Planned Missingness with Multiple Imputation: An Application in Election Day Surveys

René Bautista Marco A. Morales

Appendix

A Multiple Imputation

For the imputation, we use *Amelia II* (Honaker, King and Blackwell 2007) - which employs the bootstrapping-based Expectation-Maximization (EMB) algorithm - to make the imputations (Honacker and King 2010). Briefly, missing values are imputed linearly from the model:

$$\tilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\epsilon}_i \tag{A.1}$$

where the tildes denote random draws from appropriate posterior distributions for parameters (β) and the random term (ϵ), and the imputations are a function of the observed data ($D_{i,-j}$). Table A.1 reports the statistics on missingness in our data set.

We generated $m = 10$ imputed data set on which the analysis was performed.[1] A good imputation for survey data should account for the sample design (Rubin 1996; King, Honaker, Joseph and Scheve 2001), otherwise risking producing inconsistent estimates (Reiter, Raghunathan and Kinney 2006). To account for this fact, the imputation included "cluster effect" dummy variables for each precinct in the sample. Similarly, the survey-design weights were included in the imputation to account for the variables used to select the precincts (Judkins 1998). Were these variables irrelevant to the imputation, we would be producing inefficient but not inconsistent estimates (Reiter et al. 2006).

The point estimates are computed according to "Rubin rules" (Rubin 1976, 1996). For simplicity, we use the notation by (King et al. 2001) and define q as the quantity of interest, for which we calculate a point estimate as:

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j \tag{A.2}$$

[1] Graphic diagnostics from the imputations as suggested by Abayomi, Gelman and Levy (2008) are available upon request.

and the variance of the point estimate as the sum of the *within* and the *in-between* imputations variance:

$$\begin{aligned} SE(q)^2 =& \bar{w} + b \\ =& \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + \left(1 + \frac{1}{m}\right)\frac{\sum_{j=1}^{m}(q_j - \bar{q})^2}{m-1} \end{aligned} \tag{A.3}$$

The quantity of interest ($\bar{q}$) is distributed t with degrees of freedom equal to:

$$d.f. = (m-1)\left[1 + \frac{1}{m+1}\frac{\bar{w}}{b}\right]^2 \tag{A.4}$$

Readers interested in further details on Multiple Imputation and the algorithms developed to implement it are referred to Rubin (1987), King et al. (2001), and Horton and Kleinman (2007). Readers interested in MI applications to political science are referred to King et al. (2001) and Honacker and King (2010).

Table A.1: Descriptive statistics on MI variables

Variable	Obs	missing	Mean	Std. Dev.	Min	Max
Female	7763	1	0.483	0.500	0	1
Presidential approval	7609	155	0.822	1.839	-3	3
Split-ticket vote	7379	385	0.479	0.500	0	1
Education	7761	3	3.032	1.287	1	5
Income	6782	982	4.028	1.801	1	7
Social class	7581	183	2.149	0.945	1	5
Vote for President	6694	1070	2.082	0.970	1	5
Vote for Deputies	6670	1094	2.113	0.974	1	5
Remembers scholarship ad	2022	5742	0.253	0.435	0	1
Remembers schools ad	2023	5741	0.215	0.411	0	1
Remembers social insurance ad	2024	5740	0.284	0.451	0	1
Remembers housing ad	2021	5743	0.253	0.435	0	1
Remembers Oportunidades ad	2022	5742	0.283	0.451	0	1
Scholarship beneficiary	1977	5787	0.212	0.409	0	1
School program benef.	1975	5789	0.119	0.324	0	1
Popular insurance benef.	1976	5788	0.198	0.399	0	1
Housing program benef.	1977	5787	0.133	0.340	0	1
Oportunidades benef	1977	5787	0.281	0.449	0	1
Negative campaign	1470	6294	4.040	2.850	1	9
Positive campaign	1471	6293	3.789	2.846	1	9
Respondent ideology	1374	6390	4.749	2.077	1	7
PAN ideology	1411	6353	4.820	2.197	1	7
PRI ideology	1410	6354	4.361	2.217	1	7
PRD ideology	1397	6367	3.147	2.183	1	7
Econ retrospective eval	1771	5993	0.120	1.346	-3	3
Econ prospective eval	1505	6259	1.054	1.269	-3	3
Party ID	1648	6116	3.927	2.374	1	8
Expected winner	1354	6410	1.969	0.909	1	5
Strategic voter	1699	6065	0.068	0.252	0	1
Political interest	1790	5974	3.075	0.822	2	4
Mexico democracy	1657	6107	0.769	0.421	0	1
FCH ideology	1409	6355	4.820	2.179	1	7
RMP ideology	1405	6359	4.397	2.115	1	7
AMLO ideology	1403	6361	3.264	2.219	1	7
Prefers balance of power	1788	5976	0.195	0.396	0	1
Voted for change	1788	5976	0.634	0.482	0	1

