

**To cite this article:** Pérez-Foguet, A., Giné-Garriga, R., Ortego, M.I., 2017. Compositional data for global monitoring: The case of drinking water and sanitation. *Sci. Total Environ.* 590–591, 554–565

**To link to this article:** <http://dx.doi.org/10.1016/j.scitotenv.2017.02.220>

## **Compositional data for global monitoring: the case of drinking water and sanitation**

A. Pérez-Foguet<sup>\*,1</sup>, R. Giné-Garriga<sup>\*,2</sup> and M. I. Ortego<sup>+,3</sup>

\* Research group on Engineering Sciences and Global Development, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya · BarcelonaTech

+ Compositional and Spatial Analysis COSDA Research Group, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya · BarcelonaTech

Email<sup>1</sup>: [agusti.perez@upc.edu](mailto:agusti.perez@upc.edu)

Email<sup>2</sup>: [ricard.gine@upc.edu](mailto:ricard.gine@upc.edu)

Email<sup>3</sup>: [ma.isabel.ortego@upc.edu](mailto:ma.isabel.ortego@upc.edu)

# **Compositional data for global monitoring: the case of drinking water and sanitation**

## **Abstract**

Introduction: At a global level, access to safe drinking water and sanitation has been monitored by the Joint Monitoring Programme (JMP) of WHO and UNICEF. The methods employed are based on analysis of data from household surveys and linear regression modelling of these results over time. However, there is evidence of non-linearity in the JMP data. In addition, the compositional nature of these data is not taken into consideration. This article seeks to address these two previous shortcomings in order to produce more accurate estimates.

Methods: We employed an isometric log-ratio transformation designed for compositional data. We applied linear and non-linear time regressions to both the original and the transformed data. Specifically, different modelling alternatives for non-linear trajectories were analysed, all of which are based on a generalized additive model (GAM).

Results and discussion: Non-linear methods, such as GAM, may be used for modelling non-linear trajectories in the JMP data. This projection method is particularly suited for data-rich countries. Moreover, the ilr transformation of compositional data is conceptually sound and fairly simple to implement. It helps improve the performance of both linear and non-linear regression models, specifically in the occurrence of extreme data points, i.e. when coverage rates are near either 0% or 100%.

## **Keywords**

Water, sanitation and hygiene; Service ladder; Compositional data; Log transformation; Joint Monitoring Programme (JMP) of WHO and UNICEF

## 1. Introduction

For the past 15 years, the Millennium Development Goals (MDGs) have challenged the international community to reduce by half the proportion of the population without safe drinking water and basic sanitation. At the global level, the target for safe drinking water was met in 2010, and over 90 per cent of the world's population now has access to improved sources of drinking water. In contrast, the world has fallen short on the sanitation target, leaving 2.4 billion without access to improved sanitation facilities (Joint Monitoring Programme, 2015a).

Since 1990 and throughout this period, the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP) has monitored progress by producing national, regional and global estimates of population using improved drinking water sources and improved sanitation facilities. As the only available source of comprehensive and internationally-comparable data, the JMP has served as the UN-recognised instrument for monitoring progress towards the MDG target. Specifically, JMP has reported on two separate service "ladders" (Joint Monitoring Programme, 2008). The sanitation ladder reports on the proportion of population with: no sanitation facilities at all; reliant on technologies defined as "unimproved"; sharing sanitation facilities of otherwise acceptable technology; and using "improved" sanitation facilities. Similarly, the drinking-water ladder reports on the proportion of those: using drinking water directly collected from surface water; using other unimproved water sources; using "improved" sources other than piped household connections; and benefiting from household connections in a dwelling, plot or yard. The ladder approach provides a powerful tool for supporting decisions about planning, monitoring and evaluation, targeting and reporting, as it allows countries to strive for higher levels of service (e.g. piped on premises) while ensuring that those with no service (surface water, open defecation) are prioritized (Kayser et al., 2013; Moriarty et al., 2011; Potter et al., 2011).

The principal data sources used by JMP are national censuses and nationally representative household surveys. Yet during the MDG period of 1990 to 2015, the JMP has not merely reported on the latest survey findings but has also published model estimates using simple linear regression. Linear regressions average small differences in coverage between surveys, provide estimates for years in which no survey data are available, and are relatively easy to explain to policy makers and practitioners responsible for water and sanitation service delivery. In addition, using linear regression was an adequate response to limited availability of data at the start of the MDG period. However, there is evidence of non-linearity in the JMP data, and various non-linear patterns have been observed (Bartram et al., 2014; Fuller et al., 2015; Wolf et al., 2013). In the presence of non-linearity, non-linear trajectories can improve the accuracy of estimates and projections.

With the end of the MDG era in 2015, the emphasis has shifted to the Sustainable Development Goals with new targets for the year 2030 (United Nations General Assembly, 2015, 2014). In preparing for the new monitoring framework, the JMP has facilitated international consultations on post-2015 targets and indicators (Joint Monitoring Programme, 2012, 2011) and has also reviewed the current JMP method for deriving estimates of coverage (Joint Monitoring Programme, 2015b). A background paper presented during the JMP taskforce on methods identified patterns of curvature in the data and analyzed different modelling alternatives for these non-linear trajectories (Fuller et al., 2015, 2014). Complementary to this work, and to further support the JMP in the task of improving the reporting methods, we now have investigated the effects of considering the compositional nature of data used

for estimating service levels (e.g. proportions that sum to 1) in time interpolation models. Specifically, this article discusses the suitability of different regression approaches for estimating the population using drinking water and sanitation at the national level, against two complementary criteria: i) accuracy of estimates to report on different service levels, and ii) replicability and ease of communication to non-specialists. The ultimate aim is a more consistent identification by country of those who suffer from inadequate levels of service. In doing so, this article helps to unravel the interdependencies between social processes and the water cycle. Notably, other sectors (e.g. energy) are likely to adopt the ladder approach in their monitoring and reporting frameworks. Thus, our main findings may have the potential for wider implementation (Banerjee et al., 2013; Sustainable Development Solutions Network, 2015) and may be applicable to other spheres of the total environment.

In detail, this study:

- employs an isometric log-ratio transformation designed for compositional data (Egozcue et al., 2003; Pawlowsky-Glahn et al., 2015);
- applies linear and non-linear time regression models to both the original and the transformed data. Starting from the previous study developed by Fuller et al. (2015), we compared an ordinary least squares (OLS) linear regression—currently used by the JMP—to a generalized additive model (GAM), in which the linear form is replaced by a sum of smooth functions (Hastie and Tibshirani, 1987, 1986, Wood, 2006, 2004);
- analyses three patterns of curvature in addition to linear trajectories: i) saturation, ii) acceleration, and iii) deceleration. Other non-linear patterns, such as negative acceleration and negative deceleration, were not considered as they are very rare (Fuller et al., 2015);
- models indicators separately (e.g. improved drinking water against the other water service ladder proportions aggregated in a single value) and simultaneously (e.g. improved sanitation, shared but unimproved sanitation and open defecation in a joint analysis), in order to account for the individual populations as parts of the whole.

## **2. Background: the issue of compositional data**

Compositional data (CoDa) are arrays of positive components representing parts of a whole. Their main characteristic is that multiplication by a positive constant does not change the information contained in it, i.e. the relevant information is contained in the ratios between components (Pawlowsky-Glahn et al., 2015). Frequently, CoDa are normalized so that their components add to a constant (e.g., 100, one, a million). By definition, the JMP data are compositional, i.e. individual populations in the dataset are not independent of each other but are related by being expressed as a percentage of the total.

The problems of undertaking statistical analyses with compositional data have been widely discussed in literature, mostly in connection with multivariate data analysis (Filzmoser et al., 2009; Lloyd et al., 2012; Pawlowsky-Glahn et al., 2015). However, results from these works have not reached the wider academic community. Although some practitioners believe that the application of classical univariate

statistical methods to CoDa is methodologically sound, compositional data are inherently multivariate. If the compositional character is ignored, spurious correlations and subcompositional incoherencies appear. Thus, CoDa should never be seen as truly univariate data.

Too often, statistical analysis of CoDa focuses on one component and the remainder. This remainder is built as the sum of all remaining components (also called amalgamation). If all variables were measured, one could omit one data dimension (variable) without any loss of information, due to their compositional character. For instance, if a population sample was analysed for all possible individual populations, all of these populations would sum up to 100%. As a consequence, the data belong to a subspace of the Euclidean space, the simplex, which has its own geometrical structure, and real Euclidean geometry is thus inappropriate for such data. The special geometry of the simplex is the so-called Aitchison geometry. The simplex endowed with the operations and distance defined in Aitchison geometry is an Euclidean space (Aitchison, 1986; Pawlowsky-Glahn et al., 2015).

Euclidean geometry plays an important role in statistical data analysis, even in the univariate case. Problems arise when statistical analysis of compositional data is undertaken without considering their own structures. These problems cannot be overstated. Even in the apparently simple construction of a histogram, one counts the number of data points falling into certain intervals with equal length, measured by Euclidean distance. Therefore, a histogram may not reveal the true distribution as inherent in the compositional data. Major problems emerge when computing the Pearson correlation coefficient between parts, as contradictory correlations often appear—the so-called spurious correlation. Remarkably, a significant number of statistical methods are based on the correlation structure of data. Applying these statistical methods or tests to the raw compositional data would lead to erroneous or misleading results. In order to apply statistical procedures to compositional data, the Aitchison distance should be used. However, adequate understanding of this geometry by practitioners remains elusive. To use a real Euclidean structure, the composition may be transformed to the real scale by applying the so-called principle of working in coordinates (Pawlowsky-Glahn et al., 2015). This allows the usual statistical procedures (based on Euclidean distance) to be applied to the transformed data, and the results can then be back-transformed to the original data scale for interpretation.

