Joint evolution of access to water of urban and rural populations in South America through Compositional Data Analysis.

F.A. Quispe-Coica[1] and A. Pérez-Foguet[1]
[1]Research Group on Engineering Sciences and Global Development, Dept. of Civil and Environmental Engineering, School of Civil Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain.
filimon.alejandro.quispe@upc.edu; agusti.perez@upc.edu

Summary

In the international water and sanitation sector, it is usual to use multivariate statistical methods to monitor and report the overall progress of access to water and sanitation. However, the methods used do not take into account the compositional characteristics of the data. Recently, to overcome this problem, the application of multivariate temporal interpolation models that include this characteristic has been proposed.

On the other hand, it is usual to carry out analyzes separately between the urban and rural sectors (WHO/UNICEF, 2015), even though both parties form a whole (the general population of the country). The disaggregated work does not allow to see the crossed influence between the population evolution and the levels of service. In addition, according to (Cohen, 2004), the question arises whether disaggregated work is a good option or not, given that the definition of both categories is not homogeneous, which makes international comparisons difficult.

Therefore, this work focuses on the comparative analysis between the joint and disaggregated treatment of indicators of access to water, in urban and rural contexts, considering that the level of service is divided into four categories. Different temporal interpolation models are applied to the balances defined between the parties through the isometric logratio (ilr) transformation (Egozcue et al. 2003). The identification of outliers is included through Mahalanobis distance.

The preliminary results obtained with the data from the countries of South America show that the adjusted models may depend on the joint or disaggregated approach. The quality metrics Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) confirm that the best adjustments are obtained in one way or another depending on the case. The outliers are different in each situation.

Key words: CoDa aggregate and disaggregate, models, water, outliers.

1 Introduction

In the period (Year: 2000-2015) of the Millennium Development Goals (MDGs), the global monitoring and reporting of access to water and sanitation has been in charge of the Joint Monitoring Program (JMP), between the World Health Organization (WHO) and United Nations International Children's Emergency Fund (UNICEF). After 2015, in the context of the Sustainable Development Goals (SDG) (Year: 2015-2030), the JMP continues to make the global reports (WHO/UNICEF, 2017).

For the preparation of the report, the main sources of information are the household survey, the census, and the administrative data compiled by governmental and non-governmental entities (JMP, 2018). With this information, estimates are made separately for the urban and rural sector, while national estimates are generated as the weighted average of the two using population data (JMP, 2018). Due to this, the joint evolution in the time series is not visualized, nor the cross-influence between the evolution of the population and the levels of service.

On the other hand, the definitions of the urban and rural are in question, when they differ in many countries of Latin America and the Caribbean (Dirven et al., 2011). Therefore, the JMP is forced to rely on the existing definitions of the member countries of the SDGs. A clear example of this is Peru, in which rural population is considered to be one that does not exceed 2,000 thousand inhabitants (DS. N°031-2008-VIVIENDA, 2008); Meanwhile, in Chile it is considered rural, *a human settlement with a population less than or equal to 1,000 inhabitants, or between 1,001 and 2,000 inhabitants where more than 50% of the population that declares to have worked is dedicated to primary activities* (INE, 2018). It is in this scenario that the question arises whether disaggregated work is a good option or not, given that the definition of both categories is not homogeneous, which makes international comparisons difficult(Cohen, 2004).

Another situation is presented in the data used for national, regional and global estimates; when it is appreciated that water and sanitation service levels have been disaggregated into four parts for the urban and rural sector. Being these, the services improved (Piped on premises, Other improved) and unimproved (Surface water, Other unimproved), for the case of water. While for sanitation they are: Improved, Shared, Open defecation, Unimproved. Also called water and sanitation ladders (WHO/UNICEF, 2015). What shows that these data have compositional characteristics, because they are part of a whole, and the sum is a constant (Aitchison, 1986). Post-2015, they remain compositional, but in five water and sanitation ladders (WHO/UNICEF, 2017). Therefore, they have to be addressed as such. For this, there are already antecedents in the sector that show the importance of taking into account the compositional nature of the population data, when the estimates of access to water sanitation and hygiene (WASH) are modeled (Pérez-Foguet et al., 2017). In addition, the statistical analysis in the world of compositional data (CoDa) is a good alternative to evaluate and visualize the aggregate information of the urban and rural sector, in the time series.

This new method in the sector follows statistical procedures of compositional data, for which, certain disadvantages must be overcome, one of them being the presence of zero values; because the transformations carry proportions in which this is not possible. To address this, in literature there are different methods of work (Josep Antoni Martín-Fernández et al., 2011; Palarea-Albaladejo et al., 2008; J. A. Martín-Fernández et al., 2003; Templ et al., 2016).

Therefore, the purpose of this study is to assess the cross-influence between the evolution of the population and service levels; value whether disaggregated work is a good option or not. For which, statistical methods for compositional data will be used. In addition, the presence of outliers in the models will be analyzed, with the purpose of assessing in which of them are generated mostly. The hypothesis that estimates can vary in certain situations if they have different SBPs will be tested. The generalized additive model (GAM) will be used to generate and compare the data.

The article will be organized as follows: The following section presents the work methodology in aggregate and disaggregated data of the sector. In section 3, the results are analyzed and discussed. In section 4, the conclusions of this work and the future challenges in the WASH sector are described.

2 Materials and Method

The information used for the study is available on the JMP platform (JMP, 2017). In this there are data on access to water, sanitation and hygiene (WASH), ordered by sector (Urban and Rural).

2.1 Selection of countries and indicators.

The countries in the analysis are: Bolivia, Colombia, Ecuador, Paraguay, Peru and Uruguay. They respond to water access indicators with compositional characteristics, with a minimum data quantity of twelve (Uruguay) and a maximum of twenty-six (Peru). Other countries are not considered for lack of one or more variables that prevent the formation of compositions. In addition, they do not comply with the minimum amount of data to use GAM (Fuller et al., 2016). As for the population, they are the same ones with which the JMP makes the estimates of each country. In this is the population divided into urban and rural.

