SORT

Statistics and Operations Research Transactions

Volume
46

Number 2, July-December 2022

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Statistics and Operations Research Transactions

Volume 46, Number 2, July-December 2022

ISSN: 1696-2281
eISSN: 2013-8830

Articles

46002

9 771696 228009

www.idescat.cat/sort/

Aims

SORT (*Statistics and Operations Research Transactions*) -formerly *Qüestiió*- is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-edited by the Universitat Politècnica de Catalunya, Universitat de Barcelona, Universitat Autònoma de Barcelona, Universitat de Girona, Universitat Pompeu Fabra, Universitat de Lleida i Universitat Rovira i Virgili and the cooperation of the Spanish Section of the International Biometric Society and the Catalan Statistical Society. SORT promotes the publication of original articles of a methodological or applied nature or motivated by an applied problem in statistics, operations research, official statistics or biometrics as well as book reviews. We encourage authors to include an example of a real data set in their manuscripts. SORT is an Open Access journal which does not charge publication fees.

SORT is indexed and abstracted in the *Science Citation Index Expanded* and in the *Journal Citation Reports* (Clarivate Analytics) from January 2008. The journal is also described in the *Encyclopedia of Statistical Sciences* and indexed as well by: *Current Index to Statistics*, *Índice Español de Ciencia y Tecnología*, *MathSci*, *Current Mathematical Publications and Mathematical Reviews*, and *Scopus*.

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

Editor in Chief

Executive Editors

Production Editor

Layout manager

Responsible for the Secretary of SORT

Editorial Advisory Committee

Institut d'Estadística de Catalunya

The mission of the Statistical Institute of Catalonia (Idescat) is to provide high-quality and relevant statistical information, with professional independence, and to coordinate the Statistical System of Catalonia, with the aim of contributing to the decision making, research and improvement to public policies.

Secretary and subscriptions to SORT

Institut d’Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 557.30.76 - 93 557.30.00
Fax. +34 - 93 557.30.01
E-mail: sort@idescat.cat

Publisher: Institut d’Estadística de Catalunya (Idescat)

ISSN 1696-2281
eISSN: 2013-8830
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)

www.idescat.cat/sort/

ISSN: 1696-2281
eISSN: 2013-8830
SORT 46 (2) July-December (2022)

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Universitat Rovira i Virgili
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística

Generalitat
de Catalunya
Institut d'Estadística
de Catalunya

SORT

Volume 46
Number 2
July-December 2022
ISSN: 1696-2281
eISSN: 2013-8830

Articles

SORT 46 (2) July-December 2022, 161-188 DOI: 10.2436/20.8080.02.122

Granger causality and time series regression for modelling the migratory dynamics of influenza into Brazil

Aline Foerster Grande[1], Guilherme Pumi[1,2] and Gabriela Bettella Cybis[1]

Abstract

In this work we study the problem of modelling and forecasting the dynamics of the influenza virus in Brazil at a given month, from data on reported cases and genetic diversity collected from previous months, in other locations. Granger causality is employed as a tool to assess possible predictive relationships between covariates. For modelling and forecasting purposes, a time series regression approach is applied considering lagged information regarding reported cases and genetic diversity in other regions. Three different models are analysed, including stepwise time series regression and LASSO.

MSC: *62P10, 62J05, 62J07, 92D30.*

Keywords: *Flu, time series regression, variable selection, genetic diversity, Granger causality.*

1. Introduction

Caused by the influenza virus, the flu is one of the most prevalent diseases in Brazil and worldwide, infecting about 10% of the world's population every year and causing a toll between 250,000 and 500,000 deaths annually (Barr et al., 2010; Rambaut et al., 2008). It is characterized by an acute infection of the respiratory system. Common symptoms are cough, fever, headaches, throat and muscle pain (Eccles, 2005; Rambaut et al., 2008). Due to its severity, the World Heath Organization (WHO) actively surveys the virus through the Global Influenza Surveillance and Response System (GISRS) Network. The patterns of influenza incidence are influenced by seasonality and the emergence of new variants – new types of virus that infect humans for the first time thus managing to spread further due to reduced immunity in the population. According to the

[1] Programa de Pós-Graduação em Estatística - Universidade Federal do Rio Grande do Sul.

[2] Corresponding author.

Received: January 2022
Accepted: May 2022

WHO, extensive vaccination against influenza is the most effective measure for its prevention (Barr et al., 2010). Public vaccination policies, therefore, become a fundamental agent in preventing serious epidemics and reducing the death toll from influenza.

Severe influenza cases require hospitalization and intensive care, including the need for artificial respirators. With the emergence of COVID-19, such precious assets have become scarce in many countries. Thus, moving forward, the forecast of influenza cases can help guide public health care systems in allocating resources and planning for seasonal concomitance of the two diseases.

A global dispersion process is responsible for seeding new variants that drive yearly influenza epidemics through most of the world. A frequent pattern is that new lineages affect the northern hemisphere first during the winter season. These variants tend to arrive latter in regions of the southern hemisphere such as South America and Oceania (Lemey et al., 2014; Petrova and Russell, 2018). This movement, if mathematically well described and statistically well modelled, has the potential to allow for predictions for the incidence of influenza, as well as the description of the strains expected to circulate in Brazil from data collected in Europe, Asia, and the United States during the winter season in the northern hemisphere. Such a forecast could be of great value for planning and implementation of public vaccination policies to reduce potential epidemics and minimize deaths due to influenza in Brazil.

In this paper, we study the problem of forecasting the number of influenza cases in Brazil at a given time t from data on influenza cases, as well as data related to genetic diversity, collected in other regions of the globe in preceding months.

2. The influenza virus

There are three common types of influenza viruses, influenza A, B and C, the first two being responsible for seasonal epidemics. The evolutionary dynamics of influenza A is composed by rapid mutation, natural selection and frequent rearrangement (Rambaut et al., 2008). Of the three types of viruses, type A is the one with the highest replication capacity in humans. Most of its cases occur in winter and in countries with temperate climates.

Influenza A type is subdivided into subtypes according to their surface proteins hemagglutinin (H1, H2 and H3) and neuraminidase (N1, N2). In this study, our goal is to investigate the behaviour of the two most recurrent subtype of influenza A, H1N1 and H3N2.

The H1N1 subtype appeared in 1918 causing the Spanish flu pandemic, one of the most deadly pandemics in history, affecting about a quarter of the world's population and responsible for tens of millions of deaths (Garten et al., 2009). The H1N1 flu virus reappeared in 1977 and subsequently its epidemics showed lower mortality rates when compared to the H3N2 epidemics (Rambaut et al., 2008). Then, in 2009 a pandemic of H1N1 occurred, widely known as the swine flu pandemic. The virus was first reported in Mexico, spreading across the world in the following months and infecting anywhere between 700 million and 1.4 billion of people (Rambaut and Holmes, 2009). After the 2009 pandemic, the H1N1 virus continued circulating, being responsible for annual seasonal outbreaks with high mortality rates in Brazil. The new phylogenetic groups (of origin) of the H1N1 virus, seem to appear in the northern hemisphere, arriving in Brazil only in the seasonal outbreak of the following year (Silva, 2015).

The H3N2 subtype emerged in 1968 as the third pandemic of the 20th century called the Hong Kong flu and has dominated seasonal influenza A virus epidemics in recent years (Ibiapina, Costa and Faria, 2005). Born et al. (2016) found that the strains of the seasonal influenza A(H3N2) epidemics in South America are powered by a continuous introduction of viral variants from other geographic regions, especially from North America, and an extensive viral exchange among South American countries. They also found that the subtype tends to arrive in Brazil from neighbouring countries in South America, mainly through its south-east region.

2.1. Migratory dynamics

The source-sink model for global flu circulation states that tropical regions are the origin of new seasonal mutations. Genetic diversity is generated in these original populations, and then advances to the northern and southern hemispheres. Additionally, China is identified as the most likely epicentre of the flu A virus (Rambaut et al., 2008). More recent phylogeographic studies have found that there is substantially more viral flow between locations, and that the pattern does not adhere strictly the source-sink model. However the trunk of the phylogenetic tree, which represents the viral lineage that persists over time, was placed reliably, most of the time, in China, Southeast Asia and India. Viruses circulating in other locations do not usually last more than a season or two before being replaced by new lineages originating from the trunk (Petrova and Russell, 2018; Lemey et al., 2014; Bedford et al., 2010). Furthermore, strains are generally first spread to North America and Europe and only later to South America (Russell et al., 2008).

Influenza epidemics in temperate regions of the northern hemisphere typically occur between the months of November and March and in the southern hemisphere from May to September. Seasonality patterns in the tropics vary more according to location (Petrova and Russell, 2018). In Brazil, Almeida, Codeço and Luz (2018) identified that different regions have varying seasonality patterns, with stronger seasonality signals closer to the coast, peaks happening earlier towards the north of the country and later in the year in the southern region. Born (2013) studied the phylogeography of influenza in Brazil, identifying as unlikely that the origin of a new variant be located in Brazil. Additionally the main gateway for the H3N2 flu virus in the country would be the Southeast region followed by the South and Northeast regions. The existence of such global patterns which are repeated somewhat consistently throughout the years can be seen as motivation for seeking to forecast Brazilian incidence as a function of reported cases in other countries in the previous months.

3. Methods

3.1. Granger causality

In this study, the Granger causality method will be used to study the migratory dynamics of influenza (Granger, 1969). This method aims to determine the causal direction between two variables, stipulating that X_t Granger-causes Y_t if past values of X_t help to predict the present value of Y_t, and may provide better results than considering only the past Y_t. More specifically, it is a way of verifying whether a time series helps in predicting another series through VAR modelling. To use this method, the series need to be matched.

3.1.1. Vector Autoregressive Models (VAR)

The vector autoregressive model (VAR) is an extension of the autoregressive models (AR) and its objective is to model a vector time series considering only their past values (Sims, 1980). Mathematically, a k-dimensional $\boldsymbol{Y}_t$ stochastic process is said to be a VAR(p) process if it can be written as

$$\boldsymbol{Y}_t = c + A_1\boldsymbol{Y}_{t-1} + A_2\boldsymbol{Y}_{t-2} + \cdots + A_p\boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where $c \in \mathbb{R}^k$ is a vector of constants (intercepts), $A_1, \cdots, A_p$ are $k \times k$ matrices and $\boldsymbol{\varepsilon_t}$ is a k-dimensional error term.

3.1.2. Granger causality

The idea behind Granger causality (for univariate time series) is to consider the model

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i Y_{t-i} + \sum_{j=1}^{m} \alpha_j X_{t-j} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{\varepsilon}_t$ denotes white noise. We say that X_t Granger-causes Y_t if past values of X_t help to predict the Y_t. In view of (1), to test whether X_t Granger-causes Y_t the following test can be performed:

$$H_0 : \alpha_1 = \cdots = \alpha_m = 0 \text{ vs. } H_1 : \alpha_s \neq 0, \text{ for at least one } s \in \{1, \cdots, m\}.$$

In the above test, rejection of the null hypothesis is considered evidence that X_t Granger-causes Y_t.

3.1.3. Granger Causality and Stationarity

Before applying the Granger causality method, it is necessary to check whether the series are stationary or not. A preliminary graphical analysis can assist in this matter. The absence of visible deterministic trends and/or apparent seasonality are indications of stationary behaviour. However, they are usually not enough for decision making, which should preferably be done through appropriate tests such as the widely applied Augmented Dickey-Fuller (Dickey and Fuller, 1979) or the Phillips-Perron (Phillips and Perron, 1988) test. In these tests, the null hypothesis is that the time series has at least one unit root (i.e., the series is non-stationary) and the alternative hypothesis is the absence of unit roots. In this way, the time series will be considered stationary if the null hypothesis is rejected.

3.1.4. Granger causality for non-stationary series

One way to apply the Granger causality method if the series are not stationary is to use the Toda and Yamamoto procedure, introduced by Toda and Yamamoto (1995), which comprises the following steps:

1. Check whether the series cointegrate. Two series cointegrate if they have the same integration order, say m, and if the residual of regression from one series to the other are stationary, which can be determined using a test such as the Phillips-Perron.

2. Adjust a VAR(p) model.

3. Apply the Wald Test. In order to do so, it is necessary to fit a VAR($p+m$) model to the data. This model will certainly present several non-significant variables, given the previous steps, but this is not a problem since this model will not be used directly - it is only a device to guarantee the asymptotic theory. Rejection of the null hypothesis is evidence towards the existence of Granger causality in the tested direction.

Granger causality is a concept applied in many fields. In economics, Farias and Sáfadi (2010) employed Granger causality to study the relationship among the main stock exchanges in the world, showing how markets behave with each other and analyzing whether a market has a strong influence on the others. In agronomy, Diniz et al. (2009) studied whether certain agricultural and socio-economic variables (such as cattle and demographic density) Granger-cause deforestation in the Amazon. In biology, Chen et al. (2018) study the causal relationship between cases of influenza in humans and air pollution in Taiwan. The results indicated that pollution Granger-causes flu cases in the elderly group (over 64 years old).

3.2. Variable selection in regression models

Variable selection is a central topic in regression models involving many covariates. In this section we review some of the available techniques for variable selection which will be used here.

3.2.1. Stepwise regression

Stepwise regression is an automatic tool that aims to select the most influential independent variables in a given model. It is an iterative method that adds or removes variables according to a given selection criterion. The most popular types of stepwise regression are the forward-stepwise and backward-stepwise. In this paper, the backward-stepwise selection method was preferred due to the model size. We consider the p-value based stopping criterion, which selects variables according to their Wald statistics, eliminating non-significant terms based on the magnitude of their p-values (higher p-values are preferred in removing terms), in order to obtain a model for which all variables are significant.

3.2.2. LASSO

The LASSO (least absolute shrinkage and selection operator) regression is a penalty method that aims to provide smaller and more parsimonious models (Hastie, Tibshirani and Friedman, 2009). The penalty is applied to the coefficients to decrease the number of parameters and, consequently, reduce the dimension and uncertainty in the model. It is a regression method that aims to reduce the dimensionality and improve the accuracy of the forecast and the interpretability of the resulting model.

4. Data

In this section we provide detailed information regarding the data used in our study. The Supplementary Material presents a detailed exploratory analysis of the data.

4.1. Number of positive flu cases

The data for number of positive flu cases was taken from FluNet, an online tool maintained by the World Health Organization (WHO, 2020) whose objective is to aggregate influenza virological surveillance data, launched in 1997. FluNet data comes from weekly country reports of the number of tested cases, the number of positive cases and the type of virus. Typically the reported data refer to data collected in a few reference centres in each country, and do not represent the actual flu incidence data. Since the number of positive cases are expected to correlate with influenza incidence, for the purpose of this paper it is considered as a proxy for incidence. Thus, such data will be referred to here as influenza incidence data. For this project, data from H1N1 and H3N2 influenza were collected from January 2008 to November 2019, but due to missing data problems in 2008, the data used in the analysis cover the period from October 2008 to November 2019. However, for modelling purposes, we only use data from October 2008 to December 2018, which yields a sample size $n = 123$, while data from January 2019 to November 2019 is reserved for out-of-sample forecasting purposes.

Table 1. *Aggregated regions.*

Region	Countries
Europe	Belgium, Switzerland, Spain, Estonia, Germany, Ireland, Israel, Italy, Latvia, Netherlands, Norway, Poland, Russian Federation, Slovenia, Sweden, Turkey, Denmark, United Kingdom of Great Britain, Northern Ireland
North America	Canada and United States
South America	Argentina, Bolivia, Chile, Colombia, Ecuador, French Guiana, Paraguay, Peru
Central America	Costa Rica, Cuba, Dominican Republic, El Salvador, Guatemala, Honduras, Jamaica, Mexico, Nicaragua, Panama
South Asia	India, Thailand, Indonesia, Bangladesh, Bhutan, Nepal, Sri Lanka
Western Pacific	China, Japan, Australia, Republic of Korea, Singapore, Malaysia, Vietnam, New Caledonia, Philippines, Cambodia, Lao People's Democratic Republic

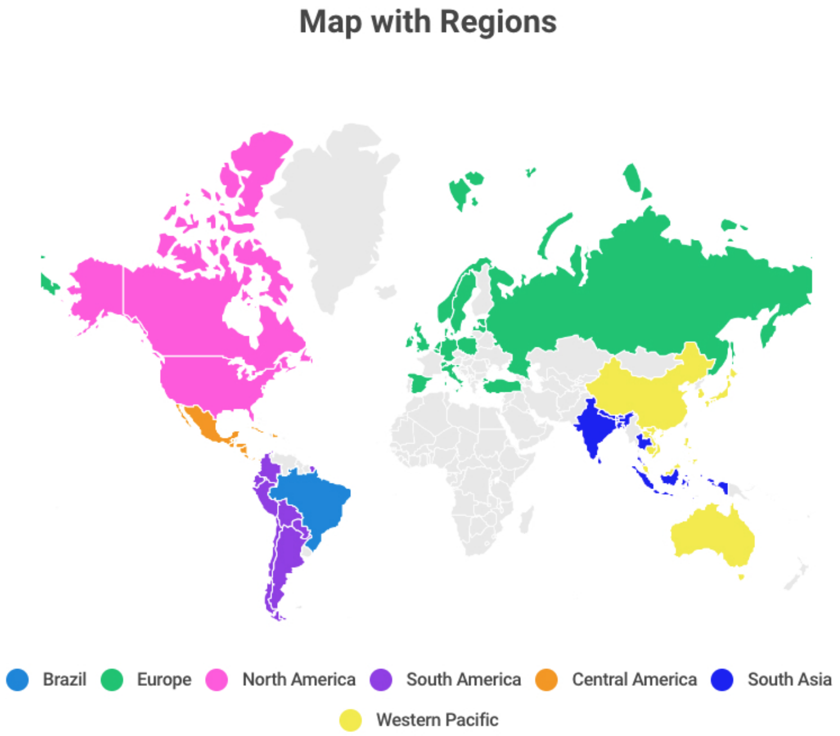


Figure 1. *Map with the regions considered in the study.*

For simplicity, we aggregate data geographically into regions based on WHO regions, with the exception of Brazil which is the focus of this study. The following locations were considered: Brazil, North America, South America (without Brazil), Central America, Europe, South Asia and Western Pacific. Note that each region is composed of a certain number of countries, responsible for reporting their data in the database. However, due to local characteristics, several countries had a high amount of missing data, often above 50%, resulting in useless local data for our purposes. In order to make the analysis feasible, all countries presenting more than 50% missing data were excluded in the construction of the database for the respective region. Table 1 and Figure 1 present the configuration of each region after applying this criterion.

The time series techniques that we used require that the time series do not contain any missing data. To resolve this, we applied an imputation method, which aims to fill in the missing data using the following criteria: multiply the average regional number of positive flu cases of the respective week by the proportion that the respective country represents in the region. In some cases, however, it happened that the week had missing data in all countries, and this was resolved by imputing it through the average between the previous and subsequent weeks. After imputation, the data were aggregated monthly.

4.2. Genetic diversity

For the genetic diversity data, viral RNA sequences were collected from the NCBI Influenza Virus Database, which compiles a comprehensive assortment of influenza sequences generated by research groups around the world (NCBI, 2020). The genetic dataset was assembled considering all complete chromosome 4 (hemagglutinin gene) sequences from human influenza A viruses in the database, from all continents and in the interval from October 2008 to September 2019. The data were retrieved in March 08, 2020. This resulted in a total of 16,008 H1N1 sequences and 15,418 H3N2 sequences. As the goal was to measure genetic diversity of viral populations, H1N1 and H3N2 sub-types were treated separately. Influenza B sequences were excluded from this study because of insufficient data at many time points.

Genetic diversity is a population measure that seeks to quantify viral variability, allowing for comparisons over time or between populations. For the sake of simplicity, throughout this paper let the genetic diversity of a viral population be defined as the average genetic distance between all sequences in the population. The distance measure considered here is the K80 distance (Kimura, 1980), which is based on nucleotide substitutions, thus the sequences must be aligned so that individual mutations can be identified. Due to the size of our dataset, the online tool MAFFT (Yamada, Tomii and Katoh, 2016) was used to generate the alignments.

The aligned sequences were then used to build a distance matrix between individual sequences in the dataset. This resulted in a symmetric $n \times n$ matrix $\mathbf{D}$ with entries $d_{i,j}$ denoting the genetic distances between the sequences i and j, where $i, j \in \{1, \cdots, n\}$.

Finally, genetic diversity was computed for temporally and geographically defined sub-populations, by averaging over all relevant entries in the distance matrix. As suggested in Jesus (2018), virus diversity was assessed using a quarterly moving average scheme, since a three-month window size best captured smooth diversity fluctuations over time for these data. This calculation was made for each month in the range from October 2008 to September 2019, separately for H1N1 and H3N2, and for the following regions: Asia, North America and global (all continents). Other regions were not considered due to insufficient data. See the Supplementary Material for a list of the countries comprising each region. Similarly to the incidence data, for modelling purposes we only consider data from October 2008 to December 2018 ($n = 123$), while data from January 2019 to November 2019 are reserved for out-of-sample forecasting purposes. All code was written in R (version 4.0.0, R Core Team, 2020) and is available (along with the relevant data) at github.com/AlineFoersterGrande/Flu_Paper.

5. Results

In this section we present the results of our analysis. We separate the different analyses by technique.

5.1. Granger causality

5.1.1. Positive influenza counts

In this section we present a Granger causality analysis of the number of positive flu cases in Brazil considering data from the other regions. Our main interest is to verify if the number of cases in Brazil can be explained by the recent historical data from other regions. In all situations, the data were considered non-stationary due to the clear seasonal pattern present (the time series plots are presented in the Supplementary Material). The global task resulted in 66 comparisons. The p-values presented in this section were corrected for false discovery rate, implemented through the function `p.adjust` in R (R Core Team, 2020). On performing step 2 of Toda and Yamamoto's procedure, we consider $p = 6$ as the maximum lag to adjust the VAR(p) model to the data.

Table 2. *Granger causality results for the number of flu cases - Brazil case.*

Null hypothesis	p-value	Lag
North America Region does not Granger-cause Brazil	0.84	–
European Region does not Granger-cause Brazil	**0.10**	3
Central America Region does not Granger-cause Brazil	0.99	–
South America Region does not Granger-cause Brazil	**0.05**	2
South Asia Region does not Granger-cause Brazil	0.78	–
Western Pacific Region does not Granger-cause Brazil	0.98	–

Table 2 presents the results of the Granger causality analysis. We conclude that among all regions, only Europe (at 10% significance) and South America (at 5% significance) Granger-cause Brazil. This result suggests that the historical data on the number of cases of flu in the European and South America Regions are helpful in predicting the present value of the incidence in Brazil.

As presented in Section 2.1, the trunk of the phylogenetic tree and source of most seasonal variation for influenza is associated to the Asian continent, particularly China, from where the virus frequently migrates to the northern hemisphere. With this in mind, we perform a Granger causality analysis to verify if the Western Pacific Region (the region that contains the majority of the data from Asia), Granger-causes the regions in the northern hemisphere.

