Clustering Approach to Detect Fabricated Answers in a Dominican Household Survey

Ian Paulino
Graduate School of Arts and Sciences
Fordham University, U.S.A.
ipaulinovelez@fordham.edu

***Abstract*—Household survey data integrity is of increasing importance for the Dominican Republic and its reported indicators, research, and public policy. A wide range of mechanisms exist to verify that survey answers have not been fabricated by the respondent or interviewer. We evaluate three unsupervised learning methods to detect these falsified answers in a simulated environment using data from the ENGIH 2018. The models were tested using purposefully altered data to verify that the classifications matched the clusters produced. Although GMM had less than optimal results, both one-class SVM and IForest proved to produce correct segmentation of fabricated answers from real data in most tested examples.**

Keywords—clustering, unsupervised learning, fabricated answers, household survey, Dominican Republic

I. Introduction

Fabrication of survey answers can have an important impact on inference, research and reporting using that data. Even very small amounts of falsified data can lead to relatively high bias in survey results, as shown by various simulations [1], [2], especially for multivariate analyses [3]. For a survey as widely used as the National Survey of Income and Expenditure of Households of the Dominican Republic (Encuesta Nacional de Gastos e Ingresos de los Hogares or ENGIH), the problem becomes even more relevant. This survey has been used in extensive economic and social research, including but not limited to topics like consumer basket analysis, income inequality, monetary poverty, educational costs, and healthcare economics and out-of-pocket expenditure estimation [4], [5], [6], [7].

This work sets out to use clustering techniques to identify fabrication survey response from altered data. Three different models are tested: the Gaussian Mixture Model, One-class SVM and Isolation Forest. During the next sections, we will aim to fulfill the following objectives:

- Explore the nuances surrounding the ENGIH 2018, the nature of its data and its susceptibility to error.
- Justify our specific approach using existing literature on survey control mechanisms.
- Review the models and implement them in different scenarios to test their efficacy in detecting falsified data.

II. Background

The ENGIH is structured in a way that makes it vulnerable to fabricated data. With the 2018 version being the fifth of its kind since 1976, it is divided into 6 extensive questionnaires which are applied to the population sample in the span of 10 days. Because of its longitudinal nature and its propensity to ask health and income-related questions, respondents might be more reluctant to answer [8], and interviewers are more prone to the tiredness, disinterest, or conflictive incentives that often provokes curbstoning or interviewer-side falsification. Education is also linked to higher survey respondent-side error [9], which is especially relevant in the context of the Dominican Republic general education indicators, such as its below average ranking in PISA 2018 [10].

The issue of transparency in the survey aggravates the matter, as it struggles to show which missing values from item non-response have been imputed or how it has been done so, especially for categorical variables which are abundant in the questionnaire we will be focusing on in this experiment. This issue is explored in [11], where the author stated real respondents are more likely to have item non-response as falsifiers will often try to fill in most if not all the questions carelessly. This analysis is obstructed for us by the current model of the ENGIH 2018 report.

As discussed, fabricated data may come from the interviewers' side or from the side of the respondent. As will be touched upon starting in the methodology section, our experiment will be undoubtedly more aligned to the randomness that is more common from the respondent [12]. It is not uncommon, however, for population surveys of this kind to see curbstoning ratios of between 0.4-6% for United States surveys and even higher rates for less developed countries [13].

The exact amount of fabricated data in surveys remains uncertain, with estimates ranging from less than 1% to over 50%. While most researchers place the share within a 1-7% range [14], [15], others suggest it could be significantly higher depending on specific survey characteristics and oversight. Specifically, smaller surveys tend to have much higher fabrication rates than large-scale population surveys [16].

III. Related work

Detection of fabricated surveys has been explored in many different settings. The approaches to detecting these falsifications can be summarized into two approaches: external control mechanisms and internal control mechanisms [11].

Although less explored in the context of this work, some external control mechanisms include recontacting respondents about their answers [17], using metadata and paradata such as the length of an interview as indicators [18], or implementing follow-up specific quality assurance falsification interviews as in [19].