Various possibilities for data transformation of compositional data have been introduced in the literature; the most widely used is the family of one-to-one log-ratio transformations (Filzmoser et al., 2009). The additive log-ratio (alr) and the centred log-ratio (clr) were the first alternatives suggested for transforming compositional data (Pawlowsky-Glahn et al., 2015). Outputs from the alr and clr transformations are however subject to some restrictions in their treatment by standard methods (Lloyd et al., 2012). The isometric log-ratio (ilr) transformation is another useful class of log-ratio transformations with good theoretical properties (Egozcue et al., 2003). It represents the inner multivariate data structure of compositions in a new geometry on the simplex that relates directly to the usual Euclidean geometry (Egozcue and Pawlowsky-Glahn, 2006). Using other transformations prior to the formalization of the structure of compositional data has been suggested in literature. For instance, the logit transformation is a valid way to analyse a two-part composition but has serious problems when there are more than two parts (Lloyd et al., 2012). For multivariate data analysis, the ilr transformation should be used (Filzmoser et al., 2009; Lloyd et al., 2012). A simple log transformation, variable by variable, or any other linear transformation of the single variables may not be an appropriate solution, as the compositional character of data is lost.

An example that illustrates the previous statement is given by a triple-nested logit approach, which can be compared to an ilr transformation, as both are applied to a compositional vector of four parts  $x = (x_1, \dots, x_4)$  such as  $x_1 + x_2 + x_3 + x_4 = 1$ . The nested logit approach is first applied to a binary partition of the whole:  $(x_1 + x_2)$  and  $(x_3 + x_4)$ , and then to the respective subparts,  $x_1$  and  $x_2$  and  $x_3$  and  $x_4$ . This leads to three transformed variables:

$$\begin{aligned} v_1 &= \log \left( \frac{x_1 + x_2}{x_3 + x_4} \right), \\ v_2 &= \log \left( \frac{x_1}{x_2} \right), \\ v_3 &= \log \left( \frac{x_3}{x_4} \right), \end{aligned}$$

with the first one proportional to the logarithm of the sum ratio of the components. The ilr approach is applied to all four variables, with a base transformation compatible with the nested approach of logit in order to facilitate comparison of both models. This leads to three balances:

$$\begin{aligned} b_1 &= \sqrt{\frac{4}{4}} \log \left( \frac{(x_1 x_2)^{1/2}}{(x_3 x_4)^{1/2}} \right) = \frac{1}{2} \log \left( \frac{x_1 x_2}{x_3 x_4} \right), \\ b_2 &= \sqrt{\frac{1}{2}} \log \left( \frac{x_1}{x_2} \right), \\ b_3 &= \sqrt{\frac{1}{2}} \log \left( \frac{x_3}{x_4} \right), \end{aligned}$$

Therefore, the first transformed variable with the ilr approach is proportional to the products of the components, and not to the sums of nested logit. The second and third transformed variables are proportional to the same ratios in both approaches. The ratios of the sums approximate the ratios of the products if only if all four variables are around 0.25. In this case, both alternatives give similar formulations. Thus, the triple logit can only be considered as a good approximation of the ilr-based approach if data are homogeneously balanced; i.e. when none of the four proportions approach limit values.

The JMP data are compositional data with a temporal evolution. In the literature, various approaches have been adopted to treat these compositional data. Some studies first use the alr transformation to express the compositions in the space of coordinates, and then apply standard statistical techniques, such as regression with a temporal covariate (Barceló-Vidal et al., 2011; Brunsdon and Smith, 1998; Mills, 2010). The isometric log-ratio transformation for compositional time series was introduced by Bergman (2008). More recently, Kynčlová et al (2015) proposed a special choice of ilr transformation and apply vector autoregressive (VAR) methods to the coordinates.

### 3. Methods

The JMP currently monitors access to sanitation and drinking-water services in 212 countries and territories. For each country, the JMP estimates are based on fitting a regression line to a series of data points from household surveys and censuses conducted by governments and other organizations. Some countries have very few data points available, while others have more than 20 (Fuller et al., 2015). An OLS linear regression was conducted to estimate the proportion of the population using the

following drinking water sources: i) on-site piped drinking water (piped water in the home, yard or plot) ii) any other improved drinking water source, and iii) water collected from a surface water source; and the following sanitation facilities: i) improved types of sanitation (including shared facilities of an improved type), and ii) open defecation. The remaining population uses unimproved drinking water sources and unimproved sanitation facilities. The percentage of the population that shares a sanitation facility of an otherwise improved type was subtracted from the trend estimates of improved sanitation facilities. The rural and urban coverage rates are reported separately, with totals based on a weighted average using the latest official data. Regression lines were extended for up to two years before the earliest and after the most recent census or survey year (but constrained within the coverage range 0–100%). For coverage in the range 5–95%, the coverage at the end of the two-year extrapolation was reported for up to four further years, after which no further data were reported (i.e., reports show “data unavailable”); for countries outside this range, the continuation was extended indefinitely (Bartram et al., 2014; Joint Monitoring Programme, 2015a).

In a previous study, a visual inspection of the data points (i.e. official JMP estimates for each country) was performed to identify the trajectory for a given country in relation to each of the indicators (Fuller et al., 2015). Data follow different patterns: 1) 100% coverage (all data points are  $\geq 98\%$ ), 2) linear growth, 3) linear decline, 4) no change (the slope for the entire period is close to zero), 5) saturation, (positive slope but plateauing near 100%), 6) acceleration (increasingly positive slope), 7) deceleration (positive slope but plateauing below 100%), 8) negative acceleration (increasingly negative slope), and 9) negative deceleration (negative but plateauing slope). Among these are non-linear trajectories (saturation, acceleration, deceleration, negative acceleration or negative deceleration), which the JMP method will not accurately model.

To illustrate the level of error when a linear projection method is applied to non-linear data, this method (i.e. OLS linear regression) was compared to a GAM. The GAM is flexible for modelling non-linear relationships and has the advantage of being completely automatic, i.e. it requires fewer subjective decisions to be made (Hastie and Tibshirani, 1986). The GAM procedure is typically applied with thin-plate regression splines with four degrees of freedom (Wood, 2003). In this study, however, the *mgcv* package (v1.8-15) for R (v3.3.1) was used, as it has an interface that enables the user to specify other limit values for degrees of freedom and other interpolation basis (Wood, 2006). Specifically, various non-linear projection methods are compared: i) thin-plate spline, four degrees of freedom; ii) thin-plate spline, six degrees of freedom; iii) thin-plate spline, eight degrees of freedom; and iv) cubic spline. All of these were applied to both the raw data and the transformed data. As suggested in the literature, the zero values were replaced by small values before computing the log-ratios (Aitchison, 1986; Lloyd et al., 2012). Different interpolations methods were compared graphically and by means of two parameters frequently used to measure the differences between values predicted by a model and the values actually observed: the root-mean-square error (RMSE) and the Nash-Sutcliffe Efficiency coefficient (NSE).

We considered two levels of analysis in the discussion. First, variables were modelled separately (with one variable, such as improved drinking-water, and the reminder); for simplicity, results only included the comparison between the OLS linear regression and the standard GAM (thin-plate regression splines with four degrees of freedom). Four different alternatives were therefore analysed: i) OLS linear regression; ii) standard GAM; iii) OLS regression after performing an isometric log-ratio transformation of the dependent variables; and iv) standard GAM after performing an isometric

log-ratio transformation of the dependent variables. The second analysis simultaneously modelled all variables in order to account for the individual populations as parts of the whole. Thus, drinking-water indicators (piped water on premises, improved water, unimproved water and surface water) and sanitation indicators (improved sanitation, unimproved sanitation and open defecation) were jointly analysed, in two different groups. In this case, a comparison was made between thin-plate regression splines with four, six and eight degrees of freedom, and also standard cubic splines. For all countries, the same projection rules as used by the JMP were applied.

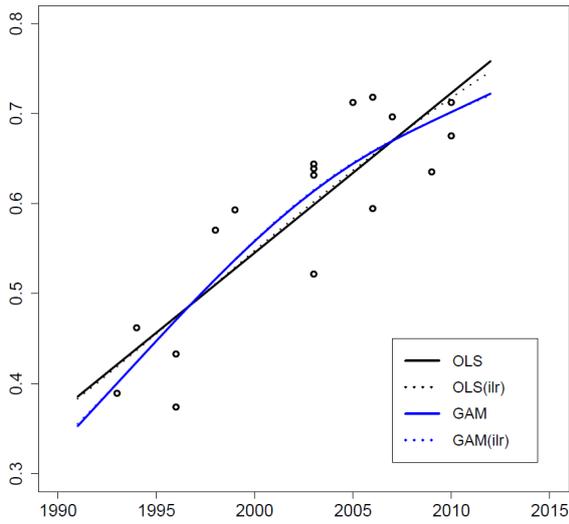
#### **4. Results and discussion**

In this section, we compare the results obtained by applying the various projection models to both the official JMP data and the ilr-transformed data. Specifically, we aimed to determine how these models performed when different patterns in the data are observed, i.e. a linear trajectory and three non-linear trajectories (saturation, acceleration and deceleration). We will first present the analysis of countries with 11 or more data points, as the performance of a GAM depends on the sample size. These countries account for roughly one-fifth of all countries (16–23%, depending on the indicator). We then discuss some examples from countries with 6–10 data points, which also represent approximately a fifth of all countries (15–25%) (Fuller et al., 2015).

