Building a Sanitary Vulnerability Map from Open Source Data. Argentina, 2010-2018

Germán Federico Rosati (✉ german.rosati@gmail.com)

Universidad Nacional de San Martín <https://orcid.org/0000-0002-9775-0435>

Tomás Alberto Olego

Bunge & Born Foundation

Antonio Vazquez Brust

Bunge & Born Foundation

Research

Keywords: sample, article, Health, Vulnerability, Census, Public Policy, Open Data

Posted Date: January 7th, 2020

DOI: <https://doi.org/10.21203/rs.2.20126/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at International Journal for Equity in Health on September 24th, 2020. See the published version at <https://doi.org/10.1186/s12939-020-01292-3>.

RESEARCH

Building a Sanitary Vulnerability Map from Open Source Data. Argentina, 2010-2018

Germán Rosati^{1,2,3*†}, Tomás Alberto Olego^{2†} and H. Antonio Vazquez Brust^{2†}

*Correspondence:

grosati@unsam.edu.ar

¹National Council of Scientific and Technical Research, CONICET, Godoy Cruz 2290, C1425FQB, Buenos Aires, Argentina

Full list of author information is available at the end of the article

[†]Equal contributor

Abstract

- **Background:** Designing public health policies to target the needs of specific places requires highly granular data. When geographic health statistics from official sources are absent or lacking in spatial detail, Sanitary Vulnerability metrics derived from Census and other georeferenced public data can be used to identify areas in particular need of attention. With that aim, a Vulnerability Map was developed, identifying areas with a substantial deficit in its population health coverage.
- **Methods:** Census, official listings of public health facilities and crowdsourced georeferenced data are used. The Vulnerability Index is built using dimensionality reduction techniques such as Autoencoders and Non-parametric PCA.
- **Main results:** The high resolution map shows the geographical distribution of a Sanitary Vulnerability Index, produced using official and crowdsourced open data sources, overcoming the lack of official sources on health indicators at the local level.
- **Conclusions:** The Sanitary Vulnerability Map's value as a tool for place specific policymaking was validated by using it to predict local health related metrics such as health coverage. Further lines of work contemplate using the Map to study the interaction between Sanitary Vulnerability and the prevalence of different diseases, and also applying its methodology in the context of other public services such as education, security, housing, etc.

First part title: Building a Sanitary Vulnerability Map from Open Source Data

Second part title: Argentina, 2010-2018

Keywords: sample; article; Health, Vulnerability, Census, Public Policy, Open Data

Background

The notion of Sanitary Vulnerability is related with health determinants. In other words, there are certain factors and variables that are strongly linked with the health status - in a broad fashion, in terms of biology, psychology and social aspects- of a person or population. These determinants can be of different types and if they are absent or weak, a state of vulnerability occurs.

We built an indicator that tries to identify places where high Sanitary Vulnerability prevails. That is, places where it is not possible to attain the minimum threshold in terms of access to health services. To do so, articulating an operative definition of Sanitary Vulnerability is an important first step that guides the later surveying and processing of available information.

The Map we built is highly disaggregated: the smallest unit of analysis is the Census Block, the smallest statistical unit in Argentina for which Census tabulations

are publicly available. In urban settings a Census Block can be as small as a single city block. In rural areas, where population density is low, Census Blocks are much larger and can cover areas of several square kilometers.

The paper is structured as follows:

- In Section 2 we discuss the operative definitions for the indicators shown in the Map.
- Section 3 explains the different analytical methods used for the estimation of the Sanitary Vulnerability Index
- In Sections 4 and 5 we report the sources of information used, and the procedures and techniques used to process and analyze the raw information.
- Section 6 describes the transformations used to combine multiple indexes into a single indicator.
- Finally, we discuss limitations and further avenues of research in Section 7.

Definition of Sanitary Vulnerability

There are well known determinants linked to health status, related to unequal access to health services, both in terms of quantity and quality. The development of an index such as the one proposed here, aims at quantifying these inequalities to complement and contextualize other sources of health related information. It also raises awareness of particularly vulnerable regions. The estimation of a Sanitary Vulnerability could help identify and eventually prioritize “hot spots” on places with poor access to sanitary services that demand focused public health policy.

The presence of inequalities in the access to health services has been thoroughly studied and documented. Not all population strata show the same chances of accessing medical treatment. In general, the most vulnerable sectors and/or residents of low population density areas have a disadvantage in terms of access to health services. ([1], [2]).

