A compositional approach for modelling SDG7 indicators: Case study applied to electricity access.

J.C. Marcillo-Delgado[a], M.I. Ortego[b], A. Pérez-Foguet[a,*]

[a]*Research group on Engineering Sciences and Global Development, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Spain*
[b]*Research group COSDA-UPC, Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya-BarcelonaTech.*

Abstract

Monitoring energy indicators has acquired a renewed interest with the 2030 Agenda for Sustainable Development, and specifically with goal 7 (SDG7), which seeks to guarantee universal access to energy. The predominant criteria to monitor SDG7 are given in a set of individual indicators. Along this line, the UN indicators proposed in the 47th session of the UN Statistical commission are a practical starting point. A relevant characteristic of these indicators is that they can be expressed as proportions from a whole, i.e., they are compositions. Notably, directly implementing traditional multivariate models onto indicators that are proportions without an intermediate process can lead to spurious analysis. Here, we aim to assess the application of compositional data analysis(CoDa) to follow up on the temporal trend indicators of the energy sector in the context of SDG7, with a case study for the most affected areas addressing the problem of electricity access. Following CoDa methodology, we first use a log-ratio transformation to bring compositions to real space and then apply three multivariate methods: linear regression, generalized additive models and support vector machine. We also address other characteristic problems of the electricity access indicators, such as data quality, which was treated by considering mod-

*Corresponding author at: Universitat Politècnica de Catalunya, Barcelona School of Civil Engineering (ETSECCPB), Campus Nord C2-310. Jordi Girona, 1-3, 08034 Barcelona, Spain.
Email addresses: juan.marcillo.delgado@gmail.com (J.C. Marcillo-Delgado), ma.isabel.ortego@upc.edu (M.I. Ortego), agusti.perez@upc.edu (A. Pérez-Foguet)

els with interactions. In sum, CoDa facilitates a controlled management of the parts that make up population based indicators, suggesting that modeling evolution of compositions as individual components – even the standard splitting of country data into rural and urban "access to" indicator – should be avoided.
Keywords: Sustainable Development Goals, SDG, compositional data analysis, trend analysis, epsilon support vector machine, generalized additive model

1. Introduction

The 2030 United Nations Sustainable Development Goals (UN SDGs) constitute an opportunity within the global community to reorient global policy and to set the world onto a sustainable trajectory. A global challenge faced by the UN SDGs is energy. Energy availability is an essential factor for economic development and human well-being [1]. However, in recent decades, the impact of providing energy on the global environment is increasing [2]. This problem is reflected in the SDG7, which seeks to "ensure access to affordable, reliable, sustainable and modern energy for all".

As a result of this objective, the use of renewable energy technologies, such as the use of biomass, wind energy, and biogas, rather than fossil fuel is being promoted [3]. Indeed, the global transition from fossil fuels to sustainable energy will accelerate at a much faster rate; an example of this is the growing use of vehicles based on clean energy [4]. The energy deficit in rural sectors is recognized as the greatest challenge, and people are beginning to look for alternative energy sources that adjust to the particularities of this sector, such as photovoltaic energy [5, 6]. These and other factors make it necessary to address access to energy as a multifaceted problem that is in constant movement [7, 8, 9, 10].

Thus, correct management for achieving the SDG7 requires a multidimensional scenario planning that includes numerous indicators [11] outside the traditional scheme that focuses on one indicator, namely, electricity installed gen-

eration capacity [12]; these indicators must be substantive, broadly indicative and effective in capturing the different dimensions of energy access [13, 14, 15].

Forecasting predicts the future based on past observations [16, 17]. Many researchers support forecasting by arguing that it provides guidelines for policymakers, although there are arguments both in favor and against using these for policy analysis [18]. Different models for energy-demand forecasting have been proposed, including time series and regression models [19]. Demand and supply have been analyzed for energy policy in different countries, settings and time horizons, such as in Pakistan, which has been disaggregated by sector until 2030 [20] and trends on consumption until 2020 [21], or in Turkey until 2026 [22], among others. Energy access scenarios for the power sector in sub-Saharan Africa are developed and analyzed in [12], which include an historical evolution of percentage of population with electricity access for selected countries from 1920 to 2010 [8].

Forecasting electricity demands with precision requires correctly determining the influencing variables for each particular country; being population one of the key factors that is generally highly correlated with the electricity demand [23]. However, population alone is not sufficient to explain the changes in the annual electricity demand, and models including socioeconomic covariates are common. Specifically, the relationships between demand, income and population growth have been analyzed for long-term planning exercises [12] On the other hand, the limitations of the traditional distinction between urban and rural, which are necessary to properly describe spatial realities due to urban growth, have already been described [24]; however, this approach is still the one that is usually included in the international agenda.

The SDG global indicator framework presented at the 47th session of the UN Statistical Commission gives a starting point for monitoring goals and targets of the 2030 agenda. The predominant criterion to measuring energy access is to use a set of individual indicators for each SDG7 target [25]. This approach promotes the disaggregated analysis of the whole. As a result, most of these indicators are given as proportions between 0 and 1. For instance, *the proportion*

of population with access to electricity (SDG 7.1.1) can be disaggregated into people from the urban sector with electricity access, people from the rural sector with electricity access and people without electricity access. *Renewable energy share in the final total energy consumption* (SDG 7.2.1) can be disaggregated in biofuels, biogas, hydro energy, wind energy and so on.

In the statistical field, these indicators are known as *compositions*. This compositional character is a very important aspect to consider, especially if stakeholders want to go beyond solely describing the data. Specifically, using a multivariate statistical method (for instance, to explain the trends, make predictions or compare different countries) could give misleading results with spurious correlations if the compositional character is not addressed [26]. Notably, most multivariate methods based on normal distribution are unable to describe compositional data by themselves [27].

The branch of statistics used for the adequate investigation and interpretation on compositions is known as compositional data analysis (CoDa) [28]. CoDa makes it possible to perform classical statistical analysis (e.g. linear regression, principal component analysis) for compositions. The common approach to working with compositions is: i) work on transformed data using an isometric log-ratio approach; ii) use an appropriate model; and iii) back-transform the results [29].

The aim of this paper is to assess CoDa application for following up temporal trend indicators of the energy sector in the context of SDG7. For this, a case study was used in one of the areas most affected by the problem of electricity access, such as the sub-Saharan region and the south of Asia [30, 31]. The World Bank (WB) database was taken as a source of information for the good acceptance of its indicators in the research field [13]. Data for creating this indicator were collected from different sources but with the majority coming from nationally representative household surveys [32]. In detail, this study:

- Applies a case study to a specific indicator: the electricity access. This four-level electricity access indicator shows the ratios of the absence or

presence of electricity service and highlights the dichotomy between urban and rural areas. This fact is very important, as energy access is predominantly a rural problem [33, 34];

- Employs an isometric log-ratio transformation to respond to the particular characteristics of the electricity access indicator;
- As these data have a grouped structure, the convenience of modelling this particularity or using a single-covariate model only time as a predictor was assessed. Interaction models allow the correct modelling of such data, as they allow effects in the slopes or in the fixed parameters to be controlled;
- Three statistical methods have been used: the classical OLS linear regression model, one based on linear predictors that involves the use of smoothing functions for covariates (generalized additive model) and a third one based on optimization algorithms (support vector machine $\epsilon - SVM$);
- The behavior of these models was evaluated both outside and within the calibration range. This process allowed predictions to be made until the year 2030;
- Root mean square error (RMSE) and mean absolute error (MAE) were used to compare the quality of the fitted models.

2. Energy access overview and data preparation

The magnitude of the electricity access problem is widely known. However, to give a brief introduction to it, we will introduce the selected countries for the applied case study. Additionally, we describe the dependent variable or variable to-be-explained as a proportion or composition that can be disaggregated from a whole.

2.1. Energy access overview

There is a huge disparity in energy use. Roughly, the poorer three-quarters of the world's population use only 10% of the world's energy [31]. About 90%

of that energy is for heating and cooking, and the rest (10%), for lighting and entertainment needs [35].

