WASH your data off: navigating statistical uncertainty in compositional data analysis

F. Ezbakhe[1] and A. Pérez-Foguet[1]
[1]Research Group on Engineering Sciences and Global Development,
Department of Civil and Environmental Engineering, School of Civil Engineering,
Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain
fatine.ezbakhe@upc.edu

Summary

International monitoring of access to drinking water, sanitation and hygiene (WASH) is essential to inform policy planning, implementation and delivery of services. The Joint Monitoring Programme for Water Supply and Sanitation (JMP) is the recognized mechanism for tracking access and progress, and it is based on household surveys and linear regression modelling over time. However, the methods employed have two substantial limitations: they do not address the compositional nature of the data, nor its statistical uncertainty (Ezbakhe & Pérez-Foguet 2018). While the first issue has been tackled previously in the literature (Pérez-Foguet et al. 2017), the effect of non-uniform sampling errors on the regressions remains ignored. This article aims to address these shortcomings in order to produce a more truthful interpretation of JMP data.

The main challenge we try to overcome is how to translate the sampling errors provided in household surveys to the space of compositional data. A Normal distribution is commonly assumed for estimates in household surveys, with a mean and its standard deviation. However, when working with binary data on households – the proportions of households that have access to WASH services – the errors cannot follow normal distributions due to the domain restrictions of proportions, limited to the range 0 to 1. Thus, the Beta distributions seems a better option to characterize the uncertainty around mean access coverage. Yet, as the Beta distribution is defined on the [0,1] interval, the zero values must be dealt with in order to employ the isometric log-ratio (ilr) transformation designed for compositional data. In this article, we investigate the use of two probability distributions (Pearson Type I and Truncated Normal) and Monte Carlo simulations to reinterpret the error in the JMP data so that compositional data analysis is possible.

With a specific focus on the WASH sector, our article shows that the importance of including the survey errors of the data – and its compositional nature – when using this information to support evidence-based policy-making. Indeed, given the current levels of statistical uncertainty in WASH, data may lead to misleading results if errors are not acknowledged (or minimized).

Key words: Demographic Data, Statistical Uncertainty, Compositional Data, Joint Monitoring Programme (JMP), WASH

1 Introduction

In 2015, the global community adopted the 2030 Agenda for Sustainable Development, a universal call to action to end poverty, protect the planet and ensure prosperity to all. The agenda compromises a set of 17 Sustainable Development Goals (SDGs) and 169 targets addressing social, environmental and economic

aspects of development. To monitor progress towards the SDGs, 232 global indicators are defined and tracked by mandated agencies (UNGA, 2017). The list includes two indicators related to SDG 6.1 and 6.2 targets related to the use of drinking water, sanitation and hygiene (WASH): *(i)* indicator 6.1.1, on the proportion of population using safely managed drinking water services; and *(ii)* indicator 6.1.2, on the proportion of population using safely managed sanitation services, including a hand-washing facility with soap and water.

The task of tracking these two global indicators is undertaken by the WHO/UNICEF's Joint Monitoring Programme for Water Supply and Sanitation (JMP). Since 1990, the JMP has produced national, regional and global estimates of population using improved drinking water sources and sanitation facilities. Specifically, the JMP uses service "ladders" to benchmark and compare across countries (JMP, 2017). For drinking water, the ladder reports on the proportion of the population using: *(i)* drinking water directly from surface water; *(ii)* other unimproved water sources; *(iii)* improved water sources that require more than 30 minutes collection time; *(iv)* improved water sources that require less than 30 minutes collection time; and *(v)* improved water sources that are located on premises, available when needed and free from contamination. Similarly, the ladder for sanitation reports on those with: *(i)* no sanitation at all (open defecation); *(ii)* other unimproved facilities; *(iii)* improved facilities shared between two or more households; *(iv)* improved facilities that are not shared; and *(v)* improved facilities that are not shared with other households and where excreta are safely disposed of in situ or transported and treated off-site.

