Mi Casa No es Tu Casa: An Agile Strategy to Generate Synthetic Data

to Overcome Security Challenges in Mexico.

José Carlos Rodríguez Pueblita

IPADE Business School [1]

CDMX, México

jcpueblita@gmail.com

Edgar Oswaldo Díaz

Instituto Nacional de Estadística y Geografía [2]

Aguascalientes, México

oswaldo.diaz@inegi.org.mx

Abstract — This document describes an agile strategy in statistical analysis to generate synthetic data to overcome increasing obstacles to carry out face-to-face surveys in Mexico, such as increasing insecurity, limited access to certain areas controlled by organized crime and budgetary constraints. We use two data sources: the 2020 Income-Expenditure Survey or ENIGH by its acronym in Spanish, and the 2020 Population and Housing Census, or CPyV, both carried out by the National Institute of Statistics and Geography of Mexico (INEGI), and several statistical learning techniques such as PCA, clustering, random forest and classification methods to generate granular synthetic data with scientific, policy and commercial uses. The result is an algorithm that allows characterizing the socioeconomic level and the income and expenditure profiles of urban households at block level for all the country, and we suggest metrics to validate the synthetic data due to the impossibility of having disaggregated data to validate our results. This agile strategy can be replicated for various contexts where data layers that satisfy certain basic conditions.

Keywords – Synthetic data, PCA, clustering, random forest, classification methods, survey, imputation, organized crime.

1. A graph of a number of people

   Description automatically generated IntroduCTION

Consumer profiling is highly data driven. Governments and companies use face-to-face surveys intensively to obtain valuable information for market analysis, segmentation, community, healthcare, and real estate planning, among others. This data provides valuable insights into the diversity and complexity of neighborhoods, helping businesses, organizations, and policymakers make data-driven decisions and better understand their target audiences. These services are in high demand in developed and developing economies as well. Recently, natural disasters -such as COVID-19- or environmental changes -for instance, increasing insecurity- have limited the access to first-hand quality data. In fact, the National Survey of Occupation and Employment (ENOE, for its acronym in Spanish), the source unemployment and informality quarterly measurement, implemented by the National Institute of Statistics and Geography of Mexico, namely Mexican Census Bureau o INEGI, had to switch to telephone interviews because of COVID-19. In 2020, 21.4% of ENOE’s interviews were done by telephone [4].

Public safety is already a chronic concern in Mexico since drug-related violence surged dramatically in early XX century due to the war on drugs. On one hand, large-scale efforts to combat trafficking and traffickers’ attempts to usurp territories have increased violence across the country. Dell (2015) showed divertion of drug traffic due to government efforts to shut down trafficking routes with the unentended effect of increasing violence along alternative drug routes. On the other hand, cartels have diversified their activities to robbery, extorsion and kidnapping, among other illegal activities, reinforcing competition among cartels and thus, violence. Violence as measured by homicide rates have increased since 2008, mainly in the last ten years (see Figure 1). This has naturally affected perception of insecurity, measured as percentage of population that reported living in unsecured cities (see Figure 2). Insecurity and exogenous shocks, such as natural disasters, have severely limited access of surveyors. In fact, there are localities were periodic institutional surveys had to be cancelled as seen in Figure 2 due to COVID-19 in the second quarter of 2020 (see pointed line), or in Acapulco in the fourth quarter of 2023 due to hurricane Otis or faced low responde rates of responses caused by extremely high perception of insecurity as in Fresnillo, Zacatecas (96.4%), as depicted in Map 1 below.

Figure 1: Homicides in Mexico, 1990-2022, Source: México Unido Contra la Delincuencia y la Impunidad with data from the Mexican Census Bureasu (INEGI)

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 2: Social Perception of Public Insecurity in Mexico

(Percentage of Population Reporting Living in Unsecured Cities), Source: INEGI

Mapa, Gráfico de dispersión

Descripción generada automáticamente

Map 1: Social Perception of Public Insecurity in Selected Mexican Cities (Percentage of Population Reporting Living in Unsecured Cities), Source: INEGI.