The lack of modern energy sources makes it common to adopt other traditional sources, as kerosene lamps, in order to meet household lighting demands [7]. It is very common to use biomass as an energy input for cooking and lighting, replacing the modern energy sources that the SDG7 seeks to promote. There are around 2500 million people in the world who depend on this energy source, which represents 38% of the world population. About 97% of biomass consumption occurs in Asian developing countries (66%) and sub-Saharan Africa (31%), whereby India (31%) and China (12%) have the population with greater dependence on this input[10].

More than two-thirds of those lacking electricity access are concentrated in few countries [36]. Eleven countries make up 67% of world population without electricity access (a total of 1.062 billion people). Sub-Saharan Africa and the region of south Asia contain 55% and 41% of the unelectrified world population, respectively. India alone contains 25% of the population lacking electricity access.

Table 1: Overview of the countries with large population without energy access in 2014

Country	Electricity access		No electricity access	
	Million	% of total	Million	% of world total
India	1024.34	79.17	269.52	25.38
Nigeria	101.73	57.65	74.73	7.04
Ethiopia**	26.48	27.2	70.88	6.68
Congo, Dem. Rep.**	9.95	13.5	63.77	6.01
Bangladesh	99.47	62.4	59.94	5.64
Tanzania	8.1	15.5	44.14	4.16
Uganda	7.92	20.4	30.91	2.91
Kenya	16.57	36	29.46	2.77
Myanmar**	27	52	24.92	2.35
Mozambique**	5.77	21.22	21.44	2.02
Sudan	16.94	44.9	20.79	1.96
World total	6179.05	85.34	1061.8	100

** These countries were not analyzed because the information availability over time was insufficient, given the aim of the paper.

Source: [32]

2.2. The case study

Many current indicators measure electricity as a lack of physical access [37]; however, this reflects electric *service needs* [15]. Making an analysis only with this single composition part increases the probability of incorrect management decisions because the indicator is systemic biased [38] by excluding the electricity access part that reflect *welfare gains*.

SDG7 indicators are generally available as individual parts or as additively added components that neglect the whole – that is, their compositional character is ignored. Amalgamation is the term that refers to the additive aggregation of parts [39] . Modelling single parts and indicators built through amalgamation can be detrimental as it may lead to spurious analyses [40].

The indicator of electricity access was constructed from the first four indicators displayed in Table 2, and using the formulas displayed in Table 3.

Table 2: Study variables

ID	Variable description
A	Access to electricity, urban (% of urban population)
B	Access to electricity, rural (% of rural population)
C	Urban population (% of total)
D	Rural population (% of total population)
E	Access to electricity (% of population)

Source: [32]

Table 3 summarizes the explained variable electricity access that has been modelled as a study case of the SDG7 indicators and its construction process. This indicator is composed of four parts $x_1, ..., x_4$ that sum one, i.e, they are proportions, while A, B, C, D are percentages. Understanding each of the parts allows the problem of electricity access service need, the welfare gains and the dichotomy between electrification of urban and rural sectors to be addressed. The characteristics of this indicator are framed within a mathematical structure called simplex, which is detailed in section 3.1.

Table 3: Formulas for the construction of the compositional variable response.

Comp.	Description: Proportion of population living ...	Pseudonym	Formula
x_1	... in urban areas with electricity access	urban	$(A \cdot C)/(100 \cdot 100)$
x_2	... in rural areas with electricity access	rural	$(B \cdot D)/(100 \cdot 100)$
x_3	... in urban areas lacking electricity	non-urban	$C/100 - x_1$
x_4	... in rural areas lacking electricity	non-rural	$D/100 - x_2$

3. Methodology

Most indicators included in SDG7 respond to a compositional structure, i.e., they provide relative information. The simplest approach is to apply "log -ratio transformations" to compositional data [26]. Three statistical methods have been applied for the modelling of the transformed series. Variable access to electricity was used as a case study. Trend analysis was performed both within and outside of the calibration range.

3.1. Compositional data analysis - CoDa

The sample space that governs compositions is termed simplex [26]:

$$\mathcal{S}^D = \left\{ x = [x_1.x_2, ..., x_D] \middle| x_i > 0, i = 1, 2, ..., D; \sum_{i=1}^{D} x_i = \mathcal{K} \right\}. \quad (1)$$

This mathematical structure describes the elementary characteristics of compositions. Compositions are usually denoted with the symbol x. The elements of a vector in S^D are called parts or components. The $D - parts$ are strictly positive real numbers and carry only relative information, and the sum of them is equal to a constant $\mathcal{K}$, which is usually 1 or 100% [26].

The constant $\mathcal{K}$ imposes a linear constraint over data. This peculiarity makes interpreting the correlations difficult [26]. If a correlation analysis is applied to these raw compositions, the obtained result might be spurious, as this closed data induces the bias towards negative correlations [39].

The particularities of the simplex have motivated a whole mathematical structure around it, specially making it possible to work with "standard" multivariate techniques designed for unconstrained data, as these techniques are

based on variances and covariances that are defined for the real Euclidean space but not for the simplex [41]. These are completely inappropriate and uninterpretable for such data [42] without prior treatment.

The Aitchison geometry of the simplex is a validated alternative for modelling compositional data that makes the $\mathcal{K}$ constant irrelevant and allows standard statistical models to be used [40]. The methodology is very easy to implement: each composition should be transformed to log-ratio vectors, after which the problem should be reformulated in terms of the corresponding log-ratio vectors. An appropriate standard multivariate procedure can then be applied to the log-ratio vectors. Finally, the log-ratio results must be reformulated in terms of proportions that sum one [42].

The log-ratio transformations are a family of transformations that express compositional data in the unconstrained real space. Most common transformations are additive log-ratio (alr), centered log ratio (clr) or isometric log ratio (ilr) [41]. These transformations exhibit important properties that help preserve the metric characteristics of the simplex [43], including invariance under scale group of transformations, subcompositional consistency and permutation invariance [44].

This paper is based on ilr. This transformation assigns the coordinates x^*, with respect to a given orthonormal basis, to the composition. Orthonormal basis can be readily obtained by the sequential binary partition (SBP) procedure. SBP is a hierarchy of the parts of a composition. In the first order of the hierarchy, all parts are split into two groups. In the following steps, each group is in turn split into two groups. The process continues until all groups have a single part [26].

Each part of the ilr is called balance, and parts are denoted with the term x^*. Vectors belonging to S^D correspond to $D-1$ balances. For this study, assuming the $4-parts$ study variable detailed in Table 3 and using the SBP

method, three real balances are obtained:

$$
\begin{aligned}
x_1^* &= \sqrt{\frac{3 \cdot 1}{3+1}} \ln \frac{(x_1 \cdot x_2 \cdot x_3)^{\frac{1}{3}}}{x_4}, \\
x_2^* &= \sqrt{\frac{1 \cdot 2}{1+2}} \ln \frac{x_2}{(x_1 \cdot x_3)^{\frac{1}{2}}}, \\
x_3^* &= \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_1}{x_3}.
\end{aligned}
\tag{2}
$$

These balances are in the space of real numbers. The first balance provides more information than the rest, as it relates the four components. Following this logic, x_3^* provides less information about the whole. Other balances can be obtained with the SBP method. As detailed in the next subsection, single covariate models and interactions models are proposed for time evolution modelling. In models with interactions, the basis selected does not have major influence in the results. However, in the case of single covariate models calibrated with data which includes two subseries, the selection affects the results in some cases because different balances generate different interactions of the subseries. In this situation, results are more robust when the subseries are parallel than when the subseries have different slopes or have opposite signs. To solve this problem, the 120 balances provided by the SBP method in S^4 compositions were reviewed, and one that produced robust estimates with both kinds of models was chosen.

3.2. Electricity access has two subseries

A benefit of additively structuring aggregated parts of electricity access as a composition was the identification of two subseries within the computed indicator. This connection was made possible by reviewing the existing consistency between access to electricity (E variable in Table 2) and the components x_1 and x_2 related to the electricity access composition, shown in Table 3. This dichotomy is characterized by z:

$$
z = \begin{cases} 0, & \text{if } E = x_1 + x_2 \\ 1, & \text{otherwise} \end{cases}
\tag{3}
$$

This anomaly is part of the frequent problems that responsible entities for collecting energy statistical information from different countries in the world have to deal with. The methodology of the data suggest that, given the low frequency and the regional distribution of some surveys, a number of countries have gaps in their available data. To develop the historical evolution and starting point of electrification rates, a simple modelling approach was adopted to fill in the missing data points around 1990, 2000 and 2010 [32].