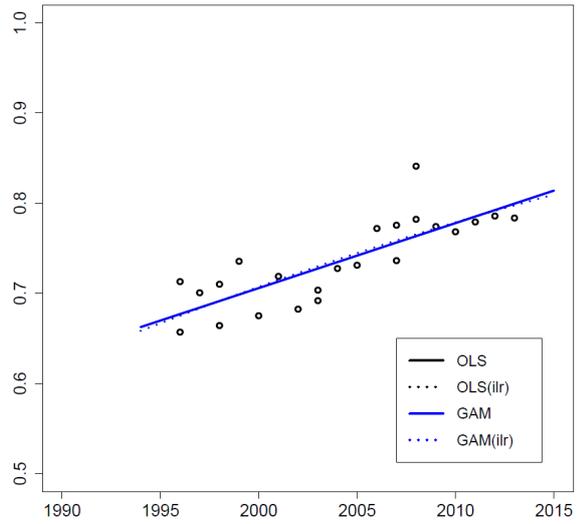
##### *4.1 Countries with more than 10 data points*

The majority of countries with 11 or more data points (62.0–85.4%, depending on the indicator) showed linear trajectories. Some countries had linear decline, and 2.1–20.6% of countries showed no evidence of any change in coverage over the MDG period. Non-linear trajectories were seen in 14.7–38.0% of countries (Fuller et al., 2015).

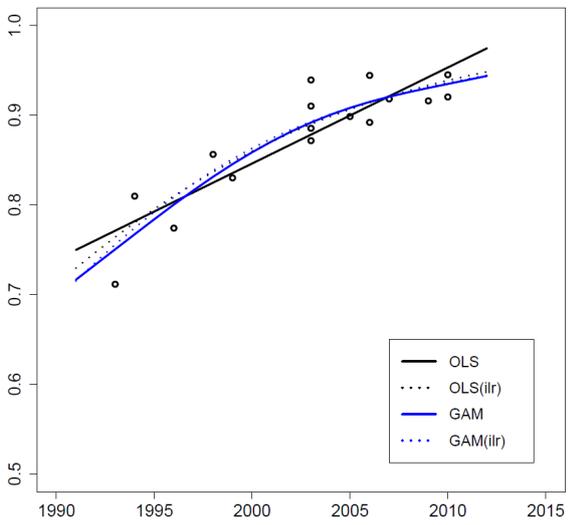
The most common linear pattern was that of linear growth, which indicates that a country is making linear progress in coverage. Specifically, between 43.8–68.6% of countries showed a pattern of linear growth, depending on the indicator. However, while a GAM was able to fit a linear trajectory, it was not any better under these circumstances than OLS, as one could expect (Fig. 1a). The ilr transformation generated a curve with slower growth at high levels.



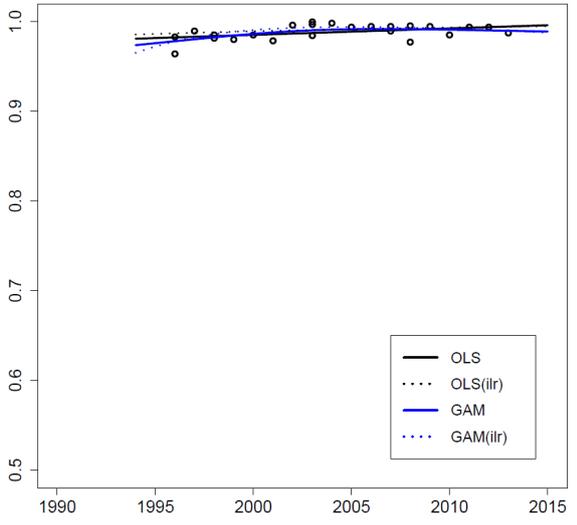
a.1) Improved water in rural Burkina Faso



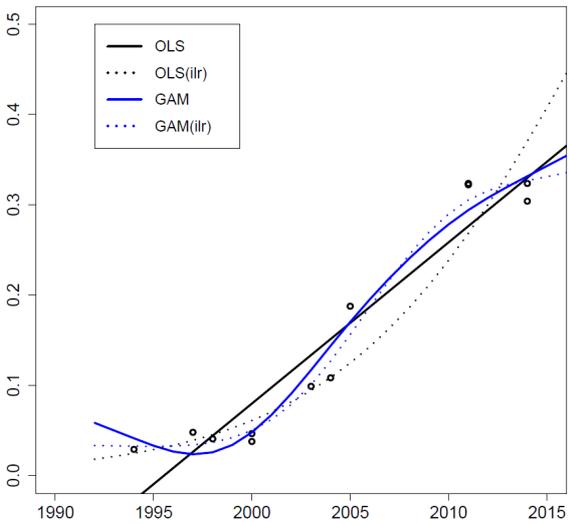
a.2) Improved water in rural South Africa



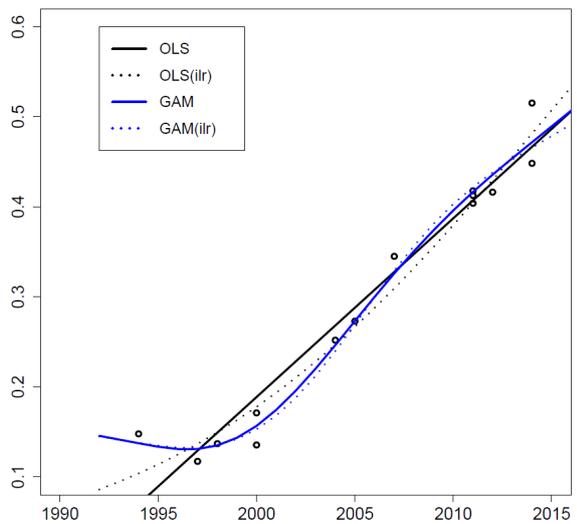
b.1) Improved water in urban Burkina Faso



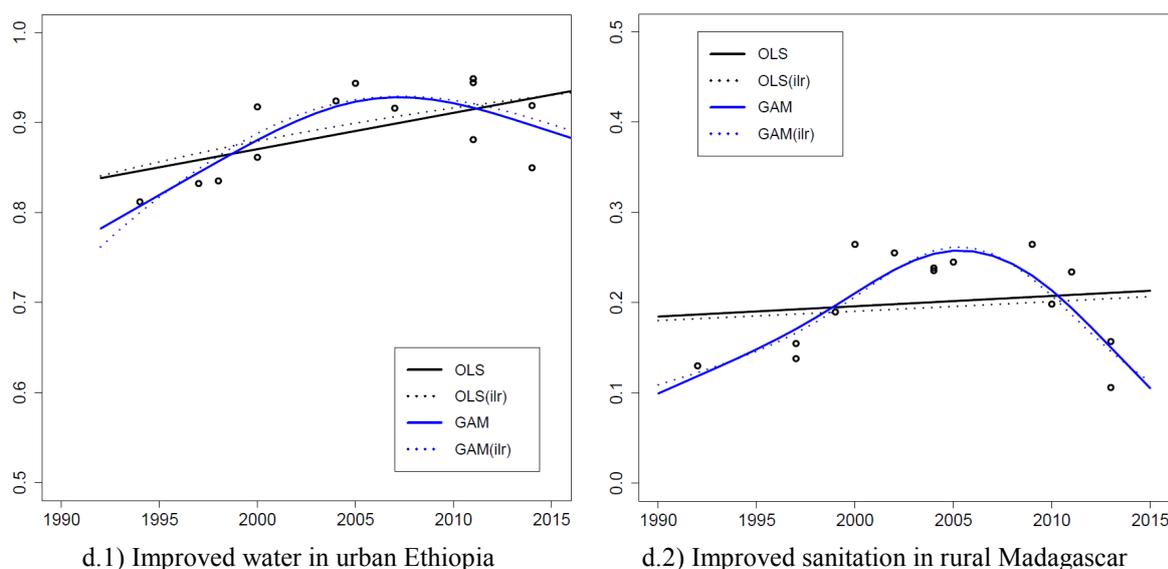
b.2) Improved water in urban South Africa



c.1) Improved sanitation in rural Ethiopia



c.2) Improved water in rural Ethiopia



**Fig. 1.** Examples in the JMP data of (a) linear growth (linear progress in coverage); (b) saturation (slowing down of progress as approaching 100% coverage); (c) acceleration (no progress followed by progress); (d) deceleration (progress followed by less or no progress). Two different models are fitted to both the original data and the transformed data: *i*) Generalized Additive Model (GAM) (blue lines), *ii*) ordinary least squares (OLS) (black lines). The projections referenced as OLS (ilr) and GAM (ilr), which account for the multidimensional compositional nature of the data, are shown as dashed lines.

Saturation occurs when progress slows as coverage reaches 100%. A country's progress may be linear until it approaches full coverage, at which point the rate of progress typically flattens out. Saturation is for instance a common non-linear pattern for the inverse of open defecation in urban areas (Fuller et al., 2015). Achieving full coverage, however, should be a priority (Joint Monitoring Programme, 2015c; United Nations General Assembly, 2014, 2010), and dedicated programmes for the poorest and most marginalized should be in place to ensure that no one is left behind as a country moves from 99% to 100% coverage. Two clear examples of saturation are shown in Figure 1b. When saturation is present, the official JMP estimates overestimate coverage in early years but underestimate coverage in intermediate years. In late years, the coverage is likely near 100%. Since the current JMP method will never predict more than 100%, JMP estimates are probably accurate for these later years, i.e. they integrate in the worst case rounding errors. The models based on transformed data perform well for saturation trajectories, and remarkably they are also able to fit minor gradual decreases of coverage (e.g. water in urban South Africa).

