Research Articles

Reconstruct Partisan Support Distribution with Multiply Imputed Survey Data: A Case Study of Taiwan's 2008 Presidential Election*

Frank C. S. Liu**

ABSTRACT

Analyzing survey data is one of the most promising methods by which to predict election results. But respondents may conceal their preferences. Hence, it has been difficult for researchers to obtain true partisan support distributions of with one survey data set. Given the constraints of cost, could we possibly predict vote shares more accurately with one sample? This paper employs

* An earlier version of this paper was presented at the 9th Conference on Survey Methods and Application, September 11, 2009, Academia Sinica, Taipei, Taiwan. The author is grateful for comments from Hong Yong-Tai (洪永泰) and Shia Ben-Chang (謝邦昌).

** Associate Professor, Institute of Political Science, National Sun Yat-Sen University; e-mail: csliu@mail.nsysu.edu.tw

Note: Received: September 15, 2009; Accepted: September 17, 2010.

multiple imputation (MI) for point estimation as a way to (re)construct the distribution of partisan supporters in Taiwan's 2008 presidential election. The findings show an identifiable difference between the biased point estimation and a better one of using MI. Although there remain other types of errors that may influence the accuracy of a prediction, readers may find this method relatively cost efficient when formulating strategies to improve point estimation pertaining to election results.

Key Word: multiple imputation, partisanship, survey research, missing values, election prediction

以多重插補法重建政黨支持比率的圖像：以 2008 總統大選前夕面訪案資料爲例*

劉正山**

摘要

透過調查樣本來推估得票率是選舉預測方式之中非常使用的方法。然而，即使抽樣過程恰當且具有母體代表性，受訪者在面對投票給誰這類問題時所產生的拒答現象卻往往造成資料的遺失，進而造成點估計的偏誤。這個情形在選舉期間的投票意願調查或敏感問題的調查格外嚴重。這個嚴重遺漏值的問題不但導致民衆對於使用調查資料的預測能力失去信心，也造成學者對於這些資料產生出來的描述統計數據感到懷疑。本研究嘗試以多重插補法進行遺失資料（missing data）補足的工作，並將此法應用到政黨得票率的點估計上。本研究使用台灣 2008 年總統選舉前蒐集的「台灣選舉與民主化調查」面訪資料（Taiwan's Election and Democratization Study for the 2008

---

* 本文初稿發表於 2009 年 9 月 11 日中央研究院主辦「第九屆調查研究方法與應用」學術研討會，承蒙國立台灣大學政治學系洪永泰教授、輔仁大學統計資訊學系暨應用統計所謝邦昌教授等多位學術先進對於本文提出的寶貴評論與意見，作者受益良多，特此致謝。

** 國立中山大學政治學研究所副教授，E-mail: csliu@mail.nsysu.edu.tw

legislative elections, TEDS2008L, N=1,240)，比較多重捕補前後與兩黨候選人支持率的差異。研究發現，使用多重捕補法將有助於修正因高度遺漏值所造成的候選人支持率的點估計偏誤。

關鍵字：多重捕補法、政黨傾向、調查研究、點估計、選舉預測

I. Introduction

A commonly recognized problem among public opinion researchers and pollsters is that respondents, when asked about sensitive issues, often hesitate to disclose their vote choices and attitudes. This item-non-response problem caused by respondents' reluctance (by saying "Don't Know", "It Depends", or simply refusing to answer some of the questions) has long been recognized as a cause of biased point estimation within a survey sample. One safe method for reconstructing partisan support distribution is to summarize a decent number of polls and survey data, and then calculate the average of the polls. The problem with this method, however, is that poll companies and survey institutes may not release their data to the public, and/or they may not release the information prior to the election. Therefore, it has been difficult for individual researchers to apply this method to upgrade their understanding about the preferences of the electorate.

Over these years, scholars of vote prediction have proposed a few approaches to deal with this item-non-response problem and hope to enhance the accuracy of description, or even prediction, with single data set (the four methods will be discussed in Section I). This

paper joints the line of effort by exploring the usefulness of multiple imputation (MI) in advancing the accuracy of preferences descriptions of the electorate.[1] The rationale of this study is straightforward: as MI has been a sophisticated method for regression analysis, it is reasonable that we use it to advance our understanding about the preference distribution of the electorate.

Note that this method is only applicable to the item-non-response situation. A number of methods that deal unit non-response (i.e., respondents not available for the survey or not being able to answering a series of questions), such as raking, are beyond the scope of this paper. Furthermore, even though this paper will show that this method has potential to empower researchers to use single data set to reconstruct partisan support distribution, the results of this preliminary attempt should not be over-expected. The method proposed here is more about advancing our understanding about the preference distribution of the population (the electorate) than pursuing a perfect match between estimated vote shares and election results, given the fact that a good proportion of the electorate may give up their votes. This paper is aimed to show that applying MI in describing the preference distribution of the population is better than not using MI.

Sections I and II will give a brief overview about current

1. There are certain other methods by which to deal with missing data, such as giving weights to compensate for those who were excluded due to missing values, and constructing the likelihood based on incompletely observed data. These methods are not covered here, for they are not directly dealing with increasing estimation accuracy. For an overview, see Raghunathan (2004) and Särndal and Lundström (2005).

methods dealing with item non-response and MI techniques, respectively. Section III will consist of a description of the data set for inspection, including its pattern of missingness and the variables selected for imputing individuals' vote choices, which are the key variables used to calculate the distribution of supporters of the two political parties in Taiwan's 2008 presidential election. This section will then introduce a software package, *Amelia II*, designed to handle these tasks. Section IV will present the results of ten experiments and suggest strategies for choosing a reasonable and manageable number of variables to achieve results that are as good as those derived from methods using all variables. This paper will conclude with a discussion (Section V) about the limitations of this method and provides a path for future research.

