Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Preprocessing alternatives for compositional data related to water, sanitation and hygiene

Alejandro Quispe-Coica*, Agustí Pérez-Foguet

Department of Civil and Environmental Engineering (DECA), Engineering Sciences and Global Development (EScGD), Barcelona School of Civil Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain

HIGHLIGHTS

• WASH data require preprocessing to adjust trends with compositional regression models.
• Outliers need to be treated differently based on amount of data from time series.
• Robust OLS of hygiene trends is recommended over standard OLS.
• Detection and exclusion of outliers improves quality measures of GAM regressions.

GRAPHICAL ABSTRACT

The Sustainable Development Goals (SDGs) 6.1 and 6.2 measure the progress of urban and rural populations in their access to different levels of water, sanitation and hygiene (WASH) services, based on multiple sources of information. Service levels add up to 100%; therefore, they are compositional data (CoDa). Despite evidence of zero value, missing data and outliers in the sources of information, the treatment of these irregularities with different statistical techniques has not yet been analyzed for CoDa in the WASH sector. Thus, the results may present biased estimates, and the decisions based on these results will not necessarily be appropriate. In this article, we therefore: i) evaluate methodological imputation alternatives that address the problem of having either zero values or missing values, or both simultaneously; and ii) propose the need to complement the point-to-point identification of the WHO/UNICEF Joint Monitoring Program (JMP) with other robust alternatives, to deal with outliers depending on the number of data points. These suggestions have been considered here using statistics for CoDa with isometric log-ratio (ilr) transformation. A selection of illustrative cases is presented to compare performance of different alternatives.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

Monitoring access to WASH services is a multiscale process involving bodies from the local level—to support the planning (Giné-Garriga et al., 2013, 2015) and implementation of government public policies (Jiménez Fdez. de Palencia and Pérez-Foguet, 2011)—to the international level (WHO/UNICEF, 2017). WASH monitoring has evolved substantially over the last 15 years. A key point is the movement from the use of single indicators of performance (such as coverage of water and sanitation by improved and unimproved technologies) to multidimensional frameworks that understand WASH in relationship with concepts such as poverty (Giné-Garriga and Pérez-Foguet, 2013a, 2019) and

* Corresponding author.
E-mail addresses: filimon.alejandro.quispe@upc.edu (A. Quispe-Coica), agusti.perez@upc.edu (A. Pérez-Foguet).
human rights (Baquero et al., 2015; Giné-Garriga et al., 2017), or from the perspective of vulnerable and marginalized groups (Redman-Maclaren et al., 2018; Ezbakhe et al., 2019; Anthony et al., 2020a). Integrating these concepts leads to a much higher complexity than simple coverage of a population by one technical solution or another.

This multidimensional nature was first measured through aggregated indicators such as the WASH poverty index (Giné-Garriga and Pérez-Foguet, 2013a, 2013b) that extended the seminal proposal of the Water Poverty Index (Sullivan, 2002; Sullivan et al., 2003; Giné-Garriga and Pérez-Foguet, 2010; Pérez-Foguet and Giné-Garriga, 2011). Likewise, some limitations of aggregated indicators, such as the compensability between dimensions and the lack of mechanisms to consider cross-influences between dimensions, has been tackled with different techniques (Ezbakhe and Pérez-Foguet, 2018; Giné-Garriga et al., 2018), mostly within the approach of supporting specific decision making processes.

Some of these ideas are currently integrated into the international WASH ladder monitoring driven by the JMP, which has moved from a coverage perspective to a service level approach. Here, a safe management category for water and for sanitation, and the hygiene ladder, are as main novelties. However, the basic framework for local and international monitoring still needs trend analysis of the percentage of population expressed by single categories (e.g., the primary indicators; WHO/UNICEF, 2018), whose particular characteristic is that they describe the parts of a whole with a sum of 100% (or 1, if they are proportions). This approach overlooks that data are compositional and require a particular statistical approach (Aitchison, 1986; Egozcue and Pawlowsky-Glahn, 2005; Lloyd et al., 2012; Pérez-Foguet et al., 2017; Ezbakhe and Pérez-Foguet, 2019). Further, the compositional nature of data is not addressed in the proposed alternatives for multidimensional monitoring or by the JMP for global WASH monitoring (WHO/UNICEF, 2018), which may lead to spurious correlations between the parties (Pérez-Foguet et al., 2017).

Fuller et al. (2016) classified the temporal evolution of access to water and sanitation according to the linearity or non-linearity of trends and proposed the use of Generalized Additive Models (GAM) when data are at a minimum. The compositional nature of the population percentages is included in the analysis presented by Pérez-Foguet et al. (2017), which concluded that using GAM for the isometric log-ratio (ilr) transformations of the usual follow-up variables is suitable. In this way, the non-linearity of the sum restrictions equal to constant is adequately treated. This is especially relevant when parts of the total tend to values near the extremes of all or nothing. However, the proposal does not address common situations, such as the presence of values reported as zero, or missing data in parts of the total, thereby preventing a direct application of the compositional approach.

Data with a value of zero are commonly presented in countries that have made significant progress in the provision of improved water and sanitation services; as a consequence, populations with access to unimproved sources have been drastically reduced, with the number in many cases at or near zero. The ilr transformations in the data therefore cannot be carried out if the zero values are not first excluded or imputed. Exclusion is an easy alternative to address the problem, but if the amount of data in the sector is low, this can affect the predictive capacity of the models. Thus, in the literature, alternatives have been proposed for the imputation of zero values in each situation according to the CoDa properties, including rounded zeros (Palarea-Albaladejo et al., 2007; Palarea-Albaladejo and Martín-Fernández, 2008; Martín-Fernández et al., 2012; Templ et al., 2016; Chen et al., 2018), count zeros (Martín-Fernández et al., 2015) and essential zeros (Aitchison and Kay, 2003).

The techniques related to rounded zeros are the more convenient imputation alternatives for the WASH sector, given that even in more developed countries, there are likely to be at least small percentages of populations that do not have access to any kind of water services. Simple replacement and multiplicative replacement have already been addressed in previous studies of the sector (Pérez-Foguet et al., 2017; Ezbakhe and Pérez-Foguet, 2019). Despite their simplicity in the application, these methods tend to underestimate the variability of data; therefore, it is advisable that they are only used when the presence of zeros is low (Palarea-Albaladejo and Martín-Fernández, 2008). In the presence of large amounts of zero values, other imputation alternatives are recommended, according to the variability of data that exist in the time series.

