

# DETECTING ELECTION FRAUD WITH MACHINE LEARNING IN BOLIVIA

**AUTHOR: BENJAMIN JUAREZ**

**MENTOR: PROFESSOR ALVAREZ**

## Abstract

With the increased demand of ensuring the legitimacy of democratic elections, the prominence of election forensics has grown as it has been applied in several circumstances in recent years. Moreover, machine learning algorithms have been utilized in order to detect instances of various types of voter fraud, from ballot box stuffing to vote stealing. In particular, Bolivia's recent presidential election in 2019 turned controversial abruptly as late-counted votes gave the impression of fraud in favor of the incumbent, resulting in a period of political instability. However, recent studies of this election are suggesting that no such fraud had occurred. With this project, once the datasets for the official and preliminary counts of the election were acquired, further validation was necessary in order to become familiar with the data and prepare it for modeling. The next step involves synthetically generating the appropriate training dataset from a combination of historical data and demographic data through regression analysis. Using this synthetic data to train a supervised machine learning model ultimately allowed for the classification of *mesas* (voting booths) as clean or at-risk through the consideration of elements such as turnout and vote shares. With only about 84% of the *mesas* from this election labeled as clean, there arises some potential concern for the presence of fraud in certain locations.

## Introduction

Fair and just representation within the government is essential to the well-being of any democratic nation. Therefore, it is critical that the integrity of the processes that elect these representatives is ensured and protected. With that being said, especially amidst the controversies in our society today regarding voter fraud, there is a greater demand for guaranteeing free, legitimate elections. There are number of ways that election processes are monitored, whether it be through in-person methods or other forms observations. However, such methods may lack the efficiency to greatly detect and hinder fraud in such elections. Even with certain statistical methods and approaches that examine the election data in its distributions and patterns, there is an important aspect that is missing: the ability to consider patterns from election data from previous years. This would be known as a *priori* knowledge, where an approach would be capable of utilizing voter data from earlier elections in order to help paint a recognizable picture of what the data is supposed to look like [1].

With this need for a better mechanism for identifying election fraud, the use of machine learning models have arisen as a relatively new technique for analyzing data from elections. These models have advantages such that they are capable of utilizing data from previous elections along with being able to handle extensive datasets. With a wide range of models to choose from, these machine learning approaches have proved to be useful techniques for identifying and measuring the extent of anomalies that may lead to the existence of election fraud.

For this particular study, the Random Forest model has been utilized. This approach with a supervised machine learning ensemble method is used for the classification of voting booths (*mesas*) to indicate the possible presence of fraud. Specifically, the potential existence of vote stealing and ballot box stuffing is measured through this approach. With the Random Forest model trained using synthetically generated data, this study was able to label each *mesa* as either clean or at-risk. The results have indicated that there are certain areas that contain *mesas* with a much higher presence of potential manipulation or error, leading to their classifications as at-risk. Further explanation and details are discussed below.

## Historical Context

This study focuses on the 2019 Bolivian presidential election in particular. Occurring in October, this election turned controversial very quickly as late-counted votes initially appeared to show evidence of fraud which favored the incumbent Evo Morales [2]. To give some context, this election was reasonably contested by the incumbent, Evo Morales, who represents the Movement Toward Socialism (MAS) political party, and Carlos Mesa, who represents the Civic Community (CC) political party. The election was decided by popular vote, such that a candidate wins the presidency if are ahead by a 10-point margin once achieving at least 40% of the votes. Otherwise, there would be a runoff election to elect the winner [2].

That said, "with more than 80% of the vote counted, incumbent Evo Morales had a 7.9-point lead over the runner-up...But the following evening, with nearly all of the vote counted, Morale's margin narrowly exceeded 10 points" [2]. Morales was subsequently accused of fraud by his contestant, Mesa. The Organization of American States (OAS) supported this call for fraud with an announcement conveying "deep concern and surprise at the drastic and hard to explain change in the trend of the preliminary results revealed after the closing of the polls" [2].

These circumstances resulted in violence as political turmoil erupted. Eventually, related to his fraud allegations, Morales resigned and fled to Mexico. Currently, Jeanine Añez Chávez, an "opposition-party senator" is the interim president [2]. Handfuls of government officials were arrested and charged with "sedition and subversion", while up to 35 others were killed as a result of the political instability and violence [3].

However, recent studies of the data from this election are finding new conclusions. Further analysis of this data was able to explain the quantitative patterns that occurred without calling for fraud [2]. Even more recent news possibly reveals a programming error that may have led to the conclusions that fraud was present [4]. At this point in time, this new, potential error has yet to be fully confirmed and verified.

Taking all of this information in, this study has decided to provide another perspective on this election through quantitative analysis. By applying a supervised machine learning technique, Random Forest, to the election data, this study intends to locate voting booths (*mesas*) from this election which may be classified as at-risk of manipulation or error. With this opportunity to detect potential anomalies through a machine learning approach, there is an opportunity to present a supplemental viewpoint to this controversial election.

## Data

The data used in this study is the official data from the 2019 Bolivian presidential election. At the point when this study first began, there was only limited data publicly available. Specifically, the website for the Plurinational Electoral Organ (OEP) - which is the independent electoral branch of the Bolivian government - only lists the election data at the department level, rather than at the *mesa* (voting booth) level [5]. While useful, this data was not sufficient enough for the desired machine learning implementations. Luckily, Professor R. Michael Alvarez, was able to connect me with a colleague, Dorothy Kronick, who has greatly studied this case in Bolivia. Her recent paper, "Do shifts in late-counted votes signal fraud? Evidence from Bolivia", re-examined this election and seemingly disproves certain aspects of the fraud allegations through quantitative analysis [2]. With her generous help, the desired data was acquired.

Particularly, two datasets are utilized in this study. As a part of the verification process, there are two systems that collect the election data: the TREP (Transmission of Preliminary Electoral Results) data and the *cómputo* data. The TREP data represents the preliminary count, while the *cómputo* data designates the "legally binding official count" [2]. For the 2019 election, there were 5,296 precincts which contained 34,555 *mesas*. Both the TREP and *cómputo* data files are at the desired *mesa* level, organized by the information for each *mesa*. To explain, each row of data contains the information about the location of each *mesa* (country, department, municipality, precinct, etc.), along with the distribution of votes for each presidential candidate. The datasets also contains the number of eligible votes, and number of blank and null votes. There are also timestamps associated with each *mesa* included in both sets of data regarding the verification process. It is appropriate to mention that these datasets were labelled in Spanish, and thus translated accordingly. The proper translations and descriptions of the data parameters are included in the following data dictionary table. Further figures are displayed in order to provide the reader with a better glimpse into the election data. See replication materials in the appendix for the full datasets.

| Term                   | Translation                        | Description  |
|------------------------|------------------------------------|--|
| Pais                   | Country                            | Country in which mesa was located  |
| Departamento           | Department                         | Department that manages respective mesa  |
| Provincia              | Province                           | Province in which mesa was located   |
| Municipalidad          | Municipality                       | Local government in which mesa was located   |
| Localidad              | Locality                           | Locality in which mesa was located   |
| Recinto                | Precinct                           | Ballots were cast in mesas located in one of 5,296 precincts                       |
| Mesa                   | Voting booth                       | Ballots were cast at each of 34,555 voting booths                                  |
| Codigo de verificacion | Verification code                  | Each voting booth had a unique verification code assigned to it                    |
| Eleccionario Acta      | Election tally sheet               | Refers to what is being voted on (president/vice president, etc.)                  |
| Estado Acta            | Report tally sheet                 | Status of votes (e.g. computed)  |
| CC                     | Civic Community                    | Number of votes for political party of presidential candidate Carlos Mesa          |
| FPV                    | Front for Victory                  | Number of votes for political party of presidential candidate Israel Rodriguez     |
| MTS                    | Third System Movement              | Number of votes for political party of presidential candidate Felix Patzi          |
| UCS                    | Civic Solidarity Union             | Number of votes for political party of presidential candidate Victor Hugo Cardenas |
| MAS                    | Movement Toward Socialism          | Number of votes for political party of presidential candidate Carlos Mesa          |
| 21 F                   | Bolivia Alliance Says No           | Number of votes for political party of presidential candidate Oscar Ortiz Antelo   |
| PDC                    | Christian Democratic Party         | Number of votes for political party of presidential candidate Chi Hyun Chung       |
| MNR                    | Revolutionary Nationalist Movement | Number of votes for political party of presidential candidate Virginio Lema        |
| PAN-BOL                | National Action Party of Bolivia   | Number of votes for political party of presidential candidate Ruth Nina            |
| Validos en Acta        | Valid votes on tally sheet         | Sum of valid votes   |
| Blancos                | Blank ballots                      | Number of blank votes  |
| Nulos                  | Null ballots                       | Number of null votes   |
| Emitidos Reales        | Total votes                        | Sum of valid votes, blank votes, and null votes                                    |
| Inscritos              | Registered voters                  | Number of eligible voters  |

TABLE 1. This table translates and explains the parameters used in the election data .