With this service ladder approach, the JMP generates rural, urban and national estimates for each country, for a total of 26 indicators related to WASH (JMP, 2018). The 8 indicators included in this paper are shown in Table 1. Simple linear regression using ordinary least squares method (OLS) is employed to estimate the proportion of the population using each service level. These estimates are used to monitor progress towards SDG targets, as well as to support informed policy and decision making by national governments, development partners and civil society.

Table 1: 8 primary indicators used by the JMP for monitoring drinking water and sanitation services.

Water	The proportion of the population that uses...
W_1	Piped water drinking water sources
W_2	Other improved drinking water sources
W_3	No drinking water facility (surface water)
W_4	Other unimproved drinking water sources
Sanitation	**The proportion of the population that uses...**
S_1	Improved sanitation facilities connected to sewers
S_2	Other improved sanitation facilities
S_3	No sanitation facilities (open defecation)
S_4	Other unimproved sanitation facilities

However, the "JMP estimation" method has two substantial limitations. First, the compositional nature of the data is not taken into account. The JMP models the service ladder proportions separately, which may derive into untenable results where the sum of the proportions is not equal to 1 (i.e., the whole population). This issue has been addressed previously by Pérez-Foguet et al. (2017), who revealed the importance of considering the compositional nature of WASH coverage estimates for statistical data analysis. Second, the large degree of uncertainty inherent within JMP estimates remains unexplored (Ezbakhe & Pérez-Foguet, Agustí, 2018). This uncertainty stems from sampling errors in the household surveys from which the JMP

draws data and, as such, should be accounted for when estimating WASH coverage.

In this context, and to further support the JMP in the task of improving the modelling of WASH data, this paper investigates how to translate sampling errors provided in household surveys to the space of compositional data.

2 Methodology

In household surveys, a Normal distribution is commonly assumed for estimates, with a mean μ and its standard deviation σ. When working with proportions, however, a Normal distribution is not appropriate, since it may yield values that exceed the 0 and 1 bounds. A Beta distribution is more suitable for the statistical modelling of proportions (Ferrari & Cribari-Neto 2004). Yet, as the Beta distribution is defined on the [0,1] interval, zero values must be dealt with in order to employ log-ratio transformations designed for compositional data.

In this paper, we test the use of two probability distributions to reinterpret JMP data: *(i)* Pearson Type I distribution, a generalization of the Beta distribution bounded to [λ,1 $-$ λ]; and *(ii)* Truncated Normal distribution, a generalization of the Normal distribution bounded to [λ_1,λ_2] (in this case [λ,1 $-$ λ]). Their densities are given by Equations 1 and 2, respectively.

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} (x-\lambda)^{\alpha-1} (1-(x-\lambda))^{\beta-1} \tag{1}$$

$$f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) \left(\Phi(\frac{1-\lambda-\mu}{\sigma}) - \Phi(\frac{\lambda-\mu}{\sigma})\right)^{-1} \tag{2}$$

where α and β are the shape parameters of the Pearson Type I distribution, and ϕ and Φ the probability density and cumulative distribution functions of the standard Normal distribution.

The shape parameters are derived from the original data by matching the first and second moments of the "extended Beta" distribution with those of the Normal distribution, as seen in Equations 3 and 4.

$$\mu = \lambda + (1-2\lambda)\frac{\alpha}{\alpha+\beta} \tag{3}$$

$$\sigma^2 = (1-2\lambda)^2 \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4}$$

As suggested by Martín-Fernández et al. (2011), λ can be defined as the "rounding-off error", which relates to the number of significant digits in the database. In this case, we assume $\lambda = 10^{-4}$.

With these two distributions, we use Monte Carlo simulations to generate n sets of JMP data ($n = 1000$). These simulated datasets are used to quantify the uncertainty of JMP data and report the confidence bounds of regressions. For each n simulation, we follow the compositional data (CoDa) methodology: *(i)* we first use a isometric log-ratio (ilr) transformation to bring the compositions to the real space, *(ii)* then apply both ordinary least squares (OLS) linear regression and generalize additive models (GAM) with 4 degrees of freedom to the transformed data, and *(iii)* back-transform the interpolated results to the original scale.

