



Quality Basic Service Indicators: A Twitter Approach

Quality Basic Service Indicators: A Twitter Approach

Yoselin Arvelaiz
Daniel Barráez
Manuel Solórzano



Programa de las Naciones Unidas para el Desarrollo (PNUD) en Venezuela

Jan Harfst, Representante Residente

Rosicler Gómez, Representante Residente Adjunta

Centro de Desarrollo Humano, Progreso Multidimensional y ODS

Daniel Barrález, Gerente

Coordinación editorial

Silvia Orozco Pabón, Asistente de Comunicaciones

Diagramación

Yessica Soto González

Diseño de portada

Esther Calderon

© 2021, Programa de las Naciones unidas para el Desarrollo

Todos los derechos reservados.



Abstract

This paper proposes a methodology to construct Twitter-based basic service quality indices. The methodology allows identifying failures in various basic services (electric power, residential water, public transport, domestic gas, basic waste collection, and automotive fuel), the type of failures (blackout, outages, and shortages), and their location by municipality or state where the disorder occurred. The indices constructed can be used as proxies for SDG indicators. The procedure was implemented using tweets reporting shortages or failures for residential utilities in Venezuela during a quartet. The goodness of fit of localization algorithms was 95%. The methodology is a less expensive alternative tool to traditional statistical surveys, can be implemented in real-time, and can be applied to any other country. The indices localization by municipality and state provides relevant information for policymakers to design and implement infrastructure plans for development, as well as identify vulnerable population and localities for implement humanitarian plans and other kinds of response. The failure indicators constructed can help to fill the gap in the sparse service failure statistics in many countries.



Quality Basic Service Indicators: A Twitter Approach

Yoselin Arvelaiz^a

Daniel Barráez^b

Manuel Solorzano^a

Primary SDG: Goal 1, “End poverty in all its forms everywhere”.

Secondary SDG: Goals 6, 7 and 11.

MAIN FINDINGS

- A methodology to construct Twitter-based basic service quality indices is proposed. The methodology was implemented to build the quality index for the following services: residential water, electrical power, domestic gas, public transport, and waste collection, as well as automotive fuel.
- The constructed indexes account for different types of outages (blackouts, outages, and shortages) as well as service complaints and are localized by municipality or state where the disruption occurred. The goodness of fit of localization algorithms was 95%.
- The residential water, electrical power, public transport, and waste collection are proxies

to SGD indicators 6.2.1, 7.1.1, 11.2.1, and 11.6.1, respectively. The aggregation of all basic service indices built is a proxy to SDG indicator 1.4.1 “Proportion of population living in households with access to basic services.”

- The methodology is a less expensive alternative tool to traditional statistical surveys and can be implemented in real-time.

POLICY RECOMMENDATIONS

- The failure indicators constructed can help fill the gap in the sparse basic service failure statistics in many countries and help the policymaker follow up the basic service quality perceptions.
 - The indices localization by municipality and state provides relevant information for policymakers to design and implement infrastructure plans for development and identify the vulnerable population and the localities for implement humanitarian plans and other kinds of response.

a) Universidad Central de Venezuela
b) UNDP Venezuela



INTRODUCTION

The quality of basic services has a clear impact on the quality of life of the population and the country's development level. The basic services are transversal to almost SDG goals, and the SDGs 1, 6, 7, and 11 has indicators on specific services. It is usual to use some indicators of basic services provision, like electricity and safe water, as indicators by multidimensional poverty indices and it is widely accepted that the good quality provision is needed for good governance. The monitoring of these services provides relevant information for policymakers to design and implement service infrastructure plans for development. Additionally, the disarray in the provision of these services does not affect the entire population equally, they have a greater impact on the vulnerable population, so monitoring quality basic services are helpful to identify the vulnerable population and localities for implement humanitarian plans and other kinds of respond.

This paper proposes and implement a methodology to build quality basic service indicator using tweets post reporting shortages, failures, or complaints for the following six basic services: residential water, residential electricity, domestic gas, gasoline, sanitation, and public transportation. This failure or dissatisfaction approach is justified. People are more prone to express dissatisfaction than satisfaction, especially when disorders happened. Failures and shortages are not unusual in developing countries and are strongly related to services quality. The Twitter approach is also justified. Although Twitter users are not a representative sample of the population, many relevant actors use this social network to express themselves on issues of public interest: heads and senior government officials, parliamentarians, political and social leaders, NGOs, media, influencers, and citizens in general.

