (BALS11),
- balance in reference to GUS variables (podm_gm, uzyt_gm, bezr_pow, wyn_pow) and the variables from the presidential election 2010 (P10_votes, P10_Komorowski, P10_KaczynskiJ, P10_Napieralski) (BALP10),

- balance in reference to GUS variables (podm_gm, uzyt_gm, bezr_pow, wyn_pow) and variables from the parliamentary election to the Sejm 2007. (S07_votes, S07_PO, S07_PiS, S07_*LiD)* (BALS07).

Additionally, each balanced sample was at the same time a stratified sample according to the above-described scheme. Within strata the precincts were sampled with the same probability of being selected, however, between strata the probabilities differed due to the allocation being disproportionate to the number of precincts in a stratum. In stratified sampling the estimators using auxiliary variables had a form of combined estimators, which means that the model is estimated for the whole sample altogether and not separately for each stratum like in the case of separate estimators. Due to small sizes of a sample in strata separate estimators would be in this case less stable.

The use of the past election results, due to the incomplete link of precincts between elections, implies the restriction of frame population to the set U2 or U3 (see Tab. 1). Even in the case of using only GUS variables, or if auxiliary variables are completely excluded, the frame population is restricted to regular precincts (set U1), which reflects the practical way of conducting the research. Nevertheless the aim of the survey is to estimate the actual fraction of the whole population (U). Thus, it seems appropriate to validate the estimates against non-included units. The correction is not necessary in the case of ratio estimators, which use the sum of additional features for the whole population, thus the estimates can be generalized to the entire population U. The estimates obtained by using estimators based on the log-linear model can be generalized only to the particular frame population (U1, U2 or U3). The same applies to the Horvitz-Thompson estimator, due to the restriction of frame population to the regular precincts. Therefore, the part of estimators was extended with the correction based on the past election results of the entire population in relation to the result of the particular frame population. General formula of the correction is as follows:

$$\widehat{\mathrm{P}}_{\mathrm{J}}^{*} = \widehat{\mathrm{P}}_{\mathrm{J}} \frac{\mathrm{P}_{\mathrm{J}\prime}^{(\mathrm{W})}}{\mathrm{P}_{\mathrm{J}\prime}^{(\mathrm{W,Ui})}} \tag{18}$$

where:

$P_{J\prime}^{(W)}$ – the actual fraction of votes cast on committee/candidate linked to the committee *J* in the election W∈{S07, P10}, in population U,

$P_{J\prime}^{(W,Ui)}$ – the actual fraction of votes cast on committee/candidate linked to the committee *J* in election W, in population Ui (Ui ∈ {U1, U2, U3}).

The correction applied only to irregular precincts (Ui=U1) is based on the assumption that the voters abroad, in prisons, on vessels, etc. are different from the rest of voters and the directions of those differences remain constant over at least a few years. The analysis of the past election results indicates the presence of some constant trends, i.a. the result of PO in irregular precincts was usually higher than the result in regular precincts (pkw.gov.pl). These trends, however, do

not have to sustain in the future, thus in case of Horwitz-Thompson estimators it was decided to test both estimators with correction or without it.

Juxtaposition of all sampling plans and methods of estimation, taking into account the fact that not all combinations are possible, gives 30 possible strategies of research. Some of the strategies use the same auxiliary variables, both at the stage of selecting the balanced sample and at the stage of estimation. Such a solution is not inconsistent due to the fact that the sample is almost never exactly balanced, thus using the complex estimators in the sample approximately balanced can bring additional benefits (Tille 2011, p. 223).

Due to the fact that in the majority of elections, three first parties usually get the vast majority of votes and estimating their results is of primary importance, the number of estimated parameters was limited to the results of three committees with the highest results. Besides, estimating separately very low fractions would artificially lower the mean absolute estimation error.

The sample size in each analysed scheme was set at the level of n=100 precincts, which (taking account of all voters in the sampled precinct) corresponds to the 50-70 thousand of elementary units. Every strategy was simulated M=1000 times. The effectiveness of a strategy was measured in two ways: separately for each of three committees and altogether. In the first case the Empirical Root Mean Squared Error (ERMSE), was used:

$$\mathrm{ERMSE}_{\mathrm{J}} = \sqrt{\frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}}\left(\widehat{\mathrm{P}}_{\mathrm{J,i}} - \mathrm{P}_{\mathrm{J}}\right)^{2}} \cdot 100 \tag{19}$$

where:

$\widehat{\mathrm{P}}_{\mathrm{J,i}}$ – the estimates of fraction of votes cast on *J* committee in i^{th} iteration.