Table 3 presents the significant cases. We find that the Western Pacific Region Granger-causes the regions of Europe and South Asia. Therefore, it can be said that the historical data related to the number of flu cases in the Western Pacific Region helps to predict the present incidence of the South Asian and European Regions.

Table 3. *Granger causality results for the number of flu cases - Western Pacific case.*

Null hypothesis	p-value	Lag
Western Pacific does not Granger-cause South Asia	0.05	3
Western Pacific does not Granger-cause Europe	0.04	3

Note that, although there is no direct evidence that the number of flu cases in the Western Pacific Region Granger-causes the incidence in Brazil, there is an indirect effect of the Pacific Region in Brazil, since the Pacific Granger-causes the European Region which in turn, Granger-causes the incidence in Brazil. This indirect effect was not directly detected because the Granger causality analysis is not transitive. Finally, we investigate whether any region Granger-causes another region. The results are presented in Table 4.

Table 4. *Granger causality results for the number of flu cases - all regions.*

Null hypothesis	p-value	Lag
Central America does not Granger-cause Europe	0.05	3
Central America does not Granger-cause Western Pacific	0.00	3

Note that the results in Table 4 indicate that the incidence in Central America Granger-causes the incidence of the European and Western Pacific Regions. The most likely justification for the Granger causality of Central America in other regions is the occurrence of the swine flu (H1N1) in the years 2009 and 2010. Mexico is considered the origin of the swine flu pandemic, which justifies the increase in the incidence of influenza first in the Central America and then in the other regions.

5.1.2. Genetic diversity

In this section, the data described in subsection 4.2 (genetic diversity) are used to examine the influence among different regions under the prism of Granger causality. The initial interest is to verify whether the number of flu cases in Brazil can be explained by the past genetic diversity data from other regions. Granger causality tests were used to assess whether H1N1 and H3N2 genetic diversity in North America, Asia and around the globe (termed All) affect Brazilian incidence. The results are shown in Table 5.

Table 5. *Granger causality results considering genetic diversity data as covariate and number of cases in Brazil as response.*

Null hypothesis	p-value
North America (H1N1) does not Granger-cause Brazil	0.85
All (H1N1) does not Granger-cause Brazil	0.84
Asia (H1N1) does not Granger-cause Brazil	0.84
North America (H3N2) does not Granger-cause Brazil	0.69
All (H3N2) does not Granger-cause Brazil	0.42
Asia (H3N2) does not Granger-cause Brazil	0.76

It can be seen that no genetic diversity Granger-causes Brazil, that is, the measurement of genetic diversity does not help in predicting the present value of the incidence of influenza in Brazil. Given these results, a second analysis was performed to verify whether the incidence of other regions can be explained by the genetic diversity data. Table 6 presents the all statistically significant results.

Table 6. *Granger causality results considering the genetic diversity data as covariate and number of cases or genetic data in other regions as responses.*

Null hypothesis	p-value	Lag
North America (H1N1) does not Granger-cause South Asia (cases)	0.05	3
Asia (H1N1) does not Granger-cause Central America (cases)	0.00	6
North America (H1N1) does not Granger-cause Asia (H1N1)	0.01	2
All (H1N1) does not Granger-cause North America (H1N1)	0.00	2
All (H1N1) does not Granger-cause Asia (H1N1)	0.00	2

We conclude that the genetic diversity of Asia (H1N1) helps in predicting the incidence of influenza in the Central American Region. Furthermore, it shows that North American genetic diversity (H1N1) Granger-causes the South Asian cases and H1N1 diversity on Asia.

5.2. Time series regression approach

In this section we present time series regression analysis of the data presented in Sections 4.1 and 4.2. A classical ARMA approach to model incidence data is presented in the Supplementary Material.

5.2.1. Number of positive flu cases

We start by considering the data described in Section 4.1 (number of positive flu cases) to represent the flu incidence in Brazil (denoted by B_t), in Europe (E_t), in North America (A_t), in Central America (C_t), in South America (S_t), in South Asia (s_t) and in Western

Pacific (W_t) at time t. We applied historical data of the regions considered in the last 11 months (lags), denoted by $B_{t-1},\cdots,B_{t-11}$ for Brazil, $A_{t-1},\cdots,A_{t-11}$ for the North America and similarly for other regions. We also considered a covariate μ_t representing the monthly average number of positive flu cases in Brazil at time t, $t \in \{1,\cdots,123\}$, calculated as

$$\mu_t = \frac{1}{\#I_t} \sum_{k \in I_t} B_k,$$

where, I_t denotes the set of time indexes in $\{1,\cdots,123\}$ corresponding to the same month as t and $\#I_t$ denotes the cardinality of I_t. Notice that only observed values were used to calculate μ_t. For modelling purposes, B_t was the response variable, while lagged data from Brazil and all other regions were used as covariates.

5.2.2. Modelling

Due to the large number of explanatory variables, we considered three different methods to fit the data, chosen because of their ability to perform variable selection. The first one was the stepwise backward regression method based on p-values with significance level 0.1, denoted simply by Stepwise model. We also applied the LASSO model considering two main schemes for model selection: first, based on cross-validation (leave-one-out), denoted LASSO CV, which yielded a model with 13 covariates plus the intercept;

Table 7. *Estimation results for the fitted Stepwise, LASSO 5 and LASSO CV models.*

Variables	Stepwise	LASSO 5	LASSO CV
Intercept	-0.843	60.821	44.089
B_{t-1}	0.84700	0.52843	0.68871
B_{t-2}	-0.34248	–	-0.15364
A_{t-2}	-0.00414	–	–
A_{t-4}	–	0.00019	0.00086
C_{t-1}	0.01515	–	0.00095
C_{t-3}	-0.02490	–	-0.01308
E_{t-1}	0.00535	–	0.00180
E_{t-2}	0.01117	0.00548	0.00650
E_{t-3}	–	0.00242	0.00084
E_{t-4}	0.00642	–	0.00230
E_{t-8}	0.00255	–	–
μ_t	–	0.01431	–
s_{t-4}	–	–	-0.00173
s_{t-9}	–	–	0.00066
W_{t-7}	–	–	-0.00158
W_{t-9}	–	–	0.00138

and second, since this model is somewhat large, a more parsimonious alternative using the "one-standard-error" rule (Hastie et al., 2009, section 7.10), selecting a model with five variables denoted by LASSO 5. Table 7 presents the covariates selected by each model and their fitted values. The intercept was always kept.

Notice that variables B_{t-1} (number of positive cases in Brazil at time $t-1$) and E_{t-2} (number of positive cases in Europe at time $t-2$) are the only variables present in all fitted models. Another way of interpreting the results is by analysing the coefficients of each variable present in the final model. It shows the direction of the impact that the explanatory variables have on the response variable B_t. For example, the explanatory variable B_{t-1}, which is included in all models, has a positive coefficient. This indicates that as the number of cases in Brazil at time $t-1$ increases/decreases, so does the number of cases in Brazil at time t. Also noteworthy is that the variable μ_t appears only in the LASSO 5 model.

5.2.3. Forecast

After modelling, we perform an in-sample and out-of-sample forecast exercise for the data considering each fitted model. Data from October 2008 to December 2018 were used for modelling purposes, while data from January 2019 to November 2019 were reserved for out-of-sample comparison. Hence, the forecast horizon in all cases is $h = 11$ steps ahead.

Notice that, since we are using a time series regression approach with several past values of regressors entering in the final model, these values must be updated if out-of-sample forecast values are to be obtained (that is, in order to obtain future values of the response variable, we need future values for the covariates as well). In order to do that we employed two approaches. The first approach employed is known as h one-step ahead forecast. In this case, for each incremental step ahead we updated the covariates with their observed values. This is only useful for small forecast horizons or to forecast short run dynamics, as in the case of flu data.

In the second approach, known simply as h-steps ahead forecast, we did not use any knowledge about future values of the covariates. Instead we forecasted their values using some plausible method. Of course, there are several ways to do that. We forecasted future values of the covariates by using their monthly average calculated from the observed data, including Brazil. This second approach can be employed for forecasting in practice. Figures 2 to 4 show both, the in-sample and out-of-sample (for $h = 11$) one-step ahead forecast values (the regressor values are updated at each step, including out-of-sample) compared to the observed ones (in black).

It can be seen that both in-sample and out-of-sample predictions appear to be reasonable for all models, except in a few epochs, such as the year 2011, where the models predicted a peak that did not occur, or the first half of 2015, where the peak was overestimated by all models. Since we are using a non-restricted time series approach to model the data, we obtain a few negative values for the incidence, located at the valleys. These negative values are not considered a problem because the main focus of the study of flu

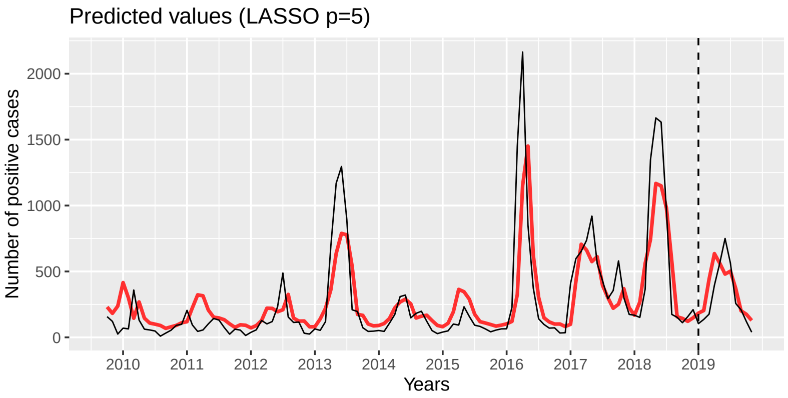


Figure 2. *In-sample and out-of-sample one-step ahead forecasts for the LASSO 5 model.*

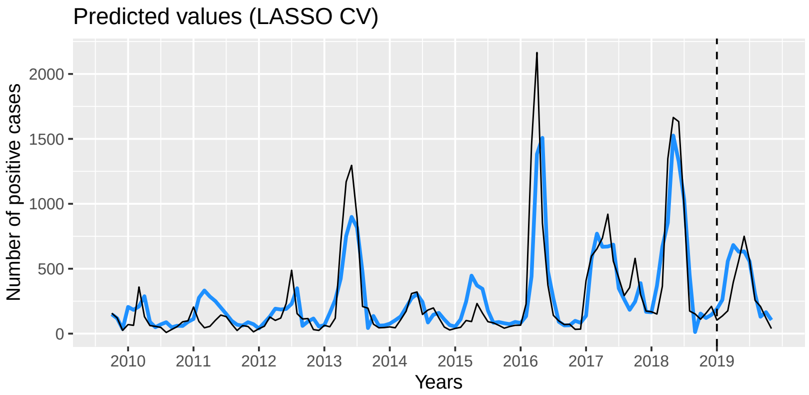


Figure 3. *In-sample and out-of-sample one-step ahead forecasts for the LASSO CV model.*

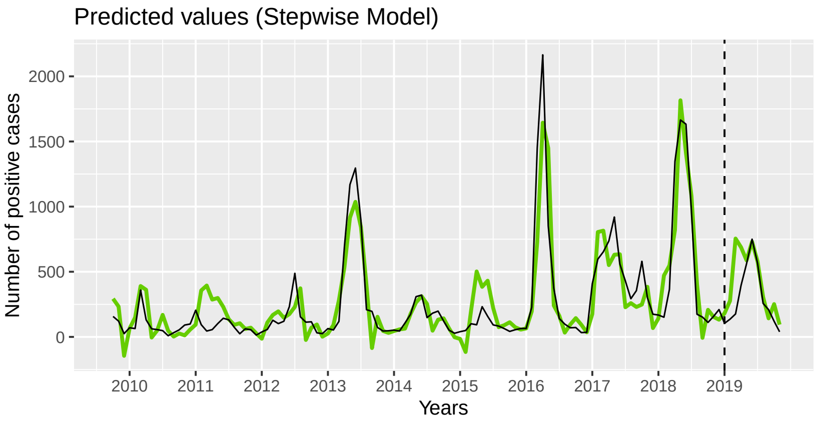


Figure 4. *In-sample and out-of-sample one-step ahead forecasts for the Stepwise model.*

pandemics are the peaks of the curve, not its valleys. Table 8 presents in-sample and out-of-sample mean square error (MSE) and mean absolute percentage error (MAPE) of forecasting. The best results in each case are highlighted in red.

Table 8. *Mean square error and mean absolute percentage error of the in-sample and 11 one-step ahead forecasts for each of the 3 fitted models.*

Measures/Models	Stepwise	LASSO 5	LASSO CV
MSE (in-sample)	31586.7	52788.1	38354.5
MSE (out-of-sample)	43228.2	21805.3	25860.9
MAPE (in-sample)	91.2	105.2	81.5
MAPE (out-of-sample)	81.2	64.3	68.7

The Stepwise model and the LASSO CV were the best performers in-sample, while for out-of-sample, the LASSO 5 model performed best both in terms of MSE and MAPE.

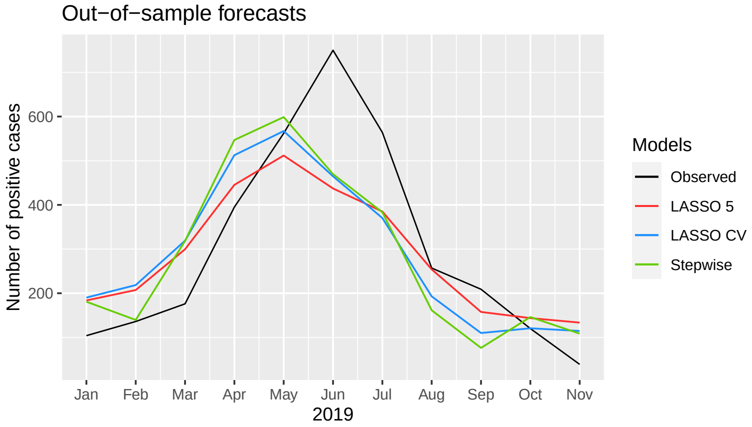


Figure 5. *11-steps ahead forecasts for the different fitted models compared to the observed values (in black).*

In a second moment, we analyze the out-of-sample h-steps ahead forecast ability of the fitted models, for $h \in \{1, \cdots, 11\}$. Figure 5 presents the forecasted values of each fitted model along with the observed values (in black). From Figure 5 we observe that all models overestimate the number of positive cases until May/April, missing the peak that occurred in June and underestimating the number of cases from June to October. For comparison purposes, Table 9 presents the mean square error for h-steps ahead forecasts for each model. The best results in each forecast horizon are highlighted in red. The model presenting the overall best performance was the LASSO 5, which presented the lowest MSE in 7 out of the 11 forecast horizons considered, followed by the Stepwise which presented overall smallest MSE in the remaining 4 forecast horizons. Interestingly, the LASSO 5 uniformly outperforms the LASSO CV in all forecast horizons.

This poor forecasting performance of the LASSO CV may be attributed to overfitting. Stepwise and LASSO CV presented similar performances.

Table 9. *Mean squared error of the h-steps ahead forecast for each fitted model. The best forecast in terms of MSE for each horizon is presented in red.*

Horizon/Models	Stepwise	LASSO 5	LASSO CV
1-step ahead	5904.13	6346.41	7417.52
2-steps ahead	2958.38	5712.86	7110.19
3-steps ahead	8736.86	8895.71	11549.42
4-steps ahead	12323.75	7305.05	12114.87
5-steps ahead	10129.28	6349.04	9697.23
6-steps ahead	21543.97	21599.59	21646.22
7-steps ahead	23130.21	23092.44	23917.95
8-steps ahead	21381.31	20206.98	21442.05
9-steps ahead	20962.64	18254.56	20149.43
10-steps ahead	18934.49	16485.37	18134.53
11-steps ahead	17653.39	15798.62	17005.14

5.2.4. Genetic diversity

In this section we consider both the number of positive flu cases and the genetic diversity data (described in sections 4.1 and 4.2) to characterize the flu incidence in Brazil. We apply similar notation to Section 5.2.1. The genetic diversity in North America at time t is denoted by N_t for the H1N1 subtype and n_t for the H3N2 subtype, in Asia by P_t for the H1N1 subtype and p_t for the H3N2, and M_t (H1N1) and m_t (H3N2) denote the global genetic diversity. Again the response variable is taken as B_t while lagged variables related to other regions, including genetic data, will be used as covariates.

We aim to explain the incidence of influenza in Brazil at time t (B_t) by using the historical incidence and genetic diversity data of the regions considered in the last 6 months (lags). The same three methods of Section 5.2.1 were considered. Table 10 presents the selected covariates and their respective coefficients, for each model.

From Table 10, we observe that the incidence variables that appear in all three models are B_{t-1} (number of positive cases in Brazil with one lag), E_{t-2} (number of positive cases in Europe with two lags) and A_{t-4} (number of positive cases in North America with four lags). Furthermore, when analyzing the variables related to genetic diversity, we observe that the covariates P_{t-4} (genetic diversity of the H1N1 flu in Asia with four lags) and P_{t-5} (genetic diversity of the H1N1 flu in Asia with five lags) appear in two of the tree models. After model fitting, we proceed with an in-sample and out-of-sample forecast analysis similar to the one presented in Subsection 5.2.3. To perform the out-of-sample analysis, it is necessary to forecast future values of covariates entering the model. Covariates related to incidence are forecasted in the same way as in Subsection 5.2.1. The genetic

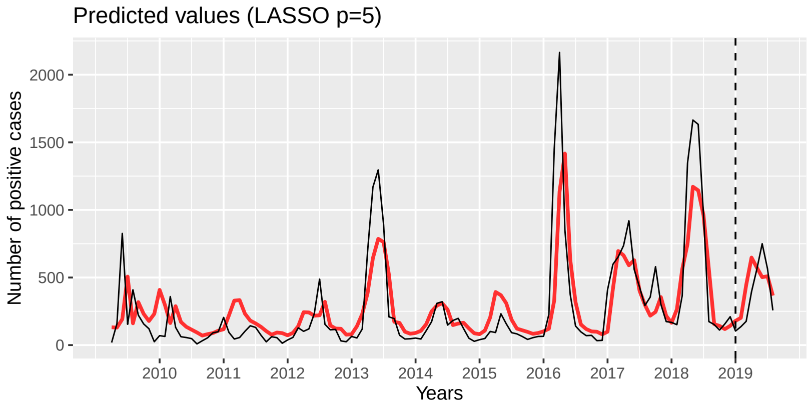


Figure 6. *In-sample and out-of-sample one-step ahead forecasts for the LASSO 5 model.*

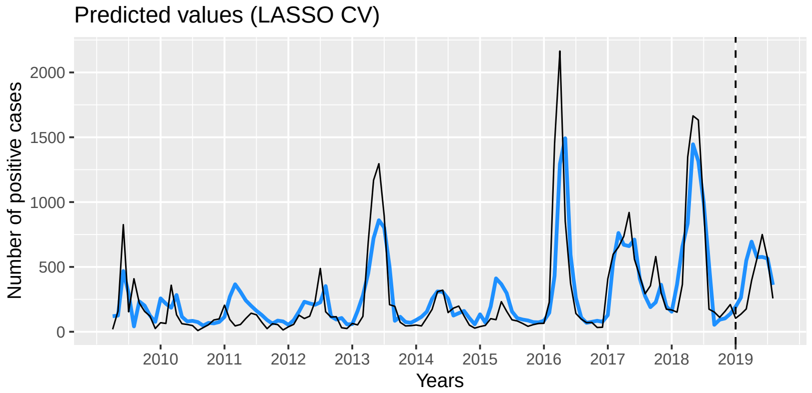


Figure 7. *In-sample and out-of-sample one-step ahead forecasts for the LASSO with cross-validation model.*

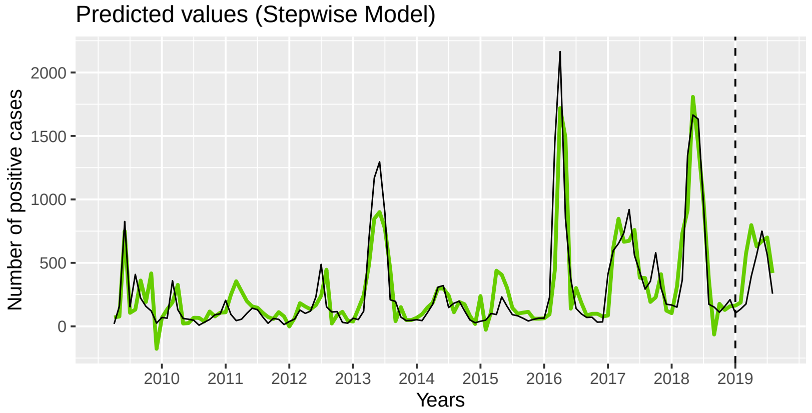


Figure 8. *In-sample and out-of-sample one-step ahead forecasts for the Stepwise model.*

Table 10. *Estimation results for the fitted Stepwise, LASSO 5 and LASSO CV models.*

Variables	Stepwise	LASSO 5	LASSO CV
Intercept	38.789	54.872	55.460
B_{t-1}	0.85717	0.50499	0.60024
B_{t-2}	-0.34293	–	-0.03383
B_{t-3}	–	–	-0.04411
A_{t-4}	0.00286	0.00037	0.00166
C_{t-2}	0.02204	–	–
C_{t-3}	-0.03545	–	-0.00946
S_{t-2}	0.02604	–	–
E_{t-1}	–	–	0.00161
E_{t-2}	0.00893	0.00535	0.00609
E_{t-3}	–	0.00219	0.00086
E_{t-4}	–	–	0.00053
s_{t-4}	–	–	-0.00957
W_{t-6}	–	–	-0.00094
μ_t	–	0.07586	0.05940
M_{t-5}	–	–	-837.80
P_{t-4}	6482.74	–	3827.28
P_{t-5}	-9221.79	–	-3156.14

diversity time series, however, do not present any evident trend or seasonality, as in the incidence data (see the time series plots presented in the supplementary material). Hence, the same approach of considering monthly averages is not adequate for the genetic diversity data. To overcome this difficulty, we consider a static approach: future values of genetic data are forecasted considering the average of the respective data observed from January to December, 2018. Figures 6 to 8 show the one-step ahead forecasted values in and out-of-sample for each model along with the observed values (in black).

Table 11. *Mean square error and mean absolute percentage error of forecast for each model.*

Measures/Models	Stepwise	LASSO 5	LASSO CV
MSE (in-sample)	34567.0	55306.2	42026.3
MSE (out-of-sample)	47623.4	27781.2	36757.7
MAPE (in-sample)	91.2	110.9	89.8
MAPE (out-of-sample)	66.5	53.1	66.7

Note that, in general, the in-sample and out-of-sample predictions appear to be reasonable for all considered models. Some peaks, such as the ones in years 2013, 2016 and 2018, are underestimated by the models, while others, such as 2011 and 2015 are overestimated. To compare the models regarding their predictive abilities, Table 11 presents the MSE and MAPE for the one-step ahead forecast for each model, in and out-of-sample. The best results in each case are highlighted in red.

Analogously to the results obtained in Section 5.2.1, analyzing the MSE we observe that the Stepwise model and the LASSO CV are the best perform in terms of in-sample forecast while the LASSO 5 is the best performer out-of-sample. Again, the main source of forecast error are a few peaks in the data not very well identified by any of the models, more noticeably, 2011, 2015 and 2017.