Within the characteristics of these two subseries, several issues stand out: first, the representativeness that the subseries occupy in the data set, as $z = 0$ is usually more representative (see Table 4); second, a lack of knowledge about the probability of occurrence of $P(z|z = 0)$ and $P(z|z = 1)$, which presents difficulties when predicting trends using z; and third, the reliability of both subseries, as $z = 0$ is more reliable because two of its components (x_1, x_2) have been validated with another indicator.

Table 4: Representativeness of subseries by country

Country	Time period	Frequency		
		$Z = 0$	$Z = 1$	Total
Bangladesh	1994-2014	11	10	21
India	1993-2014	14	8	22
Kenya	1993-2014	16	6	22
Nigeria	1990-2014	18	7	25
Sudan	1990-2014	22	3	25
Tanzania	1992-2014	12	11	23
Uganda	1991-2014	15	9	24

3.3. Interpolation methods

A large range of statistical models are applied to the energy sector, especially ones that forecast energy demand [19]. Nevertheless, only a few models explain indicators that are ratios or proportions of a whole resulting from SDG7 [45, 46, 47]. For electricity access trends, it is common to approach this relationship over time by ordinary least squares (OLS) [34] because of its linearity.

Considering the covariates time t and the factor variable z, many linear relationships or interaction types can be recognized to model electricity access.

For this paper, the following types are relevant: a) the *most general* in which every level of z has a different slope and intercept; and b) *coincident regression lines*, i.e., the two subseries are the same, in which case the mean function requires only a term for the intercept.

Following this methodology, when compositions are response variables in S^D, and the ilr method is used as an alternative for its correct management, $D-1$ regressions will be estimated for each coordinate. Thus, considering the balance x^*_{ik} and its prediction $\hat{x}^*_{ik}$, the OLS target function is expressed as the sum of the squared residuals of the $D-1$ linear regressions:

$$SSE = \sum_{k=1}^{D-1} \left\{ |\hat{x}^*_{ik} - x^*_{ik}|^2 \right\}. \tag{4}$$

A particularity of applying log-ratio transformations over the time is the smoothness of some series. The solution to the OLS models was to use orthogonal polynomials of t_i using the R *poly* command [48] rather than only t_i.

As an alternative to the OLS classic estimates, other statistical models can help to improve the estimate, especially when dealing with a smooth series and for making predictions outside the calibration range. For this, epsilon support vector machine ($_\epsilon SVM$) and generalized additive model (GAM) were chosen as the two interpolation methods whose use is easily comprehensible.

SVM is a model widely used in the energy sector [49, 50, 51]. In many cases, its estimations outperform conventional trend models, such as the ARIMA model [51]. The basic concept of the SVM is to map input vectors into a higher dimensional feature space through some nonlinear mapping, chosen *a priori* [52].

Unlike the OLS estimates, this model has the advantage that specifying the relationships between the response variable x^*_{ik} and its covariates is not mandatory. As SVM uses kernel functions, nonlinear relationship may suddenly appear to be quite linear [53]. This study used the Gaussian RBF kernel to deal with nonlinearity.

Additionally, SVM uses some parameters to control the regression quality: the cost of error $C > 0$, the insensitive tube ϵ and the parameter γ_{SVM}, which

is proper from the RBF kernel. The value of these parameters are usually arbitrary and depend on the object of analysis [53]. The R package library that was used (e1071) proposes by default $\gamma_{SVM} = 1/(\text{data dimension})$ [54]. However, in our applied case, little data were available; given that the higher this value, the more wiggly our decision boundary becomes, we opted for $\gamma_{SVM} = 1/(3 \cdot \text{data dimension})$.

To determine the proper parameters of C and ϵ, we conducted an optimization process that gives C and ϵ the most appropriate values to minimize the RSME of the regression, using a genetic algorithm (GA), from the GA library of the R package (for more details, see [55]).

We previously obtained good results with generalized additive models (GAM) for modelling linear and non-linear behaviors in water access and sanitation for the SDG6 using CoDa [56], which motivated its use for SDG7. This model is understood as a generalized linear model with a linear predictor involving a sum of smooth functions of covariates [57].

As in SVM, GAM needs a combination of parameters to improve its performance; of these, the most important are: i) the smoothing basis *bs*; ii) the *dimension* k, of the basis used to represent smooth terms; this value amounts to setting the maximum possible degrees of freedom allowed for each model term; iii) γ_{GAM} that multiplies the effective degrees of freedom in the GCV or UBRE/AIC; and iv) the smoothing parameter estimation method [57].

For smoothing basis, a grid search was made to minimize the RMSE among the most important bases: a) thin plate regression splines (tp); b) tp with shrinkage (ts); c) cubic regression spline (cr); d) cr with shrinkage (cs); and e) P-splines (ps). On the other hand, the values of k must be high enough to have enough degrees of freedom, but small enough to be adjusted with the available data. In the present study, we chose to use the default parameter $k = 10$ for single covariate models and $k = 6$ when using models with interactions. The smoothing parameter estimation method selected was the default method of generalized cross validation and Mallow's Cp (GCV.Cp). As GCV tends to overfit on occasion, it has been suggested that $\gamma_{GAM} = 1.4$ can largely correct

this without compromising model fit [58].

3.4. Fitted model quality criteria

The criteria for comparing the different models are the root mean square error (RMSE) and the mean absolute error (MAE), which are standard statistical metrics that indicate the accuracy of the fitted models. The original values and the fitted values over the raw data were used to compute these metrics. Interpretation of these metrics points out that the closer its value is to zero, the better the prediction is.

RMSE is defined as:

$$\text{RMSE} = \left[n^{-1} \sum_{i=1}^{n} |\hat{x}_i - x_i|^2 \right]^{1/2}, \tag{5}$$

MAE is defined as:

$$\text{MAE} = \left[n^{-1} \sum_{i=1}^{n} |\hat{x}_i - x_i| \right]. \tag{6}$$

4. Results and discussion

This section details the main results obtained. All calculations performed were based on the software *R* [48]. The libraries to address the CoDa methodology are *compositions* [59] and *robCompositions* [60]. GAM was applied to the package *mgcv* [61], and SVM to the package *e1071* [54].

4.1. Calibration assessment

4.1.1. CoDa trend model with a single covariate

Figure 1 shows a graphical representation of the trend model considering *coincident lines* for the Bangladesh case, i.e., using only time as a covariate. The three plots on top show the fitted and real values for the balances (model in coordinates) using LM, SVM or GAM. Ordering balances considering the criteria of amount of information used implies: balance 1 $\succ$ balance 2 $\succ$ balance 3 (see equation 2). As a result, balance 3 presents a greater challenge in modelling

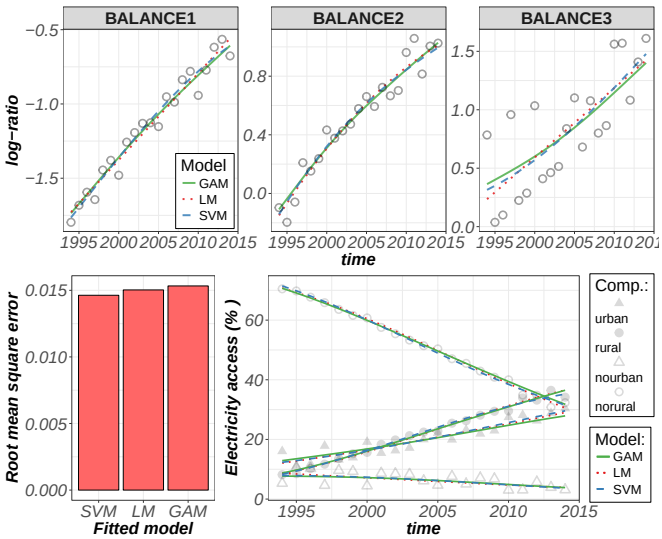

Figure 1: Single covariate model for Bangladesh. GAM use "cs" for all balances. SVMs were fitted with $(C = 5^{3.937}, \epsilon = 0.105)$; $(C = 5^{4.925}, \epsilon = 0.123)$;$(C = 5^{3.970}, \epsilon = 0.522)$ for the balances 1 to 3, respectively.

than the other two balances. The estimate of balance 3 with LM resembles a straight line, while that of SVM and GAM resemble two smooth lines.