The lack of data defining the composition is also a topic of special importance in the sector, as it affects some categories of analysis. For example, according to the national survey (PNAD17) in the rural sector of Brazil, 88.4% of the population have access to improved drinking water sources (and 82.7%, by pipe), but no information is given about access by surface sources (WHO/UNICEF, 2019a). The lack of one or more data points for a specific year means that the ilr transformation cannot be applied directly, so that the information of that year is lost in the follow-up of all parts (Quispe-Coica and Pérez-Foguet, 2018). A first alternative is to exclude incomplete data from the analysis, but this can lead to biases (Strike et al., 2001), severe loss of information, inaccurate estimates that do not help managers make the best decisions, etc. There are different alternatives based on completing the missing data, including a multiplicative replacement by Martín-Fernández et al. (2003), a modified EM algorithm by Palarea-Albaladejo and Martín-Fernández (2008) and a classic and robust method imputation by Hron et al. (2010); however, the most appropriate techniques for the specific cases of the WASH sector have not yet been determined.

Finally, the quality of available data can be classified in many cases as low or very low. The JMP validates data and metadata (data source information) one-by-one to determine what can be used. Discrepancies between data are not per se a reason for exclusion. To cite one example, the percentage of the population with access to piped water in rural Indonesia was reported to be 6.6% by the National Socio-economic survey in 2016, but another source of information reported that it was 41.5% (Performance Monitoring and Accountability; PMA16) (WHO/UNICEF, 2019b). This stems from the use of multiple sources of information and is not easily remedied automatically, yet it directly influences the estimates obtained under any model. Recently, Ezbakhe and Pérez-Foguet (2019) proposed a method to deal with uncertainties that originate in statistical sampling, using compositional models of trends as applied to water and sanitation data. However, completing the punctual validation of the JMP with techniques and procedures for the detection of outliers or other data errors other than sampling (Bain et al., 2018) is still pending. Therefore, evaluating identification alternatives for the sector’s CoDa is necessary.

When working with CoDa, outliers cannot be identified for a variable independently of the rest. Multivariate analysis methods are necessary to facilitate the adequate detection of outliers and to enable data with evident errors to be identified, which can alter the estimates (Filzmoser and Hron, 2008; Filzmoser et al., 2009, 2012). Filzmoser and Hron (2008) proposed the use of robust identification techniques based on the Mahalanobis distance (MD). The proposal applies to general regression models, such as GAM. Nevertheless, the low amount of data that some countries have can limit the use of this application. Other alternatives, such as ordinary least squares (OLS) regression, provide a better option in those cases. However, direct application of OLS is not convenient, as it can be negatively influenced by the presence of outliers. Therefore, it is necessary to apply robust estimators for linear regression models. Several methods for this exist in the literature, including M-estimation and S-estimation (Rousseeuw and Yohai, 1984), MM-estimation (Yohai, 1987; Koller and Stahel, 2011) and others (see overview in Maronna et al., 2019). In this study, the MM-type estimators are applied, based on the good results obtained with them in other studies. It should be added that robust estimates do not necessarily exclude outliers, but rather modulate their influence on the calibrated model, which gives it a strong advantage for use with limited data.
This work proposes and analyzes different coupled strategies for the treatment of zeros, missing data and outliers in compositional trend models, as applied to the international monitoring of the WASH sector, completing the previous work in this regard and facilitating its practical application to the available data. Specifically, it addresses the following objectives:

- Evaluate alternatives for the treatment of zeros or missing data, or both simultaneously, using robust methods;
- Identify and treat outliers by robust methods in a differentiated way for contexts with few or many (more than six) different temporal data, according to the Fuller et al. (2016) classification.

For this, a set of twelve trends has been selected, with different characteristics, that are internationally representative and within the set of situations in the sector for both urban and rural WASH settings.

2. Method

The algorithm proposed and shown in Fig. 1 follows statistical procedures and techniques for CoDa that can be easily applied and replicated in any sector or area of analysis. To understand them, one must first know some basic concepts, such as: i) CoDa represent vectors with D representing strictly positive components, and the sum is a constant “k”, as shown in Eq. (1); ii) its sample space is the simplex $S^D$; for statistical analysis, it is necessary to move to the Euclidean space using ilr transforms, which requires that D components be passed to $(D-1)$.

$$S^D = \{X = (X_1, \ldots, X_D) : \forall X_i > 0, \ i = 1, 2, \ldots, D \sum_{i=1}^D X_i = k \} \quad (1)$$

$k$: can be 1, 100, or any other positive constant.

These concepts and terms, although they seem simple, are not common in the WASH sector. Therefore, it is necessary to be clear about them, to understand the method of analysis in CoDa.

**STEP I. Preparation of compositional data (CoDa)**

If the information is in population (P) units, the proportions of the service categories are formed according to Eq. (2). Subsequently, vectors are constructed with the parts, in which the sum is a constant “k” (100% if it is given as a percentage, or 1 if given in proportions). In vectors for which data are missing, “NA” is used.

$$X^p_1(\% ) = \frac{P^p_1}{\sum_{r=1}^4 P^r_1} \times 100 \quad (2)$$

$$X_1 + X_2 + X_3 + X_4 = 100 \quad (3)$$

$$X_{h1} + X_{h2} + X_{h3} = 100 \quad (4)$$

Indicators are formed according to Table 1. Water and sanitation are represented in Eq. (3) and each comprise four parts, while hygiene indicators are represented in Eq. (4) and comprise three parts.