II. Methods of Dealing with Item-Non-Response

Over these years, scholars have proposed a few approaches to deal with this item-non-response problem. The first is replacing forced-choice questions with subjective probability scales. For example, instead of asking, "Whom would you vote for if the election was held today?"—a question that usually results in a high rate of "undecided", researchers could reduce this refusal rate by asking, "On a scale of 1 to 10, how likely are you to vote for each candidate on Election Day?" As this method is found to be effective primarily for elections involving more than two candidates, the accuracy of predictions formulated using such adjusted question would not be better than that of using the original question in elections with only two

candidates (K. J. Flannelly, L. T. Flannelly, & Malcolm S. Mcleod, Jr., 1998; Laura T. Flannelly, Keven J. Flannelly, & Malcolm S. Mcleod, Jr., 2000).

The second method is to change contextual settings in order to decrease respondents' anxiety levels while they are revealing their opinions regarding sensitive issues and to boost their willingness to express their true preferences. This might include using a self-administrated or "secret-ballot" questionnaire (Bishop & Fisher, 1995). There is also some inconclusive discussion related to ways such as online chat rooms that might be able to serve this purpose (for example, Ho & McLeod, 2008; McDevitt, Kiousis, & Wahl-Jorgensen, 2003). However, adoption of such techniques will inevitably increase the cost of surveys.

Thirdly, it is suggested that researchers use alternative dimensions in order to probe for hidden partisans (Coakley, 2008). Coakley, in his study of North Ireland, chose "nationalist" as an alternative dimension to probe for hidden Sinn Fe'in supporters. He found that many (approximately 30%) of the individuals whose attitudinal configurations suggested that they were likely Sinn Fe'in supporters claimed to support either another party or no party at all. The challenge and difficulty of adopting this method, nevertheless, will be that researchers need to theoretically justify and repetitively test the validity of using such alternative survey questions. Whereas there is no single convenient poll question by which to uncover hidden partisans, there remains room to further explore various theory-based alternative questions to advance this approach.

The fourth approach, and also the one that is adopted in this

paper, suggests that researchers deal with the problem of under-representation of extreme-right voters by multiply imputing missing values based on information drawn from voter profiles or characteristics (Durand, Blais, & Larochelle, 2004). While concerns exist regarding biased estimates, current development of methods is making this technique a cost-efficient and easy one to use.

III. Multiple Imputation for Electoral Studies with Missing Values

Multiple imputation (MI)—inserting new values into data cells with missing values based on information derived from other related variables—has been identified and used as one of the techniques by which to solve the problem of missing values caused by non-response in surveys.[2] Over the past few decades, a growing number of researchers have followed Rubin's (1987) recommendation to replace missing or deficient values with a number of alternative values representing a distribution of possibilities (for instance, Paul, Mason,

2. MI is a method commonly used to deal with missing data problem, including item-nonresponse (nonresponse to some, but not all, survey questions) and unit-nonresponse (nonresponse to all survey questions). A common and still useful alternative is list-wise deletion of observations due to both item-nonresponse and unit-nonresponse in the regression analysis. However, because a significant number of observations are excluded from analysis, this method may yield biased parameter estimates. While the default procedure of most statistical packages excludes the observations with missing values, list-wise deletion has been identified as a problem for most electoral studies (Gelman, King, & Liu,1998). This concern regarding biased estimates can be minimized if the loss of cases due to missing data is less than about 5%, and if pretest variables can reasonably be included in the models as covariates (see Graham, 2009).

McCaffrey, & Fox, 2008).

In contrast to single imputation or stochastic imputation, which refers to conducting one stochastic imputation on the basis of information derived from other variables, MI is a procedure whereby several data sets are created based on the original, and then the same analysis is performed separately on each complete data set. MI is preferable to single imputation because single imputation generally leads to underestimation of standard errors and overestimation of test statistics, as MI reflects sampling variability and other uncertain factors inherent in models (Rubin, 1987, p. 11-18; Weisberg, 2005, p. 143-150). While some scholars may think this technique is unrealistic, or have concerns about "making up" data, Stuart, Azur, Frangakis, and Leaf (2009) inversely argue that "complete-case analyses require stronger assumptions than does imputation" (p. 1134).

To be more specific, the practice of MI is comprised of three steps: (a) the imputation stage (creating imputed data sets), (b) the analysis stage (calculating parameter coefficients on the basis of each created data set), and (c) the combination stage (calculating final coefficients and standard errors from the numbers obtained in the analysis stage). After selecting proper variables for imputing the variable with missing values—i.e., those variables that are correlated with the targeted variable(s) of interest, the MI algorithm will take the values of the chosen variables and generate for each missing value more than one new value. The procedure will then create several new data sets (usually five or more), in which all missing values are filled. Next, a researcher runs his or her models based on each imputed data set. Supposing that there are five imputed data sets, he

or she will consequently have five sets of coefficients. In the third stage, called "combination", the researcher obtains the value of the interested coefficient by averaging the five coefficients of the same variable acquired in the five imputed data sets. The standard error of the resulting coefficient is also determined based on the variance inside each imputed data set and between imputed data sets (see King, Honaker, Joseph, & Scheve, 2001; Stuart et al., 2009).