These actors are fundamental in shaping people's perceptions on many issues, such as the quality of basic services and their options for improvement. Today, the importance of social networks in the discussion of public affairs is beyond doubt. The proposed indicators have some advantages on the traditional statistical surveys to measure basic service satisfaction. It is more rapid and less expensive analyzes tweets than survey collect and analysis information. Moreover, the twitter-based indicators would be carrying on in real-time. On the other hand, the failure indicators constructed can help to fill the gap in the sparse service failure or satisfaction statistics in many countries. Although the failure statistics are habitually record by the provider enterprises, they are not usually available to the public. Users' satisfaction provision statistics are more difficult to find because it requires specific customer surveys that the providers or other organizations do not necessarily conduct or, if they do, are not available to the public. An additional advantage of Twitter on surveys is that the computer code to process the tweets can be shared freely, as we do in this work.

Recently, works have successfully used social network analysis for development issues. The work (Solorzano, 2018) uses Twitter to study multidimensional poverty, and the report (UNDP, 2021) employs social media to analyze the information pollution impact on vulnerable populations during the covid-19 pandemic. In this work, as the cited papers, we employ a "text mining" collection of tools (Aggarwal & Zhai, 2012) (Python, 2021) (The R Project for Statistical Computing, 2020) (scikit-learn, 2021) to analysis the tweets text. Our methodology works in three steps. The first step in the methodology is to build the database of tweets which we will employ to identify the failures, using the keywords to extract the relevant

posts, like blackout, electrical, or water outages. In the second one, we identify the tweets that certainly report failures or shortages. In the last one, we try to identify the municipality or state where the disorder happened, mainly analyzing the post text or other user's information. Of course, the anonymity or any other user private information is fully respected by this work; we are only interested in the disorders reporting and place at municipalities or state levels, not in the users that report the failure. The methodology was implemented for mentioned basic services in Venezuela from December 2019 to February 2020. For each one of these services, an index and heat maps are built. The precision of the algorithms proposed to identify the failure localities analyzed using a statistical test based on quality control analysis have satisfactory results.

The main findings of this work are the following. Analyzing 1,671,869 tweets posted in Venezuela from 12/01/2019 to 02/28/2020, six quality basic services indicators were built: domestic water, domestic electric power, domestic gas, automotive fuel, public transport, and basic waste collection. The indicators were localized by municipality and stated with a 95% of goodness of fit of localization algorithms. Four of these indicators are proxies of SDG indicators:

Table 1

PROXY PROPOSED	SDG INDICATOR
Domestic water	6.2.1 Proportion of population using (a) safely managed sanitation services and (b) a hand-washing facility with soap and water
Domestic electrical power	7.1.1 Proportion of population with access to electricity
Public transportation	11.2.1 Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities
Waste collection	11.6.1 11.6.1 Proportion of municipal solid waste collected and managed in controlled facilities out of total municipal waste generated, by cities
The combination of all proposed indicators	1.4.1 "Proportion of population living in households with access to basic services."

The proposed methodology is a less expensive alternative tool to traditional statistical surveys, can be implemented in real-time and the constructed indicators can help to fill the gap in the sparse service failure or satisfaction statistics in many countries. The proposed algorithms rely on text mining techniques that are easily modifiable to recognize other municipality listings and even languages so the methodology can be used in any country

This working paper contains three additional sections. In the Methodology section, we present the algorithms for failure detection and regional failure localization and give some examples to show how they work. In the results section, we display the indices, their temporal evolution, the head maps and identify the most affected municipalities and states. The conclusions section ends the paper.

METHODOLOGY AND DATA

This section presents the methodology with which we try to identify different types of disorders in providing basic services that affect the population's quality of life. Disorders vary depending on the type of service, such as power or water outages, gasoline shortages, gas shortages resulting in irregular sale or distribution of gas in cylinders or the refilling, difficulties in public transportation, such as shortage or absence of transportation units. In what follows, to simplify the language, we will use the word "failure" to refer to all these different types of services disarrays.

Database building. The first step in the methodology is to build the database containing the "candidate" tweets to report failures or complains about the basic services. By candidate tweet we mean a post that probably report the failures, but we are not certain yet about the failure happened. This step is achieve using the key search word in the Twitter Api related with the basic services failures to extract the relevant posts, like blackout, electrical, or water outages. The size of the database used in this work is the 1,671,869 post and spans from 12/01/2019 to 02/28/2020. The selected tweets

were filtrated to consider only the tweets posted in Venezuela and focus on the following six basic services: water, electricity, gas, gasoline, sanitation, and public transportation.