The indicators of analysis in the period of the MDGs have been related to the monitoring of the population that accesses improved and unimproved water, made up of water and sanitation ladders. In this study, we followed the same classification and used the same population data from JMP for estimates of access to services, according to the indicator.

The services of access to improved and unimproved water are classified as follows:

Access to improved water: They are supplied by networks and other improved forms.

- **Piped water** (Xu,r$_1$)**:** They are considered like this, the access of the water by public network inside the house, public network outside the house but inside the building, public tap and others.
- **Other improved sources** (Xu,r$_2$)**:** Tank truck, and other forms of access to improved water that is not piped.

Access to unimproved water: They are supplied from surface sources and other unimproved sources.

- **Surface water** (Xu,r$_3$)**:** According to the country they can be, river, spring, irrigation channel, and other.
- **Other unimproved sources** (Xu,r$_4$)**:** Other non-surface water sources

The disaggregated analysis has a composition of four parts for the urban sector and four for the rural sector. While the aggregate analysis between urban and rural, they carry compositions of eight parts. These are represented as follows:

$$x_{r\_1} + x_{r\_2} + x_{r\_3} + x_{r\_4} = 1 \qquad \text{Eq. (1)}$$

$$x_{u\_1} + x_{u\_2} + x_{u\_3} + x_{u\_4} = 1 \qquad \text{Eq. (2)}$$

$$x_{r\_1} + x_{r\_2} + x_{r\_3} + x_{r\_4} + x_{u\_5} + x_{u\_6} + x_{u\_7} + x_{u\_8} = 1 \qquad \text{Eq. (3)}$$

Where: Xu,r$_1$= Piped water; Xu,r$_2$=Other improved sources; Xu,r$_3$ =Surface water; Xu,r$_4$ = Other unimproved sources.
*Xu,r: Urban or rural value.

To make CoDa in eight parts, the population of the urban and rural sector has been multiplied with the proportions in CoDa of four parts (Eq. (1) y Eq. (2)), as appropriate. Then, the population expressed in eight parts was divided among the total population. Thus forming the compositions shown in Eq. (3).

2.2 Balances and transforms

Balances are defined in a group of parties that have access to improved and unimproved services.

Given this premise, there is a need to divide the aggregate analysis (eight-part CoDa) into two scenarios (Figure 1-*V3* y Figure 2-*V4*). The first will be when the main proportion is defined by sectors (proportion of people who have access to water in the rural sector among people who access water in the urban sector (see *V3*)). After this, the proportion criterion between improved and unimproved services is applied internally. In the second scenario, balances are made between access to improved and unimproved services, in the total (see *V4*); unlike the previous one, the group of parts is not separated by sector. Regarding CoDa of four parts, in both scenarios the same balance is maintained (*V1* and *V2*)

Next, balance "V" is made according to Egozcue and Pawlowsky-Glahn, (2005).

Scenario 1: Balance of eight parts, is carried out by sectors (Rural and Urban).

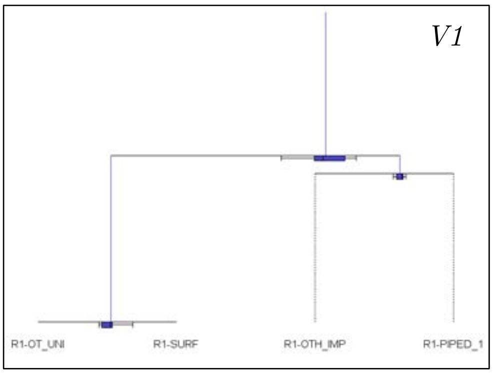
- Rural: Group of parts between the improved (Xr_1, Xr_2) and the unimproved (Xr_3, Xr_4). Balance *V1*, shows the partitions.
- Urban: Group of parts between the improved (Xu_1, Xu_2) and the unimproved (Xu_3, Xu_4). Balance *V2*, shows the partitions.
- Aggregate data urban and rural: Group of parts between rural (Xr_1, Xr_2, Xr_3, Xr_4) and urban (Xu_5, Xu_6, Xu_7, Xu_8). Internally, they will be classified as a group of access to improved and unimproved water. Balance *V3*, shows the partitions.

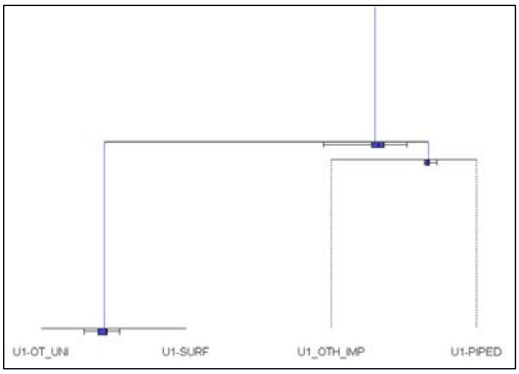
$$V_1 = \left(\begin{array}{cccc} x_{r1} & x_{r2} & x_{r3} & x_{r4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{array}\right)$$

$$V_2 = \left(\begin{array}{cccc} x_{u1} & x_{u2} & x_{u3} & x_{u4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{array}\right)$$

V2

$$V_3 = \left(\begin{array}{cccccccc} x_{r1} & x_{r2} & x_{r3} & x_{r4} & x_{u5} & x_{u6} & x_{u7} & x_{u8} \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 & 0 & 0 & 0 & 0 \\ +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & +1 & +1 & -1 & -1 \\ 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{array}\right)$$