The composition vectors that present irregular data (e.g., that are observed with values of zero, missing values or zero plus missing values simultaneously) are approached in a way or another the classification (Egozcue et al., 2003) with each particularity in the balances V. This procedure is also applied to generate the models.

$$Y = ilr(X) = \sqrt{\frac{r+s}{r+s}} \ln \frac{g_{ilr}(X_r+)}{g_{ilr}(X_r-)} \quad (5)$$

$r$ = number of positive variables in the balance $V$
$s$ = number of negative variables in the balance $V$
$g_{ilr}(-) =$ = geometric mean of variables

However, to illustrate the behavior of the models in the transformed data, a type of balance is carried out, consistent with the usual form of analysis in the WASH sector. For example, global monitoring is based on the classification of access to improved and unimproved water and sanitation services, which are subsequently subdivided into service categories (WHO/UNICEF, 2017; Turman-Bryant et al., 2018); likewise, both inequalities in access to water and sanitation (Yang et al., 2013; Bain et al., 2014; UNICEF/WHO, 2019; Anthonj et al., 2020b; Chitonge et al., 2020) as well as studies of access to WASH and its relation to health (Prüss-Ustün et al., 2014; Freeman et al., 2017; Ashole Alto et al., 2020; Hasan and Alam, 2020; Patel et al., 2020) imply in one way or another the classification of improved and unimproved services. Therefore, the order of the balances (Egozcue and Pawlowsky-Glahn, 2005) is defined under this criterion (see Fig. 2), with the breakdown of each part as follows:

The water and sanitation balances each comprise four parts and follow the same procedure ($V_i$), with the balance carried out between the proportion of the population:

i. with access to improved ($X_1 \times X_2$) and unimproved ($X_3 \times X_4$) services;
ii. next, with access to network services ($X_1$) and other improved ($X_2$) forms of access;
iii. finally, with access to services ($X_3$) and other unimproved ($X_4$) forms of access.

Hygiene balances comprise three parts and are performed under the following procedure ($V_i$), with the balance carried out between the proportion of the population:

i. with a handwashing facility on premises ($X_{h1} \times X_{h2}$) and no handwashing facility (no service) ($X_{h3}$);
ii. next, with access to basic services ($X_{h1}$) and limited service ($X_{h2}$).

The results of balances and transformations are shown in Table 2.

**STEP II. Treatment of values of zero, missing data and zero plus missing data simultaneously**

Countries with data that include values of zero, missing values or values of zero plus missing values simultaneously are approached in a differentiated way with robust statistical techniques, as the low quality of data from the sector can influence the imputations (Hron et al., 2010; Martín-Fernández et al., 2012; Maronna et al., 2019). For the three cases mentioned, two treatment alternatives are compared. The number of zeros is denoted by NZ, and the number of missing values, by NM.

i) $NZ = 0, NM = 0$: no pre-processing of data is performed.

ii) $NZ > 0, NM = 0$: treatment of zero values with two variants of the log-ratio expectation-maximization (EM) algorithm; IrEM function (Palarea-Albaladejo and Martín-Fernández, 2015) and impR2lir (Tempel et al., 2019).

iii) $NZ = 0, NM > 0$: treatment of missing values through least trimmed squares (LTS) (Hron et al., 2010), implemented in the impCoda function, or with the same log-ratio EM algorithm used for the case (ii) ($NZ > 0$ and NM = 0); IrEM function.

iv) $NZ > 0, NM > 0$: treatment of zero and missing values simultaneously, also with two alternatives. One is to consider zero values as a special type of missing values (Palarea-Albaladejo and Martín-Fernández, 2008; Martín-Fernández et al., 2011) and apply the same LTS algorithm as before (e.g., the impCoda function). The opposite should not be considered because the missing values are not necessarily zero values. The other alternative is using the extended version of the log-ratio EM algorithm, the IrEMplus function, presented by Palarea-Albaladejo and Martín-Fernández (2020).

**STEP III. Models and estimates**
Countries are classified into two groups according to the amount of data, with six being the separation limit. This classification is described in Fuller et al. (2016). However, as the low quality of data also affects the predictive capacity of the models, we opted to carry out robust models in both groups as detailed below:

**STEP I:**
Preparation of compositional data (CoDa)

- **NZ = 0, NM = 0**
- **NZ > 0, NM = 0**
- **NZ = 0, NM > 0**
- **NZ > 0, NM > 0**

**STEP II:**
Data ≥ 6

- **No**
- **Yes**

**STEP III:**
Outliers: Mahalanobis distance (MD)

- OLS (ilr)-robust

**STEP IV:**
Inverse transformation

**STEP V:**
Results and quality test

For countries with data points < 6: the models are built using the robust OLS regression method on transformed data from Table 2B, for which the lmrob function is used, which calculates a regression estimator of the MM type as previously described (Yohai, 1987; Yohai et al., 1991; Koller and Stahel, 2011). Evaluation of the influence of outliers in linear regression models is carried out using robustness weights. On the other hand, standard linear regression models are added to transformed and non-transformed data, for comparison with the robust alternative.

Countries with data points ≥ 6: the model-fitting procedure combines the outlier identification method as part of the preprocessing and then excludes these data from the analysis to generate robust models, as described below:

1. Outliers in multivariate data are identified by calculating the robust Mahalanobis distance (Eq. (6)) in isometric log-ratio coordinates of Eq. (5). For the computational calculation, the outCoDa function is applied (Templ et al., 2011).

\[
MD(Y_i^n) = \left[(Y_i - T)^T C^{-1} (Y_i - T)\right]^{1/2} \quad \text{for } i = 1, 2, ..., n
\]
where T and C are estimators of location and the covariance, respectively (Mahalanobis, 1936). Robustness is achieved by exchanging T and C for the minimum covariance determinant (MCD), which are robust estimators (Filzmoser and Hron, 2008). Potential outliers are those that have robust MD (square) greater than the cut-off value, which is the 0.975 quantile of the $\chi^2_{D-1}$ distribution with D-1 degrees of freedom (Rousseeuw and van Zomeren, 1990). In the case of water and sanitation, the degree of freedom of the chi-square distribution is three, and the cut-off value is 3.0575. Points that are above the threshold distance are not taken into account in subsequent estimates ($\text{MD}(Y_t) > \chi^2_{0.975}$).

ii. After identifying outliers, regression models are constructed with GAM, with four degrees of freedom ($k = 4$), on the transformed data in Table 2A. The analysis is performed for data both with and without the presence of outliers. The predictive capacity of the model between the two is compared with the adjusted coefficient of determination (R-adj); values near to one the predictive capacity of the model is better. The computational calculation to generate the models is done with the gam function.

**STEP IV. Inverse transformation**

The interpolation or extrapolation values in the transformed data are returned to the simplex space, for which the inverse transformation is mandatory. Therefore, performing an inverse transformation is mandatory.

**STEP V. Results and quality test**

The whole process of the algorithm described up to STEP IV allows the interpolations and extrapolations of the different alternatives in the categories of access to WASH to be evaluated and compared, using quality metrics. In order to see the impact of the alternatives in STEP II on the scale of data, the root mean square error (RMSE) metric is applied to models expressed in terms of X. On the other hand, the evaluation of the predictive capacity of the models in the data is carried out through the non-dimensional indicator goodness of fit of Nash Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) applied to the observed and estimated X of the model. If NSE = 1, the fit of the model is perfect, while NSE < 1 suggests that the observed mean is a better predictor than the model (Ritter and Muñoz-Carpena, 2013).