Failure detection. After an exploratory data analysis of the database, a "dictionary" or word collection is built to identify failures for each basic service considered, so we have six dictionaries. Some examples words contained in the dictionaries are shown in Table 1. The failure detection algorithm works as follows. For a given post, if it contains a specific word *combination* in one of the dictionaries, a failure is detected, and the tweet is assigned in the respective failure service database. It is worth noting that the failure detection algorithm identifies a word sequence and not isolated words. For example, if the text post is "without water in Maracaibo", the algorithm recognizes the post as a fault because of the expression "without water in". However, if the text is "without money you can't buy a glass of water", the algorithm does not identify a failure, although the words "without" and "water" are in the text. As a tweet can express multiples failures services (for example, "without water and gasoline"), each post is compared with the six dictionaries and may be included in more than one failure database. Of course, a given post can be no included in any of the failures databases. At the end of the procedure, we have six new databases with the failure detected post. These databases also store the posting time of the tweets to follow the failure's temporal evolution. Each failure database provides the national-level failure information. The following step is attempted to identify the municipality or state where the failure happened.

Table 2. Keyword Dictionaries For The Different Categories

Electricity	Water	Transportation	Gas	Basic Waste Collection	Gasoline
Blackout		Subway	Cylinder	Garbage(s)	Gasoline
Blackouts	Without water	Van(s)	Gas	Colecction	Fuel
Light	Hydrocapital	Car(s)	Firewood(s)	Toilet(s)	Pump station
Electricity	Is in failure	Passenger(s)	Stove	Sewer(s)	
Electric(s)	Water outages	Metrobus	Stoves	Sanitation(s)	
Corpoelec		Bus	Charcoal	Waste(s)	
		Transport(s)	Charcoals	Excreta(s)	
		Motorcycle(s)		Scrap(s)	
		Automobile(s)		Residue(s)	
		Vehicle(s)		Dumpster(s)	
		Truck(s)			

Municipality/State identification. Then, we try to identify the municipality and state. The algorithm explores three information fields in the tweet, which are hierarchized in the following order:

1. The tweet's text can provide clear information about failure location, for example, a post with the phrases "without water in the Libertador municipality" or "electrical outages in neighborhood El Valle in Caracas".
2. The tweet broadcast location: the location information that the user's device emits in the tweet generation. This location may mention the country, state, or city of the location.
3. The user city, municipality, or state location at time of creating the user account.

The two last information sources are only available if the user authorizes them. The algorithm does not use personal identification; anonymity user information is respected. A list of municipalities and states, like a municipality dictionary, is employed to identify the location in a similar way to failure dictionaries detection.

A statistical test is performed to check the goodness of the fault location and detection algorithms. Three hundred tweets with their failure detection and location results are selected randomly, and it was verified by a human if the fault location and detection result are correct. The fault identification and location were successful 95% of the time. All statistical requirements was satisfied (Gutierrez & De la Vara, 2013).

Finally, we build the failure or quality index by aggregation at the national and regional level with its respective heat maps.



The following examples illustrate how the procedure work. Example A “without electricity in Maracaibo”, example B “no electricity at home” and example C “when I hear these arguments I feel like an electric shock in my head”.

1. *Database building.* The three examples are included in the Database because of the word “electricity” in the text of each tweet.
2. *Failure detection.* The example A is identified as a failure by the expression “without electricity in”. The example B is also identified as a failure by the expression “no electricity at”. The example C is not identified as a failure because of there is not any expression in the electricity failure dictionaries to match a failure. The isolated word “electricity” is not sufficient to identify a failure. The example C is excluded from the procedure.
3. *Municipality/State identification.* In the example A, with the expression “in Maracaibo” we identify the municipality Maracaibo of Zulia state in the list of municipalities and states. In the example B, no expression about the place is identified, we try to identify tweet broadcast location or location at time of creating the user account.

RESULTS

The six service failure databases contenting 515,439 tweets provide all the information needed to build the failure indices. Table 2 shows tweets distribution by service.

Failure results

Table 3: distribution of tweets by fault detection

	Electricity	Water	Transportation	Gasoline	Gas	Sanitation
Tweets total	212,847	125,259	63,921	62,433	26,290	24,689
% share	41	24	13	12	5	5

Each index is normalized to the daily average over the three months of study; the 100 value corresponds to the sample mean. In the following figures, we show the seven days moving average (Brockwell & Davis, 1996) at the national level.