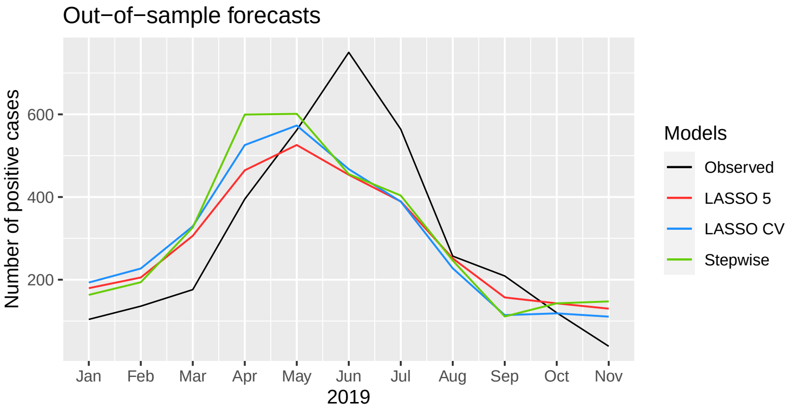


Figure 9. *11-steps ahead forecasts for the different fitted models compared to the observed values (in black).*

Table 12. *Mean squared error of the h-steps ahead forecast for each fitted model. The best forecast in terms of MSE for each horizon is presented in red.*

Horizon/Model	Stepwise	LASSO 5	LASSO CV
1-step ahead	3512.23	5701.45	7945.86
2-steps ahead	3428.68	5255.09	8122.44
3-steps ahead	9823.78	9153.06	13233.38
4-steps ahead	17825.10	8076.48	14200.18
5-steps ahead	14567.40	6721.95	11384.56
6-steps ahead	26605.43	20228.37	22796.37
7-steps ahead	26475.39	21693.13	23918.67
8-steps ahead	23179.94	18985.19	21037.13
9-steps ahead	21671.71	17174.71	19696.15
10-steps ahead	19556.80	15509.01	17726.74
11-steps ahead	18847.90	14851.20	16581.80

In a second step, we analyse the models' predictive capabilities considering forecast horizons from 1 to 11-steps ahead, in the same spirit as in Subsection 5.2.1. Figure 9 presents the forecasted values for the different models, as well as the observed values (in black). We observe that all models overestimate the number of positive cases until May/April, missing the peak that occurred in June and underestimating the number of cases from June to October. A comparison between the models is presented in Table 12, where we present mean squared error for each considered h-steps ahead forecast. The results show that no model uniformly outperforms all others. The model with the best results was the LASSO 5, which displayed the lowest MSE in 9 out of the 11 forecast horizons considered. Again the LASSO CV is uniformly outperformed by LASSO 5, while compared to Stepwise, the LASSO CV wins in middle to long horizons.

5.2.5. Comparison of forecasts

We now compare the results presented in Subsections 5.2.1 and 5.2.4. The interest lies in comparing the models with only incidence data with the models with incidence and genetic diversity data, regarding their predictive power. In the graphs below, the term "Incidence" will be used for models containing only incidence data while the term "Genetic" will be used for models considering incidence and genetic diversity data. Figure 10 shows the MSE obtained in the out-of-sample forecast for all models. It can be seen that in all cases the models based on "Incidence" presented more accurate forecasts (lower MSE). Furthermore, the LASSO 5 proved to be the overall best model in terms of prediction capabilities in all cases.

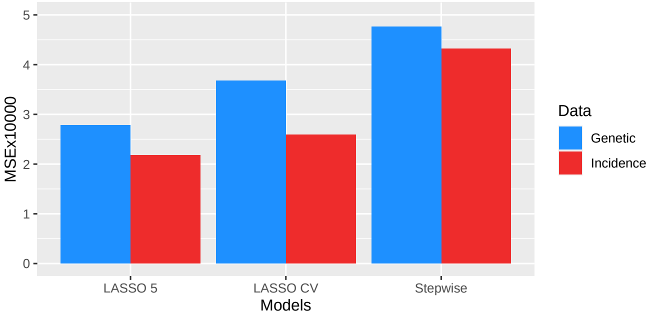
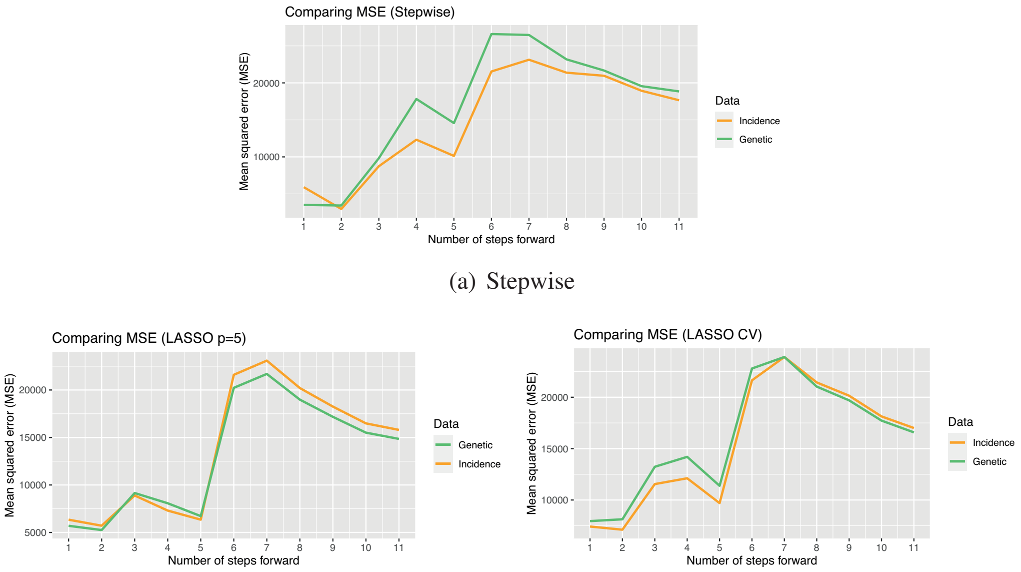


Figure 10. *Comparison of the mean squared errors (MSE) between the incidence data and the genetic data in the 11 one-step ahead forecasts.*

Figure 11 presents time series plots of the MSE for h-steps ahead forecast for each model considering the incidence and genetic data. For the Stepwise model, the out-of-sample forecasts produced using the incidence data present smaller MSE in all horizons but $h = 1$. For the LASSO, the models based on the genetic data presented smaller MSE in the long run, that is, for all horizons $h \geq 6$ for the LASSO 5 and $h \geq 7$ for the LASSO CV. In the short run, for the LASSO CV, the model based on incidence data performs best, while there is no clear pattern in the case of the LASSO 5 model. Ultimately, this indicates that including genetic diversity data, at least as measured here, does not seem to add much predictive value to the models. However, there are many other approaches to assess genetic diversity that can be explored and might prove more valuable for incidence modelling.



(a) Stepwise

(b) LASSO 5

(c) LASSO CV

Figure 11. *Comparison of the h-steps ahead forecasts mean squared errors between the incidence and genetic data for the Stepwise (upper panel), LASSO 5 and LASSO CV (lower panel) models.*

5.3. Residual analysis

Residual analysis is of paramount importance in time series analysis, being performed after model identification and fitting. In this section we present a residual analysis related to the models fitted in the previous sections, focusing mainly in portmanteau and normality tests. Observe, however, that the only model that actually requires a residual analysis is the Stepwise, as it is the only one based on p-values. Nevertheless, for the sake of exploration, we shall proceed with the residual analysis for all models. To assess the presence of correlation in the residuals, we perform the widely applied Ljung-Box test (Ljung, 1986). Recall that the null hypothesis for the Ljung-Box test is that all correlations up to a specified lag m are null. In this analysis we consider $m = 20$. We also test the residuals for normality by using Shapiro-Wilk's test (Shapiro and Wilk, 1965), for which the null hypothesis is that the tested sample comes from a normally distributed population.

The Ljung-Box test's results for the residuals of all models presented in Sections 5.2.1 (flu incidence) and 5.2.4 (genetic data) are presented in Table 13. From the results we conclude that in all cases the residuals present no correlation up to lag $m = 20$, at any reasonable significance level.

Table 13. *p-values of the Ljung-Box test applied to the fitted models' residuals with $m = 20$.*

Dataset	Models		
	Stepwise	LASSO 5	LASSO CV
Incidence	0.9561	0.2096	0.8212
Genetic	0.9922	0.3805	0.4770

As for the Shapiro-Wilk test, it is clear from the in-sample forecasts (Figures 2 to 4 and Figures 6 to 8) that the residual will present outliers due to underestimation of peak values. These outliers may substantially affect the Shapiro-Wilk test. To minimize this effect, we removed some of the outliers by using two hard thresholds: we eliminate any points with magnitude larger than 400 (threshold 1) and 200 (threshold 2), in absolute value. Table 14 summarizes the results by presenting the p-values of the Shapiro-Wilk test with and without the removal of outliers, along with the number of outliers removed in each case. From the results we observe that the residuals of all models reject the null hypothesis in the Shapiro-Wilk test with very small p-values. The Stepwise model for the incidence data is the only one that do not reject the null hypothesis in the Shapiro-Wilk's test after applying threshold 1, which trimmed out only 5 points. The LASSO CV model for all data and Stepwise with genetic data did not reject at the 0.05 significance level the null hypothesis in Shapiro-Wilk's test after applying threshold 2, at the cost of removing several points. The Shapiro-Wilk's test applied to the residuals from the LASSO 5 model rejected the null hypothesis in all cases.

Table 14. *p-values for the Shapiro-Wilk test applied to the complete residuals and upon removing points with magnitude greater than 400 and 200, in absolute value. The number of points removed for each threshold applied is presented in parenthesis.*

Dataset	Threshold	Models		
		Stepwise	LASSO 5	LASSO CV
Incidence	complete	< 0.0001	< 0.0001	< 0.0001
	400	0.3012(5)	< 0.0001(9)	0.0003(5)
	200	–	0.0089(23)	0.1881(19)
Genetic	complete	< 0.0001	< 0.0001	< 0.0001
	400	0.0046(4)	< 0.0001(9)	< 0.0001(6)
	200	0.2357(26)	0.0015(28)	0.0952(25)

Finally, another important diagnostic is the homoscedasticity of the residuals for the fitted Stepwise model, the only one of our procedures that relies on distributional assumptions for model selection. Figure 12 presents simple time series plot and observed vs. fitted values for the residuals obtained from the Stepwise model considering the Incidence and Genetic data. From the time series plot we observe a clear increase in variance in both residuals, also evident in the residual versus fitted models. These findings are corroborated by Breusch-Pagan and White's tests (see Greene, 2012, section 11.4) (not shown). The presence of heteroscedasticity in the model's residuals may affect the p-values obtained from Wald's test, implying that the fitted model may be incorrectly specified in the sense that the procedure may have excluded important variables, included unimportant ones, or both. This, however, does not diminish its applicability as a predictive model.

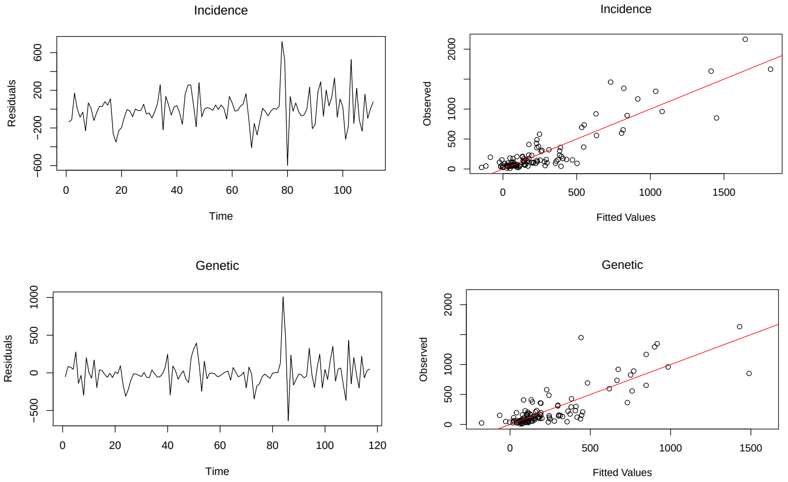


Figure 12. *Time series plot (left panel) and observed vs. fitted value (right panel) for the residuals obtained from the fitted Stepwise model. Plots related to incidence are shown in the upper panel while genetic ones are shown in the lower panel.*

6. Discussion

In this paper we considered the problem of modelling and forecasting the incidence of influenza virus in Brazil at a given month t. Here, FluNet positive flu counts were used as a proxy for incidence. The objective is to use temporal information (flu historical time series data) to model the number of cases in Brazil based on recent data on the number of cases and the genetic diversity observed in other regions. Incidentally, the study also sheds light on the migratory dynamics of the influenza virus from North America and Europe to Brazil.

In Section 5.1 (Granger causality analysis) we found evidence that past values of influenza incidence in the European and South American Regions help to predict the present value of influenza incidence in Brazil. We also discovered evidence of an indirect effect of the Western Pacific Region and Central America in Brazil. These results are intriguing when considering updating vaccines in Brazil with data related to strains from Europe from previous seasons.

As for the time series regression approach (Section 5.2), it was found that only two variables are present in all considered models, namely: B_{t-1} (number of positive flu cases in Brazil with one lag), and E_{t-2} (number of positive cases in Europe with two lags), while A_{t-4} (number of positive cases in North America with four lags) was present in five out of six models. It is interesting to note that most predictors from northern hemisphere regions appear with lags of 3–5, possibly capturing seasonal properties of the dynamic. Additionally, while Asian genetic diversity measures appear as relevant predictors, the global genetic diversities do not.

The proposed models were also evaluated regarding their forecast capabilities. Considering h-steps ahead out-of-sample forecast, in both analysis of Sections 5.2.1 and 5.2.4, the model that overall best predicted the incidence of influenza in Brazil (in terms of MSE) in the short run was the Stepwise and in the middle to long run, the LASSO with 5 variables. The LASSO CV model performed poorly in all cases. This might be a consequence of overfitting since the LASSO CV is the one with most variables included among the considered models.

The Covid19 pandemic has largely impacted human global circulation and, consequently, the global dynamics of influenza transmission. Some lineages have remained present in local circulation and others have all but disappeared (such as B/Yamagata). Overall, the FluNet numbers of positive cases have drastically decreased. It is expected that once circulation returns to prepandemic levels influenza cases will rise again, however it is still unclear to what degree the previous transmission patterns will be reestablished or if we will see new dynamics. It has even been argued that we might see more severe influenza epidemics due to changes in immunity related to low circulation periods (Dhanasekaran et al., 2021).

Ultimately, it is likely that influenza incidence will once more be largely determined by a global dynamic, and thus modelling the Brazilian cases based on the number of cases in other regions will remain relevant. Furthermore, this same approach might prove valuable to other countries, particularly those in the global south, similarly placed in the global dynamics.

Overall, our results for short and long run forecasts ($h = 1$ and $h = 11$ steps ahead) were fairly good. Together with the relationships outlined by the Granger-causality analysis they help shed light on the global determinants of influenza incidence in Brazil. Time will tell if the particular predictors selected here will remain relevant, and in this sense, this work can be seen as historical record to be compared with the postpandemic dynamics. Nevertheless, the overall approach highlights a modelling concept which can potentially be useful in the development of public health policies regarding epidemic management and immunizations.

Acknowledgments

Aline F. Grande and Guilherme Pumi gratefully acknowledge the support of CNPq and FAPERGS. Gabriela B. Cybis gratefully acknowledges the support of the Serrapilheira Institute (grant number Serra-G1709-18939). The authors are also grateful to Rafaela Gomes de Jesus for helping with the genetic diversity data assembly.

References

Almeida, A., Codeço, C., and Luz, P. M. (2018). Seasonal dynamics of influenza in Brazil: the latitude effect. *BMC infectious diseases*, 18(1):1–9.

Barr, I. G., McCauley, J., Cox, N., Daniels, R., Engelhardt, O. G., Fukuda, K., Grohmann, G., Hay, A., Kelso, A., Klimov, A., Odagiri, T., Smith, D., Russell, C., Tashiro, M., Webby, R., Wood, J., Ye, Z., and Zhang, W. (2010). Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: Basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009-2010 Northern Hemisphere season. *Vaccine*, 28(5):1156–1167.

Bedford, T., Cobey, S., Beerli, P., and Pascual, M. (2010). Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS pathogens*, 6(5):e1000918.

Born, P. S. (2013). *Análises filogenéticas e filogeográficas dos vírus influenza A(H3N2) papel do Brasil no cenário de dispersão global e ajuste temporal entre as cepas vacinais e os vírus circulantes no período de 1999 a 2012*. PhD thesis, Instituto Oswaldo Cruz.

Born, P. S., Siqueira, M. M., Faria, N. R., Resende, P. C., Motta, F. C., and Bello, G. (2016). Phylodynamics of influenza A(H3N2) in South America, 1999-2012. *Infection, Genetics and Evolution*, 43:312–320.

Chen, C. W. S., Hsieh, Y.-H., Su, H.-C., and Wu, J. J. (2018). Causality test of ambient fine particles and human influenza in Taiwan: Age group-specific disparity and geographic heterogeneity. *Environment International*, 111:354–361.

Dhanasekaran, V., Sullivan, S., Edwards, K., Xie, R., Khvorov, A., Valkenburg, S., Cowling, B., and Barr, I. (2021). Human seasonal influenza under covid-19 and the potential consequences of influenza lineage elimination. *Research Square (preprint)*.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431.

Diniz, M. B., Junior, J. N. O., Neto, N. T., and Diniz, M. J. T. (2009). Causas do desmatamento da Amazônia: uma aplicação do teste de causalidade de Granger acerca das principais fontes de desmatamento nos municípios da Amazônia Legal brasileira. *Nova Economia*, 19(1):121–151.

Eccles, R. (2005). Understanding the symptoms of the common cold and influenza. *The Lancet Infectious Diseases*, 5(11):718–725.

Farias, H. P. and Sáfadi, T. (2010). Causalidade entre as principais bolsas de valores do mundo. *RAM. Revista de Administração Mackenzie*, 11(2):96–122.

Garten, R. J., Davis, C. T., Russell, C. A., and et. al. (2009). Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. *Science*, 325(5937):197–201.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438.

Greene, W. H. (2012). *Econometric Analysis*. Pearson Education Limited, 7 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Ibiapina, C. C., Costa, G. A., and Faria, A. C. (2005). Influenza A aviária (H5N1) - a gripe do frango. *Jornal Brasileiro de Pneumologia*, 31(5):436–444.

Jesus, R. G. (2018). Caracterização e visualização da diversidade genética do vírus influenza ao longo do tempo. Monografia (Bacharel em Estatística), UFRGS (Universidade Federal do Rio Grande do Sul), Porto Alegre.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C. A., Smith, D. J., Pybus, O. G., Brockmann, D., et al. (2014). Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS pathogens*, 10(2):e1003932.

Ljung, G. M. (1986). Diagnostic testing of univariate time series models. *Biometrika*, 73(3):725–730.

NCBI (2020). National Center for Biotechnology Information. Last accessed 12 June 2020.

Petrova, V. N. and Russell, C. A. (2018). The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16(1):47–60.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. and Holmes, E. (2009). The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Currents*, 1:RRN1003.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619.

Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., and Smith, D. J. (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science*, 320(5874):340–346.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Silva, P. C. R. (2015). Dinâmica molecular dos vírus Influenza A (H1N1) pandêmico em cinco anos de circulação no Brasil. Master's thesis, Instituto Oswaldo Cruz.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Toda, H. Y. and Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66(1-2):225–250.

WHO (2020). FluNet. Last accessed 12 June 2020.

Yamada, K. D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data−reexamination of the usefulness of chained guide trees. *Bioinformatics*, 32(21):3246–3251.

DOI: 10.2436/20.8080.02.123

Compositional combination and selection of forecasters

Antonio Martín Arroyo* and Aránzazu de Juan Fernández†

Abstract

The Split-Then-Combine approach has previously been used to generate the weights of forecasts in a combination in the Euclidean space. This paper extends this approach to combine forecasts inside the simplex space, the sample space of positive weights adding up to one. As it turns out, the simplicial statistic given by the sample centre compares favourably against the fixed-weight, average forecast. Besides, we also develop a Combination-After-Selection method to get rid of redundant forecasters. We apply these approaches to make out-of-sample one-step ahead combinations and subcombinations of forecasts for several economic variables. This methodology is particularly useful when the sample size is smaller than the number of forecasts, a case where other methods (e.g., ordinary least squares or principal component analysis) are not applicable.

MSC: *54C65, 91B84, 62P20.*

Keywords: *Aitchison geometry, Combination-After-Selection, Dimensionality problem, Simplex, Split-Then-Combine.*

Competing interest: The authors do not present any conflicts of interests.

Declarations

Funding: The second author acknowledges financial support from the Spanish Government Project PID2019-108079GBC22.

* a.martin@uam.es. Universidad Autónoma de Madrid (UAM): Economía Cuantitativa, E-III-306 , Avda. Francisco Tomás y Valiente 5, 28049 Madrid (Spain). ORCID ID: 0000-0002-9071-5231

† Corresponding author: aranzazu.dejuan@uam.es. Universidad Autónoma de Madrid (UAM): Economía Cuantitativa, E-III-307, Avda. Francisco Tomás y Valiente 5, 28049 Madrid (Spain). ORCID ID: 0000-0002-7171-7434
Received: October 2021
Accepted: October 2022

Availability of data and material: The data used in this paper are obtained from a public source and are available from the authors upon request.
Code availability: The application developed in this paper was estimated using R code. The codes are available from the authors upon request.
Authors's contribution: Both authors have worked in the manuscript in a similar way. Both have prepared the technical part of the manuscript. Both authors have developed the application and both authors have contributed to the writing of the paper.

1. Introduction

There is a vast body of literature advocating the usefulness of forecast combination methods, both theoretically and empirically. A simple and widely used one consists of simply attributing equal weights to the individual predictions (neutral element of a weight combination in the simplex.) However, the idea of determining the optimal weight combination that minimize some objective criterion (e.g., the mean square forecast error) is more appealing (Conflitti, De Mol and Giannone, 2015). This is the case of the varying-weight sample centre g (Aitchison, 1982). It is the central tendency of our weight combinational sample and it is defined as the weight combination whose components are the sample geometric mean of the weights of each forecaster. Unlike ordinary least squares (OLS) and principal component analysis (PCA), this is a viable strategy even when the number of forecasters to be combined gets large, provided that we constrain weights to be positive and add up to one. Hence, the optimal combination problem reduces to a (possibly high-dimensional) constrained least-squares regression problem, where the complete covariance structure between weights is taken into account. Indeed, this enforces an implicit shrinkage on weights which ensures a reasonable out-of-sample performance of the combined forecasters. This problem turns out to be analogous to the determination of no-short minimum variance Markowitz portfolios, which are a special case of a larger family of sparse and stable portfolios that are derived through a constrained "lasso" regression problem (Tibshirani, 1996), where the weight vector has a unit L1-norm. This type of constraint is known to enforce sparsity, namely the presence of zeros in the weight vector, which means that only a small number of forecasters will be selected (subcombinations in our Combination After Selection (CAS) approach).