In specialized literature there is a differentiation between the notion of “inequality in access” with the notion of “vulnerability”, which is linked to a heightened risk to develop certain illnesses, or being exposed to certain factors of environmental risk.

An obstacle when trying to quantify “vulnerability” comes from the necessity to consider multiple factors that could explain inequalities in the access to the health system. In [3] the main dimensions and indicators linked to notions of vulnerability and inequality related to health care access are summarized.

Table 1 shows the main indicators from a number of previous studies analyzed in [3] such as those linked to experiencing poverty, belonging to a minority, suffering from chronic mental or physical illnesses, and lacking medical insurance. While these indicators are defined at the individual level, there are other factors affecting Sanitary Vulnerability that are prominently correlated with the environmental realm. [4].

We evaluated many of the aforementioned indicators. Based on the availability of public data sets providing both nation-wide coverage and high spatial resolution, our approach for the Sanitary Vulnerability Index integrates two dimensions:

- Access to public -that is, universal, Government funded- health services: for this dimension, we used physical proximity to public health facilities. A data set with the addresses of public health facilities throughout Argentina was

Vulnerability Indicators	Papers	Percentage
Poverty	21	91.3 %
Ethnic or Racial minority	18	78.3 %
Chronic physical or mental illnesses	12	52.2 %
Lack of Health Insurance	8	34.8 %
Old age	6	26.0 %
Incarceration	3	13.0 %
Migration	3	13.0 %
Low level of education	3	13.0 %
Residence in underserved areas	2	8.7 %
Unemployment	1	4.3 %
Widowhood	1	4.3 %
Homelessness	1	4.3 %

Table 1: Sanitary Vulnerability Indicators from [3]

compiled, based on data extracted from official listings published on websites ran by Municipal, Provincial and National authorities. These addresses were georeferenced to obtain their geographical coordinates. Then, for every Census Block in the country, a number of locations were sampled at random and then the shortest route to its nearest public health facility was calculated.

- Socioeconomic Status (SES): to construct a SES index we used Census information. As it will be detailed later, the indicator is based on Census 2010 data at the individual level. A series of relevant variables (education level, poverty indicators, etc.) were combined using variational autoencoders, a method for dimensionality reduction based on neuronal nets (see section 5).

Methods

The two dimensions mentioned before were combined into a georeferenced Sanitary Vulnerability Index that was projected onto the territory as a Sanitary Vulnerability Map. As discussed, the Map was designed to help identify areas with a potential deficit in sanitary services; that is, places where the local population access to health services lie below a minimum threshold. With this in mind, we built a metric to sort and classify the different zones in terms of their potential deficit.

Data sources

To construct the Map, four different classes of sources were employed:

- Census data (National Census of Population and Homes and Households 2010 [5] [6])
- Georeferenced polygons representing Census Blocks.
- Names, service level and geocoded addresses of public health facilities: Public hospitals, health centers, and health posts ($n = 16.564$)
- Street grids (national and provincial routes, municipal roads) used to find the shortest path between households and health facilities.

All information was aggregated at the Census Block level, an areal unit.

In the following sections the procedures employed to process and analyze these diverse sources are explained.

Distance to Public Hospitals and Health Centers

To build distance metrics, a unified, nation-wide health facility location data set was compiled by pooling different official data sources:

- National data set of Hospital and Primary Care Centers: based on the National Database of Hospital and Primary Care Centers (Base Nacional de Hospitales y Centros de Atención Primaria), which was compiled through the Sanitary Information System of Argentina (SISA) and published at the Argentine Republic Spatial Data Infrastructure (IDERA) site at <http://catalogo.idera.gob.ar>. This was considered the "master" data set, to be expanded with information from additional sources

Other health facility location sources:

- SUMAR health centers: SUMAR is a National program aimed at improving access to health coverage for people with no health insurance. Their official website lists thousands of public health facilities that were "scraped"^[1] and combined with those found in other sources.
- Hospitals and health care centers listed at the National Program for Sex Health and Responsible Reproduction website: Health facility lists, published as separate files for every Province, were downloaded and geocoded.
- Local government sources: information from several sources at the Provincial or Municipal level, usually from Provincial Ministries of Health websites, was also extracted and integrated in order to identify facilities not found in any of the National level sources.