Figure 1: Water Index

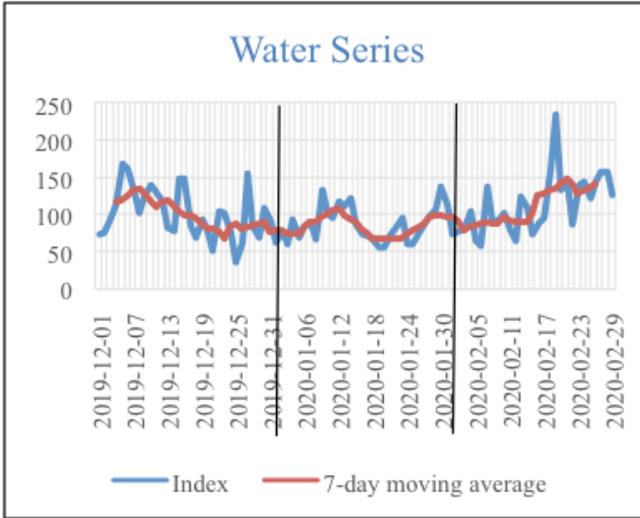
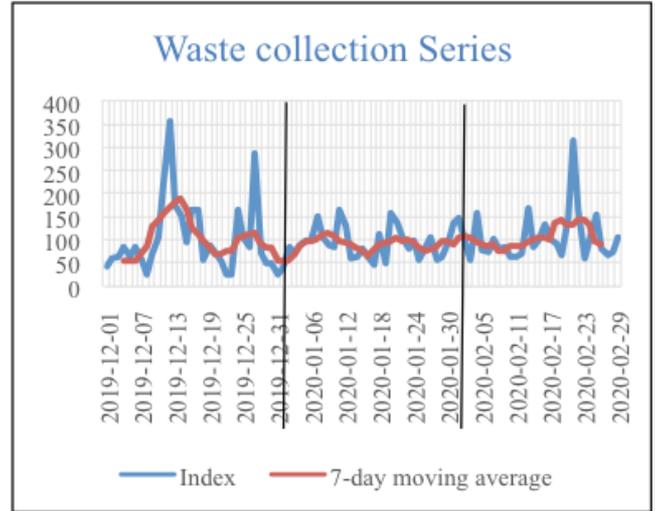


Figure 2: Sanitation Index



The range from 36 to 236 for water index in figure 1 is explained by a combination of effects the rainy season (many values below 100) and specific water cut service (values above 100). For the sanitation index in figure 2, reports show low intensity, except for the days 12-December-2019, 27-December-2019, and 21-January-2020. The moving average failure reports increasing in December, one of the months with the highest commercial activity.

Figure 3: Automotive fuel Index

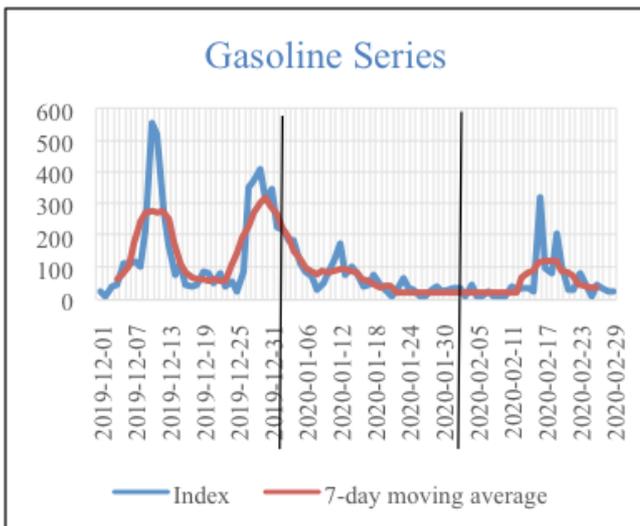
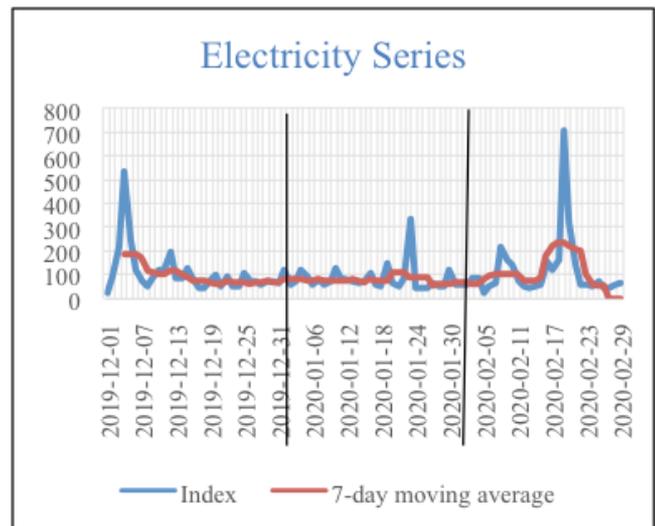