Forecasters have access to a wide variety of information and forecasting techniques, thus leading to a considerable degree of heterogeneity or redundancy among them. A weighted average forecast is expected to perform better than individual ones because this way we can diversify away idyosincratic forecast misspecifications, thus reducing the variance of the forecast. The simplest example is the (fixed weight) arithmetic average. More sophisticated methods that make use of varying weights usually do not improve the average in empirical applications because of the instability of the estimated weights (a problem known as *forecast combination puzzle,* Stock and Watson, 2004); in particular, when an increasing number of forecasters requires us to estimate an increasing number of weights (a problem known as *the curse of dimensionality*). The forecast combination puzzle has been considered by Smith and Wallis (2009), who pointed out that the failure of more sophisticated combination methods is due to the estimation of the combining weights.

With forecast (or model)-specific combinations, forecasting is often based on predicting the same variable independently by forecasters. However, analysts who are interested in forecasting a variable from a specific source should not ignore the forecasts from other competing sources. A forecast combination is in fact influenced by all the forecasts; hence, the relationship among individual forecasters is lost when forecasts are independently analysed. Only a few methods have been suggested that incorporate dependence between forecasters. Multivariate models could incorporate dependence between forecasters if we knew such a dependence. Alternatively, we can engage straightaway with weight distributions based on given individual forecast errors, as dependence between weights can be incorporated directly, thus increasing forecast accuracy.

One important difference between modeling forecast-specific combinations and weight distributions is that weights are directly dependent on each other on an aggregated level. The awareness of problems, however, arising from the use of standard statistical methods with proportions (weights) dates back to Pearson, (1897); that is, spurious effects on their covariance structure. In particular, each row or column of the variance matrix of a vector of weights sums up to zero. Given that the variances are always positive, this implies that some covariances are forced towards negative values (Chayes, 1960).

Independent modelling and forecasting with forecast-specific combinations are not only unattractive since they ignore dependence patterns among (relative) weights, but also because weights often fail to be coherent in the sense of the erratic way in which the covariance associated with two specific weights can fluctuate in sign as we move from a full combination to lower and lower dimensional subcombinations. In fact, there is no relationship between the variance matrix of a subcombination and that of the full combination. Besides, variances may display different rank orderings as we form subcombinations, which could lead to implausible forecasters.

Also, avoided forecasts in a subcombination will result in an increase of weights for some other forecasters. By definition of a weight combination, not only is there a common element in the numerator and denominator of each weight, but also all weights have a common denominator. Avoided forecasters in a subcombination thus affect both the numerator and the denominator, and the dependence between forecasters is therefore not as easy to predict.

Moreover, all combinations are subcombinations of a larger one. Since the covariance between two weights depends on which other forecasters are reported in the dataset, there is no guarantee that a plot of a subcombination exhibits similar or even compatible patterns with the plot of the original dataset, even if the forecasters not included in the subcombination are irrelevant (redundant).

There is thus incoherence of the correlation between weights as a measure of dependence. Note, however, that the ratio of two weights remains unchanged when we move from a full combination to a subcombination. Therefore, as long as we work with scale invariant functions (i.e., ratios), we shall be subcombinationally coherent (Aitchison, 1986).

Since standard descriptive statistics (e.g., arithmetic mean and standard deviation) are not informative with combinations, in this paper we propose a time-varying method to combine, select, and recombine forecasters based on Aitchison (1982, 1986), who characterized compositions as vectors having a relative scale and identified its sample space with the simplex. More crucial than the constraining property of compositional data is the scale-invariant property of this kind of data. Indeed, when we are considering only few forecasters of a full combination we are not working with constrained data but our data are still compositional. This approach has been successfully applied to various fields; see, for instance, Aitchison (1986) Billheimer, Guttorp and Fagan (2001), Egozcue and Pawlowsky-Glahn (2005, 2019), Coenders and Ferrer-Rosell (2020) and Greenacre (2021). Software packages available now to deal with compositional data are, for example, Van den Boogaart and Tolosana-Delgado (2013) and Filzmoser, Hron and Templ (2018). To our knowledge, it has not been applied to combinations of forecasts. Compositional Data Analysis (CoDA) is a well-established set of statistical methods for the analyses of compositional data, that enables coherent modelling of weight combinations where dependences between weights are explicitly modeled, so a relative improvement in the weight for one forecaster leads to a decline in the relative weight for the remaining ones.

Any statement about weight combinations can be reformulated in terms of (centred) logratios and viceversa (one-to-one transformation). Data are projected into multivariate real space, opening up all available standard multivariate techniques. Moreover, weight combinations may be represented by orthonormal coordinates (Mateu-Figueras, Pawlowsky-Glahn and Egozcue (2011); Pawlowsky-Glahn and Buccianti (2011), Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2015)) in a real Euclidean space that can be interpreted in themselves or from their representation in the simplex (Aitchison geometry).

The analysis that is presented in this paper uses the Split-Then-Combine (STC) approach of Arroyo and de Juan Fernández (2014) to generate the weights of a combination. Because they are restricted to be positive and sum up to one, we propose the sample centre g of our weight combinational sample as our basic simplicial combination vector. To get a subcombination, we develop a Combination-After-Selection (CAS) procedure to recombine the best subset of forecasters.

The paper is organized as follows: the next section describes the STC approach both in the Euclidean and simplex spaces. Then, we explain the CAS strategy. In the empirical application, in Section 4, we pull out information provided by panels of quarterly periodicity from a pool of expert forecasters for the US macroeconomy over the period 1991–2018. Forecast accuracy of simplicial combinations are compared with the uniform benchmark arithmetic average.The results obtained with CAS are clearly better than the obtained with the other combinations. Finally, some concluding remarks complete the paper.

2. The Split-Then-Combine (STC) approach

Arroyo and de Juan (2014) proposed the Split-Then-Combine approach to generate combinations for panel m across J forecasters $\left(\widehat{Y}_{t,j}\right)$, $j = 1,2,...,J$, along $t = 1,2,...,T$ periods using the expression:

$$\widehat{Y}_t^{(m)} = \omega_{t,1}^{(m)}\widehat{Y}_{t,1}^{(m)} + \omega_{t,2}^{(m)}\widehat{Y}_{t,2}^{(m)} + ... + \omega_{t,J}^{(m)}\widehat{Y}_{t,J}^{(m)},$$

where the weights $\omega_{t,j}^{(m)}$ vary in two dimensions: (1) from one period to the next; and (2) from one panel to another. We have one panel for each season. Each panel is a tableau of T rows (years) and J columns (forecasters). Each row is then closed to a positive weight combination with weights adding up to one. Finally, this weight combination is used to weight forecasters in out-of-sample forecasting exercises. For example, if we are working with monthly data, we will have 12 panels, one for each month; if we work with quarterly data, we will have four panels, one for each quarter. Panels take into account the different behaviour of the time series among seasons, but STC can also be applied to time series with lower frequency than quarterly or monthly data.[1]

The weights of the STC approach must satisfy two restrictions: be positive and sum up to one; the latter is to avoid biased combinations if individual forecasts are unbiased. Arroyo and de Juan (2014) developed the STC in the Euclidean Space. Here, we also study the STC in the so-called Aitchison geometry (Billheimer et al., 2001, and Pawlowsky-Glahn and Egozcue, 2001).

In order to see the differences between both methods, we first briefly review the STC approach in the Eucidean space; then, we expand the STC approach to the simplex space.

2.1. The STC approach in the Euclidean Space

Table 1 shows how the STC approach works in the Euclidean space. Columns 2 to 5 show the forecasts of the variable of interest for panel m. Each element of this column represents the forecast of each forecaster for a given period. For instance, $\widehat{Y}_{2,1}^{(m)}$ is the forecast of a variable of interest Y from forecaster 2 for period 1 in panel m. The 6th column shows the cross average by period for the J forecasters; that is, $\overline{\widehat{Y}}_{J,1}^{(m)}$ is the average of the J forecasters for the first forecasting period. The 6th row shows the time average by forecaster, that is, $\overline{\widehat{Y}}_{1,T_1}^{(m)}$ is the average over time of all the forecasts from the first forecaster. Column 7 reports the actual, observed data of the variable and the 7th row shows the precision of each forecast average with respect to the overall average $\left(\overline{\overline{\widehat{Y}}}_{J,T_1}^{(m)}\right)$. This measure is used to construct the weights ω that will be assigned to each forecast in the STC approach in the Euclidean space.

[1] See Bujosa-Brun et al. (2020) for an application of the STC approach to annual data with only one panel.

Table 1. *STC approach in the Euclidean Space.*

Panel m	1	2	...	J	$\overline{\widehat{Y}}^{(m)}_{J,t}$	Real data
1	$\widehat{Y}^{(m)}_{1,1}$	$\widehat{Y}^{(m)}_{2,1}$	...	$\widehat{Y}^{(m)}_{J,1}$	$\overline{\widehat{Y}}^{(m)}_{J,1}$	$Y^{(m)}_{1}$
2	$\widehat{Y}^{(m)}_{1,2}$	$\widehat{Y}^{(m)}_{2,2}$	...	$\widehat{Y}^{(m)}_{J,2}$	$\overline{\widehat{Y}}^{(m)}_{J,2}$	$Y^{(m)}_{2}$
...	...	...	...	...	...	...
T_1	$\widehat{Y}^{(m)}_{1,T_1}$	$\widehat{Y}^{(m)}_{2,T_1}$	...	$\widehat{Y}^{(m)}_{J,T_1}$	$\overline{\widehat{Y}}^{(m)}_{J,T_1}$	$Y^{(m)}_{T_1}$
$\overline{\widehat{Y}}_{j,T_1}$	$\overline{\widehat{Y}}_{1,T_1}$	$\overline{\widehat{Y}}_{2,T_1}$	...	$\overline{\widehat{Y}}_{J,T_1}$	$\overline{\overline{\widehat{Y}}}^{(m)}_{J,T_1}$	$\overline{Y}^{(m)}_{T_1}$
Fixed	$\left(\overline{\widehat{Y}}^{(m)}_{1,T_1} - \overline{\overline{\widehat{Y}}}^{(m)}_{J,T_1}\right)^{-2}$	$\left(\overline{\widehat{Y}}^{(m)}_{2,T_1} - \overline{\overline{\widehat{Y}}}^{(m)}_{J,T_1}\right)^{-2}$	...	$\left(\overline{\widehat{Y}}^{(m)}_{J,T_1} - \overline{\overline{\widehat{Y}}}^{(m)}_{J,T_1}\right)^{-2}$		

The STC weights ω are then computed with the information up to time T_1 for each panel using the precision accuracy of each forecaster based on the normalized average squared forecast error:

$$\omega^{(m)}_{j,T_1} = \frac{\left(\overline{\widehat{Y}}^{(m)}_{j,T_1} - \overline{\overline{\widehat{Y}}}^{(m)}_{J,T_1}\right)^{-2}}{\sum_{j=1}^{J}\left(\overline{\widehat{Y}}^{(m)}_{j,T_1} - \overline{\overline{\widehat{Y}}}^{(m)}_{J,T_1}\right)^{-2}}.$$

From these weights, we then form the STC combination in T_1+1 for panel m:

$$\widehat{Y}^{(m)}_{T_1+1} = \omega^{(m)}_{1,T_1}\widehat{Y}^{(m)}_{1,T_1+1} + \omega^{(m)}_{2,T_1}\widehat{Y}^{(m)}_{2,T_1+1} + ... + \omega^{(m)}_{J,T_1}\widehat{Y}^{(m)}_{J,T_1+1}.$$

This expression must be computed for each panel, $m = 1,2,...,M$. These weights satisfy two restrictions: they are positive and add up to one. Once we get forecasts at T_1+1, we re-compute the weights by rolling over another one-step-ahead combination for T_1+2, and so on, always keeping the same two restrictions.

2.2. Difficulties with the weight combinations

Standard descriptive statistics are not informative with weight combinations. In particular, the arithmetic mean and the variance of individual weights do not fit the Aitchison geometry as value of central tendency and measure of dispersion. These statistics are defined in the framework of Euclidean geometry in real space, which is not a sensible geometry for weights. Therefore, it is necessary to introduce alternatives. They are found in the concept of sample centre (Aitchison, 1997), variation matrix, and total variance (Aitchison, 1986).

The constraints of constant unit sum and relative meaning of the forecasters' weights have important implications for their statistical analysis, thus rendering direct application of multivariate statistical methods misleading or spurious when applied to combinations for various reasons: (see Chayes, 1960 and Barceló-Vidal and Martín-Fernández, 2016).

1. **Nonnormality:** due to the bounded range of values between 0 and 1, instead of $-\infty$ and $+\infty$.

2. **Spurious correlation**

3. **Singularity**: Euclidean (i.e., raw) variance matrices of random weights are always singular due to the constant sum constraint. A classical way to get rid of singularity is to erase one weight, but results will depend on which one is erased, not being an operation that is permutation invariant.

4. **Negative bias**: Some of the Euclidean covariances are forced towards negative values. Hence Euclidean correlations are not free to range over the usual interval $(-1,1)$ subject only to the non-negative definiteness of the variance matrix.

5. **Null-correlation:** With negative bias, what is the meaning of zero correlation between two components of a combination?

6. **Subcombinational Incoherence:** There is no relationship between the Euclidean variance matrix of a subcombination and that of the full combination. Besides, variances may display different and unrelatable rank orderings as we form subcombinations. Note, however, that the ratio of two components remains unchanged when we move from full combination to a subcombination so that as long as we work with scale invariant functions (i.e., ratios), we shall be subcombinationally coherent.

7. **Nonsense of scatterplots** for pairs of forecasters: Since the raw covariance between two weights depends on which other forecasters are reported in the dataset (all combinations are subcombinations of a larger one), there is no guarantee that the Euclidean plot of a closed subcomposition of forecasters exhibits similar or even compatible patterns with the Euclidean plot of the original dataset, even if the forecasters not included in the subcomposition are irrelevant. Thus, a regression line drawn in such a plot cannot be trusted.

8. Finally, the construction of a combination from a vector of Euclidean amounts is a constraining closing operation similar to that of the construction of a vector of subcombinations from the related combination. We may therefore expect the same difficulty in relating variance-covariance matrices of weights in the simplex and those in the the Euclidean space.

Weight combinations are multivariate observations carrying relative information: those following the principle of scale invariance, typically being represented in proportions and percentages. In other words, for combinations the relevant information is contained in (log-)ratios. Combinations thus need an own set of statistical methods and should not be treated with statistical methods made for interval scale data. Instead, combinations should be always treated in a log-ratio-transformed scale. It is quite evident that our dataset can only be combinational if it has at least two forecasters. Otherwise, we cannot speak of a weight in a unit total. That implies a substantial difference between combinational data and other multivariate datasets. Most multivariate analysis begin with a univariate analysis of the individual variables (the marginals), whereas each marginal forecaster of a combinational dataset has no meaning on itself, isolated from the remaining forecasters. One combinational dataset should only use proportional weight values. Therefore, results on a subset of forecasters (subcombination) do not depend on the presence or absence of other irrelevant forecasters in the dataset (subcompositional coherence).

3. The STC approach in the Simplex Space and Combination-after-Selection (CAS)

Traditional decomposition techniques provide inconsistent results when applied to compositional data as they do not recognize the implicit constraints of summing to a constant (Aitchison, 1982, 1986): mathematically, compositional data lie in the bounded space of the simplex while traditional decomposition techniques are defined for data in the real space. Aitchison (1986, pp.79) showed that by making log-ratio transformations it is possible to express compositional data in the real space where the data can be analysed with conventional models and then transformed back into the simplex. For instance, the Aitchison inner product, defined in tems of logratios, turns out to be equivalent to the Euclidean inner product in terms of centred logratios. We make use of the centred log-ratio transformation to express the weights in the real space. The *clr* transformation takes the logarithm of the ratio of each weight divided by the geometric mean of all weights. This transformation maintains the initial constraint in the weights as its elements sum to 0 by construction but resulting values are real. The inverse *clr* transformation takes the data back to the simplex with the closure operator $\mathcal{C}$ that divides the exponential of each clr entry by the sum of all entries.

Consider a $T \times J$ panel $\widehat{\mathbb{Y}}$ of T out-of-sample forecasts $\widehat{Y}_{t,j}$ produced over time by J forecasters on some variable of interest Y_t, and $\mathbb{A}$ be its related panel of prediction accuracies $a_{t,j} \equiv \left(\widehat{Y}_{t,j} - Y_t\right)^{-2} \in \mathbb{R}_+$. Then, the matrix

$$\mathbb{W} \equiv \begin{pmatrix} w_{1,1} & \dots & w_{1,J} \\ \dots & \dots & \dots \\ w_{t,1} & \dots & w_{t,J} \\ \dots & \dots & \dots \\ w_{T,1} & \dots & w_{T,J} \end{pmatrix} \equiv \begin{pmatrix} w'_{1\bullet} \\ \dots \\ w'_{t\bullet} \\ \dots \\ w'_{T\bullet} \end{pmatrix} \equiv \begin{pmatrix} w_{\bullet 1} & \dots & w_{\bullet j} & \dots & w_{\bullet J} \end{pmatrix},$$

with weights $w_{t,j} \equiv a_{t,j} / \sum_{j=1}^{J} a_{tj}$ represents T combination vectors $w_{1\bullet}$, ..., $w_{T\bullet}$ such that $w_{t,j} > 0$ for all t and j, and $\sum_{j=1}^{J} w_{t,j} = 1$ for all t. Thus, $w'_{t\bullet}$ is just a $1 \times J$ point in a simplex space $\mathbb{S}^{J-1}$ of positive weights adding up to one of dimension $J-1$ [2]. The function $\mathcal{C} : \mathbb{R}^J_+ \mapsto \mathbb{S}^{J-1}$ that transforms a vector of precisions $a_{t\bullet} \in \mathbb{R}^J_+$ into a vector of weights $w_{t\bullet} \in \mathbb{S}^{J-1}$ is called a closure transformation $w_{t\bullet} = \mathcal{C}(a_{t\bullet})$. Since this operator cancels out any constant, $\mathcal{C}(ca_{t\bullet}) = \mathcal{C}(a_{t\bullet})$, it is scale invariant. Hence, we just need to work with scale invariant functions (e.g., ratios or logratios). The ratio of two weights remains unchanged when we move from a full combination to a subcombination; that is, $a_{t,i}/a_{t,j} = w_{t,i}/w_{t,j} = s_{t,i}/s_{t,j}$ for all t. Hence, as long as we work with ratios or logratios, we guarantee scale invariance. Therefore, we only consider relative precision among forecasters: each weight in a combination vector has no meaning on itself isolated from the others. Every statement about vectors in $\mathbb{S}^{J-1}$ will be fully expressed in terms of logratios in $\mathbb{R}^{J-1}_+$ with inferences transformed back from $\mathbb{R}^{J-1}_+$ into combinational statements in $\mathbb{S}^{J-1}$. The STC sample centre g of $\mathbb{W}$ is defined as:

$$g = \mathcal{C}\left(\prod_{t=1}^{T} w_{t,1}^{1/T}, ..., \prod_{t=1}^{T} w_{t,J}^{1/T}\right) \equiv \mathcal{C}\left(g\left(w_{\bullet 1}\right), ..., g\left(w_{\bullet J}\right)\right). \quad (1)$$

That is, the point in the simplex given by the closure of the geometric averages of weights over time. It can also be viewed as the inverse function of the clr isomorphic transformation applied to the time average of the sample forecasters' weights (Pawlowsky-Glahn et al., 2015). Note that in this definition, the geometric mean is considered column-wise (i.e., by forecasters), while in the clr transformation the geometric mean is considered row-wise (i.e., by samples).

The centred logratio transformation $\texttt{clr} : \mathbb{S}^{J-1} \mapsto \mathbb{R}$, for each $t = 1, ..., T$,

$$x_{t,j} = \texttt{clr}\left(w_{t,j}\right) := \ln w_{t,j} - \frac{1}{J}\sum_{j=1}^{J} \ln w_{t,j} = \ln \frac{w_{t,j}}{\prod_{j=1}^{J} w_{t,j}^{1/J}} \equiv \ln \frac{w_{t,j}}{g\left(w_{t\bullet}\right)}, j = 1, ..., J, \quad (2)$$

where $g(w_{t\bullet})$ is the geometric average of the J weights for the t^{th} observation. This function may be interpreted as a bijection $\mathbb{S}^{J-1} \leftrightarrow \mathbb{H}^{J-1}$ between $\mathbb{S}^{J-1}$ and a vector subspace of $\mathbb{R}^J_+$ defined by the expression $\mathbb{H}^{J-1} := \left\{x_{t\bullet} \in \mathbb{R}^J_+ : \sum_{j=1}^{J} x_{t,j} = 0\right\}$, orthogonal to the vector of ones. The inverse clr transformation is then defined by

$$\texttt{clrInv}\left(x_{t\bullet}\right) := \mathcal{C}\left(\exp x_{t\bullet}\right) = \mathcal{C}\left(w_{t\bullet}/g\left(w_{t\bullet}\right)\right) = \mathcal{C}\left(w_{t\bullet}\right) = w_{t\bullet} \in \mathbb{S}^{J-1}, \quad (3)$$

that is, $\texttt{clrInv}$ allows us to go from $\mathbb{R}^{J-1}_+$ back to $\mathbb{S}^{J-1}$.

[2] Although in most CoDa papers the superscript of the simplex space is the number of parts, we prefer to emphasize its dimension which, due to the constraint, is $J-1$. This is in line with the dimension of an isomorphic subspace of the real space isometric with the simplex.

The CAS subcombination is defined as $\mathcal{C}(w_1,...,w_I) = (s_1,...,s_I) \in \mathbb{S}^{I-1}$ inside a simplex of a lower dimension $I-1$ so that $s_1 > 0,...,s_I > 0$ and $s_1 + ... + s_I = 1$. Sometimes, especially when $J >> T$, we perform another subsequent selection by choosing those forecasters inside the previous CAS selection.

A selected CAS subcombination $\mathcal{C}\mathbb{S} : \mathbb{S}^{J-1} \mapsto \mathbb{S}^{I-1}$ will be viewed as taking place in two stages: a selection of $I < J$ forecasters by a selecting $I \times J$ matrix $\mathbb{S}$, followed by its closure,

$$\mathcal{C}\mathbb{S}(g) = \mathcal{C}(\mathbb{S}g) := \frac{(w_1,...,w_I)'}{w_1 + ... + w_I} = (s_1,...,s_I)'. \tag{4}$$

For $I = 3$, the CAS subcombination can be represented in a ternary diagram by barycentric coordinates (height of the point over the side of the triangle opposite to it). Similarly, for $I = 4$, it can be represented by a tetrahedron where each possible 3-forecast subcombination vector is found by projecting every 4-forecast vector onto the side opposite to the vertex corresponding to the removed forecasters.

The performance of CAS is good just because we get rid of redundant forecasters (curse of dimensionality), thus increasing the forecast accuracy of simplicial statistics in a simplex of a lower dimension (sometimes just a tetrahedron $J = 4$).

