MASTER'S THESIS

Interuniversity Master in Statistics and Operations Research UPC-UB

Title: A Compositional Approach For Modelling SDG7 Indicators

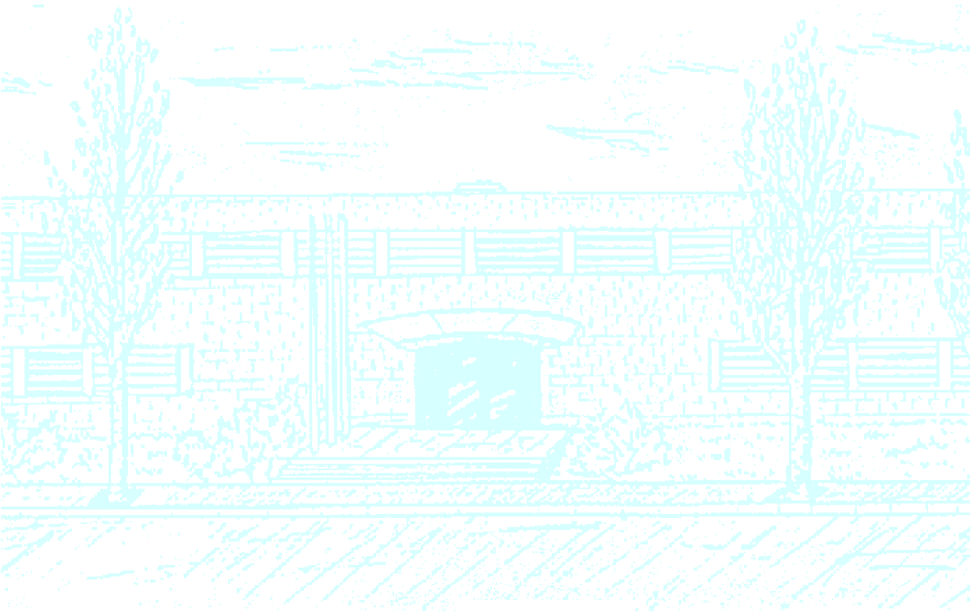
Author: Juan Carlos Marcillo Delgado

Advisors: María Isabel Ortego and Agustí Pérez Foguet

Department: Civil and Environmental Engineering. Applied Mathematics and Statistics Section.

University: Universitat Politècnica de Catalunya

Academic year: 2017/2018



UPC

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Dedicado a mi compañera de lucha, mi amiga fiel
que siempre me hace emerger de las cenizas:
Jacqueline Tatiana Hidrobo Morales
.

Acknowledgments

Agradezco a Dios, por darme la salud, la fortaleza y sobre todo su luz para culminar esta maestría en Estadística.

En segundo lugar a la UPC y la UB por generar un ambiente de calidad y estar dotada de profesores que hacen que el aprendizaje sea más efectivo.

A mi familia por brindarme su apoyo constante. A Moncerrate que siempre monitorea mis avances en las tesis, a Fátima por permitirme ser parte de su alegría, a Ángel por sus sabios consejos y a María y Mercedes que han superado la categoría de hermanas y son como una madre para mí.

A don Guillermo y doña Anita por haberme acogido en su hogar, por brindarme todas las facilidades a su alcance en este proceso y por todos los bellos momentos compartidos.

Un agradecimiento especial a Agustí y Maribel por admitirme en su grupo selecto de tesistas, por sus discursos llenos de base científica y sobre todo por ayudarme a sacarle provecho a esta etapa de aprendizaje.

Abstract

The monitoring of indicators related to the energy sector has acquired a renewed interest with the 2030 agenda on sustainable development goals (SDG), specifically with the goal seven that seeks to guarantee universal access to energy. The broad-based nature of energy has promoted the use of substantive, broadly indicative and effective metrics that allow to capture the different dimensions of energy access. A relevant characteristic of these indicators is that they can be expressed as proportions or can be disaggregated from a whole, i.e, they are compositions. This type of indicators have their own characteristics which make not suitable to implement traditional multivariate models on them. The mathematical structure and methods developed for the treatment of compositions is Compositional Data Analysis (CoDa). Following this methodology, a log-ratio transformation can be chosen to bring these indicators to the space of real numbers, and then apply any multivariate method. The scope of this TFM is to apply compositional models based on an isometric log-ratio transformation to follow up on temporary indicators of the energy sector in the context of SDG7. The electricity access indicator was selected to develop this aim. The existing dichotomy between the urban and rural sectors is considered. This dichotomy is very important since the problem of electricity access is predominantly rural. It is presented an analysis for five countries (Bangladesh, India, Kenya, Nigeria and Sudan) belonging to the areas most affected by the problem of electricity access, such as the Sub-Saharan region and the South of Asia. It is concluded 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. Three statistical methods have been used: a traditional one which the majority is related to (Linear Regression), another based on linear predictors that involve the use of smoothing functions for covariates such as (Generalized Additive Model) and the other based on optimization algorithms ($\epsilon-$SVM).

Keywords— SDG7, Electricity access, Compositional data analysis, Trend analysis.

Resumen

El monitoreo de indicadores relacionados con el sector energético ha adquirido un renovado interés con la agenda 2030 que trata los Objetivos del Desarrollo Sustentable (ODS), específicamente con el objetivo siete que busca garantizar el acceso universal a energía. La naturaleza amplia del sector energía ha promovido el uso de indicadores sustantivos, ampliamente indicativos y efectivos que permitan capturar las diferentes dimensiones de acceso a energía. Una característica relevante de estos indicadores es que pueden establecerse como proporciones o pueden desagregarse de un todo, es decir, son composiciones. Este tipo de indicadores tienen sus propias características, que entre otras cosas, hacen que no sea adecuado la implementación de los modelos multivariados tradicionales. La estructura matemática y métodos desarrollados para el tratamiento de composiciones se denomina Análisis de Datos Composicionales (CoDa). Siguiendo esta metodología, se debe escoger una transformación log-ratio que permita usar estos indicadores en el espacio de los números reales y con ello aplicar cualquier método multivariado. El objeto de este TFM es aplicar modelos de composición basados en una transformación isométrica log-ratio para hacer un seguimiento de los indicadores temporales del sector energético en el contexto del ODS 7. El indicador de acceso a la electricidad fue seleccionado para desarrollar este objetivo. Para ello se consideró la dicotomía existente entre el sector urbano y rural. Esta dicotomía es muy importante puesto que el problema de acceso a electricidad es predominantemente rural. Para efectos de la presente se realizó un análisis de cinco países (Bangladesh, India, Kenia, Nigeria y Sudán), los cuales pertenecen a las áreas más afectadas por el problema de acceso a electricidad, como es la región de África subsahariana y el sur de Asia. Se concluye que CoDa facilita una gestión más controlada de las partes que componen el indicador, especialmente cuando se trata de hacer inferencias fuera del rango de calibración. Para esto se utilizaron tres métodos estadísticos: uno tradicional con el que la mayoría está relacionada (Regresión lineal), otro basado en predictores lineales que implican el uso de funciones de suavizamiento para covariables como es el caso de (Modelo Aditivo Generalizado) y otro basado en algoritmos de optimización ($\epsilon-$MVS).

Palabras clave— ODS 7, Acceso a electricidad, Análisis de datos composicionales, Análisis de tendencias.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This research revolves around the monitoring temporary indicators of the energy sector in the context of SDG7 and the consideration of their compositional characteristics to better respond to trend modelling. It is focused on electricity access, taking as unit of analysis countries that have large population without electricity access as India, Bangladesh, Nigeria, Kenya and Sudan.

1.1 Problem Description

Monitoring energy indicators has received a renewed interest with the 2030 agenda about Sustainable Development Goals SDG that in a certain way succeed the Millennium Development Goals (MDGs) which expired at the end of 2015. They express widespread public concern about social, economic and environmental areas of sustainable development performed in a framework of 17 goals 169 targets (Barclay et al., 2015).

Within this goals, the SDG7 seeks to ”ensure access to affordable, reliable, sustainable and modern energy for all”. It includes three targets a) ensure universal access to affordable, reliable and modern energy services, b) increase substantially the share of renewable energy in the global energy mix, c) double the global rate of improvement in energy efficiency (UN, 2015). In order to correctly follow these objectives, it is necessary to use indicators that address the problem not only globally, but also represent the multidimensional nature of energy access.

Table 1.1.1 summaries some energy World-Bank indicators to monitor SDG7 where most of them share the common characteristic that they are the percentage of a whole, i.e., they are proportions between 0 and 1. In order to obtain a robust analysis with this kind of data, it is very important to consider each one of their parts. However, when modelling energy multidimensional indicators the compositional character is very neglected. In this sense, it is important to mention that traditional multivariate methods could give misleading results (Van den Boogaart and Tolosana-Delgado, 2013) if the compositional characteristics of these indicators are not addressed.

Table 1.1.1: Monitoring SDG7 indicators.

Target	World-Bank Indicator
Ensure universal access to affordable, reliable and modern energy services	Electricity access (% of Population Total) Access to clean fuels and technologies for cooking (% of population)
Increase substantially the share of renewable energy in the global energy mix	Renewable energy consumption (% TFEC)
Double the global rate of improvement in energy efficiency	Energy intensity level of primary energy (MJ / $ 2011 PPP GDP)

It is important to emphasize that most indicators respond better to one type of statistical models than others at the moment of modelling. For instance, a similar study conducted by Pérez-Foguet et al. (2017) in the case of drinking water and sanitation where the response variable was a composition and consequently Compositional Data Analysis (CoDa) was applied, it was compared between ordinary least squares (OLS) and generalized additive model (GAM) observed performance improvements when the data is transformed with CoDa but also models respond better for GAM than LM.

Other relevant characteristic of energy access is that the problem is concentrated in few countries (Bhattacharyya, 2006). The 63% of world population without electricity access is focused on 9 countries, mostly located in South Asia and Central Africa. India alone accounts with around the 25% of world population without electricity access (WorldBank, 2017).

One of the main limitations to ensure energy access in a country is the geographical distribution of the population (Chaurey et al., 2004). For instance, electricity access advances faster in urban than rural areas. World urban access rate was 96% in 2016, and the rural rate 73% (Cozzi et al., 2017). Onyeji et al. (2012) shows that the size of rural population plays an important role, especially in Sub-Saharan Africa. This implies that the compositional character of this indicator should be considered when tracking this problem.

Based on this precedents the problem is systematized with the following questions:

Does Compositional Data Analysis (CoDa) consideration improve the temporary interpolations of the SDG7?

This question gives rise to a series of sub-questions:

What statistical model should be used to model this type of data?

How do statistical models behave both outside and within the calibration range?

What happens if standard statistical techniques are applied on such data in raw form?

1.2 General objective

Assess the application of compositional models to follow up on the temporary indicators of the energy sector in the context of SDG7.

1.3 Specific objectives

The specific objectives that will contribute to develop the *general objective* of this work are the following:

- To show the compositional character of SDG7 temporary indicators
- To develop statistical models combined with compositional data methods
- To show the behavior of compositional models in both outside and within the calibration range
- To show some consequences of using standard statistical techniques on such data in raw form

Chapter 2

Literature Review

2.1 Universal Energy Access ¿ What is meant by energy access?

Without a doubt, the first step to address the energy access problem is to try to define it as clearly and consistently as possible. In the literature you can find many definitions of energy access, some perhaps more complex than others. In this section we try to show the dynamic and evolutionary character of the concept of access to energy and how it can be enriched as more contributions are generated in the energy area.

As a first advance we can mention the definition of Pachauri et al. (2012) which is very easy to understand: *"Universal access to modern energy is the physical availability of electricity and modern energy carriers and improved end-use devices such as cook stoves at affordable prices for all"*. So this approach considers that the energy access problem is determined by three indicators (energy input, improved end-use devices, affordable prices).

This definition is very practical but perhaps for some researchers it may seem very subjective for not considering another aspects such as cultural practices (Mensah et al., 2014). For this reason, it is very common for researchers to adapt the energy access definition to their research projects. For instance, Kanagawa and Nakata (2008) in their study define *energy access improvement* as:

Electrification in analyzed areas through dissemination of electric lighting appliances such as incandescent bulbs, fluorescent tubes, and Compact Fluorescent Lamps (CFL) instead of using kerosene lamps.

From the above it is clear that energy access definition is a dynamic concept that can be adjusted to the particular needs of policy makers. In addition, it is a growing concept and as applications are being made, more contributions will be given to improve their understanding. For instance, International Energy Agency IEA in 2013 established the next definition for modern energy access:

Reliable and affordable access by a household to clean cooking facilities, a first connection to electricity and then

> *an increasing level of electricity consumption over time.* (Birol et al., 2013).

This definition is centered in households and some countries like Ghana started to use a similar definition, but authors like Mensah et al. (2014) found some weaknesses associated with it. Between them, the failure to account for energy access for economic activities which is critical for development. Actually the concept of energy access has been improved for the IEA and it is considered other aspects as the economic aspects and public services.

> *Access to modern energy services includes household access to a minimum level of electricity; household access to safer and more sustainable cooking and heating fuels and stoves than traditional biomass stoves; access that enables productive economic activity; and access for public services.* (Cozzi et al., 2017).

Surely in the literature there will be many waybills to define energy access. One of them is the one proposed by Spalding-Fecher et al. (2005) which states that an energy access strategy must start with people's needs in the household (lighting, cooking, heating/cooling), at work (productive activity) and in the economy in general (industrial development), taking into account geographical location, the amount of available investment, as well as the sustainability of energy sources (renewable or non-renewable).

It is concluded that the definition of energy access is a moving concept. It does not need to remain unchanged over time. In addition, reaching a consensus on access to energy is full of difficulties, one of the problems is that bibliography is generally based on poverty literature, which would seek only to ensure a minimum supply of energy (Bhattacharyya and Ohiare, 2012) and the goal is rather to achieve certain vital services that improve human well-being (communication, illumination, thermal comfort, entertainment, etc.) (Groh et al., 2016).

2.2 Contextualization of universal energy access problem

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 (Bazilian et al., 2010b) and generally 90% of that energy is for heating and cooking and the rest (10%) for lighting and entertainment needs (Bhattacharyya and Ohiare, 2012). Most of the energy access affected people are concentrated in Sub-Saharan Africa, South Asia and some developing countries where India has the largest population without electricity access.

Recent metrics suggest that globally there are 1.060 million people lack access to the electricity grid. Where more than two-thirds of those lacking electricity access are concentrated in 12 countries (Bhattacharyya, 2006). Most of this unelectrified population resides in Sub-Saharan Africa (55%) and some developing countries from Asia (41%), where India represents the (23%) of people without adequate energy access lives (Cozzi et al., 2017).

In addition, it should be highlighted the precarity of energy use as a factor associated with low rates of electricity consumption. For instance, it is typical to have grid capacity problems in countries like Bangladesh, especially at peak load times (Groh et al., 2016). Furthermore, developing countries strongly depend on fossil fuel for electricity generation (Magnani and Vaona, 2016) which leads, among other problems, those related to health and the environment.

On the cooking fuels side, there are 2500 million people in the world relying on biomass where 66% are represented for developing countries of Asia, standing out the relevance of India (31%) and China (12%) and considering Sub-Saharan Africa (33%) . All this represents 99% of biomass world consumption (Cozzi et al., 2017).

Although Sub-Saharan Africa and South Asia are the most affected regions by energy access, it should be noted that the dimensionality of the problem is different on them. For instance, for cooking purposes in Sub-Saharan Africa it is very common the predominance of solid fuels. Coal and charcoal is more used in China (and to some extent in India) while gas (LPG, bio gas and natural gas) is more used in developing countries outside from Sub-Saharan Africa regions (Bhattacharyya and Ohiare, 2012).

On the other hand, it is interesting to highlight that one of the greatest challenges to achieving full electrification is the dichotomy between rural and urban areas (Mensah et al., 2014). In fact, some authors consider that energy access is predominantly a rural problem (Bhattacharyya and Ohiare, 2012). For instance, in Sub-Saharan Africa 23% of rural population have electricity access while in urban areas 74% of them have electrification. A study conducted by Doll and Pachauri (2010) showed that one of the causes in delay electrification process in Sub-Saharan Africa is that there are large populations living at low densities.

The lack of modern energies makes it common to adopt other traditional sources in rural areas. It is very common to adopt kerosene lamps in order to meet household lighting demand in developing countries. But maintaining this type of energy alternatives in the long term generates a stagnation in the people quality life, For instance, they do not obtain sufficient lighting for studying in a house at night, and this is one of the obstacles to achieve higher educational attainment (Kanagawa and Nakata, 2008).

It is worth noting that one of the main predominant characteristics of the region without access to electricity is poverty and it is the major obstacle for ensure universal energy access. In the case of India, energy provision has two dominant characteristics: (1) strong public sector presence and (2) prevalence of excessive subsidies and cross-subsidies (Bhattacharyya, 2006) and it will represent a great barrier for low-income countries.

Finally, in order to achieve universal energy access in 2030 could be considered experiences of those countries that have obtained good results in their energy supply process Thailand went from 25% access to 100% electricity in less than a decade (Pachauri et al., 2012). China has successfully developed rural diversification energy projects over the last few years and achieved a great feat of almost 100% power supply and projects like grid extension has emerged as the preferred mode of electrification in almost all successful cases (Bhattacharyya and Ohiare, 2012).

2.3 The context of the indicators for measuring universal energy access

The obstacles to widespread electricity access are largely well known . However, one of the main concerns from the 2030 agenda it has been the availability of indicators to monitor this target, which is why the issue of availability, versatility and the source of this type of indicators has been very approached (Mensah et al., 2014; Hailu, 2012).

In general an indicator is a quantitative or a qualitative measure derived from a series of observed facts that can reveal relative positions (e.g. of a country) in a given area. Indicators are useful in identifying trends and drawing attention to particular issues. They can also be helpful in setting policy priorities and in benchmarking or monitoring performance (Commission et al., 2008).

To ensure such a complex goal as the SDG7 for 2030 involves a robust information-base which serves as a support tool that allows tracking and comparing objectives (Nussbaumer et al., 2012). Nevertheless, the broad-based nature of energy, requires the development of energy access measures be substantive, broadly indicative and effective in capturing the different dimensions of energy access (Hailu, 2012; Groh et al., 2016; Mensah et al., 2014).

Thereby, within the existing debate on energy access indicators many researchers highligh the importance of addressing the multidimensional nature of energy access. For instance, Nussbaumer et al. (2012) in their speech about uni- versus multi-dimensionality foregrounds that complex issues such as human development are multidimensional in their very nature and their assessment therefore requires a framework in which various elements can be captured.

Despite the importance of more disaggregated indicators. The planning scenarios and forecasting methods in energy access are usually simple, focus primarily on one metric, namely, installed generation capacity (Bazilian et al., 2012) or said in a more specialized language, electricity demand. Planning process based on a single metric is commonly more appealing from a communication viewpoint but it neglect other factors (e.g. population, energy mix, efficiency) that complicate the background of the problem (Bazilian et al., 2010a).

Fortunately, with the implementation of the SDG7, many efforts have been directed towards the diversification of energy indicators and data quality. Many of these contributed by entities such as the World Bank or the International Energy Agency IEA which have the advantage of being commonly applicable across countries and allowing cross comparability.

One of the recent advances in order to assess energy access for households, productive entities, and communities along several dimensions of access is the multi-tier framework (MTF) which is used to reflect both aggregated and dissected analyses as be possible (Groh et al., 2016).

Thus, it becomes clear the need of disaggregated indicators and consequently of methods that make easier the modelling of them. Achieve this task will allow to take advantage of the information richness that this kind of data provides against the use of only uni-dimensional indicators.

Finally, one can mention as other limiting factor for the correct monitoring of access to electricity the poor and inconsistent national statistics that could block cross-country analysis and undermine efforts to implement global or regional programmes. Still, a lack of data should not be use as a justification for delaying building national energy planning capability and developing energy plans (Bazilian et al., 2012).

2.4 The compositional character of universal energy access indicators

One of the main causes that limits multidimensional indicators use is the complexity existing around the correct handling of them (Hailu, 2012). For example, if any statistical classical analysis is carried out on all the parts that make up a multidimensional indicator of proportions with unit-sum constraint, it can lead to erroneous interpretations since the hidden influence of this type of indicators is neglected (Reimann et al., 2017). Fortunately, there are significant advances in the management of multidimensional indicators, especially those that can be established as proportions, or can be disaggregated from a total.

Aitchison (1986) aware of the existence of some multidimensional indicators type expressed mathematically in vectors of proportions play an important role in many disciplines and often they display appreciable variability from vector to vector. He developed a simple and appropriate statistical methodology for the adequate investigation and interpretation of this kind of data. This branch of statistics is known as Compositional data analysis (CoDa).

Aitchison (1986) uses the term composition to refer to any vector x with non-negative elements $X_1, ..., X_n$ representing proportions of some whole and any X_i element is called component. One of the contributions of this methodology is the incorporation of the the unit-sum constraint within the statistical modelling to eliminate any distortion or doubt about the inference over compositional indicators.

Other advantage about CoDa is that this approach makes it possible to perform classical statistical analysis (e.g. linear regression, principal component analysis) which is not suitable when using only the raw compositional data and this is a clear advantage due to the large number of normally distributed available methods for multivariate phenomena and the robustness of those. The only prerequisite you need is to work on transformed data using a log-ratio approach and back-transform the results (Pawlowsky-Glahn and Egozcue, 2001).

Within the usual log-ratio transformations, the first developed transformation was the additive log-ratio (alr), which consisted in choosing an arbitrary component as a divisor (which represented a more conceptual than practical problem). To avoid this arbitrariness problem the composition is divided by the geometric mean resulting in the centered log-ratio (clr), but the disadvantage was that the clr covariance matrix is singular. Then, the recognition that compositions can be represented in coordinates with orthonormal bases (isometric log ratio - ilr) helped to avoid the arbitrariness of the alr and the singularity of the clr Pawlowsky-Glahn and Buccianti (2011).

Thus, CoDa has been the source of many discussions in practice due to the enormous importance compositional data have in applied sciences (Pawlowsky-Glahn and Egozcue, 2001). Compositional analysis has opened a space in the research field especially in researches related to chemical compositions of rocks and sediments at different depths (Flood et al., 2016) which are compositional study fields by nature (Reimann et al., 2017). However, one of Aitchison's major concerns was related to household surveys, especially when it comes to studying certain variables as expenditure composition, the consumer demand study, including fuel and light consumption (Aitchison, 1986).

In research on energy access there are many compositional indicators, among which we can mention the total primary energy (TPE) used in the paper of Parajuli et al. (2014) which is the sum of residential, commercial, transport, agricultural and others primary energies. The simple fact of difference between urban and rural population with access to electricity is already a composition. The multi-tier framework (MTF) that assesses energy access for households uses fractional measurements between tier 0 and 1 (Groh et al., 2016) and therefore it is a compositional indicator.

It is worth noting that although compositional data analysis is a practical approach to address measures of proportions, there is a high reluctance in the monitoring of energy access indicators, either due resistance to new theories or due to that energy access multidimensional character starts to be taken into account with the 2030 agenda.

2.5 Review of compositional statistical models related with Universal Energy Access

In this section we proceed to a literature review on statistical models related to universal access to energy where the compositional character of this field of study has been considered. The main objective is to highlight how compositional indicators are becoming relevant with the agenda 2030 and how compositional data analysis methodology can contribute to the robustness of these models if considered.

The first model in discussion is an econometric model proposed for Panos et al. (2016) to estimate electricity access (% of population total) based on a ordinary least squares linear regression and using the covariates: a) percentage of population living with less than $2 per day (poverty covariate) b) the population urbanization rate (urbanization rate) and c) the average electricity per capita in residential sector.

This is an interesting example because most variates are proportions (i.e. compositions). This author, in spite of not working with compositional methods, is conscious that the dependent variable responds to a compositional structure and to deal with this peculiarity the variate was transformed such a logit form $\ln(\text{electricity access}/(1 - \text{electricity access}))$ which is a very valid transformation within the vision of CoDa since it transforms the scale and considers both parties (access and no access to electricity).

Parajuli et al. (2014) performs a more complex model that mixes several compositions within it, using Cobb-Douglas log-linear models to project the primary energy consumption in Nepal. This model resembles the one proposed in this project because it uses a composition as response variable, i.e., it consist in several models to explain each components of total primary energy (residential energy, commercial energy, energy in transport and energy in agricultural sector). In addition, some of its covariates include other compositions such as the disaggregation of GDP by sector (commercial, agricultural, industrial) and population total (disaggregated by urban and rural).

If the compositional part were neglected, the model would be reduced to a linear regression with estimation problems, but approaching it from a compositional viewpoint is a important challenge because you have to deal with three compositions at the same time which is not a problem with CoDa but in this case you lose control over unit-sum constraint for the three compositions which generates many doubts when making inference since the components are out of control. The use of neperian logarithm as a transformation for urban and rural population is a bit redeemable but one must be very cautious with interpretations since when one of the parties increases the other must decrease.

By other side Magnani and Vaona (2016) performed a panel data model where it was preferred a linear model to a log-linear one not to constrain the elasticity of the dependent variable with respect to the independent ones to be constant throughout the sample. The percentage of the population with access to electricity was used as dependent variable. Five models were estimated. In this example the author chooses to include as an explanatory variable the urban or rural population but not both at the same time.

In this case, the compositional character of the variables is totally omitted. According with Van den Boogaart and Tolosana-Delgado (2013) most multivariate methods developed for data with real value give misleading results for compositional data and can lead to spurious correlations. For this example the correlation structure between electricity access, urban and rural population gave negative values, contradicting the usual interpretations of correlation and covariance, among other things that independence is usually related to zero correlation. Moreover, compositional researchers are very critical when one of the parts of the whole is omitted because you cannot model and interpret compositional indicators correctly.

As highlighted in this section, with the 2030 agenda a series of statistical models have been carried out, using known statistical tools such as the linear regression estimated by ordinary least squares. However, there is no evidence of models that use CoDa as a tool that helps improve statistical robustness.

Within the models to be used in this thesis, besides OLS, there are Generalized Additive Models (GAM) and Support Vector Machine (SVM). To which is added that there is no evidence of SVM using CoDa, although it is a widely used model in the energy field (Suganthi and Samuel, 2012; Ekonomou, 2010), especially to predict energy demand from a one-dimensional approach.