Once the data was identified, the extraction, cleanup, geocoding and deduplication process was :

- 1 Data was downloaded either as a single consolidated file (when available) or iteratively scraped from its website.
- 2 Records with missing or unusable addresses were removed.
- 3 Health facility addresses were georeferenced using the Google Geocoding API.
- 4 All geocoded locations were overlaid
- 5 For each service-level category (hospitals, health centers, sanitary posts) geocoded facility locations were projected on a map, adding 100 meter buffer zone around each one. Where locations overlapped -i.e. their buffer areas intersected- those records were considered duplicates and only one kept. In that way, it was possible to identify and unify records obtained from different sources that used different addresses or names referring to a single site.

As an example, health facility locations in Buenos Aires City, obtained from different sources, are shown. The data set was clearly enriched by combining sources: while there are some overlapping sites corresponding to different data sets (red and blue dots), some others represent facilities that were found only on one of the sources (yellow and green dots).

Figure 1: Comparison between master data set (SISA) and a municipal source. Buenos Aires City

^[1]The term "scrapping" or "web scrapping" refers to a set of techniques and technologies used to extract information from websites when no simple data extraction mechanism -like a download link- is provided.

Classification of health facilities according to their service level

After their location was established, public health facilities were classified according to their service level. Establishing their service level is important as not all health facilities are equal. Access to a hospital that offers specialized medical procedures improves the chances of receiving a better treatment, emergency care, and potentially better outcomes compared to being attended in a basic health services center (sometimes referred as a sanitary post in Argentina). As the data sources did not use an standardized system to indicate health service levels, an expert review was conducted with the help of physicians experienced in public health administration, who assessed the data and sorted the facilities into three categories, in decreasing order of complexity:

- 1 Hospital
- 2 Health Center
- 3 Sanitary Post

After removing a number of facilities that do not belong to any of the categories (i.e. nursing homes, administrative offices, etc.), 15903 records were retained from the 16654 distinct places originally identified by merging disparate sources.

Travel time to the nearest public health facility

Physical accessibility has been recently used as a proxy to unequal distribution of goods and resources. A recent paper by [7] shows that there is a positive correlation between aggregated time access to urban agglomerates and several common socioeconomic indicators: income level, educational level, child mortality rates, etc.

In order to define an accessibility metric based on household proximity to Public health facilities, distance and travel time to the nearest facility was calculated for every Census Block in Argentina.

To solve the issue of the varied shapes and areas of Census Blocks -quite different across the country, with a stark contrast between high and low density areas-, distances and travel times to the nearest facility was calculated by sampling many locations at random inside each Block. The procedure is as follows:

- 1 Five points are randomly chosen for each Census Block
- 2 For each point, the nearest health facility is identified
- 3 For each point, travel time and distance traversing the local street grid to its nearest facility is calculated.
- 4 Distances and times are averaged.

The procedure was repeated for facilities in each of the service level categories (see previous section). To identify the nearest facility, a kNN algorithm -k nearest neighbors- algorithm was used to compare the coordinates of the samples locations and the facility addresses.

To find the optimal route between the two points we used Open Source Routing Machine (OSRM), a high performance routing system that can find the shortest path between locations through the public road network, by foot and by car ([8]). Detailed road data was obtained from OpenStreetMap ([9]), an open, global, crowd-sourced database whose quality has made it a frequently used source in mobility studies([10], [11])

Travel time on foot

As an accessibility measure, walking distance and time (travel on foot) was chosen over travel by car as there is evidence -at least for certain medical treatments- that the distance by foot to the health facility is a good predictor of treatment outcome.

Results from this study of Baltimore City clients attending outpatient drug addiction treatment programs suggest that having to travel more than 1 mile from the treatment center reduces medical patients' chances of completing treatment by almost 50%, after controlling for the effects of demographic variables and type of drug problem. Moreover, living more than 4 miles away from treatment decreases the expected length of treatment by almost 13 days in comparison to clients traveling less than 1 mile.[1]

Moreover, [1] indicates that distances longer than a mile (approximately 1.6 km) affect treatment outcome, highly increasing the chances of a patient abandoning rehabilitation treatment. When the distance exceeds 4 miles (6.4 km), the expected duration of a treatment increases by almost two weeks.

For a very reduced number of Census Blocks it was not possible to determine the travel path on foot to the nearest facility (10 out of 52406 Blocks analyzed). For each of them, visual inspection of aerial imagery indicated that they comprise either rural areas near lakes or estuaries, or large plots dedicated to agriculture, where no paths are present. In those instances, the largest distance found for neighboring Blocks was used instead.