Acceleration occurs when a country makes substantial progress after a period of no, little or less progress. Acceleration is a common non-linear pattern for improved sanitation in rural areas, improved drinking water in rural areas and piped drinking water in rural areas (Fuller et al., 2015). Two examples of acceleration are shown in Figure 1c. Whenever there is an accelerated increase in coverage, the official JMP estimates tend to underestimate the level of coverage in early years, and to overestimate in the later years. In the intermediate years, the JMP estimates are mainly overstated as compared to the GAM outputs. The major difference between the two non-linear models (original GAM and GAM with ilr transformation of the data) is that the latter tends to flatten the curve at the end of the studied period.

Deceleration occurs when a country is making progress, but then progress abruptly stalls, or even declines, before reaching a high level of coverage. There were fewer examples of deceleration in the data, but to some extent it was common in urban sanitation (Fuller et al., 2015). The rural-urban migration and rapid urbanisation is a likely cause of the linear decline of coverage in urban areas. Figure 1d shows two examples of deceleration. When a country's progress in coverage stagnates, the official JMP values tend to overestimate the level of coverage in early and later years, but to underestimate it in intermediate years. At first sight, the non-linear models perform similarly well in both examples. The projections based on ilr-transformed data are however able to deal with extreme cases, i.e. when the predicted coverage values approach 0% or 100% (e.g. sanitation in rural Madagascar). In contrast, projection methods based on raw data can yield predicted values that are less than 0% or greater than 100%. These predictions have to be manually corrected.

In a complementary analysis, Table 1 summarizes the values of RMSE and NSE parameters, which have been applied to all projections depicted in Figure 1. Both indicators were computed from the sum of the square differences between the estimated parameter and its estimator. No significant improvement was observed when non-linear trajectories were applied to model a linear behaviour or saturation (Table 1). In one specific case (a.2), the RMSE and NSE present exactly the same values, as the two trajectories (GAM versus OLS) are also the same. The values for non-transformed and ilr-transformed trajectories are also very similar. In another case (b.2), RMSE shows low values for all cases, indicating that the models accurately fit the data. The NSE values are also low, indicating that model predictions are not far from considering a constant value, consistent with the observed flat pattern. At the same time, however, we observed a clear trend when data are accelerating (c.1 and c.2) or decelerating (d.1 and d.2): in both cases, the model accuracy increases considerably if a GAM is applied.

**Table 1** Values of the root-mean-square error (RMSE) and the Nash-Sutcliffe Efficiency coefficient (NSE) for results of models presented in Figure 1

	RMSE ( $\cdot 1E-2$ )				NSE ( $\cdot 1E-1$ )			
	OLS	OLS (ilr)	GAM	GAM (ilr)	OLS	OLS (ilr)	GAM	GAM (ilr)
Figure a.1	5.62	5.50	5.18	5.16	7.41	7.51	7.80	7.81
Figure a.2	2.71	2.72	2.71	2.72	6.55	6.51	6.55	6.51
Figure b.1	2.99	2.62	2.49	2.52	7.79	8.31	8.48	8.44
Figure b.2	0.73	0.78	0.65	0.68	1.98	0.75	3.72	2.97
Figure c.1	3.52	3.80	2.12	1.54	9.17	9.04	9.70	9.84
Figure c.2	3.09	2.45	1.77	1.91	9.47	9.67	9.82	9.80
Figure d.1	3.78	3.70	2.60	2.70	3.28	3.58	6.83	6.58
Figure d.2	5.22	5.26	2.74	2.84	0.19	0.04	7.29	7.09

In summary, we found that non-linear methods that account for curvature are adequate to deal with the different patterns observed in the official JMP estimates. In other words, we have shown that a GAM correctly fits a linear trajectory as well as all three forms of curvature. Importantly, the GAM-based projection provides a technical advancement with respect to the current JMP method when datasets show nonlinear behaviours. However, criteria for evaluating the suitability of this modelling approach should not only relate to the accuracy of the model, but also to its plausibility as well as its practical implementability. Different criteria should be adequately balanced before selecting the

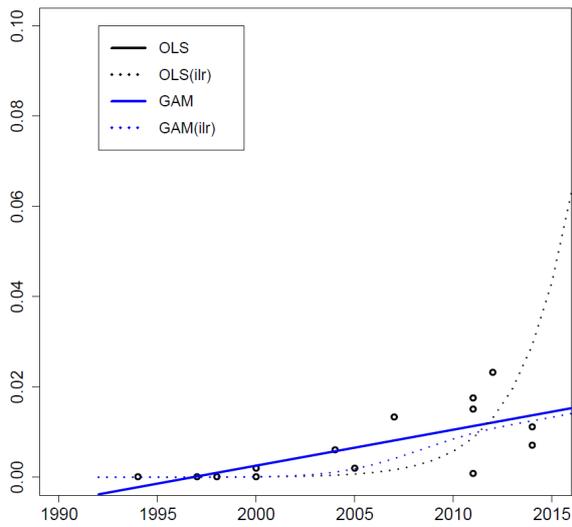
method employed for estimating the population using drinking water and sanitation (Joint Monitoring Programme, 2015b): i) accuracy of estimates and projections; ii) consistency in estimating service levels (e.g. open defecation, piped on premises) and disparities in access (e.g. wealth quintiles, urban or rural); iii) replicability by non-specialists; and iv) ease of communication to non-specialists and decision makers. There is typically a trade-off: those methods that are easiest to explain and implement may also be less accurate. It is critical to note, however, that a statistical method that accurately produces water and sanitation estimates yet cannot be easily understood by non-technical stakeholders may not be the best choice for global monitoring purposes.

In addition, the problems resulting from the analysis of compositional data using standard statistical methods should be addressed, and the compositional nature of JMP data should be considered for statistical data analysis. The examples we presented above illustrate that applying statistical tests to the raw data would, by definition, lead to erroneous results. These statistical procedures can, however, be carried out with the ilr-transformed data, and afterwards back-transformed into the original data scale. Indeed, the ilr transformation is conceptually sound and is fairly simple to implement. It thus has the potential for wider acceptability and utilization. The graphs in Figure 1 show that the GAM using ilr-transformed data generates a prediction that will not necessarily be the same as the prediction done with original data. Specifically, one advantage of basing the projection on ilr-transformed data is the ability to deal with extreme values when coverage levels approach 0% or 100%. With raw data, the level of coverage has to be manually corrected and constrained within the range 0–100%, particularly when models yield predicted values that are less than 0 or greater than 100%.

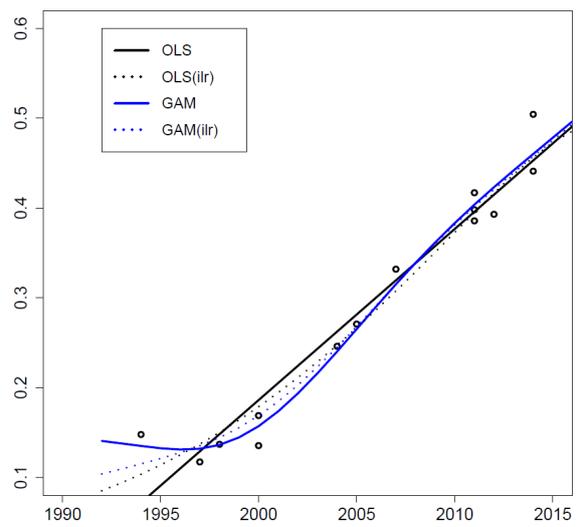
***Accounting for other compositional parts:*** The examples we have discussed so far assume that for the variable of interest  $x_1$  (e.g. improved sanitation), only the relationship to the remainder will be investigated. In other words, for a considered variable  $x_1$ , the remainder has been assumed to be  $1-x_1$ , since nothing is known about other parts, or because only a single part and its relative relationship to the rest of the whole is of interest.

If the values of additional variables  $x_2, \dots, x_n$  are known (e.g. unimproved sanitation and open defecation), and if the relationship of one variable to each of the other existing variables needs to be considered, the approach proposed can be readily extended. The ilr transformation of the “ $n$ ” variables produces a new set of  $(n-1)$  variables. Technically, the original variables are being projected onto a new base of orthonormal vectors (Pawlowsky-Glahn et al., 2015). The transformed data (called balances in the compositional vocabulary) can then be modelled either with OLS or GAM, and the interpolated results can subsequently be back-transformed to the original set of variables.

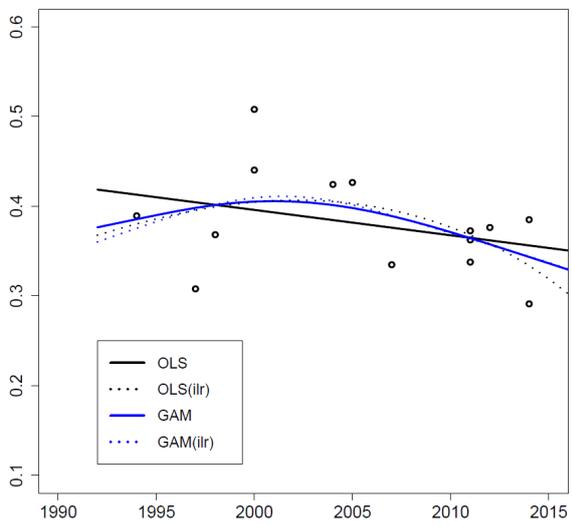
The JMP defines coverage through four different levels of service, and we present each proportion of the population separately in following examples. Note that, in Figure 2, we plotted the results obtained by performing four OLS regressions and four GAMs, using one separate projection per population group, such that the figure shows the specific contribution of each compositional part when modelled by four different projections.