V3

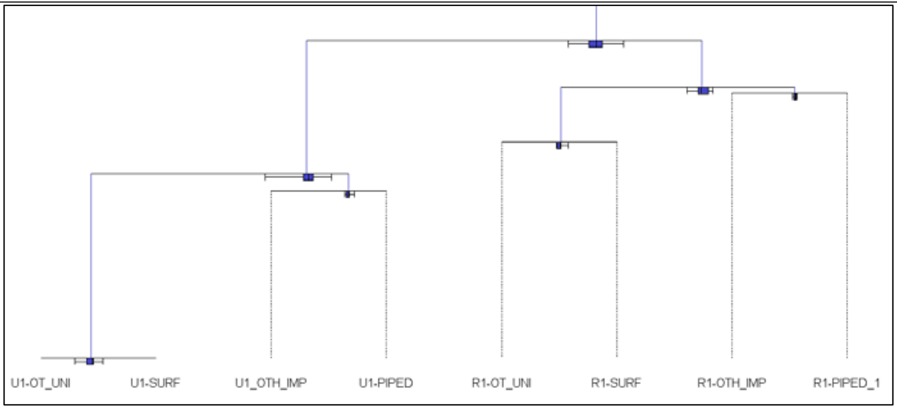


Figure 1: Balance of Colombia – Scenario 1.

Scenario 2: Balance of eight parts is made between access to improved water and not improved, in the total of variables.

- Rural: Group of parts between the improved (Xr_1, Xr_2) and the unimproved (Xr_3, Xr_4). Balance *V1*, shows the partitions.
- Urban: Group of parts between the improved (Xu_1, Xu_2) and the unimproved (Xu_3, Xu_4). Balance *V2*, shows the partitions.
- Aggregate data urban and rural: Group of parts between rural (Xr_1, Xr_2, Xu_5, Xu_6) and urban (Xr_3, Xr_4, Xr_7, Xr_8). Internally, they will be classified as a group of access to improved and unimproved water. Balance *V4*, shows the partitions.

$$V_1 = \left(\begin{array}{cccc} x_{r1} & x_{r2} & x_{r3} & x_{r4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{array}\right)$$

$$V_2 = \left(\begin{array}{cccc} x_{u1} & x_{u2} & x_{u3} & x_{u4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{array}\right)$$

$$V_4 = \left(\begin{array}{cccccccc} x_{r1} & x_{r2} & x_{r3} & x_{r4} & x_{u5} & x_{u6} & x_{u7} & x_{u8} \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & 0 & 0 & -1 & -1 & 0 & 0 \\ +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & +1 & 0 & 0 & -1 & -1 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{array}\right)$$

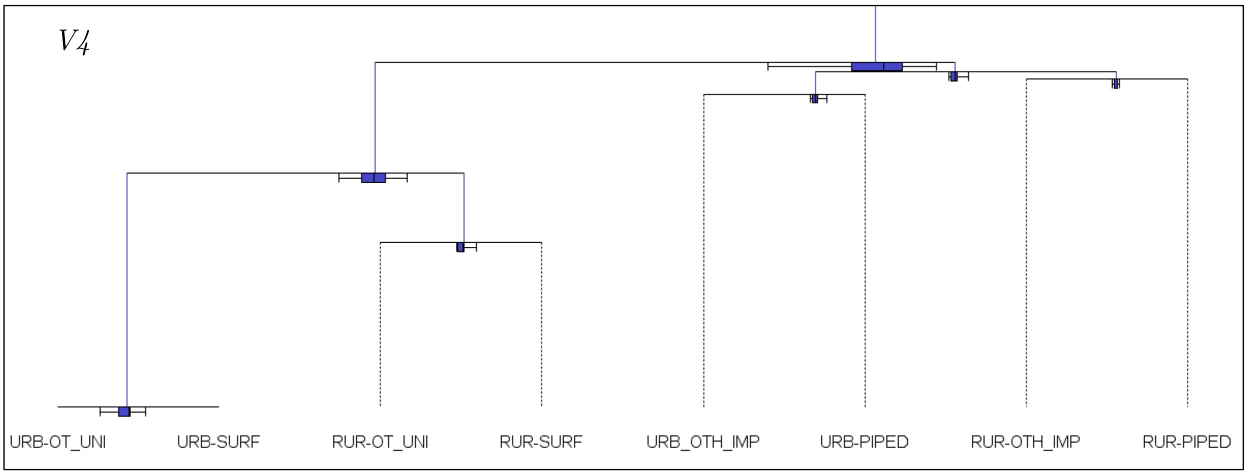


Figure 2: Balance of Colombia – Scenario 2.

After balances, isometric log-ratio transformations are performed (J. J. Egozcue et al., 2003). For this, the following equation is used.

$$\mathrm{Y}^{t} = ilr = \sqrt{\frac{r \times s}{r+s}} \ln \frac{g_m(X_r+)}{g_m(X_s-)} \qquad \text{Eq. (4)}$$

r = Number of positive variables in the balance V.

s = Number of negative variables in the balance V.

g_m(-) = It is the geometric mean of the variables.

2.3 Metrics of analysis

The existence of outliers is detected using the robust Mahalanobis distance (Filzmoser et al., 2008). Subsequently, it is compared according to the amount of data that are considered outliers in four-part CoDa and eight-part CoDa. Preliminary analysis removing outliers, show variation in estimates with aggregate and disaggregated data. Therefore, the comparison between eight-part CoDa and four-part CoDa is made without removing these values. This in order that you are comparatives are carried out under similar conditions.

For time series estimates, GAM (k = 4) is used. This is supported by the presence of non-linear data (Fuller et al., 2016). In addition, the flexibility of GAM helps to better adapt to the data in the time series.

Estimation results are compared between four-part CoDa (urban and rural) and eight-part CoDa. For this, the values estimated in eight-part CoDa are multiplied by the total population (the same population used to make eight-part CoDa in item 2.1) and then disaggregated into four-part CoDa for urban and rural. The result of these is compared with the estimates made in a disaggregated manner. This comparison is made with RMSE metric. If the value is zero, then the values estimated in eight-part CoDa generate same values when working independently in rural and urban areas (four-part CoDa). Otherwise, estimated values differ between the two. For the latter, the predictive capacity between CoDa 8 and CoDa 4 is evaluated and compared. For this, the NSE indicators are used and comparisons are made with the observed values in the rural and urban sectors (Table 3).

