Methods and Data Sources for Measuring Socio-Economic Factors: A Literature Review

Vizuete-Salazar, Yasmina, *Escuela Politécnica Nacional, Facultad de Ingeniería de Sistemas*,

Segura-Morales, Marco, *Escuela Politécnica Nacional, Departamento en Informática y Ciencias de la Computación*

***Abstract*—The compiling of the population data, to establish its socioeconomic factors, is a high-cost task for governments and regulatory organizations due to the need for financial and human resources. This limitation makes it almost impossible to count on immediate updated socioeconomic population information. This article compiles a series of alternative data sources and methods that can be applied to reduce the costs and the time required to update such information. The review focus on how these sources and methods have been used in developing countries during time, highlighting the solutions for satisfying the need of updated socioeconomic factors of the population.**

***Index Terms*— Census, Data analytics, Population, Socio-economic factors.**

I. INTRODUCTION

CENSUSES provide important information about the status of a population since they allow the establishment of its socio-economic factors and through them derive poverty indexes [1].

Poverty is a complex phenomenon that needs to be understood from many points of view in order to be taken into account adequately [2]. Nowadays, poverty is measured in a single dimension called income poverty [3], which is usually based on income or consumption. However, there are other variables with which the socio-economic situation of a family can be established (e.g. the Human Development Index [HDI] uses three dimensions: health, knowledge and life status; the Multidimensional Poverty Index [MPI] is composed of ten indicators grouped in three areas: education, health and living conditions) [3].

The method that is most widely used for collecting socio-economic data is the census. This means conducting door-to-door visits to perform surveys about family structure and the personal situation of each of its members. These visits are carried out every five to ten years depending on budgetary availability. After processing the survey data, the collected information allows the establishment of socio-economic family situations and, through them, an aggregated value for the whole country. The results allow the government and regulatory organizations to develop social policies to improve the living conditions of the population [3], [4].

However, door-to-door surveys implies a high cost that impacts the state budget, especially in developing countries [4]. Consequently, some methods have been implemented for reducing costs: the number of variables to be compiled is reduced, some census zones are not visited, or the visits are simply postponed [1].

Alternatively, some developing countries have developed alternative data collecting mechanisms to obtain information about families for determining their socio-economic status [1], [4]. These mechanisms include, among others: the use of social networks, administrative records, mobile phone call logs, basic service records. This allows to obtain good estimates of poverty indicators at a lower cost and in a shorter time [1]. Thus, this article aims to identify these alternative methods to collect data from families to determine their socio-economic status, which can be used instead of visits. This research objective will also allow the identification of alternative information sources (i.e. already existing data which is being stored in different public or private institutions) which would make it possible to measure socio-economic indicators [5].

The remainder of this paper is organized as follows: in section II, we present the research question. In section III, we describe the search process. This is followed by a presentation of our main results (section IV). In section V, we present the validity consideration of the literature review. Section VI presents a discussion of our results. The section VII presents the conclusions of the research. Finally, new opportunities of research, derived from this review, are offered.

II. RESEARCH QUESTION

In this research a systematic literature review was performed following the guidelines proposed by Kitchenham [6].

The research question is "which methods and techniques, other than the traditional door-to-door approach, are being used to predict the socio-economic conditions of population?

This question allows the identification of other methods that could be used in countries where state financial resources, for

Article history:
Received 15 April 2018
Accepted 28 May 2018

This work was supported in part by the Secretary of Higher Education, Science, Technology and Innovation of Ecuador (SENESCYT).

Vizuete-Salazar, Yasmina is with the Facultad de Ingeniería de Sistemas, Escuela Politécnica Nacional, Quito, Ecuador (e-mail: yasmina.vizuete@epn.edu.ec).

Segura-Morales Marco is with the Departamento en Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Ecuador (e-mail: marco.segura@epn.edu.ec).

use in census-type visits to the population in order to determine socio-economic status, are limited[4]. In addition, it allows for the identifying of other population information sources, already existing data which is being stored in different institutions (public or private) which would make it possible to establish behaviors and indicators that then could be related to the socio-economic situation of individuals or their family nucleus, such is the case of electricity consumption data.[5]

III. Literature Review Process

Considering the aim of the article is to identify alternative methods and data sources to measure socio-economic factors, we focused on publications reporting the use of predictive methods to establish the status, characteristics, conditions, or socio-economic factors of a population. Publications dealing with methods to determine population density, as well as economic growth, were also considered because the method used in these cases could also be used to predict socio-economic factors. It was decided that the words selected for the search were to be mentioned in the title, the keywords or in the abstract. There are the establish criteria for the research process.