About GAM, a recent study conducted by Pérez-Foguet et al. (2017) related to goal six of the SDG: *Ensure availability and sustainable management of water and sanitation for all*, where linear regression was also used, it was shown that CoDa is a useful tool that can help improve temporary interpolations for trend models, which serves as a reference for the present study.

Chapter 3

Research Methodology

3.1 Unit of Analysis

This study is replicable to most countries that are included in the database of access to electricity that is on the World Bank website. However, for the purposes of this thesis it was decided to select the most representative ones. In this sense, it was selected certain countries within the area with greater problems of energy access in the world such as the Sub-Saharan region and southern Asia, considering additionally other aspects such as the existence of data enough to estimate a statistical model. In general, most countries had information available from 1990 to 2014.

These countries are Bangladesh, India, Kenya, Nigeria and Sudan. As shown in figure 3.1.1 nine countries represent around the 63% of the population with problems of electricity access in the world where the countries selected represent around the 50%. However, it is worth noting that the models made here are replicable for the 212 countries in the World Bank database that is the source of information from which the data was taken.

3.2 Study Variables

3.2.1 World Bank methodology for collecting electricity access data

Data for monitoring access to electricity are collected among different sources: mostly data from nationally representative household surveys (including national censuses) were used. Survey sources include Demographic and Health Surveys (DHS) and Living Standards Measurement Surveys (LSMS), Multi-Indicator Cluster Surveys (MICS), the World Health Survey (WHS), other nationally developed and implemented surveys, and various government agencies (for example, ministries of energy and utilities) (WorldBank, 2017).

Given the low frequency and the regional distribution of some surveys, a number of countries have gaps in the

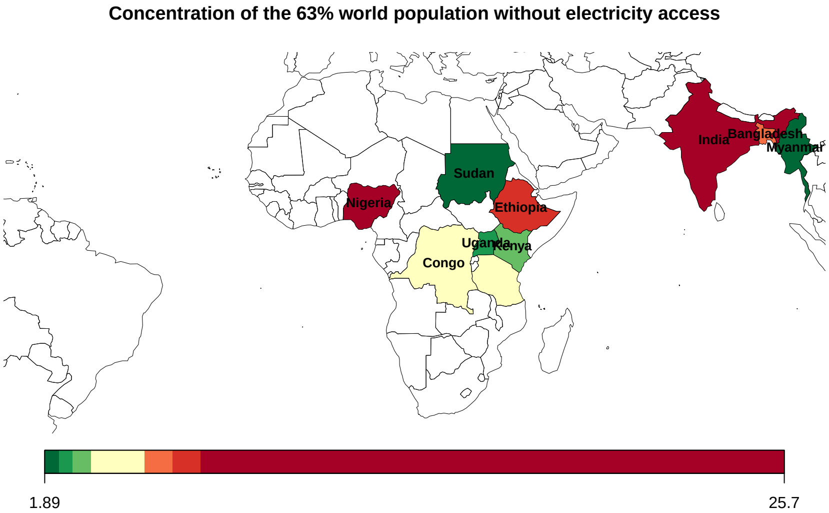

Concentration of the 63% world population without electricity access
India
Bangladesh
Myanmar
Sudan
Nigeria
Ethiopia
Uganda
Kenya
Congo
1.89
25.7

Figure 3.1.1: Countries with large population without electricity access in 2010. Source: WorldBank (2017). The colors of the palette represent the distribution of the 63% world's population without access to electricity through nine countries. The other 37% is in the rest of the world. Countries close to the green color have less population without electricity than those that there are closer to the wine color, specially India who concentrates the 25,7% of population without electricity access.

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, around 2000, and around 2010. Therefore, a country can have a continuum of zero to three data points (WorldBank, 2017).

There are 42 countries with zero data point and the weighted regional average was used as an estimate for electrification in each of the data periods. 170 countries have between one and three data points and missing data are estimated by using a model with region, country, and time variables (ibid.).

The model keeps the original observation if data is available for any of the time periods. This modelling approach allowed the estimation of electrification rates for 212 countries over these three time periods (Indicated as ”Estimate”). Notation ”Assumption” refers to the assumption of universal access in countries classified as developed by the United Nations (ibid.).

3.2.2 Dependent variable

In the present study the dependent variable or variable to explain is the composition of electricity access (access, without access), disaggregated by sector (urban, rural), in total it is a composition of four parts. Next, the description of each one of them and in parenthesis the pseudonym that will be assigned to them in this study:

x_1 : Urban population with electricity access (urban).

x_2 : Rural population with electricity access (rural).

x_3 : Urban population without electricity access (nourban).

x_4 : Rural population without electricity access (norural).

The variables used by the World Bank for the construction of this indicator were:

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)

Where, table 3.2.1 reflects how the dependent variable was created using the variables of the World Bank exposed from A to D:

Table 3.2.1: Formulas used for establishing the composition of dependent variable.

Component	Pseudonym	Formula
x_1	urban	$\frac{A \cdot C}{100 \cdot 100}$
x_2	rural	$\frac{B \cdot D}{100 \cdot 100}$
x_3	nourban	$C -$ urban
x_4	norural	$D -$ rural

Maybe if you are interested in knowing the amount of total population, the variable is registered in the World Bank with the name `Population, total`.

3.2.3 Independent variables

Since the present study focuses on representing the trend of the response variable, the explanatory variable is the time. Nevertheless, there is a variable called `Access to electricity (% of population)` within the database of World Bank that in table 3.2.1 would be the sum of $x_1 + x_2$.

This variable gives rise to the creation of a contrast variable z, which reflects the existing harmony between `Access to electricity (% of population)` and the variables `Access to electricity, urban (%`

`of urban population)` and `Access to electricity, rural (% of rural population).` With this inquiry it is detected the formation of two subseries within the components of the dependent variable. The creation of this binary explanatory variable Z is given by equation 3.1.

x_1 : Urban population with electricity access (urban).

x_2 : Rural population with electricity access (rural).

T : Access to electricity (% of population).

$$z = \begin{cases} 0, & \text{if } T = x_1 + x_2 \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

3.3 Statistical Analysis of Compositional Data

This section is dedicated to describe CoDa aspects that are related to this thesis. The first section is framed to clarifying the compositions as portions of a total. The second section, using more mathematical terms gives a clear vision of what a composition is, and relate the reader to the CoDa terminology. The third section discloses the three basic principles on which CoDa is based.

Next, it is introduced the vector space structure used in CoDa. Then it shows the compositional observations in real space and the need to transform the data, for example with an isometric log ratio (the transformation used in the results chapter), also showing a very simple method for this procedure, as is the SBP method and finally it is introduced the principle of working in coordinates to apply any standard statistical process.

It is important to mention that this section was developed using the book Pawlowsky-Glahn et al. (2015), except in those sections where another author is directly mentioned.

3.3.1 Compositions are portions of a total

A dataset is called *compositional* if it provides portions of a total. The individual parts of the composition are called *components*. Each component has an amount, representing its importance within the whole. Amounts can be measured as absolute values, in amount-type physical values like money, time, volume, mass, energy, molecules, individuals, and events (Van den Boogaart and Tolosana-Delgado, 2013).

The sum over the amounts of all components is called the *total amount* or, short, the *total*. *Portions* are the individual amounts divided by this total amount. Depending on the unit chosen for the amounts, the actual portions of the parts in a total can be different (Van den Boogaart and Tolosana-Delgado, 2013).

3.3.2 Basic concepts

In this section we discuss some fundamental concepts to understand CoDa, all of them help to understand our objective variable. For instance, the first definition applied to our study clarifies that our response variable, access to electricity, is a composition of $D = 4$ parts (urban, rural, nourban, norural).

Definition 3.1. (*D*-part composition).

A (row) vector, $x = [x_1, x_2, ..., x_D]$, is a D-part composition when all its components are strictly positive real numbers and carry only relative information.

The second definition helps us to clarify that within our study there could be compositions that are multiple of others.

Definition 3.2. ***(Compositions as equivalence classes).***

Two vectors of D positive real components $x, y \in \mathbb{R}^D_+(x_i, y_i > 0$, for all $i = 1, 2, ..., D)$ are compositionally equivalent if there exists a positive constant $\lambda \in \mathbb{R}_+$ such that $x = \lambda \cdot y$

Knowing what a closed operation is it becomes evident that in the present study, the four-part composition, electricity access, is a closed composition and always adds $k = 1$.

Definition 3.3. ***(Closure).***

For any vector of D strictly positive real components,

$$z = [z_1, z_2, ..., z_D] \in \mathbb{R}^D_+, z_i > 0 \forall i = 1, 2, ..., D$$

the closure of z to $\mathcal{K} > 0$ is defined as

$$\mathcal{C}(z) = \left[\frac{\mathcal{K} \cdot z_1}{\sum_{i=1}^{D}}, \frac{\mathcal{K} \cdot z_2}{\sum_{i=1}^{D}}, ..., \frac{\mathcal{K} \cdot z_D}{\sum_{i=1}^{D}},\right]$$

Additionally, it is clear that each country that is modeled includes a sample space, a simplex:

Definition 3.4. ***(Sample space).***

The sample space of compositional data is the simplex,

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

Furthermore, if it were decided to analyze only the population with access to electricity, leaving aside the population without access to electricity, it would be analyzing only part of the proposed whole, a subcomposition. This definition also makes clear that all compositions are subcompositions, since in the case of the dependent variable, this could be disaggregated into other factors.

Definition 3.5. ***(Subcomposition).***

Given a composition x and a selection of indices $S = \{i_1, ..., i_S\}$, a subcomposition x_S , with S parts, is obtained by applying the closure operation to the subvector $[x_{i_1}, x_{i_2}, ..., x_{i_S}]$ of x. The set of subscripts S indicate which parts are selected in the subcomposition, not necessarily the first S ones.

Finally, each time an indicator is added, it is done an amalgamation process, it means that the wealth of the indicator is being removed.

Definition 3.6. ***(Amalgamation).***

Given a composition $x \in \mathcal{S}^D$, and a selection of a indices $A = \{i_1, ..., i_a\}$ (not necessarily the first ones), $D - a \geq 1$, and the set of remaining indices $\bar{A}$, the value

$$x_A = \sum_{j \in A} x_i$$

is called amalgamated part or amalgamated component. The vector $x' = [x_{\bar{A}}, x_A]$, containing the components with subscript in $\bar{A}$ grouped in $x_{\bar{A}}$ and the amalgamated component x_A , is called amalgamated composition which is in $\mathcal{S}^{D-a+1}$.

3.3.3 Principles of compositional analysis

The principles presented in this section are the basis for a robust compositional data analysis and the reason why CoDa is a useful tool in satisfying the three fundamental principles of compositional data:

Scale invariance

In the absence of information about the total (total power production or mass of sediment), it is highly reasonable to expect analyses to yield the same results, in whichever way that total evolved. This is known as scale invariance (Aitchison, 1986).

Permutation invariance

A function is permutation invariant if it yields equivalent results when the ordering of the parts in the composition is changed. As a little example, it should be the same working with the composition $[A, B, C]$ than with $[B, A, C]$.

Subcompositional coherence

Subcompositional coherence can be practically summarized as: (i) distances between two compositions are equal or decrease when subcompositions of the original ones are considered; (ii) scale invariance of the results is preserved within arbitrary subcompositions, that is, the ratios between any parts in the subcomposition are equal to the corresponding ratios in the original composition.

3.3.4 Vector space structure

This section describes the basic operations required for a vector space structure of the simplex. The symbol of ” $\oplus$” is shown in replacement of ” $+$ ” and ” $\odot$ ” in replacement of ” $\cdot$ ”. They use the closure operation $\mathcal{C}$ that was reviewed in the previous section:

Definition 3.7. ***(Perturbation).***

Perturbation of $x \in \mathcal{S}^D$ by $y \in \mathcal{S}^D$,

$$x \oplus y = \mathcal{C}[x_1 y_1, x_2 y_2, ..., x_D y_D] \in \mathcal{S}^D$$

Definition 3.8. ***(Powering).***

Power transformation or powering of $x \in \mathcal{S}^D$ by a constant $\alpha \in \mathbb{R}$,

$$\alpha \odot x = \mathcal{C}[x_1^\alpha, x_2^\alpha, ..., x_D^\alpha] \in \mathcal{S}^D$$

3.3.5 Compositional observations in real space

Compositions in $\mathcal{S}^D$ are usually expressed in terms of the canonical basis of $\mathbb{R}^D$, $\{e_1, e_2, ..., e_D\}$. In fact, any vector $x \in \mathbb{R}^D$ can be written as

$$x = x_1[1,0,...,0] + x_2[0,1,...,0] + ... + x_D[0,0,...,1] = \sum_{i=1}^{D} x_i \cdot e_i \tag{3.2}$$

and this is the way we are used to interpret it. The problem is that the set of vectors $\{e_1, e_2, ..., e_D\}$ is neither a generating system nor a basis with respect to the vector space structure of S^D. Therefore, the aim is to find a basis that correspond to the vector space structure of $\mathcal{S}^D$ through log-ratio transformations. Once the base has been chosen we work with the coordinates of the composition and then will be applied any standard methodology. Finally, to express the results in the raw composition is necessary to back-transform de data.

Most methods from multivariate statistics developed for real valued datasets are misleading or inapplicable for compositional datasets, for various reasons (Van den Boogaart and Tolosana-Delgado, 2013):

- Independent components mixed together and closed exhibit negative correlations
- Covariance between two components depends on which other components are reported in the dataset.
- Variance matrices are always singular due to the constant sum constraints.
- Components cannot be normally distributed, due to the bounded range of values.

3.3.6 Isometric logratio transformation and coordinates

Among the most used log-ratio transformations are the additive logratio (alr), the centered logratio (clr), and the isometric logratio (ilr) which is the one used in the present study. This process was chosen since it allows to choose an orthonormal basis of $\mathcal{S}^D$ as an Euclidean space. The coordinates of compositions on this base allow to model on them. Therefore, it is convenient to study its general characteristics.

Definition 3.9. ***(Isometric logratio transformation and coordinates).***

Let $\{e_1, e_2, ..., e_{D-1}\}$ be an orthonormal basis of the simplex $\mathcal{S}^D$. The isometric logratio transformation, ilr for short, of the composition x is the function ilr: $\mathcal{S}^D \longrightarrow \mathbb{R}^{D-1}$, which assigns the coordinates x^*, with respect to the given basis, to the composition x. The vector x^* contains the $D-1$ ilr-coordinates of x. The inverse of the ilr-transformation is denoted as ilr^{-1}.

Once an orthonormal basis has been chosen, a composition $x \in \mathcal{S}^D$ is expressed as:

$$x = \bigoplus_{i=1}^{D-1} x_i^* \oplus e_i, \; x_i^* = \langle x, e_i \rangle_a \tag{3.3}$$

where $x_i^* = [x_1, x_2, ...x_{D-1}]$ is the vector of coordinates of x with respect to the selected basis. Formula 3.3 is useful at the moment of performing the back-transformation of the fitted statistical models presented in the next chapter.

3.3.7 Balances

There are several ways to define orthonormal bases in the simplex. An easy way is using the method sequential binary partition (SBP). The Cartesian coordinates of a composition in such a basis are called balances and the compositional vectors making up the basis balancing elements.

A *sequential binary partition (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.

For the kth order partition, it is possible to define the balance between the two subgroups formed at that level: if $i_1, i_2, ..., i_r$ are the r parts of the first subgroup (coded by $+1$) and $j_1, j_2, ..., j_s$ the s parts of the second (coded by -1), the balance is defined as the normalized logratio of the geometric mean of each group of parts:

$$b_k = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_{i_1} x_{i_2} ... x_{i_r})^{\frac{1}{r}}}{(x_{j_1} x_{j_2} ... x_{j_s})^{\frac{1}{s}}} \tag{3.4}$$

Table 3.3.1 shows the computed balances for the response variable x **electricity access** which consists in four components $[x_1 = urban, x_2 = rural, x_3 = nourban, x_4 = norural]$.

Table 3.3.1: Sign matrix for $D = 4$ to encode balances using SBP.

order	x_1	x_2	x_3	x_4	r	s	Balances
1	+1	+1	+1	-1	3	1	$b_1 = \sqrt{\frac{3 \cdot 1}{3+1}} \ln \frac{(x_1 \cdot x_2 \cdot x_3)^{\frac{1}{3}}}{x_4}$
2	-1	+1	-1	0	1	2	$b_2 = \sqrt{\frac{1 \cdot 2}{1+2}} \ln \frac{x_2}{(x_1 \cdot x_3)^{\frac{1}{2}}}$
3	+1	0	-1	0	1	1	$b_3 = \sqrt{\frac{1 \cdot 1}{1+1}} \ln \frac{x_1}{x_3}$

3.3.8 Principle of working on coordinates

The principle of working on coordinates is a general property of coordinates in real linear vector spaces and is based on the fact that coordinates with respect to an orthonormal basis obey standard rules of operation in real space. This means that, when performing analysis of compositional data, results that could be obtained using compositions and the Aitchison geometry are exactly the same as those obtained using the coordinates of the compositions and using the ordinary Euclidean geometry. In other words this principle justify the use of transformed compositions in $\mathcal{S}^{D-1}$ in the real space, thus facilitating to apply standard procedures as LM, GAM or SVM.

3.4 Statistical models for electricity access trend analysis

To model the temporary trend of electricity access, two types of statistical models will be differentiated: i) a bivariate model consisting of a compositional response variable x_i of $D = 4$ parts and the explanatory variable time t_i and ii) a model with mixed effects that introduces a fixed effect z_i over time. Where also is considered the interactions between time t_i and the binary variable z_i and i represents the time i.

For the above purposes, three types of statistical models will be applied: a) linear regression b) generalized additive model and c) support vector regression. Below are each of these models detailed.

3.4.1 Linear regression model - LM

Bivariate classical linear regression model

Following Weisberg (2005), a bivariate regression model consists of the mean function and the variance function

$$\begin{aligned} \text{E}(Y|T=t) &= \beta_0 + \beta_1 t \\ \text{Var}(Y|T=t) &= \sigma^2 \end{aligned} \tag{3.5}$$

The method for obtaining estimates of parameters in a model is called ordinary least squares, or OLS,in which parameter estimates are chosen to minimize a quantity called the *residual sum of squares* or RSS. The fitted value for case i is given by $\hat{\text{E}}(Y|T=t_i)$, for which it is used the shorthand notation $\hat{y}_i$,

$$\hat{y}_i = \hat{\text{E}}(Y|T=t_i) = \hat{\beta}_0 + \hat{\beta}_1 t_i \tag{3.6}$$

The residual for the ith case, denoted $\hat{e}_i$, is given by the equation

$$\hat{e}_i = y_i - \hat{\text{E}}(Y|T=t_i) = y_i - (\hat{\beta}_0 + \hat{\beta}_1 t_i) = y_i - \hat{y}_i, \text{ for } i = 1, ..., n \tag{3.7}$$

The criterion function for obtaining estimators is based on the residuals, which geometrically are the vertical distances between the fitted line and the actual y-values. The OLS estimators are those values β_0 and β_1 that minimize the function

$$RSS = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 \tag{3.8}$$

Compositional bivariate linear regression model

According to Pawlowsky-Glahn et al. (2015) the problem of regression when the response is compositional is stated as follows. A compositional sample in $\mathcal{S}^D$, denoted by $x_1, x_2, ..., x_n$, is available. The sample size is n. Each data point, x_i, $i = 1, 2, ..., n$ is associated, in the bivariate case, with one external variable or covariate t_i. The goal is to estimate the coefficients β_0 and β_1 of a curve or surface in $\mathcal{S}^D$ with equation

$$\hat{x}_i = \beta_0 \oplus (t_i \odot \beta_1) \tag{3.9}$$

The compositional coefficients $\beta_0, \beta_1 \in \mathcal{S}^D$ and the deviation of the model is defined as $\hat{x}_i \ominus x_i$ and its size is measured by the Aitchison norm $||\hat{x}_i \ominus x_i||_a^2 = d_a^2(\hat{x}_i, x_i)$. The target function (sum of squared errors, SSE) is

$$SSE = \sum_{i=1}^{n} ||\hat{x}_i \ominus x_i||_a^2 \tag{3.10}$$

to be minimized as a function of the compositional coefficients β_0, β_1, which are implicit in $\hat{x}$. The number of coefficients to be estimated in this linear model is $2 \cdot (D - 1)$. This least squares problem is reduced to $D - 1$ ordinary least squares problems when the compositions are expressed in coordinates with respect to a basis D of the simplex. Assume that an orthonormal basis has been chosen in $\mathcal{S}^D$ and that the coordinates of x_i, $\hat{x}_i$ and β_0, β_1 are $x_i^* = [x_{i1}^*, x_{i1}^*, ..., x_{i,D-1}^*]$, $\hat{x}_i^* = [\hat{x}_{i1}^*, \hat{x}_{i1}^*, ..., \hat{x}_{i,D-1}^*]$ and $\beta_j^* = \beta_{j1}^*, \beta_{j2}^*, ..., \beta_{j,D-1}^*$, for $i = 1, 2, ..., n$ and $j = 0, 1$ for the bivariant case. All this vectors are in $\mathbb{R}^{D-1}$. The model expressed in equation 3.9 is expressed as

$$\hat{x}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* t_i \tag{3.11}$$

For each coordinate, this expression becomes

$$\hat{x}_{ik}^* = \hat{\beta}_{0k}^* + \hat{\beta}_{1k}^* t_i, \ k = 1, 2, ..., D - 1 \tag{3.12}$$

Also, the Aitchison norm and distance become the ordinary norm and distance in real space. Then, using orthonormal coordinates, the target function is expressed as

$$SSE = \sum_{i=1}^{n} ||\hat{x}_i^* - x_i^*||_a^2 = \sum_{k=1}^{D-1} \left\{ |\hat{x}_{ik}^* - x_{ik}^*|^2 \right\} \tag{3.13}$$

Linear regression model with interactions

The linear regression model with interactions is a particular case of the linear regression that is very common when it is desired to differentiate between groups. In this particular case additional to the explanatory variable time t_i there is a binary variable z_i that could help to improve the goodness of fit of the model. This function is expressed as

$$y_i = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot z_i + \beta_3 \cdot t_i \cdot z_i + e_i, \text{ for } i = 1, 2, .., n, \qquad (3.14)$$
$$e_i \sim N(0, \sigma^2)$$

Where β_2 help to differentiate fixed effects between the groups formed by z_i and β_3 represents the existing interactions between the time t_i and the groups z_i.

The linear regression with interactions and compositional response it is introduced in the same way. The curve or surface in $\mathcal{S}^D$ is represented by:

$$x_i = \beta_0 \oplus (t_i \odot \beta_1) \oplus (z_i \odot \beta_2) \oplus (t_i \odot z_i \odot \beta_3) \oplus e_i \qquad (3.15)$$

and the model in $\mathcal{S}^{D-1}$ for the coordinates becomes

$$x^*_{ik} = \beta^*_{0k} + \beta^*_{1k} \cdot t_1 + \beta^*_{2k} \cdot z_i + \beta^*_{3k} \cdot t_i \cdot z_i + e^*_{ik} \qquad (3.16)$$

Polynomial linear regression with one predictor

If a mean function with one predictor t is smooth but not straight, integer powers of the predictors can be used to approximate $E(Y|T)$ (Weisberg, 2005). With one predictor, the polynomial mean function of degree d is:

$$E(y|t) = \beta_0 + \beta_1 t + \beta_2 t^2, ..., +\beta_d t^d \qquad (3.17)$$

this expression is possible to represent in compositions as expressed in the previous linear models.

$$E(x^*_{ik}|t) = \beta^*_{0k} + \beta^*_{1k} \cdot t_1 + \beta^*_{2k} \cdot t_i^2 + ... + \beta^*_{dk} \cdot t_i^d \qquad (3.18)$$

3.4.2 Generalized additive model - GAM

Following the book of Wood (2017) the generalized additive model (GAM) was originally developed by Trevor Hastie and Robert Tibshirani in 1986. It is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. In similar way than in linear regression, this section introduces the GAM model for a single smooth function $f(t_i)$ and the GAM model considering a numeric t_i and a binary z_i covariates. Furthermore, it is showed the model representations for compositional response models.

Generalized additive model for a single smooth function

In general the model for a single smooth function has a structure something like

$$g(\mu_i) = \mathrm{A}_i\theta + f_1(t_i) + e_i \tag{3.19}$$

where $\mu_i \equiv \mathbb{E}(y_i)$ and $y_i \sim EF(\mu_i, \phi)$, y_i is a response variable, $EF(\mu_i, \phi)$ denotes an exponential family distribution with mean μ_i and scale parameter, ϕ, A_i is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, the f is the smooth function of the covariate t_i and e_i are independent $N(0, \sigma^2)$ random variables. To solve this problem it is necessary both to represent the smooth functions in some way and to choose how smooth it should be.

To estimate f requires that it be presented is such way that (3.19) becomes a linear model. This can be done by choosing a *basis*, defining the space of functions of which f (or a close approximation to it) is an element. Choosing a basis amounts to choosing some *basis functions*, which will be treated as completely known: if $b_j(x)$ is the j^{th} such basis function, then f is assumed to have a presentation

$$f(t) = \sum_{j=1}^{k} b_j(t)\beta_j \tag{3.20}$$

for some values of the unknown parameters, β_j. Substituting (3.20) into (3.19) clearly yields a linear model. It should be noted that $b_j(t)$ is a *piecewise linear basis* of the univariate variable t and it is determined entirely by the locations of the function's derivative discontinuities, that is by the locations at which the linear pieces join up. Let these knots be denoted $\{x^*_j : j = 1, ..., k\}$ and suppose that $x^*_j > x^*_{j-1}$. Then for $j = 2, ..., k-1$

$$b_j(t) = \begin{cases} (t - t^*_{j-1})/(t^*_j - t^*_{j-1}), & t^*_{j-1} \leq t \leq t^*_j \\ (t^*_{i+j} - t)/(t^*_{j+1} - t^*_j), & t^*_j \leq t \leq t^*_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

So $b_j(t)$ is zero everywhere, except over the interval between the knots immediately to either side of $t_j^* \cdot b_j(t)$ increases linearly from 0 at t_{j-1}^* to 1 at t_j^*, and then decreases linearly to 0 at t_{j+1}^*. Basis functions like this, that are non zero only over some finite intervals, are said to have *compact support*.