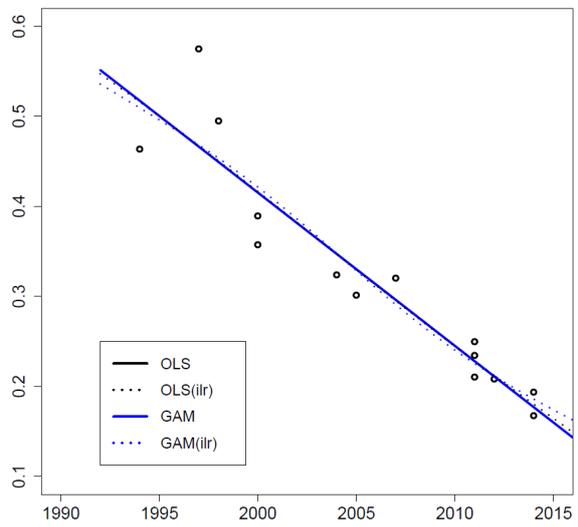
a.1) Piped water



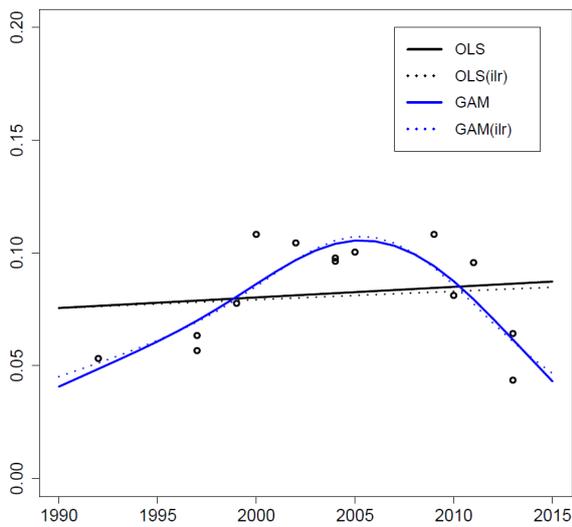
a.2) Other improved



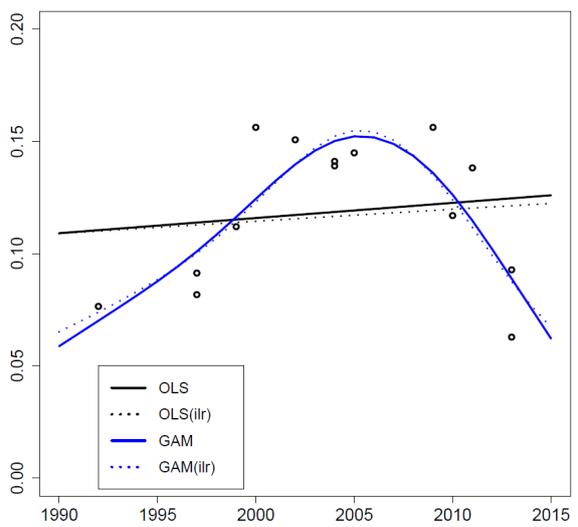
a.3) Other unimproved



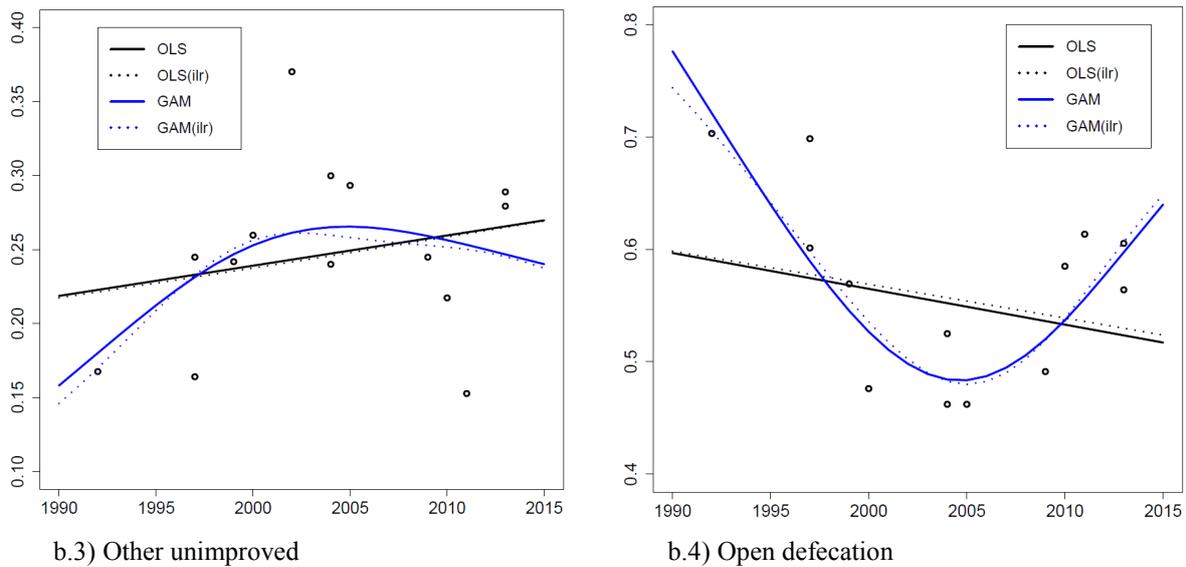
a.4) Surface water



b.1) Improved sanitation



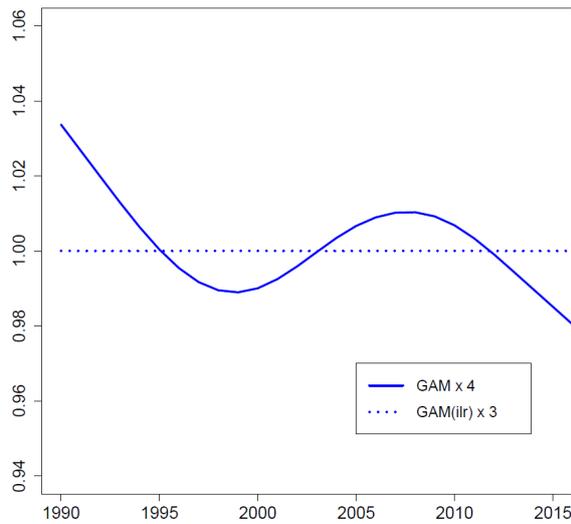
b.2) Shared sanitation



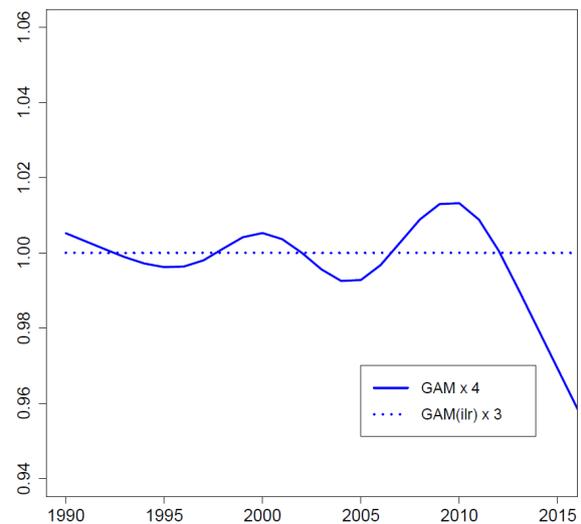
**Fig. 2.** Trajectories in the JMP data when coverage estimates are analyzed as compositional data, i.e. they are considered to be parts of a whole: (a) Drinking-water in rural Ethiopia; (b) Sanitation in rural Madagascar. Two different models are fitted to both the original data and the transformed data: i) Generalized Additive Model (GAM) (blue lines), ii) ordinary least squares (OLS) (black lines). The projections referenced as OLS (ilr) and GAM (ilr), which account for the multidimensional compositional nature of the data, are shown as dashed lines.

As previously mentioned, the outputs of a GAM are based on a small set of assumptions about the functional form of the relationship between time and coverage. It is characterized among others by the smoothing spline adopted to fit the curve and the degrees of freedom, but without an a priori hierarchy of trajectories. GAM is flexible and data-oriented. Because of this flexibility, however, GAM will fit the patterns observed for the different variables of interest separately, and nothing guarantees that the percentages related to different GAMs sum up to 100%. This situation repeats itself for all projection methods that do not account for the compositional nature of the data, except for the particular case of OLS. They will by definition add up to 100%, although there is no way to guarantee that all four values separately lie in the range from 0 to 1.

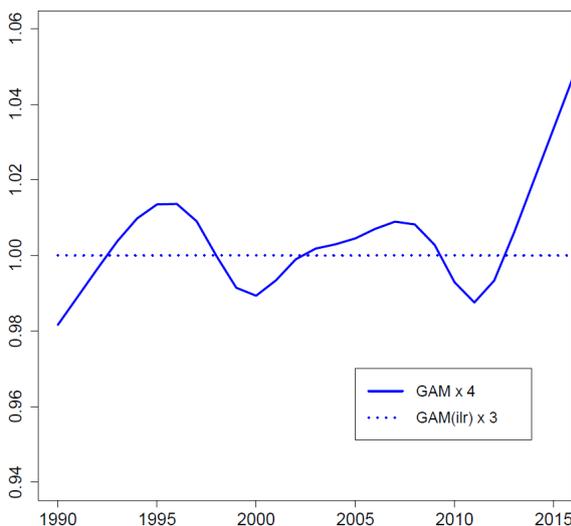
We illustrate this concept with the example shown in Figure 3: various GAMs generated percentage values that ranged from 98% to 102%. It seems intuitive that, with higher degrees of freedom, the error reported is reduced within the studied period. However, the error may increase significantly at both ends. In contrast, projections based on ilr-transformed data added up to one in all cases, as would be expected.