The statistical computation of Fig. 1 is performed through R Core Team (2020) (v.3.6.3). Preprocess of data and integration of each calculation stage are presented in Quispe-Coica and Pérez-Foguet (2020).

The following packages are used: robCompositions (v2.2.1) by Tempel et al. (2011) for impRZlir, impCoda, and outCoDa; zCompositions (v1.3.4) by Palarea-Albaladejo and Martín-Fernández (2015)) for irEm and IrEmPlus; robustbase (0.93–5) by Maechler et al. (2019) for lmrrob; mgcv (v1.8-31) of Wood (2019) for gam; and compositions (v1.40-3) by van den Boogaart et al. (2019).

### 3. Data features

To test the algorithm proposed in Fig. 1, we selected ten different countries for data on access to water and sanitation, and two countries for the hygiene case. Annual data are extracted from the JMP database from 2000 to 2019, in which both the amount of data and the presence or absence of irregularities vary in different proportions, allowing the various situations that arise in the WASH sector to be covered (see Table 3). Appendix A illustrates the implication in the correlation matrix of not using adequate techniques for CoDa.

The countries that do not present data irregularities are represented by Benin and Ghana for access to hygiene, and by Benin for access to rural water. For hygiene, the low amount of data is mainly due to the recent incorporation of this into the Sustainable Development Goals (SDG 6.2) as part of the monitoring indicators (Craven et al., 2013); in contrast, access to water and sanitation has been monitored since 1990 (Bartram et al., 2014). In this type of data, STEP II of the algorithm does not apply. Data with irregularities are presented in three different forms:

i) The first case is represented by Nigeria and Paraguay, which have values of zero in the data, of 1.14% and 7.14%, respectively. The categories of Paraguay reveal that this occurs when the provision of water services by improved sources is high (Fig. 3A); consequently, indicators of access to unimproved water have zero trends or zero values. Another peculiarity that can be seen in Paraguay is that the zero value is present only in the X3 indicator, while in Egypt, it is presented in X3 and X4.

ii) The second case concerns countries with missing values in data and are represented by South Africa, Zambia, Brazil and Benin. Brazil has the highest percentage of missing values (of 44.44%), which are distributed in the same proportions in the X3 and X4 indicators (Fig. 3B).

iii) The third case refers to countries with both values of zero and missing values in data and are represented by Bangladesh, Egypt and Uruguay. Egypt is shown as an example in Fig. 3C for data with zero values, and in Fig. 3D for data with missing values. In both graphs, data with zero values and missing values are in the categories of X3 and X4.

Data irregularities must be addressed in STEP II, by using the imputation functions most appropriate for each case.

### 4. Results and discussion

In this section, we discuss and compare work alternatives to treating values of zero, missing values and values of zero and missing values simultaneously that are usually present in the data. Subsequently, we analyze the influence of outliers on the model.
4.1. Countries with values of zero, missing values or values of zero and missing values simultaneously in the data

Of the different characteristics of the data presented in Section 3, countries with irregular data have gone through differentiated treatment methods in STEP II. For instance, Paraguay and Nigeria have zero values of zero in their data, of 7.14% and 1.14%, respectively (Table 4A). The analysis under the RMSE metric of the imputation functions \( lrEM \) and \( impRZilr \) that replace the zero values with small values do not show any significant differences that would allow us to discard any of the

### Table 3

Access to water, sanitation and hygiene (WASH).

| Region                          | Country           | Sector | Service   | Data points (X) \(^a\) | Zero value | Missing value |
|---------------------------------|-------------------|--------|-----------|-------------------------|------------|---------------|
| Sub-Saharan Africa              | South Africa      | Rural  | Sanitation| 30 (×4)                 | 0.00%      | 1.67%         |
| Latin America and the Caribbean | Brazil            | Urban  | Water     | 27 (×4)                 | 0.00%      | 44.44%        |
| Eastern and south-eastern Asia  | Indonesia         | Rural  | Water     | 26 (×4)                 | 0.00%      | 0.00%         |
| Sub-Saharan Africa              | Nigeria           | Rural  | Water     | 22 (×4)                 | 1.14%      | 0.00%         |
| Latin America and the Caribbean | Paraguay          | Urban  | Water     | 21 (×4)                 | 7.14%      | 0.00%         |
| Central and Southern Asia       | Bangladesh        | Rural  | Sanitation| 20 (×4)                 | 1.25%      | 30.00%        |
| Sub-Saharan Africa              | Zambia            | Rural  | Sanitation| 16 (×4)                 | 0.00%      | 6.25%         |
| Northern Africa and Western Asia| Egypt             | Urban  | Water     | 15 (×4)                 | 10.00%     | 30.00%        |
| Latin America and the Caribbean | Uruguay           | Urban  | Water     | 15 (×4)                 | 15.00%     | 3.33%         |
| Sub-Saharan Africa              | Benin             | Rural  | Sanitation| 10 (×4)                 | 0.00%      | 10.00%        |
| Sub-Saharan Africa              | Benin             | Rural  | Hygiene   | 5 (×3)                  | 0.00%      | 0.00%         |
| Sub-Saharan Africa              | Ghana             | Rural  | Hygiene   | 4 (×3)                  | 0.00%      | 0.00%         |

\(^a\) The year’s data points are represented by three or four levels of WASH service to which the population has access.

---

**Fig. 3.** A) Paraguay: patterns of zero values. B) Brazil: patterns of missing values. (C and D) Egypt: zero values and missing values simultaneously, with values of zero shown in C), and missing values shown in D).
two alternatives completely; therefore, both functions can be applicable, as either of them helps to overcome the problem of not being able to perform the ilr transformations of Eq. (5).

Nevertheless, when dealing with the missing values in data (Table 4B), differences in the metrics should be taken into account, which makes us choose the impCoDa function or the lrEM function depending on the country of analysis. For example, metrics in the impCoDa function are better for Benin and South Africa, while the lrEM function is better for Brazil; in contrast, no significant differences between either function (impCoDa or lrEM) are present for Zambia.

On the other hand, in countries with values of zero and missing values simultaneously (see Table 4C), the alternative of replacing zero values with “NA” and addressing them as “missing values” with the impCoDa function gives better results for Bangladesh and Egypt. This occurs when there is a higher percentage of data with missing values than zero values. However, the opposite situation occurs in the data set from Uruguay, which has 15% of zero values and 3.33% of missing values, and for which the lrEMplus function is a better alternative.