Finally, to represent the *response compositional single covariate GAM model*. The process is similar to linear regression. It only needs to be clear that it must be estimated $D-1$ models for each coordinate. Even to make the model representation easier you can use β_0 instead of $\mathrm{A}_i\theta$

$$g(\mu_{ik}^*) = \beta_{0k}^* + f_{1k}^*(t_i) + e_{ik}^* \tag{3.22}$$

where $\mu_{ik}^* \equiv \mathbb{E}(x_{ik}^*)$ and $x_{ik}^* \sim EF(\mu_{ik,\phi})$, x_{ik}^* is the balance response variable (ilr), $EF(mu_{ik}, \phi)$ denotes an exponential family distribution with mean μ_{ik} and scale parameter, ϕ_k for each $k = 1, 2, .., D-1$. f_{1k} are different smooth functions of the covariate t_i for each k and e_{ik} are independent $N(0, \sigma^2)$ random variables.

Generalized additive model with interactions between a numeric covariate and a binary covariate

The GAM model when it is included a numeric covariate t_i and a binary covariate z_i is

$$g(\mu_i) = \beta_0 + f_1(t_i) \cdot z_i + \beta_1 \cdot z_i + e_i, y_i \sim EF(\mu_i, \phi_i) \tag{3.23}$$

Where z_i is included in the model in two ways: a) it separates t_i in two smooth functions (one for $z_i = 0$ and the other for $z_i = 1$) and b) as a fixed effect of parameter β_1, and $g(\mu_i)$, β_0 and y_i preserve the properties of equation 3.19.

In the case of compositional response it is only changed y_i for x_{ik}^* where the parameter $k = 1, 2, ..D-1$ represents each coordinate of $\mathcal{S}^{D-1}$

$$g(\mu_{ik}^*) = \beta_{0k}^* + f_{1k}^*(t_i) \cdot z_i + \beta_{1k}^* \cdot z_i + e_{ik}^*, x_{ik}^* \sim EF_k(\mu_i, \phi_i) \tag{3.24}$$

Smooth predictors and estimation methods

In principle, using the R package `mgcv` it is possible to apply smooths of any number of predictors via four types of smooth:

- `s()` is used for univariate smooths, isotropic smooths of several variables and random effects.
- `te()` is used to specify tensor product smooths constructed from any singly penalized marginal smooths usable with `s()`.

- `ti()` is used to specify tensor product interactions with the marginal smooths (and their lower order interactions) excluded, facilitating smooth ANOVA models.
- `t2()` is used to specify the alternative tensor product smooth construction which is especially useful for generalized additive mixed modelling with the `gamm4` R package

All this smooth functions need a smoothing basis as it describes equation 3.20. In practice there are a lot of basis and generally they are represented by a two letter character string. Table 3.4.1 shows the main smoothing bases considered when estimating the statistical models together with their advantages and disadvantages.

Table 3.4.1: Smoothing bases built into package `mgcv`, and a summary of their advantages and disadvantages (Wood, 2006).

`bs`	Description	Advantages	Disadvantages
`'tp'`	Thin plate regression splines (TPRS)	Can smooth w.r.t. any number of covariates. Invariant to rotation of covariate axes. Can select penalty order No 'knots' and some optimality properties.	Computationally costly for large data sets. Not invariant to covariate rescaling.
`'ts'`	TPRS with shrinkage	As TPRS, but smoothness selection can zero term completely	as TPRS
`'cr'`	cubic regression spline (CRS)	Computationally cheap. Directly interpretable parameters	Can only smooth w.r.t 1 covariate. Knot based. Doesn't have TPRS optimality.
`'cs'`	CRS with shrinkage	As CRS, but smoothness selection can zero term completely.	As CRS
`'cc'`	cyclic CRS	As CRS, but start point same as end point	As CRS
`'ps'`	P-splines	Any combination of basis and penalty order possible. Perform well in tensor products	Based on equally spaced knots. Penalties awkward to interpret. No optimality properties available.

About GAM fitting methods, the package `mgcv` uses six different ways

- `GCV.Cp` is the default method from `mgcv` R package. It uses a generalized cross validation for unknown scale parameter and Mallows' Cp/UBRE/AIC for known scale.
- `GACV.Cp` is equivalent, but using generalized approximate cross validation GACV in place of GCV.
- `REML` for REML estimation, including of unknown scale. This is a method of estimation in which estimators of parameters are derived by maximizing the residual or restricted likelihood rather than the likelihood itself Everitt and Skrondal (2010).
- `P-REML` for REML estimation, but using a Pearson estimate of the scale.
- `ML` An estimation procedure involving maximization of the like lihood or the log likelihood with respect to the parameters Everitt and Skrondal (2010).
- `P-ML` similar to ML but differs from the standard likelihood.

All this methods have different properties and have some advantages and disadvantages between them. For instance GCV has the nice property that is invariant but it is not sensitive enough to over-fit. REML is similar to use Bayesian marginal likelihood and random coefficients have Gaussian distributions which most are familiar but asymptotically REML undersmooths relative to GCV resulting in higher asymptotic mean square error.

3.4.3 Support Vector Regression - SVR

Understanding Support Vector Machines

A Support Vector Machine (SVM) can be imagined as a surface that creates a boundary between points of data plotted in multidimensional and their feature values. The goal of a SVM is to create a fat boundary called a *hyperplane*, which divides the space to create fairly homogeneous partitions on either side (Lantz, 2015).

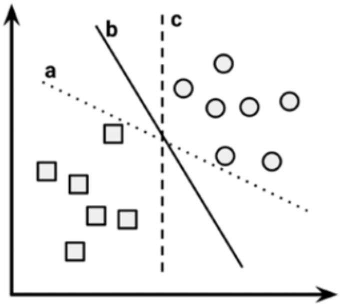


Figure 3.4.1: Behavior of the SVM algorithm (Lantz, 2015)

In two dimensions, the task of the SVM algorithm is to identify a line that separates the two classes. As shown in figure 3.4.1, there is more than one choice of dividing line between the groups of circles and squares (a,b, c). Choosing the best alternative involves to find the *Maximum Margin Hyperplane* (MMH) that creates the greatest separation between the two classes (Lantz, 2015).

The support vectors (indicated by arrows in the figure 3.4.2) are the points from each class that are the closest to the MMH; each class must have at least one support vector, but it is possible to have more than one. Using the support vectors alone, it is possible to define the MMH. This is a key feature of SVMs; the support vectors provide a very compact way to store a classification model, even if the number of features is extremely large (ibid.).

ϵ - Support Vector Regression (ϵ - SVR)

Related with SVM algorithms it was used the R package `e1071` and the used algorithm is called *epsilon-SVM*. It was introduced by Vapnik (1998). This algorithm is presented in R based on the article of Chang and Lin (2011). It considers a set of training points $\{(t_1 y_1), ...(t_\ell y_\ell)\}$, where $t_i \in R^n$ is a feature vector and $z_i \in R^1$ is the target output. Under given parameters $C > 0$ and $\epsilon > 0$,the standard form of SVR is

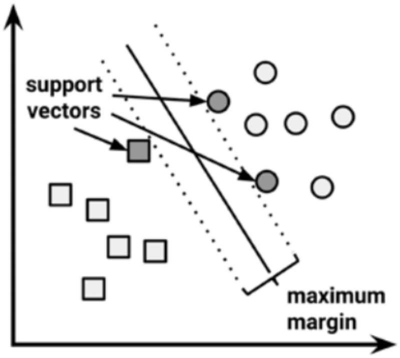


Figure 3.4.2: Interaction between support vectors and MMH (Lantz, 2015)

$$\begin{aligned}\min_{w,b,\xi,\xi^*} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{\ell}\xi_i + C\sum_{i=1}^{\ell}\xi_i^* \\ \text{subject to} \quad & w^T\phi(t_i) + b - z_i \leq \epsilon + \xi_i, \\ & z_i - w^T\phi(t_i) - b \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, ..., \ell\end{aligned} \tag{3.25}$$

The dual problem is

$$\begin{aligned}\min_{w,b,\xi,\xi^*} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{\ell}\xi_i + C\sum_{i=1}^{\ell}\xi_i^* \\ \text{subject to} \quad & w^T\phi(t_i) + b - z_i \leq \epsilon + \xi_i, \\ & z_i - w^T\phi(t_i) - b \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, ..., \ell\end{aligned} \tag{3.26}$$

Where $Q_{ij} = K(t_i, t_j) \equiv \phi(t_i)^T\phi(t_j)$

After solving problem 3.26, the approximate function is

$$y = \sum_{i=1}^{\ell}(-\alpha + \alpha^*)K(t_i, t) + b \tag{3.27}$$

where b is a constant term, (t_i, t_j) are the inner products of the support vectors as $(-\alpha + \alpha^*)$ is non-zero only when an observation is a support vector. This leads to far fewer terms in the classification algorithm and allows the use of the *kernel function* K, commonly referred to as the kernel trick (Lesmeister, 2017).

The trick in this is that the `kernel` function mathematically summarizes the transformation of the features in higher dimensions instead of creating them explicitly. This has the benefit of creating the higher dimensional, nonlinear space and decision boundary while keeping the optimization problem computationally efficient. The `kernel` functions compute the inner product in a higher dimensional space without transforming them into the higher dimensional space (Lesmeister, 2017).

The notation for popular kernels is expressed as the inner (dot) product of the features, with t_i and t_j representing vectors, gamma, and c parameters, as follows:

- linear with no transformation: $K(t_i, t_j) = t_i \cdot t_j$
- polynomial where d is equal to the degree of the polynomial: $K(t_i, t_j) = (\gamma t_i \cdot t_j + c)^d$
- radial basis function: $K(t_i, t_j) = \exp\Bigg(-\gamma\bigg|t_i - t_j\bigg|^2\Bigg)$
- sigmoid function: $K(t_i, t_j) = \tanh(\gamma t_i \cdot t_j + c)$

To conclude, equation 3.27 represent the model for a univariate response. In the case of compositional response it implies that it will be $D - 1$ regressions similarly to linear and gam models.

3.4.4 Cross validation of the statistical model indicators

Before starting with the results, it is worth mentioning that two measures will be used to contrast the models. a) Adjusted R-squared which is preferable to predicted R-square when comparing models and b) The root mean square error (RMSE) that is a standard statistical metric to measure model performance that permits compare statistical models when the R-squared cannot be computed.

Adjusted R squared

$$R^2_{adj.} = 1 - \frac{N-1}{N-k-1}[1 - R^2] \tag{3.28}$$

Root mean square error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{3.29}$$

Chapter 4

Results and discussions

This section details the main results obtained. All calculations performed were based on the software `R`. Among the considered libraries highlights `compositions` and `robCompositions` that are related with the CoDa methodology. The Generalized additive models (GAM) were applied with the package `mgcv` and the Support vector regressions with the package `e1071`. For more details, the annex C describes the script used.

About the linear models implementation it is important to mention that was used the R command `poly`. As R Core Team (2016) explains this command returns or evaluates orthogonal polynomials of degree 1 to degree over the specified set of points t: these are all orthogonal to the constant polynomial of degree 0. On other hand, it is worth noting that this polynomial feature also admits interactions with the binary variable z_i that was described on the methodology.

4.1 Electricity access modelling with a single covariate

4.1.1 Electricity access models for the under-dispersed data of Sudan

As shown in figure 3.1.1 Sudan is located on South Saharan region. In this country there are around 21 million people without electricity access that represents the 1.89% of the electricity problem over the world. The first compositional statistical model developed was done over this country because of the properties of his series that is characterized for being under-dispersed in comparison with the others modelled countries. This fact makes Sudan series easy to model and consequently its goodness of fit indicators are very high.

Tables 4.1.1 and 4.1.3 show the goodness of fit of Sudan by using a model based only in year variable (time) to explain electricity access. Comparing between LM and GAM with the R-squared approach it is appreciated that both models have results over 0.99. By means of RMSE it is possible realize a cross validation between LM, GAM and SVM where despite the good approximations of this three models, GAM stands out as the model that best approximates the temporary trends of Sudan.

Table 4.1.1: Adjusted R-squared for the Sudan single covariate model.

model	Balance1	Balance2	Balance3
LM	0.9988	0.9984	0.9923
GAM	0.9999	0.9999	0.9999

Table 4.1.2: RMSE for Sudan single covariate model.

model	Balance1	Balance2	Balance3
LM	0.00958	0.01157	0.01551
GAM	0.00011	0.00152	0.00005
SVM	0.00064	0.00297	0.00082

Figure 4.1.1 shows a graphical representation about Sudan modelling. The three plots on top show the fitted and real values for the balances (model in coordinates) and are graphical representations of previous tables contents. The right bottom represents the original values in $\mathcal{S}^4$ and the back-transformations of each balance model where it is checked that Sudan electricity access data is very homogeneous. It justifies the high r-squared obtained for this case. At last, bottom left plot shows with a bar plot the good fit of GAM opposite to LM and SVM for the raw data $\mathcal{S}^4$,i.e., for the electricity access indicator.

4.1.2 Electricity access models for the dispersed data of Bangladesh

Bangladesh is a country located in the South of Asia. The population without access to electricity was around 68 million people in 2010 that represents the 6% of the electricity problem over the world. The time series of this country is more complex than Sudan series since it has more dispersion over the time and consequently the predictive capacity of the different methods is more observable between them.

The contrast between LM and GAM is displayed in table 4.1.3. It shows that coordinate-based models have a good goodness of fit for balance 1 and 2 but the r-squared of balance 3 is around 0.5, i.e., this model does not fit very well. Table B.2.3 reaffirm the problem with balance 3 and in general the RMSE change from a value close to zero in the Sudan case to a value between 0.18 and 0.80 in the more extreme case.

Table 4.1.3: Adjusted R-squared for Bangladesh single covariate model.

model	Balance1	Balance2	Balance3
LM	0.9585	0.9428	0.5769
GAM	0.9609	0.9439	0.5430

Through figure 4.1.2 it is displayed the fitted values for LM, GAM and SVM. Plots on top evince the modelling patterns of these models and show how Balance 3 original data is more dispersed than the others. The right bottom

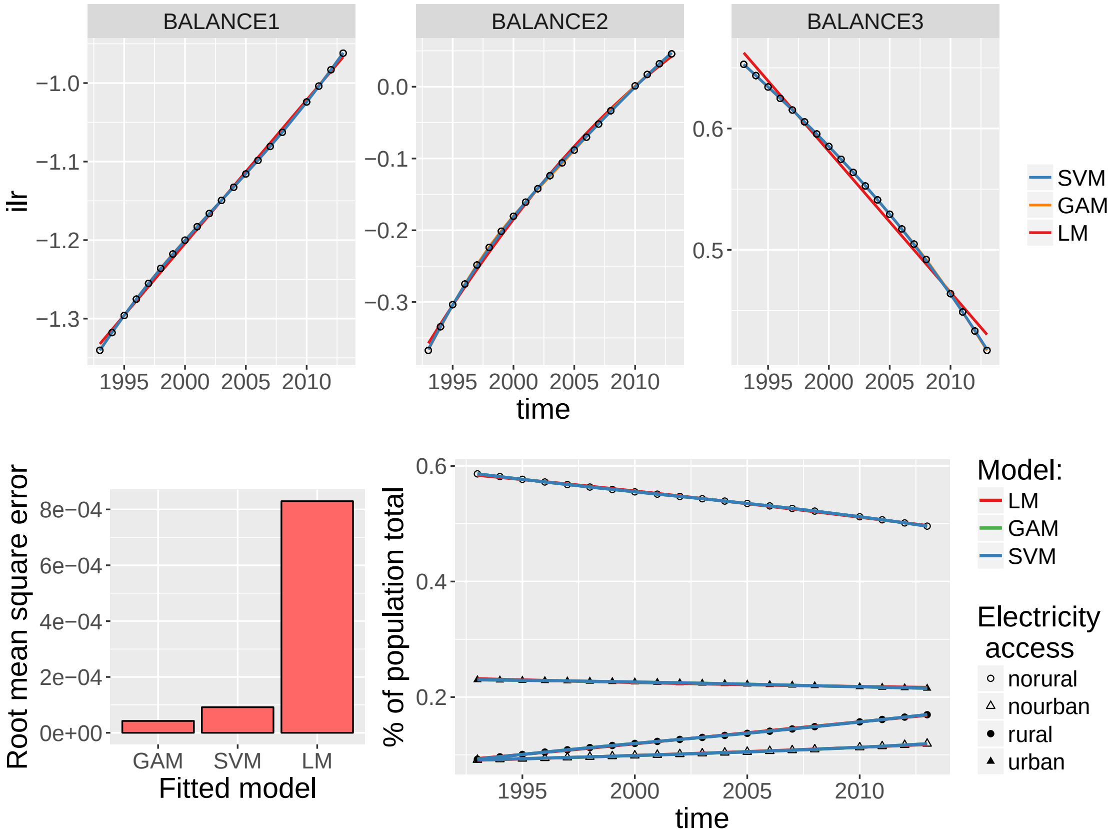


Figure 4.1.1: LM, GAM and SVM CoDa statistical models for Sudan.

Table 4.1.4: RMSE for Bangladesh single covariate model.

model	Balance1	Balance2	Balance3
LM	0.1897	0.2116	0.7803
GAM	0.1790	0.2145	0.7442
SVM	0.1865	0.2166	0.8007

plot presents little differences at the moment of back-transform the fitted models that can be appreciated in left bottom RMSE plot in a best way.

Making a comparison between the cases of Sudan and Bangladesh, it is observed that in Sudan the three balances are very aligned to a line and that line is almost the same for LM, GAM or SVM. On the other hand, in the case of Bangladesh, the first and second Balances are little dispersed while there is a greater dispersion in the third Balance, which is where there is more problems in the goodness of fit. It also highlights that when estimating the trend in Bangladesh, the three models behave differently in the third balance, the linear model estimates a straight line, while GAM and SVM estimate two very different convex curves that reflects the predictive capacity of each model.

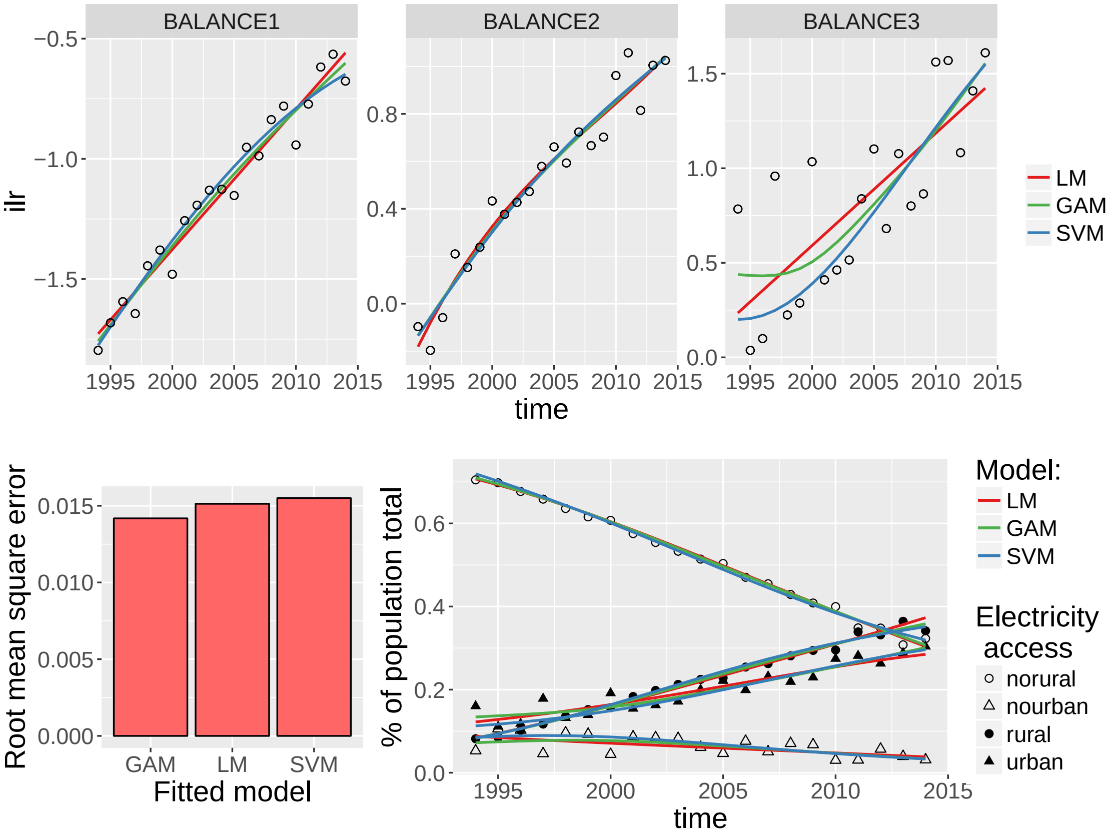


Figure 4.1.2: LM, GAM and SVM CoDa single covariate statistical models for Bangladesh.

4.2 Compositional model with interactions between a numeric covariate and a binary covariate

4.2.1 Dataset segmentation through binary covariate z_i

In the previous section it was shown that modelling electricity access trends for cases where there is dispersion in the data around time results are less robust, which translates into lower adjusted r-squares and higher RMSE. Fortunately, these estimates can be improved by introducing a control variable that helps to make a better estimate of the electricity access trends. This opportunity for improvement was detected when validating the response variable with the rest of the variables for access to electricity that are in the World Bank database, especially with the percentage of access to electricity.

Thus, it was introduced a binary covariate z_i (see equation 3.1) that validate the existing harmony between the World-Bank variable `Access to electricity (% of population)` and the modeled response variable. It permits to differentiate two subseries inside the indicator as it can be seen in figure 4.2.1 where *different* level means that there is not harmony with `Access to electricity (% of population)` and *same* level that if there is. An special feature of this series is that *same* level has under-dispersed data as Sudan but it is not the case for *different* level.

The figure 4.2.1 evidently shows the existence of two sub-series within each Balance. In figure 4.1.2 the goodness of fit was good for the first and second balance due to the fact that the sub-series were very close and parallel to each other. In the case of the third balance the two sub-series are more separated from each other, if two curves are estimated, they would not be so parallel, noting also that the previous model estimated for Bangladesh traced a curve in the middle of them, which justifies that the estimate is not so good.

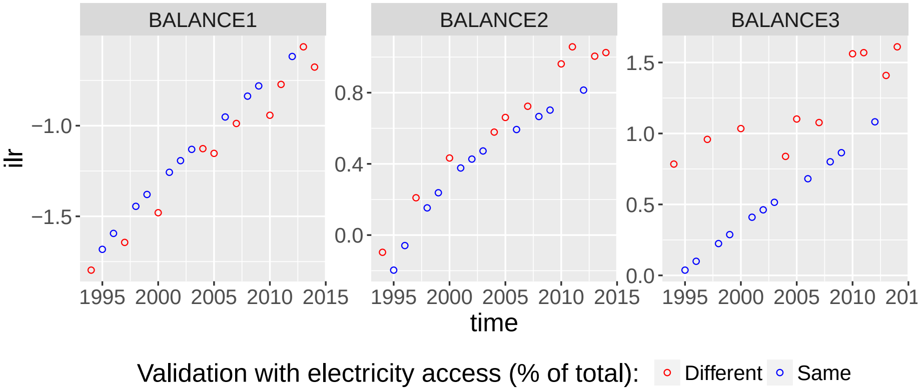


Figure 4.2.1: Electricity access balance segmented by z_i for Bangladesh.

4.2.2 Interaction model for Bangladesh

The advantages of including the binary variable z_i are translated in three ways: a) help to differentiate between two subseries, b) permit to control fixed effects between the two levels and c) facilitate the control of related slopes effects of each subseries when they are not parallels. Table 4.2.1 reflects this improvements through the increments of the adjusted r-squared, specially for balance 3 which is more dispersed than the others and table 4.2.2 with the decrements of RMSE.

Table 4.2.1: Adj. R-squared for the single covariate model and the model with interactions in Bangladesh.

Model	Numeric Covariate			Model with interactions		
	Balance 1	Balance 2	Balance 3	Balance 1	Balance 2	Balance 3
LM	0.9585	0.9428	0.5769	0.9844	0.9822	0.9444
GAM	0.9609	0.9439	0.5430	0.9833	0.9965	0.9727

Figure 4.2.2 shows the interaction models for Bangladesh. The right plot on top displays the fit capacity of LM, GAM and SVM for the most dispersed balance and it is interesting to show for *different level* that in the case of LM a line was fitted, while in SVM the curve fitted is more convex and in GAM case the *cubic spline basis* (`cs`) and the `GCV.Cp` permits linearize a cubic regression with shrinkage that pass cross all points.

Table 4.2.2: RMSE for the single covariate model and the model with interactions in Bangladesh.

Model	Numeric Covariate			Model with interactions		
	Balance 1	Balance 2	Balance 3	Balance 1	Balance 2	Balance 3
LM	0.1897	0.2116	0.78030	0.1132	0.1110	0.2677
GAM	0.1790	0.2145	0.7442	0.1069	0.0403	0.1509
SVM	0.1865	0.2166	0.8007	0.1094	0.1026	0.2562

Another remarkable aspect in the right plot of figure 4.2.2 is that while LM fitted two linear curves and SVM two convex curves, GAM represent different level with like a cubic shape and the same level with a straight line product of using two smoothing functions for each level. By other hand, the bottom right plot displays the back-transform model for electricity access where can be appreciated the temporary interpolations in almost every points. The left bottom plot shows the RMSE for real and predicted values in $\mathcal{S}^4$ and the three models have values around zero with a small superiority of GAM over the other models.