The following maps show mean traveling times on foot to the nearest hospital, health centers and sanitary posts in Argentina.

Figure 2: Travel time (on foot) to different classes of health facilities. Census Block Average.

The results for hospitals and health centers are quite similar. In contrast, the map of sanitary posts shows a clearly different pattern. There is a significant portion of the Country where travel times to the nearest sanitary post are very high. Closer inspection indicates that there exists an inverse relation between distance to hospitals / health centers and distance to sanitary posts.

This difference could be explained by two non mutually exclusive reasons:

- As local Governments decide the location (even the mere existence) of sanitary posts, the absence of sanitary posts could be explained by lack of local interest in setting up posts, or lack of means.
- Incomplete sources: that is, additional posts exist but are absent from official listings of health facilities.

Even with these caveats regarding the completeness of the data, travel distance to the nearest sanitary post was still included as part of the accessibility indicator. Since the indicator is built based it on a median function, the effect of potential outliers is minimized.

For each Census Block Δ_r is the median distance for all sampled locations inside it, to the nearest facility of each service level class: hospitals (Hosp), health centers (HCter) and sanitary posts (SPost):

$$\Delta_r = MED(\Delta_{Hosp_1}, \dots, \Delta_{Hosp_5}, \Delta_{HCTer_1}, \dots, \Delta_{HCTer_5}, \Delta_{SPost_1}, \dots, \Delta_{SPost_5}) \quad (1)$$

Socioeconomic Status -SES-

The second dimension of Sanitary Vulnerability represents a summarized measure of different SES characteristics of the population. This measure represents the population social and economic conditions, considered a key factor for the differences in sanitary risk across different populations.

Data from the 2010 National Census of 2010 at the individual level was available. Since SES is usually measured at household level, we calculated an index for each head of household in the data set from the Census. To calculate the index the following variables were used, ordinally classified:

Variable	Unit
Dwelling property condition	Dwelling
Construction Material Quality	Dwelling
Basic Services Access Quality	Dwelling
Dwelling Construction Quality	Dwelling
Overcrowding	Household
Basic Necessity Index	Household
Head of Household Educational Level	Household
Number of Unemployed persons in Household	Household
Domestic Service Existence	Household
Employment Status	Individual (head)
Educational Level	Individual (head)

Table 2: Indicators used for SE Index

To construct the SES index a "thermometer encoding" for ordinal variables was used. Let N be the number of cases and $v1, \dots, vI$ the variables. For each variable v_i , there are K_i categories. We created the following codified variables $x_{k_i}^{(i)}$ for each variable v_i and for each k_i category where $2 \leq k_i \leq K_i$. For each case j with $1 \leq j \leq N$ we have:

$$x_{k_i}^{(i)}(j) = \begin{cases} 0, & \text{if } v_i(j) < k_i \\ 1, & \text{if } v_i(j) \geq k_i \end{cases} \quad (2)$$

For example, dwelling construction quality is an ordinal variable, based on the INMAT index^[2]. INMAT values, ranging from 1 (lowest construction quality) to 4 (highest quality), were represented as:

^[2]INMAT is the acronym for the Construction Materials Index, an official statistic. It reflects the quality of the materials with which houses are built (considering the

	INMAT original	INMAT_2	INMAT_3	INMAT_4
1		0	0	0
2		1	0	0
3		1	1	0
4		1	1	1

Table 3: Example of thermometer encoding

In this example, $INMAT_1$ would be the baseline or reference category.

Construction of a SES index

The final index was computed using an autoencoder, a technique of dimensionality reduction[12], based on neuronal networks architecture. In short, an autoencoder has the objective to find a representation of input data (encoding) usually with the aim of reducing dimensionality. Usually, autoencoders learn to replicate inputs in outputs. This seems a trivial problem, but when certain restriction are introduced it becomes a complex task. For example, the size of the representation can be limited, or "noise" could be added to the inputs. Autoencoders try to learn this identity[13].

An autoencoder has two elements:

- an encoder (or recognition network) that converts inputs to an internal representations, followed by a
- decoder (or generative network) that re-converts the internal representation to outputs.

It has an architecture analogous to a Multi-Layer Perceptron, but the number of neurons in the output layer has to be equal to the number of neurons of the input layer. The trained model has a logistic function as a final layer and regularizes using a dropout (of 0.5) in the intermediate layers to improve generalization performance.