Most research using MI and discussions of this method focuses on enhancing the accuracy of estimated coefficients in regression models. For example, Penn (2007) employed MI to estimate missing income data and update a recent study examining the influence of parents' standards of living on subjective well-being. He uses data from the 1998 General Social Survey and compares results of two ordered probit models: one using complete cases only, and the other replacing missing income data with multiple imputation estimates. Consistent with earlier studies using MI, he confirms that MI allows researchers to make use of more of the available data and decreases possible biases.

Although enhancing regression analysis is the original purpose of MI, there has been little research about using multiply imputed data sets for descriptive purposes. As yet, the closest exception is a study that compares descriptive statistics derived from multiply imputed data sets to determine if two or more methods generate similar results (Bernaards et al., 2003). Still, scholars have not taken a single further step to apply this method to the field of vote prediction. If the imputed data sets that are used to run regressions yield results that are better or more informative than analyses based on

observed data, the descriptive statistics of the variables of interest in these multiply imputed data sets are certainly worth exploring.

Given the above review, I suspect that multiply imputed data sets could be equally useful in terms of inspecting the distribution of the variables of interest, not only to coefficient estimation.

IV. Data Set: TEDS2008L

1. Data Description

Taiwan's Election and Democratization Study for the 2008 legislative elections (TEDS2008L, $N=1,240$) has been chosen to examine the applicability of the MI method.[3] The TEDS2008L data was collected during mid-January and early March of 2008. Compared to other TEDS and other large-scale surveys in Taiwan that were carried out during winter or summer vacations (the periods for the executive board to recruit and train student interviewers), TEDS2008L is the only data set collected prior to a presidential election (March 22, 2008). In this investigation, the respondents were asked about voting choices in the upcoming presidential election. Thus, information about partisan support distribution is available in the original data

3. Data analyzed in this paper were from Taiwan's Election and Democratization Studies, 2008: Legislative Election (TEDS 2008L) (NSC96-2420-H-002-025). The coordinator of multi-year project TEDS is Professor Yun-Han Chu (National Taiwan University). TEDS2008L is a yearly project on the legislative election in 2008. The principal investigator is Professor Chi Huang. More information is on TEDS website (http://www.tedsnet.org). The author(s) appreciate the assistance in providing data by the institute and individual(s) aforementioned. The author(s) are alone responsible for views expressed herein.

and can be used for comparison against partisan support distributions based on multiply imputed data sets.

2. Software Package for the MI Analysis

Amelia II is a cross-operation system package designed to process EMis (Expectation Maximization with importance re-sampling), one of the suggested algorithms using Markov Chain Monte Carlo (MCMC) methods to calculate imputed values (Honaker, King, & Blackwell, 2009; Horton, & Ken P. Kleinman, 2007; Imai, Gary, & Olivia, 2009).[4] *Amelia II* is a free, cross-platform toolkit. It is compatible with R, a widely used open-source package for statistical

4. Expectation Maximization (EM) and Imputation Posterior (IP) are two primary algorithms by which to generate imputed data sets. While the IP algorithm is based on the Markov Chain Monte Carlo (MCMC) method, EM, which yields only the maximum values, is a faster and less complex alternative to IP. When using IP, a researcher needs to frequently draw an estimated mean and variance from the disputed data sets created from entire multivariate models of observed data posterior. In order to obtain an exact result as expected, a researcher is hence required to spend a substantial amount of time drawing infinitely before convergence occurs. As King et al. (2001) discuss, the trouble with EM is that it ignores estimation uncertainty and treats the estimated imputed parameters as though they were obtained from complete data sets without missing data, and, therefore, could result in biased coefficients and standard errors. Hence, EMs (EM with sampling) and EMis (EM with importance resampling) are proposed to solve the uncertainty problem in EM. As EMs is used for studies of large-N data sets and those composed of continuous variables, EMis is especially designed for samples with multiple parameters and categorical variables. King et al. recommend EMis over other algorithms because EMis possesses both the precision of IP and the speed of EM. Additionally, EMis has the capability to deal with data sets with many variables and takes the concern about the uncertainty of imputed data into account. For the discussion about the strength and limitations of alternative algorithms, such as multiple imputation by chained equations (MICE), see Stuart et al. (2009) and He and Raghunathan (2009).

programming, and has a handy graphical user interface (GUI) that allows users to intuitively set the characteristics of the variables with ease. For example, specifying whether the variable is ordinal or nominal by pointing and clicking.

A conventional difficulty in dealing with MI involves transforming categorical variables that have more than three levels into dummy variables and then analyzing them by performing a separate EM analysis with the two-level version of the variable (for example, 0 versus other) (Graham, 2009, p. 563). *Amelia II* simplifies this process considerably; the researcher simply selects the key variables and designates them as nominal variables, and *Amelia II* transforms them into dummy variables and regards them as categorical variables during the imputation process.[5] In this study, I designate two key variables as nominal, namely vote choice (Kuomingtan, KMT or Democratic Progressive Party, DPP) and party identification (the two major parties and the four small parties); the former is the target variable for imputing, while the latter is the primary variable that provides information for imputing the missing values of vote choice.[6]

---

5. For researchers following the three-step procedure of conducting MI, Zelig, another package compatible with R, is suggested for the combination stage (see Imai, King, & Lau 2009). Since hypothesis testing is not the goal of the present study, the analysis below will concentrate on using *Amelia II* for the first two stages of MI.
6. In TEDS2008L, respondents were asked about their preferred candidate for president in the upcoming presidential election and partisan orientation. The question regarding vote choice is, "Who did you vote for?" This question is preceded by the voter turnout question, "In this presidential election, many people went to vote, while others, for various reasons, did not go to vote. Did you vote?" The wording of the party identification question is as follows, "Among the main political parties in our country, includ-