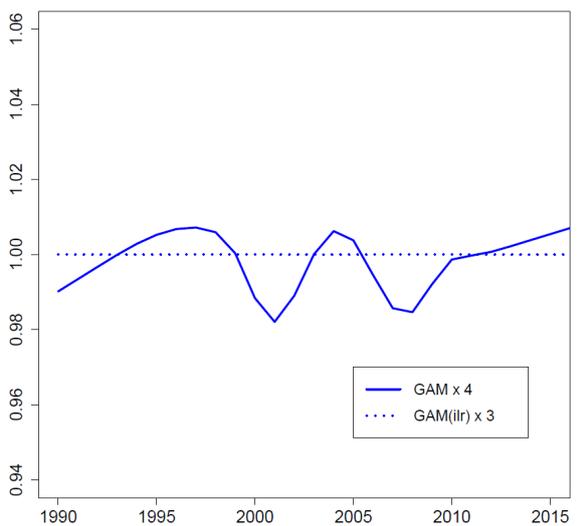
a) GAM fitted by thin-plate splines, four degrees of freedom



b) GAM fitted by thin-plate splines, six degrees of freedom



c) GAM fitted by thin-plate splines, eight degrees of freedom



d) GAM fitted by cubic splines

**Fig. 3.** Sum of population percentages when coverage levels (sanitation in rural Madagascar) are estimated separately as compositional data, i.e. they are considered to be parts of a whole.

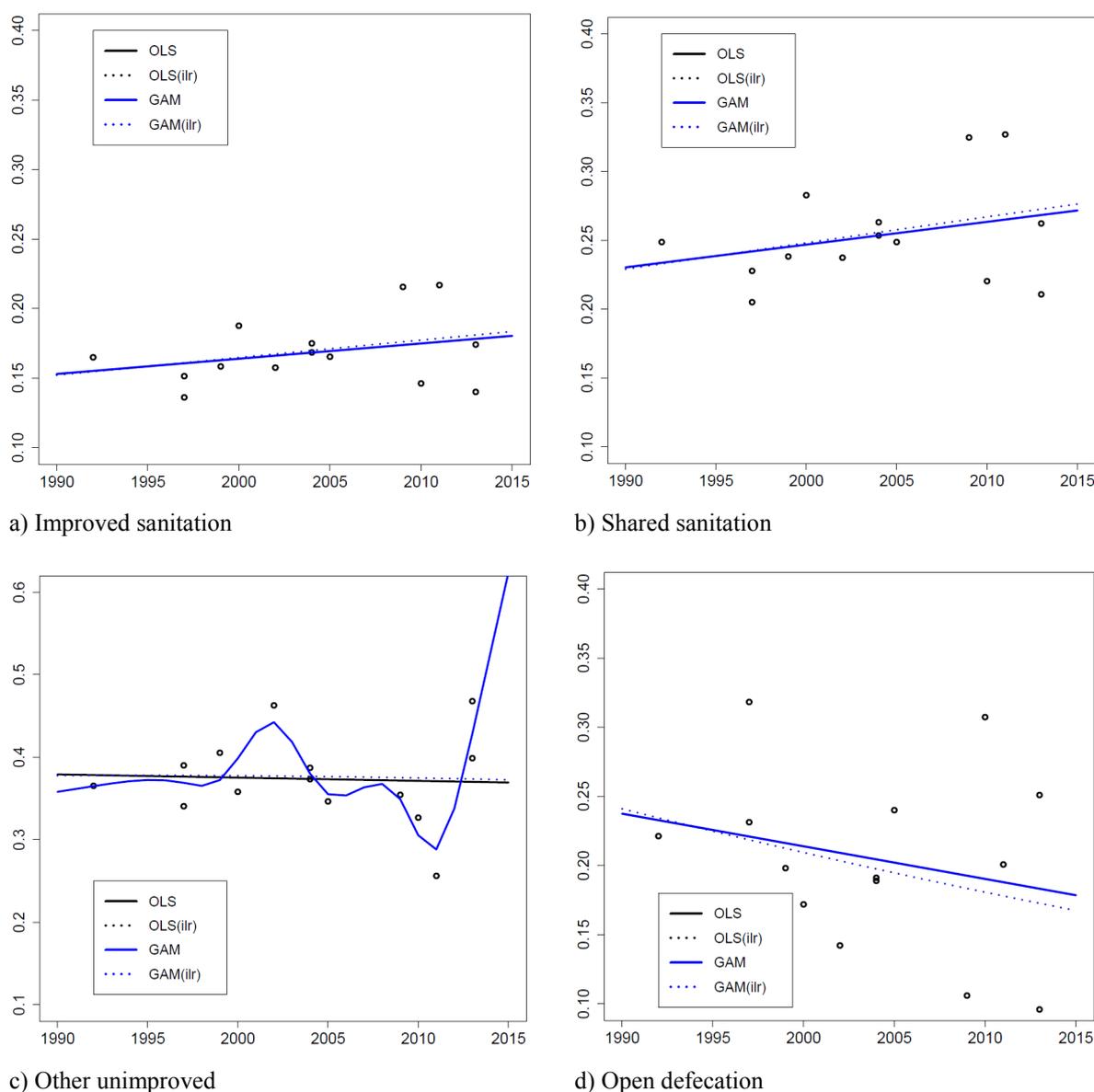
To further show the degree of error contained in the GAM-based projections, Table 2 presents the achieved results for a number of cases. We can observe that the error rate is around a few hundredths, with the two exceptions of the GAM fitted by thin-plate splines, eight degrees of freedom for i) urban sanitation in Madagascar, and ii) rural drinking-water in Ethiopia. In these two cases, polynomials of high degree produce significant errors. This is illustrated by urban sanitation in Madagascar (Figure 4): as none of the four variables give coverage levels approaching extreme values (0% or 100%), and as three of these variables are adjusted by linear regressions, the polynomial that fits the “other unimproved” data (Figure 4c) accounts for the large error detected in the last year of the projection. In this example, and for GAM models restricted to four degrees of freedom, solutions of degree one are obtained. As mentioned above, the interpolation error is zero, and only rounding errors affect the

achieved results. In an opposite case (Ethiopia, rural drinking-water), the interpolation using a polynomial of high degree produces significant errors when modelling near zero values that correspond to low piped coverage at the beginning of the studied period (Figure 2a.1). In a complementary analysis, the focus may shift to the number of population represented in previous percentages, which highlights to a certain extent the risk of overreporting or underreporting if the compositional nature of the data is not considered and a non-linear projection method is not applied. This analysis helps us to better understand the practical implications and the policy relevance of the approach adopted in this study when estimating population figures. As can be seen from the cases presented in Table 2, population gaps range from a few hundred thousand (e.g. drinking-water, urban, Burkina Faso) to millions of people (e.g. drinking-water, rural, Ethiopia).

**Table 2** Sum of population percentages when coverage levels are estimated separately as compositional data, i.e. they are considered to be parts of a whole. The values shown relate to the initial and final years of each period. Population figures are given in % and thousands of inhabitants.

Indicator	Period	Population	GAM (tps, 4)		GAM (tps, 6)		GAM (tps, 8)		GAM (cs)	
			Initial Year	Final Year						
Sanitation, rural, Madagascar	1990 – 2015	%	3.37	-1.49	0.524	-3.08	-1.84	3.36	-0.984	0.542
		·1,000	295.7	-234.6	45.9	-485.3	-161.5	530.1	-86.3	85.4
Sanitation, rural, Ethiopia	1992 – 2016	%	-0.764	0.302	-1.48	0.798	-1.16	0.559	N.A.	N.A.
		·1,000	-319.5	241.8	-619.5	639.5	-486.4	447.7	0.0	0.0
Sanitation, urban, Madagascar	1990 – 2015	%	-1.16E-08	-1.62E-08	-1.78	9.28	-2.11	25.5	-1.01	26.9
		·1,000	0.0	0.0	-49.2	787.0	-58.6	2,160.1	-28.1	2,285.6
Drinking-water, rural, Ethiopia	1992 – 2016	%	6.48	-1.58	5.92	7.48	-12.5	2.33	-7.76	-0.181
		·1,000	2,709.5	-1,269.0	2,475.4	59.9	-5,205.4	1,865.2	-3,242.3	-145.3
Drinking-water, urban Burkina Faso	1991 – 2012	%	-1.49	-1.87	-1.89	-2.25	-1.91	-2.27	-1.89	-2.25
		·1,000	-18.4	-100.7	-23.3	-120.8	-23.5	-121.8	-23.3	-120.8
Drinking-water, rural, South Africa	1994 - 2015	%	-2.46	-2.24	-1.72	-1.79	-1.95	-2.48	-1.80	-2.31
		·1,000	-435.3	-419.6	-303.9	-334.3	-343.8	-465.2	-318.3	-433.1

Notes: GAM (tps, 4): GAM fitted by thin-plate splines, four degrees of freedom; GAM (tps, 6): GAM fitted by thin-plate splines, six degrees of freedom; GAM (tps, 8): GAM fitted by thin-plate splines, eight degrees of freedom; GAM (cs): GAM fitted by cubic splines; N.A.: not applicable, as performing a cubic spline requires data points from at least ten different years.



**Fig. 4.** Sanitation in urban Madagascar. Two different models were fitted to both the original data and the transformed data: i) Generalized Additive Model (eight degrees of freedom), or ii) ordinary least squares (OLS).