All this has been carried out and implemented in R Core Team (2018), using the following statistical packages: **nlme, compositions and mgcv**, by Pinheiro et al., (2018); Boogaart et al., (2014) & Wood, (2017), respectively. To treat data with outliers in CoDa, the **robCompositions** statistical package was used (Templ et al., 2011).

3 Results and discussions

3.1 ILR transformations and outlier

Scenario 1 and scenario 2, presented the same number of outliers. According to SBP raised in Scenario 1, it shows that metrics of R-adj and the ilr of Figure 3A, and 3B are similar to six transforms of Figure 3C (ilr2, ilr3, ilr4, ilr5, ilr6 y ilr7). This occurs because they maintain proportions of CoDa 4. The only one that changes is the ilr1 of Figure 3C, because it is represented with proportions between rural and urban water (*V3*). This can be a factor so that estimates in aggregated and disaggregated data do not have significant variation, as shown in Figure 4G, 4H. On the other hand, this does not occur when the balance (*V4*) varies (Figure 4J, 4K). For this, a more exhaustive analysis will be needed.

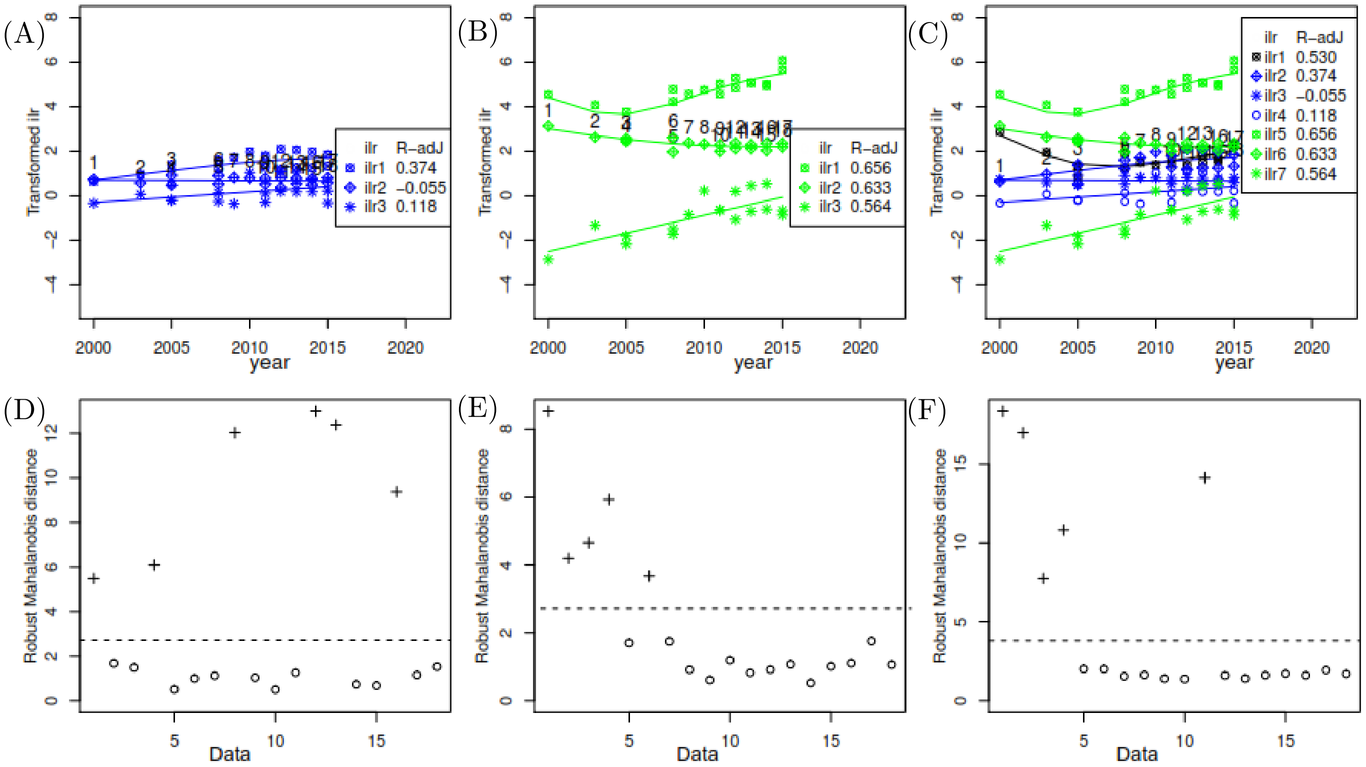


Figure 3: Outliers Colombia: (A), (D) Rural; (B), (E) Urban; (C), (F) Aggregate Urban/Rural.

The number of outliers detected has been variable. Figure 3D, 3E, and 3F show that this variation has not been punctual but in the time series. On the other hand, we expected less presence of outliers in aggregate data, due to the transforms; and a marked difference in the amount detected, between CoDa 4 and CoDa8, but this did not happen (See Table 1).

Table 1: Number of outliers detected in the countries.

Country/	Rural	Urban	Rural and Urban
N° Outlier	(CoDa4)	(CoDa 4)	(CoDa 8)
Bolivia	4	0	5
Colombia	6	5	6
Ecuador	5	4	4
Paraguay	6	5	4
Peru	5	8	7
Uruguay	1	4	ND

A vertical analysis of table 1 shows that in the rural sector there is a greater presence of outliers in Colombia and Paraguay; being this six. While Uruguay presents a single outlier. In the urban sector, a greater number of outliers have been detected in Peru, while Bolivia has no value.