3. Variables for Imputing Vote Choice

Variables that will be used in the "analysis stage", such as logistic regression, are the best choice for the MI procedure. In other words, the choice of auxiliary variables for imputing vote choice should be those theoretically associated with it, such as the dependent variables, independent variables, and control variables (Abayomi, Gelman, & Levy, 2008; King et al., 2001; Wood, White, & Royston, 2008). Self-claimed party identification is the first to be included in the list of variables for MI, because it has been well acknowledged by political scientists since Campbell, Converse, Miller, and Stokes (1960) as the most important variable for explaining vote choice.

Table 1 demonstrates a high level of missingness for these two key variables, vote choice and party identification. The proportions of missingness for these two variables are 43.0 and 37.7, respectively. If partisan support distribution is calculated by ignoring those who did not respond, the support for KMT's Ma Ying-jeou and Vincent Siew amounts to 70.3% (497/707) and that of DPP's Frank Hsieh and Su Tseng-chang comes to 29.7% (210/707). Although this "poll" predicts the victory of KMT's Ma and Siew, which is consistent with the result of the election, the point estimate is far from satisfactory and nowhere near KMT's real vote share of 58.45% and DPP's 41.55%. The partisan support distribution of KMT is overestimated, while

---

ing the KMT, DPP, NP, PFP, and TSU, do you regard yourself as leaning toward any particular party?"

Table 1 *Descriptive Statistics for TEDS2008L*

Targeted Variables	Response Items	N(%)
Vote Choice for President	KMT (Ma Ying-jeou & Vincent Siew)	497(40.1)
President	DPP (Frank Hsieh & Su Tseng-chang)	210(16.9)
	Missing (refuse to answer, don't know & skip)	533(**43.0**)
Party Identification	KMT	446(57.8)
	DPP	290(29.0)
	NP	16(2.1)
	PFP	1(0.1)
	TSU	11(1.4)
	Others	8(1.0)
	Missing (refuse to answer, depends, don't know)	468(**37.7**)

Note. Source: TEDS2008L.
N=1,240, 並請說明 KMT、DPP、NP、PFP、TSU 代表的意思

that of DPP is underestimated, implying that TEDS2008L fails to accurately describe of the partisan support distribution of each party.

Table 2 lists variables drawn from TEDS2008L for MI. These variables were put together into a new data set and loaded into *Amelia II*.[7] These variables were selected based on their theoretical association with vote choice, including party identification men-

7. Because the variables chosen here are ordinal and nominal, I conducted the chi-square test of independence. If a chosen variable is numeric, it is suggested to show the correlation between the cause(s) of the missingness or auxiliary variables, Z, and the model variable containing missingness, Y. The auxiliary variables that yield correlation rZY=0.90, or at least .50, will have a major impact on reducing the biasing effects of attrition (Graham, 2009). "One or two auxiliary variables with rZY=0.60 are better than 20 auxiliary variables whose correlations with Y are all less than rZY=0.40" (570).

Table 2 *A List of All Variables for Imputing Vote Choice (TEDS2008L)*

Variable	Description	Code	Missing	Note
vote08	Vote Choice for President in 2008	S05	533	dichotomous
partyID	Party ID	M01B	468	multinomial; $\chi^2=930.7$, $df=5$
likeKMT	The degree of liking KMT	M02A	133	a 11-point scale; $\chi^2=306.9$, $df=10$
likeDPP	The degree of liking DPP	M02B	129	a 11-point scale; $\chi^2=336.3$, $df=10$
MaScale	Evaluation of Ma Ying-Jio	G05	133	a 11-point scale; $\chi^2=389.5$, $df=10$
HsiaoScale	Evaluation of Vincent Siew	G06	177	a 11-point scale; $\chi^2=170.8$, $df=10$
HsiehScale	Evaluation of Frank Hsieh	G07	154	a 11-point scale; $\chi^2=348.4$, $df=10$
SuScale	Evaluation of Su Tseng-Chang	G03	158	a 11-point scale; $\chi^2=246.2$, $df=10$
tvNews	Most frequently watched TV news Channel (in general)	A02A	152	multinomial; $\chi^2=242.1$, $df=11$
talkShows	Most frequently watched TV news Channel (for political talk shows in particular)	B02A	123	multinomial; $\chi^2=161.9$, $df=5$
prosEcon	Prospective Economy Condition	H06	271	3-option mutinomial; $\chi^2=6.7$, $df=2$, $p=0.03$
retroEcon	Retrospective Economy Condition	H05	35	3-option multinomial; $\chi^2=22.4$, $df=2$
ChenScale	Evaluation of Chen Shui-Bian	G02	130	a 11-point scale; $\chi^2=350.2$, $df=10$
incScore	Satisfaction with Chen's administration	V01	119	a 11-point scale; $\chi^2=299.4$, $df=10$
demScore	Satisfaction with Taiwan's Democracy	U08	155	4-point ordinal; $\chi^2=21.0$, $df=3$
eth	Father's ethnicity	X02	20	multinomial; $\chi^2=43.6$, $df=4$
edu	Education level	X06	20	a 13-level scale; $\chi^2=30.0$, $df=12$, $p=.003$