INEGI reported that non-response rates official surveys have reached significant levels [5]. For instance, the 2022 Census of Centers for Women’s Justice faced a 22% non-response rate and imputation rates were 19% of questionnaires. In December 2023, the Quarterly Consumer Confidence Survey faced a 15% non-response rate [6]. The non-response or existence of missing values is present in every project to generate statistical and geographical information. It can induce bias in the estimates or statistics generated from the information. One of the main average sources of recent higher than usual rates of non-response is attributed to insecurity. In the case of extremely unsafe localities surveyors are not even allowed to operate [7]. It was reported that for the 2022 Census of Agriculture surveyors paid to organized crime to carry out their tasks [8]:

*“There are very different strategies, from, in some cases paying to enter, perhaps paying small amounts, but paying to enter (sic); to hiring personnel from the area who know very well the people of the locality or the area where the census is being carried out, and who it is also known to those people, who could be engaging in crime. And with that, well, the truth is, entry to all the places is very simple.”*

Susana Pérez Cadena, Adjunct Director-General of Economic and Agriculture Censuses, INEGI.

Another challenge faced specifically by statistical agencies is budgetary. In the case of Mexico, INEGI had to change their methodologies to adjust to budgetary restrictions. The 2024 Economic Census, that surveys all businesses in the country, has seen its budget reduced by 9.2% with respect to its latest edition in 2019. Therefore, surveyors have appealed to statistical method to impute non-accessible data. Although imputation at the observation unit level tries to identify the existence of patterns in non-responses it could induce bias, “the non-response rate thus must be calculated and compared across the different study domains, the most important variables used for stratification or the grouping that could be related to the main variables of the different projects” (INEGI, 2017).

One alternative to complement or even substitute surveys is the use of synthetic data, defined as stochastically generated data with analytical value with the virtues that maintains the confidentiality of real data. It provides an alternative to analytical science and becoming more relevant considering its life cycle from collection to storage, and useful in data science and artificial intelligence (AI) processes, allowing improvements to statistical models

with applications in financial services, retail, manufacturing, healthcare, automotive, robotics, security, social media, and marketing, with direct and measurable benefits (Koenecke and Varian, 2020; and Birisci et al. 2023).

There are many policy applications, such as the use made by the United State Census Bureau to generate Estimates of Poverty and Income of Small Areas, allowing disaggregation at the individual and/or commercial record level without compromising confidentiality, to generate public policies [9]. A similar case for commercial purposes is the Esri's Tapestry Segmentation (TS) that provides an adequate and detailed description of American neighborhoods. It divides residential areas of the United States into 67 segments based on their socioeconomic, demographic composition, classifying lifestyle groups and urbanization. It uses a segmentation system to integrate market trends that are homogeneous within and heterogeneous abroad.

We use the former examples to motivate our work with the aim to generate an agile stage-based strategy to process open data for the generation of a synthetic database. Our aim is to characterize the socioeconomic level and the income and expenditure profiles of urban households at block level in all Mexican urban localities -i.e., with more than 2,500 inhabitants-, applying a population segmentation exercise using information on a survey of income and expenses.

From an analytical point of view, the resulting Mexican Households Synthetic Database -hereby MHSD- is thought of as a tool that allows the researcher or analyst to obtain an estimate of the socioeconomic characteristics for a certain block, including the level and formation of income and expenses per household, when knowing sociodemographic variables of that block. Thus, the socioeconomic segmentation at the block level of all urbanized areas in Mexico provides a statistical overview of the behavior of household income and expenses regarding its amount and distribution, based on the use of two open data sources. It is worthwhile mentioning that the method can be replicated with all types of survey and administrative data as soon as they satisfy the basic conditions outlined below.