We carried out similar analyses by looking at the separate estimates of service levels produced by different regression approaches (Table 3). The differences between outcomes from different projections and official estimates for 2015 (OLS method) are clearly below 1% when estimating improved levels of service (i.e. improved sanitation, piped on premises, and other improved drinking water supplies). There are however differences amounting to 1.9% within variables describing unimproved levels of service (e.g. open defecation and surface water). For clarification purposes, the percentages for the 2015 MDG estimates are shown in Table 4 together with the number of the affected population.

**Table 3** Comparison of baseline and closeline estimates by different estimation methods. Population figures are given in %.

Sanitation, urban, Madagascar	Initial year (1990)				Final year (2015)			
	Improved	Shared	Other unimproved	Open Defecation	Improved	Shared	Other unimproved	Open Defecation
JMP website <sup>a</sup>	15.3	23.0	37.9	23.7	18.0	27.2	36.9	17.9
OLS <sup>b</sup> & GAM (tps, 4)	15.296	23.039	37.920	23.746	18.038	27.169	36.940	17.852
OLS – ilr & GAM (tps, 4) - ilr	15.212	22.912	37.780	24.096	18.349	27.637	37.263	16.751

Drinking-water, rural, Ethiopia	Initial year (1990)				Final year (2015)			
	Piped on premises	Other improved	Unimproved	Surface water	Piped on premises	Other improved	Unimproved	Surface water
JMP website <sup>a</sup>	0.0	3.0	42.8	54.2	1.5	47.2	35.6	15.8
OLS <sup>b</sup>	-0.539	-0.429	42.421	58.548	1.451	47.191	35.348	16.010
GAM (tps. 4)	-0.539	14.625	36.741	58.548	1.451	47.826	33.640	16.010
OLS – ilr	8.763 E-05	6.969	35.357	57.674	4.316	47.298	31.916	16.471
GAM (tps. 4) - ilr	1.654 E-05	9.353	34.451	56.196	1.329	47.538	33.727	17.406

Note: a) Data available at JMP website [accessed Feb 1st, 2017]. Data are available with precision up to one decimal (as indicated here). Notably, Joint Monitoring Programme (2015a) rounds up percentages with no decimals. All results modelled in this work are shown with three decimals in order to compare differences between them and public data. b) In theory, OLS estimates should be equal to JMP estimates, since both are computed by the same method. This applies to all cases with just one exception, of baseline estimates (1990) for drinking-water in rural Ethiopia. The differences in the “other improved” and “surface water” estimates (of 3.43% and 4.35%, respectively) were produced by minor differences in the datasets used when building the models.

**Table 4** Gap between official estimates from the JMP website and estimates produced by a GAM (tps. 4) - ilr. Population figures are given in % and thousands of inhabitants.

Sanitation, urban, Madagascar	Final year (2015)				
	Population	Improved	Shared	Other unimproved	Open Defecation
Gap (*)	%	-0.349	-0.437	-0.363	+1.149
	·1.000	-29.60	-37.07	-30.79	97.55

Drinking-water, rural, Ethiopia	Final year (2015)				
	Population	Piped on premises	Other improved	Unimproved	Surface water
Gap (*)	%	+0.171	-0.338	+1.873	-1.606
	·1.000	137.04	-270.88	1501.08	-1287.10

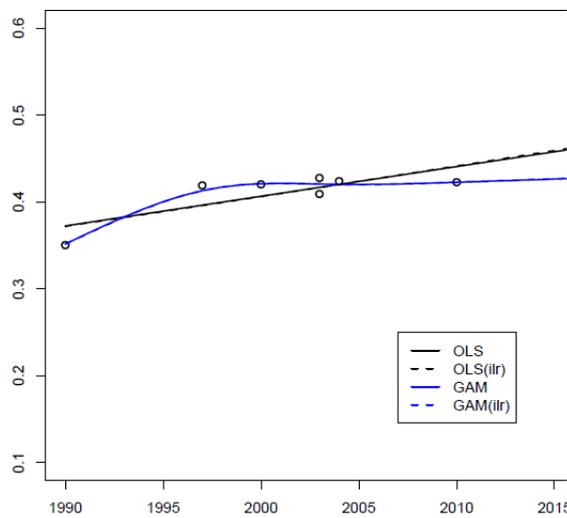
Note: \* Overestimation is presented as a positive value, and underestimation, as a negative value.

Finally, it should be noted that error handling poses a problem from the perspective of data interpretation. As the current JMP method uses differences to estimate the percentage of populations who use unimproved drinking water sources or other unimproved sanitation facilities, it ensures that the four percentages will sum up to 100%, even if they come up from GAM interpolations, for both cases—drinking-water and sanitation. In doing so, however, it eclipses the interpolation errors within the values of these specific variables. Using the ilr transformation to analyse compositional data may be thus seen as a contribution to address this gap, as it is able to accommodate an accurate projection of compositional data.

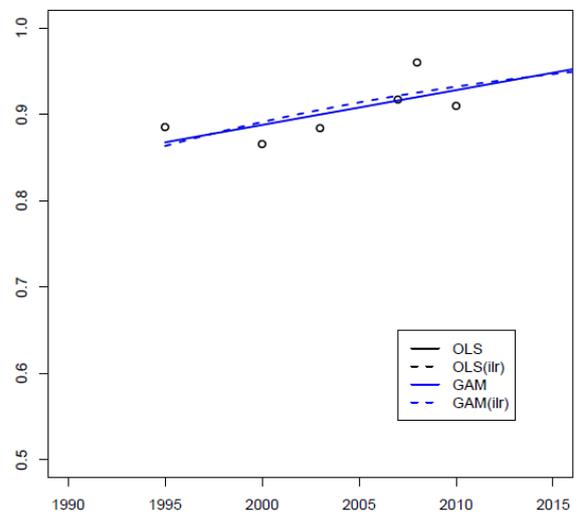
#### 4.2 Countries with 5 to 10 data points

Similar to countries with more than 10 data points, the majority of countries with 6 to 10 data points showed linear trajectories. The most common linear pattern was that of linear growth. However, we also observed non-linear trajectories.

Indeed, several patterns of linearity and non-linearity are present in the data (Figure 5). With few data points, the pattern of curvature cannot be easily defined<sup>1</sup>, but we can confirm that previous observations are equally valid, as: i) GAM-based projections are able to fit curved trajectories, and ii) the ilr-transformation allows an adequate interpretation of extreme values (such as water in urban Swaziland and sanitation in rural Chad).

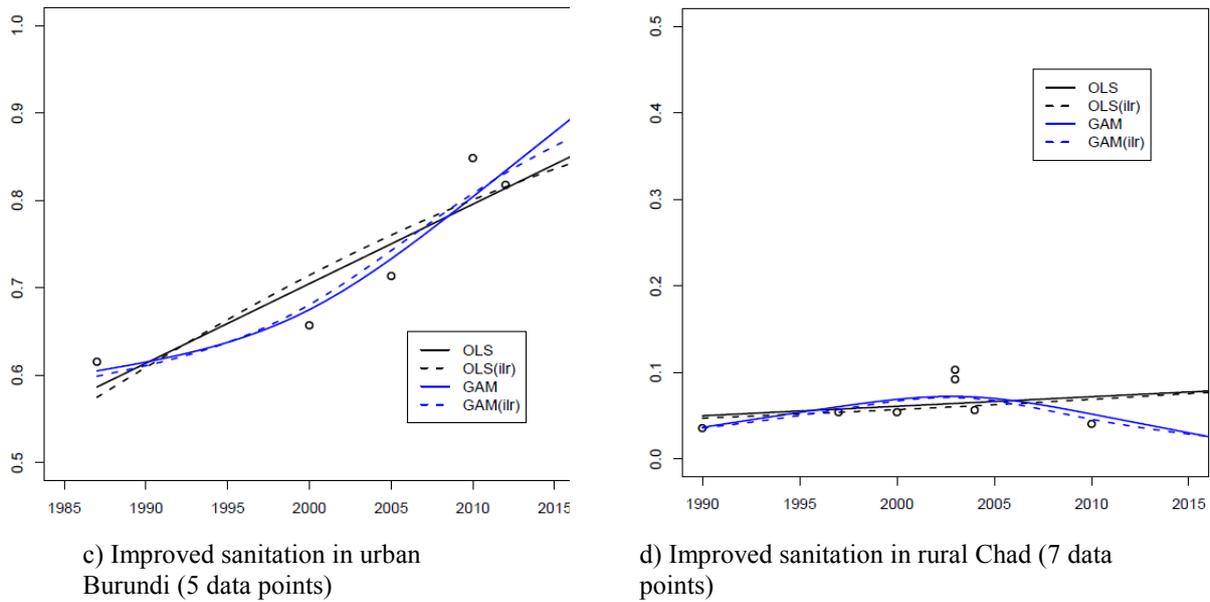


a) Improved water in rural Chad (7 data points)



b) Improved water in urban Swaziland (6 data points)

<sup>1</sup> The categorization rules of non-linear trajectories proposed by Fuller et al (2015) can only be applied for countries with 11 or more data points, as the authors acknowledge, since the  $P$ -value of the quadratic term is determined by both the magnitude of curvature and the sample size.



**Figure 5** Examples of different trajectories in the JMP data. Two different models are fitted to both the original data and the transformed data: i) Generalized Additive Model (GAM), and ii) ordinary least squares (OLS). The projections referenced as OLS (ilr) and GAM (ilr) account for the multidimensional compositional nature of the data.