Note. Source: TEDS2008L.N=1,240; chi-square test of independence against "vote08"; p values lower than .001 are not reported. In tvNews top ten most watched TV news channels are coded to 1 to 10, respectively; other channels are coded into 11; 0 is used for those saying never watch TV news.The coding of talkShows is based on a rule that re-categorizes mentioned talk show programs into corresponding news channels: 0 for never watch such programs; 1=TVBS; 2=SET TV; 3=Formosa TV; 4=Cti TV; 5=others (less than 1%).

tioned above ($\chi^2=469.8$, $df=5$, $p<.001$) and variables of evaluations of incumbents, political parties, and candidates (e.g., Erikson, Mackuen, & Stimson, 2002); habits of watching political news (Lazarsfeld, Berelson, & Gaudet, 1968); evaluation of the economy (Alvarez, Nagler, & Bowler, 2000) and democracy (Sullivan & Transue, 1999); and demographics, such as ethnicity (Kam, 2007) and education level (Zuckerman, Kotler-Berkowitz, & Swaine, 1998). These variables are chosen based on the causal relationships (at least correlations) identified in the literature. The conventional wisdom suggesting that vote choices are influenced by their ethnicity or race and the level of political knowledge. The above literature, put together, suggest that voters' choice will be influenced by (or correlated with) their liking of particular candidates or candidates, their favorite news sources that is potentially biased against certain issues, parties or candidates, and their satisfaction with current or past economy. Their satisfaction with democracy is also related to their partisanship and vote choices since the competition between parties of the two-party system will turn into the trust and distrust of electoral processes. Indeed, there shall be more variables indicated in the literature. As no data set provides the whole battery of variables for imputing vote choices, I chose the above that is theoretically and logically related to the target variable for MI.

4. Experiment Design

I utilized the partisan support distributions of the two political parties in the 2008 Presidential Election, indicated by their vote shares, as the baseline of comparison, while these figures are to be

contrasted against the partisan support distributions derived from TEDS2008L without MI and those derived from the same data set after MI is applied.

Three strategies of variable selection and nine experiments, each of which corresponds to a specific variable selection strategy, were consequently conducted. The partisan support distribution of KMT and DPP were recorded and compared in Table 2. The purpose of comparing the results is to present an efficient combination of variables that may later be used in a telephone survey, which is usually constrained by the number of questions.

The first strategy, as shown in Figure 1, includes all of the variables listed in Table 2. Note that I specified three variables, specifically partyID, vote08, and eth as nominal variables. These specifications will force *Amelia II* to impute these variables with integers in the sense of categories rather than in continuous or ordinal fashion.

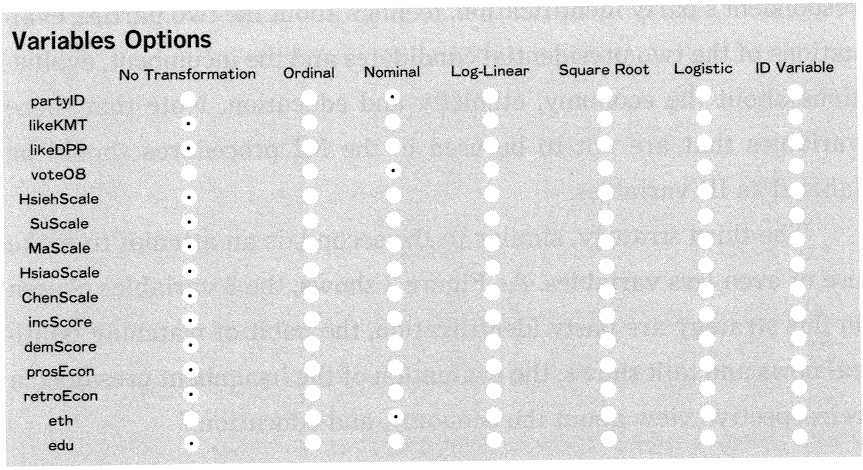

Figure 1 *Strategy 1: All variables are used*

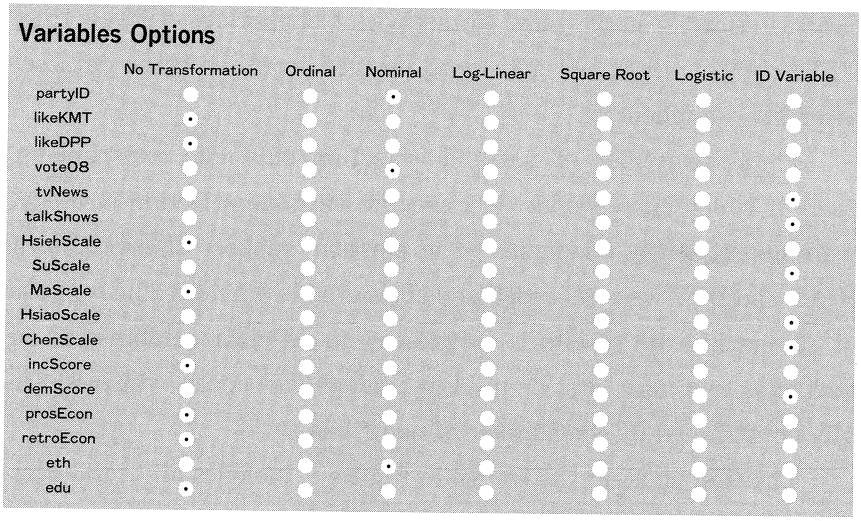

Figure 2 *Strategy 2: Some variables are used*

The second strategy entails the use of certain of the variables in the procedure. As illustrated in Figure 2, the 10 variables chosen are the respondent's party identification, feelings about the two parties, evaluations of the two presidential candidates and the incumbent, evaluations about the economy, ethnicity and education. Note that those variables that are not to be used in the MI procedures should be labeled as ID variables.