The following combination was used as a search string: (socioeconomic OR economic OR household OR poverty OR depravation OR census OR survey) AND (predictive OR predicting OR "big data" OR "data mining" OR "data analysis" OR analytics). The period for the review was established between 2013 and 2017.

Searches were carried out for collecting electronic articles published in scientific journals and conferences in English using the search engines of SCOPUS and WEB of SCIENCE.

It is necessary to mention that the terminology used to describe the predictive method changes over time, by authors and even by country. In some publications, there are combinations with the concepts of economic growth, population and housing characteristics, so that the selection of publications was difficult to determine which could be used. Accordingly, there were publications identified and selected by the search criteria that did not comply with the objectives of this study (false positive), publications that were not detected in the search but that met the objective posed (false negative or opportunistic publications), and, publications that might be of interest to the objective of this study.

The results of the search carried out are shown in Table I. It shows the order in which the bibliographic databases were consulted. Forty-three publications were found in this manner and were loaded onto ATLAS.TI for compilation and analysis. Finally, twenty-nine publications, that comply with the selection criteria and contribute to answering the posed research question, were selected. Selected articles come mainly from scientific journals (24) and the rest are from conferences (5). Most of publications correspond to the year 2017 (9 of 29).

TABLE I

Results of the Review in Bibliographic Databases

Source	Searched	Selected
Scopus	35	25
Web of Science	8	4
Total	43	29

IV. Results

A. Analysis of the evolution of the methods

The methods used, and the purpose of the selected publications are listed in Table II. A grouping of the purpose versus method used was also carried out and the results are listed on Table III.

The largest number of publications (11) is intended to predict or identify the factors of the population using various methods: correlations, decision trees, statistical distributions, regressions, and tabulation of administrative data (Referred in Table III).

There are four publications that identify the population's poverty situation using the correlation method, and three publications to identify population characteristics through correlations, decision trees and gravity models. Other methods used are machine learning, map reduce, random coefficient model and statistical means for processing administrative records.

Upon grouping the methods vs. publication purposes, as shown in Table IV, the correlation and regression methods are the most commonly used methods. Correlation has been used to determine population characteristics, poverty situation and socio-economic factors. Regression has been used to determine population group sizes, economic growth, population mobility, proxies for social data and socio-economic factors.

TABLE II.

Method and Purpose of the Selected Articles

Id	Method	Purpose
1	Correlation	Socio-economic factors
2	Correlation	Socio-economic factors
3	Decision tree	Socio-economic factors
4	Machine learning	Vacancy risks
5	Log-normal distribution with cross-regional and time variations	Socio-economic factors
6	Random coefficient model	Poverty rates
7	Regression	Proxies for social data
8	Correlation	Poverty situation
9	Regression	Socio-economic factors
10	Correlation	Poverty situation
11	Regression	Socio-economic factors
12	Correlation	Poverty situation
13	Regression	Population mobility
14	Correlation	Population characteristics
15	Register-based statistics administrative	Census
16	Regression	Economic growth and population aging level
17	Map Reduce	Predict Income
18	Resource selection probability function	Population density
19	Bayesian methodology	Environmental and socio-economic representativeness
20	Correlation	Poverty situation
21	Decision tree	Population characteristics
22	Reduction methods	Socio-economic factors
23	Regression	Crowd sizes
24	Regression	Socio-economic factors
25	Building-population gravity model	Population characteristics

Id	Method	Purpose
26	Tabulation and cross-tabulation	Socio-economic factors
27	Machine learning	Socio-economic factors
28	Correlation	Socio-economic factors
29	Regression	Patterns of co-benefit behaviors

Still in Table IV and according to the timeline, the publications show that the use of methods that allows establishing the socio-economic factors is tied to technological development and the data available at that moment.

B. Evolution of Methods

Historically, the collection of data on general populations has been halted due to financial, technological, and even ethical factors [7]. For this reason, research is being carried out to look for less expensive alternatives, which take advantage of technology.

TABLE III.