## 5. Conclusions

The “service ladder” approach adopted in the WaSH sector accounts for the different levels of service provided by various drinking water and sanitation facilities at the household level. We now have addressed the compositional nature of these data in this work. We specifically discuss the adequacy and validity of standard linear regression and generalized linear models (GAM) for modelling time dependence of proportions.

Using GAM results in more accurate estimates than using standard linear regression, particularly when datasets show nonlinear behaviours. This interpolation is particularly suited for data-rich countries. However, increased accuracy is achieved at the expense of lower plausibility, thus revealing a trade-off between these two monitoring outcomes. With respect to the various GAM-based alternatives, the GAM fitted by thin-plate splines with four degrees of freedom produced lower level of errors, particularly at both ends of the period.

Both standard linear regression and GAM may lead to erroneous and meaningless results in certain cases, although *ad hoc* post-process addresses this shortcoming from a practical viewpoint by forcing the set of proportions to sum up to one. For this reason, we have proposed an ilr-GAM approach in this study by computing the standard GAM estimates of the ilr-transformed data, back-transformed to the original scale. The ilr transformation of compositional data is conceptually sound and is fairly simple to implement. It helps improve the performance of both linear and non-linear regression models, and it avoids producing misleading results when all the compositional parts are analysed separately. From the viewpoint of achieved results, the output of the two models (GAM and ilr-GAM) may not differ significantly. Therefore, with respect to the debate about alternatives to model linear

and non-linear patterns in the data, our conclusions are in line with those expressed previously by Fuller et al (2015).

However, using the ilr data transformation helps to force the GAM to make more sensible predictions, which are constrained by the appropriate compositional nature of the data. The ilr data transformation is indeed conceptually sound and avoids producing misleading messages, specifically when extreme data points arise, i.e. when coverage rates are near 0% or 100%. Progress or regression over time in these extreme situations can be therefore computed properly and soundly, which is of relevance when for instance minor variations in coverage relate to the poorest 3% of the population. In addition, when the compositional parts are analysed separately, the ilr transformation guarantees that the different population percentages will sum up to 100%. We have shown that those projections (based on raw data) may generate percentage values whose totals typically range from 97% to 103%, although larger errors have been also observed in specific contexts if the interpolation uses higher degrees of freedom. Finally, a complementary analysis shows that differences between outcomes from different projections and official estimates for 2015 (OLS method) are clearly below 1% when estimating improved levels of service (i.e. improved sanitation, piped on premises, and other improved drinking water supplies). There are however differences amounting to 1.9% within variables describing unimproved levels of service (e.g. open defecation and surface water).

Compositional data are at the heart of quantitative population studies. With a specific focus on the WaSH sector, we now show that the compositional nature of the coverage estimates should be considered for statistical data analysis. We predict that other sectors (such as energy) are likely to adopt similar monitoring frameworks, in which case compositional data approaches will extend beyond WaSH and be applicable to other areas of global sustainable development monitoring.

## **Acknowledgments**

This work has been supported by the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP, [www.wssinfo.org](http://www.wssinfo.org)) in various ways. The discussion of the research topic was launched during the JMP Taskforce on Methods. Moreover, the JMP has provided the dataset and useful comments to improve the manuscript. This research has been partially funded by the Spanish Government (Ministerio de Economía y Competividad, "CODA-RETOS" project, ref. MTM2015-65016-C2-2-R - MINECO/FEDER); and by the Catalan Government (Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), "Engineering Sciences and Global Development" project, ref: 2014SGR1545:2014-2016; and "Compositional and Spatial Analysis" (COSDA) project, ref: 2014SGR551:2014-2016).

## **References**

- Aitchison, J., 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd, London.
- Banerjee, S.G., Bhatia, M., Azuela, G.E., Jaques, I.S., Ashok, P.E., Bushueva, I., Angelou, N., Inon, J.G., 2013. *Global Tracking Framework, Sustainable Energy for All*, The World Bank. Washington D.C. doi:10.1787/dcr-2013-20-en
- Barceló-Vidal, C., Aguilar, L., Martín-Fernández, J.A., 2011. Compositional VARIMA Time Series, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), *Compositional Data Analysis: Theory and Applications*. Wiley, Chichester, United Kingdom, pp. 87–103. doi:10.1002/9781119976462.ch7
- Bartram, J., Brocklehurst, C., Fisher, M.B., Luyendijk, R., Hossain, R., Wardlaw, T., Gordon, B., 2014. *Global*

- Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges. *Int. J. Environ. Res. Public Health* 11, 8137–8165. doi:10.3390/ijerph110808137
- Bergman, J., 2008. Compositional Time Series: An Application, in: *Proceedings of CODAWORK'08: The 3rd Compositional Data Analysis Workshop*. Univeristy of Girona, Girona.
- Brunsdon, T.M., Smith, T.M.F., 1998. The time series analysis of compositional data. *J. Off. Stat.* 14, 237–253.
- Egozcue, J.J., Pawlowsky-Glahn, V., 2006. Simplicial geometry for compositional data. *Compos. Data Anal. Geosci. from Theory to Pract.* 264, 145–160. doi:10.1144/GSL.SP.2006.264.01.11
- 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. doi:10.1023/A:1023818214614
- Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci. Total Environ.* 407, 6100–6108. doi:10.1016/j.scitotenv.2009.08.008
- Fuller, J.A., Goldstick, J., Bartram, J., Eisenberg, J.N.S., 2015. Tracking progress towards global drinking water and sanitation targets: A within and among country analysis. *Sci. Total Environ.* 541, 857–864. doi:10.1016/j.scitotenv.2015.09.130
- Fuller, J.A., Goldstick, J., Eisenberg, J.N.S., 2014. A review of the projection methods used for monitoring progress on drinking water and sanitation: Limitations and alternatives. *Ann Arbor, Michigan*.
- Hastie, T., Tibshirani, R., 1987. Generalized additive models: some applications. *J. Am. Stat. Assoc.* 82, 371–386.
- Hastie, T., Tibshirani, R., 1986. Generalized Additive Models. *Stat. Sci.* 1, 297–310. doi:10.1214/ss/1177013604
- Joint Monitoring Programme, 2015a. *Progress on Sanitation and Drinking Water: 2015 update and MDG assessment*. Geneva / New York.
- Joint Monitoring Programme, 2015b. *Report of the WHO/UNICEF JMP Task Force on Methods*. New York.
- Joint Monitoring Programme, 2015c. *JMP Green Paper: Global monitoring of water, sanitation and hygiene post-2015*. New York and Geneva.
- Joint Monitoring Programme, 2012. *Report of the Second Consultation on Post-2015 Monitoring of Drinking-Water, Sanitation and Hygiene*. WHO / UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP), The Hague.
- Joint Monitoring Programme, 2011. *Report of the First Consultation on Post-2015 Monitoring of Drinking-Water and Sanitation*. WHO / UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP), Berlin.
- Joint Monitoring Programme, 2008. *Progress on Drinking Water and Sanitation: Special Focus on Sanitation*, Joint Monitoring Programme for Water Supply and Sanitation. WHO / UNICEF, Geneva / New York.
- Kayser, G., Moriarty, P., Fonseca, C., Bartram, J., 2013. Domestic Water Service Delivery Indicators and Frameworks for Monitoring, Evaluation, Policy and Planning: A Review. *Int. J. Environ. Res. Public Health* 10, 4812–4835.
- Kynčlová, P., Filzmoser, P., Hron, K., 2015. Modeling Compositional Time Series with Vector Autoregressive Models. *J. Forecast.* 34, 303–314. doi:10.1002/for.2336
- Lloyd, C.D., Pawlowsky-Glahn, V., Egozcue, J.J., 2012. Compositional Data Analysis in Population Studies. *Ann. Assoc. Am. Geogr.* 102, 1251–1266. doi:10.1080/00045608.2011.652855
- Mills, T.C., 2010. Forecasting compositional time series. *Qual. Quant.* 44, 673–690. doi:10.1007/s11135-009-9229-8
- Moriarty, P., Batchelor, C., Fonseca, C., Klutse, A., Naafs, A., Nyarko, A., Pezon, K., Potter, A., Reddy, R., Snehalata, M., 2011. *Ladders for assessing and costing water service delivery, WASHCost*. IRC International Water and Sanitation Centre, The Hague.
- Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, Ltd, Chichester, United Kingdom. doi:10.1002/9781119003144

- Potter, A., Klutse, A., Snehalatha, M., Batchelor, C., Uandela, A., Naafs, A., Fonseca, C., Moriarty, P., 2011. Assessing sanitation service levels, WASHCost. IRC International Water and Sanitation Centre, The Hague.
- Sustainable Development Solutions Network, 2015. Indicators and a Monitoring Framework for the Sustainable Development Goals: Launching a data revolution for the SDGs. Paris and New York.
- United Nations General Assembly, 2015. Transforming our world: the 2030 Agenda for Sustainable Development. Resolution A/RES/70/1.
- United Nations General Assembly, 2014. Report of the Open Working Group of the General Assembly on Sustainable Development Goals. Resolution A/68/970.
- United Nations General Assembly, 2010. The human right to water and sanitation. Resolution A/RES/64/292.
- Wolf, J., Bonjour, S., Pruss-Ustun, A., 2013. An exploration of multilevel modeling for estimating access to drinking-water and sanitation. *J. Water Health* 11, 64–77. doi:10.2166/wh.2012.107
- Wood, S.N., 2006. Generalized additive models: an introduction with R. Chapman & Hall/CRC, United Kingdom. doi:10.1111/j.1541-0420.2007.00905\_3.x
- Wood, S.N., 2004. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *J. Am. Stat. Assoc.* 99, 673–686. doi:10.1198/016214504000000980
- Wood, S.N., 2003. Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* doi:10.1111/1467-9868.00374