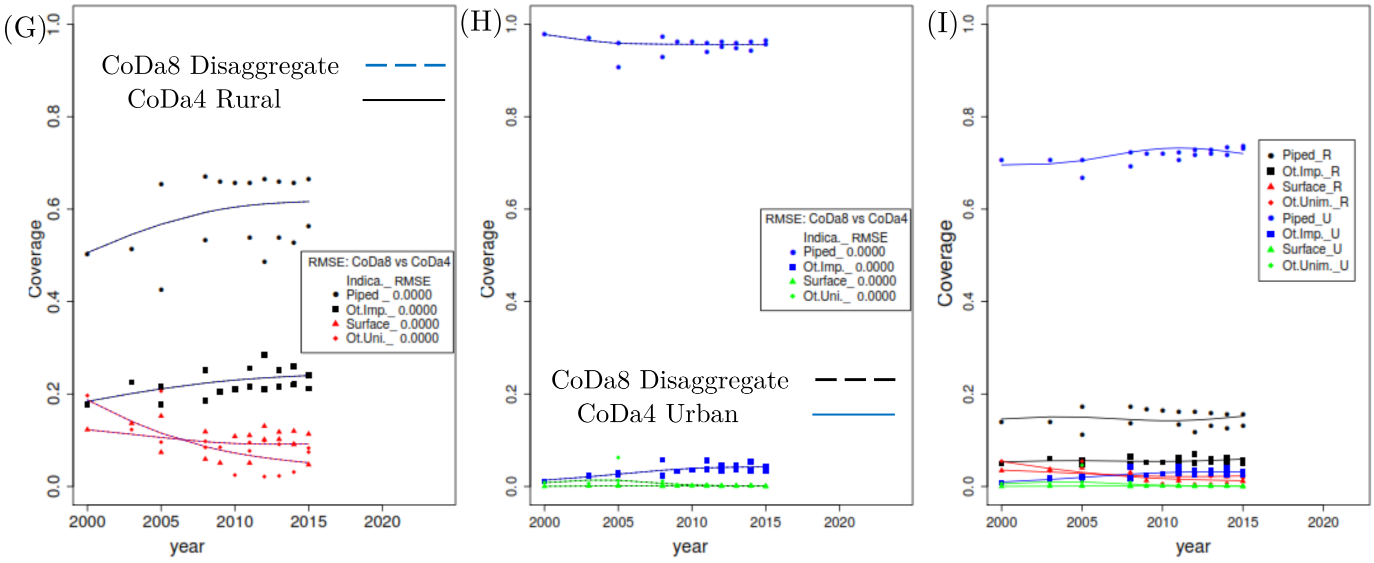
he horizontal analysis of table 1 shows the low quantity of outliers in disaggregated data (Urban and Rural) of Bolivia, then in aggregate data (CoDa 8). In Colombia, less is seen in the urban sector. In Ecuador, it could be assessed with less presence in aggregate data. In Paraguay there is less in aggregate data.

In summary, there has been no significant difference in the amount of data with outliers that allows us to assess whether four-part CoDa or eight-part CoDa is better.

3.2 Aggregated data vs. disaggregated data

The aggregate data analysis (Figure 4I) for scenario 1 shows the predominance in the proportion of the urban population that access piped water services. In addition, it can be seen that the proportion of people with access to unimproved water is low compared to the total.

In Figure 4G and 4H, it can be seen that the estimated values in CoDa 4 and CoDa 8 show the same trends in all indicators (solid line equal to the dashed line). The same analysis was carried out for the countries under evaluation, observing that in all of them the RMSE is equal to zero (See table 2). Which leads us to conclude that under the partition performed (Figure V1, V2 and V3) in scenario 1, it does not matter if we do the analysis in CoDa of four or eight parts, the result will always be the same.



(J) (K) (L)