Prior to the construction of the MHSD, there was no database or methodology to know or estimate the socioeconomic composition of all the blocks in urbanized areas in the country without conducting face-to-face surveys. Because the security challenges and budget limitations explained above, we consider that having a socioeconomic segmentation of households at the most disaggregated level possible -without compromising privacy- useful for companies and governments to understand the composition of income and expenditure throughout Mexico without the need to conduct face-to-face surveys and without compromising quality of information. The MHSD is an asset that has already been used for policy and commercial purposes, for instance, to know the socioeconomic profile of users of public services, forecasting expenditure on certain goods and services, and enriching models of credit score [10].

The rest of the paper is organized as follows: Section II describes the statistical strategy, the basic conditions and the construction of the algorithm that allows obtaining a synthetic database of household at block level. Section III presents validation metrics that consider the impossibility of having disaggregated data to validate the synthetic data. Section IV discusses the technological strategy, and Section V concludes.

1. STATISTICAL STRATEGY.

The fundamental methodology for the construction of the MHSD is formed by the following building blocks: (a) Two layers of data that satisfy certain conditions, and (b) three main stages listed below. The two basic conditions the two layers of data must satisfy are:

Condition 1 (Common Geography but Different Aggregation): Availability of two layers of data, namely datasets X and Y, that cover certain common geographical space G, where the statistical representativeness of the later is more granular or less aggregated than the former. Formally, X is an q-x-i matrix with entries xqi, where q is a state or municipality and i is a socioeconomic label; and Y is an l-x-j matrix with entries yli, where l is a household in locality s and j is a sociodemographic label.

Condition 2 (Common Vector): There must a vector c of “common” variables or labels between two layers of data, and a set of objective variables v non-available at the input layer, more formally:

being s the set of sociodemographic labels and u the set of socioeconomic variables, the intercept of the two sets of variables or c common vector in non-empty. In our experiment, the two layers of data are: (i) The 2020 income-expenditure survey -hereby ENIGH for its acronym in Spanish-, that provides social, economic and demographic information statistically representative at state and, in some cases, municipal level; yet does not provide information at the block level; and (ii) the 2020 Population Census microdata, that provides information on the social and demographic characteristics of the households at a block level, but does not include any economic data information, nor does it carry out a segmentation exercise. In the case of the MHSD, the common vector c or characteristics are social and demographic, and the objective variables v are economic. Therefore, Common Geography but Different Aggregation Condition and Common Vector Condition are satisfied.

The three main stages of the statistical strategy listed in detail below are: (i) Collection, cleaning and integration of databases, (ii) selection of relevant variables or features, (iii) definition of segment or groups, (iv) construction of predictive models for households, and (v) classification of households per block.

After the first stage is concluded -collecting, cleaning, and integrating the two main data sources-, the second stage consists of carrying out an exploratory analysis of the ENIGH (i.e., dataset X) and applying unsupervised learning methodologies, namely principal component analysis (PCA), to choose the most relevant subset of variables. In our experiment, the variables current income, number of household members, and number of working members were selected as principal components (see Verdhan, 2020). Then, in the third state, clustering techniques are applied as in Gong and Xu (2007), in the grouping of households to define three-dimensional groups or segments which make up a segmentation of the ENIGH, making natural cuts, which are obtained with the Jenks algorithm, an equilibrium method between equality interval, natural cuts (North, 2009). The result is the MHSD segmentation formed by 58 groups of households of the ENIGH defined by current income, number of household members and number of working members.

Diagrama

Descripción generada automáticamente

Figure 3. Diagram showing the defined MHSD segments or groups.

We apply supervised learning techniques in the fourth stage to obtain a function that classifies households surveyed in the CPyV (i.e., Census or dataset Y) in their respective group using the set common variables c, plus complementary variables for the classification process, as Aidi et al (2022) used in Indonesian data to predict economic outcomes based on online data. We use random forest and random over sampler to balance the classes as in Aidi et al (2022) and Albrecht et al (2023), respectively. For the case of missing observations, we use frequent imputer for categorical variables and mean imputer for continuous variables, as in Loong and Rubin (2017), as well as recursive feature elimination as in Chen and Jeong (2008) for specific training for each of the Mexican states. The outcome is 32 income classification models, meaning one per state, allowing for better training results.