The third strategy, similar to the second, is an attempt to make use of even less variables. As Figure 3 shows, the 8 variables chosen in this strategy are party identification, the habit of watching political news and talk shows, the evaluation of the incumbent president, a retrospective view about the economy, and education.[8]

8. Indeed, other strategies for combining variables and determining the number of vari-

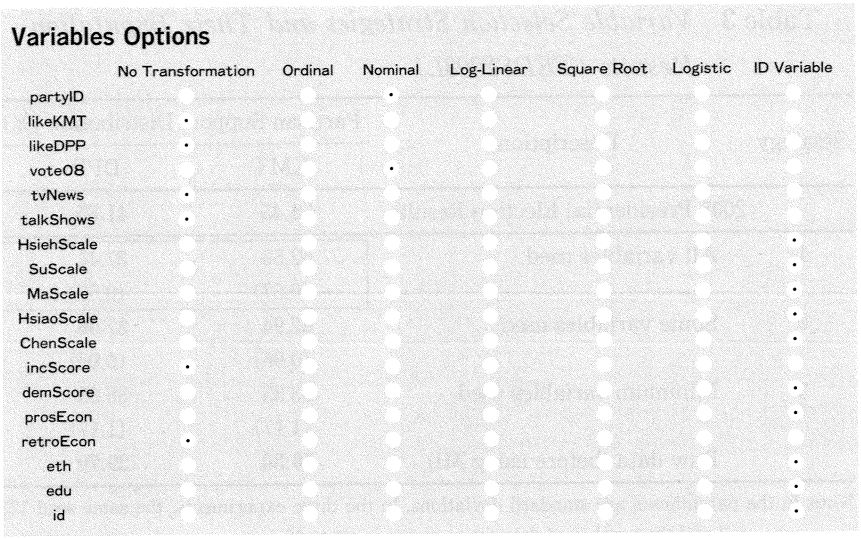

Figure 3 *Strategy 3: Minimum variables are used*

V. Experiment Results

Table 3 lists the MI results of the three strategies of variable selection. The same seed 123 was used for reiteration, and the number of data sets to be created is set to 10. The proportions shown in the table are the average of partisan support distributions derived separately from the 10 imputed data sets. The number 10 is not a fixed number; for this study, I assumed that the means of 10 samples would be more stable than that of other numbers of samples less than 10.

---

able that should be chosen also exist. These two topics exceed the scope of this paper and await future inspection.

Table 3 *Variable Selection Strategies and Their Imputation Results (TEDS2008L)*

Strategy	Description	Partisan Support Distribution (%)	
		KMT	DPP
	2008 Presidential Election Results	**58.45**	**41.55**
1	All variables used	62.59	37.41
		(0.57)	(0.57)
2	Some variables used	62.94	37.06
		(0.96)	(0.96)
3	Minimum variables used	63.35	36.65
		(1.17)	(1.17)
	Raw data (before using MI)	**70.30**	**29.70**

Note: In the parentheses are standard deviations. In the three experiments, the same seed 123 was used and the number of data set to create is set to 10.

The first finding indicates that the more variables used the better and more stable the results are, as evidenced by a comparison of standard deviations. The second finding is that, although the strategies vary, a pattern of partisan support distribution emerges for the two parties: KMT around 62% and DPP around 37%. This result, as a point estimate, is much closer than that before applying MI. The gap shrinks from 11% to about 4% for each party. Note that the closing of the gap does not imply that MI is the best tool for election result prediction. Instead, it suggests that MI provides a better way to represent the uncovered (true) distribution of the support rates of the two pairs of presidential candidates. In other words, the MI method provides an indirect way to predict election results. Readers should be reminded again not to take the figures generated by this method as an empirically valid vote shares.

VI. Conclusion and Discussion

Analyzing survey data is one of the most promising methods by which to predict election results. A commonly perceived problem with this method is that a significant proportion of respondents conceal their preferences; this is a particular problem in Taiwan, when survey questions are sensitive or when the survey is about political preferences during a campaign season. Therefore, it is difficult to employ a data set, even when the observations are well-sampled, to inspect the unknown distribution of their vote choices.

This study argues that multiple imputation (MI) can be a tool for advancing point estimation, particularly when the cost of polling is a concern. The method proposed in this paper helps to increase the predictability of single survey data. The preliminary finding derived from pre-election face-to-face survey data collected during Taiwan's 2008 presidential electoral campaign confirms this perspective, suggesting that the averaged partisan support distributions based on MI data sets provide a better guess for election results than those simply derived from the original data set.

MI is a method commonly used to deal with problems of missing data. The other types of errors that lead to biased point estimation, such as systematic non-response error, measurement error, sampling error, and frame error (also called frame imperfection), are beyond the scope of this paper. These types of errors are worth mentioning here because they account for the inaccuracy of adjusted point estimates and should be dealt with in future studies.