CoDa8 Disaggregate
CoDa4 Rural

CoDa8 Disaggregate

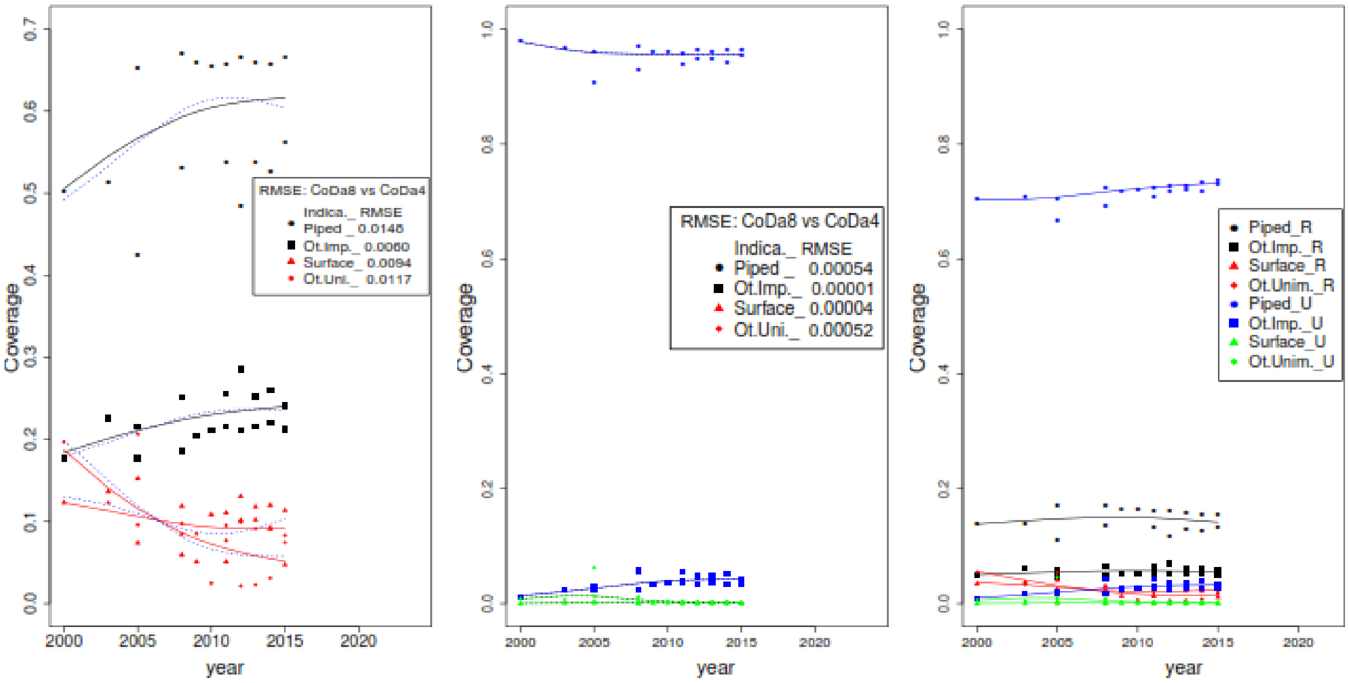


Figure 4: Comparison of models in aggregate and disaggregated data - Colombia. **Scenario 1**: (G) Rural disaggregate; (H) Urban disaggregate; (I) Rural and Urban aggregate. **Scenario 2**: (J) Rural disaggregate; (K) Urban disaggregate; (L) Rural and Urban aggregate.

Opposite case it happens when the partition changes (Scenario 2, V4). This shows that there is variation when making estimates with aggregate data (Figure 4L) and individual estimates of four parts (Figure 4J and 4K). The comparison of both gives us RMSE values different from one (Figure 4J and 4K). This does not happen in all countries or in all indicators (See table 2). In this exception is Paraguay, in which the RMSE remains zero in both scenarios (See table 2).

Table 2: Comparative of estimates between four-part CoDa and eight-part CoDa

País/(RMSE x 10^{-2})		Xr_1	Xr_2	Xr_3	Xr_4	Xu_1	Xu_2	Xu_3	Xu_4
Bolivia	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.2761	0.1381	0.285	0.1214	0.0507	0.0027	0.0065	0.0471
Colombia	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	1.4868	0.6074	0.9495	1.1742	0.0549	0.0013	0.0045	0.0524
Ecuador	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.5997	0.1628	0.4873	0.2751	0.154	0.0092	0.016	0.1479
Paraguay	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Peru	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.4504	0.1625	0.3023	0.3054	0.1416	0.0098	0.0135	0.1379

Uruguay	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.7032	0.1292	0.089	0.7448	0.0479	0.0002	0.0017	0.0464

In Paraguay the change of SBP has had no effect on the transformations; consequently, the estimates in CoDa 4 are the same as when disaggregated estimates of CoDa eight.

The variation between the two (aggregate, disaggregated data) has not been significant either, due to the fact that most of the subcompositions have RMSE values close to zero. After this analysis, we compare the efficiency of the prediction models in disaggregated data with the results of estimates in scenario 2.

Table 3: Comparison of the best efficiency of the models in aggregate and disaggregated data (Scenario 2).

País/NSE		**Xr$_1$**	**Xr$_2$**	**Xr$_3$**	**Xr$_4$**	**Xu$_1$**	**Xu$_2$**	**Xu$_3$**	**Xu$_4$**
Bolivia	CoDa4	-0.014	**0.508**	**0.375**	0.274	0.207	0.589	0.35	0.639
	CoDa8	**-0.008**	0.503	0.361	**0.278**	**0.215**	0.589	**0.358**	**0.653**
Colombia	CoDa4	0.15	**0.252**	**0.089**	0.512	**0.087**	0.49	**0.078**	**0.281**
	CoDa8	**0.169**	0.238	0.083	**0.528**	0.078	0.49	0.074	0.269
Ecuador	CoDa4	**0.453**	0.335	**0.607**	**0.476**	0.35	**0.287**	0.016	-0.008
	CoDa8	0.421	**0.343**	0.597	0.427	**0.392**	0.284	**0.089**	**0.002**
Paraguay	CoDa4	0.915	0.535	0.816	0.918	0.653	0.573	0.158	0.52
	CoDa8	0.915	0.535	0.816	0.918	0.653	0.573	0.158	0.52
Peru	CoDa4	**0.528**	0.398	**0.383**	**0.665**	-0.001	-0.036	-0.023	**0.674**
	CoDa8	0.52	**0.408**	0.367	0.659	**0.005**	-0.036	**-0.011**	0.666
Uruguay	CoDa4	**0.609**	0.366	**0.55**	**0.576**	0.138	**0.92**	**-0.112**	0.027
	CoDa8	0.594	**0.373**	0.546	0.548	**0.165**	0.919	-0.135	**0.056**

Horizontal analysis of table 3 shows that the predictive capacity of CoDa 8 in Bolivia is slightly better than CoDa 4. NSE values are higher in five out of eight indicators. The opposite situation occurs in Colombia, in which CoDa 4 presents better predictive capacity than CoDa 8. In Ecuador, both show improvements in four out of eight indicators. Therefore, it would have to be analyzed with other metrics that help to clarify whether CoDa 8 or CoDa 4 is better. Paraguay, is not affected by the variation in the partitions. In addition, it gives the same result if CoDa 8 or CoDa 4 (Table 2) is used. In Peru, CoDa 4 has a better predictive capacity. In Uruguay, CoDa 4 is better than CoDa 8.

In summary, the comparison of analyzing data in aggregate or disaggregated form, indicates that by doing analysis with CoDa 4, better predictive capacity will be presented in Uruguay, Peru, and Colombia.

For Paraguay, the use of either of them is indifferent. In Ecuador a more exhaustive analysis must be done, a priori NSE values are better in four out of eight indicators in rural and urban areas. Bolivia is the only country in which the eight-part CoDa analysis is better because its predictive capacity is better in most indicators.