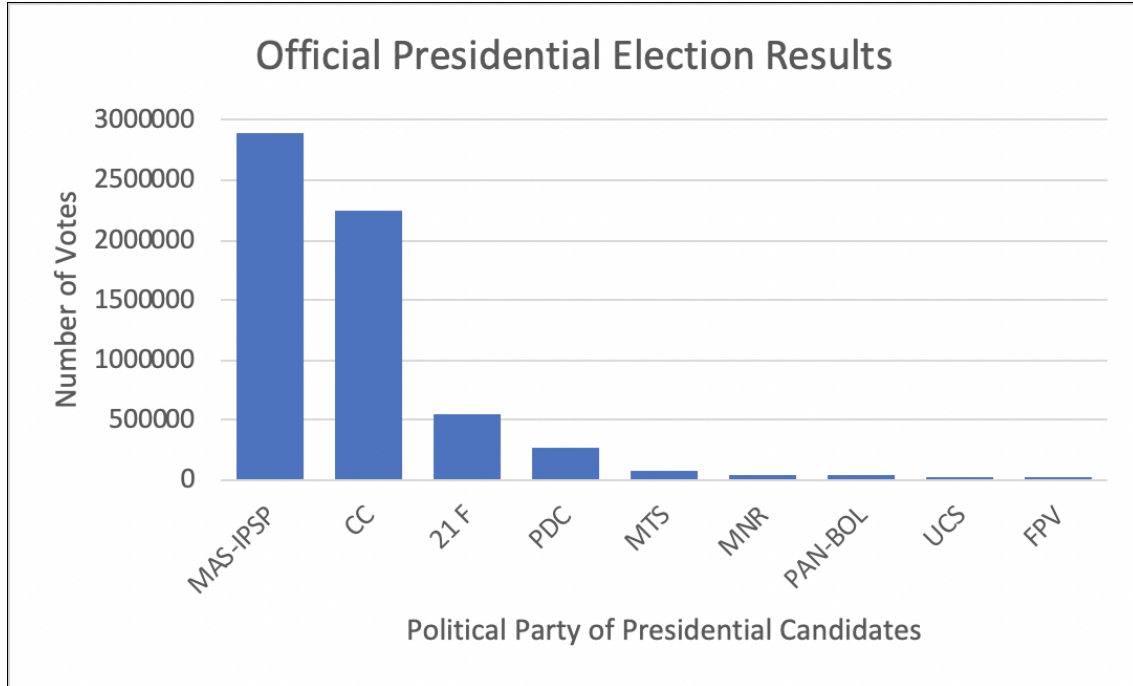
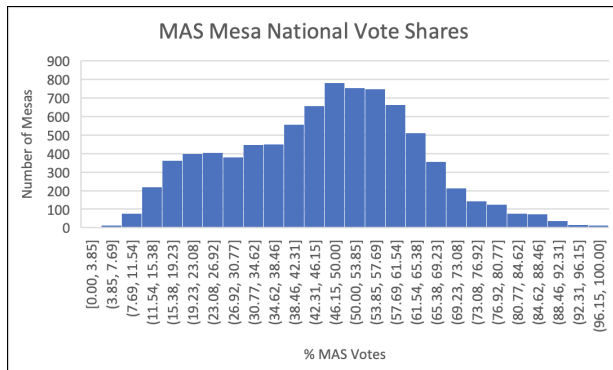
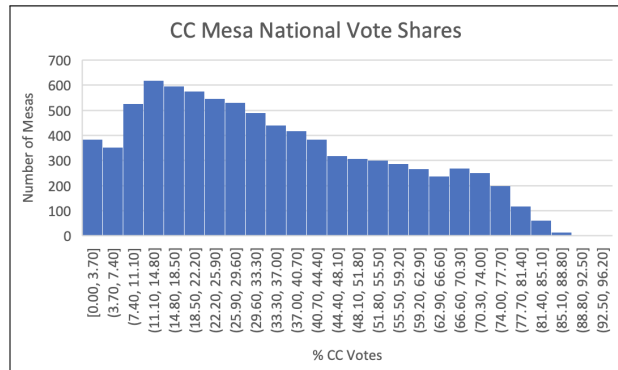


FIGURE 1. This figure shows the number of votes per party from the official count data set. The controversy over the election came from the two leading parties, CC and MAS, who are represented by Carlos Mesa and Evo Morales, respectively.



(A) Evo Morales - MAS



(B) Carlos Mesa - CC

FIGURE 2. These histograms show the frequency of votes within each *mesa* (voting booth) at the national level. These plots take into account the data from null and blank ballots.

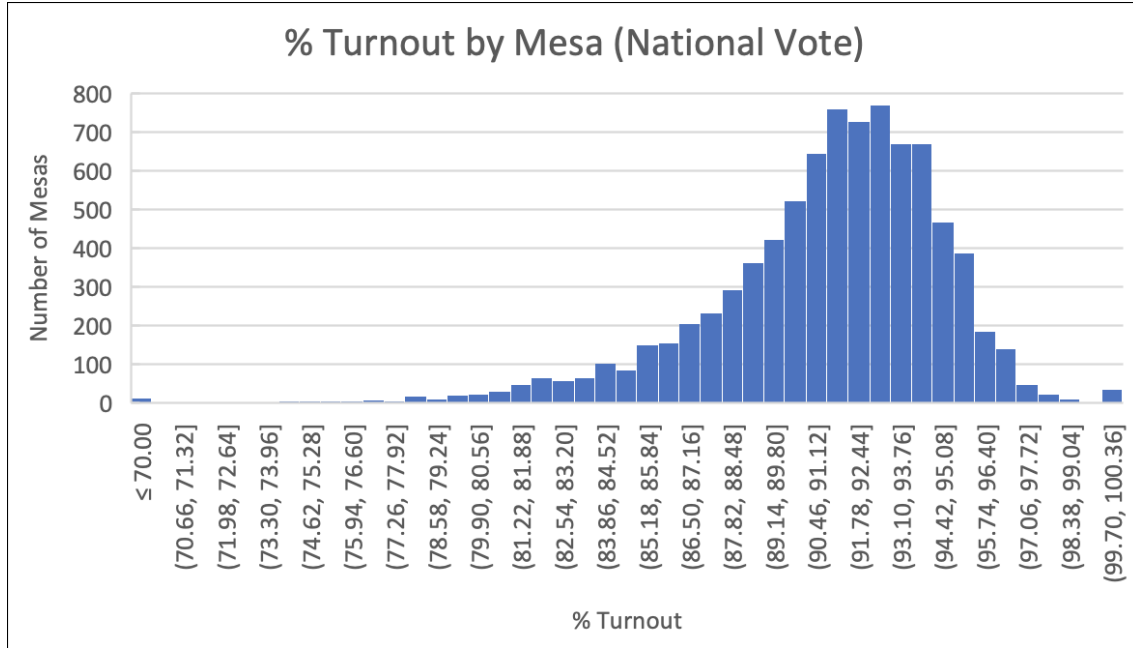
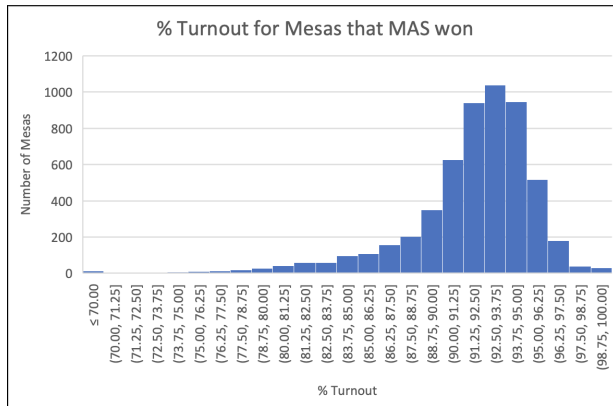
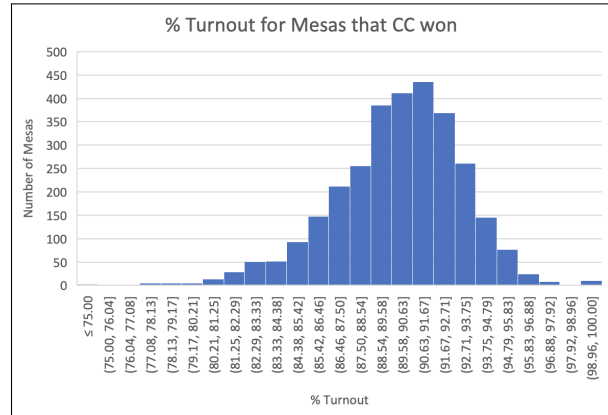


FIGURE 3. This histogram shows the turnout percentage for each *mesa* (voting booth) at the national level.