Figure 3: **Autoencoder model used**

The estimated model was the following:

$$h_1 = \tanh(W_1(x * r_1) + b_1)$$

$$h_2 = \tanh(W_2(h_1 * r_2) + b_2)$$

$$h_3 = \tanh(W_3(h_2 * r_3) + b_3)$$

$$h_4 = \tanh(W_4(h_3 * r_4) + b_4)$$

$$h_5 = \tanh(W_5(h_4 * r_5) + b_5)$$

$$\hat{x} = \sigma(W_6 h_5 + b_6)$$

predominant material used for floors and ceilings). It takes into account building strength, proofing, and finishing.

where W_i is the coefficient matrix of layer i , and b_i is the mean of layer h_i . In this way, the log-likelihood for each case can be written as

$$\Pi(\hat{x}, x) = \sum_{i=1}^I \sum_{k_i=2}^{K_i} w_{k_i}^{(i)} \left(x_{k_i}^{(i)} \log(\hat{x}_{k_i}^{(i)}) + (1 - x_{k_i}^{(i)}) \log(1 - \hat{x}_{k_i}^{(i)}) \right)$$

where $w_{k_i}^{(i)}$ is defined for each variable v_i with its K_i categories as:

$$w_{k_i}^{(i)} = \frac{1}{K_i} \cdot \left(\frac{\sum_{\ell} K_{\ell}}{I} \right)$$

The loss function was defined as a weighted likelihood:

$$L(\hat{x}, x) = \operatorname{argmin}_{W, b} \sum_{j=1}^N \Pi(\hat{x}(j), x(j))$$

The model was trained with ADAM [13] defining batches of data resampled with repetition over the empirical distribution to ensure convergence.

Since the model include the population data (the data set is the Census itself), the aim was the generation of a descriptive measure. So, the model evaluation criteria was the weighted average of the probability of each variable-category:

$$\text{Error}(\hat{x}, x) = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^I \sum_{k_i=2}^{K_i} \frac{1}{\sum_{\ell} K_{\ell}} e^{x_{k_i}^{(i)} \log(\hat{x}_{k_i}^{(i)}) + (1 - x_{k_i}^{(i)}) \log(1 - \hat{x}_{k_i}^{(i)})}$$

The final model preserves 85% of the total information of the input.

In this way, from h_3 on, each household head, and therefore, each household, is classified with a value resulting from the autoencoder, which will be called s_i , SES.

Using SES, a socioeconomic indicator was created from an aggregated measure for each Census Block using data from the head of household head and it SES. So, for each household head i with SES s_i living in the Census Block r with a population of n_r head of households η as

$$\eta_r = \frac{1}{4} Q_{.25}(s_{r1, \dots, n_r}) + \frac{1}{2} Q_{.5}(s_{r1, \dots, n_r}) + \frac{1}{4} Q_{.75}(s_{r1, \dots, n_r}) \quad (3)$$

where Q_p is the p quartile and η_r the result of applying the summary measure known as "Tukey Trimean", which delivers a compromise between robustness and efficiency, compared to the mean.

Figure 4: **SSE by Province density plot**

The previous density plots show that the index seems to capture well known disparities between Provinces:

- The city of Buenos Aires presents a distribution clearly skewed to the right (higher values of SES index)
- Provinces such as Chaco, Formosa, Jujuy, Salta, show left skewed distributions (lower SES index).

Estimation of a Sanitary Vulnerability Index

The index of Sanitary Vulnerability design took into consideration the following dimensions:

- Socioeconomic Level
- Access to health center

As said, Socioeconomic level can be considered a determinant of the population situation related to sanitation and social resources (such as education, and to conditions of overcrowding and habitability of the dwelling). Some of these resources are related individual and social trajectories and histories. Nevertheless they also have a strong relationship with residence place. In other words, the Socio Economic Level is linked to the distribution problem of and therefore it is an intensive variable.

To construct the index of Sanitary Vulnerability it is necessary to combine both variables at Census Block level. The strategy adopted was a Principal Components Analysis (PCA), searching for a linear combination of the variables that maximizes explained variance.

When the variables show atypical distributions, for example multimodal, asymmetric or fat-tailed, PCA interpretation can be difficult: the method is sensitive to the variable scaling. [14]. One possible solution is to transform variables using ranges of data [15] [16].