PURPOSE VS METHOD OF SELECTED ARTICLES

Purpose	Method	2011	2013	2014	2015	2016	2017	Total
Census	Register-based statistics					1		1
Crowd sizes	Regression						1	1
Economic growth & population aging level	Regression					1		1
Environmental & socio-economic representativeness	Bayesian methodology					1		1
Patterns of co-benefit behaviors	Regression						1	1
Population characteristics	Correlation				1			1
	Decision tree						1	1
	Building-population gravity model						1	1
Population density	Resource selection probability function					1		1
Population mobility	Regression				1			1
Poverty rates	Random coefficient model		1					1
Poverty situation	Correlation			2	1	1		4
Predict Income	Map Reduce					1		1
Proxies for social data	Regression		1					1
Socio-economic factors	Correlation	1	1				1	3
	Decision tree		1					1
	Log-normal distribution with cross-regional & time variations		1					1
	Machine learning						1	1
	Reduction methods						1	1
	Regression			1	1		1	3
	Tabulation and cross-tabulation						1	1
Vacancy risks	Machine learning		1					1

TABLE IV.

METHOD VS PURPOSE OF SELECTED ARTICLES

Method	Purpose	2011	2013	2014	2015	2016	2017	Total
Bayesian methodology	Environmental and socio-economic representativeness					1		1
Correlation	Population characteristics				1			1
	Poverty situation			2	1	1		4
	Socio-economic factors	1	1				1	3
Decision tree	Population characteristics						1	1
	Socio-economic factors		1					1
Log-normal distribution with cross-regional and time variations	Socio-economic factors		1					1
Machine learning	Socio-economic factors						1	1
	Vacancy risks		1					1
Map Reduce	Predict Income					1		1
Random coefficient model	Poverty rates		1					1
Reduction methods	Socio-economic factors						1	1
Regression	Crowd sizes						1	1
	Economic growth and population aging level					1		1
	Patterns of co-benefit behaviors						1	1
	Population mobility				1			1
	Proxies for social data		1					1
	Socio-economic factors			1	1		1	3
Resource selection probability function	Population density					1		1
Tabulation and cross-tabulation	Socio-economic factors						1	1
Register-based statistics administrative	Census					1		1
Building-population gravity model	Population characteristics						1	1

The door-to-door visitation method implies that the data remains the same until the next visit. For this reason, in order to estimate the current situation of the population [4], simulations are used as a very useful tool to simulate the changes that occur in an annual basis.

Data mining methods have become the most widely used in financial and commercial sectors to identify customer patterns and behaviors [8]. The reviewed articles show that these methods are being used to look for relevant information of the

populations that allows socio-economic factors to be inferred, opening a new area of application for these methods.

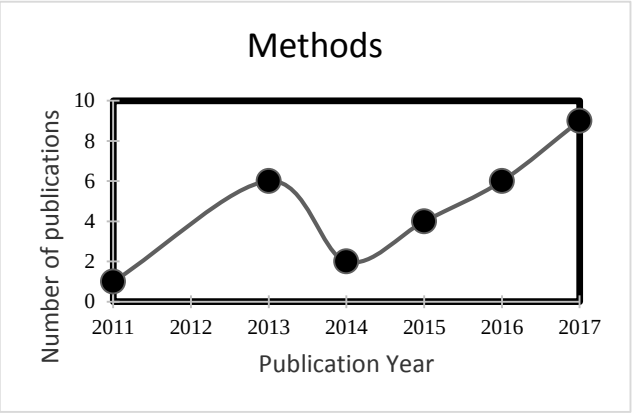

Fig. 1. Time line of the number of the articles per year, showing the different methods used to determine the socio-economic factors, poverty conditions and social status, beyond others.

C. Data Sources

Satellite data and images from Google Maps are used in several studies (e.g. [3], [9]–[12]) to establish behavior patterns among families in selected areas.

Another data source is the one generated from Twitter (e.g. [13]–[15]). The use of "tweet data" in poor sectors allows to estimate the population in those areas as well as the relationships that are established among Twitter users.

Mobile telephony is also one of the most used data sources to establish poverty maps used by academic contributions (e.g. [1]–[3], [16], [17]). The authors of these articles agree that this source has a huge potential for determining socio-economic status since it generates a large volume of data that would allow establishing relations between users by means of a "who-calls-whom" search.

D. Relationship between Data Sources and Methods

As shown in Table III, correlation is one of the most widely used methods for establishing the socio-economic status of the population. In the case of the use of telephone records, the analysis of certain characteristics of the records allows to establish the degree of poverty of the families. This is done by establishing networks of communication between users with similar socio-economic characteristics. This method establishes behavior patterns among users (assembling a virtual communication network). Thus, it is possible to correlate the way in which the users communicate and with whom they do it with poverty levels established for the user residence area [1], [2].