B Econometric analysis

The desirable properties of our multiply imputed data set would be partially wasted if we do not use an econometric model that more closely resembles the assumptions made by the theoretical model we are testing. So we advocate the use of an econometric model that makes the most efficient use of all available information, instead of discarding it by assuming it irrelevant. We do so based on two main concerns: not imposing an unwarranted Independence of Irrelevant Alternatives (IIA) assumption on voters, and failing to take advantage of readily available specifications that take into account individual and candidate-specific characteristics.

Common accounts of the 2006 election assume that the presence of a third candidate altered the probability of voting for either of the remaining two candidates. As the campaign was reaching its end, pollsters tried to forecast PRI's voting share knowing that it would modify the distribution of votes for PAN and PRD's presidential candidates. Thus, it was not uncommon to read that "had Madrazo been a better candidate" or "if Madrazo drops from the race" we would have observed a different outcome. If we were to ignore this feature, we would need to assume that Madrazo was indistinguishable from Calderón or López Obrador in the voter's mind (Hausman and Wise 1978). This would imply that the ratio of the probabilities *of an individual* voting for Calderón relative to López Obrador does *not* change whether Madrazo appears as a candidate or not (Alvarez and Nagler 1998). Most likely an unrealistic assumption, or at least one in need of empirical verification. Incorrectly assuming IIA may lead to inconsistent estimates and to incorrect conclusions on the 2006 election (Alvarez and Nagler 1998). An additional point is that some of the arguments advanced to explain vote choice are related to features of the candidates, such as better image, negative ads, and the like. To address these problems and explicitly accounting for both candidate

and individual-specific features while relaxing the IIA assumption, we present estimates from a multinomial probit model.

The multinomial probit is motivated as a Random Utility Model (RUM) where utility is determined by a *systemic component* that reflects the average behavior of individuals given a set of observed characteristics related to individuals and choices, and by a *stochastic (random) component* that accounts for deviations from the average behavior and is assumed to be determined by unobserved differences in tastes across individuals as well as unobserved characteristics of the alternatives. Note that on Eq. B.1 below, $\beta X_{ij} + \psi_j a_i$ is the systemic component and ϵ_{ij} the random component.

To apply the model to our case, we assume that individuals seek to maximize the utility they obtain from a candidate and choose from the set of available alternatives according to this criterion. The utility (U_{ij}) that each voter derives from the alternatives is defined as:

$$U_{ij} = \beta X_{ij} + \psi_j a_i + \epsilon_{ij} \tag{B.1}$$

where a_i contains characteristics of individual i, X_{ij} contains characteristics of candidate j according to individual i, ϵ_{ij} is the random component. Note that β is a vector of candidate-specific parameters and ψ_i is a vector of individual-specific parameters to be estimated. We estimate a set of β and two sets of ψ_j. ϵ_{ij} are assumed to be distributed multivariate normal, which requires specifying the correlation between each alternative's random components:

$$\epsilon_{ij} \sim MVN(\mathbf{0}, \mathbf{\Sigma}) \tag{B.2}$$

The IIA assumption is overcome by allowing the covariance matrix $\mathbf{\Sigma}$ to have non-zero correlations terms between the ϵ_{ij} and estimating it. To identify the estimation of parameters, the coefficients for PRI are normalized to zero, thus producing coefficients for PAN and PRD

relative to PRI. To identify and facilitate the estimation of the elements in $\mathbf{\Sigma}$, the disturbances are assumed to be homoscedastic ($\sigma^2_{PAN} = \sigma^2_{PRI} = \sigma^2_{PRD} = 1$) and the correlation between PAN and PRI's random component is assumed to be zero ($\sigma_{PAN,PRI} = 0$). This leads to the estimation of the covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & & \\ 0 & 1 & \\ \sigma_{PAN,PRD} & \sigma_{PRI,PRD} & 1 \end{bmatrix} \tag{B.3}$$

Estimates are presented on Table B.1 below:

Table B.1: Multinomial Probit results for Presidential election

	PAN/PRI	**PRD/PRI**
Candidate distance	-0.004***	
	(0.001)	
Negative ad	-0.161***	
	(0.039)	
Positive Ad	0.369***	
	(0.050)	
Ad scholarship	0.062	0.055
	(0.066)	(0.053)
Ad schools	0.010	0.014
	(0.072)	(0.050)
Ad insurance	0.032	0.012
	(0.089)	(0.047)
Ad housing	0.004	-0.050
	(0.080)	(0.047)
Ad Oportunidades	0.016	-0.005
	(0.079)	(0.042)
Scholarship	0.055	-0.035
	(0.077)	(0.048)

Continued on next page

	PAN/PRI	**PRD/PRI**
Schools	-0.043	-0.059
	(0.079)	(0.046)
Insurance	0.091	-0.067
	(0.131)	(0.080)
Housing	0.092	0.304***
	(0.104)	(0.066)
Oportunidades	-0.269***	-0.166***
	(0.069)	(0.046)
ID PAN	0.174***	0.023
	(0.033)	(0.022)
ID PRI	-0.239***	-0.192***
	(0.030)	(0.019)
ID PRD	0.002	0.142***
	(0.032)	(0.021)
Econ Retro Good	0.111**	0.007
	(0.047)	(0.030)
Econ Retro Bad	-0.073*	-0.006
	(0.040)	(0.024)
Econ Prosp Good	0.117	0.132**
	(0.088)	(0.059)
Econ Prosp Bad	-0.106	-0.100**
	(0.073)	(0.046)
Pres Approval Good	0.847***	0.052
	(0.074)	(0.048)
Pres Approval Bad	-0.730***	0.000
	(0.089)	(0.049)
Uncertainty FCH	-0.028***	-0.015***
	(0.006)	(0.004)
Uncertainty AMLO	-0.014**	0.014***
	(0.006)	(0.004)
Uncertainty RMP	0.005	0.003
	(0.008)	(0.005)
Age 18-24	-0.047	-0.062
	(0.107)	(0.072)
Age 25-40	0.048	-0.055
	(0.067)	(0.045)
Age 41-60	-0.027	-0.051
	(0.055)	(0.037)

Continued on next page

	PAN/PRI	**PRD/PRI**
Ed primary	0.171***	0.099
	(0.096)	(0.064)
Ed secondary	0.415***	0.301***
	(0.070)	(0.047)
Ed highschool	0.513***	0.362***
	(0.076)	(0.052)
Ed college	0.801***	0.487***
	(0.073)	(0.050)
Female	0.043	-0.066*
	(0.053)	(0.035)
Low class	-0.433***	0.201**
	(0.125)	(0.091)
Middle Class	-0.372***	0.085**
	(0.057)	(0.039)
Northwest	0.226***	-0.186***
	(0.085)	(0.057)
Northeast	0.037	-0.409***
	(0.071)	(0.051)
Southeast	-0.278***	-0.167
	(0.064)	(0.043)
Southwest	0.511***	0.492
	(0.081)	(0.050)
Urban	0.160	0.081
	(0.108)	(0.076)
Rural	-0.161***	-0.008
	(0.058)	(0.039)
Intercept	-0.196***	-0.205***
	(0.048)	(0.020)
$\sigma_{PAN,PRD}$	0.297**	
	(0.152)	
$\sigma_{PRD,PRI}$	0.338***	
	(0.140)	
Log-Likelihood	-5824.371	
LR-test	$\chi^2_{[79]}$=901.421***	
n	6,455	
MI sets	10	

Significance: 1% *** / 5% ** / 10%* two-tailed.