In the fifth and last stage, then we classify all households contained in the CPyV data where the Common Vector Condition is satisfied, using the model created in the previous stage, obtaining the probability that each household in the block belongs to each one of the state-specific income segments. The probability of belonging to the segments of number of persons per household is estimated following the distribution of the number of persons per household in each state. The average number of persons per household in the block is obtained by dividing the number of people by the number of households in the block, scaling the mean of the distribution obtained by entity, obtaining the probability that the households in each block belong to each of the household-size segments. In a similar fashion, we obtain the segment of the number of working persons by dividing the number of working persons in the block by the number of households in said area. It is assumed that this average value is how each of the households in the area behave. Therefore, the probability that any block belongs to each of the segments is written using the following joint probability formula:

.

Where P(m∈i) represents the probability that a household in block m belongs to segment i= {0,1,2,3,4,5} of current income, value obtained by the model trained in the stage above applied to the Census data. P(m∈j) denotes the probability that a household in block m belongs to segment j= {0,1,2,3} of the number of people per household, obtained by the distribution of the total number of members corresponding to the state, adjusting the average to the corresponding one for the block. Finally, P(m∈k) refers to a binary value, corresponding to the probability that a household in block m belongs to segment k={0,1,2,3}, taken from the mean of the variable for said zone. This joint probability establishes the assumption that the household of each block behaves like the average household of the block.

After obtaining the probabilities of each segment for each states, we proceed to estimate other target variables of the ENIGH, for example total income, credit card payments, coffee spending, retirement income, to which instead of making a different model for each of the variables (regression or classification), we use the MHSD segmentation estimates, assuming that for a specific locality q and a specific block, the probabilities of the m segments are:

.

Considering that a particular variable x has the following averages within the segments of that state:

,

and that the averages of the variable x are calculated over the entire state to which the block analyzed belongs, the estimation of x in the block m is

.

If we fix a particular locality q, we have the following general procedure:

* n is the number of blocks in locality q
* m is the number of segments
* v is the number of ENIGH variables
* p\_ij is the segment probability j by block i
* x\_kl is the average of the variable l for households in the segment k

then, the ENIGH variable state-specific matrix for the average household is given by:

Forma

Descripción generada automáticamente con confianza media

Thus, the resulting scalar M\_ij approximates the value of variable j for the average household in block i, whwte the matrix M has n×v dimensions, given that there are n blocks per state with v estimated variables.

III VALIDATION METRICS.

Due to the impossibility of having disaggregated data to validate our estimated variables v, for instance, income and expenditure, we suggest metrics is to assess the quality of the synthetic data. These metrics study the number (or proportion) of households in each segment, divided by the number of real data in each segment, against the number of data predicted in the same segment.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 4. Graph showing the statistical behavior of the actual and predicted data by segment.

The following closeness statistic is proposed based on metrics such as F1 or recall, used to validate classification models (Bachinger and Kronberger, 2023):

,

were segment i portrays each of the ENIGH 58 segments.

Note that the metric ranges from -1 to 1, thus it takes extreme values only when there are no actual or predicted values in the class. The best possible result is for the metric to be equal to 0, which occurs only when the number of predicted elements is equal to the number of real elements. The value will be positive when there are more real values than predicted and it will be negative otherwise.

where the proportion is understood as:

Generalized metric (absolute)

To obtain a generalized metric over all segments, we propose the following:

The resulting metric is a measure of the absolute difference between the amount of data between the segments. It is worth mentioning that this metric has values between 0 and 1, where zero represents that there is no difference between the amount of data in the classes and 1 refers to the opposite.

Balanced generalized metric

Giving the importance of the segments with a greater number of data, a balanced generalized metric seems convenient, in which the imbalance of the segments is considered, meaning:

where Pi is denoted as the proportion of elements in segment i, that is,

The metric takes values between 0 and 1, so that smaller values indicate a greater closeness between the amount of data for each segment in the two open databases compared.

Table I. Validation Results

HERE

IV TECHNOLOGY STRATEGY.