The proposed approach is tested the case of sanitation in rural Madagascar (data in Table 2). The components of the populations are: y_1 sewer, y_2 other improved sanitation facilities, y_3 open defecation, and y_4 other unimproved sanitation facilities. Standard deviations are generated randomly between 0.001 and 0.1, which is the common sampling error in households surveys.

Table 2: JMP data for sanitation in rural Madagascar.

Year	sd	y_1	y_2	y_3	y_4
1992	0.0940	0.0000	0.1300	0.7000	0.1700
1993	0.0720	0.0100	0.2300	0.7300	0.0300
1997	0.0540	0.0000	0.1400	0.7000	0.1600
2000	0.0620	0.0010	0.0865	0.4760	0.4365
2001	0.0600	0.0000	0.0800	0.2600	0.6600
2001	0.0960	0.0000	0.0900	0.2800	0.6300
2002	0.0560	0.0027	0.0887	0.3750	0.5336
2004	0.0730	0.0030	0.0962	0.4620	0.4388
2004	0.0110	0.0012	0.0791	0.5250	0.3947
2005	0.0750	0.0000	0.0909	0.4620	0.4471
2009	0.0040	0.0000	0.0998	0.4910	0.4092
2010	0.0630	0.0360	0.0566	0.5850	0.3224
2011	0.0080	0.0004	0.0843	0.6135	0.3018
2013	0.0540	0.0004	0.1055	0.6052	0.2889
2013	0.0070	0.0010	0.1560	0.5640	0.2790
2016	0.0560	0.0046	0.2317	0.4159	0.3477

3 Results and Discussion

In this section, we compare the coverage estimates obtained by: *(i)* modelling the statistical uncertainty of JMP data with Pearson Type I (aka extended Beta) and Truncated Normal distributions; and *(ii)* applying OLS and GAM regression models.

The importance of translating the sampling errors of JMP data prior to its modelling in the space of compositional data is evidenced in Figure 1. The Normal distribution (Figure 1.a) yields estimates outside the [0,1] interval, specially when proportions of populations are close to the extremes (e.g. in y_1 and y_2). The Pearson Type I distribution (Figure 1.b) may seem suitable to re-interpret the JMP data, as it is delimited at 0 and 1. However, in some cases, it may not be possible to find shape parameters (α and β) that estimate the moments of an Extended Beta distribution. This happens when the mean coverage reported is significantly lower than its standard deviation (e.g. in y_1). Therefore, this approach can only be useful to model uncertainties when standard deviations are smaller than the means. On the contrary, the Truncated Normal distribution (Figure 1.c) is more appropriate to construe the data: it does not only produce estimates between 0 and 1, but also allows for all sets of mean and standard deviation values. Therefore, we choose this latest approach to reproduce the JMP data and model its statistical uncertainty.

On the other hand, when comparing OLS and GAM regressions models (Figure 2), it becomes palpable the need to characterize and represent uncertainty around JMP estimates. In both cases, the 95% confidence interval in the period of JMP data is slightly wide (similar to the errors in the data): (i) with OLS, 0.033, 0.069, 0.097 and 0.105 for y_1,y_2,y_3 and y_4, respectively; and (ii) with GAM, 0.044, 0.089, 0.099 and 0.093. These confidence intervals become much wider in the period beyond the data collected. For instance, in 2020, we can be 95% confident that the expected percentage of the population without access to improved sanitation facilities (aside from open defecation) will be between 40.3% and 61.5% with OLS, or 14.1% and 40.1% with GAM. That it why it is essential to include the survey errors of the data when performing

statistical analysis.

