

ESTIMATING THE SOURCES OF BIAS: A PREDICTIVE MODEL TO ANALYZE POLLING PERFORMANCE FOR THE CONGRESSIONAL RACE IN MEXICO

Miguel Angel Torhton
 The George Washington University
 ✉ mtorhton@email.gwu.edu

Abstract: *This paper proposes a Montecarlo simulation analysis to estimate the sources of bias in the main national surveys for the Congressional Race in Mexico from 1997 to 2012. According to similar approaches for other electoral systems, the main sources of error analyzed were the pollster bias. The results of such analysis may be indicative of some of the main problems facing electoral polls in Mexico. The direction and patterns of bias for each pollster were identified. The sources and the characteristics of systematic errors could be of great value to understand the complexities of the Mexican electoral system and the challenges for polling agencies.*

Keywords: *bias, polls, Montecarlo simulation, Mexican Congress, elections*

Resumen: *Este artículo propone un análisis de simulación Montecarlo para estimar las fuentes del sesgo en las principales encuestas nacionales para las elecciones legislativas en México de 1997 a 2012. A partir de aproximaciones similares para otros sistemas electorales, este artículo analiza las fuentes de error producidas por el tiempo antes de la elección y el sesgo de la encuestadora. Los resultados de este análisis pueden ser indicativas de algunos de los principales problemas de las encuestas electorales en México. La dirección y los patrones de sesgo de cada encuestador fueron identificados. Las fuentes y las características de estos errores podrían ser valiosos para entender las complejidades del sistema electoral mexicano y los retos para las agencias encuestadoras.*

Palabras clave: *sesgo, encuestas, simulaciones Montecarlo, Congreso Mexicano, elecciones*

Introduction

The morning of July 2nd of 2012, Mexicans from each corner of the country woke up to receive the highly expected news of the election of Enrique Peña Nieto as the new President of this young and colorful democracy. For most public commentators, the outcome of this election was almost a confirmation of the predictions of the polls. One month

before the election, the average in some of the most important polls showed that the difference between the first and the second place was 12.1% (see table below). However, the surprise for most television pundits, political parties and citizens were that the actual difference between the first and the second place was only of 6.6%.

Table 1. Electoral Forecast for Mexican Presidential Elections in June 2012

	PAN	PRI	PRD	PANAL	Final gap between 1st and 2nd
	3 rd	1 st	2 nd	4 th	
Mitofsky	21%	36%	25%	1.90%	10.7%
GEA/ISA-Milenio	23%	45%	29%	3.50%	15.9%
Parametria	24%	43%	30%	3%	13.0%
Berumen y asociados	21%	36%	32%	-	4.1%
Covarrubias y Asociados	26%	41%	30%	3%	11.0%
Uno TV/María de las Heras	23%	40%	32%	4.55%	7.7%
Ipsos/Bimsa	24%	41%	34%	1%	7.0%
El Universal/Buendía & Laredo	24%	45%	28%	2.70%	17.1%
BGC-Excelsior	25%	44%	28%	3%	16.0%
GEA/ISA-Milenio	22%	47%	29%	2.20%	18.4%
Average	23%	42%	30%	3%	12.1%

Source: Compilation based on polling agencies reports and final electoral count

Among political parties, the gap between the predictions and the actual results kindred the debate about the role of the polls for Mexican democracy. For the party in the second place (PRD), the polls were biased to improve the public perception of the first place (PRI). The victory of Enrique Peña Nieto was questioned for the supposed coalition of pollsters to favor his campaign. The professional ethics of some of the most important pollsters went on the adverse trial of public opinion. The credibility of these institutions, a key factor in the business of surveying people, plummeted. After the so considered scandalous failure of the polls, there were some serious attempts to regulate the exposure and information contained in the polls. The main pollsters avoided this

new set of electoral reforms by arguing that their polls were not predictions and that the biases were unintentional.

This fascinating scenario — that reunites probabilistic predictions with political outcomes and policy discussion — is the main reason why this paper explores the biases in some of the main electoral polls in Mexico and proposes a predictive model of the electoral outcomes of the 2015 Congressional race¹. This aims to address the concerns of the citizenship interested in keeping track of the performance of candidates in the current electoral cycle.

Electoral Policies in New Democracies

New democracies around the world are discovering the principles of political competition by trial and error. The oldest democracies did it that way —it would be hard to assert that they have finished their learning process. Political competition is ever changing. For that reason, the uncertainty about the effects of electoral rules on electoral outcomes supposes a fundamental challenge for nascent electoral institutions. How to regulate the unknown? Seems like a common question. How to be certain that institutions aimed to protect political competition are not likely to demolish the relatively recent trust of people in a democratic system? This broad question will certainly be at the core of many political changes in new democracies for the next decades.