Frame error should be the major cause of the inaccuracy of survey data. This refers to situations in which a sample collected at a specific time fails to encompass certain elements of the target population. Coverage error is the most important form of frame error. It occurs when the elements in the sampling frame do not correspond correctly to the target population in which the researcher wants to make inferences. Coverage error, which occurs most commonly in telephone surveys, specifically refers to "the mathematical difference between a statistic calculated for the target population studied and the same statistic calculated for the target population. It occurs when there is bias due to the omission of non-covered units, such as omitting people who do not have phones in a telephone survey (Weisberg, 2005). Coverage error includes under coverage, over coverage, and duplicate listings (For example, a target population element is listed more than once in the frame). Frame error occurs in almost every survey and poll. It can be solved only when sampling techniques are advanced. "Although imperfect frames are a reality in most large surveys, the remedies employed in statistical agencies and elsewhere vary considerably. In the absence of 'firmly established methodology', the procedures in use may seem *ad hoc*" (Särndal & Lundström, 2005, p. 179; original emphasis).

Furthermore, this method is based on the assumption that every voter has a true partisan preference, if he or she is forced to reveal such a preference honestly. This assumption can find its theoretical root in most American studies of voter behavior, such as the concept of belief system. While this assumption remains valid on a theoretical level, it has not been empirically tested, so readers should be cau-

tious about applying the method to a situation where respondents may not have true preferences.

Besides the above factors that influence the quality of sampling, there are a few other topics that are associated with the MI method worthy of inspection. First, this study compares three strategies of variable choice (as shown in Table 3) by their standard divinations. This selection by standard deviation, however, is insufficient for the selection of best strategy. Future studies investigating on this strategy-selection issue. Second, *Amelia II* is one of a few software packages available for MI. A project that compares results derived from other software packages and/or other algorithms, such as nearest neighbor imputation, will be a welcome addition for scholars using MI for election prediction. Third, the entire method of using MI for election prediction is built upon the assumption that respondents of the imputed data sets turn out to vote, and that their answers to auxiliary variables are safe for predicting their voting choices. In other words, future studies should take a closer look at each of these assumptions and evaluate the extent to which the violation of these assumptions affects the results.

The forth issue is far beyond the scope of this paper, but is an important constraint that applies to all studies using MI, including the present one. That is, scholars tend to assume that the pattern of item-non-response lying behind data is "missing at random" (MAR) or "conditionally missing at random" (King et al., 2001). Specifically, suppose non-response to a party identification question (Var1) is associated with one's party orientation (Var2). The missingness on Var

1 is conditioned on the Var2 and is thus MAR.[9] This is an assumption held in most MI studies, but one that is very hard to verify, because this would require knowledge of the missing values themselves (Little & Rubin, 1987). Particularly, as G. David Garson states, "For purposes of univariate analysis (e.g., understanding the frequency distribution of how subjects respond to an opinion item) imputation can reduce bias and often is used for this purpose if data are missing at random."[10] As MAR is an important assumption embedded in the design of *Amelia II*, the analytical tool used in this paper, and because this problem awaits a solution, readers shall fol-

9. In contrast to the other two assumptions about item-missing data that are regarded as unrealistic, i.e., missing completely at random (MCAR) and not missing at random (NMAR), MAR is most commonly assumed for large-scale data sets. MCAR means that the missingness is unrelated to the variable under study, therefore, the missing data are considered to be "ignorable". NMAR, also called nonignorable (NI), means that the probability of missingness depends on both observed and unobserved values. "MAR, while empirically unverifiable, is often a reasonable assumption to make unless substantive knowledge about the data or data collection process indicates that the missingness may depend on unobserved values. . . . The MAR assumption is also sometimes made more reasonable by including 'auxiliary variables' that are related to the missingness but may not be of interest in the analyses themselves; in fact, this strategy can greatly improve the imputations" (Stuart et al., 2009, p. 1134). MCAR and MAR result in unbiased parameter estimates, while MNAR missingness is considered a problem because it yields biased parameter estimates (Graham, 2009). It is very difficult to identify the pattern and the mechanism of missing data, "since the MAR condition cannot be tested empirically, the analyst must decide on theoretical grounds whether to use imputation techniques appropriate for MAR missing data or to model the missing data as NI" (Weisberg, 2005, p. 151). In other words, MNAR would be classified as MAR if the variables were correlated, and MNAR would be classified as MCAR if the variables were uncorrelated (see Templ & Filzmoser, 2008 for their clear descriptions using visualization).

10. See http://faculty.chass.ncsu.edu/garson/PA765/missing.htm

low the current development of methods to improve the robustness of analysis, such as identifying the patterns of missing data visually (e. g., using R package "Visualization and Imputation of Missing Values", VIM). One method that is worth consideration is applying sensitivity analysis to multiple imputation. Simply put, one can use importance sampling to re-weight the parameter estimates obtained from the imputed data sets, making MI results represent the distribution of imputations under a NMAR mechanism, by which researchers can judge whether the results are likely to be sensitive to the MAR assumption, as well as the likely direction and magnitude of the effect (Carpenter, Kenward, & White, 2007).

REFERENCES

Abayomi, Kobi, Andrew Gelman, & Marc Levy
2008 Diagnostics for Multivariate Imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 57*, 273-291.

Alvarez, R. Michael, Jonthan Nagler, & Shaun Bowler
2000 Issues, Economics, and the Dynamics of Multiparty Elections: The British 1987 General Election. *American Political Science Review, 94*(1), 131-149.

Bernaards, Coen A., Melissa M. Farmer, Karen Qi, Gareth S. Dulai, Patricia A. Ganz, & Katherine L. Kahn
2003 Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey. *Journal of Data Science, 1*(3), 293-312.