Therefore, following [17], each X_j was standardized using a rankit transformation:

$$rankit(X_{i,j}) = \frac{r_j(x_{i,j}) - 0.5}{n} \quad (4)$$

For each observation i of each variable X_j the rank is calculated $r_j(X_{i,j})$ -ranging from 1 and n -, 0.5 units are subtracted and it is divided by the total number of registries n .

As a result, differences in units between variables, changes in scale, displacements and monotone transformations are removed.

Figure 5: SES without and after the rankit transformation. City of Buenos Aires and Chaco Province.

Figure 6: Distances to health centers, before and after the rankit transformation. CABA and Chaco Province.

In the previous plot it can be shown the rankit transformation effect in SES $-\eta_r-$ and access to health units $-\Delta_r-$.

It is easy to note that rankit is simply a non parametric estimator of the c.d.f. Following Egger et al. [18] y a Han y Liu [19], an Semiparametric Principal Component Analysis was performed.

Algorithm 1 Semiparametric Principal Component Analysis

```

1: procedure S-PCA(X)
2:   for j en variables do
3:     for i en cases do
4:        $z_{i,j} = \Phi^{-1}\left(\frac{r_j(x_{i,j}) - 0.5}{n}\right)$  ▷  $\Phi^{-1}$  is the inverse of a Gaussian c.d.f.
5:     end for
6:   end for
7:   calculate  $\Sigma_z$  ▷ Covariance Matrix
8:   solve for  $U, S$  such as  $\Sigma_z = \frac{1}{n}US^2U^t$  ▷ SVD
9:   return  $ZU^t, S$  ▷ Eigenvectors and Eigenvalues
10: end procedure

```

To combine both variables, it was necessary to calculate Spearman's correlation between rankits [19]. Then the non-parametric correlation matrix was decomposed. The first autovector absorbed 72% of the variability and was oriented in the opposite direction of mutual growth. This first autovector was considered the optimal combination between both variables. And after successive application of non parametric transformations, and index tolerant of contamination and above all independent to the scaling characteristics of each entry variable.

The aim was to obtain an index between 0 and 1 whose distribution were homogeneous for all units. There was a new application of a c.d.f - type estimated through log-splines using AIC as criterion of regularization over the main direction, deriving in this way the index VS_r for each Block.

In difference to rankit, when modeling the distribution with log-spline, relative distance are respected.

Figure 7: Sanitary Vulnerability Index

In the first graph of figure 9, result of S-PCA is shown in the space of Z , and the colors show the value of the resulting index. In the second graph, the result is observed in the space of X .

Results

Finally in the to graphs below, the index is shown as a function of SES index and distances, respectively.

Figure 8: Sanitary Vulnerability Index. Argentina.

Now, this visualization needs to be taken with care due to the scale employed. In fact, the previous map seems to represent an expected situation where two zones can be delimited:

- “Central” Region: Province of Buenos Aires, City of Buenos Aires, and the large urban agglomerates of each Province, characterized by higher values of Sanitary Vulnerability.

- The rest of the country, mainly in less populated regions, with critical values.

When aggregating information one level higher (census fraction) the result is slightly different.

Figure 9: **Sanitary Vulnerability Index at Census Fraction level. Argentina.**

This map shows a mixed version: the regions previously detected are not so stark. They are observed in the central zone with critic values. In the non-central zone, there are better ranked sectors.

Spotting certain areas (but keeping the level of disaggregation by Census Block), there are patterns that are invisible in the aggregated map. If the aggregated map is displayed for the Provinces of Chaco and CABA the following result arises:

Figure 10: **Index of Sanitary Vulnerability: CABA (left.), Chaco (right)**

Then, in CABA, it is possible to detect higher Sanitary Vulnerability pockets, and in Chaco there are places presenting satisfactory values for this indicator.

Validating the Sanitary Vulnerability Index

This section will attempt a first validation of the Sanitary Vulnerability Index (SVI) by complementing it with other statistical sources.

If the SVI is valid, it should be safe to assume that higher proportions of population without medical coverage are associated with higher SVI values. In fact, the medical coverage of the population, that is, the access they have to the health system should show a negative correlation with the index values

In this regard, it is necessary to clarify one point: the variable linked to the type of medical coverage of the population was not used for the SVI because the Population Census did not include a question for it in its basic questionnaire (which allows disaggregation at the Census Block level), only in the expanded one.^[3] The latter allows for a disaggregation only at the Department (an aggregation level comparable to the County level in the United States) level. That is why we will compare both distributions (SVI and health coverage) at the Department level.