Internal control mechanisms, on the other hand, use the existing survey answers to implement statistical analysis and draw answers. This is usually more cost and time-efficient than external control mechanisms but requires more mathematical nuance, as well as more domain and context awareness. We will discuss three common approaches: Benford's Law, supervised machine learning algorithms, and unsupervised machine learning algorithms.

Benford's Law states that the first digit of all numbers in survey answers should follow a declining monotonic distribution, and thus the proportion of 1's should be higher than the 2's, and so on [20]. This is a generally reliable way to identify falsification and has applications in other kinds of anomaly detections as well. It is widely used in survey data analysis with ambiguous results, where [21] calculated the deviation from the Benford distribution for every interviewer cluster to identify about 50% of falsifiers.

Benford's Law was also explored in [22], but it is suggested a more efficient approach would be to adapt Benford's distribution to one where not only the proportion of 1's is highest, but also has a higher proportion of 5's than other numbers due to respondent's tendency to round to that number. As we will see in the next section, our data contains a relevant number of categorical features, and thus Benford's law is not ideal for this specific experiment.

Supervised learning algorithms have also been employed in survey research, with their more prominent applications in predicting survey non-response [23]. Some of the more popular classification algorithms were implemented in different simulated datasets to detect survey fabrication in [24], often achieving more than 90% accuracy for both logistic regression and random forest. A similar approach was derived from this study in [13] using a random forest classifier with high prediction rates. While these approaches tend to have very good performance, they also tend to fall short in real, non-simulated scenarios, as we will not have the correct labels for which answers are fabricated.

Finally, clustering analysis is a very common approach to solving the problem of survey falsification. These unsupervised methods have been shown to have positive results for both form-level and interviewer-level data [24]. With a more sophisticated combination of Benford's law and multivariate cluster analysis, [25] managed to successfully identify falsified information from real answers. In [12], an aggregation of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Isolation Forest and Principal Component Analysis was implemented to cluster survey data, achieving more than 95% accuracy, true positive rates and true negative rates when the clustering was compared to the real labels in simulated data.

IV. Methodology

The data consists of 28,389 records and 20 features. 12 of these features are categorical (nominal), while 8 are numerical. While only 5 samples with missing data were dropped, the rest of missing values came from a survey question that was dependent on the answer of another: if an answer to question 2 would only exist if question 1 was answered in a certain way, question 2 would have missing values where an answer is not applicable. These missing values were encoded as a separate category for the purposes of this experiment. Our data is also limited to form level, as we have no data on the specific interviewers to use their information in our fabrication predictions. This must be taken into consideration when defining our approach.

Only data from Questionnaire A about household and member characteristics and income was used. Because of our experimental design, the specific details of each feature inside the questionnaire are not too interesting. In general, the categorical features consist of answers to physical living conditions, material property, health insurance, and personal and geographic identifier questions. Numeric variables consist of income questions and amount of material property. The categorical features will be encoded using Frequency Encoding, as the probability of two categories holding the same frequency is very low. The numeric features will be scaled using min-max normalization.

The data from the survey was limited to Questionnaire A due to a couple of reasons: it is the first questionnaire and thus early detection of falsification will be more useful in this context. It is also the most ambiguous in its data collection processes, but also the one that will be most likely to be taken into account by an interviewer that would curbstone the next set of questionnaires [11] in a systematic way, having a snowball effect on the totality of the survey results. Because of extremely high dimensionality, using only a subset of questionnaire A proved better for model performance, and in practice may be more beneficial with correct identification of relevant features more prone to fabrication. This identification is outside of the scope of this experiment because our randomization is simulated, and thus would be expected to behave similarly to this subset in a real scenario.