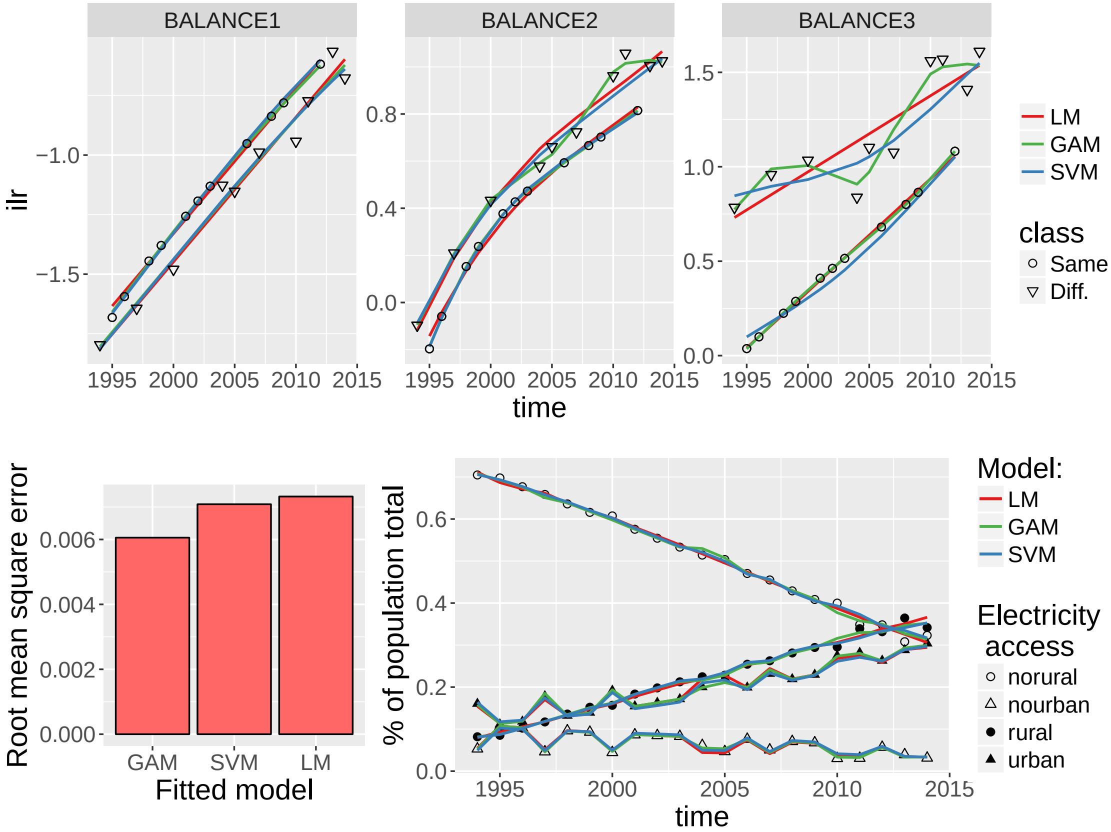


Figure 4.2.2: Interaction models for Bangladesh.

Figure 4.2.3 displays graphically the relevance of including the binary variable z_i to have a more robust model for electricity access temporary trend representation. Using the single covariate *time* draws a line for the middle of dispersed series parts whilst the interaction model facilitates get at the extreme points.

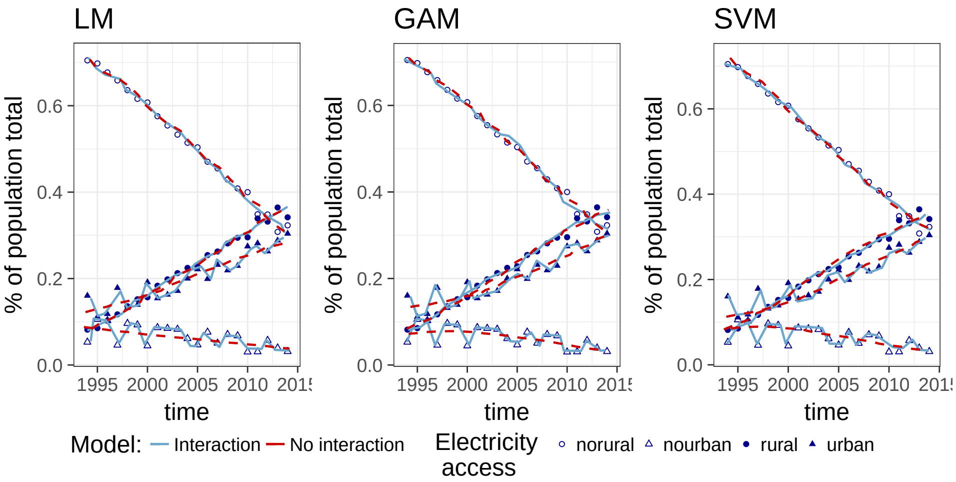


Figure 4.2.3: Comparison between single covariate model versus interaction model for Bangladesh.

4.2.3 Temporary trend models for another countries

It is worth mentioning that this methodology is replicable to any country of World-Bank electricity access database. Tables 4.2.3 and 4.2.4 exhibits comparisons between LM, GAM and SVM methods for the different CoDa balances and show that including the binary variable z_i in the model plus CoDa give good fits for India, Kenya and Nigeria.

By means of the table 4.2.3 analysis it is observed that in the case of the application of the linear model (LM) for India the inclusion of the covariate z_i was not significant. However, z_i did not have the same effect in GLM, where it is observed that the model improves with the inclusion of this binary covariate. In the case of Kenya, both the linear model and GAM gave adjusted squares above 0.9. While in Nigeria there were difficulties with the third balance.

In general, it can be said that the third balance presents more dispersion over the time when there are two sub-series. It can be justified by the fact that as the balances are created with the SBP method where one variable is excluded in each step and only two of the four variables are related at the end. This can be a cause for the hight dispersion over the third balance. In some cases like the one in Kenya or Bangladesh the dispersion is easily controllable, while contrasting this results with the R-squares of Nigeria the adjustments could be seem bad. However, in the literature it is considered a good R-squared when it is is greater than 0.70.

Table 4.2.4 allows making comparisons between LM, GAM and SVM, where it can be seen that in most of the balances, with the exception of India, most estimates that consider the covariable z_i give RMSE below 0.17 which reflects the significant contribution of this variable.

Table 4.2.3: Adj. R-squared for the single covariate model and the model with interactions for India, Kenya and Nigeria.

Model	Numeric Covariate			Model with interactions		
	Balance 1	Balance 2	Balance 3	Balance 1	Balance 2	Balance 3
LM						
India	0.8263	0.8536	0.8243	0.8263	0.8536	0.8243
Kenya	0.9383	0.8715	0.7369	0.9934	0.9667	0.9834
Nigeria	0.9780	0.5294	0.2859	0.9911	0.8009	0.8658
GAM						
India	0.8323	0.8561	0.8486	0.8833	0.8653	0.8829
Kenya	0.9400	0.8605	0.6686	0.9849	0.9625	0.9999
Nigeria	0.9715	0.5345	0.2751	0.9999	0.9288	0.8597

Table 4.2.4: RMSE for the single covariate model and the model with interactions in India, Kenya and Nigeria.

Model	Numeric Covariate			Model with interactions		
	Balance 1	Balance 2	Balance 3	Balance 1	Balance 2	Balance 3
LM						
India	0.1987	0.2910	0.4774	0.1987	0.2910	0.4774
Kenya	0.1193	0.2618	0.5537	0.0339	0.1155	0.1321
Nigeria	0.1092	0.2391	0.4101	0.0660	0.1518	0.1737
GAM						
India	0.1876	0.2869	0.4319	0.1513	0.2657	0.3591
Kenya	0.1142	0.2587	0.5382	0.0513	0.1184	0.0007
Nigeria	0.1025	0.2353	0.4127	0.0018	0.0839	0.1687
SVM						
India	0.1979	0.2840	0.3979	0.1515	0.2606	0.3232
Kenya	0.1384	0.2637	0.6636	0.0429	0.1360	0.1600
Nigeria	0.1196	0.2506	0.4677	0.0723	0.0804	0.1611

4.3 Prediction outside the calibration range

So far, models have been presented within the calibration range but ¿ what happens outside it?. This section seeks to give a brief introduction to this aspect. It is worth mentioning that for the purposes of this section, the last six years were excluded and the RMSE was calculated based on raw data, that is, in $\mathcal{S}^4$.

Figure 4.3.1 displays the summary five RMSE plots developed for each analyzed country. These graphs was ordered from lowest to highest. In some cases like Bangladesh or India results are very close between SVM and GAM. But generally, it was found that SVM presents *little* differences compared to GAM and LM and it is used the expression *little* because all models have good fits around the zero.

However, the good predictions outside the calibration range it is not a surprise due to it is well recognize their properties for many authors that use it. For instance, Ekonomou (2010) points out that compared to most learning techniques, SVM leads to better performance in prediction patterns.

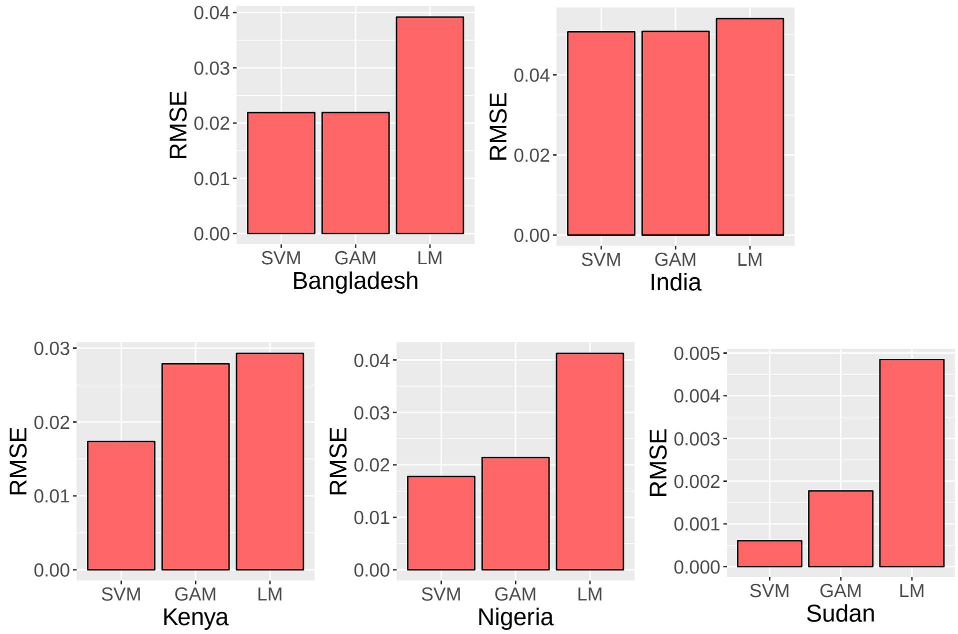


Figure 4.3.1: Comparison between LM, GAM and SVM outside the calibration rank through RMSE.

Table 4.3.1 displays a summary of Adjusted R-squared for different *balance models*, excluding the last values after 2008 to make the above plot.

Table 4.3.1: Adjusted R-squared considering the data until 2008 year.

Model/Country	BALANCE1	BALANCE2	BALANCE3
LM			
Bangladesh	0.9912	0.9959	0.9633
India	0.9385	0.8755	0.8651
Kenya	0.9999	0.9998	0.9977
Nigeria	0.9908	0.9571	0.9885
Sudan	0.9978	0.9987	0.9971
GAM			
Bangladesh	0.9930	0.9913	0.9536
India	0.9746	0.8709	0.8394
Kenya	0.9999	0.9996	0.9999
Nigeria	0.9999	0.9791	0.9989
Sudan	0.9999	0.9999	0.9999

Table 4.3.2 displays a summary through RMSE for different *balance models* where it is possible to make comparison between countries.

Table 4.3.2: RMSE considering the data until 2008 year.

Model/Country	BALANCE1	BALANCE2	BALANCE3
LM			
Bangladesh	0.0546	0.0325	0.1314
India	0.0648	0.1401	0.2079
Kenya	0.0022	0.0048	0.0339
Nigeria	0.0426	0.0652	0.0389
Sudan	0.0088	0.0076	0.0059
GAM			
Bangladesh	0.0387	0.0445	0.0997
India	0.0337	0.1337	0.2010
Kenya	0.0017	0.0071	0.0003
Nigeria	0.0010	0.0413	0.0103
Sudan	0.0000	0.0010	0.0000
SVM			
Bangladesh	0.0502	0.0245	0.1384
India	0.0378	0.1496	0.2065
Kenya	0.0032	0.0548	0.0994
Nigeria	0.0468	0.0180	0.0097
Sudan	0.0002	0.0022	0.0002

To understand the modeled process carry out in this section it is going to be presented the case of Kenya for GAM and SVM where in advance SVM responded better than GAM based on the RMSE indicator.

4.3.1 Kenya GAM prediction outside the calibration range

Figure 4.3.2 displays Kenya prediction for the last six values considering a GAM interaction model. The plots on top represent the prediction for the balance where it is remarkable that GAM predicts very well for the subseries associated with the level *same* of the binary variable z_i and not so good for the subseries with class *different*.

After fit the balance models, applying the ilr^{-1} it is obtained the bottom plot and it is observable that this model predicts almost all the electricity access points as can be seen with the painted red line and the non-predicted values in the *balances* are translated into predictions above or below real values.

As an interesting detail it is appreciated that predictions within the calibration range is excellent. Both the values for class *same* and *different* are very well predicted, which was what was shown in previous result sections.

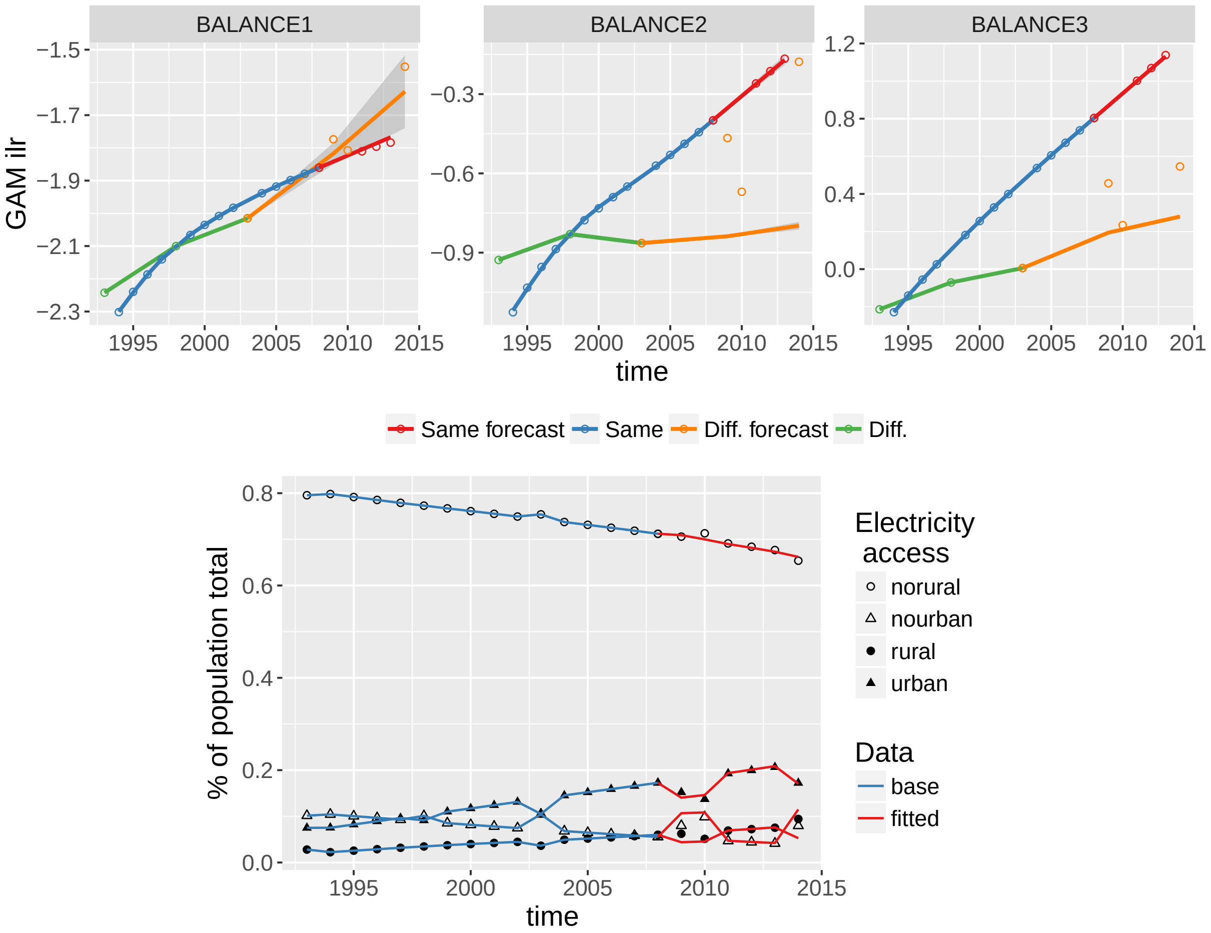


Figure 4.3.2: Kenya GAM prediction outside the calibration range.

4.3.2 Kenya SVM prediction outside the calibration range

Observing figure 4.3.3 can be appreciated that SVM predictions do not fit electricity access so exactly as GAM for subseries of class *same* but curves passes near the original values and the same for the subseries of class *different* but passing near the original values of two subseries give a slight advantage over GAM (for the Kenya modelling case).

Comparing the predictions within the calibration range with GAM, evidently GAM fits better than SVM. However, outside the calibration range, once the back-transformation is made the bottom plot shows that SVM obtained a better RMSE than GAM which is because it fitted better for the class/category *different*. This analysis leaves in evidence that SVM in average gives the best estimations but as we are modelling two subseries it should be considered that subseries with class "*different*" will give better predictions with GAM.

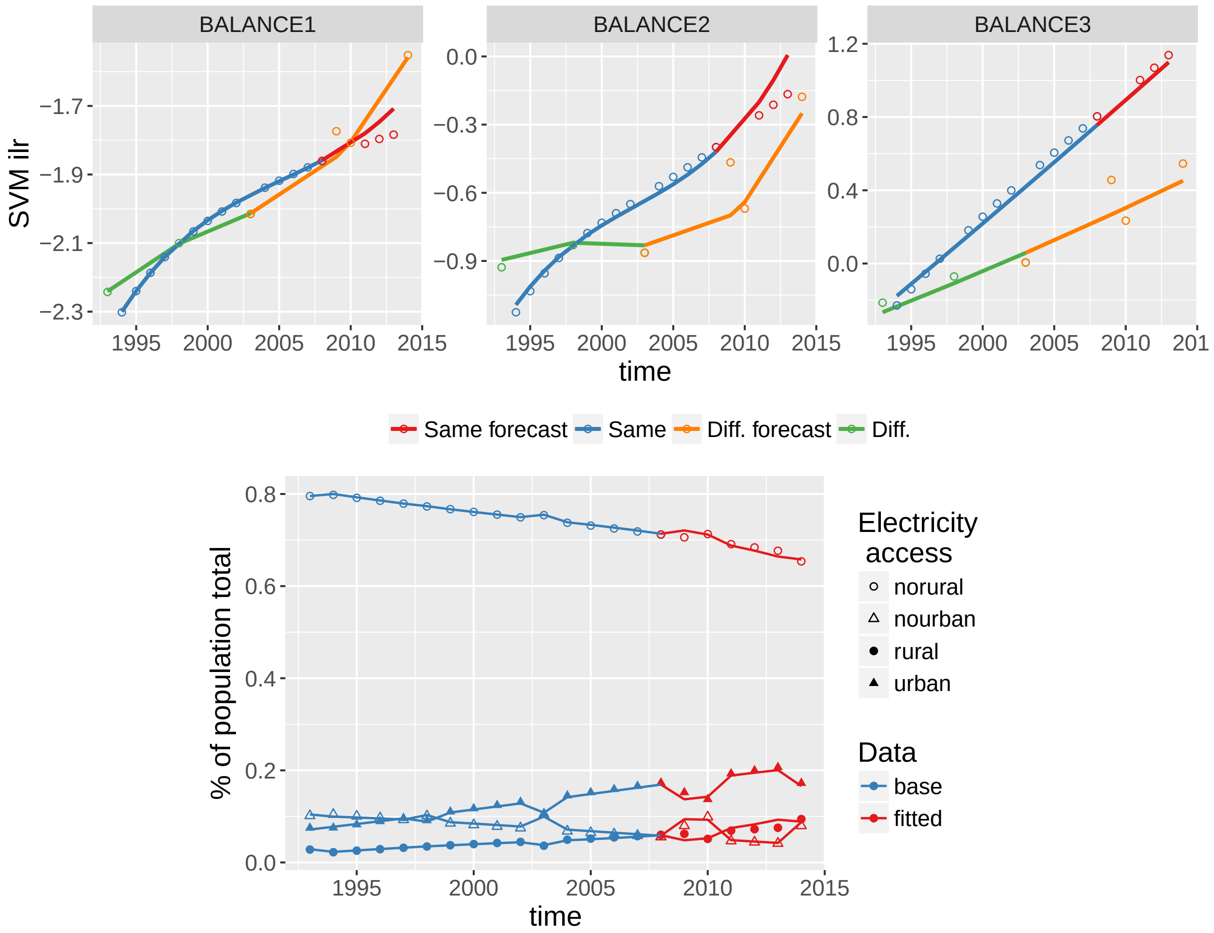


Figure 4.3.3: Kenya SVM prediction outside the calibration range.

4.4 Consequences of using standard statistical techniques over compositional data in raw form

Figure 4.4.1 shows the main problem of working in raw form for a compositional variable response. For this plot it was fitted unidimensional models for each part of electricity access indicator (urban, rural, nourban, norural). Also, it was taken into account predictions inside and outside the calibration range (for the last six observations). If a CoDa model would be done all the parts have unit sum (blue line), but as it was not considered, the plot shows the error across time for the metric.

Inside the calibration range there is less error than outside the calibration range independently of the fitted model. This is because when estimating a model if the model is good, the predictions will be around the original value and obviously the sum will be near to one. Nevertheless, outside the calibration range each forecasting does not have information about what is happening in the other parts of the electricity indicator. Consequently, predictions outside the calibration rank leads to a high error over the unit-sum constraint.

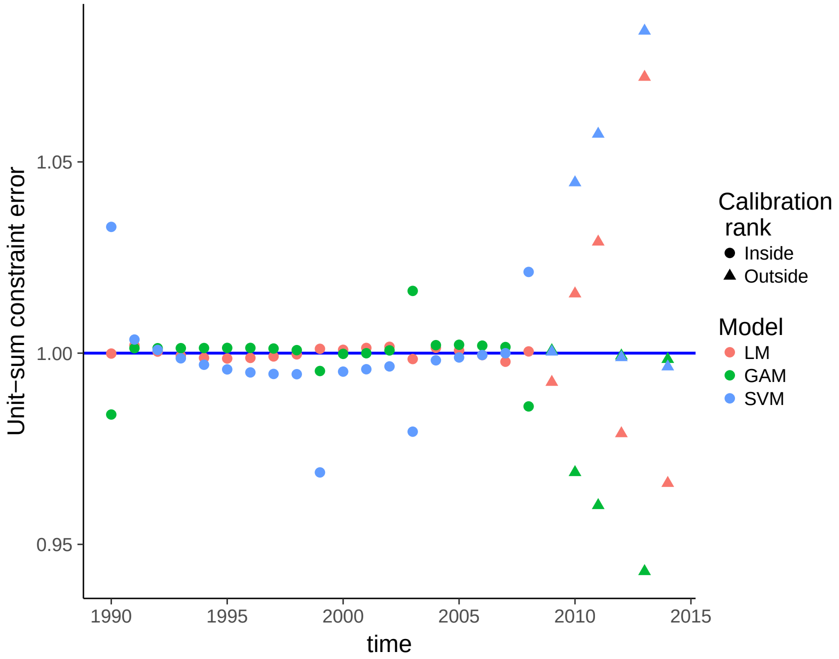


Figure 4.4.1: Cross validation unit-sum constraint error for Nigeria

By other hand, Figure 4.4.2 displays a cross validation between a model using a Coda transformation and a model in raw form (i.e., without using any transformation) inside the calibration rank. In principle, if this graph is judged without considering the previous information provided in the figure 4.4.1, it can be affirmed that the models with a raw data form are better than the compositional models. But, this view point is an spurious appreciation because it is not considered the distribution of the unit-sum constraint error across the different indicators. Adding that error to Raw models it is a way of reaching a less spurious conclusion.

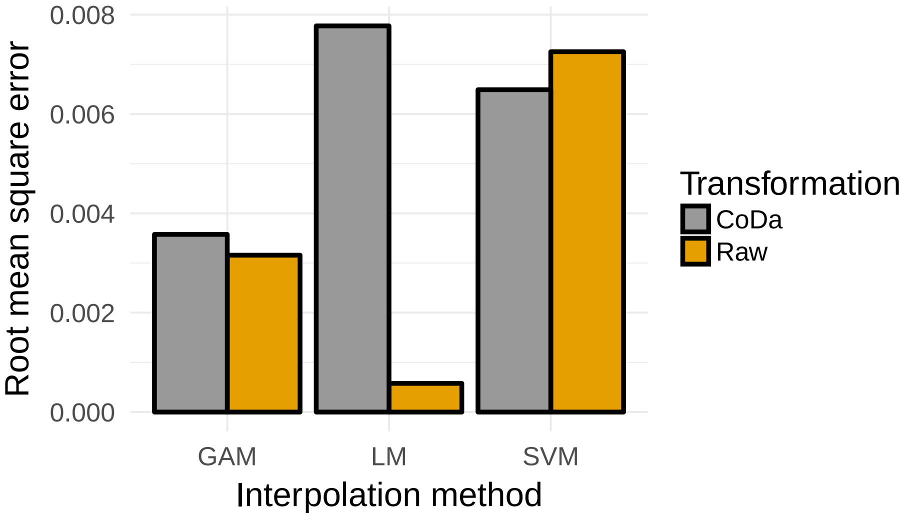


Figure 4.4.2: Cross validation between CoDa and Raw model inside the calibration rank for Nigeria using the RMSE

Outside the calibration range the problem is more serious because fitted models for majority of countries shows information like the presented in figure 4.4.3, i.e., the estimates in a raw form are lower than those of CoDa (see appendix A) without adding the unit-sum constraint error which, also, is larger than within the calibration range analysis done.