This exercise will have the following characteristics:

- 1 it will be based on census radio information
- 2 different aggregate measures will be calculated at the department level
- 3 it will be compared with the proportion of the population with medical coverage (that is, prepaid or social work) at the departmental level

The following plots show the geographical distribution of both variables and a scatterplot.

^[3]The expanded questionnaire is applied to a sample of the total population and contains occupational, health coverage and other variables. For more information, consult [5]

Figure 11: **Sanitary Vulnerability Index, by population proportion with no health care coverage (county level)**

Figure 12: **Sanitary Vulnerability Index, by population proportion with no health care (county level)**

When analyzing the correlation between both variables, it is observed that at higher average values of the SVI, greater proportions of medical coverage in the population are observed: Pearson's R correlation is about 0.70.

However, it can be seen that the departments with higher values of the SVI, correspond to a greater dispersion in the non-medical coverage of the population. In turn, these departments are the smallest in terms of population size.

Discussion and Conclusions

Having validated the Sanitary Vulnerability Map's capacity to predict local health coverage, it seems to be a valuable tool for policymaking, as it provides a spatially granular estimate of the population's access to public health services.

A current limitation of the Sanitary Vulnerability Index is that it uses data from the latest Argentina Census in 2010. As several years have passed since the Census was completed, having access to up-to-date information is crucial to improve its effectiveness. Because of that, the Index should be re-calculated as soon as future Census data is made public.

Additionally, the Index could also be improved by including other data highly granular spatial data sources such as climate and environmental metrics.

As it is, the Map has opened up two interesting further lines of work. First, Sanitary Vulnerability can be considered a traversal dimension that affects the evolution, transmission and prevalence of different pathologies. So, a further line of work could use the Sanitary Vulnerability Map as an input to study the correlation of Sanitary Vulnerability with the spread of infectious diseases or the prevalence of chronic ones. Secondly, the methodology can be applied to different contexts. For example, the estimation of travel distance to health facilities can be replicated for other Public services: education, security, social assistance, housing, etc. This opens up the possibility to generate highly disaggregated information that could be useful to take cost-effective decisions in the allocation of public resources.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable. Data has been obtained from public sources.

Availability of data and material

This study was a re-analysis of existing data, which is openly available at locations cited in the reference section. The code used during the current study are available from the corresponding author on reasonable request

Competing interests

The authors declare that they have no competing interests.

Author's contributions

All authors made the same contributions to this paper.

Acknowledgements

The authors would like to thank to the Bunge y Born Foundation directives, Alejandra Candia and Brenda Walter for the material support for this research. Also, to Guillermo Bozzoli who assisted with the article translation.

Author's information

Germán Rosati is sociologist, Assistant Researcher at CONICET and Quantitative Methods professor at University of San Martín. His research interests are social and agrarian structure, machine learning and natural language processing applications to social sciences. Tomás A. Olego is a professor of Public Opinion Methodology at University of Buenos Aires and a research fellow at the Bunge & Born Foundation, in Argentina. His area of interest is the study of political phenomena combining computational and statistical tools. H. Antonio Vazquez Brust is a professor of Urban Analytics at Torcuato Di Tella University and a research fellow at the Bunge & Born Foundation, in Argentina. His area of interest is the study of urban phenomena combining computational tools and large scale data sets.

Author details

¹National Council of Scientific and Technical Research, CONICET, Godoy Cruz 2290, C1425FQB, Buenos Aires, Argentina. ²Bunge I& Born Foundation, 25 de Mayo 501 6° Piso, C1425FQB, Buenos Aires, Argentina. ³Institute for High Social Studies, University of San Martín, 25 de Mayo y Francia, 1650 San Martín, Argentina.