The right bottom of Figure 1 represents the original values in S^4 for electricity access and the back-transformations of each balance model; here, the interpolated evolution of each component can be appreciated. The most important thing about this figure, however, is that its linear restriction $\mathcal{K} = 1$ is controlled over the time; thus, this analysis is not spurious, despite using standard models that are not properly designed for compositions.

Based on the RMSE displayed on the left bottom of Figure 1, a comparison between the LM, GAM and SVM prediction capabilities can be made. This analysis is based on the raw composition. Despite the good approximations of the three proposed models, SVM stands out as the model that best approximates the temporary trends of Bangladesh when considering *coincident lines* $F(time)$.

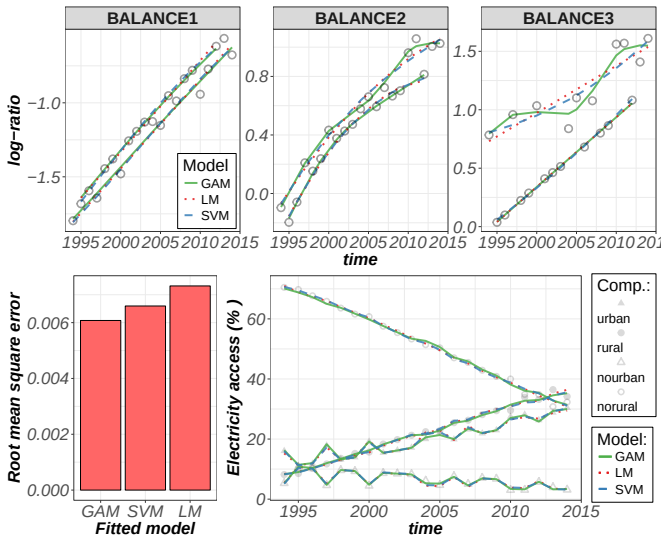

Figure 2: Model with interactions for Bangladesh. GAMs use "cs" for balances 1 and 2, and "cr" for balance 3. SVMs were fitted with $(C = 5^{4.968}, \epsilon = 0.0291)$, $(C = 5^{4.958}, \epsilon = 0.0712)$, or$(C = 5^{4.185}, \epsilon = 0.0448)$ for the balances 1, 2 or 3, respectively.

4.1.2. CoDa trend model with interactions

The estimated model displayed in Figure 2 allows two subseries within each component to be differentiated. In this particular case, it was considered *different lines situation*, $F(time, Z)$, as the three plots on top show that each subseries has different slopes and intercepts. Additionally, the different linearities of each subseries, as well as how each model acts to deal with this problem, are evident. For example, in balance 3, LM estimated straight lines for the two subseries, while GAM and SVM were adapted to the particularities of each curve.

By introducing the covariable factor z_i, it is possible to improve the model and allows almost every point of the compositions to be reached (Figure 2, bottom right). This improvement is evident with the RMSE. Comparing the single covariate model and the model with interactions for the GAM model, RMSE improves from 0.0153 to 0.0061.

A comparison was made between the single covariate model and the interaction models for all countries inside the calibration range (Table 5). Based on

the MAE and RMSE, models with interactions $F(t, z)$ were determined to be better than single covariate models, $F(t)$. According to RMSE, SVM responds better using single covariate models; however, according to MAE, GAM is the best model inside the calibration range in both cases (e.g., with single covariate or with the interaction model).

Table 5: Fitting capability inside the calibration range between single covariate model (time) and interaction model (time and z_i).

Country	$F(t)$			$F(t, z)$		
	GAM	LM	$\epsilon - SVM$	GAM	LM	$\epsilon - SVM$
RMSE						
Bangladesh	0.0153	0.0150	0.0146	0.0061	0.0073	0.0066
India	0.0165	0.0167	0.0146	0.0142	0.0149	0.0143
Kenya	0.0110	0.0107	0.0106	0.0032	0.0041	0.0034
Nigeria	0.0122	0.0129	0.0131	0.0049	0.0109	0.0073
Sudan	0.0093	0.0124	0.0118	0.0008	0.0080	0.0080
Tanzania	0.0055	0.0059	0.0054	0.0035	0.0048	0.0042
Uganda	0.0066	0.0067	0.0063	0.0041	0.0049	0.0049
MAE						
Bangladesh	0.0124	0.0122	0.0118	0.0040	0.0053	0.0045
India	0.0090	0.0101	0.0076	0.0070	0.0081	0.0069
Kenya	0.0073	0.0075	0.0074	0.0017	0.0025	0.0020
Nigeria	0.0097	0.0106	0.0101	0.0027	0.0060	0.0043
Sudan	0.0052	0.0064	0.0059	0.0005	0.0037	0.0053
Tanzania	0.0038	0.0044	0.0039	0.0025	0.0038	0.0026
Uganda	0.0049	0.0053	0.0049	0.0031	0.0039	0.0036

4.2. Prediction assessment

So far, we have presented models within the calibration range, but what happens outside of it? To illustrate this, the last six observations were excluded. The RMSE and MAE were calculated over the excluded observations based on raw data S^4, and a comparison between the single covariate model and a model with interactions was again made using LM, GAM or SVM. Figure 3 reflects the process carried out for this purpose in the case of Bangladesh, where the debate centers on GAM with interactions or SVM with interactions.

Table 6 shows the results achieved outside the calibration range. It can not be concluded that the interaction model is strictly better than the single

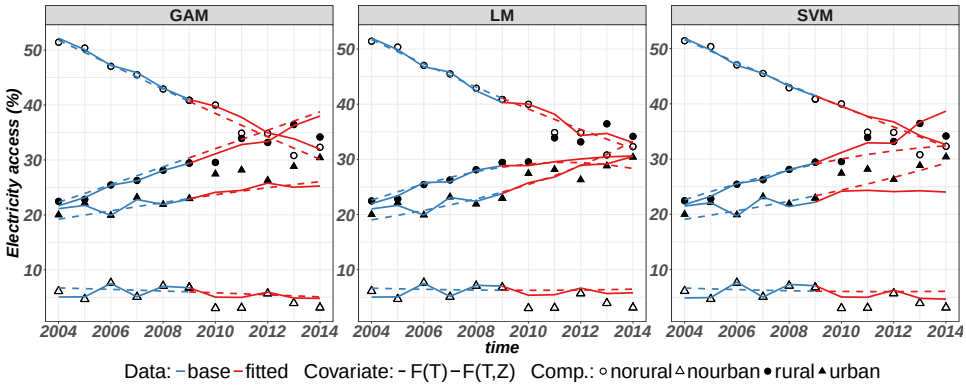

Figure 3: Predictions for the last six observations using CoDa and GAM, SVM or LM for Bangladesh

covariate model. Based on MAE estimations with time as a single covariate, Bangladesh, India, Nigeria, Sudan and Uganda were best predicted with SVM; Kenya was best predicted with GAM; and Tanzania, with OLS. On the side of models with interactions, GAM provided a better estimate for Bangladesh, Nigeria, Sudan and Uganda, and SVM provided a better estimate for India, Kenya and Tanzania. Finally, a model with interactions ($F(t_i, z_i)$) fits better than a single covariate model $F(t_i)$.

4.3. Electricity access trend projections for 2030

This section has been introduced as a discussion. Once comparative validations were made using the last six observations, these configurations were used with all data to make predictions for the 2030 agenda. The discussion revolves around whether to use CoDa with the single covariate t_i or with interactions over z_i. It is worth mentioning that these two alternatives give three electricity access types of predictions: one with $F(t_i)$ and two with $F(t_i, z_i)$ (of which, one is for $z = 0$, and one for $z = 1$).

If we only care about the reliability of data, we would always use the GAM interaction model with the most reliable substring $z = 0$, "same" (see interactions models in figures a), d), e) and g); this statement also is true for India,

Table 6: Fitting capability outside the calibration range between single covariate model (time) and interaction model (time and z_i).