Given the high missingness in our data, and in order to avoid a higher estimation error derived from using imputed vote choices, we discard them from the analysis. Therefore

the the presidential election analysis is performed with $n = 6,455$. All available information from these cases was used for the imputation process. Discarding imputed y's from the analysis has been shown to produce at least as good estimates as those produced when using all - observed and imputed - y's, but discarding imputed y's produces more efficient estimates with high missingness or a low m (von Hippel 2007). That is because cases with missing y's contain no information about the parameters we are trying to estimate in the models.

For further details on the multinomial probit model, readers are directed to Hausman and Wise (1978), and Greene (2003). For specific applications to multicandidate elections in political science, readers are directed to Alvarez and Nagler (1995, 1998) and examples cited therein.

C Simulating vote probabilities

Following Hausman and Wise (1978) and Alvarez and Nagler (1995), the probabilities of voting for a given candidate for three choices are given by:

$$P_{i,PAN} = \Phi\left(\frac{(\bar{U}_{i,PAN} - \bar{U}_{i,PRI})}{\sqrt{\sigma^2_{PAN} + \sigma^2_{PRI} - 2\sigma_{PAN,PRI}}}\right) \tag{C.1}$$

$$P_{i,PRD} = \Phi\left(\frac{(\bar{U}_{i,PRD} - \bar{U}_{i,PRI})}{\sqrt{\sigma^2_{PRD} + \sigma^2_{PRI} - 2\sigma_{PRD,PRI}}}\right) \tag{C.2}$$

$$P_{i,PRI} = 1 - P_{i,PAN} - P_{i,PRD} \tag{C.3}$$

where:

$$\bar{U}_{ij} = \beta X_{i,j} + \psi_j a_i, \quad j \in \{PAN, PRI, PRD\} \tag{C.4}$$

Note that $\sigma^2_{PRD} = \sigma^2_{PRI} = \sigma^2_{PRD} = 1$ since we assumed homoscedasticity, and that $\sigma_{PAN,PRI} = 0$ also by assumption to identify the parameters in $\mathbf{\Sigma}$. Similarly, we normalized $\bar{U}_{i,PRI} = 0$ to identify the estimation of the parameters.

Simulations are produced as per the procedure set forth in King, Tomz and Wittenberg (2000). Briefly, the algorithm consists of:

a) obtain estimates for β and its covariance matrix $\text{Var}(\beta)$ from the m models using Eqs. A.2 and A.3. Generate $d = 8000$ draws from the distribution

$$\tilde{\beta} \sim MVN(\hat{\beta}, Var(\hat{\beta})) \tag{C.5}$$

which reflects the estimation uncertainty that derives from not having infinite observations for the estimation.

b) Determine a value for each explanatory variable (X_{ij}, a_i), compute the utility $(\bar{U}_{ij})$ as defined in Eq. C.4 using a draw from $\tilde{\beta}$ in Eq. C.5. Plug this value into Eqs. C.1 or C.2 to obtain the probability of voting for a given candidate (P_{ij}). Repeat the process d times and compute the expected value $\tilde{E}[P_{ij}] = \sum_{k=1}^{d} P_{ij}/d$.

c) *First differences* (King 1998) require computing the probabilities of voting for a given candidate as defined in b), with the particularity that it is computed twice. Once with the variable of interest set at the low value (P_{ij}^L) and once at the high value (P_{ij}^H). The first difference is simply $\mathcal{D}_i = (P_{ij}^H - P_{ij}^L)$ and its point estimate $\tilde{E}[\mathcal{D}] = \sum_{k=1}^{d} \mathcal{D}_i/d$.