We have also carried out Q-mode clustering (Filzmoser et al., 2018) and biplot (Gabriel, 1971) analyses. The main goal is to achieve highly homogeneous clusters of forecasters' weights; i.e., the weights within a cluster should be very similar to each other. On the other hand, different clusters should be dissimilar, because otherwise they should have been merged into one cluster. The variation matrix Υ with elements given by the sample variance over time, $\Upsilon_{i,j} \equiv var\left(\ln \frac{w_{\bullet i}}{w_{\bullet j}}\right)$, with diagonal elements all 0, will be used to define the total variation in $\mathbb{W}$ as $\upsilon^2 := \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \Upsilon_{i,j}$. Then, υ will be a proper measure of distance among forecasters in cluster analysis, with limit cases of perfect association ($\upsilon = 0$) to perfect independence ($\upsilon = +\infty$). The variation matrix (Aitchison, 1986, or its normalized version) is suitable to express the association between weights. Low values express a high association, and all ratios in a sample are nearly perfectly proportional to each other, while large values express that the ratios are very different from each other. A measure of global dispersion of the weight combinational sample is the total variance (sum of all components of the variation matrix divided by 2J), which turns out to be the time average squared Aitchison distance of each weight combination to the sample centre, also called metric variance (Pawlowsky-Glahn & Egozcue, 2001).

The CAS approach that selects forecasters from the sample centre g of $\mathbb{W}$ can be summarized in the following steps:

1. Given a $T \times J$ table $\widehat{\mathbb{Y}}$ of J forecasters over T time periods in a given season (month or quarter in our cases), compute the related $T \times J$ table $\mathbb{A}$ of $1 \times J$ vectors $a'_{t\bullet}$ of prediction accuracies for each time period $t \in [1,T]$.

2. Convert $\mathbb{A}$ into a $T \times J$ table $\mathbb{W}$ of combination vectors $w'_{t\bullet}$ of weights inside the simplex; that is, weights in each row of $\mathbb{W}$ are positive and add up to one.

3. Calculate the sample centre g of $\mathbb{W}$.

4. Select the CAS subcombination of those forecasters with simplicial weights larger than $1/J$. [3].

5. Repeat steps 1-4 for all panels.

6. Add the next row of out-of-sample accuracy forecasts to the tableau, re-compute the matrix of weights, and update the sample centre and CAS subcombination. Continue this way until the end of the forecast period.

4. Empirical application

We apply the *STC* in the simplex and *CAS* to the variables defined in Table 2, where we include their definition and the samples used to form the combinations of forecasts. Here, we deal with forecasters obtained from the Survey of Professional Forecasters (SPF) from the Federal Reserve Bank of Philadelphia (2018). Blanks in the Survey due to the entry and exit of forecasters are fulfilled following the same strategy as in Poncela et al. (2011), that is, we only consider one-step-ahead forecasts and select only those forecasters without missing data. When there is a missing datum, we use the two-steps-ahead forecast to fill it. Forecasters with more than four consecutive missing data are excluded. For each sample, we only take into account balanced panels. This strategy is also used in Lahiri, Peng and Zhao (2017). Because of the entry and exit of forecasters in the survey, we also analyse different sample sizes, depending on the number of included forecasters. In Table 3, we show, for each variable, the number of forecasters chosen in each subsample. The combinations of forecasters are computed for the periods 2015 to 2018. Note that, in some samples, the number of forecasters is larger than the number of observations, a fact that cannot be treated with other methods (e.g., regression and PCA).

[3] When $J >> T$, we made a first subselection by applying cluster and biplot analyses. Redundant forecasts were defined, with the former, as those whose weights belong to the same cluster; and, with the latter, as those lying on a common line. The sample centre of the remaining weights were then chosen prior to using the CAS strategy.

Table 2. *Definition of the main variables used in the application. Source: Survey of Professional Forecasters documentation. SA = Seasonal Adjusted.*

Variable	Definition	Sample
NGDP	Forecasts for the quarterly level of nominal GDP. SA. billions $	1991 Q1 - 2018 Q4
PGDP	Forecasts for the quaterly level of the chian-weighted GDP price index. SA. Index. Base year 1992	1991 Q1 - 2018 Q4
UNEMP	Forecasts for the quarterly average unemployment rate. SA. % points	1991 Q1 - 2018 Q4
EMP	Forecasts for the quarterly average level of nonfarm payroll employment. SA. Thousands of jobs.	2004 Q1 - 2018 Q4
INDPROD	Forecasts for the quarterly average level of the index of industrial prod. SA. Index.	1991 Q1 - 2018 Q4
HOUSING	Forecasts for the quarterly average level of housing starts. SA. millions.	1991 Q1 - 2018 Q4
TBILL	Forecasts for the quarterly average 3-months Treasury Bill rates. % points	1991 Q1 - 2018 Q4
BOND	Forecasts for the quarterly average level of Moody's Aaa corporate. Bond yield. % points	1991 Q1 - 2018 Q4
RGDP	Forecasts for the quarterly chain-weighted real GDP. SA. annual rate. Base years 1992 - 1995, fixed weighted real GDP	1991 Q1 - 2018 Q4
RCONSUM	Forecasts for the quarterly chain-weighted real personal consumption expenditures. SA, annual rate, base years 1992 - 1995.	1991 Q1 - 2018 Q4
RNRESIN	Forecasts for the quarterly chain-weighted real nonresidential fixed investment. SA. annual rate, base years 1992 - 1995.	1991 Q1 - 2018 Q4
RRESINV	Forecasts for the quarterly chain-weighted real residential fixed investment. SA., annual rate, base years 1992 - 1995	1991 Q1 - 2018 Q4
RFEDGOV	Forecasts for the quarterly chain-weighted real federal government consumption and gross investment. SA, annual rate, base years 1992-95	1991 Q1 - 2018 Q4
RLSGOV	Forecasts for the quarterly level of chain-weighted real state and local government consumption and gross investment. SA. annual rate. base years 1992 - 1995	1991 Q1 - 2018 Q4
CPI	Forecasts for the headline CPI inflation rate. SA, annual rate, % points. Quarterly forecasts are annualized quarter-overquarter percent changes of the quarterly average price index level	1991 Q1 - 2018 Q4

Table 3. *Variables, samples and number of forecasters.*

	Samples									
Variable	*Sample* (1)		*Sample* (2)		*Sample* (3)		*Sample* (4)		*Sample* (5)	
	T	*J*	*T*	*J*	*T*	*J*	*T*	*J*	*T*	*J*
NGDP	24	3	$20^{a)}$	6	$15^{d)}$	10	$9^{g)}$	18	5	22
PGDP	24	3	$20^{a)}$	6	$15^{d)}$	10	$9^{g)}$	20	5	25
UNEMP	24	4	$20^{a)}$	6	$15^{d)}$	12	$9^{g)}$	22	5	27
EMP			11	16	10	20	8	22	5	28
INDPROD	24	4	$19^{b)}$	8	$15^{d)}$	12	$9^{g)}$	21	5	26
HOUSING	24	4	$19^{b)}$	10	$15^{d)}$	15	$10^{f)}$	19	5	26
TBILL	24	5	$19^{b)}$	8	$15^{d)}$	11	$9^{g)}$	19	5	24
BOND	24	3	$19^{b)}$	5	$14^{e)}$	7	$9^{g)}$	13	5	17
RRESINV	24	5	$20^{a)}$	9	$15^{d)}$	13	$9^{g)}$	19	5	28
RGDP	24	5	$20^{a)}$	9	$15^{d)}$	14	$9^{g)}$	25	5	31
RCONSUM	24	5	$20^{a)}$	9	$16^{c)}$	13	$10^{f)}$	20	5	29
RNREIN	24	5	$20^{a)}$	9	$16^{c)}$	13	$10^{f)}$	20	5	29
RFEDGOV	24	5	$20^{a)}$	9	$16^{c)}$	13	$10^{f)}$	19	5	28
RLSGOV	24	5	$20^{a)}$	9	$16^{c)}$	13	$10^{f)}$	19	5	28
CPI	24	5	$20^{a)}$	8	$16^{c)}$	12	$10^{f)}$	19	5	29

T = number of periods, J = number of forecasters, Sample (1): 1991 - 2014; Sample (2) a) 1995 - 2014; b) 1996 - 2014; Sample (3) c) 1999 - 2014; d) 2000 - 2014; e) 2001 - 2014; Sample (4) f) 2005 - 2014; g) 2006 - 2014; Sample (5) 2010 - 2014; For the EMP variable the samples are: (1) 2004-2014; (2) 2005-2014; (3) 2007-2014 and (4) 2010-2014

To analyse the prediction accuracy of combinations, we look at four well-known measures: Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Median Absolute Percentage Error (MdAPE). The definitions of the accuracy measures are:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right);\ RMSE = \sqrt{\frac{\sum\limits_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n}};$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right|;\ MdAPE = Median\left(\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right|\right).$$

Although in general these measures produce similar results, there are some differences depending on the type of the combination considered.[4]

[4] In previous studies, we also used Mean Absolute Scaled Error (MASE), and it made no difference with MAPE as to which method generates better results.

We compute four kinds of combinations, three with varying weights: Euclidean space, E_STC, simplex space, S_STC and CAS and the fixed-weight arithmetic average, AVE.

4.1. General results

Table 4. *Summary depending on J and T.*

	AVE	E_STC	S_STC	CAS	$EUCLIDEAN$	$SIMPLEX$	$TOTAL$
$J < T$	215	171	73	289	386	362	758
(%)	(28.3)	(22.6)	(9.6)	(39.5)	(50.4)	(49.1)	(59.9)
$J > T$	115	102	65	227	217	292	509
(%)	(22.6)	(20.0)	(12.8)	(44.6)	(42.6)	(57.4)	(40.2)
$TOTAL$	330	273	138	526	603	664	1266
(%)	(26.0)	(21.6)	(10.9)	(41.5)	(47.6)	(52.4)	

Number of times that accuracy measures favored a combination procedure.Percentages in parenthesis.

We have analyzed 1266 values of accuracy measures. General results are shown in Table 4. According to the type of weights, they favored fixed weights in 330 cases (26.0%) and varying weights in 937 (74.0%). With respect to the latter, 273 (21.6%) favored E_STC, 138 (10.9%) S_STC and 526 (41.5%) CAS. Although the CAS procedure is clearly favored, there is not a clear difference when we compare the results between Euclidean and simplex spaces. In fact, when the combinations are done in a sample with more observations than forecasters, the Euclidean combinations (AVE and E_STC) generate results as good as those obtained with the simplex (50.4% vs 49.1%); but clearly CAS is the best, with a 39.5% of the cases.

When we focus on the results for $J > T$, simplex is better (57.4% vs 42.6%), CAS works very well precisely when some other methods have little to say.

Table 5. *Results for each combination procedure by variable and accuracy criteria. Percentages of beats*

	Mean Error				RMSE				MAPE				MdAPE			
	AVE	*E_STC*	*S_STC*	*CAS*	*AVE*	*E_STC*	*S_STC*	*CAS*	*AVE*	*E_STC*	*S_STC*	*CAS*	*AVE*	*E_STC*	*S_STC*	*CAS*
NGDP	40.0	0.0	10.0	50.0	52.4	0.0	4.8	42.9	30.0	45.0	5.0	20.0	30.0	45.0	5.0	20.0
PGDP	13.6	0.0	45.5	40.9	40.0	0.0	30.0	30.0	5.0	30.0	35.0	30.0	10.0	30.0	30.0	30.0
UNEMP	35.0	0.0	5.0	60.0	40.0	0.0	20.0	40.0	19.0	42.9	0.0	38.1	25.0	40.0	0.0	35.0
EMP	75.0	0.0	6.3	18.8	81.3	0.0	0.0	18.8	68.8	6.3	0.0	25.0	68.8	25.0	0.0	6.3
INDPROD	30.0	0.0	5.0	65.0	25.0	0.0	5.0	70.0	20.0	10.0	0.0	70.0	10.5	0.0	10.5	78.9
HOUSING	0.0	0.0	19.0	76.2	5.0	0.0	20.0	75.0	0.0	10.0	5.0	85.0	0.0	10.0	5.0	85.0
TBILL	10.0	90.0	0.0	0.0	5.0	90.0	0.0	5.0	28.6	42.9	9.5	19.0	15.0	60.0	10.0	15.0
BOND	10.0	0.0	5.0	85.0	18.2	0.0	9.1	72.7	15.0	5.0	15.0	65.0	15.0	0.0	10.0	75.0
RRESIN	30.0	0.0	5.0	35.0	55.0	0.0	15.0	30.0	30.0	45.0	5.0	20.0	23.8	61.9	0.0	14.3
RGDP	15.0	0.0	0.0	85.0	10.0	0.0	20.0	70.0	10.1	35.0	5.0	50.0	10.1	40.0	5.0	45.0
RCONSUM	15.0	0.0	15.0	70.0	30.0	0.0	10.0	60.0	35.0	0.0	5.0	60.0	30.0	0.0	0.0	70.0
RNRESIN	20.0	0.0	25.0	55.0	25.0	0.0	25.0	50.0	5.0	90.0	0.0	5.0	4.8	85.7	4.8	4.8
RFEDGOV	40.0	0.0	25.0	356.0	60.0	5.0	25.0	10.0	57.1	14.3	19.0	9.5	50.0	30.0	15.0	5.0
RLSGOV	40.0	0.0	10.0	50.0	50.0	0.0	20.0	30.0	45.0	25.0	10.0	20.0	15.0	35.0	25.0	25.0
CPI	0.0	30.	15.0	25.0	15.0	35.0	0.0	50.0	0.0	50.0	10.0	40.0	10.0	55.0	5.0	30.0
MEAN	26.9	10.3	12.7	50.1	34.1	8.7	13.6	43.6	24.6	30.1	8.2	37.1	21.2	34.5	8.4	35.9

4.2. Results by method of combination, variable and accuracy criteria

Table 5 shows the percentage of beats by variable and accuracy criteria for each combination procedure. The following comments are worth mentioning:

1. Results about Euclidean and simplex spaces vary depending on the accuracy measure considered. Whereas combinations in the former are clearly better with MAPE and MdAPE, those in the latter are better with MAE and RMSE.
2. When we analyse combinations according to the type of weights, fixed weights are always the worst, therefore it is worthwile to use varying weights.
3. *CAS* is on average the best, reaching 50% of the cases with ME.
4. *S_STC* is only the best for the *PGDP* considering ME, MAPE and MdAPE, whereas *E_STC* is the best for several variables when we consider MAPE and MdAPE.
5. *AVE*'s best results occur with RMSE.

4.3. Results by number of forecasters and accuracy criteria

Tables 6 and 7 show the results of each combination by the number of forecasters and accuracy criteria.

Table 6. *Number of beats of each combination by accuracy criteria and number of forecasts.*

	Mean Error				*RMSE*				*MAPE*				*MdAPE*			
	AVE	*E_STC*	*S_STC*	*CAS*	*AVE*	*E_STC*	*S_STC*	*CAS*	*AVE*	*E_STC*	*S_STC*	*CAS*	*AVE*	*E_STC*	*S_STC*	*CAS*
$J < T$	52	21	23	95	66	18	26	75	51	65	12	66	46	67	12	63
(%)	(27.3)	(10.8)	(11.9)	(50.0)	(35.7)	(9.9)	(14.0)	(40.4)	(26.3)	(33.5)	(6.25)	(34.1)	(24.3)	(35.8)	(6.4)	(33.5)
$J > T$	31	13	17	63	37	10	20	61	27	36	15	50	20	43	13	53
(%)	(25.2)	(10.4)	(13.9)	(50.4)	(29.1)	(7.7)	(15.4)	(47.9)	(21.2)	(28.0)	(11.9)	(39.0)	(15.1)	(33.6)	(10.1)	(41.18)
TOTAL	83	34	40	158	103	28	46	135	78	101	27	116	65	111	25	116
(%)	(26.4)	(10.7)	(12.7)	(50.2)	(33.0)	(9.3)	(14.6)	(43.4)	(24.2)	(31.3)	(8.4)	(36.0)	(20.6)	(34.9	(7.9)	(36.6)

Table 7. *Number of beats of EUCLIDEAN and SIMPLEX combinations by accuracy criteria and number of forecasts.*

	Mean Error		*RMSE*		*MAPE*		*MdAPE*	
	EUCLIDEAN	*SIMPLEX*	*EUCLIDEAN*	*SIMPLEX*	*EUCLIDEAN*	*SIMPLEX*	*EUCLIDEAN*	*SIMPLEX*
$J < T$	73	118	85	101	116	78	113	75
(%)	(38.1)	(61.9)	(45.6)	(54.4)	(59.8)	(40.2)	(60.1)	(39.9)
$J > T$	44	80	47	80	63	65	63	66
(%)	(35.7)	(64.4)	(36.8)	(63.3)	(49.2)	(50.9)	(48.7)	(51.3)
TOTAL	117	198	131	181	179	143	176	141
(%)	(37.1)	(62.9)	(42.0)	(58.0)	(55.6)	(44.4)	(55.5)	(44.5)

In Table 6, we present the number of beats of each combination according to the accuracy criteria and the number of forecasters. Only in 2 cases, E_STC beats CAS and always when the number of forecasters is lower than that of observations. In the rest of the cases, CAS is always the best reaching 50% of ME, almost twice of AVE combination.

In Table 7, we show the results attending to the space where the combination is formed. As much as $J > T$ simplex is always the best, reaching more than 60% of the cases for ME and RMSE. When $J < T$, for MAPE and MdAPE, Euclidean combinations are better. These good results are obtained because E_STC works very well depending on these measures.

4.4. Results according to the variability of the forecasts

The basic idea under this section is the following: a fixed-weight combination assigns the same weight to forecasts, so if variability among them is small, then the average will work well in the same direction, however wrong it may be ('precisely' wrong) unless they are unbiased. On the other hand, when variability is high, it is better to assign different weights. This is in line with the results obtained by Jose and Winkler (2008) by comparing the accuracy of the average with trimmed and Winsorized averages and the results by Genre et al. (2013) by using the European Central Bank (ECB) survey of professional forecasters. In this latter paper, they find that some combination methods outperform the simple average of forecasts in variables with heterogeneity of forecasters and apparent bias.

In order to verify this hypothesis, we compute the variation coefficient (VC) of each variable for each combination and forecast period from 2015 to 2018. We also plotted the forecasts for each period [5]. In fact, this issue forms part of the selection procedure presented in this paper, i.e. to select those forecasters that do not share common information. In this empirical application, the forecasters come from the Survey of Professional Forecasters (SPF) and may have common information in forming their forecasts. This is the reason why we expect some forecasts to be highly correlated (even redundant) and others with low correlation. Then, CAS takes advantage of this situation and usually generates better results.

The main comments that can be pointed out are the following:

1. When all the forecasts included in the sample are highly correlated and their plots show a similar behaviour, AVE is usually the best combination. A clear example of this situation is shown in Figure 1 where we plot the forecasts for $NGDP$ for all the samples.

[5] In order to save space, these results are available upon request.