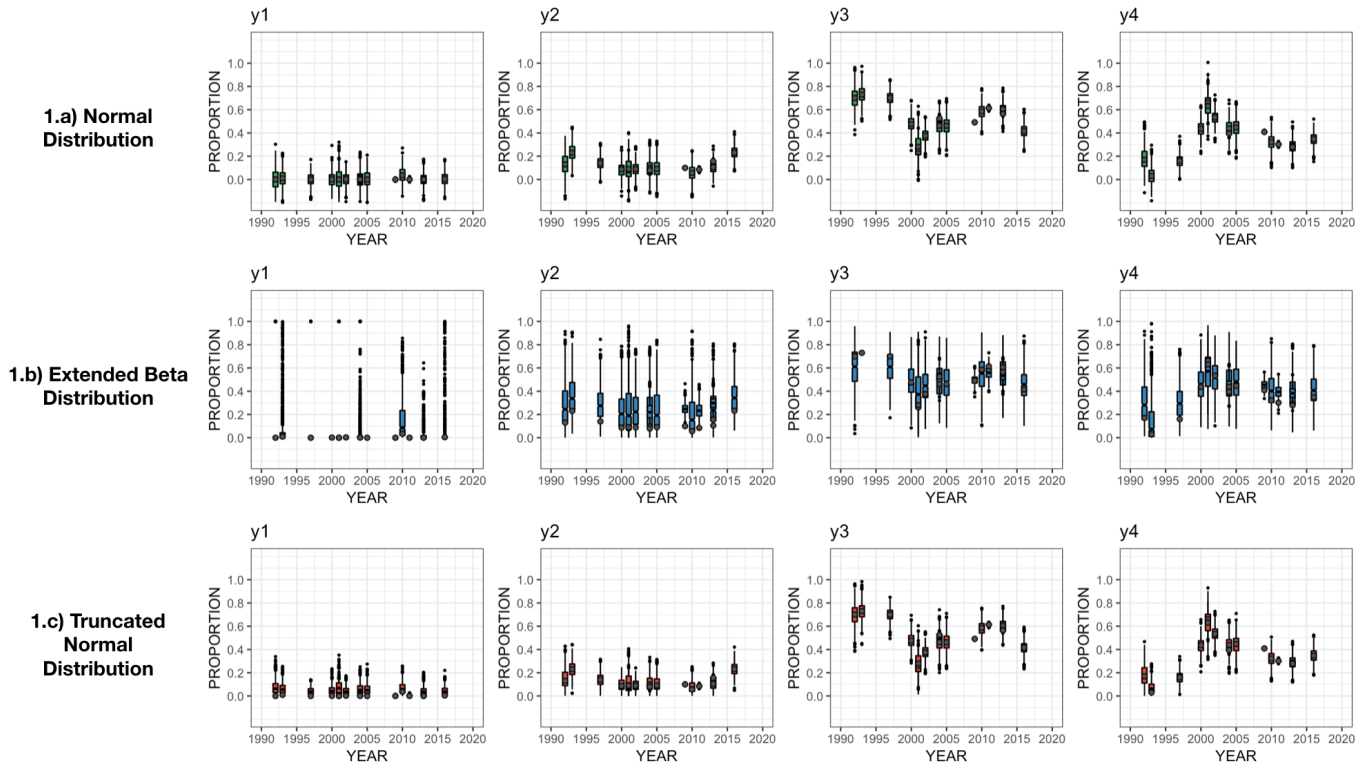

Figure 1: Boxplots of simulated JMP data considering: (1.a) Normal, (1.b) Pearson Type I (aka Extended Beta) and (1.c) Truncated Normal distributions.