Country	$F(t)$			$F(t,z)$		
	GAM	LM	$\epsilon - SVM$	GAM	LM	$\epsilon - SVM$
RMSE						
Bangladesh	0.0230	0.0291	0.0210	0.0218	0.0242	0.0251
India	0.0264	0.0267	0.0245	0.0356	0.0328	0.0262
Kenya	0.0213	0.0243	0.0210	0.0246	0.0313	0.0189
Nigeria	0.1327	0.0307	0.0278	0.0189	0.0344	0.0260
Sudan	0.0235	0.0234	0.0267	0.0291	0.0292	0.0272
Tanzania	0.0105	0.0106	0.0163	0.0152	0.0214	0.0125
Uganda	0.0141	0.0122	0.0145	0.0206	0.0224	0.0217
MAE						
Bangladesh	0.0186	0.0229	0.0169	0.0154	0.0186	0.0190
India	0.0195	0.0198	0.0156	0.0193	0.0225	0.0180
Kenya	0.0138	0.0181	0.0165	0.0161	0.0202	0.0135
Nigeria	0.1023	0.0258	0.0229	0.0143	0.0240	0.0202
Sudan	0.0145	0.0155	0.0187	0.0152	0.0162	0.0155
Tanzania	0.0089	0.0087	0.0128	0.0116	0.0167	0.0097
Uganda	0.0104	0.0102	0.0102	0.0138	0.0170	0.0145

Kenya and Tanzania). Another factor that is determinant is the frequency or representativeness of the subseries over the total data series (see Table 4). For example, in the case of Sudan, there are 3 values in the period 1990 - 2014 for the subseries of class $z = 1$; this suggests that other options should be taken into account, such as considering these to be missing points, descrambling them to make a single covariate model or using the prediction 2030 for $z = 0$ of the model with interactions.

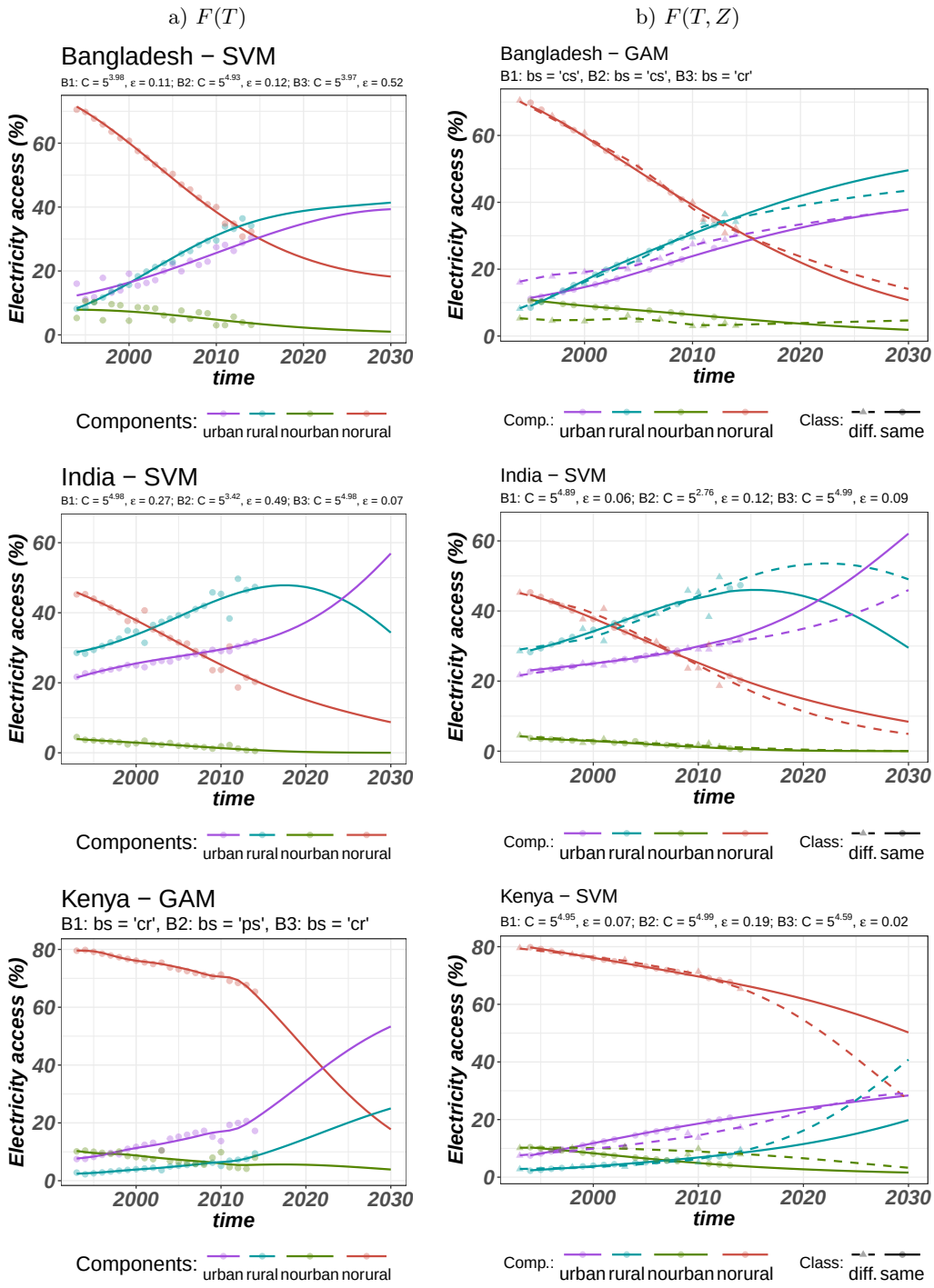

Figure 4: Comparison between a single covariate model (a) and a model with interactions (b). Bangladesh, India, Kenya.