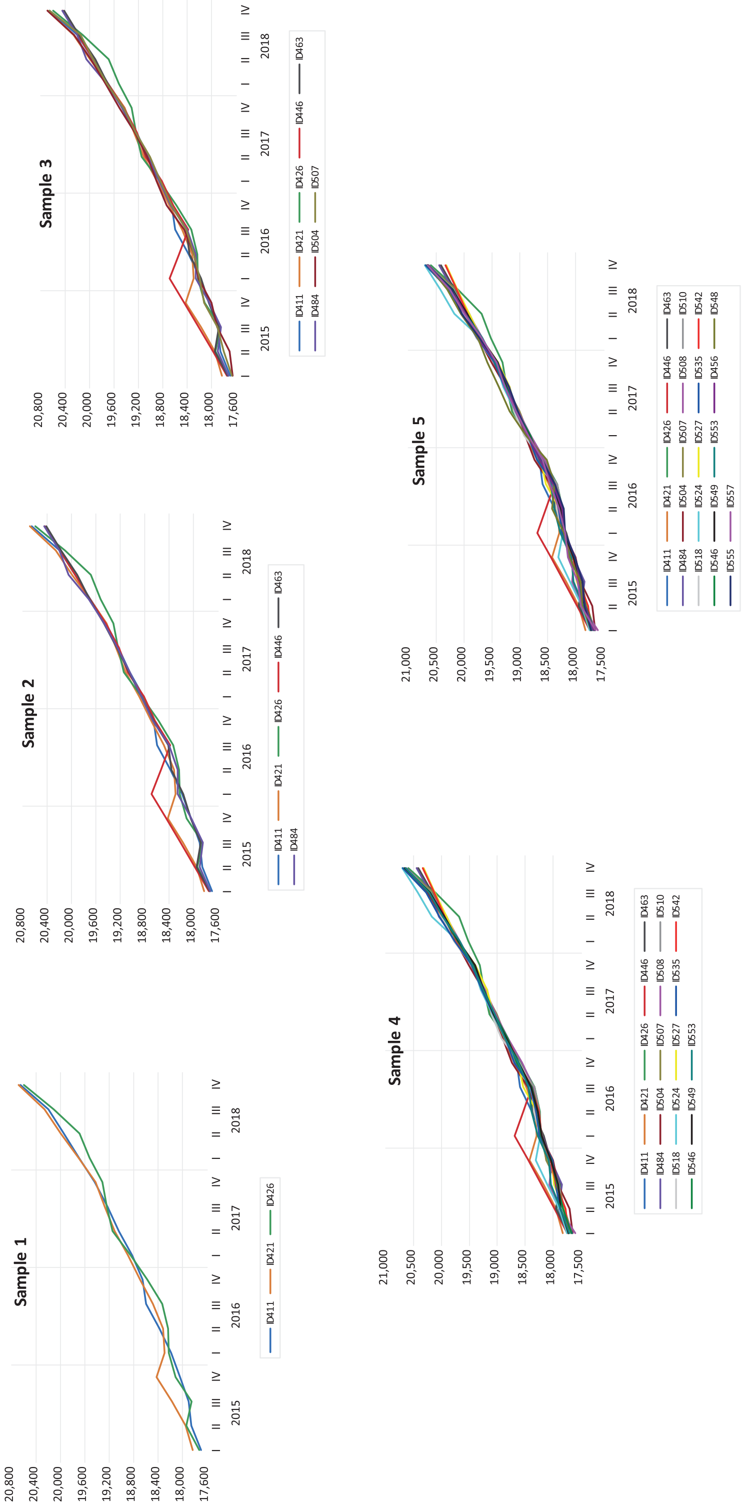


Figure 1. *NGDP forecasts by samples.*

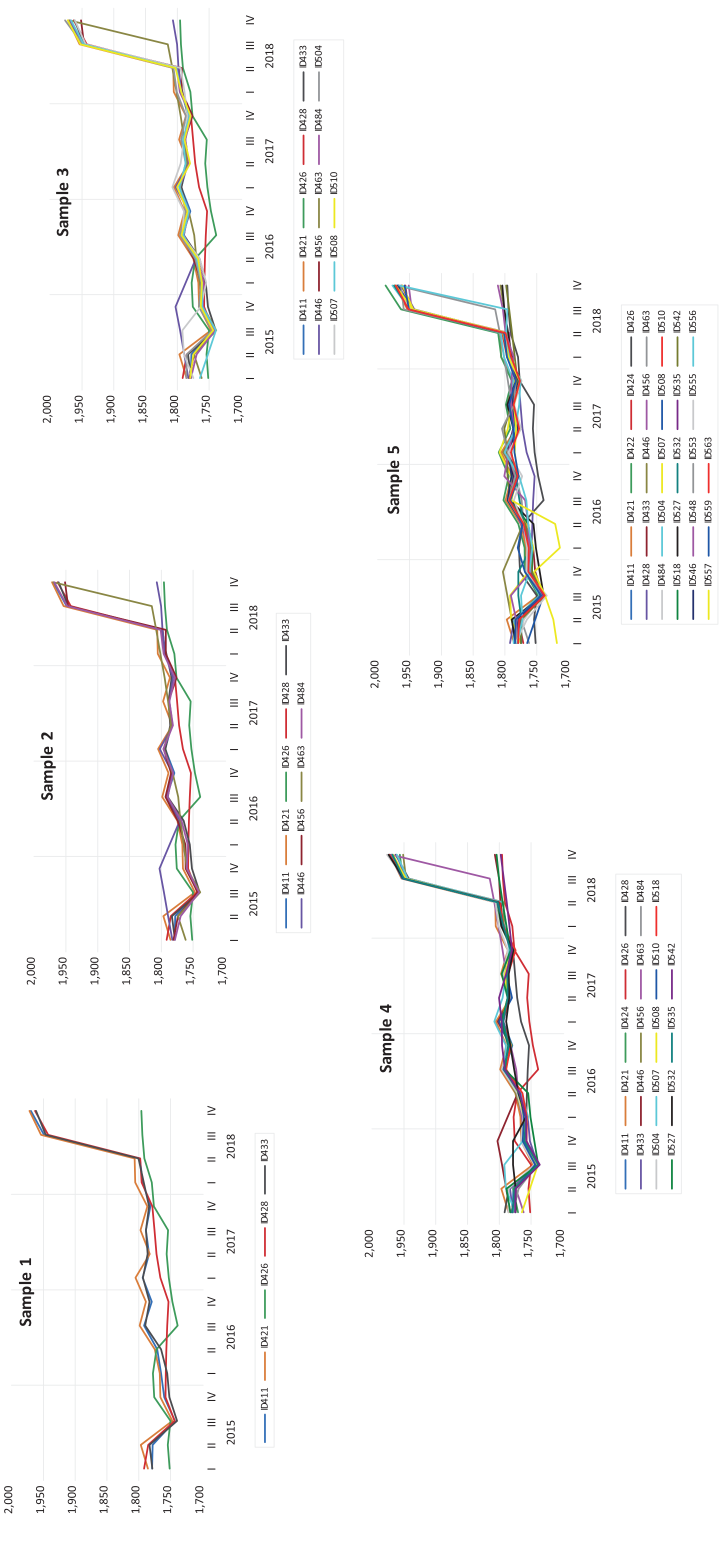


Figure 2. *RLSGOV forecasts by samples.*

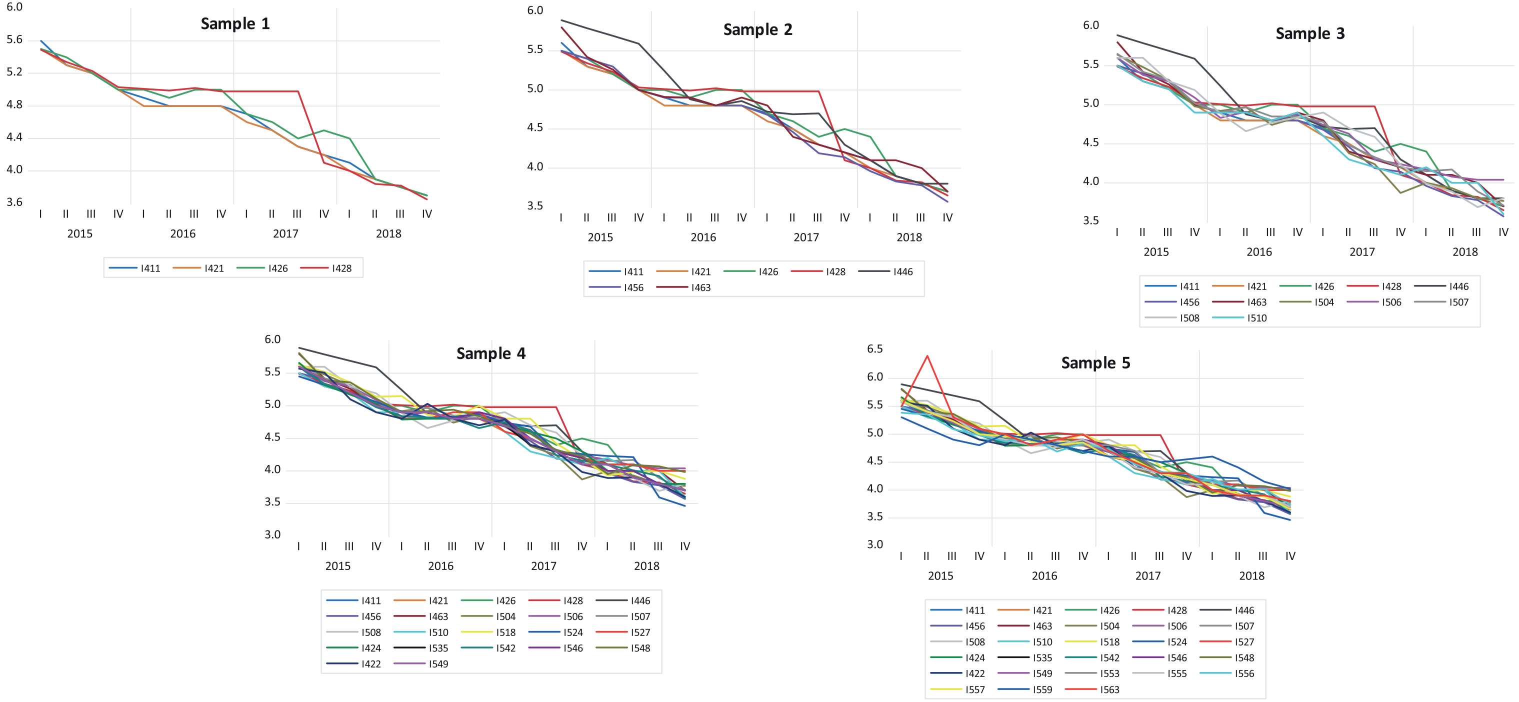


Figure 3. *UNEMP forecasts by samples.*

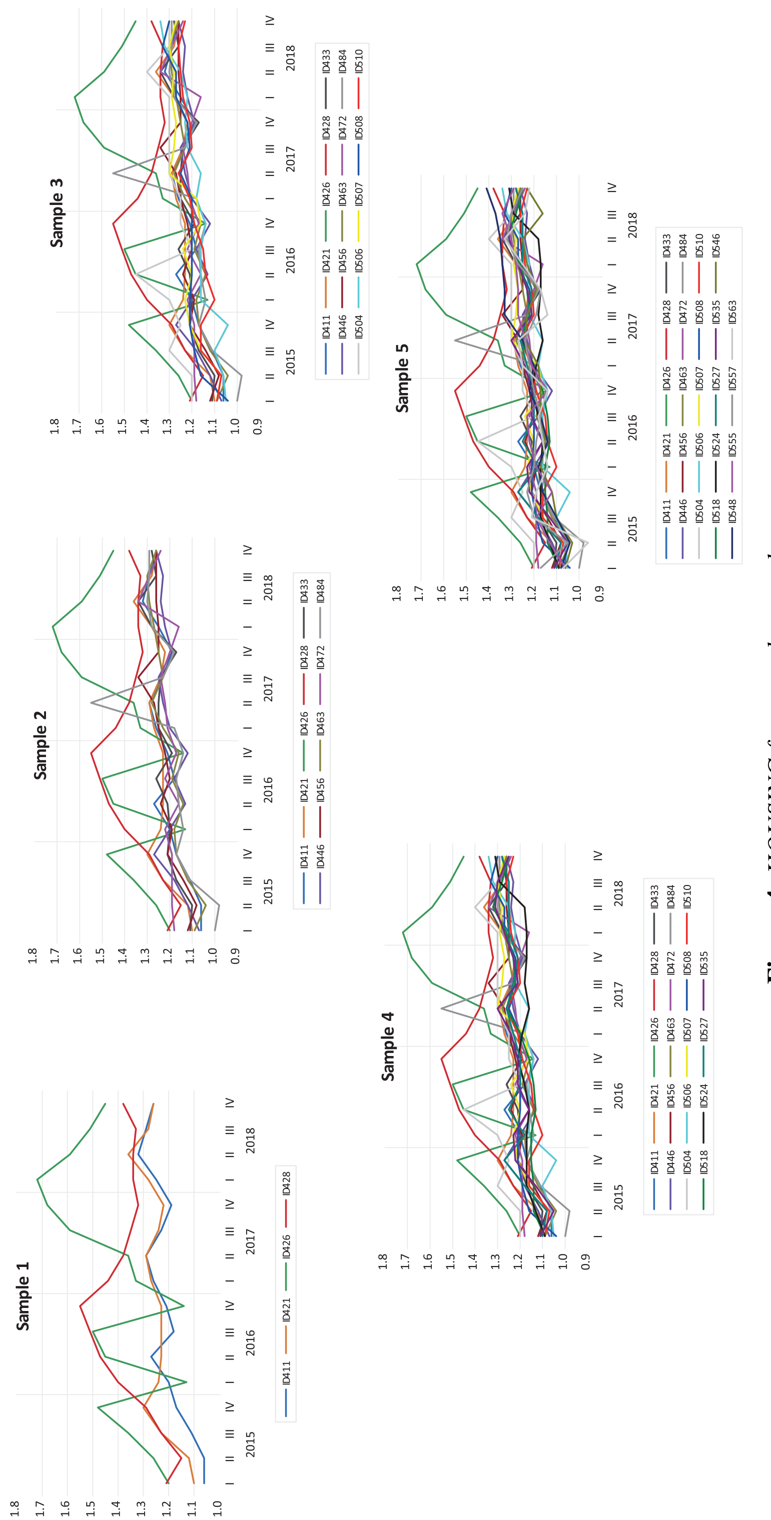


Figure 4. *HOUSING forecasts by samples.*

2. When some of the forecasts are correlated but their plots differ somewhat, AVE is better because of its varying-weight allocation. Figure 2 shows this situation for $RLSGOV$.

3. In a mixed situation with some forecasts highly correlated and some others not so, *CAS* is the best because it only selects non-redundant forecasters. In Figure 3 we show this situation for $UNEMP$.

4. In general, with low correlated forecasts, varying-weight combinations generate better results: the E_STC and S_STC, when the forecasts show a similar behaviour, and *CAS*, when they don't. Figure 4 shows a clear example of this situation for $HOUSING$.

Table 8 shows the variation coefficient (VC) and results for the aforementioned variables[6]. The analysis of the VC will be done jointly with Figures 1 to 4.

1. **NGDP**: All the graphs in Figure 1 show very little variation between forecasts. The VC in each sample is very low, suggesting that AVE should be used. Looking at the combination results, AVE is the winner in all the samples with the exception of sample 4. In this case, *CAS* generates the best forecasts for all the forecasting periods. Notice that in the graph for sample 4, although the forecasts follow a similar behaviour, there are some of them with different patterns that can be used to improve the forecast combination through *CAS*.

2. **RLSGOV**: The behaviour of the forecasts for this variable is different from the one observed before. In this case, the forecasts seem to have a similar behaviour, but the correlation between them is not too high. Then, assigning different weights generates better combinations. Looking at Figure 2, we can see that S_STC obtains very good results in 2017 and perhaps in 2016. Our perception from the graph is confirmed in Table 7: varying-weight combinations outperform the fixed-weight one. This situation is also supported by the VC, which shows higher values than the observed for *NGDP*. So, in this case, the fact that not all the forecasts show the same pattern leads to better forecasting results with varying-weight methods.

3. **UNEMP**: The VC of this variable in Table 7 clearly shows higher values than the observed for the previous variables. This fact can indicate that the average forecast may not be the best combination in this case. Looking at Figure 6, not all the forecasts have the same pattern. This favors the varying-weight combinations, E_STC, S_STC and *CAS*, the latter being the one that beats more times. Therefore, in this case, selection is better than a full combination either fixed AVE or varying E_STC.

[6] The VC, figures and results for the other variables are available upon request. They have been omitted to save space.

Table 8. *Coefficients of variation for selected variables and number of beats of the combination procedures by samples.*

	NGDP					*UNEMP*					*RLSGOV*					*HOUSING*				
	CV	*AVE*	*E_STC*	*S_STC*	*CAS*	*CV*	*AVE*	*E_STC*	*S_STC*	*CAS*	*CV*	*AVE*	*E_STC*	*S_STC*	*CAS*	*CV*	*AVE*	*E_STC*	*S_STC*	*CAS*
Sample (1)		9	2	3	2		5	4	2	5		9	5	0	2		0	0	0	16
2015	0.667	2	2	0	0	0.603	2	2	0	0	0.588	3	1	0	0	7.970	0	0	0	4
2016	0.405	0	0	3	1	2.150	0	2	2	0	0.724	3	1	0	0	11.287	0	0	0	4
2017	0.313	3	0	0	1	4.927	0	0	0	4	0.695	0	2	0	2	9.604	0	0	0	4
2018	0.409	4	0	0	0	1.575	3	0	0	1	0.324	3	1	0	0	9.850	0	0	0	4
Sample (2)		8	1	3	4		7	3	1	5		4	1	4	7		1	2	4	9
2015	0.830	4	0	0	0	3.447	1	1	0	2	0.834	0	0	0	4	6.813	0	0	0	4
2016	0.365	1	0	2	1	2.011	3	0	1	0	0.627	3	0	1	0	9.079	0	0	1	3
2017	0.245	0	1	1	2	4.080	0	2	0	2	0.566	1	0	3	0	8.518	0	0	2	2
2018	0.465	3	0	0	1	2.434	3	0	0	1	1.666	0	1	0	3	7.486	1	2	1	0
Sample (3)		6	6	1	3		5	4	1	6		2	4	3	7		0	1	0	11
2015	0.741	2	2	0	0	2.743	0	2	0	2	0.807	0	1	0	3	6.575	0	0	0	4
2016	0.342	1	0	1	2	1.829	3	1	0	0	0.550	2	1	1	0	8.528	0	0	0	4
2017	0.217	1	2	0	1	3.804	0	1	0	3	0.504	0	0	0	4	7.485	0	1	0	3
2018	0.392	2	2	0	0	2.951	2	0	1	1	1.418	0	2	2	0	6.476	0	0	0	4
Sample (4)		2	5	0	10		6	4	1	5		9	3	2	2		0	3	1	13
2015	0.590	0	2	0	2	2.253	0	0	0	4	0.725	3	0	0	1	6.188	0	0	0	4
2016	0.306	2	2	0	1	1.766	2	2	0	0	0.499	3	1	0	0	7.665	0	0	1	3
2017	0.248	0	1	0	3	3.104	1	1	1	1	0.446	3	0	1	0	6.814	0	2	0	2
2018	0.502	0	0	0	4	3.132	3	1	0	0	1.751	0	2	1	1	6.247	0	0	0	4
Sample (5)		6	5	0	5		1	2	0	13		6	1	4	5		0	0	6	10
2015	0.546	1	2	0	1	2.384	0	0	0	4	0.803	3	1	0	0	5.869	0	0	1	3
2016	0.360	1	2	0	1	1.358	0	0	0	4	1.920	1	0	3	0	6.403	0	0	1	3
2017	0.233	2	0	0	3	3.043	0	0	0	4	0.429	0	0	0	4	6.320	0	0	0	4
2018	0.459	3	1	0	0	3.463	1	2	0	1	1.708	2	0	1	1	5.825	0	0	4	0

4. **HOUSING**: Figure 4 is a clear example for *CAS* to form a combination. Different behaviour of some forecasts and high VC are the clues to select forecasters to obtain better forecasting results. Although there is a common behaviour of some forecasts, the selection of orthogonalized forecasts improves the results.

Similar results are confirmed for the other variables analysed in the empirical application. As a matter of fact, high VC and different behaviour might be the clues to consider CAS as the best subcombination to forecast a variable.

4.5. Results according to the forecast ability

When the Diebold and Mariano (1995) or Giacomini and White (2006) tests are not appropriate, it might be interesting to break down the Mean Squared Forecast Error (MSFE) into three components (bias, variance, and covariance) to assess which of them holds sway over a given MSFE:

$$\begin{aligned} MSFE := \frac{1}{H}\sum_{h=1}^{H}\left(\widehat{Y}_{T+h}-Y_{T+h}\right)^2 &\equiv \left(\overline{\widehat{Y}}_H-\overline{Y}_H\right)^2+\left(sd(\widehat{Y}_H)-sd(Y_H)\right)^2 \\ &+2\left(sd(Y_H)\right)\left(sd(Y_H)\right)\left(1-corr[\widehat{Y}_H,Y_H]\right), \end{aligned} \tag{5}$$

where $\overline{\widehat{Y}}_H$ is an H-period average forecast, $\overline{Y}_H$ is the corresponding average for the realized values (Y_H), $sd(\widehat{Y}_H)$ is the standard deviation of the forecasts, $sd(Y_H)$ is the standard deviation of the realized values for the forecast period, and $corr[\widehat{Y}_H, Y_H]$ is the correlation between forecasts and realized values. Then, proportions are defined as follow:

$$\text{Bias proportion: } \frac{\left(\overline{\widehat{Y}}_H-\overline{Y}_H\right)^2}{MSFE},$$

$$\text{Variance proportion: } \frac{\left(sd(\widehat{Y}_H)-sd(Y_H)\right)^2}{MSFE},$$

$$\text{Covariance proportion: } \frac{2\left(sd(Y_H)\right)\left(sd(Y_H)\right)\left(1-corr[\widehat{Y}_H,Y_H]\right)}{MSFE},$$

We study which one constributes more to the MSFE. A ranking of preferences may be given by the following four situations:

1. CASE 1: The best will be when there are little bias and variance (hence, high covariance proportion).

2. CASE 2: The next one will be when there is little bias, but high variance (hence, low covariance proportion).

3. CASE 3: Bad situations happen when the bias is high: either with high variance,

4. CASE 4: Or the worst, with low variance ('precisely' wrong).

Using this classification, we show in Table 9 the bias, variance, and covariance proportions for the combination procedures with lowest MSFE[7] and in Table 10 we summarize this information according to Euclidean and simplex combinations.

Table 9. *Classification according to their forecast ability.*

	AVE		*E_STC*		*S_STC*		*CAS*		*TOTAL*	
	#	%	#	%	#	%	#	%	#	%
Case 1	10	13.3	3	4.0	1	1.4	3	4.0	17	5.67
Case 2	0	0.0	18	24.0	28	37.8	18	24.0	65	21.67
Case 3	8	10.7	34	45.3	39	52.7	44	58.7	125	41.67
Case 4	57	76.0	20	26.7	6	8.1	10	13.3	93	31.0

#: Proportions of the best MSFE procedure included in specific cases.

Table 10. *Classification according to Euclidean or Simplex.*

	Euclidean		*Simplex*	
	#	%	#	%
Case 1	13	8.7	4	2.7
Case 2	18	12.0	46	30.9
Case 3	42	28.0	83	55.7
Case 4	77	51.3	16	10.7

#: Proportions of the best MSFE procedure included in specific cases.

From Table 9, we can conclude that AVE is mainly classified in the worst situation: high bias and low variance (76% of the cases), but it is also the first method classified in the best situation (13.3% of the cases).

In general, the other methods are classified most of the times in cases 2 and 3 (low bias and hig variance or high bias and high variance).

From Table 10, the case 3 is the most often with the simplex representing more than 50% of the cases, being case 2 the second best situation that happens almost 31%.

Considering the different methods of combination, we obtain that for AVE, case 4 is the most often whith MSFE. For all the others, case 3 is the one that happens most often.

[7] The specific values for the bias, variance, and covariance proportions for each variable, each sample, and each combination procedure are available upon request.

5. Conclusions

In this paper, we have used the Split-Then-Combine (STC) approach to build positive weights that sum up to one. Because of these two restrictions, most methods from multivariate statistics are inapplicable for combinational datasets, giving rise to a number of issues that make inappropiate the Euclidean geometry. Instead, the Aitchison geometry considers combinations of forecasters inside the simplex, the sampling space of positive weights adding up to one. A one-to-one transformation between the simplex and real spaces allows us to use the sample centre of the simplex, with time-varying weights, to find a Combinations after Selection (CAS) simplicial subcombinations that selectis those forecasters in a full combination that assign higher weights than the one allocated by the benchmark average.

The methodology can be summarized in these steps: first, we split experts' forecasts by seasons to assess their relative forecast performance that periodically evolves over time. Second, we choose as a combination vector the sample centre of the simplex. Then, we select forecasters inside a simplex of lower dimension by means of a centred logratio transformation. Finally, we make rolling, truly out-of-sample, one-step-ahead combinations of forecasts, even in cases where the sample size is smaller than the number of forecasters. Once a new observation is known, we recalculate the weights that we then keep one-step-ahead to form a new out-of-sample combination.

We present experimental results with a pool of expert forecasters of the US macroeconomy over the period 1991–2018. In most cases, the Combination after Selection strategy improves the average (neutral combination in the simplex space) with different criteria of forecasting accuracy, and works very well even when the number of forecasters is greater than the number of observations. Forecast combination can improve forecasting accuracy, provided that the sets of forecasters contain some independent information.

As a general rule, we can conclude that when there are a high number of heterogeneous forecasters to be combined, the best way to form a combination is by selecting a CAS simplicial subcombination formed by the most weighted, non-redundant forecasters.

For combinations of forecasts, the relevant information is contained in the clr coefficients between forecasts. This by itself might also be interesting to symmetrize possible right-skewed distributions of forecaster's precisions. Further research, therefore, will focus on pivot (or more general orthonormal) coordinates that aim to extract all relative information about a particular forecast in the combination. Moreover, exploratory and preprocessing issues may also be discussed: visualization, outlier detection, missing values, and zeros form a touchstone of the logratio analysis. Finally, many popular statistical methods, such as principal component analysis, cluster analysis, classification and regression analysis, may be adapted for dealing with combinations carrying relative information.

Ackowledgements The second author acknowledges financial support from the Spanish Government Project PID2019-108079GBC22.

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Sociesty, B*, 44, pp. 139-177.

Aitchison, J. (1986). *The statistical analysis of compositional data*, Monographs on Statistics and Applied Probability. London: Chapman & Hall, (Reprinted in 2003 with additional material by The Blackburn Press).

Aitchison, J. (1992). On criteria for measures of compositional difference, *Mathematical Geology*, 24, 365-379.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), Proceedings of IAMG'97—The third annual conference of the International Association for Mathematical Geology, Volume I, II and addendum (pp. 3-35). Barcelona: International Center for Numerical Methods in Engineering (CIMNE), pp. 1100.