(A) In *mesas* where MAS had the majority vote.



(B) In *mesas* where CC had the majority vote.

FIGURE 4. These histograms show the turnout percentage for each *mesa* (voting booth) at the national level.

In addition to the 2019 election data, this study required data from the previous presidential election in 2014 in order to train the Random Forest model. Ideally, this training data would be at the *mesa* level. Unfortunately, the OEP website only has the departmental results listed which leads this study towards the need to generate synthetic data such that *mesa*-level data can be produced. These department results also had statistics at the *circunscripción* (electoral district) level. A glimpse at this departmental data can be seen in appendix A. The process of the synthetic data generation will be detailed below, but more predictive data is necessary for this process.

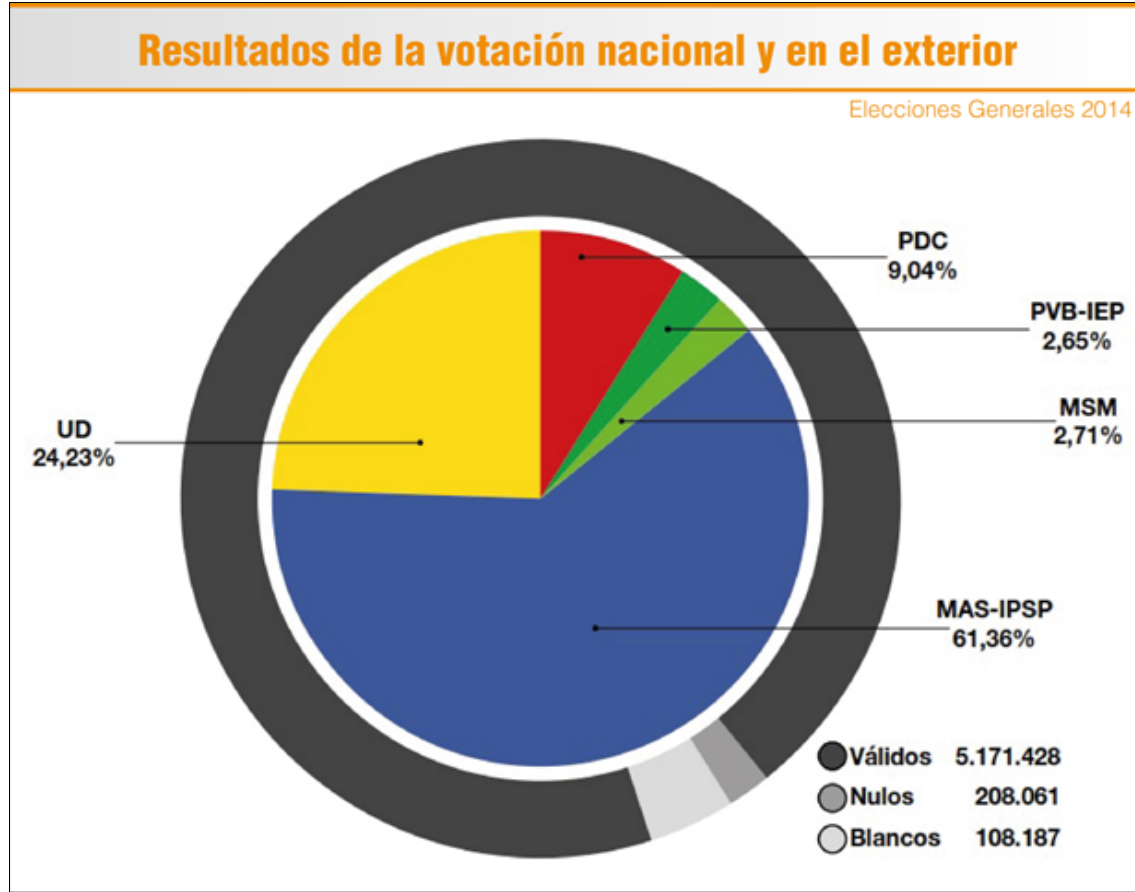


FIGURE 5. This pie chart from OEP website provides a glimpse at the total distribution of votes for the 2014 Bolivian presidential election, with Evo Morales (MAS-IPSP) holding the majority of the votes with 61.36% [5].

With that said, demographic information from the 2012 census and the National Institute of Statistics of Bolivia was collected at the department and municipality level [6]. This data included information regarding male/female populations, age distributions, literacy rates, and urbanization which are all useful factors for predicting elements of the *mesa* data. Details on the incorporation of this demographic data will be included below.

### Synthetic Data Generation

As mentioned above, the generation of synthetic data was necessary such that a proper training dataset can be constructed. The concept of the *priori* knowledge relates to this use of data prior to the 2019 election. This synthetic data, simulating trends from the 2014 presidential election, was first used to generate clean, untainted data at the *mesa*-level. The next step involved turning this clean data into manipulated data, such that mesas are categorized and labelled as either clean or at-risk of fraud. The types of fraud simulated in this study are vote stealing and ballot box stuffing. Furthermore, this labelled, manipulated dataset is what will be used as the training data for the supervised Random Forest classification model, which will be able to classify each *mesa* in the real data as either clean or at-risk. This synthetic data generation plan will now be described and illustrated as follows.

Taking the departmental 2014 election data and the demographic data associated with each department, the clean synthetic data was produced with a mixed effects regression model. Using the elements of the election data, integrated with components of the demographic data, the dependent variables of turnout

rates and vote shares for Morales (MAS) were predicted. By simulating these predictions a large number of times and plucking values for these variables from their distributions, the appropriate *mesa*-level dataset was constructed.

Now that the synthetic, clean data has been generated, it must be altered in a way such that a random portion of the *mesas* simulate either vote stealing or ballot box stuffing. Let vote stealing (labelled as VS) be defined as when a certain proportion of the votes of one or more parties is transferred to another party, which in this case would be MAS. On the other hand, let ballot box stuffing (labelled as BBS) be defined as when one party adds a certain proportion of votes to their totals, while not taking away votes from any other parties. These fraud simulations were performed by taking into account separate parameters for both types of fraud that define the probability of either vote stealing or ballot box stuffing occurring in a mesa and also define the extent to which the fraud occurred. Given these parameters, the necessary alterations to the turnout rate and MAS vote share were performed to properly simulate their respective fraud in the appropriate *mesas*. Furthermore, the *mesas* were labelled accordingly as either clean, at-risk of BBS, or at-risk of VS. Once these alterations were performed, the labelled manipulated dataset was produced. It is important to acknowledge that this procedure of altering and labelling the clean data follows a similar process performed by Professor R. Michael Alvarez and his colleagues [1]. See the figures below for more information on the distributions for turnout and MAS vote shares for the synthetic clean and at-risk data.