In the second case, the estimates for the three main committees altogether were taken into account and for every iteration the Average Manhattan Distance (AMD) was calculated and subsequently the Mean AMD was computed (MAMD):

$$\mathrm{MAMD} = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}} \mathrm{AMD}_{\mathrm{i}} \cdot 100 \tag{20}$$

where:

$$\mathrm{AMD}_{\mathrm{i}} = \frac{1}{3}\sum_{\mathrm{J}=1}^{3}\left|\widehat{\mathrm{P}}_{\mathrm{J,i}} - \mathrm{P}_{\mathrm{J}}\right| \tag{21}$$

Both measures were multiplied by 100, thus the obtained values can be interpreted in categories of percentage points. The simulation analysis was conducted in the R environment.

7. Simulation results

In Table 2 the ERMSE for all strategies for the three subsequent committees with the highest final result are presented. The estimators marked with asterisk (*) were corrected according to the formula (18). It turned out that the best strategy in the case of all three committees was the strategy {BALP10, Q_P1t0}, in which the stratified, balanced against the chosen official statistics at the level of

municipality and powiat, and against the results of presidential election 2010 sample is drawn. This sample also uses the ratio estimator in which the auxiliary variable is the result of the candidate associated with a given party, also during the 2010 election. Distribution of the effectiveness of other strategies is similar in the case of PO and PiS, whereas it differs slightly in the case of Ruch Palikota. Nevertheless, the best sampling design in all cases, irrespective of the method of estimation, turned out to be BALP10.

Table 2. ERMSE for fraction of votes cast on three winning parties

Platforma Obywatelska (Civil Platform)		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	1.458	1.112	0.980	x	x
	HT*	1.468	1.130	1.000	0.601	0.637
	Q-S11	2.672	1.904	1.071	x	x
	Q-P10	0.589	0.624	0.560	0.557	x
	Q-S07	0.615	0.593	0.576	x	0.586
	Poiss-S11*	1.201	1.054	1.024	x	x
	Poiss-P10*	1.030	0.898	0.759	0.679	x
	Poiss-S07*	1.128	0.884	0.807	x	0.757
Prawo i Sprawiedliwość (Law and Justice)		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	1.152	1.020	0.963	x	x
	HT*	1.151	1.020	0.966	0.454	0.518
	Q-S11	1.604	1.721	1.000	x	x
	Q-P10	0.444	0.460	0.452	0.416	x
	Q-S07	0.447	0.504	0.478	x	0.468
	Poiss-S11*	0.921	0.943	0.947	x	x
	Poiss-P10*	0.676	0.686	0.629	0.560	x
	Poiss-S07*	0.722	0.740	0.627	x	0.590
Ruch Palikota (Palikot's Movement)		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	0.359	0.332	0.322	x	x
	HT*	0.355	0.323	0.314	0.267	0.323
	Q-S11	0.625	0.499	0.319	x	x
	Q-P10	0.367	0.348	0.354	0.265	x
	Q-S07	0.453	0.405	0.414	x	0.323
	Poiss-S11*	0.335	0.317	0.312	x	x
	Poiss-P10*	0.296	0.288	0.287	0.268	x
	Poiss-S07*	0.385	0.355	0.352	x	0.331

Source: Own calculation.

In the case of two main parties, the ratio estimators have proved greater efficiency as they correct the direct estimation only against the past election

results of the same party or the candidate associated with the party, which presumably results from the fact that the electorates of those parties remain almost unchanged. In the case of Ruch Palikota, which was the new party on the political scene, supporting the estimation only with the past PO or Bronisław Komorowski results did not work out; generally, the simple estimators or estimators based on log-linear model (the exception is the best strategy, which like in the case of two other parties was {BALP10, Q-P10}) would be a better choice. The mean square error is an absolute measure, thus the differences in values between three committees result mainly from the differences between the values of estimated parameters.

The simulation results with respect to the second criteria of evaluation of strategies are presented in Table 3. The table includes the mean of average absolute differences (for three first parties) between the actual result and estimations in each iteration. In the case of this criteria, the strategy {BALP10, Q-P10} again turned out to be the most effective. Taking into consideration only the sampling design, irrespective of the estimator, the best solution turned out to be BALP10 – the design using the information from the presidential election 2010. BALS07, i.e. the design using the information from the previous election 2007, turned out to be slightly worse. BALS11, which balanced the sample only on data referring to municipalities and powiats, showed similar effectiveness to BALS07 in case of complex estimators, however, in case of simple estimators the effectiveness was worse. The plan using the relatively little additional information, i.e. the stratified sampling performed poorly in terms of drawing the most representative sample. The simple random sampling turned out to be the least effective.