Bishop, George F., & Bonnie S. Fisher
1995 'Secret Ballots' and Self-reports in an Exit-Poll Experiment. *Public Opinion Quarterly, 59*(4), 568-588.

Campbell, Angus, Phillip Converse, Warren Miller, & Donald Stokes
1960 *The American Voter.* New York: Wiley.

Carpenter, James R., Michael G. Kenward, & Ian R. White
2007 Sensitivity Analysis after Multiple Imputation under Missing at Random: A Weighting Approach. *Statistical Methods in Medical Research, 16*(3), 259-275.
Coakley, J.
2008 Militant Nationalist Electoral Support: A Measurement Dilemma. *International Journal of Public Opinion Research, 20*(2), 224-236.
Durand, Claire, André Blais, & Mylène Larochelle
2004 The Polls in the 2002 French Presidential Election: An Autopsy. *Public Opinion Quarterly, 68*(4), 602-622.
Erikson, Robert S., Michael B. Mackuen, & James A. Stimson
2002 *The Macro Polity*. New York: Cambridge University Press.
Flannelly, Keven J., Laura T. Flannelly, & Malcolm S. McLeod, Jr.
1998 Comparison of Election Predictions, Voter Certainty and Candidate Choice on Political Polls. *Journal of the Market Research Society, 40*(4), 337-346.
Flannelly, Laura T., Keven J. Flannelly, & Malcolm S. McLeod, Jr.
2000 Comparison of Forced-Choice and Subjective Probability Scales Measuring Behavioral Intentions. *Psychological Reports, 86*(1), 321-332.
Gelman, Andrew, Gary King, & Chuan-Hai Liu
1998 Not Asked and Not Answered: Multiple Imputation for Multiple Surveys. *Journal of the American Statistical Association, 93*(443), 846-857.
Graham, John W.
2009 Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology, 60*, 549-576.
He, Yulei , & Trivellore E. Raghunathan
2009 On the Performance of Sequential Regression Multiple Imputation Methods with Non Normal Error Distributions. *Communications in Statistics-Simulation and Computation, 38*(4), 856-883.
Ho, Shirley S., & Douglas. M. McLeod
2008 Social-psychological Influences on Opinion Expression in Face-to-Face and Computer-Mediated Communication. *Communication Research, 35*(2), 190-207.
Honaker, James, Gary King, & Matthew Blackwell
2009 Amelia II: A Program for Missing DATA. Retrieved April 24, 2009, from Amelia Software Web Site: http://gking.harvard.edu/amelia/.
Horton, Nicholas J., & Ken P. Kleinman
2007 Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *American Statistician, 61*(1),

79-90.

Imai, Kosuke, Gary King, & Olivia Lau

2009 Zelig: Everyone's Statistical Software. Retrieved May 5, 2009, from http://gking.harvard.edu/zelig/.

Kam, Cindy D.

2007 Implicit Attitudes, Explicit Choices: When Subliminal Priming Predicts Candidate Preference. *Political Behavior*, *29*(3), 343-367.

King, Gary, James Honaker, Anne Joseph, & Kenneth Scheve

2001 Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, *95*(1), 49-69.

Lazarsfeld, Paul Felix, Bernard Berelson, & Hazel Gaudet

1968 *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.

Little, Roderick J. A., & Donald B. Rubin

2002 *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, N.J.: Wiley.

McDevitt, Michael, Spiro Kiousis, & Karin Wahl-Jorgensen

2003 Spiral of Moderation: Opinion Expression in Computer-Mediated Discussion. *International Journal of Public Opinion Research*, *15*(4), 454-470.

Paul, Christopher, William M. Mason, Daniel McCaffrey, & Sarah A. Fox

2008 A Cautionary Case Study of Approaches to the Treatment of Missing Data. *Statistical Methods and Applications*, *17*(3), 351-372.

Penn, David A.

2007 Estimating Missing Values from the General Social Survey: An Application of Multiple Imputation. *Social Science Quarterly*, *88*(2), 573-584.

Raghunathan, Trivellore E.

2004 What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annual Review of Public Health*, *25*, 99-117.

Rubin, Donald B.

1987 *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Särndal, Carl-Erik, & Sixten Lundström

2005 *Estimation in Surveys with Nonresponse*. Hoboken, N.J.: Wiley.

Stuart, Elizabeth A., Melissa Azur, Constantine Frangakis, & Philip Leaf

2009 Multiple Imputation with Large Data Sets: A Case Study of the Children's Mental Health Initiative. *American Journal of Epidemiology*, *169*(9), 1133-1139.

Sullivan, J. L., & Transue, J. E.

1999 The Psychological Underpinnings of Democracy: A Selective Review of Re-

search on Political Tolerance, Interpersonal Trust, and Social Capital. *Annual Review of Psychology*, *50*, 625-650.

Templ, M., & P. Filzmoser

2008 Visualization of Missing Values Using the R-package VIM. Retrieved April 2, 2010, from http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-1complete.pdf.

Weisberg, Herbert F.

2005 *The Total Survey Error Approach: A Guide to the New Science of Survey Research.* Chicago: University of Chicago Press.

Wood, Angela M., Ian R. White, & Patrick Royston

2008 How Should Variable Selection be Performed with Multiply Imputed Data? *Statistics in Medicine*, *27*(17), 3227-3246.

Zuckerman, Alan S., Laurence A. Kotler-Berkowitz, & Lucas A. Swaine

1998 Anchoring Political Preferences: The Structural Bases of Stable Electoral Decisions and Political Attitudes in Britain. *European Journal of Political Research*, *33*(3), 285-321.