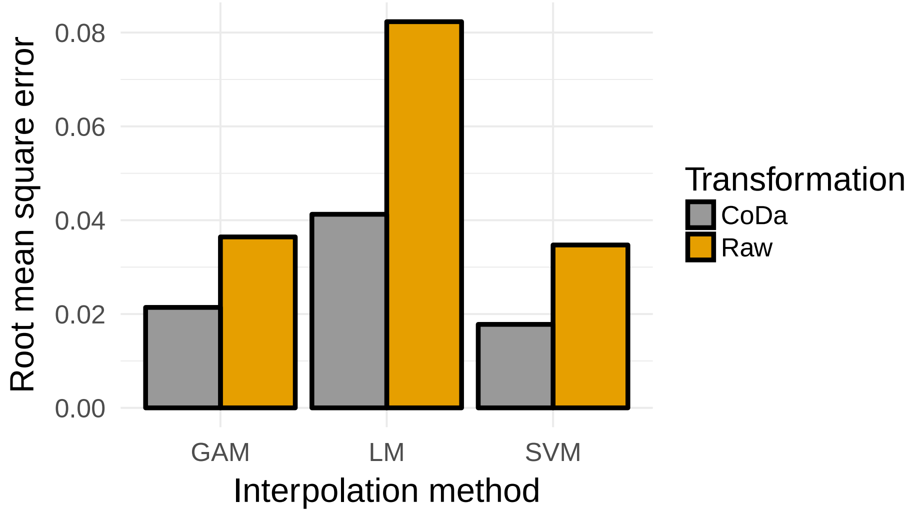


Figure 4.4.3: Cross validation between CoDa and Raw model outside the calibration rank for Nigeria using the RMSE

The next table displays a summary of estimations inside the calibration range for considering the raw data:

Table 4.4.1: RMSE on Estimated models with raw data and with time interval until 2008.

Model/Country	urban	rural	nourban	norural
LM				
Bangladesh	0.00403	0.00459	0.00361	0.00524
India	0.00343	0.01079	0.00327	0.01119
Kenya	0.00028	0.00051	0.00066	0.00033
Nigeria	0.00040	0.00067	0.00069	0.00050
Sudan	0.00003	0.00037	0.00034	0.00036
GAM				
Bangladesh	0.00262	0.00199	0.00268	0.00209
India	0.00263	0.00843	0.00320	0.00752
Kenya	0.00001	0.00032	0.00002	0.00001
Nigeria	0.00003	0.00291	0.00021	0.00560
Sudan	0.00002	0.00000	0.00000	0.00000
SVM				
Bangladesh	0.00389	0.00254	0.00322	0.00272
India	0.00328	0.01120	0.00292	0.01124
Kenya	0.00286	0.00279	0.00256	0.00228
Nigeria	0.00070	0.01412	0.00091	0.00311
Sudan	0.00002	0.00009	0.00001	0.00003

4.5 Forecasting of access to electricity by 2030

Once it has been established the outside calibration model it is possible to make predictions for example to the 2030 agenda. This section is based on India and Nigeria countries and uses the models GAM and SVM.

4.5.1 Forecasting of electricity access in India

India represents around the 25% electricity access problem in the world. Therefore it is interesting to display what happens with this country if the actual conditions remains. Figure 4.5.1 represents both scenarios of binary variable z_i, the scenario for data under-dispersed (class=same) and for dispersed data scenario (class=different). The top plots displays the predictions based on balances. This information it is not interpretable but it gives account of what happens in coordinates.

Through the SVM, the left bottom plot displays a positive scenario for rural population in detriment of the urban part which seems don't have sense with the historic data where both urban and rural population grew at the same time. It is remarkable that in the medium term (2014-2020) both models (GAM and SVM) have similar results.

By means of the right bottom plot for SVM forecasting the trend remains in favor of rural population with difference that this scenario increases faster than the above scenario to the point that by 2030 almost every rural people have electricity in detriment of urban population. Thus, SVM 2030 scenarios are not very realistic.

By other hand, the GAM forecasting displayed in the left bottom plot scenario seems more realistic than SVM. The main reason is that urban population by 2030 does not increase it remains almost constant, which comparing with urban population without electricity access has sense and it means that by 2030 around 100% of urban population has electricity and there is gonna be a 8% of people without electricity access that belongs to rural areas, i.e., this scenario proposes that by 2030 around 90% of India population will have electricity.

The GAM bottom right scenario is more optimistic than the left bottom scenario but with few percentage. Therefore, generalized additive model is a good tool to follow up temporary trends in India.

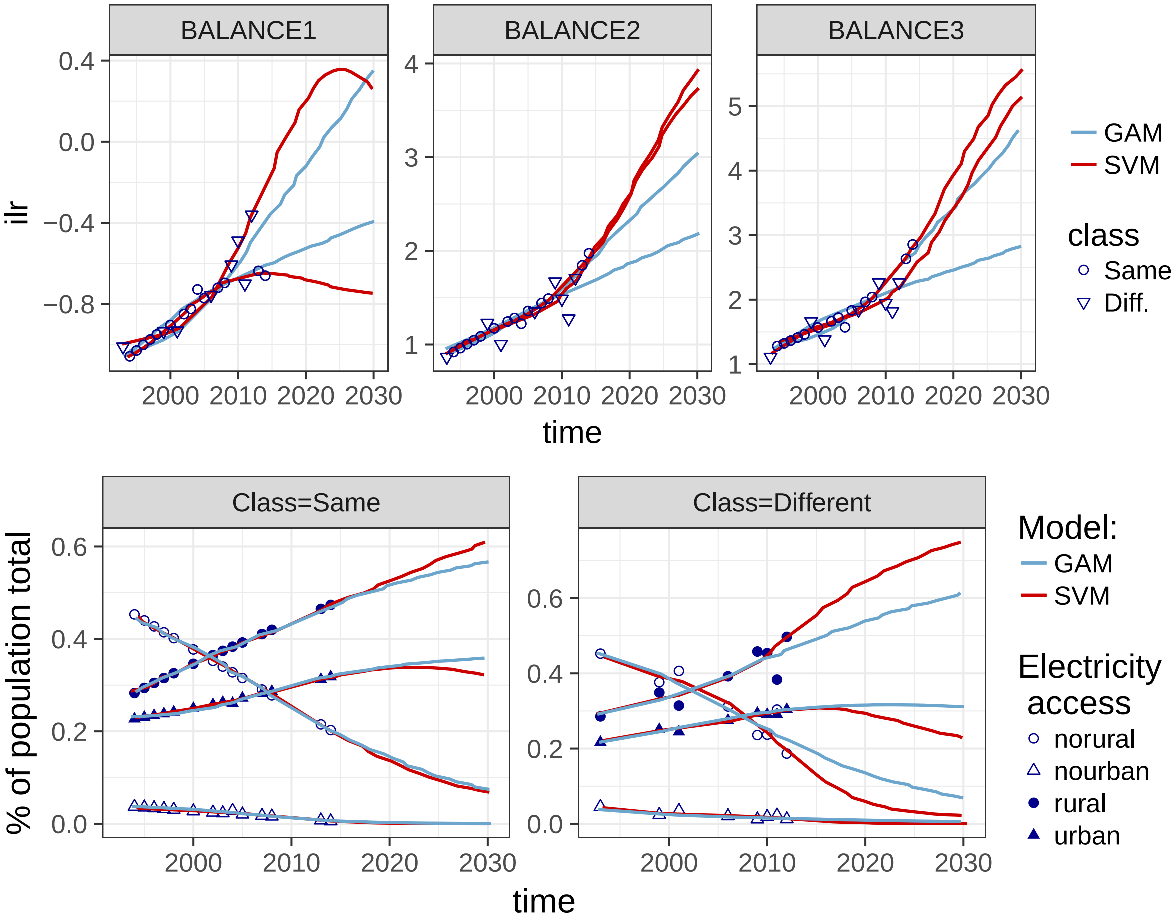


Figure 4.5.1: Forecasting of electricity access in India by 2030.

4.5.2 Forecasting of electricity access in Nigeria

Figure 4.5.2 is about 2030 trend model for Nigeria. The three balances exposed on top plots are an indication of the complexity of this data where the two subseries have patterns totally different across the time and maybe for this aspect GAM and SVM models are very different between them.

The left bottom plot shows similar results for both GAM and SVM and and you can say that both predictions are good. However, in the bottom right plot the GAM estimations has a big slope which give rise to have a bad perception of this prediction. Therefore, to have a reasonable analysis is preferable to use SVM over GAM. As a positive aspect, it is noteworthy the role of CoDa in keeping the unit-sum constraint for GAM.

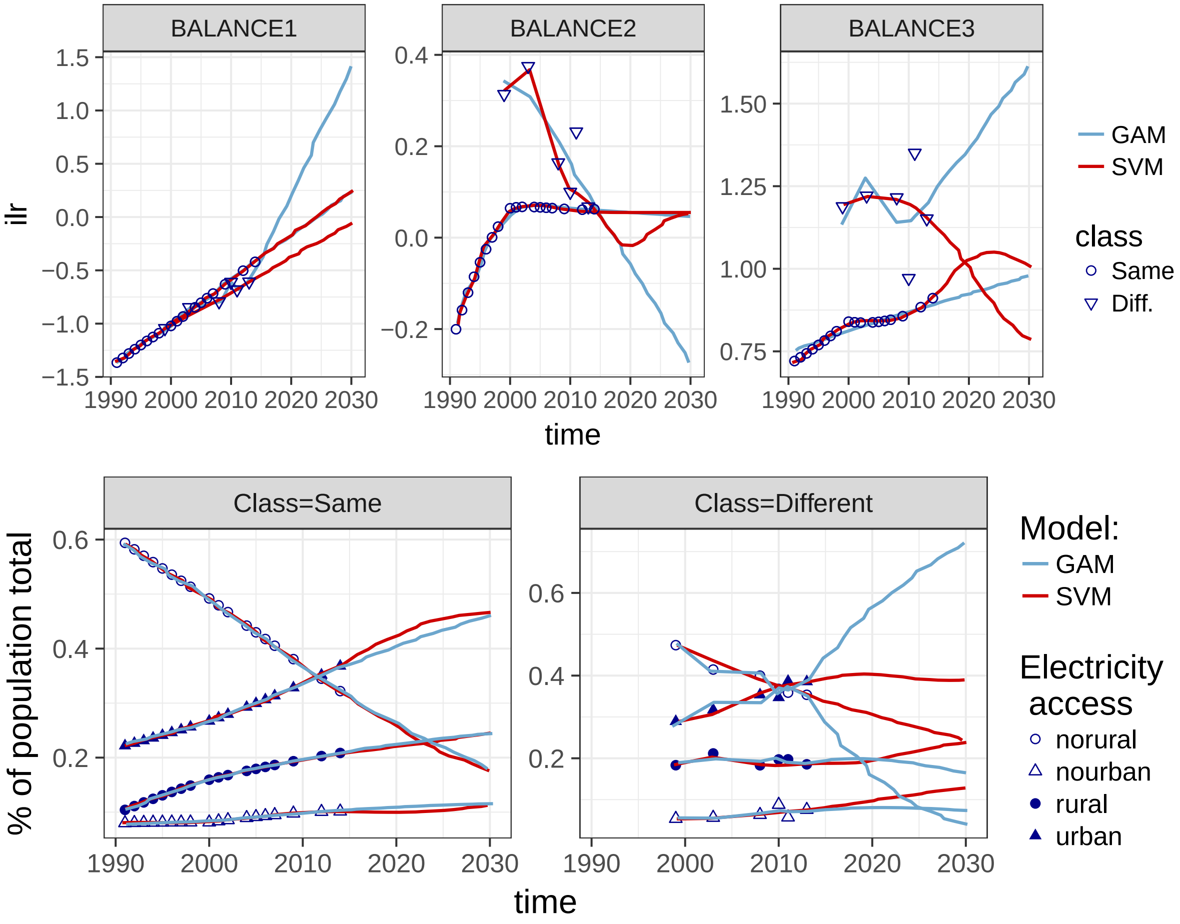


Figure 4.5.2: Forecasting of electricity access in Nigeria by 2030.

Chapter 5

Conclusion

Based on what is detailed in the section 4.4 it is confirmed that in order to model tendencies of SDG7 compositional indicators it is advisable to use LM-ilr, or GAM-ilr, 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 detailed analysis of the electricity access World-Bank indicator confirms that the data series include, in some cases, two subseries, characterized by different temporal evolution coefficients. This differentiation was possible through the validation with other World-Bank indicator `Electricity access (% of population total)` and the two parts of the proposed indicator related to electricity access in the urban and rural sectors.

This event very possibly responds to the fact that in the process of data collection there were gaps in available data and a simple modelling approach was adopted to fill in the missing data points as it is detailed in the research methodology chapter.

By other hand the relation between electricity access in rural and urban sectors it is handled individually as `Access to electricity, urban (% of urban population)` and `Access to electricity, rural (% of rural population).` Managing a multidimensional indicator in this way is very difficult, and neglects the sum of the whole. Therefore, the use of CoDa is recommended for an improvement in the management of this type of indicators. Currently, CoDa has made great progress in imputation techniques and handling of missings, if this recommendation is implemented it would be a great support for this indicator.

Based on the RMSE within the calibration range the GAM model provides a better fit. It is worth mentioning that this differences are minimal. It is worth mentioning that these differences are minimal and depending on the accuracy desired or the approach that you want to give the analysis. A linear regression model with CoDa can be an excellent option since it is very illustrative and most people are related to this statistical model.

Based on the RMSE outside the calibration range for the predictions of last six observations after the year 2008 and with observation of the predictions to 2030, GAM or SVM can be considered to make predictions of the temporary trend of electricity access. The conclusion about SVM is seen in the figure 4.3.1 where SVM is the model that best predicts. However, these differences in some cases are almost nil and as displayed figure 4.5.1 GAM can lead to very stable predictions.

Additionally, considering the two subseries identified in most cases of the subseries where there is harmony with the indicator `Electricity access (% of total population)`,i.e `class=same`, a better estimate with GAM is achieved which is very interesting in case you are more interested in this subseries.

Finally, the present work took as a unit of analysis five representative countries in the problem electricity access, but this analysis is easily replicable to the 212 countries that are in the World-Bank database the only you need is to chose an orthonormal base, transform your compositional data and run your model.

Bibliography

Aitchison, J. (1986). The statistical analysis of compositional data. *Chapman and Hall London*. 9, 17

Barclay, H., Dattler, R., Lau, K., Abdelrhim, S., Marshall, A., and Feeney, L. (2015). Sustainable Development Goals. *International Planned Parenthood Federation*. 2

Bazilian, M., Nussbaumer, P., Cabraal, A., Centurelli, R., Detchon, R., Gielen, D., Rogner, H., Howells, M., McMahon, H., Modi, V., et al. (2010a). Measuring energy access: Supporting a global target. *Earth Institute, Columbia University, New York*. 8

Bazilian, M., Nussbaumer, P., Rogner, H.-H., Brew-Hammond, A., Foster, V., Pachauri, S., Williams, E., Howells, M., Niyongabo, P., Musaba, L., et al. (2012). Energy access scenarios to 2030 for the power sector in sub-Saharan Africa. *Utilities Policy*, 20(1):1–16. 8

Bazilian, M., Sagar, A., Detchon, R., and Yumkella, K. (2010b). More heat and light. *Energy Policy*, 38(10):5409–5412. 6

Bhattacharyya, S. C. (2006). Energy access problem of the poor in India: Is rural electrification a remedy? *Energy policy*, 34(18):3387–3397. 3, 6, 7

Bhattacharyya, S. C. and Ohiare, S. (2012). The Chinese electricity access model for rural electrification: Approach, experience and lessons for others. *Energy Policy*, 49:676–687. 6, 7

Birol, F. et al. (2013). World energy outlook. *Paris: International Energy Agency*, 23(4):329. 6

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27. 27

Chaurey, A., Ranganathan, M., and Mohanty, P. (2004). Electricity access for geographically disadvantaged rural communities—technology and policy insights. *Energy policy*, 32(15):1693–1705. 3

Commission, J. R. C.-E. et al. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. 8

Cozzi, L. et al. (2017). World energy outlook special report. *France International Energy Agency (IEA)*. 3, 6, 7

Doll, C. N. and Pachauri, S. (2010). Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. *Energy Policy*, 38(10):5661–5670. 7

Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. *Energy*, 35(2):512–517. 11, 37

Everitt, B. and Skrondal, A. (2010). *The Cambridge dictionary of statistics*. Cambridge University Press Cambridge. 26

Flood, R., Bloemsma, M., Weltje, G., Barr, I., O'Rourke, S., Turner, J., and Orford, J. (2016). Compositional data analysis of Holocene sediments from the West Bengal Sundarbans, India: Geochemical proxies for grain-size variability in a delta environment. *Applied Geochemistry*, 75:222–235. 9

Groh, S., Pachauri, S., and Narasimha, R. (2016). What are we measuring? An empirical analysis of household electricity access metrics in rural Bangladesh. *Energy for Sustainable Development*, 30:21–31. 6, 8, 10

Hailu, Y. G. (2012). Measuring and monitoring energy access: Decision-support tools for policymakers in Africa. *Energy Policy*, 47:56–63. 8, 9

Kanagawa, M. and Nakata, T. (2008). Assessment of access to electricity and the socio-economic impacts in rural areas of developing countries. *Energy Policy*, 36(6):2016–2029. 5, 7

Lantz, B. (2015). *Machine learning with R*. Packt Publishing Ltd. 1, 27, 28

Lesmeister, C. (2017). *Mastering machine learning with r*. Packt Publishing Ltd. 28, 29

Magnani, N. and Vaona, A. (2016). Access to electricity and socio-economic characteristics: Panel data evidence at the country level. *Energy*, 103:447–455. 6, 11

Mensah, G. S., Kemausuor, F., and Brew-Hammond, A. (2014). Energy access indicators and trends in Ghana. *Renewable and Sustainable Energy Reviews*, 30:317–323. 5, 6, 7, 8

Nussbaumer, P., Bazilian, M., and Modi, V. (2012). Measuring energy poverty: Focusing on what matters. *Renewable and Sustainable Energy Reviews*, 16(1):231–243. 8

Onyeji, I., Bazilian, M., and Nussbaumer, P. (2012). Contextualizing electricity access in sub-Saharan Africa. *Energy for Sustainable Development*, 16(4):520–527. 3

Pachauri, S., Brew-Hammond, A., Barnes, D., Bouille, D., Gitonga, S., Modi, V., Prasad, G., Rath, A., and Zerrifi, H. (2012). Energy access for development. *Cambridge University Press and IIASA*. 5, 7

Panos, E., Densing, M., and Volkart, K. (2016). Access to electricity in the World Energy Council's global energy scenarios: An outlook for developing regions until 2030. *Energy Strategy Reviews*, 9:28–49. 10

Parajuli, R., Østergaard, P. A., Dalgaard, T., and Pokharel, G. R. (2014). Energy consumption projection of Nepal: An econometric approach. *Renewable Energy*, 63:432–444. 10

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons. 9

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5):384–398. 9

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons. 15, 22

Pérez-Foguet, A., Giné-Garriga, R., and Ortego, M. I. (2017). Compositional data for global monitoring: The case of drinking water and sanitation. *Science of The Total Environment*, 590:554–565. 3, 11

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 30

Reimann, C., Filzmoser, P., Hron, K., Kynčlová, P., and Garrett, R. (2017). A new method for correlation analysis of compositional (environmental) data–a worked example. *Science of The Total Environment*, 607:965–971. 9

Spalding-Fecher, R., Winkler, H., and Mwakasonda, S. (2005). Energy and the World Summit on Sustainable Development: what next? *Energy Policy*, 33(1):99–112. 6

Suganthi, L. and Samuel, A. A. (2012). Energy models for demand forecasting—A review. *Renewable and sustainable energy reviews*, 16(2):1223–1240. 11

UN (2015). 70/1. Transforming our world: the 2030 Agenda for Sustainable Development-A. Technical report, RES/70/1. New York, USA: United Nations. 2

Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*, volume 122. Springer. 2, 11, 15, 19

Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons. 21, 23

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. CRC press. x, 26

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press. 24

WorldBank (2017). World Bank Open Data free and open access to global development data. 1, 3, 12, 13

Appendix A

Cross validation between CoDa and Raw model outside the calibration rank

This appendix displays some bar plots that complement the section 4.4 related to the cross validation between CoDa and raw compositions. These figures show the importance of CoDa instead of using raw compositions outside the calibration rank. The main conclusion is: *Outside the calibration rank CoDa is better than work with the raw compositions*.

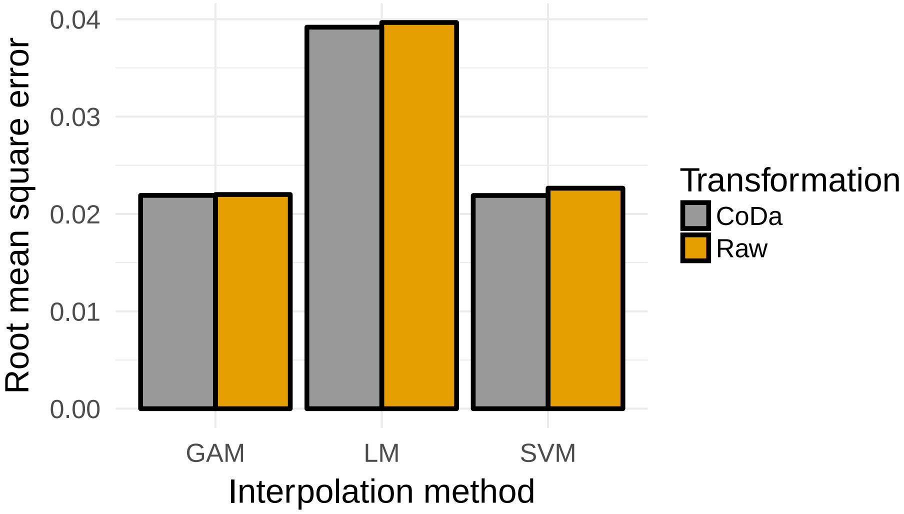


Figure A.0.1: Cross validation between CoDa and Raw model outside the calibration rank for Bangladesh using the RMSE

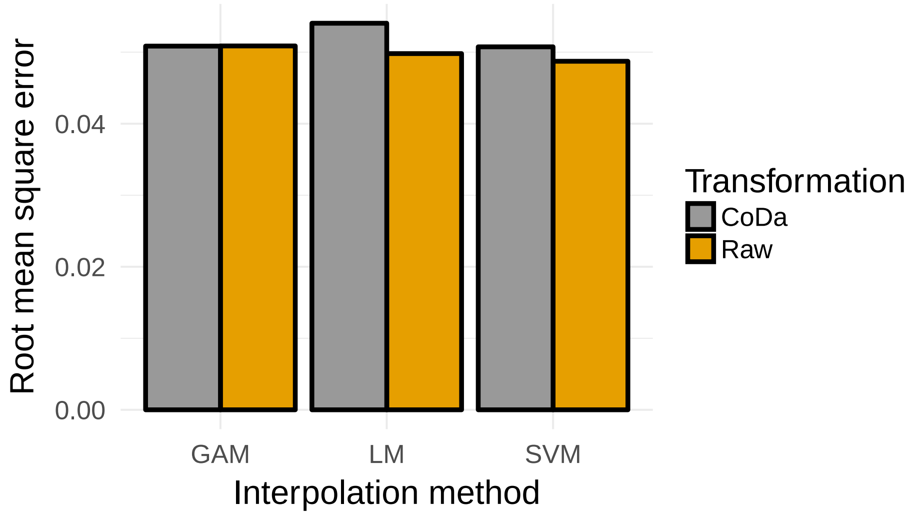


Figure A.0.2: Cross validation between CoDa and Raw model outside the calibration rank for India using the RMSE

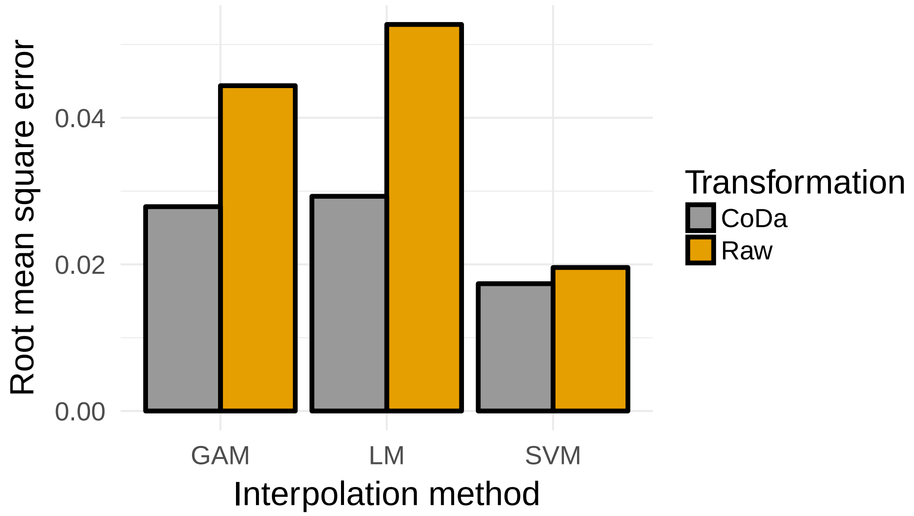


Figure A.0.3: Cross validation between CoDa and Raw model outside the calibration rank for Kenya using the RMSE

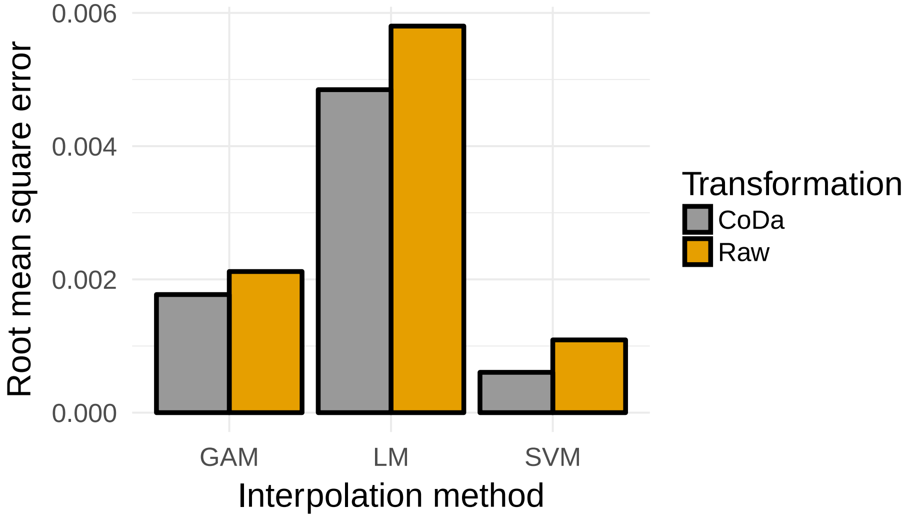


Figure A.0.4: Cross validation between CoDa and Raw model outside the calibration rank for Sudan using the RMSE

Appendix B

Parameters used in the configuration of the models

This appendix shows information related to the parameters of the different applied statistical models in the results chapter.