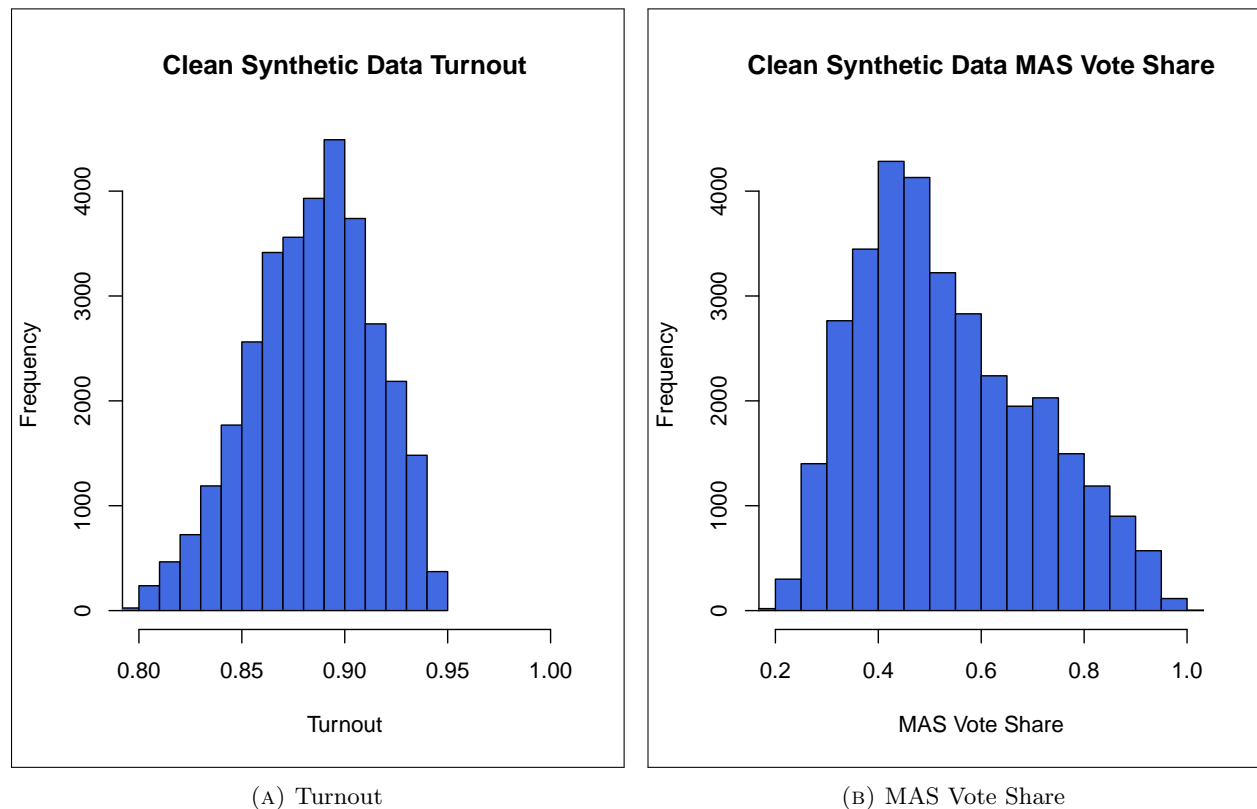


FIGURE 6. These histograms show the distribution of vote shares for Morales (MAS) and turnout for each *mesa* in the clean, synthetic dataset.

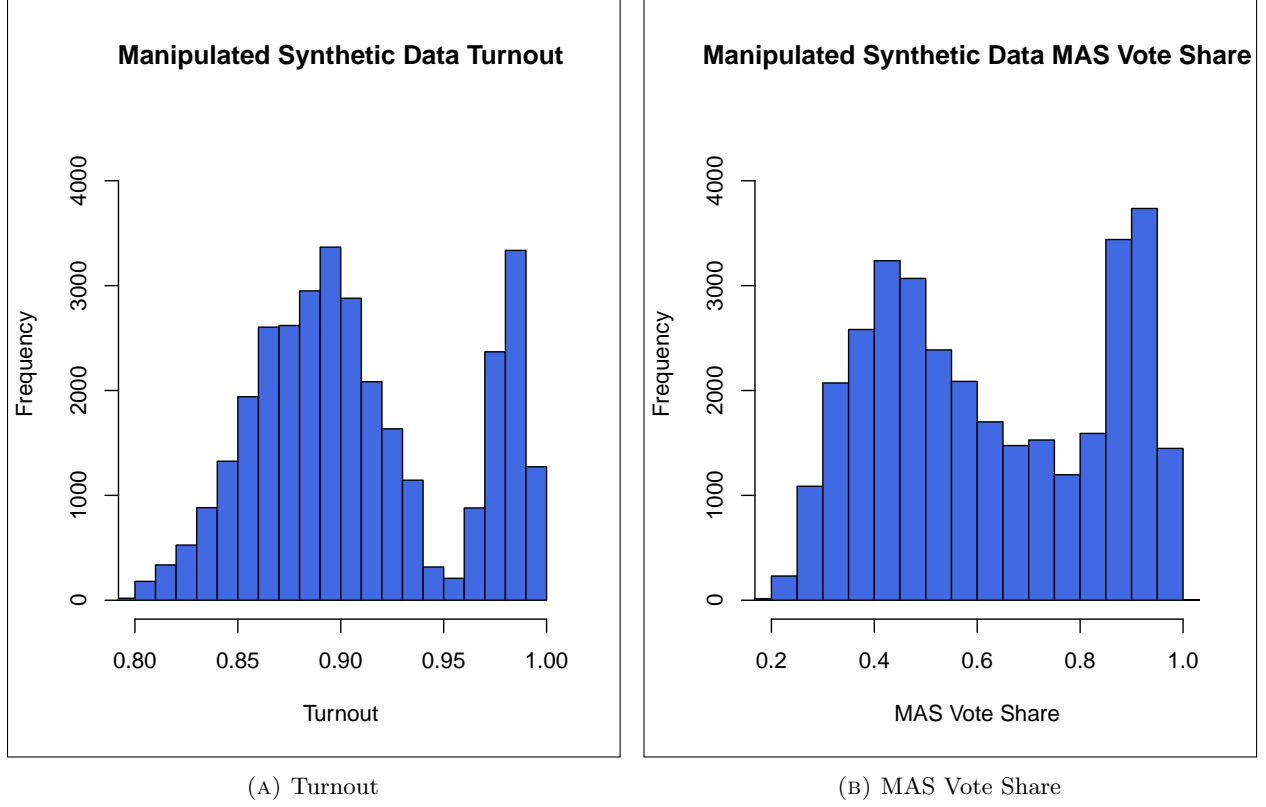
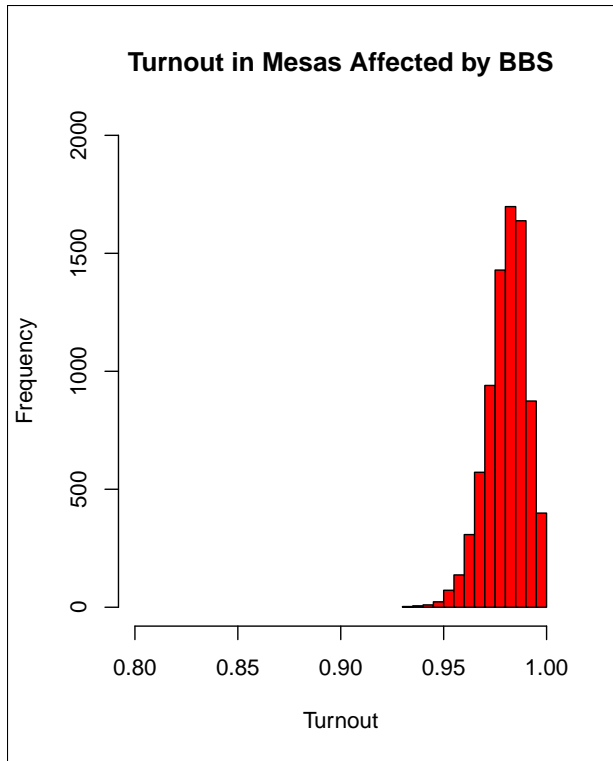


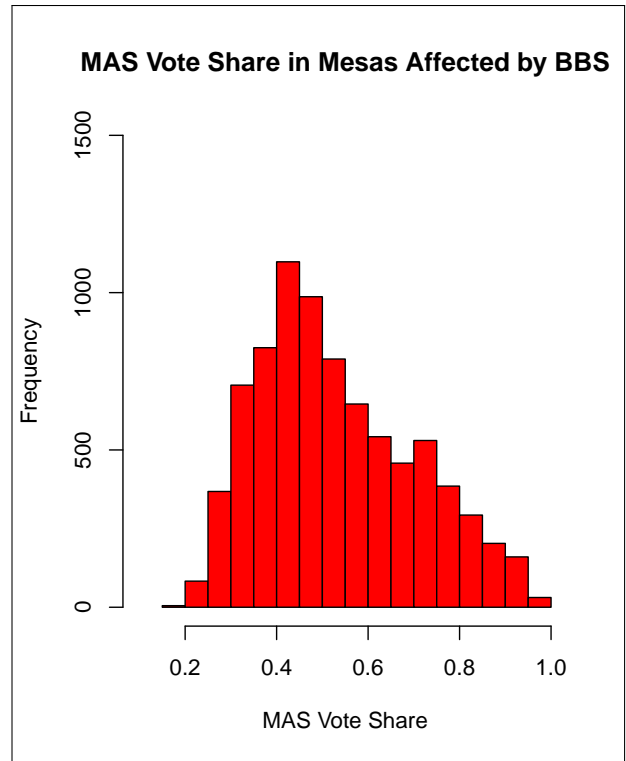
FIGURE 7. These histograms show the distribution of vote shares for Morales (Morales) and turnout for each *mesa* in the manipulated, synthetic dataset.

Following these histograms, there is a clear distinction between the distributions of the clean and manipulated synthetic data. These evident differences are attributed to the alterations from the two types of fraud. Examining subsets of the manipulated data allows for a deeper understanding of the effects of vote stealing and ballot box stuffing. By looking at the histograms below, which selectively take into account either *mesas* affected by VS or BBS, it is quite recognizable that BBS significantly increases turnout while VS significantly increases MAS vote shares. Analyzing and understanding these distributions within the manipulated, synthetic data played a key role in generating a model that is accurate in predicting the potential presence of fraud in the real data.