Figure 4: Electrical energy Index

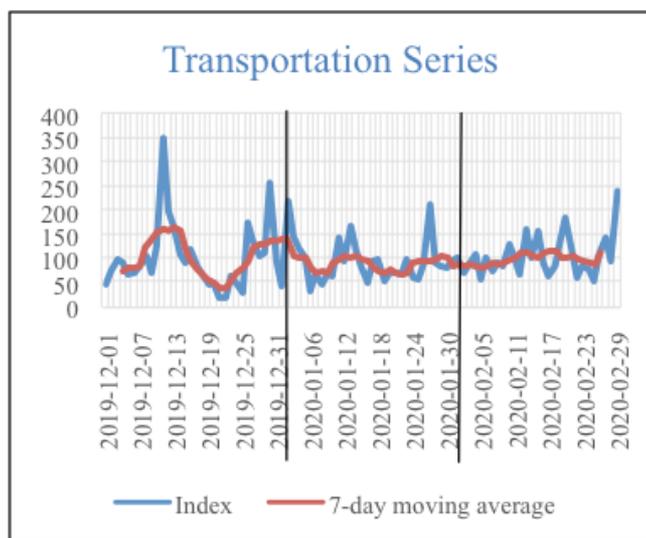
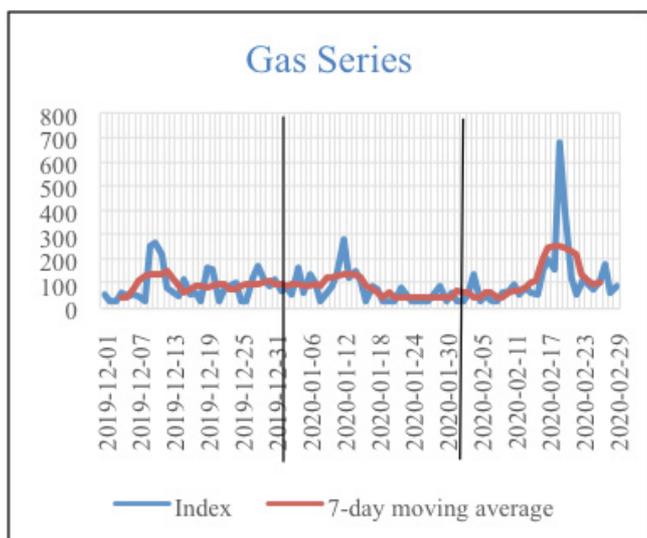


In Figure 3, the index identifies the gasoline shortage of mid-December 2019. Long lines at gas stations were observed. During January 2020, fuel supply partially normalized. In mid-February 2020, shortages come back again, especially in the border states of Táchira, Barinas, and Mérida, generating many complaints on Twitter. The behavior of the index is consistent with the news in the media.

We can see in figure 4 that this index catches the electrical blackouts. When the is above 300, there are outages reports at least in five states. Note that the gasoline and electrical indices are less volatile than water and sanitation.