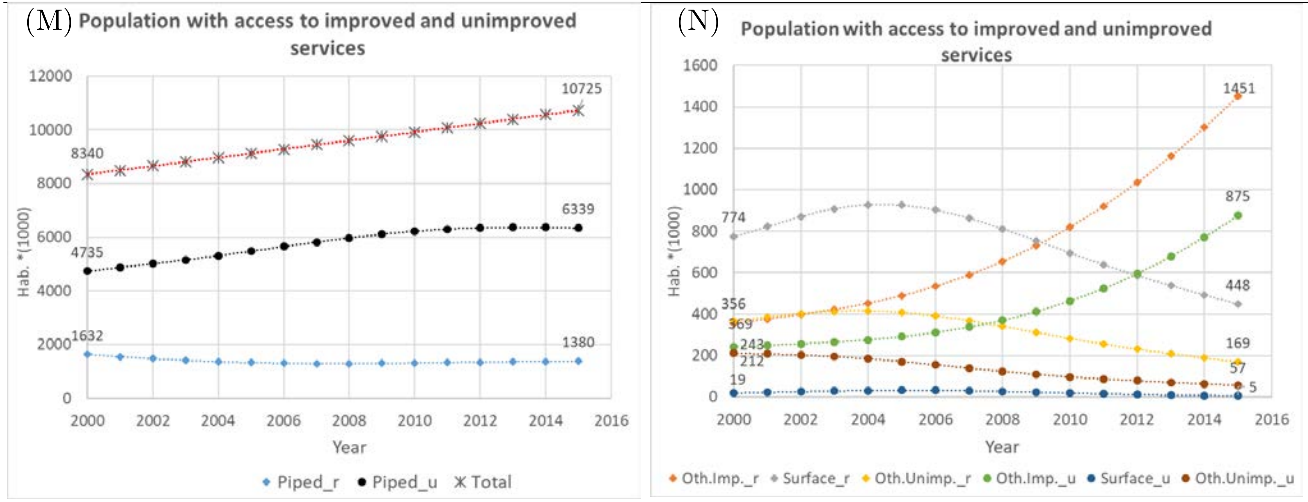


Figure 5: Temporal evolution in aggregate data of the population that accesses water services in Bolivia.

The temporal evolution of the population that accesses different water sources in Bolivia are shown in Figure 5 (Results of analysis in CoDa 8). The urban population that accesses piped water is increasing (Figure 5M). While the rural population in the same indicator shows a slight decrease. This decrease is offset by the increase in the population accessing water through other improved forms (Figure 5N). Of these, there was a greater increase in the rural population, reaching 1.451 million people who access this service. On the other hand, the urban and rural population that accesses surface water shows significant decreases in recent periods. Despite this, there is a large number of people from rural areas who continue to drink from these sources (448 thousand – Year 2015). The population that accesses water from unimproved sources decreased in the urban sector to 57 thousand people; while in the rural sector 169 thousand people were reached. Despite these decreases in both indicators (access to water by other improved forms, and access to surface water), in the rural sector more efforts are needed to close the existing gaps. Moreover, it is the area in which there is a high poverty rate and in the new SDGs they are focused on "nobody being left behind" (ONU, 2015).

In summary, the analysis in aggregate data has allowed us to analyze the temporal evolution of the population that accesses different water sources. In addition, with this methodology of work in the sector, national estimates are simple to perform, because it is the simple sum of the subcompositions; which does not happen at the moment (JMP, 2018).

Estimated values in this document may differ from the international JMP report, because linear regression methods are used, whereas GAM is used in this study. This has already been discussed extensively in the literature (Fuller et al., 2016; Bartram et al., 2014; Wolf et al., 2013; Pérez-Foguet et al., 2017).

4 Conclusions

It has been shown that in scenario 1 with an SBP1, using CoDa 4 or CoDa 8, the estimates give the same results (Figure 4G and 4H). While in scenario 2 with an SBP2, they present small differences (Figure 4J and 4K). This leads us to conclude that the selection of the PBS will influence the estimates. As a result, it opens a lot of possibilities to do analysis with different SBP and make the selection that best predictive capacity present. It is suggested to do these tests only in case you do not look for interpretations in the transforms. Otherwise, the appropriate group of parts should be selected to help us interpret better in the transforms. Because these carry proportions that contain information.

Evidence shows that CoDa 4 usually fares better in Uruguay, Peru, and Colombia (Table 3). While, in Bolivia, CoDa 8 presents better predictive capacity in five out of eight indicators. In Ecuador, you cannot infer, which of them is better.

The aggregate analysis in the data (CoDa 8), has allowed us to know in full, the temporal evolution of the population that accesses different water sources. A particular case is the one addressed in Bolivia. In which, there is an increase in the rural and urban population that accesses water by other improved forms. This was compensated by the decrease in access to unimproved water. On the other hand, it was found mostly in the rural sector, populations that access surface water sources. Consequently, Bolivia's agenda should be aimed at closing gaps in water, sanitation and hygiene. Taking as criteria, the poorest and most vulnerable populations.

On the other hand, the use of CoDa in aggregate data has certain disadvantages. The main one is the loss of information because it cannot complete the composition if it lacks data in some variables of the total. In the disaggregated analysis (urban and rural), the possibility of affecting only one sector is presented; what leads to not losing information in the other and consequently do the common analysis.

Regarding outliers, it cannot be inferred whether CoDa 8 or CoDa 4 has a lower quantity (Table 1) because there has not been significant variation. In later studies, comparisons of the models will be made by removing the outliers in each scenario.

Acknowledgements and appendices

The programming script is posted in GitHub (https://github.com/fquispec/Congress_CoDa_2019)

References

Aitchison, John. (1986). "The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability: Chapman and Hall, London London (UK)."