B.1 CoDa Models

B.1.1 Linear model with interactions

Table B.1.1: Configurations of the code used to estimate linear CoDa models with interactions.

Balance/country	Numeric covariate	Factor covariate	Interaction
Balance 1			
Bangladesh	`time`	`class`	`time:class`
India	`time`	-	-
Kenya	`poly(time,3)`	`class`	`poly(time,2):class`
Nigeria	`time`	`class`	`time:class`
Sudan	`time`	-	-
Balance 2			
Bangladesh	`poly(time,3)`	`class`	`time:class`
India	`poly(time,2)`	-	-
Kenya	`poly(time,3)`	`class`	`poly(time,2):class`
Nigeria	`poly(time,2)`	-	`time:class`
Sudan	`poly(time,2)`	-	-
Balance 3			
Bangladesh	`time`	`class`	`time:class`
India	`time`	-	-
Kenya	`time`	`class`	`time:class`
Nigeria	`time`	`class`	-
Sudan	`time`	-	-

B.1.2 GAM with interactions

Table B.1.2: Configurations of the code used to estimate GAM CoDa models with interactions.

Balance/country	Smooth function parameters s(time)					Factor	Complementary parameters		
	bs	k	fx	m	by=class	(class)	gamma	method	select
Balance 1									
Bangladesh	cs	4	FALSE	2	Yes	Yes	0.3585824	GCV.Cp	FALSE
India	cs	5	FALSE	3	Yes	No	0.6564086	GACV.Cp	FALSE
Kenya	cr	5	FALSE	3	Yes	No	0.6997835	GACV.Cp	TRUE
Nigeria	cs	9	TRUE	2	Yes	Yes	1.081015	REML	FALSE
Sudan	ts	-5	TRUE	2	No	No	0.7762536	REML	FALSE
Balance 2									
Bangladesh	cr	6	FALSE	3	Yes	Yes	0.4861186	P-ML	TRUE
India	cs	4	FALSE	2	Yes	No	0.6878384	P-REML	TRUE
Kenya	cs	4	TRUE	2	Yes	Yes	1.35335553	GACV.Cp	TRUE
Nigeria	cs	6	FALSE	3	Yes	Yes	1.0234638	P-REML	TRUE
Sudan	ts	5	TRUE	-	No	No	0.8804668	GCV.Cp	FALSE
Balance 3									
Bangladesh	cs	5	TRUE	2	Yes	Yes	1.136932	GCV.Cp	TRUE
India	cs	2	TRUE	3	Yes	No	1.556758	GCV.Cp	FALSE
Kenya	cs	7	TRUE	2	Yes	No	0.3396195	ML	FALSE
Nigeria	cs	4	FALSE	2	Yes	No	1.96798	GACV.Cp	TRUE
Sudan	ts	13	TRUE	-	No	No	0.8256237	P-REML	FALSE

B.1.3 SVM with interactions

Table B.1.3: Configurations of the code used to estimate SVM CoDa models with interactions.

Balance/country	Formula $f(t, z)$	cost	gamma	epsilon	kernel	coef0	degree
Balance 1							
Bangladesh	time*class	422.6625	0.01174233	0.04613106	radial		
India	time*class	373.8075	0.1138616	0.06965857	radial		
Kenya	time*class	1229.608	0.9642096	0.01225341	polynomial	0.06258453	3.814607
Nigeria	time*class	324.1104	0.004177228	0.004892558	radial		
Sudan	time	1461.952	0.02709443	0.001242271	radial		
Balance 2							
Bangladesh	time*class	1057.652	0.04646314	0.02467979	radial		
India	time*class	1873.09	0.04931593	0.1086807	radial		
Kenya	time*class	1792.476	0.9283877	0.14908	polynomial	0.5630647	3.466137
Nigeria	time*class	59.69329	0.9930897	0.05828638	radial		
Sudan	time	922.2505	0.07612343	0.0121935	radial		
Balance 3							
Bangladesh	time*class	896.6357	0.06013468	0.1295139	radial		
India	time*class	1613.796	0.05276283	0.1193572	radial		
Kenya	time*class	1757.871	0.09128533	0.1477945	polynomial	5.767135	4.045442
Nigeria	time*class	1590.859	0.1612484	0.03170586	radial		
Sudan	time	1343.712	0.01330174	0.0008476018	radial		

B.1.4 Considerations in the case of models with a single numeric covariate

Table B.1.4: Configurations of the code used to estimate linear CoDa models with a single covariate.

Balance/country	Numeric covariate structure
Balance 1	
Bangladesh	`time`
India	`time`
Kenya	`time`
Nigeria	`time`
Sudan	`time`
Balance 2	
Bangladesh	`poly(time,3)`
India	`poly(time,2)`
Kenya	`time`
Nigeria	`poly(time,2)`
Sudan	`time`
Balance 3	
Bangladesh	`time`
India	`time`
Kenya	`time`
Nigeria	`time`
Sudan	`time`

B.2 Models for Raw data

B.2.1 Linear configuration

Table B.2.1: Configurations of the code used to estimate linear Raw models with interactions.

Balance/country	Numeric covariate	Factor covariate	Interaction
Urban s. with electricity			
Bangladesh	`time`	`class`	`time:class`
India	`time`	-	-
Kenya	`time`	`class`	`time:class`
Nigeria	`poly(time,2)`	`class`	`poly(time,2):class`
Sudan	`poly(time,2)`	-	-
Rural s. with electricity			
Bangladesh	`time`	`class`	`time:class`
India	`time`	-	-
Kenya	`time`	`class`	`poly(time,2):class`
Nigeria	`poly(time,2)`	`class`	`poly(time,2):class`
Sudan	`poly(time,2)`	-	-
Urban s. without electricity			
Bangladesh	`time`	-	-
India	`time`	-	-
Kenya	`time`	`class`	`time:class`
Nigeria	`poly(time,2)`	`class`	`poly(time,2):class`
Sudan	`time`	-	-
Rural s. without electricity			
Bangladesh	`time`	-	-
India	`time`	-	-
Kenya	`time`	`class`	`poly(time,2):class`
Nigeria	`poly(time,3)`	`class`	`poly(time,3):class`
Sudan	`poly(time,2)`	-	-

B.2.2 GAM configuration

Table B.2.2: Configurations of the code used to estimate GAM Raw models with interactions.

Balance/country	Smooth function parameters s(time)					Factor	Complementary parameters		
	bs	k	fx	m	by=class	(class)	gamma	method	select
Urban s. with electricity									
Bangladesh	cs	5	TRUE	2	Yes	Yes	1.0517752	P-REML	FALSE
India	cs	6	FALSE	3	Yes	No	0.8236297	P-REML	FALSE
Kenya	cr	4	FALSE	3	Yes	Yes	1.251807	ML	TRUE
Nigeria	cs	8	FALSE	2	Yes	Yes	1.0010443	REML	FALSE
Sudan	cr	4	FALSE	2	No	No	1.967258	GACV.Cp	TRUE
Rural s. with electricity									
Bangladesh	cs	6	FALSE	3	Yes	Yes	1.199093	GACV.Cp	TRUE
India	cs	4	FALSE	3	Yes	No	0.9344174	P-REML	TRUE
Kenya	cs	7	FALSE	3	Yes	Yes	0.7277974	P-REML	FALSE
Nigeria	cs	4	FALSE	3	Yes	No	1.474995	P-ML	FALSE
Sudan	cs	14	TRUE	3	No	No	1.492431	REML	TRUE
Urban s. without electricity									
Bangladesh	cs	5	TRUE	3	Yes	Yes	1.058473	P-REML	FALSE
India	cs	2	FALSE	3	Yes	No	0.8125455	GACV.Cp	FALSE
Kenya	cs	5	TRUE	3	Yes	Yes	0.7114296	GCV.Cp	FALSE
Nigeria	cs	6	TRUE	3	Yes	Yes	1.1093769	P-REML	FALSE
Sudan	cs	15	FALSE	3	No	No	0.9578136	GACV.Cp	FALSE
Rural s. without electricity									
Bangladesh	cr	6	FALSE	3	Yes	Yes	0.6979624	REML	FALSE
India	cs	8	FALSE	3	Yes	No	1.127840	P-REML	TRUE
Kenya	cs	7	FALSE	3	Yes	Yes	0.6442111	REML	FALSE
Nigeria	cs	5	FALSE	1	Yes	Yes	0.3833242	P-ML	FALSE
Sudan	cs	15	TRUE	3	No	No	1.395701	P-REML	TRUE

B.2.3 SVM configuration

Table B.2.3: Configurations of the code used to estimate SVM Raw models with interactions.

Balance/country	Formula $f(t, z)$	`cost`	`gamma`	`epsilon`	`kernel`
Urban s. with electricity					
Bangladesh	`time*class`	670.2429	0.06072633	0.0896873	`radial`
India	`time*class`	447.3538	0.06656046	0.0752192	`radial`
Kenya	`time*class`	115.2261	0.006169162	0.009422912	`radial`
Nigeria	`time*class`	959.0335	0.009250104	0.02365988	`radial`
Sudan	`time`	676.0736	0.06562984	0.008478297	`radial`
Rural s. with electricity					
Bangladesh	`time*class`	662.8796	0.01353806	0.0238125	`radial`
India	`time*class`	924.5399	0.06264566	0.01549738	`radial`
Kenya	`time*class`	89.90413	0.003396583	0.02539831	`radial`
Nigeria	`time*class`	17.1627	0.03082246	0.907276	`radial`
Sudan	`time`	737.6026	0.04106443	0.006594902	`radial`
Urban s. without electricity					
Bangladesh	`time*class`	627.6893	0.1176374	0.01581494	`radial`
India	`time*class`	0.0724129	0.0724129	0.1487976	`radial`
Kenya	`time*class`	101.7273	0.009708038	0.02240919	`radial`
Nigeria	`time*class`	29.49511	0.06702711	0.08153087	`radial`
Sudan	`time`	332.1218	0.02526219	0.002168561	`radial`
Rural s. without electricity					
Bangladesh	`time*class`	539.0089	0.0418369	0.008133287	`radial`
India	`time*class`	614.8501	0.06708856	0.0158196	`radial`
Kenya	`time*class`	302.4322	0.002080184	0.02074777	`radial`
Nigeria	`time*class`	521.5111	0.02427693	0.008321363	`radial`
Sudan	`time`	889.0924	0.03564625	0.001016959	`radial`