Arroyo, A.S.M. and de Juan Fernández, A. (2014). Split-then-Combine method for out-of-sample combinations of forecasts, *Journal of Business Administration Research*, 3 (1), pp. 19-37.

Barceló-Vidal, C. and Martín-Fernández, J.A. (2016). The mathematics of Compositional Analysis, Austrian *Journal of Statistics,* 45, pp. 57-71, doi: 10.17713/ajs.v45i4.142.

Barrow, D.K. and Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: implications for inventory management. *International Journal of Production Economics,* 177, pp. 24-33.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition, *Journal of the American Statitictical Asociation*, 96 (456), 2001, pp. 1205-1214.

Van den Boogaart, K.G. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*, Berlin: Springer.

Bujosa-Brun, M., García-Ferrer, A., de Juan, A. and Martín Arroyo, A. (2020). Evaluating early warning and coincident indicators of business cycles using smooth trends, *Journal of Forecasting, 39, pp. 1-17. DOI: 10.1002/for.2601.*

Chayes, F. (1960). On correlation between variables of constant sum, *Journal of Geophysical Research* 65 (12), pp. 4185-4193.

Coenders G, and Ferrer-Rosell, B. (2020). Compositional data analysis in tourism. Review and future directions. *Tourism Anal* 25: pp. 153-168.

Conflitti, C., De Mol, C. and Giannone, D. (2015). Optimal combination of survey forecasts, *International Journal of Forecasting* 31 pp. 1096-1103.

Diebold, F.X. and Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, vol 13(3). pp. 253-263.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37 (7), 2005, pp. 795-828.

Egozcue, J.J. and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its struture. *Test*, 28(3), pp. 599-638.

Federal Reserve Bank of Philadelphia (2018). Survey of Professional Forecasters documentation. August 29, 2018.

Filzmoser, P., Hron, K., and Templ, M. (2018). *Applied Comopositional Data Analysis*, Springer.

Gabriel, K.R. (1971). The biplot—graphic display of matrices with application to principal component analysis. *Biometrika*, 58 (3), pp. 453-467.

Genre, V., Kenny, G., Meyler, A. and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average?. *International Journal of Forecasting,* 29, pp. 108-121.

Giacomini, R. and White, H. (2006). Test of conditional predictive ability, *Econometrica*, 74(6), pp. 1545-1578.

Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and its Application*, 8, pp. 271-299.

Lahiri, K., Peng, H. and Zhao, Y. (2017). Online learnig and forecast combination in unbalanced panels. *Econometric Reviews,*vol. 36, pp. 257-288.

Jose, V.R.R. and Winkler, R.L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, 24, pp. 163-169.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J.J. (2011). The Principle of Working on Coordinates, in *Compositional data analysis: Theory and Applications*, eds. V. Pawlowsky-Glahn and A. Buccianti, 29-42, Wiley and sons.

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

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Wiley.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. Chichester: Wiley.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On form of spurious correlation which may arise when indices are used in the measurement of rorgans. *Proceedings of the Royal Society of London,* LX, pp. 489-502.

Poncela, P., Rodríguez, J., Sánchez-Mangas, R. and Senra, E. (2011). Forecast combination through dimension reduction techniques. *International Journal of Forecasting,* 27, pp. 224-237.

Smith, J. and Wallis, K.F. (2009). A simple explanation of the forecast combination puzzle. *Oxford bulletin of economics and statistics*, 71(3), pp. 331-355.

Stock, J.H. and Watson, M.W. (2004). Combination forecasts of output growth in a seven country dataset, *Journal of Forecasting,* vol. 23, pp. 405-430.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society, Series B*, 58, pp. 267-288.

SORT 46 (2) July-December 2022, 217-244 DOI: 10.2436/20.8080.02.124

Missing data analysis and imputation via latent Gaussian Markov random fields

Virgilio Gómez-Rubio[1], Michela Cameletti[2] and Marta Blangiardo[3]

Abstract

This paper recasts the problem of missing values in the covariates of a regression model as a latent Gaussian Markov random field (GMRF) model in a fully Bayesian framework. The proposed approach is based on the definition of the covariate imputation sub-model as a latent effect with a GMRF structure. This formulation works for continuous covariates but for categorical covariates a typical multiple imputation approach is employed. Both techniques can be easily combined for the case in which continuous and categorical variables have missing values. The resulting Bayesian hierarchical model naturally fits within the integrated nested Laplace approximation (INLA) framework, which is used for model fitting. Hence, this work fills an important gap in the INLA methodology as it allows to treat models with missing values in the covariates. As in any other fully Bayesian framework, by relying on INLA for model fitting it is possible to formulate a joint model for the data, the imputed covariates and their missingness mechanism. In this way, it is possible to tackle the more general problem of assessing the missingness mechanism by conducting a sensitivity analysis on the different alternatives to model the non-observed covariates. Finally, the proposed approach is illustrated in two examples on modeling health risk factors and disease mapping.

MSC: *62D10, 62F15, 62M30.*

Keywords: *Imputation, missing values, GMRF, INLA, sensitivity analysis.*

[1] Department of Mathematics, School of Industrial Engineering, Albacete, Universidad de Castilla-La Mancha (Spain)

[2] Department of Economics, University of Bergamo, Bergamo, IT.

[3] Department of Epidemiology and Biostatistics, Imperial College London, London, UK.

Received: October 2022
Accepted: November 2022

1. Introduction

Missing data is an important issue a researcher needs to deal with in any statistical analysis; failing to properly account for it can result in a reduction of statistical power, or even in biased statistical inference. Consequently, countless methods have focused on how to deal with missing data (see, for example, ?Enders, 2010; van Buuren, 2012; Trivellore, 2015; Little and Rubin, 2019).

Missing data can occur for a number of reasons, as described in Little and Rubin (2019). Sometimes, the missingness mechanism is ignorable and inference can rely on the observed data alone, appropriately coupled with a suitable imputation or data augmentation model if needed. When the missingness mechanism is not ignorable, a joint approach is required to fit the analysis model, impute the missing values and assess the missingness mechanism. Under this scenario, it is recommended that a sensitivity analysis is carried out to assess the impact of the missingness mechanism on the model parameters estimates (Mason et al., 2012).

The Bayesian paradigm has gained popularity for dealing with missing data, making no distinction between parameters and missing data which are considered as additional unknown parameters. For these reasons, and differently from other ad-hoc methods (Nakagawa, 2015), with a fully Bayesian approach it is possible to combine the analysis and imputation model in a joint estimation framework (Erler et al., 2016). For instance, Mason (2009) and Mason et al. (2012) developed a fully Bayesian missing imputation framework in order to adjust for several missing covariates in longitudinal or cross-sectional studies; each of the missing covariates is assigned an imputation model, all jointly modelled with the analysis model.

The approach we propose in this paper is based on recasting the imputation model to define it as a latent Gaussian Markov random field (GMRF, Rue and Held, 2005) which is part of a larger Bayesian hierarchical model. This fits naturally within the integrated nested Laplace approximation (INLA, Rue, Martino and Chopin, 2009) methodology, as an alternative to Markov chain Monte Carlo (MCMC, see, for example, Brooks et al., 2011). This approach is suitable for continuous covariates and can be also extended to impute categorical variables. This makes model fitting with missing covariates possible in INLA, and our new approach fills an important gap, as INLA has always required that the data in the latent GMRF defining the model to be fully observed. Here we focus on the case of missing values in the covariates as INLA can easily fit models with missing values in the response variable, simply computing the corresponding posterior predictive distribution derived from the analysis model to be fit (see, for example, Gómez-Rubio, 2020).

A previous attempt to solve the issue of missing values in the covariates in the INLA framework can be found in Gómez-Rubio and Rue (2018). They adopt a Gaussian prior for the imputation of the missing values in the covariates and sample from the missing data posterior distribution through INLA within MCMC. A different approach is proposed in Chapter 8 of Blangiardo and Cameletti (2015), where a bivariate model for spatially misaligned data is estimated by adopting the stochastic partial differential equations (SPDE) approach Lindgren, Rue and Lindstrom (2011). Covariate values are imputed (in new locations) by assuming a spatial Gaussian field which is also included in the linear predictor of the response model. See also Barber et al. (2016); Forlani et al. (2020) for model examples on the use of spatial models for misalignment. Alternatively, Gómez-Rubio (2020) proposes a multiple imputation (MI) approach (Rubin, 1987, 1996; Carpenter and Kenward, 2012): the covariates are imputed multiple times through resampling, so that *N* complete datasets are used in the analysis model. All the results are then combined to obtain the final estimates of the model parameters (see Rubin, 1987, for details). The approach introduced here differs from previous approaches in that a joint framework is proposed, similarly to Mason et al. (2012). Through the joint model, the uncertainty about the imputation of the missing covariates propagates throughout the model so that it also reflects on the model parameters estimates in the analysis. At the same time, information from the outcome in the analysis model feedbacks on the imputation, making it unnecessary to include the outcome in the imputation model, as commonly done in the classic MI approach. This new approach fits naturally within the INLA framework, can be extended to consider different types of problems (i.e., not only spatial models) and can be easily fit with the associated `R-INLA` package for the `R` programming language (Gómez-Rubio, 2020).

The paper is structured as follows. In Section 2 we review methods for missing values, while in Section 3 we introduce our novel method for missing values imputation. Section 4 presents a brief summary of the INLA approach to Bayesian inference and how our novel approach fits within this framework. Section 5 shows two examples for the application of our proposed method and Section 6 presents discussion points.

2. Approaches to deal with missing data

In their seminal book, Little and Rubin (2019) identify three possible mechanisms of missingness. If the probability of being missing is the same for all the observations, we can assume that the missing data distribution does not depend on any of the observed or missing variables. In this case the data are said to be *missing completely at random* (MCAR). If the distribution of the missing data depends on completely observed variables (i.e., observed for all the subjects) and it does not depend on the varibles with missing values, the data are called *missing at random* (MAR). An example of MAR is that women are less likely to answer questions related to their income than men, but this has nothing to do with the income itself. Finally, if neither MCAR or MAR holds, the *missing not at random* (MNAR) case occurs and the missing values distribution depends on both missing and observed variables. For instance, in a neurological questionnaire, a subject is less likely to answer questions related to the disease if this is severe.

Under MCAR or MAR, the missing data mechanism is *ignorable*. As reported in Seaman and White (2013), this means that inferences obtained from a parametric model for the observed data alone are the same as inferences obtained from a joint model for the data and missingness mechanism. On the contrary, if the data are MNAR the missing data mechanism is not ignorable and a model for the missingness mechanism is required. It is important to note that we cannot gather evidence from the data at hand about the missing data mechanism (MCAR, MNAR or MAR). On the basis of the knowledge regarding the data collection methods and the assumed relationship among the collected variables, it is possible only to make assumptions about the reasons for missing data, choose the best corresponding strategy for data analysis (Pigott, 2001) and conduct a sensitivity analysis on these assumptions (Mason et al., 2012).

The simplest and most popular ad-hoc method to deal with missing information consists in replacing the missing data with a plausible value, such as the mean or median calculated over the observed cases (or the mode if the variable is categorical) or to perform a complete cases analysis (i.e., removing the observations with one or more missing values). However, while the first method has the potential of distorting the data distribution and of underestimating their variability, the second one has the major drawback of reducing the power of the study (as the dataset for the analysis will have a reduced size) and of producing biased estimates if the MCAR assumption is not valid. To overcome this issue, inverse probability weighting was developed, based on the idea of assigning different weights to the different complete cases based on specific characteristics which are relevant for the missing data; in two reviews Carpenter, Kenward and Vansteelandt (2006) and Seaman and White (2013) showed advantages and drawbacks of the approach.

In the last three decades model-based methods have been preferred to account for missing data in the case of an ignorable missing data mechanism; see, for instance, the papers by Little 1992, Little and Rubin 2019 and Schafer and Graham 2002. Regression mean imputation is the simplest of the model-based methods, where the variable with missing data is predicted based on a regression model which includes the other variables as regressors. To overcome the issue of unreasonable lack of uncertainty for the imputed values, stochastic regression imputation was proposed to generate imputed values adding some random noise (Nakagawa, 2015).

A well established and increasingly popular model-based approach to dealing with missing data occurring in more than one variable is MI proposed by Rubin (1987, 1996). Through Monte Carlo simulation, it produces several versions of the complete dataset which only differ in the imputed missing values. Then, for each complete dataset the estimates of interest are computed by fitting the analysis model (also called *substantive model*) and the results are pooled together into a final estimate which takes into account the uncertainty of the imputed data. The imputation of the missing values can be done using mainly two strategies (van Buuren, 2012): i) *joint modeling*, when missing values are imputed by sampling from a multivariate model fitted to the data, for which usually a multivariate Gaussian is used (**?**Mason et al., 2012); ii) *fully conditional specification* (also known as multiple imputation using chained equations, MICE (van Buuren and Groothuis-Oudshoorn, 2011)), when conditional univariate distributions are used to impute the missing values iteratively through a variable-by-variable approach (see White, Royston and Wood 2011 for a thorough review of this method).

2.1. Bayesian inference

Bayesian inference provides a suitable framework for dealing with missing data, as it treats missing data similarly to model parameters, making no distinction between them. For these reasons and differently from other methods, with a fully Bayesian approach, it is possible to include the analysis model, imputation model and missingness mechanism model in a joint estimation framework (Erler et al., 2016).

Let $\mathcal{D}$ denote the *complete* set of data, which will include the response variable and the covariates. It is assumed that $\mathcal{D} = (\mathcal{D}_{obs}, \mathcal{D}_{mis})$, where $\mathcal{D}_{obs}$ denotes the observed values while $\mathcal{D}_{mis}$ refers to the missing values. Moreover, let $\boldsymbol{M}$ be the missing data indicator variable, i.e., a vector or matrix with the same length or dimension as $\mathcal{D}$ with values equal to 1 if the corresponding values of $\mathcal{D}$ is missing (and 0 otherwise).

Following the selection model approach (Nakagawa, 2015), the joint distribution of $\mathcal{D}$, $\boldsymbol{M}$, the model parameters $\boldsymbol{\theta}_{\mathcal{D}}$ and the parameters in the missingness model $\boldsymbol{\theta}_M$ can be expressed as

$$\begin{aligned}\pi(\mathcal{D}, \boldsymbol{M}, \boldsymbol{\theta}_{\mathcal{D}}, \boldsymbol{\theta}_M) &= \pi(\mathcal{D}, \boldsymbol{\theta}_{\mathcal{D}})\pi(\boldsymbol{M} \mid \mathcal{D}, \boldsymbol{\theta}_M)\pi(\boldsymbol{\theta}_M) = \\ &= \pi(\boldsymbol{M} \mid \mathcal{D}, \boldsymbol{\theta}_M)\pi(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{D}})\pi(\boldsymbol{\theta}_{\mathcal{D}})\pi(\boldsymbol{\theta}_M).\end{aligned}$$

This formulation assumes that parameters $\boldsymbol{\theta}_{\mathcal{D}}$ and $\boldsymbol{\theta}_M$ are distinct and with independent priors and that the distribution of $\mathcal{D}$ (given $\boldsymbol{\theta}_{\mathcal{D}}$) does not depend on the parameters of the missingness model θ_M. Note that term $\pi(\boldsymbol{M} \mid \mathcal{D}, \boldsymbol{\theta}_M)$ represents the missingness model and $\pi(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{D}})$ the likelihood of the data.

Following this, $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M)$ depends on a set of parameters $\boldsymbol{\theta}_M$, and models the missing data mechanism for the three cases introduced above (Little and Rubin, 2019):

MCAR, if the distribution does not depend on any of the fully or partially observed variables, i.e., $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M) = \pi(\boldsymbol{M} \mid \boldsymbol{\theta}_M)$.

MAR, if the distribution depends only on fully observed variables, which means that $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M) = \pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \boldsymbol{\theta}_M)$. This implies that, given the observed data, the missingness mechanism does not depend on the unobserved data.

MNAR, if the distribution $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M)$ depends on fully and partially observed variables.

If the data are MCAR or MAR and the parameters $\boldsymbol{\theta}_M$ are distinct of the parameters of the data generating process, $\boldsymbol{\theta}_{\mathcal{D}}$, and with independent priors, then the missing data mechanism is *ignorable* and $\pi(\boldsymbol{M} \mid \mathcal{D}_{obs}, \mathcal{D}_{mis}, \boldsymbol{\theta}_M)$ can be omitted (Seaman and White, 2013). On the contrary if the data are MNAR, the missing data mechanism is not ignorable and a model for missingness is required (i.e., a logistic model) and has to be jointly estimated with the main model, that will include an imputation model for the missing values.

Note that it is not possible to tell from the data at hand whether the missing observations are MCAR, MNAR or MAR and at the same time it is not trivial to specify a model of missingness. In this case, a sensitivity analysis needs to be carried out to assess the impact of different scenarios for the missing data on the estimates of the model parameters (Carpenter, Kenward and White, 2007; Mason et al., 2012).

2.2. Missing data in the response variable

Let now $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{x})$ be the set of data including the response $\boldsymbol{y}$ and the covariates $\boldsymbol{x}$. If it is assumed that the covariates are fully observed and that the response variable $\boldsymbol{y}$ contains missing values, i.e., the response variable $\boldsymbol{y}$ can be split into observed values, $\boldsymbol{y}_{obs}$, and unobserved values, $\boldsymbol{y}_{mis}$. Hence, $\mathcal{D}_{obs} = (\boldsymbol{y}_{obs}, \boldsymbol{x})$ and $\mathcal{D}_{mis} = (\boldsymbol{y}_{mis})$. In this case likelihood $\pi(\mathcal{D}_{obs}, \mathcal{D}_{mis} \mid \boldsymbol{\theta}_{\mathcal{D}})$ corresponds to the distribution of $\pi(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis} \mid \boldsymbol{x}, \boldsymbol{\theta}_y)$, with $\boldsymbol{\theta}_y$ the hyperparameters in the likelihood.

If we assume that the missing data mechanism is ignorable, the imputation of the missing data values $\boldsymbol{y}_{mis}$ is simply done through the posterior predictive distribution $\pi(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{x})$. In general, we will have the observation model by defining an appropriate distribution for the likelihood. In addition, the mean of observation i, ϕ_i ,will be linked to a linear predictor η_i on the covariates and other effects using an appropriate link function $g(\cdot)$, i.e.,

$$g(\phi_i) = \eta_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pi} + \sum_{l=1}^{L} f_l(u_{li}). \tag{1}$$

Here, β_0 is an intercept, $\{\beta_p\}_{p=1}^{P}$ the coefficients of the P covariates available $\{\boldsymbol{x}_p\}_{p=1}^{P}$ and $\{f_l(\cdot)\}_{l=1}^{L}$ represents L different non-linear effects on covariates $\{\boldsymbol{u}_l\}_{l=1}^{L}$ (which are also part of the observed data $\mathcal{D}_{obs}$ now).

If the data are MNAR, a missing mechanism model $\pi(\boldsymbol{M} \mid \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}_M)$ is required in addition to the previous model, e.g.,

$$\begin{aligned} M_i \mid p_i &\sim Bernoulli(p_i) \\ logit(p_i) &= \gamma_0 + \sum_{r=1}^{R} \gamma_r x_{ri} + \delta y_i \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_M = [\gamma_1, \gamma_1 \cdots \gamma_R \ \ \delta]^{\top}$ and M_i is a missingness indicator for y_i. In addition, an imputation model for the missing values will be required. Furthermore, δ is a coefficient that measures the effect of the response variable on the missingness mechanism.

However, in this work we will assume that there are no missing observations in the response or that the missingness mechanism is ignorable, which means that posterior inference is based on the predictive distribution.

2.3. Missing data in the covariates

We now consider the case when $\mathcal{D}_{obs} = (\boldsymbol{y}, \boldsymbol{x}_{obs})$ and $\mathcal{D}_{mis} = (\boldsymbol{x}_{mis})$, with $\boldsymbol{x}_{obs}$ the observed values of the covariates and $\boldsymbol{x}_{mis}$ the missing ones. Henceforth, the likelihood

$\pi(\mathcal{D}_{obs}, \mathcal{D}_{mis} \mid \boldsymbol{\theta}_{\mathcal{D}})$ can be written as

$$\pi(\boldsymbol{y}, \boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} \mid \boldsymbol{\theta}_{\mathcal{D}}) = \pi(\boldsymbol{y} \mid \boldsymbol{x}_{obs}, \boldsymbol{x}_{mis}, \boldsymbol{\theta}_y)\pi(\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} \mid \boldsymbol{\theta}_x)$$

assuming that $\boldsymbol{\theta}_{\mathcal{D}} = [\boldsymbol{\theta}_y^\top \; \boldsymbol{\theta}_x^\top]^\top$ is the vector of conditionally independent parameters. The distribution $\pi(\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis} \mid \boldsymbol{\theta}_x)$ represents the joint distribution of observed and missing covariates and it includes the imputation model. For example, the joint distribution can be a multivariate normal distribution (taking into consideration correlation between covariates) for continuous covariates, or a discrete distribution if the covariate is categorical.

In general, we will have the observation model with a linear predictor as in Equation (1) together with the imputation model and the missingness model (described in Section 3) as in Equation (2) but only if the missing data are MNAR.

3. Imputation of continuous missing covariates

Differently from Section 2, let $\boldsymbol{z} = [\boldsymbol{z}_{obs}{}^\top \; \boldsymbol{z}_{mis}{}^\top]^\top$ denote now the complete set of values of a covariate, which will typically be a column vector. The response values $\boldsymbol{y}$ will be written separately where needed. This is done for simplicity, so that the imputation of a single covariate with missing observations will be considered now. However, this approach can be easily extended to consider the imputation of missing values in several continuous covariates using a multivariate model.

Let $\boldsymbol{z}^*$ be a latent effect that is split in two parts, i.e., $\boldsymbol{z}^* = [\boldsymbol{z}_{obs}^*{}^\top \; \boldsymbol{z}_{mis}^*{}^\top]^\top$. The main idea is to define latent effect $\boldsymbol{z}^*$ as a GMRF with mean $\boldsymbol{\mu}^*(\boldsymbol{\theta}_I)$ and precision $\boldsymbol{Q}^*(\boldsymbol{\theta}_I)$ so that $\boldsymbol{z}_{obs}^*$ is as close as possible to the actual values $\boldsymbol{z}_{obs}$ and so that $\boldsymbol{z}_{mis}^*$ is obtained using a particular imputation model for $\boldsymbol{z}_{mis}$ that depends on observed covariates $\boldsymbol{z}_{obs}$ and some parameters $\boldsymbol{\theta}_I$.