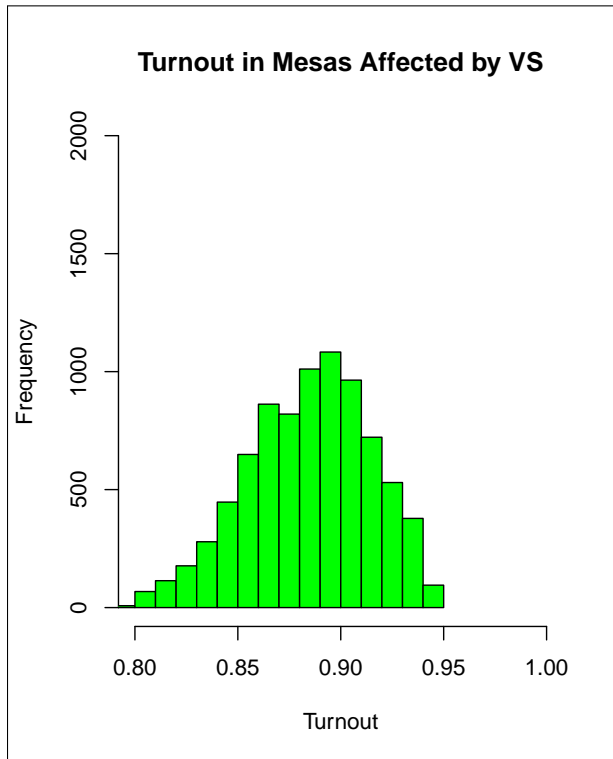




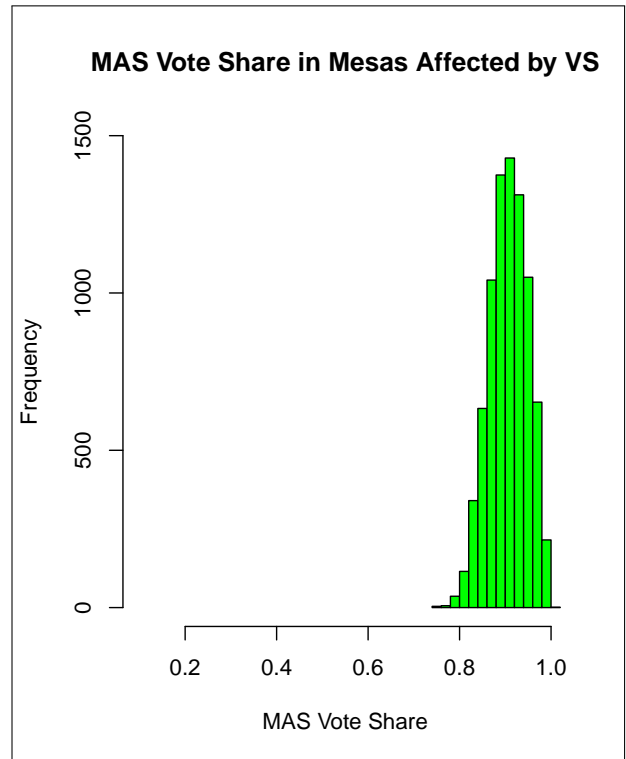
(A) BBS Turnout



(B) BBS MAS Vote Share



(C) VS Turnout



(D) VS MAS Vote Share

**FIGURE 8.** These histograms show the distribution of turnout and vote shares for Morales (MAS) for each *mesa* affected by vote stealing (VS) and ballot box stuffing (BBS) in the manipulated, synthetic dataset.

## Methods

### Microsoft Excel.

To elaborate on the methods utilized in this study, the first tool that was useful was Microsoft Excel. This software was suitable for the initial data analysis of the election datasets along with the demographic data. The graphing and visualization tools were responsible for some of the early summary tables and graphs, specifically Figures 1-4. These visuals were essential in the data validation step which helped produce a deeper understanding of the data while painting a picture of what the election looked like in regards to the distribution of votes and turnout.

Moreover, Excel was used for the manual data collection of the demographic data and department data. It was decided that the manual collection of this particular was more efficient than constructing a data scraper due to the relatively low volume of the data being collected. The demographic data was collected from a very useful interactive website: <https://www.citypopulation.de/en/bolivia/> [6]. On the other hand, the departmental election data was collected from pdf files from the OEP website [5]. See appendix A and C for more information. As mentioned, the demographic data was broken down at the department and municipality level, while the departmental data was organized at the department and *circunscripción* level. Fortunately, the OEP website provides the information that matches the municipalities to their respective *circunscripción* within each department which allowed for a fairly straightforward merging of the 2014 election and demographic data. An example of this information can be seen in appendix B. The synthetic election data was generated from this collected and formulated data.

### RStudio.

Further data analysis, synthetic data generation, and machine learning implementations were performed using the R programming language in the RStudio integrated development environment. R is commonly utilized in the field of data science, making it a great tool for analyzing, manipulating, and visualizing large datasets such as the ones used in this study. Through R's many accessible packages, including nlme, tidyverse, caret, and xtable, this study was able to conduct all the necessary operations for carrying out the election forensic analysis of this Bolivian election. In particular, the randomForest package allowed for the implementation of the supervised machine learning ensemble method, Random Forest, which is detailed below.

### Mixed Effects Modeling.

As previously mentioned, the synthetic training data was generated from the 2014 departmental data and demographic data through a mixed effects regression. This regression model was developed in R, using the nlme package which provides all the necessary tools for implementation. Specifically, two linear mixed effects regression models were generated which handled the predictions of turnout and MAS vote shares, key elements for each *mesa*. Both models had variables for literacy rates, male population, urbanization rates, age distributions, null votes, and blank votes which were used as fixed effects, while department was treated as a random effect. With these models developed, predictions for turnout and MAS vote share collected for each of the 63 *circunscripciones* within each department. By drawing the fitted, upper and lower interval values from the normal, sampling distribution simulated from the labelled parameters, the synthetic dataset could then be produced. This was done by simulating these predictions thousands of times and retrieving random values from these distributions at each simulation in order to represent *mesa*-level data. Once this was accomplished, the full synthetic, clean dataset had been constructed.

To further explain the process of altering the clean data into the manipulated, at-risk data, specific parameters determining the probability and extent of vote stealing and ballot box stuffing were selected. Following the similar procedure regarding the selection of these parameters from a paper by Professor R. Michael Alvarez, titled "Election forensics: Using machine learning and synthetic data for possible election anomaly detection", vote stealing and ballot box stuffing were randomly designated to certain *mesas* through multinomial distribution, assuming that these types of fraud were exclusive events [1]. From there, these designated, at-risk *mesas* were altered according to their assigned type of fraud and to the extent at which this fraud was to be applied. Again, these *mesas* were labelled with their appropriate type of fraud (VS or BBS).

### *Random Forest Modeling.*

The next step involved training a model with the labeled, manipulated dataset using Random Forest. By constructing decisions trees on subsets of the training data, Random Forest is used as a classification algorithm which categorizes each mesa as either clean or at risk of fraud (BBS or VS). This approach has been implemented in various other cases regarding the detection of election anomalies where it has proven its effectiveness, showing why it one of the most commonly used machine learning algorithms. The ultimate goal of testing the real election data with the trained model is to produce a better grasp on the results of this controversial election through the knowledge of how many *mesas* are at risk, where these *mesas* are located, and any conclusions that can be made about why they are at-risk.

Before testing the Random Forest model on the real election data, the accuracy of the model is determined by training the model on a portion of the labelled, training data and then testing the model on the rest of the training data with hopes that the model would accurately predict the type of fraud for each mesa. Since the correct labels for the training data are known, this is a good method for testing the performance of the model. For this study, the model was trained on a random portion of data, consisting of 90% of the training data, meaning that it was subsequently tested with the remaining 10% of the data. This process showed that the model correctly categorized about 98% of the remaining portion of the training data, indicating that it is very accurate at classifying the *mesas*. See the confusion matrix below for more information on the performance testing.

| Actual | Predictions |       |     |
|--------|-------------|-------|-----|
|        | BBS         | Clean | VS  |
| BBS    | 830         | 0     | 0   |
| Clean  | 0           | 1594  | 14  |
| VS     | 0           | 62    | 789 |

**TABLE 2. This confusion matrix portrays the performance of the Random Forest model in its accuracy in classifying the last 10% of the training data, after being trained on other 90% of the data. This table is displayed such that a small portion of *mesas* that were originally labeled as clean or VS were misclassified, while all *mesas* affected by BBS were correctly classified.**

With the knowledge that the model is very accurate with turnout and MAS vote shares as predictors, it was time to move onto the next step of training the model on the full training dataset, followed by the testing of the model with the real data to discover any possibly anomalies. R allows for the measurement of the importance of each predictor in the Random Forest model. As shown by the plot below, turnout was the most important predictor, followed closely by MAS vote shares. At this point, this model was implemented with the 2019 real, election data.