Finally, while it is true that any of the methods evaluated is adequate for at least one of the cases (depending on each case), the methods are all already better than the multiplicative imputation alternatives or other simple alternatives, as they allow variability to exist in the imputed data. This advantage is more significant when the data points show a higher percentage of these irregularities. If no alternative is applied (either simple or one of those shown in this paper), many countries in the sector should be excluded from the analysis. This is especially important if the loss of information is significant (as happens in South American countries; Quispe-Coica and Pérez-Foguet, 2018). On the other hand, once the new Sustainable Development Goals were agreed (United Nations General Assembly, 2015; UN Water, 2016), each country assumed the responsibility of reducing the population’s access to unimproved services of WASH. To this end, many countries are defining and implementing public policies that close these gaps, in which case data will tend to go to extreme values, making it even more necessary to use imputation alternatives for zero values.

4.2. Outliers

4.2.1. Countries with data points < 6

This section addresses the case of countries with little data, where the influence of outliers is penalized in the coupled model. The access of rural populations to the different levels of hygiene services in Benin and Ghana illustrates this situation. In Fig. 4A and B, we present the model fit in data transformed by standard and robust linear regression. The regression lines of both methods are similar in the transformations of ilr2, and differ for both methods in ilr1 (with more drastic changes in Fig. 4E). The difference is mainly due to the fact that, in the robust method, points 1 and 5 of Ghana and Benin, respectively, have a strong degree of negative influence on the model, so that it assigns zero value robustness weights.

How the influence of data is modulated creates significant differences in the estimates of the categories of hygiene services that the population accesses. In the case of Ghana, the effect in each category is even greater if we compare it with the other alternatives (Fig. 4F–H). Likewise, in both Benin and Ghana, the curve generated by robust OLS (ilr) best fits the data. On the other hand, looking at the results qualitatively, it is more reasonable to exclude point 1 in Ghana and point 5 in Benin, which supports the affirmation that the robust linear regression alternative is an excellent alternative for regression models in the presence of data with outliers.

Another feature to consider is that with OLS (ilr) or robust OLS (ilr), extrapolations of the service categories in 2000 and 2017 never exceeded the extreme limits of 0 and 100% (Table 5). This happens because the inverse transformation has a closing value (Eq. (4)) that allows estimates to be made in the time series without restrictions. The opposite occurs when extrapolation is performed with standard OLS. Here, the negative values of Benin (−0.726) and Ghana (−1.594) in the basic service category of the year 2000 exemplify this situation; in these cases, the JMP applies restrictions of 0 (WHO/UNICEF, 2018).

On the other hand, the results of the 2017 estimates with robust OLS (ilr) differ significantly from the other linear alternatives in all Benin categories. In Ghana, only robust OLS (ilr) and JMP regression give very similar results in all three categories. Although the estimation alternatives differ, there is a high percentage of the rural population that does not have handwashing facility (specifically, 43.07% in Benin, and 20% in Ghana), if we take into account the results of robust OLS (ilr). In both countries, this rate is expected to decrease, given the positive effects of handwashing with soap and water in the reduction and prevention of diseases, such as diarrhoea, coronavirus disease 2019 (COVID-19), acute respiratory infection and impetigo, among others (Luby et al., 2005; Cairncross et al., 2010; Hirai et al., 2017; Prüss-Ustün et al., 2019; Brauer et al., 2020; Ma et al., 2020).

4.2.2. Countries with data points ≥ 6

The possible reasons for outliers in data can be diverse. However, in the data analyzed here, it is evident that outliers commonly occur when there are different sources of information. To better illustrate this point, we present the case of the rural population of South Africa, for which information for the sewer categories in 2011 comes from three different sources: the Census (CEN) reported 6.03% access, the Income and Expenditure of Homes survey (IES) reported 44.16% access, and the General Household survey (GHS) reported 5.07% access. Based on the significant difference between data from IES and that from the other two sources of information (CEN and GHS), it is normal to assume that it is an outlier without needing to apply any validation methods.

| Table 4 |
| --- |
| Quality metrics to select the method. |

| Country | Sector | Service | Zero value | Missing value | Method – RMSE (%) | Selected method |
| --- | --- | --- | --- | --- | --- | --- |
| **A. Case II: Data with zero values** |
| Paraguay | Urban | Water | 7.14% | 0.00% | – | 0.0026 | impCoda |
| Nigeria | Rural | Water | 1.14% | 0.00% | – | 0.0060 | lrEMplus |
| **B. Case III: Data with missing values** |
| Benin | Rural | Sanitation | 0.00% | 10.00% | 1.826 | 3.094 | – | impCoda |
| Brazil | Urban | Water | 0.00% | 44.44% | 0.648 | 0.321 | – | lrEMplus |
| South Africa | Rural | Sanitation | 0.00% | 1.67% | 0.015 | 0.026 | – | impCoda |
| Zambia | Rural | Water | 0.00% | 6.25% | 8.435 | 8.227 | – | lrEMplus |
| **C. Case IV: Data with zero values and missing values simultaneously** |
| Bangladesh | Rural | Sanitation | 1.25% | 30.00% | 7.621a | – | 8.690 | impCoda |
| Egypt | Urban | Water | 10.00% | 30.00% | 0.254a | – | 0.269 | impCoda |
| Uruguay | Urban | Water | 15.00% | 3.33% | 0.052a | – | 0.048 | lrEMplus |

* Data with values of zero are considered missing values (“0” → “NA”); therefore, imputation methods with the impCoda function are applied.
On the other hand, as the census data and the EEG survey only differ by 0.96%, it is difficult to know if either value is atypical or not. Given the doubt that is generated, robust MD can apply to the country’s time series. The results obtained show that only the IES data point is an outlier (Fig. 5A.2), which confirms the previous assumption.

The punctual validation carried out by the JMP (2019) (see Excel tab “Data Summary/Sanitation for 2011”) identifies and excludes the CEN and IES data points from the model. These differences in identification that are manifested for a specific country and year can also occur for other countries when a time series is analyzed.

The coherence and contradictions in the number of outliers identified through the two methods, the robust MD and the JMP, are shown in Table 6. The number of outliers identified by the robust MD is higher than that identified by JMP in nine of ten countries, with Paraguay showing the greatest difference, while the opposite is seen for Zambia in categories X1 and X3. In contrast, in both South Africa and Zambia, the number of identified outliers is the same between the two alternatives (robust MD and JMP) in categories X1 and improved, respectively.