Bartram, Jamie, Clarissa Brocklehurst, Michael Fisher, Rolf Luyendijk, Rifat Hossain, Tessa Wardlaw, and Bruce Gordon. (2014). "Global Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges." *International Journal of Environmental Research and Public Health* 11 (8): 8137–65. https://doi.org/10.3390/ijerph110808137.

Boogaart, K. Gerald van den, Raimon Tolosana-Delgado, and Matevz Bren. (2014). "Compositions: Compositional Data Analysis." Boston. https://cran.r-project.org/package=compositions.

Cohen, Barney. (2004). "Urban Growth in Developing Countries: A Review of Current Trends and a Caution Regarding Existing Forecasts." *World Development* 32 (1): 23–51. https://doi.org/10.1016/J.WORLDDEV.2003.04.008.

Dirven, Martine, Rafael Echeverri, Cristina Sabalain, David Candia Baeza, Sergio Faiguenbaum, Adrián G. Rodríguez, and Carolina Peña. (2011). "Hacia Una Nueva Definición de 'Rural' Con Fines Estadísticos En América Latina." CEPAL. https://repositorio.cepal.org/handle/11362/3858.

DS. N°031-2008-VIVIENDA. (2008). "Decreto Supremo Que Modifica El Texto Único Ordenado Del Reglamento de La Ley General de Servicios de Saneamiento." Lima. http://www.sedapal.com.pe/contenido/031-2008-VDA (30.11.2008).pdf.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology* 35 (3): 279–300. https://doi.org/10.1023/A:1023818214614.

Egozcue, Juan Jose, and Vera Pawlowsky-Glahn. (2005). "Groups of Parts and Their Balances in Compositional Data Analysis." *Mathematical Geology* 37 (7): 795–828. https://doi.org/10.1007/s11004-005-7381-9.

Filzmoser, Peter, and Karel Hron. (2008). "Outlier Detection for Compositional Data Using Robust Methods." *Mathematical Geosciences* 40 (3): 233–48. https://doi.org/10.1007/s11004-007-9141-5.

Fuller, James A., Jason Goldstick, Jamie Bartram, and Joseph N.S. Eisenberg. (2016). "Tracking Progress towards Global Drinking Water and Sanitation Targets: A within and among Country Analysis." *Science of The Total Environment* 541 (January): 857–64. https://doi.org/10.1016/J.SCITOTENV.2015.09.130.

INE. (2018). "MEMORIA CENSO 2017 - GLOSARIO." http://www.censo2017.cl/memoria/descargas/memoria/libro_glosario_censal_2017.pdf.

JMP. (2017). "Joint Monitoring Programme for Water Supply, Sanitation and Hygiene." 2017. https://washdata.org/data.

JMP. (2018). "JMP Methodology 2017 Update & Sdg Baselines." https://washdata.org/sites/default/files/documents/reports/2018-04/JMP-2017-update-methodology.pdf.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn. (2003). "Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation." *Mathematical Geology* 35 (3): 253–78. https://doi.org/10.1023/A:1023866030544.

Martín-Fernández, Josep Antoni, Javier Palarea-Albaladejo, and Ricardo Antonio Olea. (2011). "Dealing with Zeros." In *Compositional Data Analysis*, 43–58. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch4.

ONU. (2015). "Transformar Nuestro Mundo: La Agenda 2030 Para El Desarrollo Sostenible." *Asamblea General. Septuagésimo Período de Sesiones de La Asamblea General de Las Naciones Unidas, Del 11 Al 18 de Septiembre Del 2015 (Resolución A/RES/70/1)* 16301: 40. https://doi.org/http://unctad.org/meetings/es/SessionalDocuments/ares70d1_es.pdf.

Palarea-Albaladejo, J., and J.A. Martín-Fernández. (2008). "A Modified EM Alr-Algorithm for Replacing Rounded Zeros in Compositional Data Sets." *Computers & Geosciences* 34 (8): 902–17. https://doi.org/10.1016/J.CAGEO.2007.09.015.

Pérez-Foguet, A., R. Giné-Garriga, and M.I. I. Ortego. (2017). "Compositional Data for Global Monitoring: The Case of Drinking Water and Sanitation." *Science of the Total Environment* 590–591 (July): 554–65. https://doi.org/10.1016/j.scitotenv.2017.02.220.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Sarkar Deepayan, and {R Core Team}. (2018). "Linear and Nonlinear Mixed Effects Models." https://cran.r-project.org/package=nlme.

R Core Team. (2018). "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Templ, Matthias, Karel Hron, and Peter Filzmoser. (2011). "RobCompositions: An R-Package for Robust Statistical Analysis of Compositional Data." In *Compositional Data Analysis: Theory and Applications*, 341–55. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch25.

Templ, Matthias, Karel Hron, Peter Filzmoser, and Alžběta Gardlo. (2016). "Imputation of Rounded Zeros for High-Dimensional Compositional Data." *Chemometrics and Intelligent Laboratory Systems* 155 (July): 183–90. https://doi.org/10.1016/J.CHEMOLAB.2016.04.011.

WHO/UNICEF. (2015). "Progress on Sanitation and Drinking Water – 2015 Update and MDG Assessment." World Health Organization and UNICEF. 2015. https://washdata.org/reports.

WHO/UNICEF. (2017). "Progress on Drinking Water, Sanitation and Hygiene – 2017 Update and SDG Baselines." World Health Organization and UNICEF. 2017. https://washdata.org/reports.

Wolf, Jennyfer, Sophie Bonjour, and Annette Prüss-Ustün. (2013). "An Exploration of Multilevel Modeling for Estimating Access to Drinking-Water and Sanitation." *Journal of Water and Health* 11 (1): 64–77. https://doi.org/10.2166/wh.2012.107.

Wood, Simon N. (2017). "Generalized Additive Models: An Introduction with R (2nd Edition)." *Chapman and Hall/CRC*.