Science of the Total Environment xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Compositional data for global monitoring: The case of drinking water and sanitation

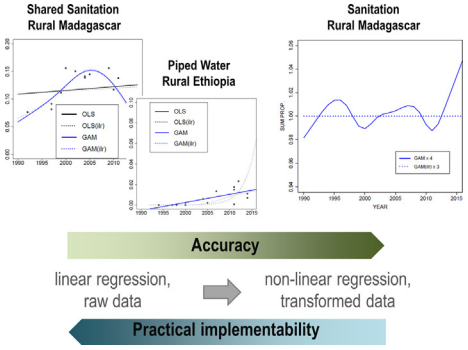
A. Pérez-Foguet [a], R. Giné-Garriga [a,*], M.I. Ortego [b]

[a] Research group on Engineering Sciences and Global Development, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Spain
[b] Compositional and Spatial Analysis COSDA Research Group, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Spain

HIGHLIGHTS

- There is evidence of non-linear time evolution in the JMP data.
- The compositional nature of population data is relevant when modelling WaSH estimates.
- Standard GAM results in more accurate estimates than linear regression models.
- The ilr transformation of CoDa improves results of time regression models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 16 December 2016
Received in revised form 27 February 2017
Accepted 27 February 2017
Available online xxxx

Editor: D. Barcelo

Keywords:
Water
Sanitation and hygiene
Service ladder
Compositional data
Log transformation
Joint Monitoring Programme (JMP) of WHO and UNICEF

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%.

1. Introduction

For the past 15 years, the Millennium Development Goals (MDGs) have challenged the international community to reduce by half the

* Corresponding author at: Universitat Politècnica de Catalunya - Barcelona School of Civil Engineering, Campus Nord - Edif. C2, C. Jordi Girona, 1-3, 08034 Barcelona, Spain.
E-mail addresses: agusti.perez@upc.edu (A. Pérez-Foguet), ricard.gine@upc.edu (R. Giné-Garriga), ma.isabel.ortego@upc.edu (M.I. Ortego).

http://dx.doi.org/10.1016/j.scitotenv.2017.02.220
0048-9697/

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% 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 analysed 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 a 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

Please cite this article as: Pérez-Foguet, A., et al., Compositional data for global monitoring: The case of drinking water and sanitation, Sci Total Environ (2017), http://dx.doi.org/10.1016/j.scitotenv.2017.02.220