![Fig. 4. Model and estimations in CoDa of hygiene. (A, E) Two different models are fitted in transformed data: i) standard OLS (blue and black solid lines) and ii) robust OLS (blue and black dashed lines). (B–D, F–H) Three different models are fitted in CoDa: i) standard OLS in original data (black solid line), ii) inverse of standard OLS of transformed data (blue solid line) and iii) inverse of robust OLS of transformed data (blue dashed line).](image)

**Table 5**

| Hygiene Method | country | Estimation year (2000) | Estimation year (2017) |
|---------------|---------|------------------------|------------------------|
|               |         | Basic service | Limited service | No service | Basic service | Limited service | No service |
| Basic service |         | 0.162         | 2.740           | 97.098     | 13.862       | 13.868       | 72.270     |
| Limited service |         | 1.059         | 1.752           | 98.142     | 27.959       | 28.968       | 43.073     |
| No service    |         | -0.726        | 3.638           | 97.088     | 6.466        | 17.351       | 76.183     |
| OLS (ilr)     |         | 4.712         | 36.629          | 58.659     | 23.431       | 32.464       | 44.104     |
| OLS (ilr)-robust |         | 0.009         | 2.912           | 97.088     | 6.043        | 16.544       | 77.413     |
| OLS a |         | -1.594        | 44.893          | 56.701     | 24.890       | 30.010       | 45.100     |
| OLS JMP website |         | NA           | NA              | NA         | NA           | NA           | 36.576     |

a OLS regression on untransformed data.

b Data available at JMP website (Benin and Ghana, Excel tab “Regressions”). Negative values are underlined. NA: not available.
Fig. 5. (A.2, B.2 and C.2) Robust Mahalanobis distance. Distances greater than the cut-off value (dashed lines) are considered outliers. (A.3, B.3 and C.3) Two different models are fitted in transformed data: i) GAM with outliers (solid lines) and ii) GAM without outliers (dashed lines).
These differences suggest that identifying outliers under the usual JMP analysis method is insufficient and requires additional tools. Therefore, the robust MD method both reinforces and complements the usual form of analysis. Furthermore, it allows current and other atypical values to be methodically identified, which reduces the identification bias. The disadvantage of the MD method is that the calculated distance represents the four parts (see Fig. 5A.2, B.2 and C.2), and therefore the exclusion of points that exceed the threshold leads to the loss of information for all four categories of the year. This does not happen with either the JMP method or the univariate statistics identification methods.

Following the sequence of the algorithm (Fig. 1), STEP III can be applied (Fig. 5). In Indonesia and South Africa, exclusion of outliers improved the quality of the models of all transformed data (Fig. 5A.3 and B.3). R-adjusted quality metrics confirm this affirmation. However, in Uruguay, quality metrics only improved in ilr3 transformations; this demonstrates the flexibility of GAM, which seeks to adjust to the data, regardless of whether it has outliers or not. On the other hand, although the models are generated in transformed data, it is more important in the WASH sector to see the quality of the predictive capacity in each category of analysis. Therefore, it is necessary to return the interpolations and extrapolations of the model to the space of the simplex, without ruling out that everything that happens in the transformed data will influence the results of the different levels of service.

The results of applying the inverse transformation in STEP IV of the algorithm are shown in Fig. 6. The presence of outliers influenced the fit of the models in a differentiated way; this affected the estimates. In Indonesia, the estimate of the percentage of rural population that have access to piped water in 2020 is 10.8% if the model was generated with data that includes outliers; however, this value decreased to 5.7% if outliers were excluded from the analysis, resulting in a 5.1% difference between the two estimates. In South Africa, this difference increased to 7.2% if we analyzed the category of the rural population that has access to sanitation through other improved forms.

For the 2020 estimates for Uruguay, there is no significant difference between the two alternatives (models with outliers included and without outliers). For example, for the category of access to piped water service, the difference between the two models was 0.004%. The remaining three categories also did not differ from these statements. It appears that in countries that have covered almost all water service provision, modeling and comparison are no longer relevant. Nonetheless, it cannot be ruled out that modeling is necessary for trend data to extreme values, as small proportions passed to population units can have significant effects, such as in China and India. On the other hand, we must emphasize two things: i) the estimates cannot exceed the extreme values of 0 and 100% in any service category; and ii) it is very important to use adequate statistical techniques, such as in STEP II, to treat values of zero, according to the variability of the time series data, as this allows models to be built without excluding data.

The quality metrics in Table 7 reinforce the hypothesis that outliers influence the quality of the models. The metrics of the four indicators are the same or better when outliers are excluded in six of the ten countries (namely, South Africa, Brazil, Indonesia, Nigeria, Uruguay and Benin). Of these countries, South Africa, Indonesia and Uruguay have NSE2 metrics near to 1, which indicates the high predictive capacity of the models in these countries, according to the indicator. The opposite is seen in Egypt, where the observed average is a better predictor than the model in the four analysis categories, in both models with or without outliers. In Bangladesh, Paraguay and Zambia, the improvement was only present in some categories.

On the other hand, the temporal trends of the service categories show the inequalities that exist in access to water and sanitation between the urban and rural sectors. In Indonesia and South Africa, access to water and sanitation by other improved forms is increasing (Fig. 6D.2 and E.2); however, in Uruguay, this category tends towards values of zero (Fig. 6F.2). If we compare only Indonesia and Uruguay, the rural-urban gap in the category of access to piped water is further increased, mirroring the world situation reported in the literature with respect to disparities that exist in access to water and sanitation in both sectors (Bain et al., 2014; Chitonge et al., 2020). That said, and in the context of the SDGs that seeks to ensure that no one is left behind (United Nations General Assembly, 2015), the rural sector in both Indonesia and South Africa is faced with a greater challenge in the provision and safe management of water and sanitation services.

Finally, after outliers have been identified, it is not recommended to eliminate them automatically, as this can lead to loss of relevant information that helps explain the specific situation or time series of the country. Additionally, there are other factors that the analyst does not value when excluding data (such as the cost of obtaining data through a survey, census or other alternatives that is representative of the country); therefore, the essential thing before excluding outliers would be to understand why the values are anomalous. An alternative that would help to understand the presence of these data could be to consult the institutions of origin for the information sources. Nevertheless, obtaining answers becomes complicated when it depends on third-party institutions (for instance, for reports to the SDG, the associated countries generally have statistical or other specialized institutions that are responsible for collecting, processing, and sharing information to interested parties). In these cases, exclusion is simply a necessity because of the improvements it brings to the models.