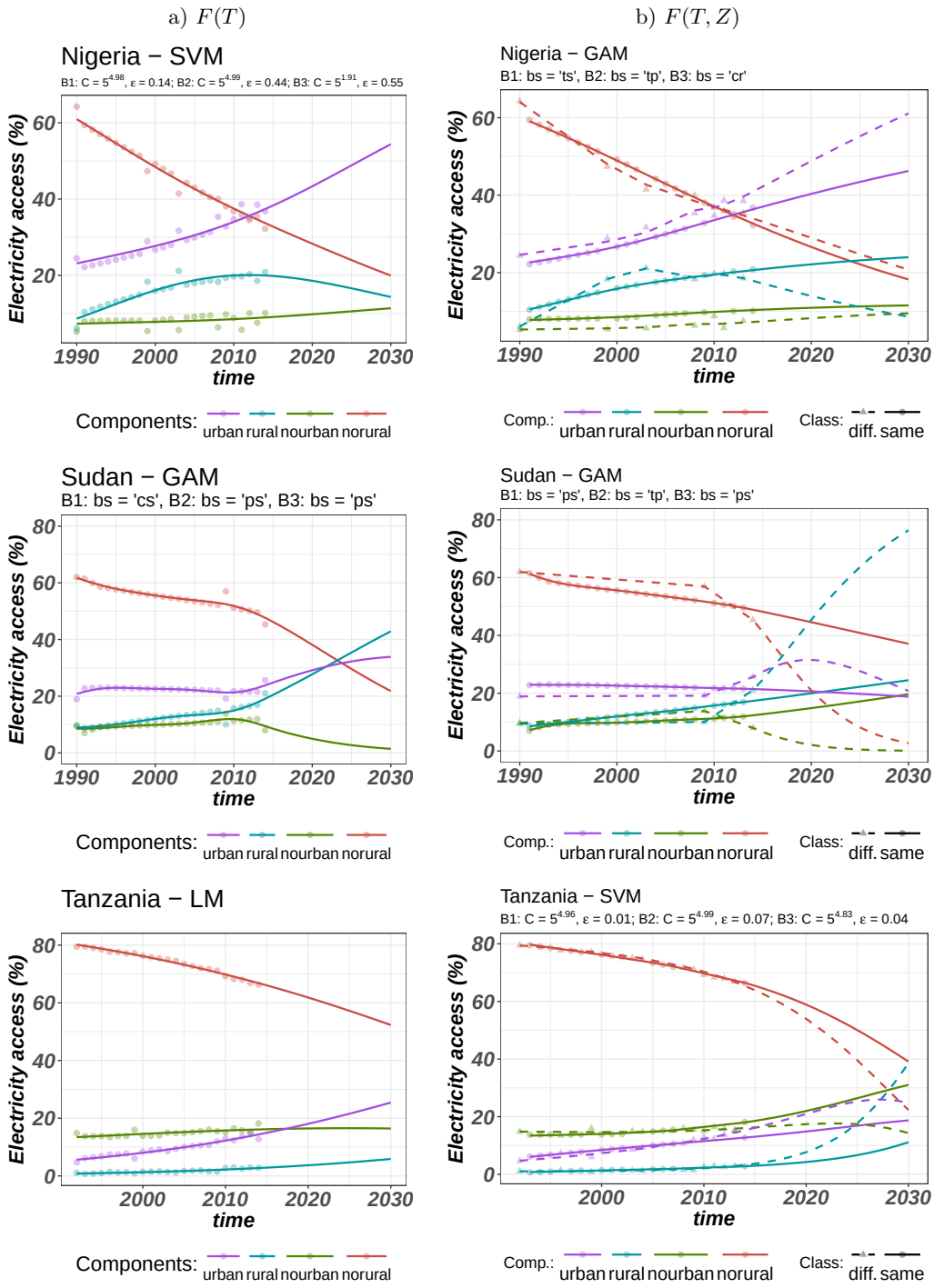

Figure 5: Comparison between a single covariate model (a) and a model with interactions (b). Nigeria, Sudan, Tanzania.

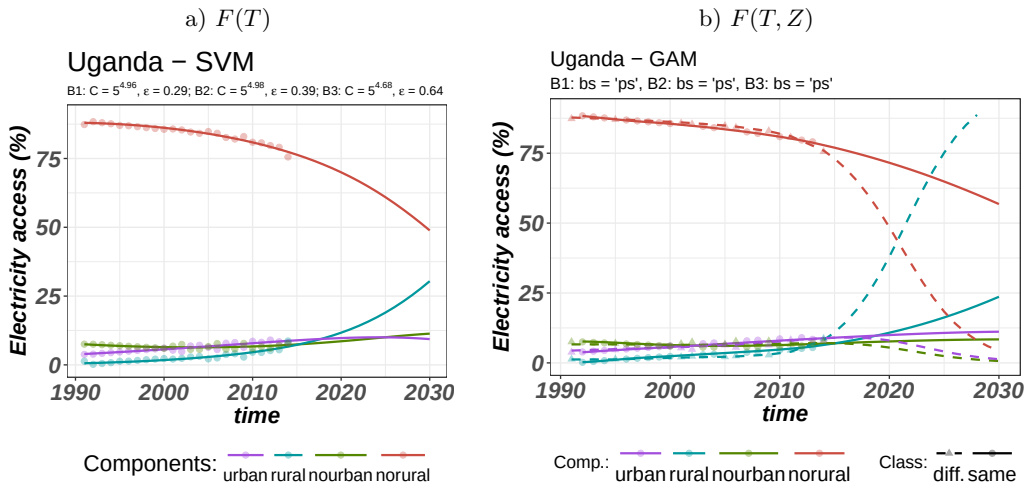

Figure 6: Comparison between a single covariate model (a) and a model with interactions (b). Uganda.

Another determinant factor is the overlap of the subseries. For India, the two subseries are overlapped, making it irrelevant whether a single covariate model or an interaction model is used. However, for Nigeria, the subseries are clearly separated, and the interaction model gives a better appreciation of the trends. Depending on how separated the subseries are, using a single covariate model can generate problems of atypicality in the series; consequently, the predictions are not very good in most single covariate models.

The 2030 predictions for the analyzed countries are shown in figures from 4 to 6. All of these predictions are based on the best MAE of Table 6. Two figures are shown for each country: the single covariate model (on the left) and the interaction model (on the right).

Finally, it is important to emphasize that standard models should not be used to analyze compositions without prior treatment. Figure 7 highlights one of the main consequences of treating compositional data as univariate. When applying standard models, the sum-to-one restriction is not satisfied, which forces one to perform a manual adjustment when the values are below 0% or above 100%. For instance, the total sum of the composition can vary between 99% and 108% for

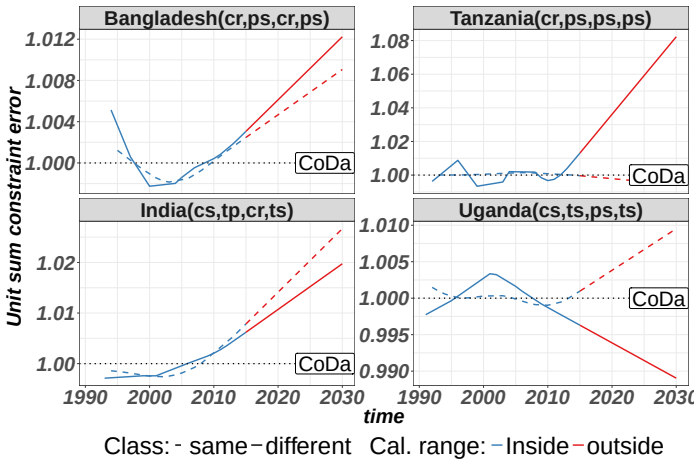

Figure 7: Unit sum constraint error for components (x_1, x_2, x_3, x_4) using standard GAM models with interactions for Bangladesh, India, Tanzania and Sudan.

GAMs; this highlights that the error is greater outside the calibration range than within it, regardless of the interactions. In addition, an error is generated from the point of view of the data interpretation, especially close to the lower and upper limits. The best way of performing analysis of compositional data by using models related with the ordinary euclidean distance as the models proposed in this paper is through *the principle of working in coordinates* [26, 59], i.e., using coordinates with respect to a basis.

5. Conclusions

In order to model tendencies of SDG7 compositional indicators, it is advisable to use CoDa with LM, GAM or SVM as opposed to standard models (including GAM or SVM). The main argument is that CoDa facilitates a more controlled management of the parts that make up the indicator, especially when it comes to making inferences outside the calibration range.

The compositional approach over the electricity access indicator developed in section 2.2 confirms that the data series include in some cases two subseries, characterized by different temporal evolution coefficients. Among the

seven countries analyzed, the subseries were: very different (Nigeria, Sudan and Uganda); partially overlapping, especially in the components related to the rural areas (Bangladesh, Kenya, Tanzania); and overlapping in all components (India). Based on the different levels of overlap, it is best to avoid using compositions with amalgamated indicators but rather to use the simplex structure for analysis.

This anomaly very possibly responds to the fact that, in the process of data collection, there were gaps in available data, and that a simple modelling approach was adopted to fill in the missing data points by considering an individual part methodology. Therefore, the use of CoDa is recommended for improving the management of this type of indicators.

Comparing the single covariate model and the model with interactions for the 2030 projection, we observed cases in which: the projections did not differ from each other (India); differed only slightly from each other (Bangladesh, Uganda and Tanzania); or differed greatly from each other (Kenya, Nigeria and Sudan).

At least three decisive factors were found for the decision to opt for a 2030 projection based on either the single covariate model or the interaction model: a) the representativeness of the subseries within the total set (Sudan case), b) the overlapping of the subseries (India versus Nigeria case), and c) the influence of the outliers that are generated by using a single covariate model (Kenya, Sudan and Bangladesh cases).

Based on RMSE and MAE for the analyses made within the calibration range (section 4.1), we can conclude that the best estimates are obtained using models with interactions. The behavior of the three representative models suggests that GAM can be used for this purpose, with SVM in second place and OLS in third. Based on MAE and RMSE of the models outside the calibration range, SVM and LM are recommended for single covariate model, and GAM and SVM, for interaction models when making predictions outside the calibration range.