For the generation of the algorithm presents, a controlled sandbox-type scenario was implemented, connected to a data lake, using IDE Jupyter notebooks and python packages. Data Engineering codes are created for extraction, loading, and transformation of open data that serve as an input for training with data science techniques, according to best practices (see Villaseñor García et al., 2022). The MHSD can be consulted per geographical coordinate in the following webpage: www.pondera.xyz. Access to the collaborative repository for the project is available at: https://github.com/jcpueblita/AnalyticSyntheticData

V. CONCLUSIONS.

The literature on synthetic data highlights its value as a data source that allows analyzing and forecasting behaviors without violating the privacy of the individuals under study. The increasing insecurity, limited access to certain areas controlled by organized crime and budgetary constraints faced in Mexico by public agencies such as INEGI and private companies, as well, as the evolution of personal data and information protection regulation that seeks to give individuals more control over their personal data and forces companies to ensure that the way they collect, process and store information is secure, we forecast that the use of face-to-face surveys, individual transactional or operational data for the purposes of development and evaluation of public and commercial strategies will encounter greater limitations in the near future. In this document we present a strategy to generate synthetic data to segment the population based on their socioeconomic profile using public databases and we propose an algorithm that allows very granular forecasts, either for scientific, public policy or commercial purposes. Additionally, we propose metrics to assess the quality of synthetic data given the impossibility of using traditional metrics due to the nature of the challenge. This methodology makes it possible to estimate point values of economic variables that facilitate answers to specific questions of relevance for designers and evaluators of public policies or commercial or marketing strategies. The approach presented can be replicated based on relatively undemanding basic conditions, so it can be widely applied as soon as there are population censuses with accessible microdata and surveys that reflect the behavior of household income and spending.

ACKNOWLEDGMENTS

The authors thank Roberto Cabal, Alain Cabrera, Alonso Medina, Fernando Rivas, and Saskia van Der Werff for their support in research assistance and implementation of the estimation strategy. We also appreciate the sponsorship of Pondera S.A.P.I. de C.V. who financed the development of this method.

ReferENCES

[1] Department of Business Economics, IPADE Business School, Floresta 20, Clavería, Azcapotzalco 02080 Mexico City, Mexico, Email: [jcpueblita@ipade.mx](mailto:jcpueblita@ipade.mx).

[2] Science Data Laboratory, INEGI, Av. Héroe de Nacozari no 2301, Aguascalientes, México, Email: [oswaldo.diaz@inegi.org.mx](mailto:oswaldo.diaz@inegi.org.mx).

[3] Market optimization companies like Nielsen primarily collects data from persons through several methods, with a focus on measuring media consumption habits such as television viewing, radio listening, online streaming, and more. Some common methods Nielsen uses to obtain data from individuals are panel Measurement, surveys and diaries, metering devices, digital tracking, partnerships and data aggregation, and retail measurement. ESRI (Environmental Systems Research Institute), a leading provider of geographic information system (GIS) software and data analytics solutions., developed Tapestry Segmentation as a geodemographic classification system that categorizes U.S. residential neighborhoods into distinct segments based on demographic and socioeconomic characteristics, lifestyle preferences, and behavioral patterns.

[4] Source: INEGI, <https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=889463909743>

[5] Non-response rates are defined as <https://extranet.inegi.org.mx/calidad/wp-content/uploads/2018/03/FT_IC_P_TASA-NO-RESPUESTA-unidad_act_171205.pdf>

[6] Source: <https://www.inegi.org.mx/rnm/index.php/catalog/839>

<https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2024/ENSU/ENSU2024_01.pdf>

[7] “Two INEGI surveyors missing in Chihuahua are found dead”, Animal Politico, 2015, March 11th 2015, <https://www.animalpolitico.com/2015/03/dos-encuestadores-del-inegi-desaparecidos-en-chihuahua-son-hallados-muertos>.