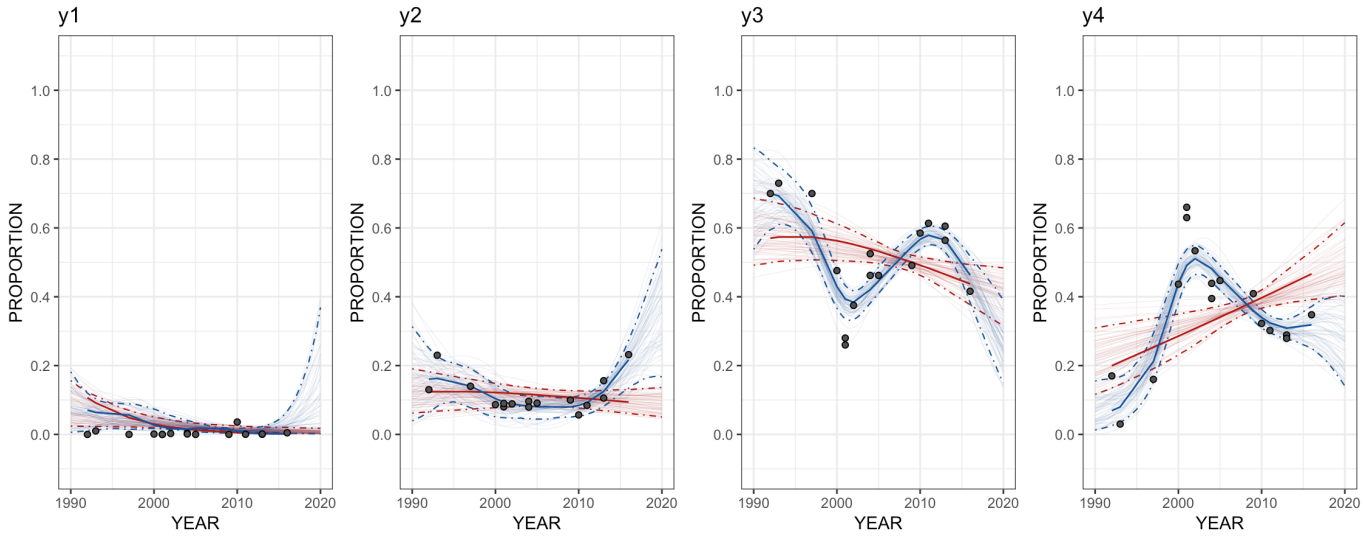

Figure 2: OLS (in red) and GAM (in blue) regressions of JMP data, after ilr-transformation (with 95% confidence intervals using Truncated Normal distributions).

Finally, when comparing which regression model is more appropriate, it can be seen that GAM fits better when datasets show nonlinear behaviours. According to the trajectory categorization methodology proposed by Fuller et al. (2016), these components present the following patterns: y_1 "no change" (i.e. the slope for the entire period is close to zero); y_2 and y_3 (i.e. negative but plateauing slope); and y_4 deceleration (positive slope but plateauing below 1). As shown in Table 3, significant improvement is observed when GAM regression is applied to components y_2, y_3 and y_4. Therefore, using GAM results (after ilr transformation) in JMP can lead to more accurate coverage estimates.

Table 3: Values of root-mean-square error (RMSE) for results of models presented in Figure 2.

Model	y_1	y_2	y_3	y_4
OLS	0.0391	0.0522	0.1433	0.1699
GAM	0.0238	0.0238	0.0648	0.0701

References

Ezbakhe, F., Pérez-Foguet, Agustí (2018). Multi-criteria decision analysis under uncertainty: two approaches to incorporating data uncertainty into Water, Sanitation and Hygiene planning. *Water Resources Management 32*(15), pp. 5169–5182.

Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), pp. 799–815.

Fuller, J.A., Goldstick, J., Bartram, J., Eisenberg, J.N.S (2016). Tracking progress towards global drinking water and sanitation targets: A within and among country analysis. *Science of the Total Environment 541*, pp. 857–864.

JMP (2017). Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines. World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), Geneva.

JMP (2018). JMP Methodology: 2017 Update and SDG baselines. World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), Geneva.

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A. Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 43–58.John Wiley & Sons.

Pérez-Foguet, A., Giné-Garriga, R., Ortego, M.I. (2017). Compositional data for global monitoring: The case of drinking water and sanitation. *Science of the Total Environment 590*,pp. 554–565.

UNGA (2017). Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. *Resolution A/RES/71/313*. United Nations General Assembly, New York.