Appendix C

R Code

This appendix displays the script related to the LM, SVM and GAM estimated models for Bangladesh and the code used to reproduce the different plots showed in the Results chapter. To reproduce other country should be changed the parameters of the different models (see appendix B) and change the country name. The data set of each country is in appendix D.

```
# ================
# Functions
# ----------------

multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use ’cols’ to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}

g_legend<-function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)}

#---------------------

load("CODA.RData") # the data
head(CODA)

load("SBP.RData") # the binary matrix
head(SBP)

V=balances(CODA[,c("urban","rural","nourban","norural")],t(SBP))$V # change matrix
rm(SBP)

pais <- "Bangladesh" #selecting a country between: Bangladesh, India, Kenya, Nigeria, Sudan

# compute the ILR and selection of some variables:

ILR <- cbind(ilr(CODA[CODA$Country.Name == pais, c("urban", "rural", "nourban", "norural")], V = V
             CODA[CODA$Country.Name == pais, c("class", "time", "data", "I", "plot")])

colnames(ILR) <- c("BALANCE1", "BALANCE2", "BALANCE3", "class", "time", "data", "I", "plot")
head(ILR)

## ############################################################################################
## Exercise a)
#
## ========================================================================================== #
##        Computing LM, GAM and SVM CoDa single covariate statistical models.
#
## ------------------------------------------------------------------------------------------ #
###############################################################################################

## ------------------
## Linear model .-
## ==================

# Parameter configurations:

LM1S <- lm(BALANCE1 ~ time, data = ILR[ILR$plot == 0, ])
LM2S <- lm(BALANCE2 ~ poly(time, 3), data = ILR[ILR$plot == 0, ])
LM3S <- lm(BALANCE3 ~ time,data = ILR[ILR$plot == 0, ])

# Balance Prediction:

BALANCE1 <- cbind( predict(LM1S, interval = "confidence", ILR),
                   ILR = ILR$BALANCE1, ILR[ ,c("time" , "data","I","plot","class")], BALANCE = "BA
BALANCE2 <- cbind( predict(LM2S, interval = "confidence", ILR),
                   ILR = ILR$BALANCE2, ILR[ ,c("time" , "data","I","plot","class")], BALANCE = "BA
BALANCE3 <- cbind( predict(LM3S, interval = "confidence", ILR),
                   ILR = ILR$BALANCE3, ILR[ ,c("time" , "data","I","plot","class")], BALANCE = "BA

BALANCES.LMS <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.LMS<-with(BALANCES.LMS, BALANCES.LMS[order(BALANCE, data, time), ])
BALANCES.LMS$model <- "LM"
rm(LM1S, LM2S, LM3S, BALANCE1, BALANCE2, BALANCE3)

## ------------------
## GAM model .-
## ==================

# Parameter configurations:

gam1S <- gam(BALANCE1 ~ s(time, bs = "cs", k = 4, fx = F, m = 2),
           gamma = 0.3585824, method = "GCV.Cp", select = F,data = ILR[ILR$plot == 0, ])
gam2S <- gam(BALANCE2 ~ s(time, bs = "cr", k = 6, fx = F,m = 3),
           gamma = 0.4861186, method = "P-ML", select = T, data = ILR[ILR$plot == 0, ])
gam3S <- gam(BALANCE3 ~ s(time, bs = "cs", k = 5, fx = T, m = 2),
           gamma = 1.136932, method = "GCV.Cp", select = T, data = ILR[ILR$plot==0,])

# Balance Prediction:

pre1 <- predict(gam1S , ILR, type = "link", se.fit = TRUE)
BALANCE1 <- cbind.data.frame(fit = pre1$fit, lwr = pre1$fit - (2*pre1$se.fit),
                             upr = pre1$fit + (2*pre1$se.fit),
                             ILR = ILR$BALANCE1, ILR[ , c("time", "data", "I", "plot", "class")],
                             BALANCE = "BALANCE1")
rm(pre1)

pre2 <- predict(gam2S, ILR, type = "link", se.fit = TRUE)
BALANCE2 <- cbind.data.frame(fit = pre2$fit, lwr = pre2$fit - (2*pre2$se.fit),
                             upr = pre2$fit + (2*pre2$se.fit),
                             ILR = ILR$BALANCE2, ILR[ ,c("time", "data", "I", "plot", "class")],
                             BALANCE = "BALANCE2")
rm(pre2)

pre3 <- predict(gam3S, ILR, type = "link", se.fit = TRUE)
BALANCE3 <- cbind.data.frame(fit = pre3$fit, lwr = pre3$fit - (2*pre3$se.fit),
                             upr = pre3$fit + (2*pre3$se.fit),
                             ILR = ILR$BALANCE3, ILR[ ,c("time", "data", "I", "plot", "class")],
                             BALANCE = "BALANCE3")
rm(pre3)

BALANCES.GAMS <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.GAMS <- with(BALANCES.GAMS, BALANCES.GAMS[order(BALANCE, data, time), ])
BALANCES.GAMS$model <- "GAM"
rm(gam1S, gam2S, gam3S, BALANCE1, BALANCE2, BALANCE3)

## ------------------
## SVM model .-
## ==================

# Parameter configurations:

svm1S <- svm(BALANCE1 ~ time, cost = 422.6625, gamma = 0.01174233, epsilon = 0.04613106,
            kernel = "radial", data = ILR[ILR$plot == 0, ], type = 'eps-regression')
svm2S <- svm(BALANCE2 ~ time, cost = 1057.652, gamma = 0.04646314, epsilon = 0.02467979,
             kernel = "radial", data = ILR[ILR$plot == 0, ], type = 'eps-regression')
svm3S <- svm(BALANCE3 ~ time, cost = 896.6357, gamma = 0.06013468, epsilon = 0.1295139,
            kernel = "radial", data = ILR[ILR$plot == 0, ], type = 'eps-regression')

# Balance Prediction:

BALANCE1 <- cbind( fit = predict(svm1S, ILR), ILR = ILR$BALANCE1,
                   ILR[ , c("time", "data", "I", "plot", "class")],
                    BALANCE = "BALANCE1")
BALANCE2 <- cbind( fit = predict(svm2S, ILR), ILR = ILR$BALANCE2,
                   ILR[ ,c("time", "data", "I", "plot", "class")],
                    BALANCE = "BALANCE2")
BALANCE3 <- cbind( fit = predict(svm3S, ILR), ILR = ILR$BALANCE3,
                   ILR[ ,c("time", "data", "I", "plot", "class")],
                   BALANCE = "BALANCE3")

BALANCES.SVMS <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.SVMS <- with(BALANCES.SVMS, BALANCES.SVMS[order(BALANCE, data, time), ])
BALANCES.SVMS$model <- "SVM"
rm(svm1S, svm2S, svm3S, BALANCE1, BALANCE2, BALANCE3)

# ==============================================
# Global plot configuration
# ----------------------------------------------

# Match over LM, GAM and SVM balance predictions

GTS <- rbind(BALANCES.LMS[ , -c(2,3)], BALANCES.GAMS[ , -c(2,3)], BALANCES.SVMS)
GTS <- GTS[GTS$plot == 0, ]
rm(BALANCES.LMS, BALANCES.GAMS, BALANCES.SVMS)

# ilr^-1 to return to the raw form

PredF <- spread(GTS[ , c("time", "fit", "BALANCE", "class", "model")], key = BALANCE, value = fit)
ilrinv <- ilrInv(PredF[ , 4:6], V = V)
PredT <- cbind(ilrinv, PredF[ , 1:3])
colnames(PredT) <- c("urban", "rural", "nourban", "norural", "time", "val", "model")
rm(PredF, ilrinv)

# RAW original variate

Y <- CODA[CODA$Country.Name == pais & CODA$plot==0,
          c("urban", "rural", "nourban", "norural", "time", "class")]

Yg <- gather(Y, key = access, value = value, urban:norural)

# Match of original and fitted raw variate

YS <- gather(PredT, key = access, value = value, urban:norural)
YS2 <- cbind(YS[YS$model == "LM", ], GAM = YS[YS$model == "GAM", "value"],
             SVM = YS[YS$model == "SVM", "value"])
colnames(YS2) <- c("time", "class", "model", "access", "LM", "GAM", "SVM")
YS2$model <- NULL
YS2$total <- Yg$value
rm(YS, PredT)

# RMSE without interactions

FIT.W <- gather(YS2, key = model, value = fit, LM:SVM)
FIT.W$SE <- (FIT.W$fit - FIT.W$total)^2
RMSE.W <- aggregate( SE ~ model, data = FIT.W, FUN = function(x)sqrt(sum(x)/length(x)))
RMSE.W <- RMSE.W[order(RMSE.W$SE), ]
RMSE.W$position <- 1:3
RMSE.W$position <- as.factor(RMSE.W$position)
levels(RMSE.W$position) <- RMSE.W$model
RMSE.W #RMSE for GAM, LM and SVM ordered from lowest to highest
rm(FIT.W)

# ===================================================================================
# Figure 4.1.1 - 4.1.2: LM, GAM and SVM CoDa single covariate statistical models for ...
# -----------------------------------------------------------------------------------

W1 <- ggplot(GTS, aes(time, fit, colour = model)) +
  geom_line(size = .8) +
  geom_point(aes(time, ILR), colour = "black", shape = 21, size = 2) +
  facet_wrap( ~ BALANCE, nrow = 1, scales = "free_y", dir = "v") +
  ylab("ilr") +
  guides(colour = guide_legend(title = NULL, reverse = F, ncol = 1)) +
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(1, 3, 2, 5)]) +
  theme(legend.position = "right") +
  theme(text = element_text(size = 18), axis.text.x = element_text(angle = 0)) +
  coord_fixed( )
rm(GTS)

W2 <- ggplot(YS2, aes(time, total, shape = access)) +
  geom_point(size = 1.5, colour = "blue4") +
  geom_line(aes(y = LM, colour = "LM"), size = .8) +
  geom_line(aes(y = GAM, colour = "GAM"), size = .8) +
  geom_line(aes(y = SVM, colour = "SVM"), size = .8) +
  ylab("%␣of␣population␣total") +
  scale_shape_manual(values = c(21, 2, 19, 17)) +
  scale_colour_manual("Model:",
                      breaks = c("LM", "GAM", "SVM"),
                      values = brewer.pal(8, "Set1")[c(3, 1, 2)]) +
  labs(shape = "Electricity␣\n␣access") +
  theme(aspect.ratio = 7 / 11, text = element_text(size = 18))

W3 <- ggplot(RMSE.W, aes(position, SE)) +
      geom_bar(colour = "black", stat = "identity", fill = "#FF6666") +
      labs(x = ’Fitted␣model’, y = ’Root␣mean␣square␣error’) +
      theme(aspect.ratio = 9 / 9, text = element_text(size = 18))
rm(RMSE.W)

layout <- matrix(c(3, 3, 3,1, 2, 2), nrow = 2, byrow = TRUE)
pdf("Bangladesh_ILR_sin.pdf",width=9.8,height=7.5)
multiplot(W3,W2,W1,cols = 2,layout = layout)
dev.off()
rm(layout,W1,W2,W3)

## ##################################################################################
## Exercise b)
#
## ================================================================================ #
##       Figure 4.2.1: Electricity access balance segmented by z for ...
#
## -------------------------------------------------------------------------------- #
#####################################################################################
```

```
ILR6 <- gather(ILR[ILR$plot == 0, ], key=ILR, value = value, BALANCE1:BALANCE3)

colours <- c( `1` = "red", `0`="blue")
gc <- ggplot(ILR6, aes(time, value, colour = as.factor(class))) +
  geom_point(shape=21) +
  facet_wrap( ~ ILR, nrow = 1, scales = "free_y", dir = "v") +
  ylab("ilr") +
  guides(colour = guide_legend(title = 'Validation␣with␣electricty␣acces␣(%␣of␣total):␣',reverse =
  scale_colour_manual(values = colours, labels = c("Same", "Different")) +
  theme(legend.position = "bottom",text = element_text(size = 16))

pdf("ILR_Bang_class.pdf", width=8, height = 4)
gc
dev.off()
rm(gc, colours, ILR6)

## #######################################################################################
## Exercise c)
#
## ===================================================================================== #
##        Figure 4.2.2: LM, GAM, SVM Interaction models for ...
#
## ------------------------------------------------------------------------------------- #
## #######################################################################################

## ------------------
## Linear model .-
## ==================

# Parameter configurations:

LM1T <- lm(BALANCE1 ~ time + class, data = ILR[ILR$plot == 0, ])
LM2T <- lm(BALANCE2 ~ poly(time,3) + class + time:class, data = ILR[ILR$plot == 0, ])
LM3T <- lm(BALANCE3 ~ time + class + time:class, data = ILR[ILR$plot == 0, ])

# Balance Prediction:

BALANCE1 <- cbind( predict(LM1T, interval = "confidence", ILR),
                   ILR = ILR$BALANCE1, ILR[ , c("time", "data", "I", "plot", "class")],
                   BALANCE = "BALANCE1")
BALANCE2 <- cbind( predict(LM2T, interval = "confidence", ILR),
                   ILR = ILR$BALANCE2, ILR[ ,c("time", "data", "I", "plot", "class")],
                   BALANCE = "BALANCE2")
BALANCE3 <- cbind( predict(LM3T, interval = "confidence", ILR),
                   ILR = ILR$BALANCE3, ILR[ ,c("time", "data", "I", "plot", "class")],
                   BALANCE = "BALANCE3")

BALANCES.LMT <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.LMT <- with(BALANCES.LMT, BALANCES.LMT[order(BALANCE, data, time), ])
BALANCES.LMT$model <- "LM"
rm(BALANCE1, BALANCE2, BALANCE3)

## ------------------
## GAM model .-
## ==================

# Parameter configurations:

gam1T <- gam(BALANCE1 ~ s(time, bs = "cs", k = 4, fx = F, m = 2,
         by = factor(class)) + factor(class), gamma = 0.3585824,
         method = "GCV.Cp", select = F, data = ILR[ILR$plot == 0,])
gam2T <- gam(BALANCE2 ˜ s(time, bs = "cr", k = 6, fx = F, m = 3,
         by = factor(class)) + factor(class), gamma = 0.4861186,
         method = "P-ML", select = T, data = ILR[ILR$plot == 0, ])
gam3T <- gam(BALANCE3 ˜ s(time, bs = "cs", k = 5, fx = T, m = 2,
         by = factor(class)) + factor(class), gamma = 1.136932,
         method = "GCV.Cp", select = T, data = ILR[ILR$plot == 0, ])

# Balance Prediction:

pre1 <- predict(gam1T, ILR, type = "link", se.fit = TRUE)
BALANCE1 <- cbind.data.frame(fit = pre1$fit, lwr = pre1$fit - (2*pre1$se.fit),
                             upr = pre1$fit + (2*pre1$se.fit), ILR = ILR$BALANCE1,
                             ILR[ , c("time", "data", "I", "plot", "class")], BALANCE = "BALANCE1"
rm(pre1)

pre2 <- predict(gam2T, ILR, type="link", se.fit = TRUE)
BALANCE2 <- cbind.data.frame(fit = pre2$fit, lwr = pre2$fit - (2*pre2$se.fit),
                             upr = pre2$fit + (2*pre2$se.fit), ILR = ILR$BALANCE2,
                             ILR[ , c("time", "data", "I", "plot", "class")], BALANCE = "BALANCE2"
rm(pre2)

pre3 <- predict(gam3T, ILR, type = "link", se.fit = TRUE)
BALANCE3 <- cbind.data.frame(fit = pre3$fit, lwr = pre3$fit - (2*pre3$se.fit),
                             upr = pre3$fit + (2*pre3$se.fit), ILR=ILR$BALANCE3,
                             ILR[ , c("time", "data", "I", "plot", "class")], BALANCE = "BALANCE3"
rm(pre3)

BALANCES.GAMT <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.GAMT <- with(BALANCES.GAMT, BALANCES.GAMT[order(BALANCE, data, time), ])
BALANCES.GAMT$model <- "GAM"
rm(BALANCE1, BALANCE2, BALANCE3)

## ------------------
## SVM model .-
## ==================

# Parameter configurations:

svm1T <- svm(BALANCE1 ˜ time*factor(class), cost = 422.6625, gamma = 0.01174233,
             epsilon = 0.04613106, kernel = "radial", data = ILR[ILR$plot==0,], type = ’eps-regres
svm2T <- svm(BALANCE2 ˜ time*factor(class), cost = 1057.652, gamma = 0.04646314, epsilon = 0.02467
             kernel = "radial", data = ILR[ILR$plot == 0, ], type = ’eps-regression’)
svm3T <- svm(BALANCE3 ˜ time*factor(class), cost = 896.6357, gamma = 0.06013468, epsilon = 0.12951
             kernel = "radial", data = ILR[ILR$plot == 0, ], type = ’eps-regression’)

# Balance Prediction:

BALANCE1 <- cbind(fit = predict(svm1T, ILR), ILR = ILR$BALANCE1,
                  ILR[ , c("time", "data", "I", "plot", "class")], BALANCE = "BALANCE1")
BALANCE2 <- cbind(fit = predict(svm2T, ILR), ILR = ILR$BALANCE2,
                  ILR[ , c("time", "data", "I", "plot", "class")], BALANCE = "BALANCE2")
BALANCE3 <- cbind(fit = predict(svm3T, ILR), ILR = ILR$BALANCE3,
                  ILR[ ,c("time", "data", "I", "plot", "class")], BALANCE = "BALANCE3")

BALANCES.SVMT <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.SVMT <- with(BALANCES.SVMT, BALANCES.SVMT[order(BALANCE, data, time), ])
BALANCES.SVMT$model <- "SVM"
```

```
rm(BALANCE1, BALANCE2, BALANCE3)

# ==============================================
# Global plot configuration
# ----------------------------------------------

# Match over LM, GAM and SVM balance predictions

GTF <- rbind(BALANCES.LMT[ , -c(2,3)], BALANCES.GAMT[ , -c(2,3)], BALANCES.SVMT)
GTF <- GTF[GTF$plot==0, ]
rm(BALANCES.LMT, BALANCES.GAMT, BALANCES.SVMT)

# ilrˆ-1 to return to the raw form

PredF <- spread(GTF[, c("time", "fit", "BALANCE", "class", "model")], key = BALANCE, value = fit)
ilrinv <- ilrInv(PredF[ , 4:6], V = V)
PredT <- cbind(ilrinv, PredF[ , 1:3])
colnames(PredT) <- c("urban", "rural", "nourban", "norural", "time", "val", "model")
rm(PredF, ilrinv)

# RAW original variate

#Y <- CODA[CODA$Country.Name == pais & CODA$plot==0,
#          c("urban", "rural", "nourban", "norural","time","class")]

#Yg<-gather(Y, key = access, value=value, urban:norural)

# Match of original and fitted raw variate

YF <- gather(PredT, key = access, value=value, urban:norural)
YF2 <- cbind(YF[YF$model == "LM", ], GAM = YF[YF$model == "GAM", "value"],
             SVM = YF[YF$model == "SVM", "value"])
colnames(YF2) <- c("time", "val", "model", "access", "LM", "GAM",  "SVM")
YF2$model <- NULL
YF2$total <- Yg$value
rm(YF, PredT)

# RMSE without interactions

FIT.I <- gather(YF2, key=model, value = fit, LM:SVM)
FIT.I$SE <- (FIT.I$fit - FIT.I$total)ˆ2
RMSE.I <- aggregate(SE ˜ model, data = FIT.I, FUN = function(x)sqrt(sum(x)/length(x)))
RMSE.I <- RMSE.I[order(RMSE.I$SE) , ]
RMSE.I$position <- 1:3
RMSE.I$position <- as.factor(RMSE.I$position)
levels(RMSE.I$position) <- RMSE.I$model
RMSE.I #RMSE for GAM, LM and SVM ordered from lowest to highest
rm(FIT.I)

# =======================================================================================
# Figure 4.2.2: LM, GAM, SVM Interaction models for ...
# ---------------------------------------------------------------------------------------

I1 <- ggplot(GTF, aes(time, fit, colour = model, shape = factor(class))) +
  geom_line(size = .8) +
  geom_point(aes(time, ILR), colour = "black", size = 2) +
  facet_wrap( ˜ BALANCE, nrow = 1, scales = "free_y", dir = "v") +
  ylab("ilr") +
  guides(colour = guide_legend(title = NULL, reverse = F, ncol = 1)) +
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(3, 1, 2, 5)]) +
  theme(legend.position = "right") +
  theme(text = element_text(size = 18), axis.text.x = element_text(angle = 0)) +
  scale_shape_manual(values = c(21, 6), labels = c("Same", "Diff.")) +
  labs(shape = "class") +
  coord_fixed( )
rm(GTF)

I2 <- ggplot(YF2, aes(time, total, shape = access)) +
  geom_point(size = 1.5, colour = "blue4") +
  geom_line(aes(y = LM, colour = "LM"), size = .8) +
  geom_line(aes(y = GAM, colour = "GAM"), size = .8) +
  geom_line(aes(y = SVM, colour = "SVM"), size = .8) +
  ylab("%␣of␣population␣total") +
  scale_shape_manual(values = c(21, 2,19, 17)) +
  scale_colour_manual("Model:",
                      breaks = c("LM", "GAM", "SVM"),
                      values = brewer.pal(8, "Set1")[c(3, 1, 2)]) +
  labs(shape = "Electricity␣\n␣access") +
  theme(aspect.ratio = 7 / 11, text = element_text(size = 18))

I3 <- ggplot(RMSE.I, aes(position, SE)) +
  geom_bar(colour="black", stat ="identity", fill = "#FF6666") +
  labs(x = 'Fitted␣model', y = 'Root␣mean␣square␣error') +
  theme(aspect.ratio = 9 / 9,text = element_text(size = 18))
rm(RMSE.I)

layout <- matrix(c(3, 3, 3,1, 2, 2), nrow = 2, byrow = TRUE)
pdf("Bangladesh_ILR_con.pdf", width = 9.8, height = 7.5)
multiplot(I3, I2, I1, cols = 2, layout = layout)
dev.off( )
rm(layout, I1, I2, I3)

## ##############################################################################################
## Exercise d)
#
## ============================================================================================ #
## Figure 4.2.3: Comparison between single covariate model versus interaction model for ...
#
## -------------------------------------------------------------------------------------------- #
## ##############################################################################################

g1 <- ggplot(YF2, aes(time, total, shape = access)) +
  geom_point(size = 1.5, colour = "blue4") +
  geom_line(aes(y = LM, colour = "ILM"), size = .8, position = 'jitter') +
  geom_line(aes(y = YS2$LM, colour = "LM"), linetype = "dashed", size = .8, position = 'jitter') +
  ylab("%␣of␣population␣total") +
  scale_shape_manual(values = c(21, 2,19, 17)) +
  scale_colour_manual("Model:", breaks = c("ILM", "LM"),
                      values =c("skyblue3", "red3"), label=c("Interaction", "No␣interaction"))+
  labs(shape = "Electricity␣\n␣access")+
  ggtitle("LM")+
  theme_bw( )+
  theme(legend.position = "bottom")

g2 <- ggplot(YF2, aes(time, total, shape = access)) +
  geom_point(size = 1.5, colour = "blue4") +
  geom_line(aes(y = GAM, colour = "ILM"), size=.8, position = 'jitter') +
  geom_line(aes(y = YS2$GAM, colour = "GAM"), linetype = "dashed", size=.8, position = 'jitter') +
  ylab("%␣of␣population␣total") +
  scale_shape_manual(values = c(21, 2, 19, 17)) +
  scale_colour_manual("Model:", breaks = c("ILM", "GAM"),
                      values = c("red3", "skyblue3"), label = c("Interaction", "No␣interaction"))
  labs(shape = "Electricity␣\n␣access") +
  ggtitle("GAM") +
  theme_bw() +
  theme(legend.position = "bottom")


g3 <- ggplot(YF2, aes(time, total, shape = access)) +
  geom_point(size = 1.5, colour = "blue4") +
  geom_line(aes(y = SVM, colour = "ILM"), size = .8, position = ’jitter’) +
  geom_line(aes(y = YS2$SVM, colour = "SVM"), linetype = "dashed", size = .8, position = ’jitter’)
  ylab("%␣of␣population␣total") +
  scale_shape_manual(values = c(21, 2, 19, 17))+
  scale_colour_manual("Model:",breaks = c("ILM", "SVM"),
                      values =c("skyblue3", "red3"),label=c("Interaction", "No␣interaction"))+
  labs(shape = "Electricity␣\n␣access") +
  ggtitle("SVM") +
  theme_bw( ) +
  theme(legend.position = "bottom")

hlay <-rbind(c(1, 1, 2, 2, 3, 3),
             c(1, 1, 2, 2, 3, 3))

mylegend<-g_legend(g1)

# ==========================================================================================
# Figure 4.2.3: Comparison between single covariate model versus interaction model for ...
# ------------------------------------------------------------------------------------------

p3 <- grid.arrange(arrangeGrob(g1 + theme(legend.position="none"),
                               g2 + theme(legend.position="none"),
                               g3 + theme(legend.position="none"),
                               layout_matrix = hlay),
                   mylegend,heights=c(10, 1))

ggsave(file="Bangladesh_versus.pdf", p3, width = 10, height = 5)
rm(YS2, YF2,g1, g2, g3, p3, hlay, mylegend)

## #####################################################
## Exercise e)                                         #
## ==================================================  #
## Figure: Prediction outside the calibration range.   #
## ----------------------------------------------------#
## #####################################################

## ------------------
## Linear model .-
## ==================

# Parameter configurations:

LM1 <- lm(BALANCE1 ˜ time + class, data = ILR[ILR$data == "base", ])
LM2 <- lm(BALANCE2 ˜ poly(time,3) + class + time:class, data = ILR[ILR$data == "base", ])
LM3 <- lm(BALANCE3 ˜ time + class + time:class, data = ILR[ILR$data == "base", ])

# Balance Prediction:
```

```
BALANCE1 <- cbind( predict(LM1, interval = "confidence", ILR),
                   ILR = ILR$BALANCE1, ILR[ , c("time", "data", "I", "plot")], BALANCE = "BALANCE1
BALANCE2 <- cbind( predict(LM2, interval = "confidence", ILR),
                   ILR = ILR$BALANCE2, ILR[ ,c("time", "data", "I", "plot")], BALANCE = "BALANCE2"
BALANCE3 <- cbind( predict(LM3, interval = "confidence", ILR),
                   ILR = ILR$BALANCE3, ILR[ ,c("time", "data", "I", "plot")], BALANCE = "BALANCE3"

BALANCES <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES <- with(BALANCES, BALANCES[order(BALANCE, data, time), ])
rm(LM1, LM2, LM3, BALANCE1, BALANCE2, BALANCE3)

# ==============================================
# Global plot configuration
# ----------------------------------------------

# Match over LM, GAM and SVM balance predictions

fitted <- spread(BALANCES[ , c("time", "data", "BALANCE", "fit")], key = BALANCE, value = fit)
fitted <- with(fitted, fitted[order(data, time), ])
ilrinv <- as.data.frame(ilrInv(fitted[ , c("BALANCE1", "BALANCE2", "BALANCE3")], V = V))
rm(fitted)

## arreglo en visualizacion de los datos
LM.CODA <- cbind(gather(CODA[CODA$Country.Name == pais, ], key = access, value=real, urban:norural
                 fitted = gather(ilrinv, key = access, value = value, Var1:Var4)$value)
rm(ilrinv)

# ==================================================================================
# LM Interaction model outside the calibration range
# ----------------------------------------------------------------------------------

OLM1 <- ggplot(BALANCES,aes(time, fit, ymin = lwr, ymax = upr, colour=I))+
  geom_line()+
  geom_smooth(stat = "identity")+
  geom_point(aes(time, ILR, colour=I))+
  facet_wrap(~BALANCE, nrow = 1,scales = "free_y",dir="v")+
  ylab("LM␣ilr")+
  guides(colour = guide_legend(title = NULL, reverse = T, nrow = 1))+
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(3,5,2,1)])+
  theme(legend.position = "bottom")+
  theme(text = element_text(size = 18),axis.text.x = element_text(angle = 0))#-90
rm(BALANCES)

OLM2 <- ggplot(LM.CODA, aes(time, real, colour = data, shape = access))+
  geom_point( )+
  geom_line(aes(time, fitted), size = .6) +
  ylab("%␣of␣population␣total") +
  guides(colour = guide_legend(title = "Data", reverse = F)) +
  scale_shape_manual(values = c(21, 2, 19, 17)) +
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(2, 1, 4, 5, 6, 7)], labels=c("base", "fitte
  labs(shape = "Electricity␣\n␣access") +
  theme(text = element_text(size = 18))#-90

hlay <-rbind(c(1, 1, 1, 1, 1, 1),
             c(NA, 2, 2, 2, 2, NA))

OLM <- grid.arrange(arrangeGrob(OLM1, OLM2, layout_matrix = hlay))
rm(OLM, OLM1, OLM2, OLM3, hlay)
```

```
## ------------------
## GAM model .-
## ==================

# Parameter configurations:

gam1 <- gam(BALANCE1 ~ s(time, bs = "cs", k = 4, fx = F, m = 2,
            by = factor(class)) + factor(class), gamma = 0.3585824,
            method = "GCV.Cp", select = F, data = ILR[ILR$data == "base", ])
gam2 <- gam(BALANCE2 ~ s(time, bs = "cr", k = 6, fx = F, m = 3,
            by = factor(class)) + factor(class), gamma = 0.4861186,
            method = "P-ML", select = T, data = ILR[ILR$data == "base", ])
gam3 <- gam(BALANCE3 ~ s(time, bs = "cs", k = 5, fx = T, m = 2,
            by = factor(class)) + factor(class), gamma = 1.136932,
            method = "GCV.Cp", select = T, data = ILR[ILR$data == "base", ])

# Balance Prediction:

pre1 <- predict(gam1, ILR, type = "link", se.fit = TRUE)
BALANCE1 <- cbind.data.frame(fit = pre1$fit, lwr = pre1$fit - (2*pre1$se.fit),
                             upr = pre1$fit + (2*pre1$se.fit), ILR = ILR$BALANCE1,
                             ILR[ ,c("time", "data", "I")], BALANCE = "BALANCE1")
rm(pre1)

pre2 <- predict(gam2, ILR, type = "link", se.fit = TRUE)
BALANCE2 <- cbind.data.frame(fit = pre2$fit, lwr = pre2$fit - (2*pre2$se.fit),
                             upr = pre2$fit + (2*pre2$se.fit), ILR = ILR$BALANCE2,
                             ILR[ ,c("time", "data", "I")], BALANCE = "BALANCE2")
rm(pre2)

pre3 <- predict(gam3, ILR, type = "link", se.fit = TRUE)
BALANCE3 <- cbind.data.frame(fit = pre3$fit, lwr = pre3$fit - (2*pre3$se.fit),
                             upr = pre3$fit + (2*pre3$se.fit), ILR = ILR$BALANCE3,
                             ILR[ , c("time", "data", "I")], BALANCE = "BALANCE3")
rm(pre3)

BALANCES2 <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES2 <- with(BALANCES2, BALANCES2[order(BALANCE, data, time), ])
rm(gam1, gam2, gam3, BALANCE1, BALANCE2, BALANCE3)

# ==============================================
# Global plot configuration
# ----------------------------------------------

fitted <- spread(BALANCES2[, c("time", "data", "BALANCE", "fit")], key=BALANCE, value=fit)
fitted <- with(fitted, fitted[order(data, time), ])
ilrinv <- as.data.frame(ilrInv(fitted[, c("BALANCE1", "BALANCE2", "BALANCE3")], V = V))
rm(fitted)

## arreglo en visualizacion de los datos

GAM.CODA <- cbind(gather(CODA[CODA$Country.Name == pais, ], key = access, value = real, urban:noru
                  fitted = gather(ilrinv, key = access, value = value, Var1:Var4)$value)
rm(ilrinv)

## arreglo en visualizacion de los datos

GAM.CODA <- GAM.CODA[!(GAM.CODA$time <
                         max(CODA$time[CODA$data == "base"]) & GAM.CODA$data == "fitted"), ]
rm(ilrinv)

# ==========================================================================
# GAM Interaction model outside the calibration range
# --------------------------------------------------------------------------

OGAM1 <- ggplot(BALANCES2, aes(time, fit, ymin = lwr, ymax = upr, colour = I)) +
  geom_line() +
  geom_smooth(stat = "identity") +
  geom_point(aes(time, ILR, colour = I)) +
  facet_wrap( ~ BALANCE, nrow = 1, scales = "free_y", dir = "v") +
  ylab("GAM␣ilr") +
  guides(colour = guide_legend(title = NULL, reverse = T, nrow = 1)) +
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(3, 5 , 2 ,1)]) +
  theme(text = element_text(size = 15),
        axis.text.x = element_text(angle = 0),legend.position = "bottom")
rm(BALANCES2)

OGAM2 <- ggplot(GAM.CODA, aes(time, real, colour = data, shape = access)) +
  geom_point() +
  geom_line(aes(time, fitted), size = .6) +
  #geom_smooth(stat = "identity")+
  ylab("%␣of␣population␣total") +
  guides(colour = guide_legend(title = "Data", reverse = F))+
  scale_shape_manual(values = c(21, 2,19, 17)) +
  scale_colour_manual(values = brewer.pal(8,"Set1")[c(2, 1, 4, 5, 6, 7)],
                      labels=c("base", "fitted")) +
  labs(shape = "Electricity␣\n␣access")+
  theme(text = element_text(size = 15))

# ================================================================
# Figure 4.3.2: ... GAM prediction outside the calibration range.
# ----------------------------------------------------------------

hlay <- rbind(c(1, 1, 1, 1, 1, 1),
             c(NA, 2, 2, 2, 2, NA))

OGAM <- grid.arrange(arrangeGrob(OGAM1, OGAM2, layout_matrix = hlay))

ggsave(file="Bangladesh_gam_p6.pdf", OGAM, width = 9, height = 7)
rm(OGAM, OGAM1, OGAM2, hlay)

## ------------------
## SVM model .-
## ==================

# Parameter configurations:

svm1 <- svm(BALANCE1 ~ time*factor(class), cost = 422.6625, gamma = 0.01174233, epsilon = 0.046131
            kernel = "radial", data = ILR[ILR$data == "base", ], type = ’eps-regression’)
svm2 <- svm(BALANCE2 ~ time*factor(class), cost = 1057.652, gamma = 0.04646314, epsilon = 0.024679
            kernel = "radial", data = ILR[ILR$data == "base", ], type = ’eps-regression’)
svm3 <- svm(BALANCE3 ~  time*factor(class), cost = 896.6357, gamma = 0.06013468, epsilon = 0.12951
            kernel = "radial", data = ILR[ILR$data == "base", ], type = ’eps-regression’)


# Balance Prediction:
```

```
BALANCE1 <- cbind( fit = predict(svm1, ILR),
                   ILR = ILR$BALANCE1, ILR[ , c("time", "data", "I")], BALANCE = "BALANCE1")
BALANCE2 <- cbind( fit=predict(svm2, ILR),
                   ILR = ILR$BALANCE2, ILR[ ,c("time", "data", "I")], BALANCE = "BALANCE2")
BALANCE3 <- cbind( fit = predict(svm3, ILR),
                   ILR = ILR$BALANCE3, ILR[ ,c("time", "data", "I")], BALANCE = "BALANCE3")

BALANCES3 <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES3 <- with(BALANCES3, BALANCES3[order(BALANCE, data, time), ])
rm(svm1, svm2, svm3, BALANCE1, BALANCE2, BALANCE3)

# ==============================================
# Global plot configuration
# ----------------------------------------------

fitted <- spread(BALANCES3[, c("time", "data", "BALANCE", "fit")], key = BALANCE, value = fit)
fitted <- with(fitted, fitted[order(data, time), ])
ilrinv <- as.data.frame(ilrInv(fitted[ , c("BALANCE1", "BALANCE2", "BALANCE3")], V = V))
rm(fitted)

## arreglo en visualizacion de los datos

SVM.CODA <- cbind(gather(CODA[CODA$Country.Name == pais,], key = access, value=real, urban:norural
                  fitted = gather(ilrinv, key = access, value=value, Var1:Var4)$value)

## arreglo en visualizacion de los datos
SVM.CODA <- SVM.CODA[!(SVM.CODA$time < max(CODA$time[CODA$data == "base"]) & SVM.CODA$data=="fitte
rm(ilrinv)

# ==========================================================================
# SVM Interaction model outside the calibration range
# --------------------------------------------------------------------------

OSVM1 <- ggplot(BALANCES3, aes(time, fit, colour = I)) +
  geom_line( ) +
  geom_smooth(stat = "identity") +
  geom_point(aes(time, ILR, colour = I)) +
  facet_wrap( ~ BALANCE, nrow = 1, scales = "free_y", dir="v") +
  ylab("SVM␣ilr") +
  guides(colour = guide_legend(title = NULL, reverse = T, nrow = 1)) +
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(3, 5, 2, 1)]) +
  theme(legend.position = "bottom") +
  theme(text = element_text(size = 15),axis.text.x = element_text(angle=0))#-90
rm(BALANCES3)

OSVM2 <- ggplot(SVM.CODA, aes(time, real, colour = data, shape = access)) +
  geom_point( ) +
  geom_line(aes(time, fitted), size = .6) +
  #geom_smooth(stat = "identity")+
  ylab("%␣of␣population␣total") +
  guides(colour=guide_legend(title="Data",reverse = F)) +
  scale_shape_manual(values = c(21, 2, 19, 17)) +
  scale_colour_manual(values = brewer.pal(8, "Set1")[c(2, 1, 4, 5, 6, 7)],
                      labels=c("base", "fitted"))+
  labs(shape = "Electricity␣\n␣access")

# ==========================================================
# Figure 4.3.3: ... SVM prediction outside the calibration range.
# ----------------------------------------------------------
hlay <- rbind(c(1, 1, 1, 1, 1, 1),
              c(NA, 2, 2, 2, 2, NA))

OSVM <- grid.arrange(arrangeGrob(OSVM1, OSVM2, layout_matrix = hlay))


ggsave(file = "Bangladesh_svm_p6.pdf", OSVM, width = 9, height = 7)
rm(OSVM1, OSVM2, hlay, OSVM)

# ============================================================================================
# Figure 4.3.1: Comparison between LM, GAM and SVM outside the calibration rank throught RMSE.
# --------------------------------------------------------------------------------------------

LM.CODA$model <- "LM"
GAM.CODA$model <- "GAM"
SVM.CODA$model <- "SVM"
FIT.O <- rbind(LM.CODA, GAM.CODA, SVM.CODA)
FIT.O <- FIT.O[FIT.O$plot == 0, ]
FIT.O$SE <- (FIT.O$real - FIT.O$fit)^2
rm(LM.CODA, GAM.CODA, SVM.CODA)

RMSE.O <- aggregate(SE ~ model + data, data = FIT.O, FUN = function(x)sqrt(sum(x)/length(x)))
RMSE.O <- RMSE.O[order(RMSE.O$SE), ]
RMSE.O$position <- 1:3
RMSE.O$position <- as.factor(RMSE.O$position)
levels(RMSE.O$position) <- RMSE.O$model
RMSE.O

G.RMSE.O <- ggplot(RMSE.O[RMSE.O$data == "fitted", ], aes(position, SE)) +
geom_bar(colour = "black", stat = "identity", fill = "#FF6666") +
labs(x = 'Fitted␣model', y = 'Root␣mean␣square␣error') +
theme(aspect.ratio = 9 / 9, text = element_text(size = 18))

ggsave(file = "Bangladesh_RMSE_outside.pdf", G.RMSE.O, width = 9, height = 7)
rm(FIT.O, G.RMSE.O)

## #####################################################################
## Exercise f)                                                         #
## ================================================================== #
##    4.4. Consequences of using standard statistical techniques over   #
##    compositional data in raw form                                    #
## ------------------------------------------------------------------ #
#######################################################################

##  ====================================================================
##      RAW MODELS.- LM
##  --------------------------------------------------------------------

RAW <- CODA[CODA$Country.Name == pais, -7]

lm1R <- lm(urban   ~ time*class, data = RAW[RAW$data == "base", ])
lm2R <- lm(nourban ~ time*class, data = RAW[RAW$data == "base", ])
lm3R <- lm(rural   ~ time, data = RAW[RAW$data == "base", ])
lm4R <- lm(norural ~ time, data = RAW[RAW$data == "base", ])

# ==============================================
# Prediccion 1.- PREDICCIONES lm().- RAW MODELS
# ----------------------------------------------
```

```
urban <- cbind(predict(lm1R, interval = "confidence", RAW), raw = RAW$urban,
                RAW[ , c("time", "data", "I", "plot", "class")], access = "urban")
nourban <- cbind(predict(lm2R, interval = "confidence", RAW), raw = RAW$nourban,
                RAW[ ,c("time", "data", "I", "plot", "class")], access = "nourban")
rural <- cbind(predict(lm3R, interval = "confidence", RAW), raw = RAW$rural,
                RAW[ ,c("time", "data", "I", "plot", "class")], access = "rural")
norural <- cbind(predict(lm4R, interval = "confidence", RAW), raw = RAW$norural,
                RAW[ ,c("time", "data", "I", "plot", "class")], access = "norural")

RAW.LM <- rbind(urban, nourban, rural, norural)
RAW.LM <- with(RAW.LM, RAW.LM[order(access, data, time),])
RAW.LM$model <- "LM"
rm(lm1R, lm2R, lm3R, lm4R, urban, nourban, rural, norural)

##  =============================================================================================
##      RAW MODELS.- GAM
##  ---------------------------------------------------------------------------------------------

gam1R <- gam(urban ~ s(time, bs = "cs", k = 5, fx = T, m = 2, by = factor(class)) +
             factor(class), gamma = 1.0517752, method = "P-REML", select = F,
             data = RAW[RAW$data == "base", ])
gam2R <- gam(nourban ~ s(time,bs = "cs", k = 5, fx = T, m = 3, by = factor(class)) +
               factor(class), gamma = 1.058473, method = "P-REML", select=F,
               data = RAW[RAW$data == "base", ])
gam3R <- gam(rural ~ s(time, bs = "cs", k = 6, fx = F, m = 3, by = factor(class)) +
             factor(class), gamma = 1.199093, method = "GACV.Cp", select = T,
             data = RAW[RAW$data == "base", ])
gam4R<- gam(norural ~ s(time, bs = "cr", k = 6, fx = F, m = 3, by = factor(class)) +
              factor(class), gamma = 0.6979624, method = "REML", select = F,
              data = RAW[RAW$data == "base", ])

# ===============================================
# Prediccion 1.- PREDICCIONES gam().- RAW MODELS
# -----------------------------------------------

pre1 <- predict(gam1R, RAW, type = "link", se.fit = TRUE)
urban <- cbind.data.frame(fit = pre1$fit, lwr = pre1$fit - (2*pre1$se.fit),
                          upr = pre1$fit + (2*pre1$se.fit), raw = RAW$urban,
                          RAW[ , c("time", "data", "I", "plot", "class")], access = "urban")
rm(pre1)

pre2 <- predict(gam2R, RAW, type = "link", se.fit = TRUE)
nourban <- cbind.data.frame(fit = pre2$fit, lwr = pre2$fit - (2*pre2$se.fit),
                            upr = pre2$fit + (2*pre2$se.fit), raw = RAW$nourban,
                            RAW[ , c("time", "data", "I", "plot", "class")], access = "nourban")
rm(pre2)

pre3<-predict(gam3R, RAW, type = "link", se.fit = TRUE)
rural <- cbind.data.frame(fit = pre3$fit, lwr = pre3$fit - (2*pre3$se.fit),
                          upr = pre3$fit + (2*pre3$se.fit), raw = RAW$rural,
                          RAW[ ,c("time", "data", "I", "plot", "class")], access = "rural")
rm(pre3)

pre4 <- predict(gam4R, RAW, type = "link", se.fit = TRUE)
norural <- cbind.data.frame(fit = pre4$fit, lwr = pre4$fit - (2*pre4$se.fit),
                            upr = pre4$fit + (2*pre4$se.fit), raw = RAW$norural,
                            RAW[ ,c("time", "data","I", "plot", "class")], access = "norural")
rm(pre4)
```

```
RAW.GAM <- rbind(urban, nourban, rural, norural)
RAW.GAM <- with(RAW.GAM, RAW.GAM[order(access, data, time), ])
RAW.GAM$model <- "GAM"
rm(gam1R, gam2R, gam3R, gam4R, urban, nourban, rural, norural)

##  ============================================================================================
##      RAW MODELS.- SVM
##  --------------------------------------------------------------------------------------------

svm1R <- svm(urban ~ time*factor(class), data = RAW[RAW$data == "base",], cost=670.2429,
             gamma = 0.06072633, epsilon = 0.0896873, coef0 = 0, kernel = "radial")

svm2R <- svm(nourban ~ time*factor(class), data = RAW[RAW$data == "base", ], cost = 627.6893,
             gamma = 0.1176374, epsilon = 0.01581494, kernel = "radial")

svm3R <- svm(rural ~ time*factor(class), data = RAW[RAW$data == "base", ], cost = 662.8796,
             gamma = 0.01353806, epsilon = 0.0238125, coef0 = 0, kernel = "radial")

svm4R <- svm(norural ~ time*factor(class), data = RAW[RAW$data == "base", ],  cost = 539.0089,
             gamma = 0.0418369, epsilon = 0.008133287, kernel = "radial")

# ==========================================================
# Prediccion 2.- PREDICCIONES SVM().- inside calibration rank
# ----------------------------------------------------------

urban <- cbind(fit = predict(svm1R, RAW), raw = RAW$urban,
               RAW[ ,c("time", "data", "I", "plot", "class")], access = "urban")

nourban <- cbind( fit = predict(svm2R, RAW), raw = RAW$nourban,
                  RAW[ ,c("time", "data", "I", "plot", "class")], access = "nourban")

rural <- cbind( fit = predict(svm3R, RAW), raw = RAW$rural,
                RAW[ ,c("time", "data", "I", "plot", "class")], access = "rural")

norural <- cbind( fit = predict(svm4R, RAW), raw = RAW$norural,
                  RAW[ , c("time", "data", "I", "plot", "class")], access = "norural")

RAW.SVM <- rbind(urban, nourban, rural, norural)
RAW.SVM <- with(RAW.SVM, RAW.SVM[order(access, data, time), ])
RAW.SVM$model <- "SVM"
rm(svm1R, svm2R, svm3R, svm4R, urban, nourban, rural, norural)

RAW.FIT <- rbind(RAW.LM[, -c(2,3)], RAW.GAM[, -c(2,3)], RAW.SVM)
RAW.FIT <- RAW.FIT[RAW.FIT$plot == 0, ]
rm(RAW.LM, RAW.GAM, RAW.SVM)

### SVM calibracion

RAW.FIT2 <- spread(RAW.FIT[ , -c(2,5,6,7)], key = access, value = fit)
RAW.FIT2$suma <- RAW.FIT2$urban + RAW.FIT2$nourban + RAW.FIT2$rural + RAW.FIT2$norural

# =========================================
# Cross validation unit-sum constraint error
# -----------------------------------------

USCE <- ggplot(RAW.FIT2, aes(time, suma, shape = data, colour = model)) +
  geom_hline(yintercept = 1, colour = "blue", size = 1) +
  geom_point(size = 3) +
  scale_shape_manual(values = c(19, 17), labels = c("Inside", "Outside")) +
  labs(shape = "Calibration␣\n␣rank", colour = "Model") +
  ylab("Unit-sum␣constraint␣error")+
  theme_classic(base_size = 18)

ggsave(file = "Bangladesh_USCE.pdf", USCE, width = 9, height = 7)

RAW.FIT$SE <- (RAW.FIT$raw - RAW.FIT$fit)^2
RMSE.RAW <- aggregate(SE ~ model + data, data = RAW.FIT, FUN = function(x)sqrt(sum(x)/length(x)))
RMSE.RAW$transformation <- "Raw"
RMSE.O$transformation <- "CoDa"
RMSE.O$position <- NULL
rm(RAW.FIT, RAW.FIT2, USCE)

RMSE.T <- rbind(RMSE.O, RMSE.RAW)
RMSE.T <- with(RMSE.T, RMSE.T[order(SE, data, model), ])
RMSE.T$position <- 1:nrow(RMSE.T)
RMSE.T$position <- as.factor(RMSE.T$position)
levels(RMSE.T$position) <- RMSE.T$model


# ===================================================================
# Figure 4.4.2: Cross validation between CoDa and Raw model inside the
#               calibration rank for Nigeria using the RMSE
# -------------------------------------------------------------------

hc1 <- ggplot(data = RMSE.T[RMSE.T$data == "base", ],
              aes(x = position, y=SE, fill = transformation)) +
  geom_bar(stat = "identity", color = "black", position = position_dodge( ), size = 1.2) +
  scale_fill_manual(values=c(’#999999’,’#E69F00’))+
  labs(x = ’Interpolation␣method’, y = ’Root␣mean␣square␣error’, fill = "Transformation") +
  theme_minimal(base_size = 18)

ggsave(file="Bangladesh_versus_inside.pdf", hc1, width = 7, height = 4)

# ====================================================================
# Figure 4.4.3: Cross validation between CoDa and Raw model outside the
#               calibration rank for Nigeria using the RMSE
# --------------------------------------------------------------------

hc2 <- ggplot(data = RMSE.T[RMSE.T$data == "fitted", ],
              aes(x = position, y = SE, fill = transformation)) +
  geom_bar(stat = "identity", color = "black", position = position_dodge( ), size = 1.2) +
  scale_fill_manual(values = c(’#999999’, ’#E69F00’)) +
  labs(x = ’Interpolation␣method’, y = ’Root␣mean␣square␣error’, fill = "Transformation") +
  theme_minimal(base_size = 18)
hc2

ggsave(file = "Bangladesh_versus_outside.pdf", hc2, width = 7, height = 4)
rm(hc1, hc2, RMSE.T, RMSE.O, RMSE.RAW)

## #################################################################
## Exercise g)                                                     #
## ============================================================== #
##    2030 forecasting                                             #
## -------------------------------------------------------------- #
###################################################################

# ===================
# Pequenos arreglos 1
# --------------------

ILR30 <- rbind(ILR, ILR)
ILR30 <- ILR30[1:32 , ]
ILR30$BALANCE1 <- ILR30$BALANCE2 <- ILR30$BALANCE3 <- NA
ILR30$class <- c(rep(0,16), rep(1, 16))
ILR30$time <- c(2015:2030, 2015:2030)
ILR30 <- rbind(ILR[ILR$plot == 0, ], ILR30)
ILR30$data <- ifelse(ILR30$time <= 2014, "base", "fitted")
ILR30$plot <- ILR30$I <- NULL

# ---

# LM2030

BALANCE1 <- cbind( predict(LM1T, interval = "confidence", ILR30), ILR = ILR30$BALANCE1,
                   ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE1")
BALANCE2 <- cbind( predict(LM2T, interval = "confidence", ILR30), ILR = ILR30$BALANCE2,
                   ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE2" )
BALANCE3 <- cbind( predict(LM3T, interval = "confidence", ILR30), ILR = ILR30$BALANCE3,
                   ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE3" )

BALANCES.LMF <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.LMF <- with(BALANCES.LMF, BALANCES.LMF[order(BALANCE, data, time), ])
BALANCES.LMF$model <- "LM"
rm(LM1T, LM2T, LM3T, BALANCE1, BALANCE2, BALANCE3)

# GAM2030

pre1 <- predict(gam1T, ILR30, type = "link", se.fit = TRUE)
BALANCE1 <- cbind.data.frame(fit = pre1$fit, lwr = pre1$fit - (2*pre1$se.fit),
            upr = pre1$fit + (2*pre1$se.fit), ILR = ILR30$BALANCE1,
            ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE1")
rm(pre1)

pre2 <- predict(gam2T, ILR30, type = "link", se.fit = TRUE)
BALANCE2 <- cbind.data.frame(fit = pre2$fit, lwr = pre2$fit - (2*pre2$se.fit),
            upr = pre2$fit + (2*pre2$se.fit), ILR = ILR30$BALANCE2,
            ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE2")
rm(pre2)

pre3 <- predict(gam3T, ILR30, type = "link", se.fit = TRUE)
BALANCE3 <- cbind.data.frame(fit = pre3$fit,lwr = pre3$fit - (2*pre3$se.fit),
            upr = pre3$fit + (2*pre3$se.fit), ILR = ILR30$BALANCE3,
            ILR30[ ,c("time", "data", "class")], BALANCE = "BALANCE3")
rm(pre3)

BALANCES.GAMF <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.GAMF <- with(BALANCES.GAMF, BALANCES.GAMF[order(BALANCE, data, time), ])
BALANCES.GAMF$model <- "GAM"
rm(gam1T, gam2T, gam3T, BALANCE1, BALANCE2, BALANCE3)

# SVM2030

BALANCE1 <- cbind( fit = predict(svm1T, ILR30[ , c("time", "class")]), ILR = ILR30$BALANCE1,
                   ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE1")
BALANCE2 <- cbind( fit = predict(svm2T, ILR30[ , c("time", "class")]), ILR = ILR30$BALANCE2,
                   ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE2")
```

```
BALANCE3 <- cbind( fit = predict(svm3T, ILR30[, c("time", "class")]), ILR = ILR30$BALANCE3,
                   ILR30[ , c("time", "data", "class")], BALANCE = "BALANCE3")

BALANCES.SVMF <- rbind(BALANCE1, BALANCE2, BALANCE3)
BALANCES.SVMF <- with(BALANCES.SVMF, BALANCES.SVMF[order(BALANCE, data, time), ])
BALANCES.SVMF$model <- "SVM"
rm(svm1T, svm2T, svm3T, BALANCE1, BALANCE2, BALANCE3)

GT <- rbind(BALANCES.LMF[ , -c(2:3)], BALANCES.GAMF[ , -c(2:3)], BALANCES.SVMF)
rm(BALANCES.LMF, BALANCES.GAMF, BALANCES.SVMF)

g1 <- ggplot(GT[GT$model != "LM", ], aes(time, fit, colour = model, shape = factor(class))) +
  geom_line(size = .8, position = ’jitter’) +
  geom_point(aes(time, ILR), colour = "blue4", size = 2) +
  facet_wrap( ˜ BALANCE, nrow = 1,scales = "free_y", dir = "v") +
  ylab("ilr") +
  guides(colour = guide_legend(title = NULL, reverse = F, ncol = 1)) +
  scale_colour_manual(values = c("skyblue3", "red3")) +
  scale_shape_manual(values = c(21, 6), labels = c("Same", "Diff.")) +
  labs(shape = "class") +
  coord_fixed( ) +
  theme_bw(base_size = 18)
g1

# Valores Predichos
PredF <- spread(GT[ , c("time", "fit", "BALANCE", "class", "model")], key = BALANCE, value = fit)
ilrinv <- ilrInv(PredF[ , 4:6] , V = V)
PredTF <- cbind(ilrinv, PredF[ , 1:3])
colnames(PredTF) <- c("urban", "rural", "nourban", "norural", "time", "class", "model")
rm(PredF, ilrinv)

# ===================
# Pequenos arreglos 2
# -------------------

Y30 <- rbind(Y, Y)
Y30 <- Y30[1:32 , ]
Y30$urban <- Y30$rural <- Y30$nourban <- Y30$norural <- NA
Y30$class <- c(rep(0,16), rep(1,16))
Y30$time <- c(2015:2030, 2015:2030)
Y30<-rbind(Y, Y30)

# ---

YT <- gather(PredTF, key = access, value = value, urban:norural)
YT2 <- cbind(YT[YT$model == "LM", ], GAM = YT[YT$model == "GAM", "value"],
             SVM = YT[YT$model == "SVM", "value"])
colnames(YT2) <- c("time", "class", "model", "access", "LM", "GAM", "SVM")
YT2$model <- NULL

Yg <- gather(Y30, key = access, value = value, urban:norural)

YT2$total <- Yg$value
YT2$class <- factor(YT2$class)
levels(YT2$class) <- c("Class␣=␣Same", "Class␣=␣Different")
rm(Yg, Y30, YT, PredTF)

g2 <- ggplot(YT2, aes(time, total, shape = access)) +
  geom_point(size = 2, colour = "blue4") +
  geom_line(aes(y = SVM, colour = "SVM"), size = .8, position = ’jitter’) +
  geom_line(aes(y = GAM, colour = "GAM"), size = .8, position = ’jitter’) +
  facet_wrap( ˜ class, nrow = 1, scales = "free_y", dir="v") +
  ylab("%␣of␣population␣total") +
  scale_shape_manual(values = c(21, 2, 19, 17)) +
  scale_colour_manual("Model:",
                      breaks = c("GAM", "SVM"),
                      values = c("skyblue3", "red3")) +
  labs(shape = "Electricity␣\n␣access") +
  theme_bw(base_size = 18)
g2

hlay <-rbind(c(1, 1, 1, 1, 1, 1),
             c(2, 2, 2, 2, 2, 2))

# ##########################################################
# Figure 4.5.1: Forecatisting of electrcity access by 2030
# ----------------------------------------------------------

p3 <- grid.arrange(arrangeGrob(g1, g2, layout_matrix = hlay))
ggsave(file="Bangladesh_2030.pdf", p3, width = 9, height = 7)
rm(p3,g1,g2,hlay,GT,YT2)

rm(CODA, ILR, ILR30, RAW, V, Y, pais, g_legend, multiplot)
```