The following experiment will focus on a methodology akin to [12]. We will randomize the answers on different proportions of the total samples, simulating completely fabricated responses based on no prior knowledge to fill in non-response. The variability in the proportion of fabricated data in surveys according to literature prompted an approach where a different size of the data was randomized and tested for each model. We will be testing three models on data with 1%, 5% and 20% of its samples fabricated.

The three models to be explored are the Gaussian Mixture Model (GMM), One-Class Support Vector Machine, and Isolation Forest. The models were chosen because of their success in the field of anomaly detection, which bears similarity with the problem of survey fabrication detection. A brief description of each algorithm will be discussed, followed by the overall architecture.

Finally, we will be evaluating each algorithm based on their precision, recall and F1-scores when compared to the true labels of our fabricated data, as defined by:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

A. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) implemented here is a mathematical model that represents the distribution of data as a mixture of several bell-shaped curves. Each curve, called a Gaussian component, represents a cluster of data points with similar features. The model estimates the weights and properties of these components using a special algorithm called Expectation-Maximization (EM), a maximum-likelihood estimator [26]. The weighted sum of the components is given by:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i)$$

Where:

x is the feature-dimensional data vector.

w_i is the mixture of weights.

$g(x|\mu_i, \Sigma_i)$ = Component gaussian densities

The EM first calculates the posterior probability of each data point belonging to each Gaussian component. Then it updates the parameters of each Gaussian component based on the posterior probabilities.

B. One-Class SVM

A novel approach to adapt traditional SVM techniques for one-class classification was proposed in [27], where only data from the target class is available. Their method involves:

- Mapping the data into a higher-dimensional space using a kernel function.
- Treating the origin as the sole member of a hypothetical "second class."
- Employing parameters to control how strictly the one-class data is separated from the origin.
- Applying standard two-class SVM algorithms to achieve the desired separation.

The authors frame the problem as finding a simple subset of the feature space where the probability of encountering a test point (from the target class) outside this subset is bound by a pre-specified value. This essentially translates to estimating a function that takes positive values within the "safe" region containing the target data and negative values outside it. This makes the algorithm especially useful in anomaly detection problems.

C. Isolation Forest

Isolation forests are a type of unsupervised machine learning algorithm used for anomaly detection. Unlike other anomaly detection techniques that focus on learning "normal" behavior, isolation forests work by isolating anomalies by "randomly partitioning" the data.

The algorithm defined in [28] starts by taking a set of data points and randomly selecting a feature. It then randomly chooses a split value within the range of that feature. This split value divides the data into two subsets: one containing data points with values less than the split value and another containing those with values greater than the split value. This process of random splitting is repeated recursively for each subset until a certain stopping criterion is met (e.g., maximum depth reached).

Once this process is done, the algorithm assigns anomaly scores to each data point based on the average path length it takes to isolate said point. Ideally, anomalies will be characterized by having shorter path lengths as they will be isolated earlier on tree construction. The anomaly score s are given by:

$$s(x,n) = 2^{\frac{-E(h(x))}{c(n)}}$$

Where:

$E(h(x))$ is the average height of x in the collection of isolation trees.

$c(n)$ is the average path length of unsuccessful searches in binary search trees.

ALGORITHM 1: RANDOMIZATION AND MODELING

```
    Input: survey data
    Output: three sets of clustered data for each algorithm
1   tailored_data ← feature_selection(survey_data)
2   encoded_data ← frequency_encoding(tailored_data)
3   normalized_data ← normalization(encoded_data)
4   for i in random_sample(normalized_data) // i ← records
5       for x in normalized_data(i) do // x ← features
6           If (i,x) = continuos then
7               (i,x) ←random integer in range x
8           If (i,x) = categorical then
9               (i,x) ←random choice from list x
10  Precision, recall, f1-score ←GMM(normalized_data)
11  Precision, recall, f1-score ← OCSVM(normalized_data)
12  Precision, recall, f1-score ← IF(normalized_data)
```

V. Results

Table 1 shows the precision, recall and F1-score of the three models attained at each level of fabricated data for the class representing that the sample was altered. Overall, the One-Class Support Vector Machine (OCSVM) was more consistent in its high performance at identifying fabricated data, with F1-scores ranging from 79% to 88% for all three samples. With the minimal proportion of fabricated data of 1%, the Isolation Forest

(IForest) seemed to struggle at correctly predicting the few anomalies. However, it becomes considerably better when exposed to higher shares of fabricated data, jumping to 86% at 5% and 89% at 20%. It is worth noting that although OCSVM is more consistent across the board, it seemed to be less reliable when exposed to more categorical features than we have in the current setup.