In the case of Mexico, the policy outcome of this research is very clear. It is fundamental to learn more about the biases in polls. Information about the probabilistic nature of polls will help to expand our comprehension of the uses and misuses of these instruments. This could be relevant in the case of future attempts to regulate the role of

¹ The first version of this model was presented for the Models for Public Policy Analysis course at The George Washington University on June 2015, one month before the actual electoral process.

polling agencies in electoral races. Also, the attempt to develop a predictive model is aimed to shed light to some characteristics of the Mexican electoral system and to provide more accurate estimates of the outcomes of the elections. It is important to stress that this is a novel approach for the Mexican electoral system. This attempt to generate a predictive model is focused on understanding the characteristics of the Mexican electoral system and in finding the possible traps for predicting results in this context.

Theoretical Approach

The main inspiration for this work is the electoral forecasts developed by Nate Silver in the blog Five Thirty Eight (FIVETHIRTYEIGHT.COM, 2015). This website forecasts the probabilities of certain outcomes in politics, sports and entertainment. According to Silver, it is a way to differentiate the signals from the noise (SILVER, 2010). This is a thought-provoking metaphor to understand the way we approach knowledge. In a world full of information, which are the specific data that can be used to predict an outcome? Although there are important critiques of this kind of works, the success of Silver's electoral forecasts for the U.S. Presidential and Senate races are basic to understand the assumptions followed in this research. The most important assumptions taken from Nate Silver's work are the following:

Assumption 1

A key element of Nate Silver's models is that the average results of polls are weighted. This is used to correct for the biases in the polls. However, this bias correction cannot be the only component to understanding an electoral outcome. According to Nate Silver, models that are only based on the results of polls tend to show

enormous disparities. For that reason, his models are corrected for certain political, financial, religious, ethnic and demographic factors in each state to generate an “ideal” poll to be compared with the actual poll. This is particularly relevant in the case of Mexico, in the sense that this could be useful to test the biases in the polls and compare the results of these instruments to actual predictions.

Assumption 2

Time plays an important role in the precision of a poll. The closest to the election, the most precise the estimates of the polls. This will be a second variable to measure the precision of the polls in Mexico.

Assumption 3

Nate Silver claims that models based on fundamentals tend to be bad predictors about the outcomes of Presidential Elections. Fundamentals are contextual variables that American political scientists used to predict the outcomes of an election: presidential approval rates, the state of the economy and the approval of exterior conflicts. Due to the need to explore the precision of polls instead of other demographic factors, this model will ignore this kind of variables.

Nate Silver, however, is far from being the only person behind electoral forecasts. The monograph *Methods of Forecasting American Election Outcomes*, compiled by Chandrasekhar Putcha (2010), describes some of the most important approaches for modeling the outcomes of elections. From this reunion of different perspective, it is possible to conclude that most models are based in two components:

1. Projections based on polls
2. Time-series forecasting models.

The idea of tracking some variables for a long period of time (time series) is a basic condition to define a model. Even if democratic competition in Mexico is relatively recent, it is possible to find some patterns in the electoral outcomes that could provide a more accurate estimate of electoral outcomes than those portrayed by polls.

Data and parameters

The data about previous elections was taken from the electoral dataset of the National Electoral Institute of Mexico (IFE, 2012). It comprises information about electoral turnout from 1997 to 2012 (3 intermediate elections in total). Also, the analysis of electoral districts from 2009 to 2012 is based on previous works by Javier Aparicio (APARICIO, 2010). The information about polls results was taken from the historical results of the polling agencies analyzed and taken directly from its websites.

It is important to consider that the American and the Mexican political system are completely different, which radically changes some of the assumptions for the model. For example, there is no reelection, no Electoral College and polls are only rarely conducted at the state level for national races. It is also relevant to stress that the literature about forecasting models for Mexican elections is almost inexistent. The closest to this is a very interesting geographical forecast of electoral outcomes in Mexico City performed by Carlos Vilalta (VILALTA, 2008).

Due to the conditions of the Mexican political system, the most interesting electoral competition to analyze the bias of pollsters and to build a predictive model is the **Congressional Race for the Federal**

Chamber of Deputies. It is very likely that there are more variables affecting the Presidential Election, which is a single member district race for the whole country. Instead, the Chamber of Deputies offers each three years a single member district competition for 300 districts with a proportional formula that ends up over representing small parties in the Lower House of the bicameral Legislative regime in Mexico.

Based on the literature research described above and in the conditions of the electoral system studied, the predictive model for the 2015 Congressional Race in Mexico based on these two factors:

- a) a correction of the biases in polls- A
- b) the historical results in electoral districts- B

$$\textit{Electoral Outcome} = P(\text{estimate of A}) + P(\text{estimate of B})$$

Both factors are expected to yield an average result for each party in the electoral race. Also, each factor will be weighted to predict the actual average for each party. The initial weight will be A (.5) / B (.5). The outcomes of different weights will be compared through sensitivity analysis. For instance, A (.6) / B(.4), A (.7) /B (.30), A(1) / B(0). In the following section, the paper will explore further the characteristics and assumptions to model these factors.