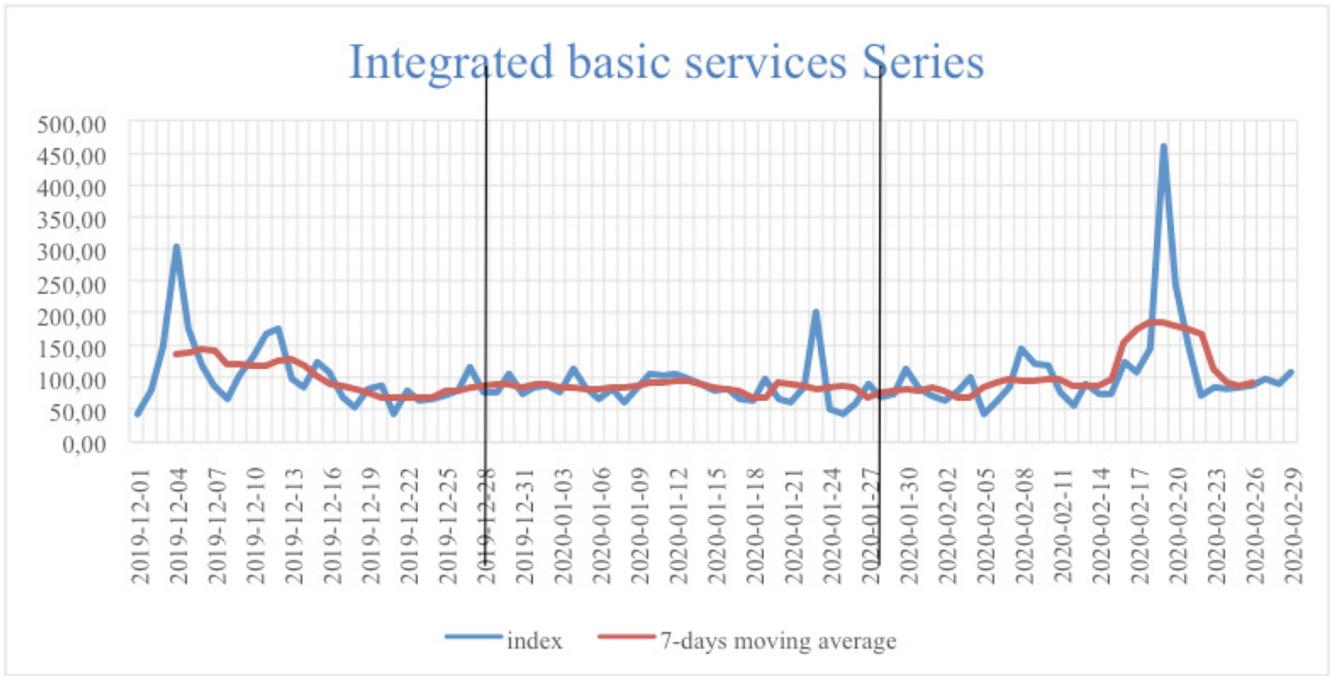
Figure 5: Gas domestic Index

Figure 6: Public transportation index



When we localize the gas index in figure 5, the municipalities with the highest sales or irregular distribution of gas cylinders reports are those with low incomes, as expected. It can see in Figure 6 that the transportation services index is highly correlated with the Automotive fuel index in figure 3 because almost all transport units in the country use fuel. As seen in figures from 1 to 6, all the indexes have a natural and intuitive behavior.

Figure 7: Integrated basic service Index



To build a proxy to SDG indicator 1.4.1 “Proportion of population living in households with access to basic services”, we make an index with the aggregation of all service failure databases, except the gasoline database, the integrated basic service index that is displayed in figure 7. The exclusion of gasoline shortage is justified because this service is not included in the metadata of 1.4.1 SDG indicator (UNSTATS).

LOCALIZATION RESULTS

The localization algorithm was able to identify the location 71% and 64% by state and municipality respectively. These results allow us to identify the areas with major difficulties, that we display in the following heat maps in the figures 8 and 9.