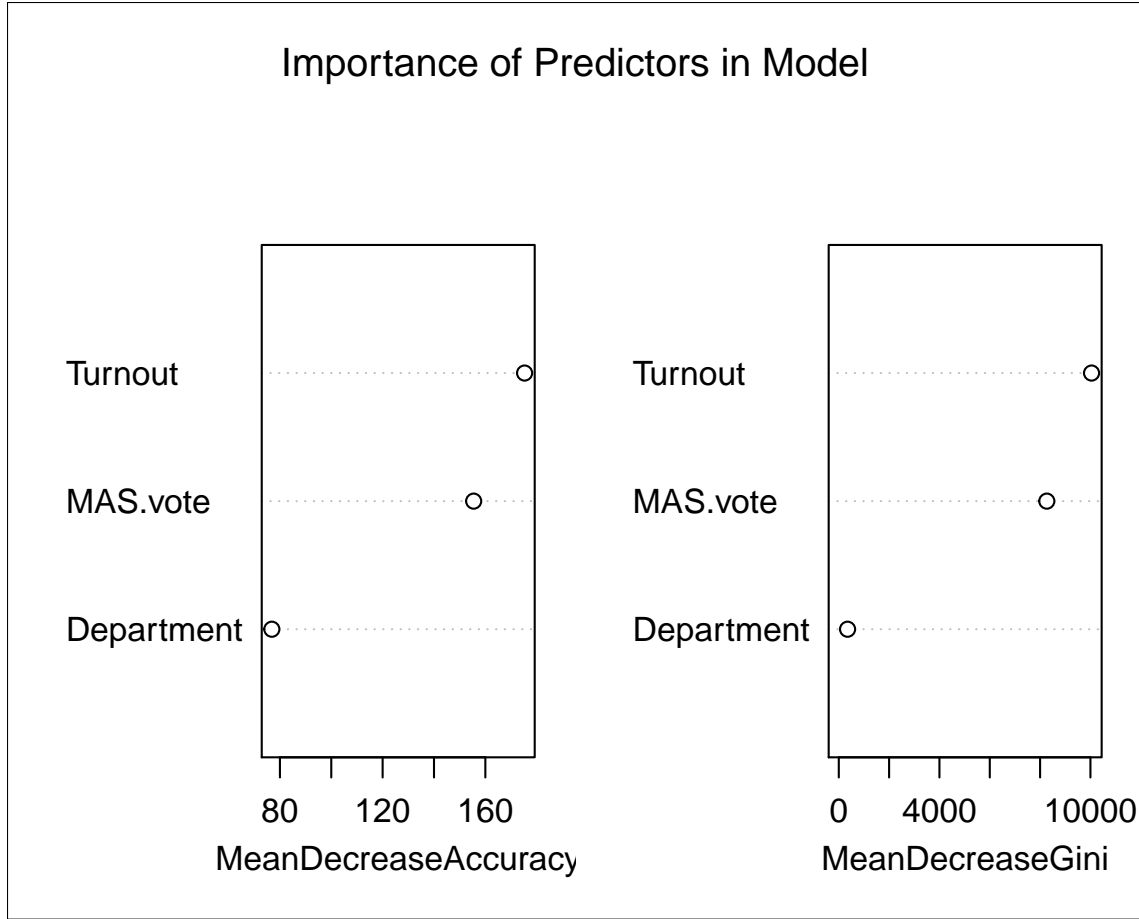


FIGURE 9. This figure demonstrates the importance of turnout, MAS vote share, and department as predictors in the Random Forest model. Importance is measured by the "average decrease in the Gini index across decision trees upon permutation of the values of each predictive feature" [1].

## Results

Applying the real data with the trained model successfully allowed for the prediction and classification of potential fraud risk for each real *mesa*. Under the assumptions that selected parameters, the results were that about 84% of the *mesas* were predicted to be clean, leaving the other 16% to be classified as at-risk (with about 11% being at-risk of BBS and about 5% being at-risk of VS). See the table below for more details.

|           | Clean | At.Risk | BBS.Risk | VS.Risk | Avg.Turnout | Avg.MAS.vote |
|-----------|-------|---------|----------|---------|-------------|--------------|
| All Mesas | 84.30 | 15.70   | 11.10    | 4.60    | 0.90        | 0.46         |

TABLE 3. This table displays the predictions from the Random Forest model.

A more detailed look within each department allows for the determination of departments that are potentially at a higher risk of fraud. For each department, the percentage of clean and at-risk (BBS or VS) *mesas* along with the turnout and MAS vote share data is portrayed. From this data, the departments of Cochabamba, La Paz, and Potosí are relatively the most at-risk.

| Department | Clean | At.Risk | BBS.Risk | VS.Risk | Avg.Turnout | Avg.MAS.vote |
|------------|-------|---------|----------|---------|-------------|--------------|
| Cochabamba | 78.20 | 21.80   | 10.10    | 11.70   | 0.90        | 0.57         |
| La Paz     | 82.50 | 17.50   | 14.40    | 3.00    | 0.92        | 0.53         |
| Potosí     | 83.00 | 17.00   | 7.90     | 9.10    | 0.89        | 0.46         |
| Santa Cruz | 86.20 | 13.80   | 12.20    | 1.60    | 0.89        | 0.34         |
| Oruro      | 88.50 | 11.50   | 9.40     | 2.10    | 0.91        | 0.47         |
| Tarija     | 90.30 | 9.70    | 6.80     | 2.90    | 0.87        | 0.39         |
| Chuquisaca | 92.00 | 8.00    | 2.30     | 5.80    | 0.88        | 0.40         |
| Pando      | 95.80 | 4.20    | 2.70     | 1.50    | 0.84        | 0.43         |
| Beni       | 96.80 | 3.20    | 2.00     | 1.20    | 0.86        | 0.34         |

TABLE 4. This table displays the predictions from the Random Forest model for each department.

Looking at this data, it can be seen that most of the departments have mostly clean *mesas*. From the table above, there is a clear trend that a higher average turnout and MAS vote share relates to departments with a higher percentage of at-risk *mesas*. Specifically, the *mesas* the Cochabamba department particularly has a much higher rate of being classified as at-risk of vote stealing.

The departments of Cochabamba, La Paz, and Potosí were further examined in order to possibly discover some reasoning for their higher rates of potential fraud. The table below portrays the most at-risk municipalities from these departments, specifically identifying municipalities with at least 25 *mesas* that have more than 20% of these *mesas* classified as at-risk.