Table 3. MAMD for the three committees with the highest results

		Sampling design				
		SRS	STRAT	BALS11	BALP10	BALS07
Estimator	HT	0.794	0.653	0.600	x	x
	HT*	0.798	0.656	0.603	0.348	0.393
	Q-S11	1.301	1.088	0.632	x	x
	Q-P10	0.369	0.381	0.362	0.326	x
	Q-S07	0.401	0.400	0.389	x	0.365
	Poiss-S11*	0.652	0.612	0.604	x	x
	Poiss-P10*	0.516	0.498	0.445	0.401	x
	Poiss-S07*	0.582	0.522	0.478	x	0.445

Source: Own calculation.

The most favourable assessment of the effectiveness of the methods of estimations for the ratio estimators is, as in the first criteria, using the presidential election 2010 results (Q-P10) as the auxiliary variables. The same estimators using the parliamentary election to the Sejm 2007 results (Q-S07) as the auxiliary variables performed slightly worse. Subsequently, the estimators based on the

log-linear model, which irrespective of the higher number of auxiliary variables used do not surpass the ratio estimators in terms of effectiveness, are ranked. This results from the strong correlation between the estimated parameter and the same parameter in the previous election and not so strong link to the other auxiliary variables, and also from the relatively small sample size which leads to the less stable estimations of the model with many variables. The least effective estimator turned out to be Q-S11 estimator, in which the sum of votes in total was estimated by the ratio estimator and the sum of votes cast on *J* committee was estimated by the HT estimator.

The correction in Horwitz-Thompson estimator (HT*), calculated with respect to exclusion of irregular precincts from the frame population in plans SRS, STRAT and BALS11, led to minimal change in the estimate. The result of this correction for estimating the particular parties results (Table 2) is not unequivocal, however, as far as MAMD is concerned it is negative for every sampling plan. This confirms the above-mentioned assumptions that the electorate in irregular precincts can differ from the voters across country, however, the directions of those differences do change in time, thus they do not qualify for the correction of estimates from regular precincts. Consequently, the restriction of frame population to regular precincts should not systematically bias the results of research.

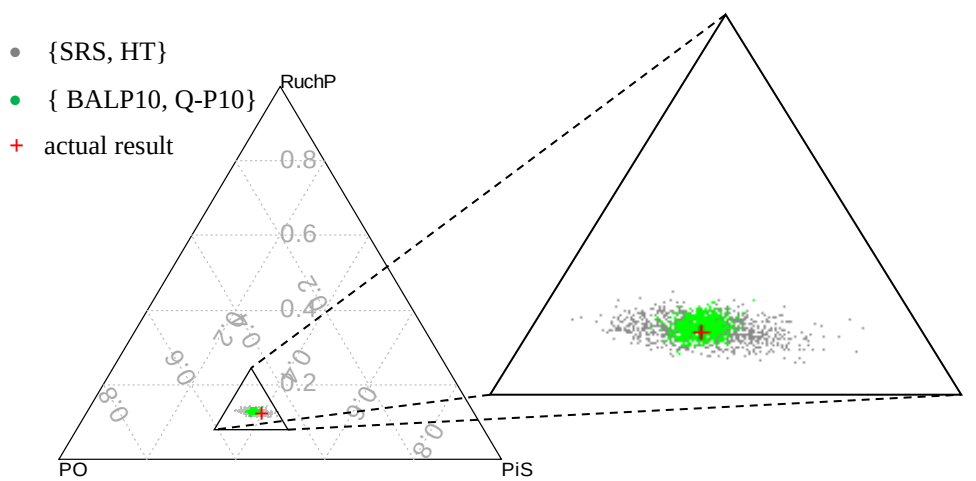

Figure 2. Ternary plot of the simulation results for strategies {SRS, HT} and {BALP10, QP-10}

Source: Own calculation.

To illustrate the difference between the results of the best strategy using auxiliary variables and the results of the classical strategy, i.e. without any auxiliary variables, the ternary plot was created, which is presented in the

Figure 2. The ternary plot is a type of scatter diagram for three variables adding up to a constant. In order to be able to present the estimates for three main parties in this way, the estimates were transformed. Thanks to the transformation, the sum of the estimates equalled 1, that is, as if only these three parties took part in the election. Each point represents the result of one out of M simulations. The location of the point indicates the distribution of votes over three parties – the closer to the vertex of a triangle, the larger part of votes is distributed to the committee described on the particular vertex. Smaller scatter of points for strategy {BALP10, Q-P10} compared to the strategy {SRS, HT} is the reflection of higher effectiveness of the first one.