It is worth mentioning that if the aim is to obtain predictions for the most "reliable" and "representative" substring of electricity access indicators, the

model that best estimates both within and outside the calibration range is GAM. Thus, based on this estimate, and using all the factors related to access to electricity (ceteris paribus) up to the cut-off date (2014), none of the seven countries studied will reach 100% access to electricity until the year 2030, as the rural sector is that with the greatest need for electricity access service.

The most optimistic cases indicate that: i) India could increase access to electricity (welfare gain) from 79% to 93%; ii) Bangladesh could improve from 64.5% to 87.42% and iii) Nigeria could reach to 70.12%. On the other hand, Tanzania is expected to be the country with the greatest deficiency in access to electricity, 77%, and Uganda the second, with a 65% of service need.

Considering the dichotomy between urban and rural sectors, an improvement in electrification in both sectors is foreseen for all the analyzed countries, except Sudan with a decrease in urban access to electricity of 7%. The rate of improvement is more significant in the rural sector in the cases of Bangladesh and Uganda, with an increase of 15%, and in the urban sector in the cases of India (10%), Kenya (16%), Nigeria (9%) and Tanzania (5%). It is interesting to note that despite the decreasing overall trend in Sudan, an increase of 3% in access to electricity in the rural sector is expected.

The present work used seven representative countries as a unit of analysis for the electricity access problem, but this analysis is easily applicable to the 212 countries that are in the World - Bank database; it is as simple as choosing an orthonormal base, transforming the compositional data and running the model. This extension, which is out of the scope of this contribution, would increase the understating of global access to electricity evolution, including coupling effects between evolving rural and urban populations, and would address one of key inequities in terms of sustainable development.

The approach presented here overcomes the problems of consistency within data of historical series, which undermines comparability between countries and therefore the implementation and monitoring of coordinated programs as needed in the international energy agenda. Notably, this technique could be extended to further analyze other forms of energy provision and be used at other geographical

scales.

Acknowledgments

This research has been partially funded by the Ministerio de Economía y Competitividad del Gobierno de España (MINECO/FEDER, Ref: MTM2015-65016-C2-2-R); and by the Agència de Gestió d'Ajuts Universitaris i de Recerca de la Generalitat de Catalunya (Ref. 2017 SGR 656 and 2017 SGR 1496).

Bibliography

[1] M. Nilsson, P. Lucas, T. Yoshida, Towards an integrated framework for SDGs: Ultimate and enabling goals for the case of energy, Sustainability 5 (10) (2013) 4124–4151. doi:https://doi.org/10.3390/su5104124.

[2] D. van Vuuren, N. Nakicenovic, K. Riahi, A. Brew-Hammond, D. Kammen, V. Modi, M. Nilsson, K. Smith, An energy vision: the transformation towards sustainability—interconnected challenges and solutions, Current Opinion in Environmental Sustainability 4 (1) (2012) 18 – 34. doi:https://doi.org/10.1016/j.cosust.2012.01.004.

[3] L. P. Ghimire, Y. Kim, An analysis on barriers to renewable energy development in the context of Nepal using AHP, Renewable Energy 129 (2018) 446 – 456. doi:https://doi.org/10.1016/j.renene.2018.06.011.

[4] N. Edomah, Governing sustainable industrial energy use: Energy transitions in Nigeria's manufacturing sector, Journal of Cleaner Production 210 (2019) 620 – 629. doi:https://doi.org/10.1016/j.jclepro.2018.11.052.

[5] N. Moksnes, A. Korkovelos, D. Mentis, M. Howells, Electrification pathways for kenya–linking spatial electrification analysis and medium to long term energy planning, Environmental Research Letters 12 (9) (2017) 095008. doi:https://doi.org/10.1088/1748-9326/aa7e18.

[6] A. Giwa, A. Alabi, A. Yusuf, T. Olukan, A comprehensive review on biomass and solar energy for sustainable energy generation in Nigeria, Renewable and Sustainable Energy Reviews 69 (2017) 620 – 641. doi:https://doi.org/10.1016/j.rser.2016.11.160.

[7] M. Kanagawa, T. Nakata, Assessment of access to electricity and the socio-economic impacts in rural areas of developing countries, Energy Policy 36 (6) (2008) 2016 – 2029. doi:https://doi.org/10.1016/j.enpol.2008.01.041.

[8] S. Pachauri, A. Brew-Hammond, D. Barnes, D. Bouille, S. Gitonga, V. Modi, G. Prasad, A. Rath, H. Zerriffi, The Global Energy Assessment: Toward a More Sustainable Future, Cambridge University Press, 2012. URL http://www.iiasa.ac.at/web/home/research/Flagship-Projects/Global-Energy-Assessment/Chapte19.en.html

[9] F. Birol, et al., World energy outlook, International Energy Agency, 2013. URL https://www.iea.org/publications/freepublications/publication/WEO2013.pdf

[10] L. Cozzi, et al., World energy outlook special report (2017). URL https://www.iea.org/weo/weospecialreports/

[11] P. Nussbaumer, M. Bazilian, V. Modi, Measuring energy poverty: Focusing on what matters, Renewable and Sustainable Energy Reviews 16 (1) (2012) 231 – 243. doi:https://doi.org/10.1016/j.rser.2011.07.150.

[12] M. Bazilian, P. Nussbaumer, H.-H. Rogner, A. Brew-Hammond, V. Foster, S. Pachauri, E. Williams, M. Howells, P. Niyongabo, L. Musaba, B. Ó. Gallachóir, M. Radka, D. M. Kammen, Energy access scenarios to 2030 for the power sector in sub-Saharan Africa, Utilities Policy 20 (1) (2012) 1 – 16. doi:https://doi.org/10.1016/j.jup.2011.11.002.

[13] Y. G. Hailu, Measuring and monitoring energy access: Decision-support tools for policymakers in Africa, Energy Policy 47 (2012) 56 – 63, universal

access to energy: Getting the framework right. doi:https://doi.org/10.1016/j.enpol.2012.03.065.

[14] G. S. Mensah, F. Kemausuor, A. Brew-Hammond, Energy access indicators and trends in Ghana, Renewable and Sustainable Energy Reviews 30 (2014) 317 – 323. doi:https://doi.org/10.1016/j.rser.2013.10.032.

[15] S. Groh, S. Pachauri, N. D. Rao, What are we measuring? an empirical analysis of household electricity access metrics in rural Bangladesh, Energy for Sustainable Development 30 (2016) 21 – 31. doi:https://doi.org/10.1016/j.esd.2015.10.007.

[16] C. W. Granger, P. Newbold, Forecasting Economic Time Series, 2nd Edition, Elsevier Monographs, Elsevier, 1986. doi:https://doi.org/10.1016/B978-0-12-295183-1.50012-1.

[17] R. J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, 2nd Edition, OTexts, 2018.
URL https://otexts.org/fpp2/

[18] C. A. Sims, Are forecasting models usable for policy analysis?, Fed. Reserve Bank. Minneap. Q. Rev. 10 (1986) 2 – 16.
URL https://ideas.repec.org/a/fip/fedmqr/y1986iwinp2-16nv.10no.1.html

[19] L. Suganthi, A. A. Samuel, Energy models for demand forecasting—a review, Renewable and Sustainable Energy Reviews 16 (2) (2012) 1223 – 1240. doi:https://doi.org/10.1016/j.rser.2011.08.014.

[20] U. Perwez, A. Sohail, S. F. Hassan, U. Zia, The long-term forecast of pakistan's electricity supply and demand: An application of long range energy alternatives planning, Energy 93 (2015) 2423 – 2435. doi:https://doi.org/10.1016/j.energy.2015.10.103.

[21] A. Hussain, M. Rahman, J. A. Memon, Forecasting electricity consumption in pakistan: the way forward, Energy Policy 90 (2016) 73 – 80. doi:https://doi.org/10.1016/j.enpol.2015.11.028.

[22] K. Kavaklioglu, Modeling and prediction of Turkey's electricity consumption using Support Vector Regression, Applied Energy 88 (1) (2011) 368 – 375. doi:https://doi.org/10.1016/j.apenergy.2010.07.021.

[23] M. E. Günay, Forecasting annual gross electricity demand by artificial neural networks using predicted values of socio-economic indicators and climatic conditions: Case of Turkey, Energy Policy 90 (2016) 92 – 101. doi:https://doi.org/10.1016/j.enpol.2015.12.019.