### 5. Conclusions

The existence of values of zero, missing values or both simultaneously makes it necessary to treat data in a differentiated manner, for which distinct treatment options are available. While these options are not equivalent, no clear criteria exist for choosing exactly which one to use, with all alternatives potentially equally good. Further, these options are suitable for analyzing data with variations in temporal evolution, which is not possible if we apply the multiplicative replacement (Martín-Fernández et al., 2003).

In countries with low amounts of data, we concluded that robust linear regression (robust OLS (ilr)) is suitable for the analysis of WASH sector data, since it limits the influence of outliers on the calibrated model. Both quantitatively and qualitatively, the declaration of outliers can be validated.

In countries with ≥ 6 data points, the identification of outliers with the robust Mahalanobis distance tends to give us more than the qualitative classification made with the JMP (and specifically, for nine of the ten countries evaluated), which reinforces the usual classification of the JMP. However, we must bear in mind that, in the robust Mahalanobis distance method, all parts of the year are excluded, while in JMP, only part of the composition is likely to be excluded. This conclusion goes

---

Table 6: Identification of outliers in WASH.

| Country           | Sector     | Service  | Data points (X) | Method      |
|-------------------|------------|----------|-----------------|-------------|
| South Africa      | Rural      | Sanitation| 30              | Improved X1 | X2 |
| Brazil            | Urban      | Water    | 27              |             | 0 0 |
| Indonesia         | Rural      | Water    | 26              |             | 3 4 |
| Nigeria           | Rural      | Water    | 22              |             | 1 0 |
| Paraguay          | Urban      | Water    | 21              |             | 8 0 |
| Bangladesh        | Rural      | Sanitation| 20              |             | 7 2 |
| Zambia            | Rural      | Water    | 16              |             | 3 6 |
| Egypt             | Urban      | Water    | 15              |             | 1 0 |
| Uruguay           | Urban      | Water    | 15              |             | 4 1 |
| Benin             | Rural      | Sanitation| 10              |             | 3 0 |

a: Robust Mahalanobis distance represents all parts at a single point, and those that exceed the threshold are considered outliers.
b: The JMP performs the puntual validation of data for each country. Data available at the JMP website (Country/Excel tab “Data Summary”).
Fig. 6. (D, E and F) Two different models are fitted in CoDa: i) Inverse of GAM transformed data with outliers (dashed lines) and ii) inverse of GAM transformed data without outliers (solid lines).
The statistical techniques applied for international monitoring of WASH are performed using CoDa. The peculiarities of this type of data suggest that there may be a spurious correlation if the appropriate statistical techniques are not applied in the analysis. However, in the WASH sector, this statement has not yet been tested. Thus, a correlation analysis of the population with access to different levels of WASH services should be carried out, as exemplified by the following situation.

As already discussed previously (Pérez-Foguet et al., 2017; Ezbakhe and Pérez-Foguet, 2019), the statistical techniques applied for international monitoring of WASH are performed using CoDa. The peculiarities of this type of data suggest that there may be a spurious correlation if the appropriate statistical techniques are not applied in the analysis. However, in the WASH sector, this statement has not yet been tested. Thus, a correlation analysis of the population with access to different levels of WASH services should be carried out, as exemplified by the following situation.

| Country                          | Data points (×) | Metric NSE |
|----------------------------------|----------------|-----------|
|                                  |                | With outliers (NSE1) | Without outliers (NSE2) |
|                                  |                | X1 | X2 | X3 | X4 | X1 | X2 | X3 | X4 |
| South Africa–rural sanitation    | 30 (×4)        | -0.002 | 0.59 | 0.84 | 0.12 | 0.39 | 0.91 | 0.93 | 0.70 |
| Brazil–urban water               | 27 (×4)        | 0.27 | 0.28 | 0.13 | 0.08 | 0.29 | 0.28 | 0.51 | 0.20 |
| Indonesia–rural water            | 26 (×4)        | -0.02 | 0.17 | 0.65 | 0.61 | 0.66 | 0.95 | 0.92 | 0.95 |
| Nigeria–rural water              | 22 (×4)        | 0.17 | 0.29 | -0.30 | -0.09 | 0.24 | 0.61 | 0.79 | 0.08 |
| Paraguay–urban water             | 21 (×4)        | 0.75 | 0.63 | 0.19 | 0.66 | 0.75 | 0.75 | 0.14 | 0.64 |
| Bangladesh–rural sanitation       | 20 (×4)        | 0.15 | 0.24 | 0.83 | 0.41 | 0.38 | 0.41 | 0.81 | 0.68 |
| Zambia–rural sanitation           | 16 (×4)        | -0.04 | 0.03 | 0.05 | 0.16 | 0.07 | 0.07 | 0.01 | -0.003 |
| Egypt–urban water                | 15 (×4)        | -0.06 | -0.04 | -0.78 | -0.02 | -0.07 | -0.08 | -1.61 | -0.02 |
| Uruguay–urban water              | 15 (×4)        | 0.15 | 0.54 | -0.03 | 0.06 | 0.88 | 0.75 | 0.44 | 0.88 |
| Benin–rural sanitation            | 10 (×4)        | 0.22 | 0.54 | 0.68 | 0.45 | 0.22 | 0.74 | 0.70 | 0.52 |

NSE2 values less than NSE1 are shown in gray; negative values are underlined.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was developed within the framework of a grant from the Peruvian government (Reference PRONABEC-President of the Republic Scholarship), through a full scholarship awarded to Alejandro Quispe Coica, and was partially funded by the Ministry of Science, Innovation and Universities of Spain (Ref: RTI2018-095518-B-C22) and by the Agència de Gestió d’Ajuts Universitaris i de Recerca de la Generalitat de Catalunya (Ref: 2017 SGR 1496).

Appendix A. Spurious correlation

As already discussed previously (Pérez-Foguet et al., 2017; Ezbakhe and Pérez-Foguet, 2019), the statistical techniques applied for international monitoring of WASH are performed using CoDa. The peculiarities of this type of data suggest that there may be a spurious correlation if the appropriate statistical techniques are not applied in the analysis. However, in the WASH sector, this statement has not yet been tested. Thus, a correlation analysis of the population with access to different levels of WASH services should be carried out, as exemplified by the following situation.