Appendix D

Data sets

D.1 Bangladesh

urban	rural	nourban	norural	class	time	data	I	plot
0.1603866	0.0818189	0.0528934	0.7049011	1	1994	base	different	0
0.1113173	0.0852079	0.1056127	0.6978621	0	1995	base	same	0
0.1180614	0.1023988	0.1025786	0.6769612	0	1996	base	same	0
0.1783821	0.1171186	0.0459979	0.6585014	1	1997	base	different	0
0.1320362	0.1358612	0.0961438	0.6359588	0	1998	base	same	0
0.1392646	0.1520918	0.0927554	0.6158882	0	1999	base	same	0
0.1915508	0.1566405	0.0443492	0.6074595	1	2000	base	different	0
0.1544480	0.1834141	0.0865120	0.5756259	0	2001	base	same	0
0.1628356	0.1981513	0.0847244	0.5542887	0	2002	base	same	0
0.1714996	0.2125785	0.0827904	0.5331315	0	2003	base	same	0
0.2000332	0.2246134	0.0611068	0.5142466	1	2004	base	different	0
0.2214691	0.2282827	0.0466209	0.5036273	1	2005	base	different	0
0.1991574	0.2543835	0.0760126	0.4704465	0	2006	base	same	0
0.2318258	0.2626526	0.0505442	0.4549774	1	2007	base	different	0
0.2190674	0.2812544	0.0706126	0.4290656	0	2008	base	same	0
0.2318258	0.2626526	0.0505442	0.4549774	1	2007	fitted	diff. forecast	1
0.2190674	0.2812544	0.0706126	0.4290656	0	2008	fitted	same forecast	1
0.2295073	0.2943449	0.0675827	0.4085651	0	2009	fitted	same forecast	0
0.2744626	0.2954670	0.0301574	0.3999130	1	2010	fitted	diff. forecast	0
0.2816495	0.3390608	0.0306005	0.3486893	1	2011	fitted	diff. forecast	0
0.2629793	0.3318139	0.0569207	0.3482861	0	2012	fitted	same forecast	0
0.2882264	0.3644787	0.0393036	0.3079913	1	2013	fitted	diff. forecast	0
0.3039901	0.3417278	0.0311699	0.3231122	1	2014	fitted	diff. forecast	0

D.2 India

urban	rural	nourban	norural	class	time	data	I	plot
0.2168615	0.2856408	0.0450485	0.4524492	1	1993	base	different	0
0.2269190	0.2828653	0.0370710	0.4531447	0	1994	base	same	0
0.2305576	0.2938792	0.0355124	0.4400508	0	1995	base	same	0
0.2342571	0.3047239	0.0339129	0.4271061	0	1996	base	same	0
0.2380095	0.3153803	0.0322705	0.4143397	0	1997	base	same	0
0.2418108	0.3258257	0.0305892	0.4017743	0	1998	base	same	0
0.2506459	0.3489511	0.0238841	0.3765189	1	1999	base	different	0
0.2495396	0.3460828	0.0271304	0.3772472	0	2000	base	same	0
0.2445617	0.3142775	0.0346183	0.4065425	1	2001	base	different	0
0.2579812	0.3650189	0.0244588	0.3525411	0	2002	base	same	0
0.2627159	0.3741471	0.0230041	0.3401329	0	2003	base	same	0
0.2607838	0.3832162	0.0282462	0.3277538	0	2004	base	same	0
0.2720415	0.3922780	0.0203085	0.3153720	0	2005	base	same	0
0.2752874	0.3923007	0.0204026	0.3120093	1	2006	base	different	0
0.2815929	0.4105075	0.0174671	0.2904325	0	2007	base	same	0
0.2864735	0.4196790	0.0159865	0.2778610	0	2008	base	same	0
0.2752874	0.3923007	0.0204026	0.3120093	1	2006	fitted	diff. forecast	1
0.2864735	0.4196790	0.0159865	0.2778610	0	2008	fitted	same forecast	1
0.2939411	0.4581258	0.0119289	0.2360042	1	2009	fitted	diff. forecast	0
0.2907420	0.4537899	0.0185580	0.2369101	1	2010	fitted	diff. forecast	0
0.2905540	0.3834799	0.0222060	0.3037601	1	2011	fitted	diff. forecast	0
0.3039739	0.4970426	0.0123361	0.1866474	1	2012	fitted	diff. forecast	0
0.3124037	0.4649715	0.0075363	0.2150885	0	2013	fitted	same forecast	0
0.3180586	0.4736340	0.0056014	0.2027060	0	2014	fitted	same forecast	0

D.3 Kenya

urban	rural	nourban	norural	class	time	data	I	plot
0.0749912	0.0280007	0.1014588	0.7955493	1	1993	base	different	0
0.0753547	0.0223203	0.1041653	0.7981597	0	1994	base	same	0
0.0822865	0.0256515	0.1003435	0.7917185	0	1995	base	same	0
0.0892422	0.0288506	0.0965478	0.7853594	0	1996	base	same	0
0.0962193	0.0318929	0.0927607	0.7791271	0	1997	base	same	0
0.0913045	0.0347345	0.1009155	0.7730455	1	1998	base	different	0
0.1102290	0.0374190	0.0852710	0.7670810	0	1999	base	same	0
0.1172522	0.0399241	0.0816678	0.7611559	0	2000	base	same	0
0.1242715	0.0423217	0.0781185	0.7552883	0	2001	base	same	0
0.1312771	0.0446605	0.0746329	0.7494295	0	2002	base	same	0
0.1051590	0.0363639	0.1043210	0.7541561	1	2003	base	different	0
0.1452126	0.0493524	0.0678874	0.7375476	0	2004	base	same	0
0.1521259	0.0518002	0.0646241	0.7314498	0	2005	base	same	0
0.1589949	0.0543757	0.0614551	0.7251743	0	2006	base	same	0
0.1658176	0.0571106	0.0583824	0.7186894	0	2007	base	same	0
0.1726054	0.0599853	0.0553946	0.7120147	0	2008	base	same	0
0.1051590	0.0363639	0.1043210	0.7541561	1	2003	fitted	diff. forecast	1
0.1726054	0.0599853	0.0553946	0.7120147	0	2008	fitted	same forecast	1
0.1520805	0.0622218	0.0797495	0.7059482	1	2009	fitted	diff. forecast	0
0.1371832	0.0512074	0.0985268	0.7130826	1	2010	fitted	diff. forecast	0
0.1928974	0.0691360	0.0467726	0.6911940	0	2011	fitted	same forecast	0
0.1996863	0.0722546	0.0440137	0.6840454	0	2012	fitted	same forecast	0
0.2065088	0.0753572	0.0412912	0.6768428	0	2013	fitted	same forecast	0
0.1723475	0.0942518	0.0796225	0.6537782	1	2014	fitted	diff. forecast	0

D.4 Nigeria

urban	rural	nourban	norural	class	time	data	I	plot
0.2217454	0.1042073	0.0800146	0.5940327	0	1991	base	same	0
0.2263702	0.1111070	0.0803998	0.5821230	0	1992	base	same	0
0.2311074	0.1178674	0.0807126	0.5703126	0	1993	base	same	0
0.2359526	0.1244655	0.0809574	0.5586245	0	1994	base	same	0
0.2409028	0.1308775	0.0811472	0.5470725	0	1995	base	same	0
0.2459546	0.1370796	0.0812954	0.5356704	0	1996	base	same	0
0.2510939	0.1430590	0.0813761	0.5244710	0	1997	base	same	0
0.2563173	0.1487919	0.0814127	0.5134781	0	1998	base	same	0
0.2891827	0.1832918	0.0538573	0.4736682	1	1999	base	different	0
0.2669994	0.1595152	0.0814006	0.4920848	0	2000	base	same	0
0.2731870	0.1638437	0.0835030	0.4794663	0	2001	base	same	0
0.2795194	0.1679549	0.0855606	0.4669651	0	2002	base	same	0
0.3171524	0.2117367	0.0564076	0.4147033	1	2003	base	different	0
0.2926206	0.1756796	0.0894994	0.4422004	0	2004	base	same	0
0.2993917	0.1793687	0.0913483	0.4298913	0	2005	base	same	0
0.3063211	0.1829830	0.0931089	0.4175870	0	2006	base	same	0
0.3134195	0.1865415	0.0947705	0.4052685	0	2007	base	same	0
0.3536330	0.1830557	0.0633870	0.3999243	1	2008	base	different	0
0.3134195	0.1865415	0.0947705	0.4052685	0	2007	fitted	same forecast	1
0.3536330	0.1830557	0.0633870	0.3999243	1	2008	fitted	diff. forecast	1
0.3281600	0.1934120	0.0977200	0.3807080	0	2009	fitted	same forecast	0
0.3469704	0.1972548	0.0878296	0.3679452	1	2010	fitted	diff. forecast	0
0.3863930	0.1975149	0.0572270	0.3588651	1	2011	fitted	diff. forecast	0
0.3516010	0.2028690	0.1007390	0.3447910	0	2012	fitted	same forecast	0
0.3853458	0.1854366	0.0755942	0.3536234	1	2013	fitted	diff. forecast	0
0.3679789	0.2085425	0.1014411	0.3220375	0	2014	fitted	same forecast	0

D.5 Sudan

urban	rural	nourban	norural	class	time	data	I	plot
0.2299422	0.0924143	0.0913278	0.5863157	0	1993	base	same	0
0.2294503	0.0966453	0.0923397	0.5815647	0	1994	base	same	0
0.2289504	0.1008031	0.0933696	0.5768769	0	1995	base	same	0
0.2284342	0.1048690	0.0944058	0.5722910	0	1996	base	same	0
0.2278991	0.1088185	0.0954709	0.5678115	0	1997	base	same	0
0.2273380	0.1126316	0.0965620	0.5634684	0	1998	base	same	0
0.2267439	0.1162993	0.0976761	0.5592807	0	1999	base	same	0
0.2261175	0.1198490	0.0988325	0.5552010	0	2000	base	same	0
0.2254549	0.1233234	0.1000251	0.5511966	0	2001	base	same	0
0.2247539	0.1267637	0.1012561	0.5472263	0	2002	base	same	0
0.2240128	0.1302108	0.1025272	0.5432492	0	2003	base	same	0
0.2232295	0.1337054	0.1038405	0.5392246	0	2004	base	same	0
0.2224025	0.1372883	0.1051975	0.5351117	0	2005	base	same	0
0.2215303	0.1410000	0.1065997	0.5308700	0	2006	base	same	0
0.2206129	0.1448699	0.1080471	0.5264701	0	2007	base	same	0
0.2196567	0.1488836	0.1095333	0.5219264	0	2008	base	same	0
0.2196567	0.1488836	0.1095333	0.5219264	0	2008	fitted	same forecast	1
0.2177885	0.1571119	0.1130115	0.5120881	0	2010	fitted	same forecast	0
0.2168991	0.1612688	0.1149809	0.5068512	0	2011	fitted	same forecast	0
0.2160563	0.1654140	0.1170937	0.5014360	0	2012	fitted	same forecast	0
0.2152691	0.1695213	0.1193309	0.4958787	0	2013	fitted	same forecast	0