Figure 8

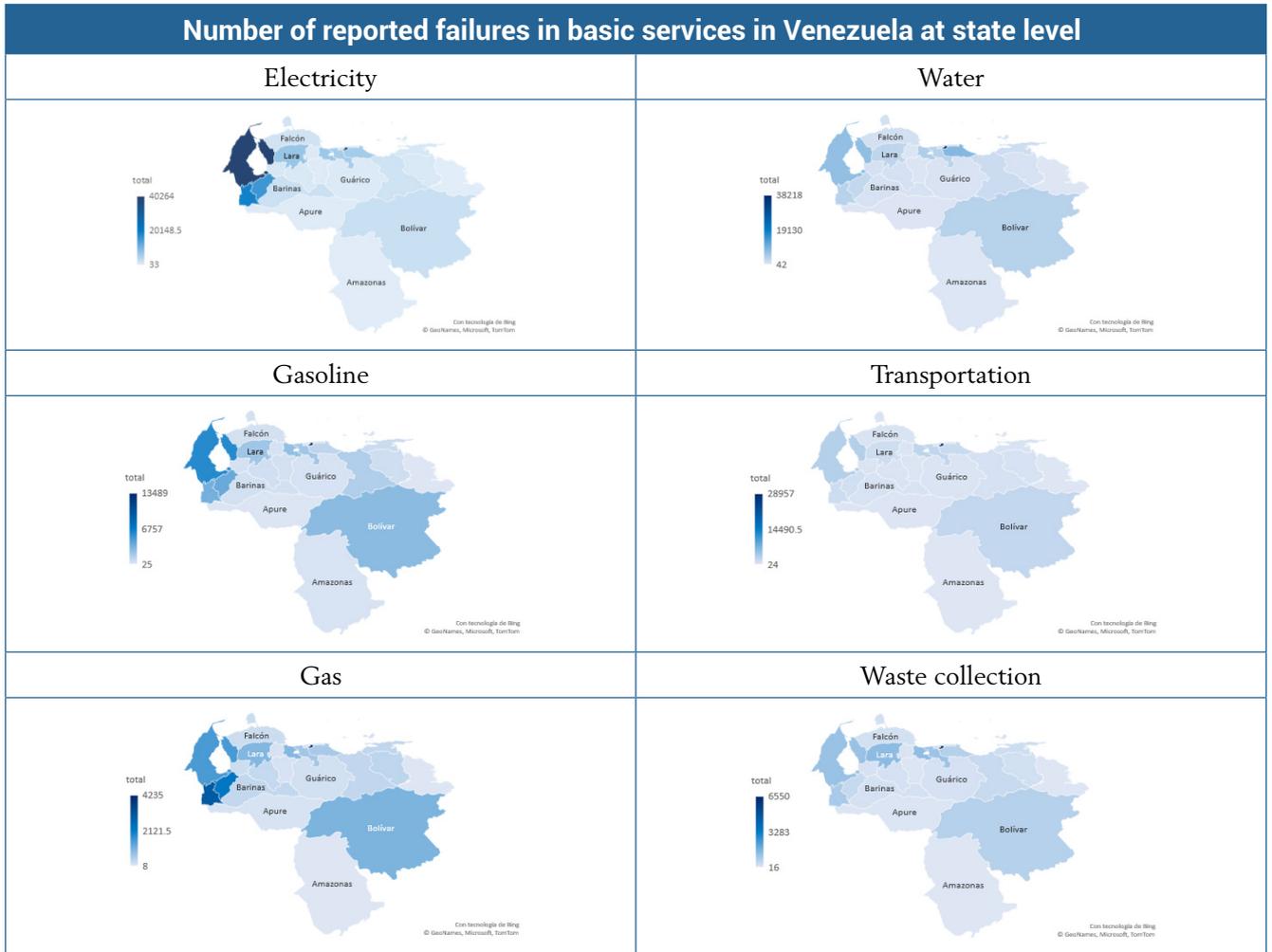
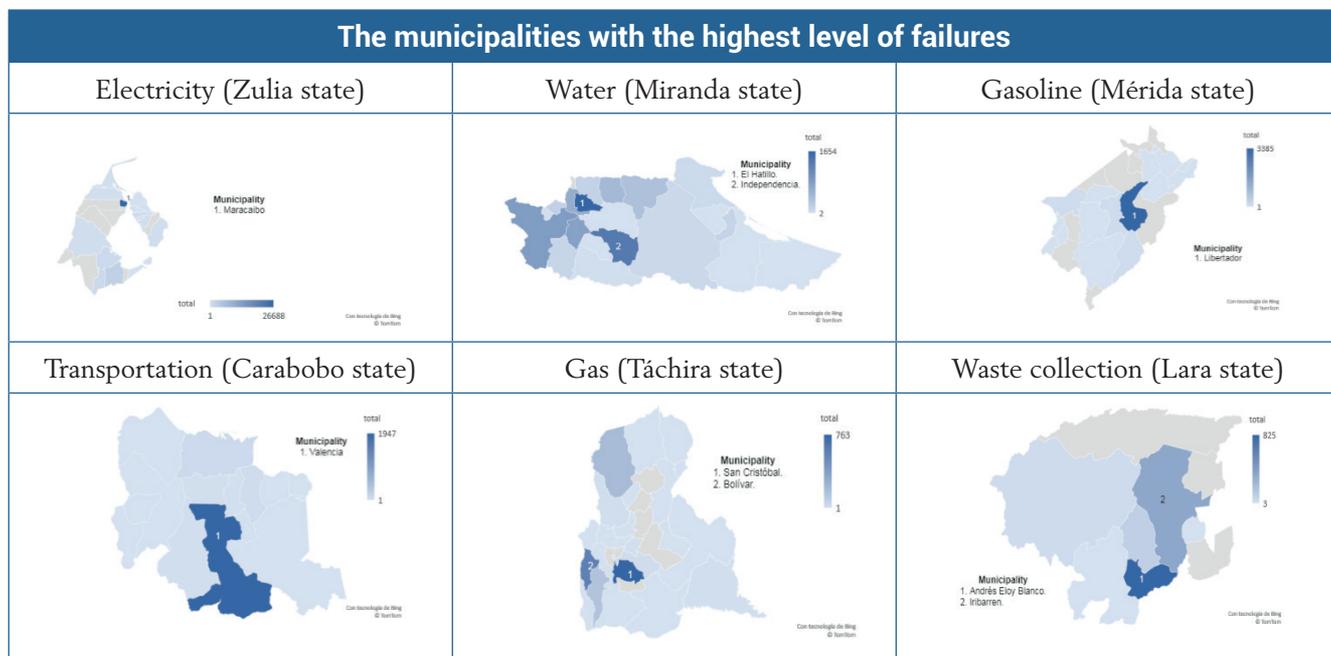