Twitter is also used on correlation studies by exploiting the geographical metadata inserted within each message at the time of being sent to the recipient. This establishes a communication network among the social network that allows the identification of patterns for inferring socio-economic factors. As indicated by Botta, Moat and Preis [15], the geographic location of the tweets allows inferring where a person resides (location under normal conditions) and his/her socio-economic status based on the characteristics of his/her contacts.

Correlations studies are also performed using geospatial data or satellite images. Similarities can be established for the population that lives in a given area while it is compared with areas already defined as being poor. Thus, it can be inferred that the population in a certain area has an established poverty level [11], [18].

All these studies highlight that, using correlation as a method; it is possible to use different data sources to infer the socio-economic factors of the population.

V. Validity Considerations

The literature review followed the guidelines proposed by Kitchenham. In Section III, it is mentioned the research planning process, identified the protocol used for searching the publication, the period for the bibliographic review, and the search engines used.

The research allows identifying the alternative methods used to collect data from families that is generated daily and continuously. This point is important to mention, since it would imply that it is possible to determine the socio-economic situation of families with the use of this data. Therefore, the research question was answered because we find alternative methods instead of the door-to-door approach.

In addition, the literary review also showed which methods were used to process the data and predict the socio-economic factors. Correlations and regressions are the most used methods.

VI. Discussion

The review shows various alternatives of data sources and methods that can be applied to infer the socio-economic conditions of the population.

However, it is necessary to point out that in none of the selected publications there is a comparison of the costs incurred by each one of these alternatives and what the equivalent door-to-door visits method implies. This could be a good measure of what the budgetary benefit of using these data sources and alternative methods is.

Another point that remains pending is determining whether the data used in each case was available for the studies carried out. There are publications in which the mobile service operators have made their data available to the scientific community to be analyzed and to determine its use.

The use of the methods for estimating poverty is subject to the variables chosen for the construction of the MPI or HDI indexes. It is important to establish if the available data can be related to the measures of these indexes and thus achieve a correct measurement.

VII. Conclusions

The results show there are alternative methods to collected socio-economic data from families instead of the door-to-door approach, using data, which is available from satellite data, images from Google Maps, the consumption of basic services

like mobile telephony and electricity, and from social networks. Each data source generates a large volume of data that allows establishing behavioral patterns and relationships of the population. Therefore, it is possible to estimate the current situation of the population and to make simulations of how the socio-economic situation will vary over time.

Based on our analysis, it is crucial to determine the kinds of variables that can contribute to establish a multidimensional index to measure the socio-economic status of the families. The input data influences the selection of the dimensions to compose the aggregated index.

For this reason, the information technology tools are used to simulate the changes, merging the data and choosing the variables that can contribute to find the better approach to measure the socio-economic factors of the families.

Correlation method allows analyzing characteristics of the collected data and establish the patterns of the population, thus it is possible to find similarities between people and to infer the socio-economic situation.

Regression method also is used to measure socio-economic factors, determining the population group sizes and associating with the economic growth of the country. It contributes to use the historical data to establish the behaviors of people and communities.

The combination of these methods and the IT tools can be a great contribution to predict socio-economic factors in less time and at lower cost

VIII. Future Studies

Methods and data sources that would allow inferring the socio-economic factors of the population have been considered in this study, but it is possible to consider the inclusion of other sources that have not been mentioned in the revised publications.

It is possible to consider institutional databases from registry offices, labor ministries, health ministries, or even internal revenue services, because they can contribute to better adjust the inference of the socio-economic factors. Some of this data is already available in open data repositories.

With the inclusion of phone and Tweeter logs, poverty indicators could be built with data that is updated periodically. This would mean that new methods could be applied to provide low-cost poverty indicators, which would represent significant savings for the government budget and would allow updated information for the creation and adjustment of public policies.

References

[1] C. Smith-Clarke, A. Mashhadi, and L. Capra, "Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks," *Proc SIGCHI Conf Hum Factors Comput Syst*, pp. 511–520, 2014.

[2] N. Pokhriyal, W. Dong, and V. Govindaraju, "Virtual Networks and Poverty Analysis in Senegal," 2015.

[3] N. Pokhriyal and D. C. Jacques, "Combining disparate data sources for improved poverty prediction and mapping," *Proc Natl Acad Sci*, no. 12, p. 201700319, 2017.

[4] A. Mathiassen, "Testing Prediction Performance of Poverty Models: Empirical Evidence from Uganda," *Rev Income Wealth*, vol. 59, no. 1, pp. 91–112, 2013.

[5] B. Anderson, S. Lin, A. Newing, A. B. Bahaj, and P. James, "Electricity consumption and household characteristics: Implications for census-taking in a smart metered future," *Comput Environ Urban Syst*, vol. 63, pp. 58–67, 2017.