References

1. Beardsley, K., Wish, E., Fitzelle, D., O'Grady, K., Arria, A.: Distance traveled to outpatient drug treatment and client retention (2003)
2. Timyan, J., SJ, G.B., Measham, D.M., Ogunleye, B.: Access to care: more than a problem of distance. (1993)
3. Grabovschi, C., Loignon, C., Fortin, M.: Mapping the concept of vulnerability related to health care disparities: a scoping review. *BMC Health Service Research*, 13, 94 (2013).
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626765/pdf/1472-6963-13-94.pdf>
4. Prüss-Üstün, A., Neira, M.: Preventing Disease Through Healthy Environments: a Global Assessment of the Burden of Disease from Environmental Risks. World Health Organization, ??? (2016)
5. INDEC: Censo nacional de población, hogares y viviendas 2010. base de datos. definiciones de la base de datos. Technical report, Instituto Nacional de Estadística y Censos (2011).
<http://200.51.91.245/redarg/CENSOS/CPV2010rad/Docs/base.pdf>
6. Censo Nacional de Población y Vivienda 2010. Cuestionario básico.
<https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-6>. Accessed: 2018-04-14
7. Weiss, D., Nelson, A., Gibson, H.e.a.: A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 25 (2018)
8. Huber, S., Rust, C., et al.: Calculate travel time and distance with openstreetmap data using the open source routing machine (osrm). *Stata J* **16**, 416–423 (2016)
9. OpenStreetMap contributors: Planet dump retrieved from <https://planet.osm.org>.
<https://www.openstreetmap.org> (2017)
10. Haklay, M.: How good is volunteered geographical information? a comparative study of openstreetmap and ordnance survey datasets. *Environment and planning B: Planning and design* **37**(4), 682–703 (2010)
11. Juran, S., Broer, P.N., Klug, S.J., Snow, R.C., Okiro, E.A., Ouma, P.O., Snow, R.W., Tatem, A.J., Meara, J.G., Alegana, V.A.: Geospatial mapping of access to timely essential surgery in sub-saharan africa. *BMJ global health* **3**(4), 000875 (2018)
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, ??? (2016).
<http://www.deeplearningbook.org>
13. Gerón, A.: Hands-On Machine Learning with Scikit-Learn and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly, ??? (2017). <http://shop.oreilly.com/product/0636920052289.do>
14. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics. Springer, ??? (2006)
15. Baxter, M.J.: Standardization and transformation in principal component analysis, with applications to archaeometry. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **44**(4), 513–527 (1995)
16. Bacon-Shone, J.: Ranking methods for compositional data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **41**(3), 533–537 (1992)
17. Solomon, S., Sawilowsky, S.: Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods* (2009)
18. Egger, B., Kaufmann, D., Schönborn, S., Roth, V., Vetter, T.: Copula eigenfaces - semiparametric principal component analysis for facial appearance modeling. In: Magnenat-Thalmann, N., Richard, P., Linsen, L., Telea, A., Battiato, S., Imai, F.H., Braz, J. (eds.) *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 1: GRAPP, Rome, Italy, February 27-29, 2016.*, pp. 50–58. SciTePress, ??? (2016). doi:10.5220/0005718800480056.
<https://doi.org/10.5220/0005718800480056>
19. Han, F., Liu, H.: Scale-invariant sparse pca on high-dimensional meta-elliptical data. *Journal of the American Statistical Association* **109**(505), 275–287 (2014). doi:10.1080/01621459.2013.844699

hospitales

centros de salud

postas sanitarias

vulnerabilidad
sanitaria

vulnerabilidad
sanitaria

4 km

vulnerabilidad
sanitaria

100 km

vulnerabilidad
sanitaria

SVI mean by county

% health care

Figures

Figure 1

Comparison between master data set (SISA) and a municipal source. Buenos Aires City

hospitales

centros de salud

postas sanitarias

Figure 2

Travel time (on foot) to different classes of health facilities. Census Block Average.

Figure 3

Autoencoder model used

Figure 4

SSE by Province density plot

Figure 5

SES without and after the rankit transformation. City of Buenos Aires and Chaco Province.

Figure 6

Distances to health centers, before and after the rankit transformation. CABA and Chaco Province.

Figure 7

Sanitary Vulnerability Index

Figure 8

Sanitary Vulnerability Index. Argentina.

Figure 9

Sanitary Vulnerability Index at Census Fraction level. Argentina.

Figure 10

Index of Sanitary Vulnerability:. CABA (left.), Chaco

Figure 11

Sanitary Vulnerability Index, by population proportion with no health care coverage (county level)

Figure 12

Sanitary Vulnerability Index, by population proportion with no health care (county level)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [biblio.bib](#)
- [bmcartbiblio.sty](#)
- [Manuscript.tex](#)
- [bmcmathphys.bst](#)
- [bmcart.cls](#)
- [Manuscript.bbl](#)