Models

This model is the first attempt to generate a predictive forecast for the outcomes of the 2015 Mexican Legislative Election. This model considers the polls for the 2015 Congressional Race as an average and assumes that the distribution of electoral preferences for each party is normal. Only five of the most important public opinion polls will be selected due to data availability: Reforma, Mitofsky, GEA/ISA, Parametría, and BGC. The time before the election of these polls varies between two and three months from the election of June 7th, 2015.

Table 2. Forecast for 2015 Congressional Election

Pollster	Reforma	Mitofsky	GEA/ISA	Parametria	BGC
PAN	22%	23.5%	28%	24%	22%
PRI	32%	32.7%	33%	32%	36%
PRD	14%	16%	14%	13%	15%
MORENA	8%	11.2%	9%	10%	7%
PVEM	7%	7.1%	7%	8%	8%
Date of the poll	mar-15	abr-15	mar-15	abr-15	abr-15

Source: Compilation based on polling agencies report

The model weights the bias-in-polls model with the most likely results of a model derived from two most recent elections at district level. The bias in polls model was performed with a Montecarlo simulation comprised by the average biases of polling agencies. The district level model compares three types of districts: Bastion 1, where the ruling party has never lost an election; Unsafe Bastion 2, where the ruling party has lost before but still has a victory with a large difference from the second place; and, Competitive 3, where there have been victories from the opposition and the difference between the first place and the second is less than 5%.

By comparing the results for each party, the average of the 1000 runs allows to identify the direction and the size of the bias for each pollster. It also offers a useful estimate of the “actual” electoral preference despite the bias of the pollster and the bias caused by the date in which the poll was fielded.

1.1. The pollster bias

The pollster bias refers to the average difference between the predicted result of a pollster and the actual result. This average measures the bias for polls performed by each pollster in different elections from 1997 to 2012. It is possible to notice below that the average bias for pollster goes from 28% to only 3%. This is an interesting indicator for

the precision of each pollster. Each pollster bias is considered as an independent standard deviation of the average (the pollster's estimate for 2015) in the Montecarlo simulation.

Table 4. Average Pollster Bias (1997-2012)

Pollster	Average bias
BGC-Excelsior	28%
Buendía y Laredo	24%
Parametría	20%
GEA/ISA	20%
Mitofsky	16%
Covarrubias y Asociados	15%
Grupo Impacto Inteligente 360°	13%
Reforma	13%
Ipsos/Bimsa	7%
María de las Heras	3%

Source: Montecarlo simulation based on data from 1997-2012.

b. Historical results in electoral districts

The most accurate historical records of electoral outcomes can be found at the electoral districts. Each congressional race represents a competition in 300 districts. When comparing the two last elections, it is possible to notice that most of them are competitive districts (61%). Nearly 39% of the districts have only had a winner with an average gap between the first and the second place of 18% —so to say, not competitive elections. For this reason, it would be relevant to categorize different districts in terms of its competitiveness: Bastion 1, Unsafe Bastion 2 and Competitive 3. According to the historical results from 2009 up to date, it will be possible to calculate the likely number of districts won for each party. This will provide an estimate of the electoral preference that will be weighted with the biases in polls to reach an “actual” estimate of the electoral preferences for each party.

After these estimates, an approximate calculus of the proportional representation seats will be provided. The results will be followed by a sensitivity analysis according to different possible turnouts and different weights for the model.

Table 5. Turnout for intermediate elections (1997-2012)

	1997	2003	2009
Turnout	58%	42%	45%
Average	48%		
Maximum	58%		
Minimum	42%		

Source: Compilation based on turnout counts from Instituto Nacional Electoral.

Table 6. Probability of winning according to the type of district

	N	%	P of winning
Bastion 1	119	40%	1
Unsafe Bastion 2	115	38%	0.75
Competitive	66	22%	0.5
Total	300		

Source: Compilation based on district analysis from IFE

Results

In terms of the initial policy problem described in this paper, it is possible to establish that there is certainly some bias in public opinion polls. After the Montecarlo simulation with 1000 runs, the bias-in-polls model may allow us to conclude the following assertions about the polling companies in Mexico, an information that could be of great value for the Mexican electorate and for the polling agencies themselves. In more detail, some of the main characteristics of the bias are the following:

- It is more likely that the leftist parties (PRD and Morena) may be underestimated.
- The polling company that is more likely to overestimate PAN is GEA/ISA.
- The polling company that is more likely to overestimate PRI is BGC.
- Mitofsky is the most likely agency to overestimate MORENA.
- The PVEM is not likely to be underestimated by any of the polling agencies.