The Gaussian Mixture Model held very low precision scores at both 1% and 5% fabricated data, which means the model had trouble identifying the anomalies correctly and overclassified many samples as fabricated data, which explains its very high recall as well. Given this behavior, its higher performance with 20% fabricated data seems understandable, and in terms of F1-score it performed the best out of the three models with this dataset.

TABLE I
PRECISION, RECALL AND F1-SCORE FOR FABRICATED CLASS

Model	FABRICATED DATA PROPORTION	Precision	Recall	F1-score
GMM	1%	13%	97%	22%
	5%	57%	99%	72%
	20%	86%	98%	92%
OCSVM	1%	79%	80%	79%
	5%	86%	86%	86%
	20%	88%	88%	88%
IForest	1%	66%	66%	66%
	5%	86%	86%	86%
	20%	89%	89%	89%

Lastly, it is worth taking into consideration that out of the three models, GMM is the only one that does not require a threshold for the decision function to be defined beforehand, which is contradictory with its performance. IForest and OCSVM, on the other hand, benefit from having an idea of what the proportion of fabricated data is, and thus may require more extensive knowledge or analysis in non-simulated environments despite their higher performance.

VI. CONCLUSION

The unsupervised learning methods explored in this work, often used in anomaly detection, can generally be used to identify fabricated answers in the large-scale household surveys. After implementing GMM, one-class SVM and Isolation Forests on ENGIH 2018 data, the latter two had success in correctly predicting which samples were falsified. This research shows machine learning applications can aid the data collection process in large-scale surveys of the Dominican Republic and boost the research engine in the country.

Some of the limitations of this research include a lack of interviewer-side data, which has been discussed to be a powerful predictor of fabrication. This is especially relevant when dealing with curbstoning, as interviewers may be smarter about how they fill in the missing data. This provokes the idea that fabricated data may follow other patterns besides the random distribution explored in this experiment. Although these different distributions and a more sophisticated modeling are beyond the scope of this project, an expansion focusing on these issues is planned for the future.

VII. REFERENCES

[1] O. Castorena, M. Cohen, N. Lupu and E. Zechmeister, "How worried should we be? The implications of fabricated survey data for political science.," *International Journal of Public Opinion Research,* vol. 35, no. 2, p. edad007, 2023.

[2] R. Gomila, R. Littman, G. Blair and E. L. Paluck, "The audio check: A method for improving data quality and detecting data fabrication.," *Social Psychological and Personality Science,* vol. 8, no. 4, pp. 424-433, 2017.

[3] K.-H. Reuband, "Interviews that are not interviews: "Successes" and "failures" in cheating on interviews," *kölner zeitschrift für soziologie und sozialpsychologie,* 1990.

[4] I. Santana, "La canasta de consumo de los hogares dominicanos, su evolución histórica y," *Ciencia, Economía y Negocios, 5(2), 129–149.,* vol. 5, no. 2, pp. 129-149, 2021.

[5] F. Alvaredo, M. de Rosa, I. Flores and M. Morgan, "Desigualdad del ingreso en la República Dominicana 2012-2019: una revisión a partir de la combinación de fuentes de datos," CEPAL, 2022.

[6] R. C. Alonso, C. Polanco, R. Mancebo and E. de León, "Diferencias metodológicas entre el cálculo de pobreza monetaria oficial de República Dominicana en 2020 y los publicados por la CEPAL en el Panorama Social 2021.," Ministerio de Economía, Planificación y Desarrollo de la República Dominicana, Santo Domingo, 2022.