| Municipality      | Clean | At.Risk | BBS.Risk | VS.Risk | Avg.Turnout | Avg.MAS.vote | Mesas |
|-------------------|-------|---------|----------|---------|-------------|--------------|-------|
| Entre Ríos        | 2.80  | 97.20   | 33.30    | 63.90   | 0.93        | 0.91         | 108   |
| Villa Tunari      | 8.10  | 91.90   | 40.10    | 51.80   | 0.94        | 0.94         | 197   |
| Puerto Villarroel | 8.60  | 91.40   | 36.40    | 55.00   | 0.93        | 0.91         | 151   |
| Tapacarí          | 11.90 | 88.10   | 19.00    | 69.00   | 0.91        | 0.93         | 42    |
| Palca             | 14.70 | 85.30   | 8.80     | 76.50   | 0.92        | 0.87         | 34    |
| Totora            | 18.20 | 81.80   | 21.20    | 60.60   | 0.92        | 0.89         | 33    |
| Cocapata          | 24.10 | 75.90   | 20.70    | 55.20   | 0.90        | 0.86         | 29    |
| Sacaca            | 27.30 | 72.70   | 45.50    | 27.30   | 0.94        | 0.85         | 33    |
| Laja              | 34.00 | 66.00   | 60.00    | 6.00    | 0.95        | 0.78         | 50    |
| Sapahaqui         | 35.70 | 64.30   | 14.30    | 50.00   | 0.92        | 0.84         | 28    |
| Pucarani          | 38.40 | 61.60   | 56.20    | 5.50    | 0.94        | 0.71         | 73    |
| Tiraque           | 38.50 | 61.50   | 7.70     | 53.80   | 0.91        | 0.88         | 52    |
| Colquechaca       | 40.00 | 60.00   | 10.00    | 50.00   | 0.89        | 0.79         | 40    |
| Pocoata           | 45.70 | 54.30   | 20.00    | 34.30   | 0.89        | 0.81         | 35    |
| San Pedro         | 46.50 | 53.50   | 4.70     | 48.80   | 0.86        | 0.84         | 43    |
| Chimoré           | 47.70 | 52.30   | 27.70    | 24.60   | 0.92        | 0.86         | 65    |
| Mecapaca          | 52.50 | 47.50   | 30.00    | 17.50   | 0.93        | 0.77         | 40    |
| Calamarca         | 52.90 | 47.10   | 38.20    | 8.80    | 0.94        | 0.73         | 34    |
| Shinahota         | 53.10 | 46.90   | 31.20    | 15.60   | 0.91        | 0.86         | 64    |
| Chayanta          | 53.30 | 46.70   | 26.70    | 20.00   | 0.92        | 0.76         | 30    |
| Arbieto           | 59.50 | 40.50   | 8.10     | 32.40   | 0.87        | 0.82         | 37    |
| Colquiri          | 59.50 | 40.50   | 32.40    | 8.10    | 0.93        | 0.68         | 37    |
| Sicasica          | 60.00 | 40.00   | 14.00    | 26.00   | 0.93        | 0.77         | 50    |
| Ayopaya           | 61.40 | 38.60   | 13.60    | 25.00   | 0.88        | 0.81         | 44    |
| Colomi            | 64.30 | 35.70   | 7.10     | 28.60   | 0.91        | 0.79         | 56    |
| Tiahuanacu        | 64.30 | 35.70   | 14.30    | 21.40   | 0.93        | 0.76         | 28    |
| Aiquile           | 67.90 | 32.10   | 3.60     | 28.60   | 0.88        | 0.75         | 56    |
| Mizque            | 68.10 | 31.90   | 6.40     | 25.50   | 0.89        | 0.77         | 47    |
| Achocalla         | 72.90 | 27.10   | 25.70    | 1.40    | 0.93        | 0.71         | 70    |
| Betanzos          | 73.60 | 26.40   | 1.90     | 24.50   | 0.86        | 0.73         | 53    |
| Patacamaya        | 74.50 | 25.50   | 12.70    | 12.70   | 0.90        | 0.68         | 55    |
| Batallas          | 75.60 | 24.40   | 20.00    | 4.40    | 0.93        | 0.68         | 45    |
| Capinota          | 77.80 | 22.20   | 14.80    | 7.40    | 0.91        | 0.70         | 54    |
| Sipesipe          | 77.80 | 22.20   | 14.60    | 7.60    | 0.91        | 0.71         | 144   |
| Coripata          | 78.00 | 22.00   | 22.00    | 0.00    | 0.92        | 0.46         | 50    |
| El Alto           | 78.10 | 21.90   | 21.90    | 0.00    | 0.93        | 0.55         | 3022  |
| Copacabana        | 78.40 | 21.60   | 8.10     | 13.50   | 0.92        | 0.74         | 37    |
| Irupana           | 78.60 | 21.40   | 19.00    | 2.40    | 0.91        | 0.49         | 42    |

TABLE 5. This table displays the predictions from most at-risk municipalities from the departments of Cochabamba, La Paz, and Potosí.

The displayed municipality-level data is organized by the percentage of clean *mesas* within their respective municipality. Clearly, there is a large difference the most, at-risk municipality (Entre Ríos) and the least, at-risk municipality (Irupana). There exist of variety of explanations for these stark differences between the various parameters portrayed above. Taking a deeper look into the demographic information related to these municipalities helped to generate a better understanding of the results. The following table incorporates demographic information such as urbanization and literacy rates, along with the number of *mesas* within each municipality. It is important to note that the demographic data is same dataset used for the synthetic data, meaning that it is data from a few years prior to this 2019 election. These demographics should still be relatively accurate in depicting the certain trends identified.

| Municipality      | Department | Rural | Literate | Mesas |
|-------------------|------------|-------|----------|-------|
| Entre Ríos        | Cochabamba | 0.56  | 0.95     | 108   |
| Villa Tunari      | Cochabamba | 0.88  | 0.94     | 197   |
| Puerto Villarroel | Cochabamba | 0.72  | 0.94     | 151   |
| Tapacarí          | Cochabamba | 1.00  | 0.82     | 42    |
| Palca             | La Paz     | 1.00  | 0.91     | 34    |
| Totora            | Cochabamba | 1.00  | 0.88     | 33    |
| Cocapata          | Cochabamba | 1.00  | 0.90     | 29    |
| Sacaca            | Potosí     | 0.88  | 0.85     | 33    |
| Laja              | La Paz     | 1.00  | 0.91     | 50    |
| Sapahaqui         | La Paz     | 1.00  | 0.91     | 28    |
| Pucarani          | La Paz     | 1.78  | 0.90     | 73    |
| Tiraque           | Cochabamba | 0.89  | 0.89     | 52    |
| Colquechaca       | Potosí     | 0.88  | 0.75     | 40    |
| Pocoata           | Potosí     | 1.00  | 0.82     | 35    |
| San Pedro         | Potosí     | 1.00  | 0.82     | 43    |
| Chimoré           | Cochabamba | 0.71  | 0.96     | 65    |
| Mecapaca          | La Paz     | 1.00  | 0.94     | 40    |
| Calamarca         | La Paz     | 1.00  | 0.94     | 34    |
| Shinahota         | Cochabamba | 0.73  | 0.94     | 64    |
| Chayanta          | Potosí     | 0.85  | 0.85     | 30    |
| Arbieto           | Cochabamba | 0.79  | 0.94     | 37    |
| Colquiri          | La Paz     | 0.70  | 0.92     | 37    |
| Sicasica          | La Paz     | 0.65  | 0.94     | 50    |
| Ayopaya           | Cochabamba | 1.00  | 0.84     | 44    |
| Colomi            | Cochabamba | 0.80  | 0.89     | 56    |
| Tiahuanacu        | La Paz     | 0.83  | 0.90     | 28    |
| Aiquile           | Cochabamba | 0.66  | 0.88     | 56    |
| Mizque            | Cochabamba | 0.87  | 0.84     | 47    |
| Achocalla         | La Paz     | 0.16  | 0.95     | 70    |
| Betanzos          | Potosí     | 0.86  | 0.83     | 53    |
| Patacamaya        | La Paz     | 0.51  | 0.94     | 55    |
| Batallas          | La Paz     | 0.87  | 0.89     | 45    |
| Capinota          | Cochabamba | 0.53  | 0.90     | 54    |
| Sipesipe          | Cochabamba | 0.66  | 0.93     | 144   |
| Coripata          | La Paz     | 1.00  | 0.96     | 50    |
| El Alto           | La Paz     | 0.00  | 0.97     | 3022  |
| Copacabana        | La Paz     | 0.62  | 0.89     | 37    |
| Irupana           | La Paz     | 1.00  | 0.94     | 42    |

TABLE 6. This table portrays the department, demographic info, and number of *mesas* related to the at-risk municipalities. The demographic data is from the 2012 census.

With the consideration of the data above, there are some possible conclusions that can be determined. First off, the majority of these at-risk municipalities come from the departments of Cochabamba and La Paz, while the most at-risk municipalities are located in the Cochabamba department. With the exception of El Alto (a municipality which encases a dense urban population), these at-risk municipalities appear to be mainly consist of rural populations. With some basic contextual knowledge of the Bolivian election, it is known that lower-class, rural populations were more likely to vote for Evo Morales (MAS).

With that being said, it is important to remember that just because the model classified *mesas* in these particular municipalities as at-risk, it does not directly imply that there is a definite existence in fraud in these *mesas*. Rather, it would be more accurate to say that the data from these *mesas* suggests that there

exist potential anomalies in the distributions, which certainly could be as a result of either vote stealing or ballot-box stuffing. There are other factors which can contribute to a municipality having election data that does not follow the trends seen in the synthetic, training data. The classification of either clean or at-risk done by the Random Forest model is based upon the previously mentioned *prior* knowledge from the incorporation of the 2014 data and the demographic data. So, again, there are a variety of factors that would have occurred before the 2019 election that would affect the trends seen in the synthetically generated data.

Another meaningful point to acknowledge is that procedure for producing the synthetic data may have lacked some accuracy in predicting the full details of the 2014 presidential election. As mentioned, the only real data from the 2014 election was at the department level, leaving some room for possible inaccuracies in generating data at the *mesa*-level.

To emphasize, the Random Forest model predicted that a large majority of the *mesas* in the 2019 Bolivian presidential election were clean of possible vote stealing and ballot box stuffing. The remaining portion of *mesas* that was classified as at-risk of these types of fraud simply shows that the data potentially indicates the existence of anomalies which could be due to fraud or some other form of manipulation or error.

## Conclusions

In an ideal scenario when using a supervised machine learning approach such as Random Forest, the training data is readily available and labeled properly. However, this circumstance is usually not the case with studies such as the one detailed in this paper. In these cases, the use of synthetically generated training data is vital to the construction of an accurate and functional machine learning model. Following a culmination of similar procedures used in election forensics, this study utilizes both demographic and election data in combination with concepts of various types of election fraud in order to produce the labeled, synthetic data that was used to train the Random Forest model. This ability to fabricate datasets like these which simulate some event under analysis (which in this case was the 2019 Bolivian presidential election) will continue to be an essential component in research, particularly in the field of election forensics.