Two external organizations want to know the access of a population to water or to rural sanitation with different levels of services in Indonesia and South Africa, respectively. For this, organization A requests population data from the four categories \([X_1, X_2, X_3, X_4]\), while organization B only needs to know data for three of the four categories \([X_1^*, X_2^*, X_4^*]\). With the information obtained, organization A makes proportions to visualize the percentage of the population that each category represents in the composition. Organization B follows a similar procedure for all three categories of services. Subsequently, each organization performs a correlation analysis of the selected categories, reaching consistent and contradictory conclusions, according to the category of analysis.
From Table A1-1, organization A infers that the correlation of the categories of access to water between X1w and X3w is positive and low (correlation = 0.27), while organization B concludes that the correlation between X1w and X3w* has a high degree of relationship, but negative (correlation = −0.75). For the population with access to sanitation (Table A1-2), an analysis similar to that described, organization A concludes that the relationship between X3s and X4s is positive (correlation = 0.40), and organization B, that it is negative (correlation = −0.30). Therefore, for the same categories of analyses of both cases, the different methods of analysis give different conclusions.

| Category | Source  | Full composition “A” | Ratio | Subcomposition “B” | Ratio |
|----------|---------|----------------------|-------|--------------------|-------|
| Water    | SUS01   | [6.51, 60.45, 5.72, 27.32] | 1.14  | 4.78               | 1.14  | 4.78   |
|          | SUS02   | [6.17, 62.31, 5.46, 26.06] | 1.13  | 4.77               | 1.13  | 4.77   |
|          | DHS03   | [7.90, 57.09, 29.01]    | 1.32  | 4.83               | 1.32  | 4.83   |
| Sanitation| ESS00  | [5.05, 46.93, 21.88]     | 0.24  | 0.87               | 0.24  | 0.87   |
|          | CEN01   | [7.10, 47.23, 17.97]      | 0.26  | 0.65               | 0.26  | 0.65   |
|          | WHS03   | [30.59, 43.49, 12.92]     | 2.35  | 0.99               | 2.35  | 0.99   |

*To exemplify both water and sanitation, only data from the first three rows are shown.

On the other hand, when ratios are compared between categories for both organization A and organization B, the result is the same (Table A2). For example, for water access data, the ratio of [X4w/X3w] equals [X4w*/X3w*], and the same is true with the ratios of access to sanitation. Therefore, the alternatives for statistical analysis of CoDa are based on a log-ratio relationship approach (Aitchison, 1986; Egozcue et al., 2003), each with its own particularities. As the WASH sector is familiar with the assumptions made, the application of adequate statistics for CoDa becomes mandatory.

**References**

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

Aitchison, J., Kay, J.W., 2003. Possible solutions of some essential zero problems in compositional data analysis. In Proceedings of CoDaWork’03. The 1st Compositional Data Analysis Workshop. University of Girona, Girona (Spain).

Anthonj, C., Tracy, J.W., Fleming, L., Shields, K.F., Tikoisuva, W.M., Kelly, E., Thakkar, M.B., Christenson, E., Bartram, J.K., 2014. Rural:urban inequalities in post-2015 targets and indicators for drinking-water. Sci. Total Environ. 490, 509–513. https://doi.org/10.1016/j.scitotenv.2014.05.007.

Ashole Alto, A., Godana, W., Gedamu, G., 2020. Impact of community-led total sanitation composition and subcomposition of WASH data. Int. J. Environ. Health. 2020. https://doi.org/10.1101/2020.04.07.20057117.

Baquero, Ó.F., Jiménez Fdez. de Palencia, A., Pérez-Foguet, A., 2015. Reporting progress on the human right to water and sanitation through JMP and GLAAS. J. Water Sanit. Hyg. Dev. 5, 310–321. https://doi.org/10.2166/washdev.2015.151.

Brauer, M., Zhao, J.T., Bennett, F.B., Stanaway, J.D., 2020. Global Access to Handwashing: Implications for COVID-19 Control in Low-income Countries. medRxiv 2020.04.07.20057117. https://doi.org/10.1101/2020.04.07.20057117.

Cairncross, S., Hunt, C., Boisson, S., Bostoen, K., Curtis, V., Fung, I.C.H.H., Schmidt, W.-P., 2010. Water, sanitation and hygiene for the prevention of diarrhoea. Int. J. Epidemiol. 39, i193–i205. https://doi.org/10.1093/ije/dyp105.

Chen, J., Zhang, X., Hron, K., Templ, M., Li, S., 2018. Regression imputation with Q-mode clustering for rounded zero replacement in high-dimensional compositional data. J. Appl. Stat. 45, 2067–2080. https://doi.org/10.1080/02664763.2017.1410524.

Chitonge, H., Mokoena, A., Kongo, M., 2020. Water and Sanitation Inequality in Africa: Challenges for SGD 6. Springer, Cham, pp. 207–218 https://doi.org/10.1007/978-3-030-14857-7_20.

Craven, J., Giné–Garrriga, R., Jiménez Fdez. de Palencia, A., Pérez-Foguet, A., 2013. Introducing Hygiene Elements into Sanitation Monitoring. Loughborough University.

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795–828. https://doi.org/10.1007/s11004-005-7381-9.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric Logratio transformations for compositional data analysis. Math. Geol. 35, 279–300. https://doi.org/10.1023/A:1023818214614.

Ezhakhe, F., Pérez-Foguet, A., 2018. Multi-criteria decision analysis under uncertainty: two approaches to incorporating data uncertainty into water, sanitation and hygiene planning. Water Resour. Manag. 32, 5169–5182. https://doi.org/10.1007/s11269-018-2152-9.

Ezhakhe, F., Pérez-Foguet, A., 2019. Estimating access to drinking water and sanitation: the need to account for uncertainty in trend analysis. Sci. Total Environ. 696, 133830. https://doi.org/10.1016/j.scitotenv.2019.133830.

Ezhakhe, F., Giné–Garrriga, R., Pérez-Foguet, A., 2019. Leaving no one behind: evaluating access to water, sanitation and hygiene for vulnerable and marginalized groups. Sci. Total Environ. 683, 537–546. https://doi.org/10.1016/j.scitotenv.2019.05.207.

Filzmoser, P., Hron, K., 2008. Outlier detection for compositional data using robust methods. Math. Geosci. 40, 233–248. https://doi.org/10.1007/s11004-007-9141-5.