To define our "typical" individual we set all continuous variables at their means and categorical variables at their mode rendering a 41 year-old, primary-educated, middle class, urban resident of the southwest, who does not remember seeing any ads and is not a beneficiary of government programs, who evaluates positively the president as well as the economy in the past year and the next year, strongly identifies with PRI, thinks López Obrador generated the most negative campaign and Calderón the most positive one, and has the mean distance to all candidates as well as the mean uncertainty levels about candidates' positions.

D Potential bias in the projection of election results

In order to show the particular advantage of exit polls, which extends to planned missingness-multiple imputation (PM-MI) on this regard, a simple comparison of vote estimates produced by our exit poll and pre- and post-electoral surveys for this same election are shown. Figure D.1 compares point estimates and their associated theoretical sampling error for each of these estimates. The actual election results are denoted by the vertical line with the official percentage of vote marked above. The first estimate (Pre-election) corresponds to the one generated by the Mexico 2006 Panel study[2] conducted over the month prior to the election. The second estimate (PM-MI) is produced by the exit poll data. The third estimate (Post-election) corresponds to the raw estimate of the post-electoral survey of the Mexico 2006 Panel Study. The fourth estimate (Post-election rev) corresponds to the same Panel estimate but "corrected" to exclude non-voters. Since registered Mexican voters are issued a special ID or "electoral card" that is marked every time an individual casts a ballot, the survey included an indicator for those cases where the interviewer could directly verify the existence of the mark on the voter's ID.[3]

Figure D.1 shows that the realized vote shares from the exit poll estimates are closer to the official results than estimates from pre- and post-electoral survey data. While the estimates were accurate for the PAN candidate (winner of the election) in the post-election survey, the estimates and margins of error were notoriously off for the PRI and PRD can-

[2]Senior Project Personnel for the Mexico 2006 Panel Study include (in alphabetical order): Andy Baker, Kathleen Bruhn, Roderic Camp, Wayne Cornelius, Jorge Domínguez, Kenneth Greene, Joseph Klesner, Chappell Lawson (Principal Investigator), Beatriz Magaloni, James McCann, Alejandro Moreno, Alejandro Poiré, and David Shirk. Funding for the study was provided by the National Science Foundation (SES-0517971) and *Reforma* newspaper; fieldwork was conducted by *Reforma* newspaper's Polling and Research Team, under the direction of Alejandro Moreno. `http://web.mit.edu/polisci/research/mexico06`.

[3]This is by no means a perfect measure, as some actual voters might not carry their ID with them at the time of the interview, thus potentially discarding actual voters from the sample. Doing so, reduces the sample nearly by half, but this only ensures that the voters included in the sample are actually voters and cannot affect the accuracy of reported vote choice.

didates, whose votes were under and overestimated respectively in the post-election wave. The pre-election wave only produced a good estimate for the PRD candidate, but was off for the other two candidates. This is not unexpected since exit polls and other surveys draw their sample from different populations: exit polls sample from actual voters on Election Day, while pre and post-electoral surveys sample from potential voters and try to screen voters from among them. We would expect exit polls to be more accurate than post-electoral surveys simply as a result of survey design.