[6] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature review," *Inf Softw Technol*, vol. 51, no. 1, pp. 7–15, 2009.

[7] D. Helbing and S. Balietti, *From social data mining to forecasting Socio-Economic crises*, vol. 195, no. 1. 2011.

[8] V. Atalay, S. Ustun, and S. Bulbul, "The Determination of Socio-economic Factors Affecting Student Success by Data Mining Methods," *2013 12th Int Conf Mach Learn Appl*, vol. 2, pp. 540–542, 2013.

[9] E. M. Weber, V. Y. Seaman, R. N. Stewart, T. J. Bird, A. J. Tatem, J. J. McKee, B. L. Bhaduri, J. J. Moehl, and A. E. Reith, "Census-independent population mapping in northern Nigeria," *Remote Sens Environ*, vol. 204, no. February, pp. 786–798, 2018.

[10] Y. Yao, X. Liu, X. Li, J. Zhang, Z. Liang, K. Mai, and Y. Zhang, "Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data," *Int J Geogr Inf Sci*, vol. 31, no. 6, pp. 1220–1244, 2017.

[11] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei, "Using Deep Learning and Google Street View to Estimate the Demographic Makeup of the US," vol. 0, 2017.

[12] D. Quercia and D. Saez, "Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use," *IEEE Pervasive Comput*, vol. 13, no. 2, pp. 30–36, 2014.

[13] R. O. Sinnott and W. Wang, "Estimating micro-populations through social media analytics," *Soc Netw Anal Min*, vol. 7, no. 1, 2017.

[14] C. J. Vargo and T. Hopp, "Socioeconomic Status, Social Capital, and Partisan Polarity as Predictors of Political Incivility on Twitter: A Congressional District-Level Analysis," *Soc Sci Comput Rev*, vol. 35, no. 1, pp. 10–32, 2017.

[15] F. Botta, H. S. Moat, and T. Preis, "Quantifying crowd size with mobile phone and *Twitter* data," *R Soc Open Sci*, vol. 2, no. 5, p. 150162, 2015.

[16] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science (80-)*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[17] B. Aragona and D. Zindato, "Counting people in the data revolution era: challenges and opportunities for population censuses*," *Int Rev Sociol*, vol. 26, no. 3, pp. 367–385, 2016.

[18] P. R. Choudhury and M. K. Behera, "Using Administrative Data for Monitoring and Improving Land Policy and Governance in India," *Proc 10th Int Conf Theory Pract Electron Gov - ICEGOV '17*, pp. 127–135, 2017.

Yasmina Vizuete-Salazar was born in Quito, Ecuador in 1973. She received the B.S. in informatics engineering from the Central University of Ecuador, Quito, in 2000. She received the M.S. degree in business information management from the Central University of Ecuador, Quito, in 2006. She also received the M.S. degree in auditing in quality management, in 2011 from the Technical Particular University of Loja (Loja). Currently, she is a doctoral student of the Informatics Doctorate Program at the National Polytechnic School of Quito, Ecuador.

She is a professional of computer science with experience in technological administration, formulation and management of projects, management information systems, redefinition of processes and strategic planning for the organization. From 2001 to 2015, she was Expert Data Analyst and Project Manager in the social area, with the specialty in conditional and unconditional cash transfers, with the Ministry of Economic and Social Inclusion of Ecuador. The last position was UNICEF consultant as a Project Manager, leading the development team and process for the delivery of the database-processing platform according to the guidelines and requirements made by the Ministry of Economic and Social Inclusion of Ecuador. The aim of the project was to ensure the proper functioning of the cash transfers without affecting the continuity of the business, with the highest standards of safety and quality. Currently, she is researching about the measurement of multidimensional poverty in Ecuador.

Marco Segura-Morales was born in Quito, Ecuador, in 1978. He received the B.S. degree in Computer Science from the Escuela Politécnica Nacional of Ecuador in 2004. He received his M.Sc. degree in Engineering Management in 2011 and his PhD degree in Systems Engineering in 2015 from the George Washington University, Washington D.C., USA.

He has more than fifteen years of solid experience in the fields of Software Engineering, Data Management, and Systems Engineering, leading the successful implementation of projects in local and international environments. Since 2016, he is a full-time professor at the Facultad de Ingeniería de Sistemas of the Escuela Politécnica Nacional, and his areas of research include Software Engineering and Information Systems. He has participated also as a revisor for the Latin American Journal of Computing and the IEEE ETCM.