Table 8. Expected Results for 2015 Congressional Race after Corrections

	Reforma	Mitofsky	GEA/ISA	Parametría	BGC	Average
PAN	22.30%	21.33%	22.46%	21.77%	21.63%	21.9
PRI	32.1%	31.7%	32.0%	31.5%	32.2%	31.9
PRD	14.2%	14.3%	14.0%	14.4%	14.1%	14.2
MORENA	8.0%	9.3%	7.2%	7.6%	7.8%	8.0
PVEM	6.7%	6.9%	7.3%	7.1%	6.2%	6.9

Table 9. Patterns of Bias for pollsters after inclusion of the bias-in-polls model

	Reforma	Mitofsky	GEA/ISA	Parametría	BGC
PAN	0.054	1.554	6.054	2.054	0.054
PRI	0.071	0.771	1.071	0.071	4.071
PRD	-0.200	-1.800	0.200	1.200	-0.800
MORENA	-0.114	3.086	0.886	1.886	-1.114
PVEM	0.009	0.109	0.009	1.009	1.009

The model of district competition focused on the three major parties, since most of the districts have been traditionally won by them.

For this reason, this model overestimates the actual average for each party. However, it allows to measure the effect of previous district competition. As shown in a table above, the patterns of district competition allowed this paper to assign a probability of 100% of winning to Bastion 1, a 75% to Bastion 2 and 50% to Bastion 3. A decision tree with this information was used to calculate the possible electoral outcomes.

Table 10. Decision Tree- Modeled Distribution of Districts Won

Final Modeled Distribution of Districts won		
PAN	115	35%
PRI	136	42%
PRD	73	23%

Discussion

Although the predictive capacity of the model may be biased due to the complexities of the Mexican political system, this paper proposes an interesting challenge to address the complexities of electoral outcomes within Mexican political system and to try to approach them in a systematic, probabilistic way. From this initial analysis, it is clear that there is much work to be done. A deeper analysis on the district model may be necessary to take into account the formulas of proportional representation that are embedded in the definition of seats for the Mexican Chamber of Deputies. Further exploration on the causes of pollster bias may be necessary to improve the performance of polling agencies.

However, it is clear that the corrections proposed by the predictive model (bias-in-poll) allowed a more accurate representation of the actual results. As shown in the table below, the predictive model corrected

most of the polling bias that was present in the 2015 election. This paper suggests as a policy recommendation to encourage the use of this kind of models to improve the performance of polling agencies for coming elections.

**Table 11. Final Electoral Results (Jul. 2015) vs Predicted Model
(Jun. 2015)**

Party	Final Results	Predictive Model
PAN	21.01 %	21.9%
PRI	29.18 %	31.9%
PRD	10.87 %	14.2%
MORENA	8.39 %	8.0%
PVEM	6.91 %	6.9%

Source: Final Results from Instituto Nacional Electoral

References

- APARICIO, J. Un modelo Monte Carlo para la Cámara de Diputados en México, *Política y Gobierno*, v. XVII, n. 2, p. 351-379, 2010.
- FIVETHIRTYEIGHT.COM. Politics – FiveThirtyEight. [online] Available at: <http://fivethirtyeight.com/politics/> [Accessed 19 Aug. 2015].
- KUSCHIK, M. México: elecciones y el uso de las encuestas preelectorales. *Revista Mexicana de Ciencias Políticas y Sociales*, v. 45, n. 184, p. 103-127, 2002.
- PUTCHA, C. *Methods of Forecasting American Election Outcomes*. New York: Edwin Mellen Press, 2010.
- SILVER, N. *The Signal and the Noise*, New York: Penguin Press, 2012.
- VILALTA PERDOMO, C. J. ¿Se pueden predecir geográficamente los resultados electorales?: Una aplicación del análisis de clusters y outliers espaciales. *Estudios Demográficos y Urbanos*, v. 23, n. 3, sep./dic., p.571-613, 2008.

Data Sources

INSTITUTO NACIONAL ELECTORAL, *Estadísticas Lista Nominal y Padrón Electoral*
http://ine.mx/archivos3/portal/historico/contenido/Estadisticas_Lista_Nominal_y_Padron_Electoral/
 ———, *Memoria histórica sobre encuestas electorales*.
http://www.ine.mx/archivos2/portal/Elecciones/PEF/2014-2015/pdf/Memoria_Historica_2012-2015.pdf

MIGUEL ANGEL TORHTON
ESTIMATING THE SOURCES OF BIAS: A PREDICTIVE MODEL TO ANALYZE
POLLING PERFORMANCE FOR THE CONGRESSIONAL RACE IN MEXICO

———, *Sistema de Consulta de la Estadística de las Elecciones Federales 1991-2012*
<http://siceef.ife.org.mx/pef2012/SICEEF2012.html>