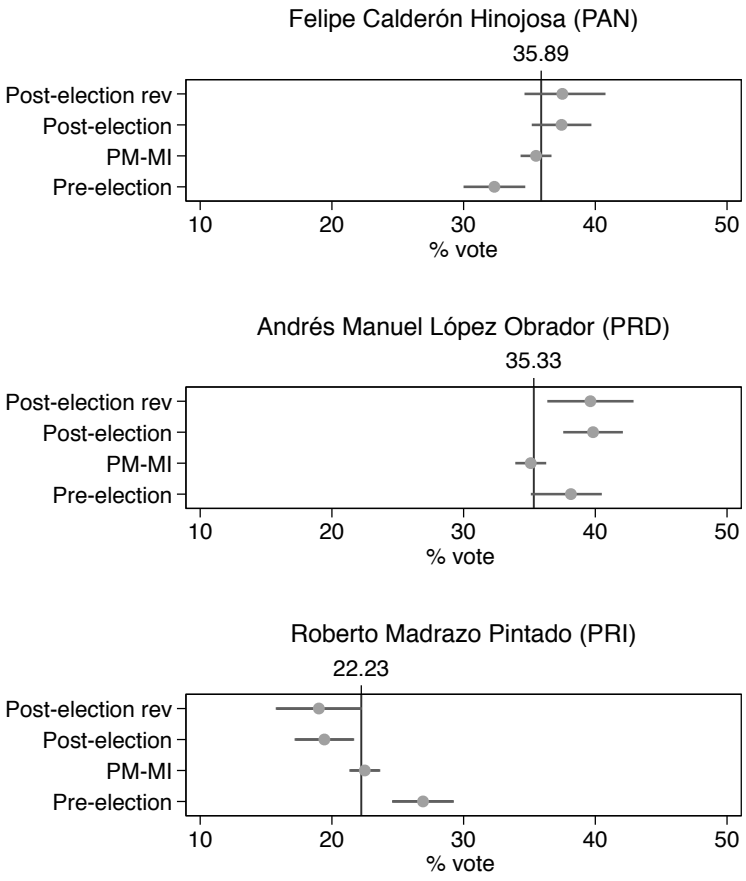

Figure D.1: Point estimates - and associated 95% confidence intervals - of vote shares for each candidate generated with data from each one of the row sources. *Pre-election* estimates come from the Mexico 2006 Panel Study pre-election wave. *PM-MI* estimates come from the *Parametría*'s 2006 exit poll. *Post-election* estimates come from the Mexico 2006 Panel Study post-election wave. *Post-election rev* estimates come from the Mexico 2006 Panel Study pre-election wave, corrected for verified voters. Titles on each graph correspond to the corresponding candidate.

References

Abayomi, K., Gelman, A., and Levy, M. (2008), "Diagnostics for Multivariate Imputations," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3), 273–291.

Alvarez, R. M., and Nagler, J. (1995), "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election," *American Journal of Political Science*, 39(3), 714–744.

Alvarez, R. M., and Nagler, J. (1998), "When Politics and Models Collide: Estimating Models of Multiparty Elections," *American Journal of Political Science*, 42(1), 55–96.

Greene, W. H. (2003), *Econometric Analysis*, fifth edn, New York, NY: Prentice Hall.

Hausman, J. A., and Wise, D. A. (1978), "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogenous Preferences," *Econometrica*, 46(2), 403–427.

Honacker, J., and King, G. (2010), "What to do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science*, 54(2), 561–581.

Honaker, J., King, G., and Blackwell, M. (2007), AMELIA II. A program for missing data. Version 1.1.27,. Cambridge, MA: Harvard University [GKing.Harvard.edu/amelia/].

Horton, N. J., and Kleinman, K. P. (2007), "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician*, 61(1), 79–90.

Judkins, D. R. (1998), "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys: Comment," *Journal of the American Statistical Association*, 93(443), 861–864.

King, G. (1998), *Unifying Political Methodology. The Likelihood Theory of Statistical Inference*, Ann Arbor, MI: University of Michigan Press.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001), "Analyzing Incomplete Political Science Data: An alternative algorithm for multiple imputation," *American Political Science Review*, 95(1), 49–69.

King, G., Tomz, M., and Wittenberg, J. (2000), "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science*, 44(2), 341–355.

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006), "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data," *Survey Methodology*, 32(2), 143–149.

Rubin, D. B. (1976), "Inference and Missing data," *Biometrika*, 63(3), 581–592.

Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys*, New York, NY: Wiley & Sons.

Rubin, D. B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91(434), 473–489.

von Hippel, P. (2007), "Regression with Missing Ys: An Improved Strategy for Analyzing Multiply-Imputed Data," *Sociological Methodology*, 37(1), 83–117.