[24] B. Cohen, Urban growth in developing countries: A review of current trends and a caution regarding existing forecasts, World Development 32 (1) (2004) 23 – 51. doi:https://doi.org/10.1016/j.worlddev.2003.04.008.

[25] UN, 71/313. Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development, Tech. rep., A/RES/71/313. New York, USA: United Nations (2017).
URL https://digitallibrary.un.org/record/1291226

[26] V. Pawlowsky-Glahn, J. J. Egozcue, R. Tolosana-Delgado, Modeling and analysis of compositional data, John Wiley & Sons, 2015.
URL https://www.wiley.com/en-es/Modeling+and+Analysis+of+Compositional+Data-p-9781118443064

[27] K. G. Van den Boogaart, R. Tolosana-Delgado, Analyzing compositional data with R, Vol. 122, Springer, 2013. doi:10.1007/978-3-642-36809-7.

[28] J. Aitchison, The statistical analysis of compositional data, Journal of the Royal Statistical Society. Series B (Methodological) (1982) 139–177.
URL https://www.jstor.org/stable/2345821

[29] V. Pawlowsky-Glahn, J. J. Egozcue, Geometric approach to statistical analysis on the simplex, Stochastic Environmental Research and Risk Assessment 15 (5) (2001) 384–398. doi:https://doi.org/10.1007/s004770100077.

[30] C. N. Doll, S. Pachauri, Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery, Energy Policy 38 (10) (2010) 5661 – 5670, the socio-economic transition towards a hydrogen economy - findings from European research, with regular papers. doi:https://doi.org/10.1016/j.enpol.2010.05.014.

[31] M. Bazilian, A. Sagar, R. Detchon, K. Yumkella, More heat and light, Energy Policy 38 (10) (2010) 5409 – 5412, the socio-economic transition towards a hydrogen economy - findings from European research, with regular papers. doi:https://doi.org/10.1016/j.enpol.2010.06.007.

[32] World-Bank, World Bank Open Data Free and open access to global development data (2017).
URL https://data.worldbank.org/

[33] S. C. Bhattacharyya, S. Ohiare, The chinese electricity access model for rural electrification: Approach, experience and lessons for others, Energy Policy 49 (2012) 676 – 687, special Section: Fuel Poverty Comes of Age: Commemorating 21 Years of Research and Policy. doi:https://doi.org/10.1016/j.enpol.2012.07.003.

[34] I. Onyeji, M. Bazilian, P. Nussbaumer, Contextualizing electricity access in sub-Saharan Africa, Energy for Sustainable Development 16 (4) (2012) 520 – 527. doi:https://doi.org/10.1016/j.esd.2012.08.007.

[35] S. C. Bhattacharyya, Energy access programmes and sustainable development: A critical review and analysis, Energy for Sustainable Development 16 (3) (2012) 260 – 271. doi:https://doi.org/10.1016/j.esd.2012.05.002.

[36] S. C. Bhattacharyya, Energy access problem of the poor in India: Is rural electrification a remedy?, Energy Policy 34 (18) (2006) 3387 – 3397. doi:https://doi.org/10.1016/j.enpol.2005.08.026.

[37] S. Pachauri, D. Spreng, Measuring and monitoring energy poverty, Energy Policy 39 (12) (2011) 7497 – 7504, clean Cooking Fuels and Technologies in Developing Economies. doi:https://doi.org/10.1016/j.enpol.2011.07.008.

[38] R. L. Bingham, L. A. Brennan, B. M. Ballard, Discrepancies between Euclidean distance and compositional analyses of resource selection data with known parameters, Journal of Wildlife Management 74 (3) (2010) 582–587. doi:https://doi.org/10.2193/2009-119.

[39] C. D. Lloyd, V. Pawlowsky-Glahn, J. J. Egozcue, Compositional data analysis in population studies, Annals of the Association of American Geographers 102 (6) (2012) 1251–1266. doi:https://doi.org/10.1080/00045608.2011.652855.

[40] J. J. Egozcue, V. Pawlowsky-Glahn, Groups of parts and their balances in compositional data analysis, Mathematical Geology 37 (7) (2005) 795–828. doi:https://doi.org/10.1007/s11004-005-7381-9.

[41] P. Filzmoser, K. Hron, Correlation analysis for compositional data, Mathematical Geosciences 41 (8) (2009) 905. doi:https://doi.org/10.1007/s11004-008-9196-y.

[42] J. Aitchison, Principles of compositional data analysis, Lecture Notes-Monograph Series (1994) 73–81.
URL https://www.jstor.org/stable/4355794

[43] J. Aitchison, Logratios and natural laws in compositional data analysis, Mathematical Geology 31 (5) (1999) 563–580. doi:https://doi.org/10.1023/A:1007568008032.

[44] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barcelo-Vidal, Isometric logratio transformations for compositional data analysis, Mathematical Geology 35 (3) (2003) 279–300. doi:https://doi.org/10.1023/A:1023818214614.

[45] E. Panos, M. Densing, K. Volkart, Access to electricity in the world energy council's global energy scenarios: An outlook for developing regions until 2030, Energy Strategy Reviews 9 (2016) 28 – 49. doi:https://doi.org/10.1016/j.esr.2015.11.003.

[46] R. Parajuli, P. A. Østergaard, T. Dalgaard, G. R. Pokharel, Energy consumption projection of Nepal: An econometric approach, Renewable Energy 63 (2014) 432 – 444. doi:https://doi.org/10.1016/j.renene.2013.09.048.

[47] N. Magnani, A. Vaona, Access to electricity and socio-economic characteristics: Panel data evidence at the country level, Energy 103 (2016) 447 – 455. doi:https://doi.org/10.1016/j.energy.2016.02.106.

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

[49] S. Fan, L. Chen, W.-J. Lee, Machine learning based switching model for electricity load forecasting, Energy Conversion and Management 49 (6) (2008) 1331 – 1344. doi:https://doi.org/10.1016/j.enconman.2008.01.008.

[50] W.-C. Hong, Electric load forecasting by support vector model, Applied Mathematical Modelling 33 (5) (2009) 2444 – 2454. doi:https://doi.org/10.1016/j.apm.2008.07.010.

[51] J. Wang, W. Zhu, W. Zhang, D. Sun, A trend fixed on firstly and seasonal adjustment model combined with the ϵ-svr for short-term forecasting of

electricity demand, Energy Policy 37 (11) (2009) 4901 – 4909. doi:https://doi.org/10.1016/j.enpol.2009.06.046.

[52] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 1995.
URL https://www.springer.com/br/book/9780387987804

[53] B. Lantz, Machine learning with R, Packt Publishing Ltd, 2015.
URL https://www.packtpub.com/big-data-and-business-intelligence/machine-learning-r

[54] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, r package version 1.6-8 (2017).
URL https://CRAN.R-project.org/package=e1071

[55] P. Cortez, Modern optimization with R, Springer, 2014.
URL https://www.springer.com/la/book/9783319082622

[56] A. Pérez-Foguet, R. Giné-Garriga, M. Ortego, Compositional data for global monitoring: The case of drinking water and sanitation, Science of The Total Environment 590-591 (2017) 554 – 565. doi:https://doi.org/10.1016/j.scitotenv.2017.02.220.

[57] S. N. Wood, Generalized additive models: an introduction with R, CRC press, 2017.

[58] Y.-J. Kim, C. Gu, Smoothing spline Gaussian regression: more scalable computation via efficient approximation, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66 (2) (2004) 337–356. doi:https://doi.org/10.1046/j.1369-7412.2003.05316.x.

[59] K. G. van den Boogaart, R. Tolosana, M. Bren, compositions: Compositional Data Analysis, r package version 1.40-1 (2014).
URL https://CRAN.R-project.org/package=compositions

[60] M. Templ, K. Hron, P. Filzmoser, robCompositions: an R-package for robust statistical analysis of compositional data, John Wiley and Sons, 2011. doi:https://doi.org/10.1002/9781119976462.ch25.

[61] S. N. Wood, Thin-plate regression splines, Journal of the Royal Statistical Society (B) 65 (1) (2003) 95–114. doi:https://doi.org/10.1111/1467-9868.00374.