[8] “INEGI paid to organized crime to carry out the census”, Aristegui Noticias, July, 28th, 2023, <https://aristeguinoticias.com/2807/dinero-y-economia/inegi-pago-al-crimen-organizado-para-poder-realizar-censo/>

[9] Source: h <https://www.census.gov/about/what/synthetic-data.html>

[10] See a brief demo of different commercial uses here: https://pondera.xyz/. Pondera was founded by one of the authors in 2015.

[11] Athey, Susan; Imbens, Guido; “Machine Learning Methods That Economists Should Know About”, Annual Review of Economics, 2019. 11:685–725

[12] Bachinger, Florian & Kronberger, Gabriel. (2023). Comparing Shape-Constrained Regression Algorithms for Data Validation. 10.1007/978-3-031-25312-6\_17.

[13] Chen, Xue-wen & Jeong, Jong Cheol. (2008). Enhanced recursive feature elimination. 429 - 435. 10.1109/ICMLA.2007.35.

[14] “Cómo se hace la ENOE. Métodos y procedimientos”, INEGI, 2023 <https://www.inegi.org.mx/app/biblioteca/ficha.html?upc=702825190613>

[15] Gong, Yihong & Xu, Wei. (2007). Data Clustering Techniques. 10.1007/978-0-387-69942-4\_3.

[16] Gürsakal, Necmi & çelik, Sadullah & Birisci, Esma. (2023). An Introduction to Synthetic Data. 10.1007/978-1-4842-8587-9\_1.

[17] Khairani, Fitri & Kurnia, Anang & Aidi, Muhammad & Pramana, Setia. (2022). Predictions of Indonesia Economic Phenomena Based on Online News Using Random Forest. SinkrOn. 7. 532-540. 10.33395/sinkron.v7i2.11401.

[18] Koenecke, Allison & Varian, Hal. (2020). Synthetic Data Generation for Economists.

[19] Loong, Bronwyn & Rubin, Donald. (2017). Multiply-Imputed Synthetic Data: Advice to the Imputer. Journal of Official Statistics. 33. 10.1515/jos-2017-0047.

[20] North, Matthew. (2009). A Method for Implementing a Statistically Significant Number of Data Classes in the Jenks Algorithm. 6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009. 1. 35-38. 10.1109/FSKD.2009.319.

[21] Population and Housing Census CPyV, INEGI 2020

<https://www.inegi.org.mx/programas/ccpv/2010/default.html#Datos_abiertos>

<https://www.inegi.org.mx/programas/ccpv/2020/#Datos_abiertos>

[22] Verdhan, Vaibhav. (2020). Supervised Learning with Python: Concepts and Practical Implementation Using Python. 10.1007/978-1-4842-6156-9.

[23] Zhong, Ming & Shen, Kai & Heidkamp, John & Kosti, A & Rodriguez, J & Albrecht, A & Bhirud, A & Vah,. (2023). A Random Membership Sampler Based on Guided Random Walks.

[24] Synthetic Data, United States Census Bureau

<https://www.census.gov/about/what/synthetic-data.html>

[25] Survey National Income Expenditure in Households ENIGH, INEGI, 2020

<https://inegi.org.mx/contenidos/programas/enigh/nc/2020/doc/enigh2020_ns_descriptor_archivos_fd.pdf>

<https://www.inegi.org.mx/programas/enigh/nc/2020/#Datos_abiertos>

[26] Tapestry Segmentation (TS) by ESRI

<https://storymaps.arcgis.com/stories/6e8f2d8c08d8427892e816d1aeb373f8>

[27] Tasa de no respuesta antes de imputación a nivel unidad de observación, Formato para presentar propuestas de indicadores de calidad. Versión 3.0, Comité de Aseguramiento de la Calidad Indicadores de calidad del INEGI, October, 2017

[28] Villaseñor García, E.A. et al., "Data Lake Strategy for Data Science Workflows," 2022 11th International Conference on Software Process Improvement (CIMPS), Acapulco, Guerrero, Mexico, 2022, pp. 219-223, Doi: 10.1109/CIMPS57786.2022.10035694.