The results from this study succeed in adding another perspective to an already multi-layered, controversial historical event. As detailed earlier, this presidential election and its aftermath had significant effects on Bolivia, both socially and politically. With this analysis of this election as a whole, it is difficult to say whether or not the conclusions contradict any of the recent findings on this election, since this study focuses on the data quantitatively without the consideration of timestamps and does not make any definitive claims against any groups involved. However, these results indeed bring about some cause for concern. The predictions from the Random Forest model do not classify all of the *mesas* as clean from fraud, so it would prove useful to test this data with other machine learning approaches in order to come to a stronger conclusion regarding the potential fraud in certain locations. Further work on this project would allow for the implementation of other methods along with a deeper analysis of these at-risk locations. Overall, this study has proven to be another useful application of a machine learning approach in the context of election forensics.

Furthermore, this analysis of the 2019 Bolivian presidential election deepens the knowledge of this event which is continuing to be strongly examined as more information is being released during the time frame in which this paper is being written. The methods used in this analysis will continue to play an important role in ensuring the integrity of elections. Given the current circumstances in many countries, it appears that these election forensic tools will be more meaningful than ever.

In regards to other work done by Professor Alvarez and the rest of the Caltech/MIT Voting Technology Project, the techniques and methods used in this study, such as Random Forest, have already been applied in multiple circumstances in countries such as Venezuela, Argentina, etc. These previous studies provided a strong foundation to the work conducted in this research project. The future possibilities with election forensic research and applications are vast and growing in importance, and this study hopes to prove to be another significant utilization of these computational social science methods. These techniques can and will be applied in many other scenarios in the context of elections and many other situations involving large sets of data. Beyond the implications of election forensics, machine learning will continue to be a growing field with great importance.



## References


- [1] Alvarez, Levin, & Zhang. *Election forensics: Using machine learning and synthetic data for possible election anomaly detection*. 2019.
- [2] Idrobo, Kronick, & Rodriquez. *Do shifts in late-counted votes signal fraud? Evidence from Bolivia*. 2020.
- [3] Monkey Cage. *Bolivia dismissed its October elections as fraudulent. Our research found no reason to suspect fraud..* 2020.  
<https://www.washingtonpost.com>
- [4] Beeton. *Major Coding Error Reveals Another Fatal Flaw in OAS Analysis of Bolivia's 2019 Elections*. 2020.  
<https://cepr.net>
- [5] Plurinational Electoral Organ. *Electoral Processes*. 2020.  
<https://www.oep.org.bo>
- [6] City Population. *Bolivia*. 2020.  
<https://www.citypopulation.de/en/bolivia/>

## Acknowledgements

I, Benjamin Juarez, graciously thank Professor R. Michael Alvarez for all of his time and work involved with this project. By sharing a substantial amount of knowledge through both his recent papers and his own advice through discussion on a weekly basis, Professor Alvarez has proved to be an exceptional mentor in my project. His guidance truly allowed for a very enjoyable, enriching experience with this project through all of its obstacles and learning curves.

I would also like to thank Dorothy Kronick for not only her very insightful contributions through her paper on this particular election which helped to give me a deeper understanding of this event, but also for her generous help with providing the necessary datasets which made this project possible. She also provided further insight on certain aspects of the data further into the project. Without her help, this project would not have been nearly as productive, so with that being said, I fully thank her for all of her contributions

Furthermore, I would like to thank the John Randolph Haynes and Dora Haynes Foundation for their generous support through their research grant to Professor R. Michael Alvarez.



**4. RESULTADOS DEPARTAMENTALES**

**4.1.- DEPARTAMENTO DE CHUQUISACA**

| DETALLE  | TOTAL   | Porcentaje |
|--|---------|------------|
| Ciudadanos (as) Habilitados (as)   | 323.129 |            |
| Votos emitidos   | 284.218 | 100,00%    |
| Votos válidos  | 261.562 | 92,03%     |
| Votos nulos  | 13.348  | 4,70%      |
| Votos en blanco  | 9.308   | 3,27%      |
| <b>VOTOS VALIDOS OBTENIDOS POR CADA PARTIDO POLITICO, AGRUPACION CIUDADANA O ALIANZA</b> |         |            |
| Partido Demócrata Cristiano (PDC)  | 44.671  | 17,08%     |
| Partido Verde de Bolivia-Instrumento de la Ecología Política (PVB-IEP)                   | 4.438   | 1,70%      |
| Movimiento Sin Miedo (MSM)   | 4.405   | 1,68%      |
| Movimiento al Socialismo-Instrumento Político por la Soberanía de los Pueblos (MAS-IPSP) | 165.785 | 63,38%     |
| Unidad Demócrata (UD)  | 42.263  | 16,16%     |

**Resultados por circunscripción uninominal**

**CIRCUNSCRIPCIÓN N° 1**

| DETALLE  | TOTAL  | Porcentaje |
|--|--------|------------|
| Ciudadanos (as) Habilitados (as)   | 96.222 |            |
| Votos emitidos   | 85.987 | 100,00%    |
| Votos válidos  | 73.535 | 85,52%     |
| Votos nulos  | 3.030  | 3,52%      |
| Votos en blanco  | 9.422  | 10,96%     |
| <b>VOTOS VALIDOS OBTENIDOS POR CADA PARTIDO POLITICO, AGRUPACION CIUDADANA O ALIANZA</b> |        |            |
| Partido Demócrata Cristiano (PDC)  | 30.570 | 41,57%     |
| Partido Verde de Bolivia-Instrumento de la Ecología Política (PVB-IEP)                   | 1.436  | 1,95%      |
| Movimiento Sin Miedo (MSM)   | 2.087  | 2,84%      |
| Movimiento al Socialismo-Instrumento Político por la Soberanía de los Pueblos (MAS-IPSP) | 27.770 | 37,76%     |
| Unidad Demócrata (UD)  | 11.672 | 15,87%     |

FIGURE 10. This screenshot shows a page from the pdf (from the OEP website) containing the departmental results from the 2014 Bolivian presidential election. This data was used along with the appropriate demographic information in order to produce the synthetic, training data [5].

## APPENDIX B.

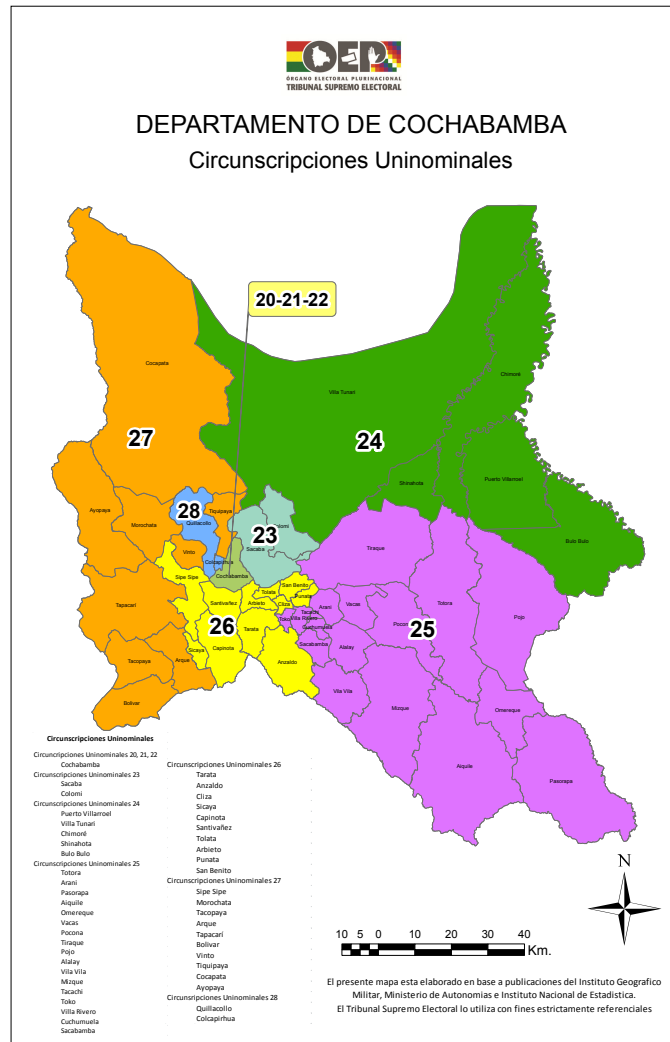


FIGURE 11. This map from OEP website provides a view of the region containing the department of Cochabamba. This portrayal further illustrates the *circunscripciones* located within the department, along with the respective municipalities which reside within each *circunscripción*. The OEP has a map similar to this one for each department. This information was used to match the municipalities to the *circunscripciones* for each department [5].

## APPENDIX C.

Link to github repository for replication materials: <https://github.com/benjuarez8/Bolivia-SURF-2020.git>