7. Conclusions

The subject of this paper was the evaluation of the usefulness of available additional data to strengthen the process of estimating the distribution of votes cast during the election in exit poll survey. The additional data were taken from two sources: the Central Statistical Office and the National Electoral Commission. A priori information was included in the strategy of survey both at the stage of selecting a sample and at the stage of estimating parameters. The proposed strategies were tested on the detailed results of the parliamentary election to the Sejm 2011. The results of the conducted simulation indicate that drawing a sample balanced against the selected auxiliary variables as well as the use of those variables in the estimation process significantly improves the effectiveness of the survey. This conclusion was not obvious in the beginning, as the auxiliary features did not refer to the units of research directly; data from GUS refer to the higher aggregation level and data from PKW are not linked to the current research and somehow force the restriction of frame population. Out of two past elections tested as a reference point for the correction of current estimates, the chronologically nearest presidential election 2010 turned out to be best.

Acknowledgement

The research was supported by the grant number 538-2320-0842-12 from the Faculty of Management of University of Gdańsk. I am grateful to professor Mirosław Szreder for his valuable comments.

REFERENCES

BARRETO, M. A., GUERRA, F., MARKS, M., NUÑO, S. A., WOODS, N. D. (2006). Controversies in exit poll, Political Science and Politics, Vol. 39, No. 3, pp. 477–483.

BAUTISTA, R., CALLEGARO, M., VERA, J. A., ABUNDIS, F., (2006). Nonresponse in Exit Poll Methodology: A Case Study in Mexico, Paper presented at the annual meeting of the American Association For Public Opinion Association, Fontainebleau Resort, Miami Beach, FL. Available at: <http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000612.pdf> [Accessed on 12 January 2014].

BISHOP, G. F., FISHER, B. S., (1995). 'Secret ballots' and self-reports in an exit-poll experiment, Public Opinion Quarterly, Vol. 59, No. 4, pp. 568–588.

BUSCH, R. J., LIESKE, J. A. (1985). Does time of voting affect exit poll results?, Public Opinion Quarterly, Vol. 49, No. 1, pp. 94–104.

DEVILLE, J.-C., TILLÉ, Y. (2004). Efficient Balanced Sampling: The Cube Method, Biometrika, Vol. 91, No. 4, pp. 893–912.

DOMAŃSKI, H., MARKOWSKI, R., SAWIŃSKI, Z., SZTABIŃSKI, P. B., (2010). Assessment of methodology and the results of research conducted before the first and second round of the presidential elections in 2010, Warsaw: OFBOR.

Election Code, Act of 5 January 2011, Journal of Laws 2011 No. 21, item 112, Article 12.

HILMER, R., (2008). Exit polls in Germany, Berlin: 3MC Conference Proceedings.

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?, Public Opinion Quarterly, Vol. 40, No. 2, pp. 182–193.

KOZŁOWSKI, A., (2012). The usefulness of past data in sampling design for exit poll surveys, Economic Studies, University of Economics in Katowice, Vol. 120, pp. 45–57.

LANGEL, M., TILLÉ, Y., (2011). Corrado Gini, a pioneer in balanced sampling and inequality theory, METRON – International Journal of Statistics, Vol. LXIX, No. 1, pp. 45–65.

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

LEVY, M. R., (1983). The methodology and performance of election day polls, Public Opinion Quarterly, Vol. 47, No. 1, pp. 54–67.

MERKLE, D. M., EDELMAN, M., (2002). Nonresponse in exit polls: a comprehensive analysis, in: Groves, R. M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (Eds.), Survey Nonresponse, New York: John Wiley & Sons.

MOON, N., (2008). Predicting the Election Result from an Exit Poll – the UK example, Berlin: 3MC Conference Proceedings.

MOORE, D. W., (2003). New Exit Poll Consortium Vindication for Exit Poll Inventor, Inside the polls, Gallup. Available at: <http://www.gallup.com/poll/9472/new-exit-poll-consortium-vindication-exit-poll-inventor.aspx> [Accessed on 24 January 2013].

Ośrodek Badania Opinii Publicznej, archival site, <www.obop.com.pl> [Accessed on 24 January 2013].

Państwowa Komisja Wyborcza, <pkw.gov.pl> [Accessed on 24 January 2013]

SCHEUREN, F., ALVEY, W., (2008). Elections and Exit Polling, New Jersey: John Wiley & Sons.

SZREDER, M., (2010). Methods and techniques of opinion polls surveys, Warsaw: PWE.

SZREDER, M., (2011). Emotions and the truth of election night, Rzeczpospolita, 28.09.2011, No. 227.

The Election code, Act of 5 January 2011, Journal of Laws 2011, No. 21, item 112, Article 12.

TILLÉ, Y., (2006). Sampling Algorithms, New York: Springer.

TILLÉ, Y., (2011). Ten years of balanced sampling with the cube method: An appraisal, Survey Methodology, Vol. 37, No. 2, pp. 215–226.