[7] I. Acevedo, E. Castro, R. Fernandez, I. Flores, M. P. Alfaro, M. Szekely and P. Zoido, "Los Costos Educativos de la Crisis Sanitaria en América Latina y el Caribe.," BID, 2020.

[8] F. Rogers and M. Richarme, "The honesty of online survey respondents: Lessons learned and prescriptive remedies," *Decision Analyst, Inc White Papers,* pp. 1-5, 2009.

[9] J. J. Hox, "Hierarchical regression models for interviewer and respondent effects," *Sociological methods & research,* vol. 22, no. 3, pp. 300-318, 1994.

[10] OECD, "PISA 2018 Assessment and Analytical Framework," OECD Publishing, Paris, 2019.

[11] S. Walzenbach, "Do Falsifiers Leave Traces? Finding Recognizable Response Patterns in Interviewer Falsifications," *methods, data, analyses,* vol. 15, no. 2, p. 36, 2021.

[12] N. M. Jebreel, R. Haffar, A. K. Singh, D. Sánchez, J. Domingo-Ferrer and A. Blanco-Justicia, "Detecting bad answers in survey data through unsupervised machine learning.," in *Privacy in Statistical Databases:*

UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, 2020.

[13] I. Hernandez, T. Ristow and M. Hauenstein, "Curbing curbstoning: Distributional methods to detect survey data fabrication by third-parties.," *Psychological Methods*, vol. 27, no. 1, p. 99, 2022.

[14] A. Finn and V. Ranchhod, "Genuine fakes: The prevalence and implications of data fabrication in a large South African survey," *The World Bank Economic Review*, vol. 31, no. 1, pp. 129-157, 2017.

[15] S. Bredl, N. Storfinger and N. Menold, "A literature review of methods to detect fabricated survey data," 2011.

[16] S. De Haas and P. Winker, "Detecting fraudulent interviewers by improved clustering methods-the case of falsifications of answers to parts of a questionnaire," *Journal of Official Statistics* , vol. 32, no. 3, p. 643, 2016.

[17] A. Koch, "Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994," *ZUMA-Nachrichten*, vol. 19, no. 36, pp. 89-105, 1995.

[18] C. Hood and J. Bushery, "Getting more bang from the reinterviewer buck: Identifying 'at risk' interviewers," in *Proceedings of the American Statistical Association (Survey Research Methods Section)*, 1997.

[19] E. A. Krejsa, M. C. Davis and J. M. Hill., "Evaluation of the quality assurance falsification interview used in the census 2000 dress rehearsal," in *Proceedings of the American Statistical Association (Survey Research Methods Section)*, 1999.

[20] F. Benford, "The law of anomalous numbers.," in *Proceedings of the American Philosophical Society*, 1938.

[21] J. Schraepler and G. G. Wagner, "Characteristics and impact of faked interviews in surveys–An analysis of genuine fakes in the raw data of SOEP.," *Allgemeines Statistisches Archiv*, vol. 89, pp. 7-20, 2005.

[22] Y. Wang and S. Pedlow, "Detecting falsified cases in scf 2004 using Benford's Law," in *Proceedings of the American Statistical Association (Survey Research Methods Section)*, 2005.

[23] C. Kern, T. Klausch and F. Kreuter., "Tree-based machine learning methods for survey research," *Survey research methods*, vol. 13, no. 1, p. 73, 2019.

[24] B. Birnbaum, "Algorithmic approaches to detecting interviewer fabrication in surveys," University of Washington, 2012.

[25] S. Bredl, P. Winker and K. Kötschau, "A statistical approach to detect cheating interviewers," 2008.

[26] D. A. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, vol. 741, no. 659-663, 2009.

[27] B. Schölkopf, R. Williamso, A. Smola, J. Shawe-Taylor and J. Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.

[28] F. T. Liu, K. M. Ting and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, 2008.