Figure 9



It is worth noting that failure detection is not a simple function of population or internet users in each state; we have a low correlation between them, as we can see in Table 4.

Table 4

Failure detection by service and population or internet users correlation						
	Electricity	Water	Gasoline	Public transport	Gas	Waste collection
Poblacion	0.64	0.41	0.5	0.28	0.4	0.39
Internet	0.57	0.62	0.56	0.49	0.48	0.57

On the other hand, even though we do not have real fault data to validate our failures indexes, the lack of such data is one of the motivations for this work, the public service disorders reported by twitter users are widely confirmed by media reports and informative tweets from service provider institutions.



CONCLUSIONS

The results show that proposed methodology allows identifying failures in various services (electric power, potable water, transportation, domestic gas, sanitation, and automotive fuel), the type of failures (blackout, outages, and shortages), and their location by municipality or state where the disorder occurred. The localization algorithm allows identify the most vulnerable municipalities and states. This identification is valuable information for policy makers in prioritizing areas where investments in infrastructure services and humanitarian respond plans are needed. The methodology is a less expensive alternative tool to traditional statistical surveys and can be implemented in real time. The failure and localization dictionaries employed by the algorithms, are easily modifiable and even translated into other languages, hence that the methodology can be applied in any country. The failure indicators constructed can help to fill the gap in the sparse service failure statistics in many countries.

BIBLIOGRAPHY

- Aggarwal, C., & Zhai, C. (2012). *Mining Text Data*. Springer Science+Business Media.
- Brockwell, P., & Davis, R. (1996). *Introduction to Time Series and Forecasting*. New York: Springer-Verlag, New York Inc.
- Gutierrez, H., & De la Vara, R. (2013). *Control Estadístico de la calidad y Seis Sigma* (Third ed.). México: McGraw-Hill/Interamericana Editores, S.A. DE C.V.
- Python. (2021). *re — Regular expression*. Retrieved from Python Documentation: <https://docs.python.org/3/library/re.html>
- scikit-learn. (2021). *scikit-learn*. Retrieved from Machine Learning in Python: <https://scikit-learn.org/stable/>
- Solorzano, M. (2018). *Las voces de Twitter sobre las dimensiones de pobreza en Venezuela: Un enfoque de Big Data*. Caracas: Central University of Venezuela.
- The R Project for Statistical Computing. (2020, November 18). *Text Mining in R*. Retrieved from Text Mining Package: <https://cran.r-project.org/web/packages/tm/tm.pdf>
- UNDP. (2021). *Narratives and Information Pollution on Vulnerable Groups During The COVID-19 Pandemic: Panama and Venezuela cases*. Retrieved from UNDP: https://www.ve.undp.org/content/venezuela/es/home/library/human_development/narrativas-y-contaminacion-informativa-sobre-grupos-vulnerables-.html
- United Nations. (n.d.). *Transforming our world: the 2030 Agenda for Sustainable Development*. Retrieved from UN, Department of Economic and Social Affairs: <https://sdgs.un.org/2030agenda>
- UNSTATS. (n.d.). *SDG Indicators*. Retrieved from Sustainable Development Goals: <https://unstats.un.org/sdgs/indicators/indicators-list/>
- UNSTATS. (n.d.). *SDG Indicators Metadata repository*. Retrieved from Sustainable Development Goal indicators: <https://unstats.un.org/sdgs/metadata/>

 @PNUD_Venezuela

 PNUD Venezuela

 PNUD Venezuela

 www.ve.undp.org



Centro de Desarrollo Humano,
Progreso Multidimensional y ODS
del PNUD en Venezuela



Programa de las Naciones Unidas para el Desarrollo
Av Fco. De Miranda, Torre Parque Ávila, piso 6, oficina 6-A.
Los Palos Grandes, Caracas, Venezuela.
Teléfono: (58) 212 208 4444
Fax: (58) 212 263 8179