To guarantee that the distribution of $\boldsymbol{z}_{obs}^*$ is taken to be as close as possible to the observed covariate data $\boldsymbol{z}_{obs}$, the mean of $\boldsymbol{z}_{obs}^*$ is set equal to $\boldsymbol{z}_{obs}$ and its associated sub-block in $\boldsymbol{Q}^*(\boldsymbol{\theta}_I)$ equal to a diagonal matrix with high values (e.g., 10^{10}) in the diagonal. In this way, the values of $\boldsymbol{z}_{obs}^*$ are centered at observed values $\boldsymbol{z}_{obs}$ and have a negligible variation about these observed values. Regarding the distribution of $\boldsymbol{z}_{mis}^*$ (with mean $\boldsymbol{\mu}_c$ and precision $\boldsymbol{Q}_c$), it will be based on an imputation model on observed covariates $\boldsymbol{z}_{obs}$ and parameters $\boldsymbol{\theta}_I$. Finally, we will also assume that $\boldsymbol{z}_{obs}^*$ and $\boldsymbol{z}_{mis}^*$ are independent because the marginal distribution of $\boldsymbol{z}_{mis}^*$ will include all dependence of the missing values on the observed data $\boldsymbol{z}_{obs}$.

Consequently, the joint distribution of $\boldsymbol{z}^*$ is given by

$$\boldsymbol{z}^* \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{z}_{obs} \\ \boldsymbol{\mu}_c \end{bmatrix}, \begin{bmatrix} 10^{10}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & Q_c \end{bmatrix}\right), \tag{3}$$

where $\boldsymbol{I}$ represents the identity matrix. The distribution of $\boldsymbol{z}^*$ will be used later when defining the imputation model for the missing values as a latent effect for `R-INLA` in Section 4.

3.1. Imputation latent effect

To derive the distribution for the imputation model, $\pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I)$, a multivariate Normal distribution is assumed for the joint distribution of the complete set of covariates $\boldsymbol{z}$:

$$\boldsymbol{z} \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{\mu}_{obs} \\ \boldsymbol{\mu}_{mis} \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q}_{obs,obs} & \boldsymbol{Q}_{obs,mis} \\ \boldsymbol{Q}_{mis,obs} & \boldsymbol{Q}_{mis,mis} \end{bmatrix}\right) = \text{Normal}\left(\boldsymbol{\mu}, \boldsymbol{Q}\right), \qquad (4)$$

where both the mean and the precision matrix can depend on $\boldsymbol{\theta}_I$. It follows that the *imputation model* is defined by the following conditional distribution (Rue and Held, 2005):

$$\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I \sim \text{Normal}\left(\boldsymbol{\mu}_c, \boldsymbol{Q}_c\right)$$

where $\boldsymbol{\mu}_c = \boldsymbol{\mu}_{mis} - \boldsymbol{Q}^{-1}_{mis,mis}\boldsymbol{Q}_{mis,obs}\left(\boldsymbol{z}_{obs} - \boldsymbol{\mu}_{obs}\right)$ and $\boldsymbol{Q}_c = \boldsymbol{Q}_{mis,mis}$. Note that $\boldsymbol{\mu}_c$ and $\boldsymbol{Q}_c$ are necessary to define the distribution of the new latent effect given in Equation (3).

As stated above, the distribution of $\boldsymbol{z}^*_{mis}$ will play the role of the imputation model of the missing values. This imputation model will, in practice, be a sub-model in a larger model that will be defined using the conditional distribution of the missing values $\boldsymbol{z}_{mis}$ given the observed data $\boldsymbol{z}_{obs}$ and hyperparameters $\boldsymbol{\theta}_I$. Note that in this sub-model $\boldsymbol{z}_{obs}$ can be regarded as the data while $\boldsymbol{z}_{mis}$ and $\boldsymbol{\theta}_I$ are the parameters to estimate. Because this sub-model will be included as part of a fully Bayesian larger model, posterior inference on $\boldsymbol{z}_{mis}$ and $\boldsymbol{\theta}_I$ will be based on all observed data in the model (i.e., response variable and observed covariates) so that there is feedback from other parts of the model to make inference on $\boldsymbol{z}_{mis}$ and $\boldsymbol{\theta}_I$.

Considering only the data and parameters in the sub-model, the way in which the imputation sub-model is defined relies on the distribution of $\boldsymbol{z}_{mis}$ given $\boldsymbol{z}_{obs}$. This can be written as

$$\pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}) = \int_{\Theta_I} \pi(\boldsymbol{z}_{mis}, \boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs}) d\boldsymbol{\theta}_I = \int_{\Theta_I} \pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I)\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs}) d\boldsymbol{\theta}_I.$$

where Θ_I is the parametric space of $\boldsymbol{\theta}_I$.

Here, $\pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I)$ is the conditional distribution of the missing values given the observed data and the hyperparameters of the imputation model introduced above. Also, note that $\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs})$ can be regarded as the distribution of the hyperparameters in the imputation sub-model given the observed data. Note that this distribution is estimated only from the observed data $\boldsymbol{z}_{obs}$, so it can be regarded as an *informative prior* for $\boldsymbol{\theta}_I$. Moreover, it can be rewritten as

$$\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs}) \propto \pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\theta}_I)\pi(\boldsymbol{\theta}_I)$$

where $\pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\theta}_I)$ is obtained by integrating $\boldsymbol{z}_{mis}$ out in the distribution of $\boldsymbol{z}$, that is, $\pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\theta}_I) = \int \pi(\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis} \mid \boldsymbol{\theta}_I) d\boldsymbol{z}_{mis}$. Finally, the hyperparameters $\boldsymbol{\theta}_I$ are typically modelled as exchangeable a priori.

Next, two particular examples of imputation with a typical linear regression and a spatial model (useful when the covariate is spatially correlated) are described. It is worth noting that the principles presented below can be extended to a wide range of models, including longitudinal data, time series and other smooth terms.

3.2. Imputation with a linear regression model

The first imputation model that we describe is based on the linear regression model. We assume that the mean of the multivariate Normal distribution in Equation (4) is defined, considering the n observations, as $\boldsymbol{X\beta}$. Here, $\boldsymbol{X}$ is a matrix of P fully observed covariates (columnwise) with associated coefficient vector $\boldsymbol{\beta} = [\beta_0 \cdots \beta_P]^\top$. To match the structure of $\boldsymbol{z} = [\boldsymbol{z_{obs}}^\top \, \boldsymbol{z_{mis}}^\top]^\top$, matrix $\boldsymbol{X}$ can be rewritten as a block matrix as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{obs} \\ \boldsymbol{X}_{mis} \end{bmatrix}$$

Under the linear regression model, we assume that the mean of $\boldsymbol{z}$ depends on a linear combination of the fully observed covariates, i.e., $\boldsymbol{\mu} = E(\boldsymbol{z}) = \boldsymbol{X\beta}$. By adopting the block notation, we thus assume that the joint distribution of Equation (4) is given by

$$\boldsymbol{z} \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{X}_{obs}\boldsymbol{\beta} \\ \boldsymbol{X}_{mis}\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \tau \boldsymbol{I}_{obs} & \boldsymbol{0} \\ \boldsymbol{0} & \tau \boldsymbol{I}_{mis} \end{bmatrix} \right),$$

where τ is the precision hyperparameter and $\boldsymbol{I}_{obs}$ and $\boldsymbol{I}_{mis}$ are identity matrices whose dimensions depend on the number of missing and observed data in $\boldsymbol{z}$. In this case the vector of hyperparameters is given by $\boldsymbol{\theta}_I = [\boldsymbol{\beta}^\top \, \tau]^\top$. Note that, given $\boldsymbol{\theta}_I$, observations are assumed independent of each other, which simplifies the model.

Following the approach presented in Section 3.1, we obtain that the conditional distribution of $\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I$ (i.e., the imputation model) has the following mean and precision:

$$\boldsymbol{\mu}_c = \boldsymbol{X}_{mis}\boldsymbol{\beta}, \qquad \boldsymbol{Q}_c = \tau \boldsymbol{I}_{mis},$$

As stated above, note that $\boldsymbol{\beta}$ and τ are informed by $\pi(\boldsymbol{\beta}, \tau \mid \boldsymbol{z}_{obs})$, which is proportional to $\pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}, \tau)$. Note that $\pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\beta}, \tau)$ can be easily derived from the multivariate normal distribution of $\boldsymbol{z}$ above and that it will also be a multivariate normal distribution with mean $X_{obs}\boldsymbol{\beta}$ and precision $\tau \boldsymbol{I}_{obs}$.

Finally, priors must be set on the hyperparameters. For simplicity, each of the elements in $\boldsymbol{\beta}$ is assigned a Normal distribution with zero mean and small precision. Parameter τ has a vague prior (e.g., a Gamma distribution with small precision). All hyperparameters are independent a priori, so that $\pi(\boldsymbol{\theta}_I) = \pi(\tau)\Pi_{i=0}^{P}\pi(\beta_i)$. Note that other priors could be easily considered here.

3.3. Imputation with a spatial model

When the covariate to be imputed is spatially correlated we can assume a conditional autoregressive (CAR) specification (Held and Rue, 2010) so that the mean is $\boldsymbol{\mu} = \boldsymbol{\alpha} = [\alpha \cdots \alpha]^\top$ and the precision is $\boldsymbol{Q} = \tau(\boldsymbol{I} - \rho \boldsymbol{W})$. Here, α is the intercept of the linear predictor, ρ is a spatial autocorrelation parameter, and $\boldsymbol{W}$ is an adjacency matrix, defining the sets of neighbours. This is often scaled dividing it by its largest eigenvalue as this will allow us to take ρ in the $(0,1)$ interval. Note that $\boldsymbol{W}$ can be rewritten as a block matrix with four sub-matrices according to missing and observed values, as done with $\boldsymbol{Q}$ in Equation (4). The vector of hyperparameters is now given by $\boldsymbol{\theta}_I = [\tau \;\; \rho \;\; \alpha^{\top}]^{\top}$.

Adopting block notation, under the CAR specification for imputation the following joint distribution is assumed for $\boldsymbol{z} = [\boldsymbol{z}_{obs}{}^{\top} \; \boldsymbol{z}_{mis}{}^{\top}]^{\top}$:

$$\boldsymbol{z} \mid \boldsymbol{\theta}_I \sim \text{Normal}\left(\begin{bmatrix} \boldsymbol{\alpha}_{obs} \\ \boldsymbol{\alpha}_{mis} \end{bmatrix}, \begin{bmatrix} \tau(\boldsymbol{I}_{obs} - \rho \boldsymbol{W}_{obs,obs}) & -\tau\rho \boldsymbol{W}_{obs,mis} \\ -\tau\rho \boldsymbol{W}_{mis,obs} & \tau(\boldsymbol{I}_{mis} - \rho \boldsymbol{W}_{mis,mis}) \end{bmatrix} \right).$$

It then follows that the conditional distribution of $\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I$ (i.e., the imputation model) is characterised by the following mean and precision matrix:

$$\boldsymbol{\mu}_c = \boldsymbol{\alpha}_{mis} - (\boldsymbol{I}_{mis} - \rho \boldsymbol{W}_{mis,mis})^{-1}(-\rho \boldsymbol{W}_{mis,obs})(\boldsymbol{z}_{obs} - \boldsymbol{\alpha}_{obs})$$

$$\boldsymbol{Q}_c = \tau\,(\boldsymbol{I}_{mis} - \rho \boldsymbol{W}_{mis,mis})$$

Again, τ, ρ and α are informed by $\pi(\tau, \rho, \alpha \mid z_{obs})$, which is proportional to the product $\pi(z_{obs} \mid \tau, \rho, \alpha)\pi(\tau, \rho, \alpha)$. As in the previous case, $\pi(z_{obs} \mid \tau, \rho, \alpha)$ can be easily derived from the multivariate normal distribution of $\boldsymbol{z}$ above and that it will also be a multivariate normal distribution with mean α_{obs} and precision $\tau(\boldsymbol{I}_{obs} - \rho \boldsymbol{W}_{obs,obs})$.

Finally, α is given a Gaussian prior with zero mean and small precision, τ is assigned a vague prior (e.g., a Gamma distribution with a small precision), while logit(ρ) is assigned a Gaussian prior with zero mean and small precision (see, for example, Gómez-Rubio, 2020, Chapter 5, for details on why this parameterisation is used).

3.4. Extension to the imputation of categorical missing covariates

The imputation of the missing values in categorical variables does not fit into the GMRF framework described in Section 3 as these variables are defined in a discrete space. For this reason, a different approach will be considered for defining the imputation model $\pi(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\theta}_I)$ and for estimating the model. In particular, as imputation model we will consider a multinomial likelihood which can be fit with INLA by using the multinomial-Poisson transformation (Baker, 1994).

Note that in this case the procedure is similar to the multiple imputation approach: the imputation model is specified where the categorical variables with missing values are considered as the response variables, so that the predictive distribution of the missing observations can be computed. Similarly to the case of missing data in the response, values are sampled to fill the missing values in the covariates. Then, the analysis model is run by using the imputed covariates as completely known. This procedure is repeated by simulating several samples and estimating the corresponding models; finally, all the resulting models are pooled by using Bayesian model averaging (Gómez-Rubio and Rue, 2018). Note that this approach does not produce feedback in the estimation of the parameters of the imputation model as in the joint approach, given that it is done in two-stages rather than jointly. For this reason, and similarly to the classical MI, the outcome $\boldsymbol{y}$ should be included in the imputation model. Alternatively, INLA within MCMC can be used to fit the joint model using a fully Bayesian approach (see the example in Gómez-Rubio and Rue, 2018).

Inference on the model parameters when multiple imputation of a categorical covariate can be summarised as follows. Considering the generic parameter θ_k we can write its posterior marginal distribution as:

$$\pi(\theta_k \mid z_{obs}, \boldsymbol{y}) = \sum_{z_{mis} \in \Theta_{mis}} \pi(\theta_k, z_{mis} \mid z_{obs}, \boldsymbol{y}) = \sum_{z_{mis} \in \Theta_{mis}} \pi(\theta_k \mid z_{obs}, z_{mis}, \boldsymbol{y}) \pi(z_{mis} \mid z_{obs}, \boldsymbol{y}).$$

Here, Θ_{mis} represents the parametric space of the missing values of the categorical covariate, which in a Bayesian framework are considered to be random variables.

Given L samples $\{\boldsymbol{z}_{mis}^{(l)}\}_{l=1}^{L}$ from $\pi(\boldsymbol{z}_{mis} \mid z_{obs}, \boldsymbol{y})$, the previous marginal can be approximated as

$$\pi(\theta_k \mid z_{obs}, \boldsymbol{y}) \simeq \frac{1}{L} \sum_{l=1}^{L} \pi(\theta_k \mid z_{obs}, \boldsymbol{z}_{mis}^{(l)}, \boldsymbol{y}),$$

where $\pi(\theta_k \mid z_{obs}, \boldsymbol{z}_{mis}^{(l)}, \boldsymbol{y})$ is the marginal of θ_k obtained from fitting the original model with the observed data and the imputed covariate $\boldsymbol{z}_{mis}^{(l)}$.

Note that when continuous covariates with missing values are also present both approaches can be combined. For example, an imputation model can be combined for the continuous covariate which is part of the joint model that is fit to every simulated dataset where only the missing values of the categorical covariate are filled in. Furthermore, a missingness model for the categorical variables can be incorporated into the model similarly to the one used for the continuous variables.

4. The Integrated Nested Laplace Approximation approach (INLA)

The approach presented in the previous sections can be implemented using a number of methods for Bayesian inference. However, it overcomes a major limitation in the INLA method as, at present, it cannot cope with missing values in covariates. An introduction to the INLA method and the computational details is presented here; then we focus on how to implement our proposed framework.

INLA (Rue et al., 2017; Martino and Riebler, 2019; Gómez-Rubio, 2020) is a deterministic approach for Bayesian inference. It is designed for the class of latent Gaussian Markov random field models, where the distribution of the response y_i (observed for the i-th unit) is assumed to belong to a distribution family (usually part of the exponential family). This is often characterized by a parameter ϕ_i (i.e., the mean of y_i) defined as a function of a structured additive predictor η_i through a link function such that $g(\phi_i) = \eta_i$ (e.g. the logarithm function is used for Poisson data). The linear predictor is defined as in equation (1).

Regarding the tems in the linear predictor, recall that β_0 is the intercept, coefficients $\boldsymbol{\beta} = [\beta_1 \cdots \beta_P]^\top$ quantify the (linear) effect of some covariates $\boldsymbol{x} = \{\boldsymbol{x}_p\}_{p=1}^{P}$ on the response, and $\boldsymbol{f} = \{f^{(1)}(\cdot), \ldots, f^{(L)}(\cdot)$ is a set of functions defined in terms of some covariates $\boldsymbol{u} = \{\boldsymbol{u}_l\}_{l=1}^{L}$.

Through functions $f(\cdot)$ it is possible to include in the model random effects (perhaps indexed in space and time), smooth and non-linear effects of the covariates. For this reason, the class of latent GMRF models can accommodate a wide range of models, from standard generalized linear models (GLM) to generalized linear mixed models (GLMM), including data for time series, lattice data, point pattern and geostatistical data.

As stated, the set of latent effects $\boldsymbol{\chi} = \{\boldsymbol{\eta}, \beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$ is a latent GMRF in the model, which depends on some hyperparameters $\boldsymbol{\theta}_2$. Moreover, observations are assumed to be independent given the latent effects $\boldsymbol{\chi}$ and the likelihood hyperparameters denoted by $\boldsymbol{\theta}_1$. For convenience, in the following the vector of hyperparameters wil be denoted as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^{\intercal}\ \boldsymbol{\theta}_2^{\intercal}]^{\intercal}$.

The outputs of Bayesian inference with INLA are the marginal posterior distributions for each element of the latent effects and hyperparameters vector denoted by $p(\boldsymbol{\chi}_{\bullet} \mid \boldsymbol{y})$ and $p(\boldsymbol{\theta}_{\bullet} \mid \boldsymbol{y})$, respectively. INLA provides deterministically accurate approximations to these distributions in a short computing time by using the Laplace approximation and numerical integration.

Each latent GMRF model can be rewritten hierarchically with three levels:

1. The model for the observed data $\boldsymbol{y} = [y_1 \cdots y_n]^{\intercal}$ (i.e., the likelihood) defined as a function of some parameters $\boldsymbol{\chi}$ and hyperparameters $\boldsymbol{\theta}$:

$$\boldsymbol{y} \mid \boldsymbol{\chi}, \boldsymbol{\theta} \sim \pi(\boldsymbol{y} \mid \boldsymbol{\chi}, \boldsymbol{\theta}) = \prod_{i \in \{1, \ldots, n\}} \pi(y_i \mid \chi_i, \boldsymbol{\theta}).$$

2. The model for the latent effects $\boldsymbol{\chi}$:

$$\boldsymbol{\chi} \mid \boldsymbol{\theta} \sim \text{Normal}\left(\mathbf{0}, \boldsymbol{Q}(\boldsymbol{\theta})\right)$$

   where $\boldsymbol{Q}(\boldsymbol{\theta})$ is a sparse precision matrix given the GMRF assumption.

3. The model for the complete vector of hyperparameters: $\pi(\boldsymbol{\theta})$. As usually hyperparameters are assumed to be independent a priori, $\pi(\boldsymbol{\theta})$ will be defined as the product of different univariate prior distributions.

Given all these models and components the joint posterior distribution of the random effects and the hyperparameters is given by

$$\pi(\boldsymbol{\chi}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{y} \mid \boldsymbol{\chi}, \boldsymbol{\theta}) \pi(\boldsymbol{\chi} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}).$$

As stated above, INLA computes the posterior marginals of the hyperparameters and latent effects using that representation by means of numerical integration and the Laplace approximation (see Rue et al., 2009, for details).

4.1. Computational details

The INLA approach is implemented through an `R` package named `R-INLA`, which is available from the INLA website (http://www.r-inla.org/home). The model to be fit is defined by setting a formula with all the additive latent effects in the model, which includes fixed and random effects. The `R-INLA` package includes a good number of implemented latent effects but others can be implemented as well (see, for example Gómez-Rubio, 2020). Note that by default, when `R-INLA` finds missing values in the covariates (which have the value `NA` in `R`) they are replaced by zeros so that the effect of the covariate does not affect the linear prediction of that subject. However, this is an issue that could result in biased estimates of the coefficients of the covariates. This is described in the `R-INLA` list of frequently asked questions (FAQ) in the package website. If the missing value is found in the response variable, the predictive distribution is computed.

Generic latent effects can be implemented by defining their structure as a latent GMRF. This means defining the mean, precision, hyperparameters and the priors of the hyperparameters. These are known as `rgeneric` latent effects in `R-INLA` (see, for example Gómez-Rubio, 2020, Chapter 11). Once a new latent effect is defined, it can be easily incorporated as any other additive effect in the model formula.

For the new latent effects described in this paper and defined in Equation (3) we have to specify the mean $\boldsymbol{\mu}_c$ and precision $\boldsymbol{Q}_c$ of the block of the missing values. Remember that the block of the observed covariates is simply there to make those values of the latent effect to be as close as possible to the observed values and that it does not depend on any hyperparameter or other data. Furthermore, the role of the prior on the hyperparameters of the imputation model $\boldsymbol{\theta}_I$ is now taken by distribution $\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs})$. Hence, the actual prior used in the latent effects is taken as

$$\pi(\boldsymbol{\theta}_I \mid \boldsymbol{z}_{obs}) \propto \pi(\boldsymbol{z}_{obs} \mid \boldsymbol{\theta}_I)\pi(\boldsymbol{\theta}_I)$$

and the normalizing constant is ignored as it is not needed. In a standard implementation of a latent effect, the prior of $\boldsymbol{\theta}_I$ would be a typical distribution density that depends on a set of fixed hyperparameters, but now the prior of $\boldsymbol{\theta}_I$ is made of the product of the two terms above. For this reason, it can be regarded as an informative prior as it is essentially estimated from a model fit to $\boldsymbol{z}_{obs}$. This is what will allow the latent effect to produce good estimates of the missing values (if the imputation model is correct). In general, there is no way to assess this, but the more covariates used in the imputation model the better (see Gelman and Hill, 2007, Chapter 25). The actual prior of the model hyperparameters is $\pi(\boldsymbol{\theta}_I)$ and this can take different forms depending on the number and type of hyperparameters in the model. Usually, this will be split into the product of several univariate prior distributions.

Note also that `R-INLA` works with unbounded hyperparameters, so that the parameters in $\boldsymbol{\theta}_I$ may need to be transformed when the latent effect is defined. This may also require to include additional terms in the prior (see, for example Gómez-Rubio, 2020, Chapter 11). A typical example is to use internally the log-precision instead of the precision.

Once the imputation latent effect is included in the model formula, it will be part of the joint latent effect χ and incorporated into the Bayesian model, so that a full Bayesian approach is used to estimate all the model parameters.

As stated in previous sections, a missingness model can be included (in addition to an imputation one) for the case in which missingness is MAR or MNAR. Including a missingness model requires defining a model with two likelihoods: one for the main model and a binomial model for the missingness indicator variables. Note that under MCAR and MAR both models are independent, hence the latter is not needed; however, under MNAR it is necessary to explicitly include it and to make it dependent on the variables with imputed values. Hence, there will be feedback between both models that may affect the imputation process and the estimation of the other model parameters.

Full details about how to fit these models in `R` are provided in the associated R code (see Section 5 below). A new `MIINLA` R package which implements the approach proposed here and that can be easily used together with the `R-INLA` package is available at https://github.com/becarioprecario/MIINLA.

5. Examples

In this section we develop two examples to show how the imputation method proposed above can be used with INLA under MCAR, MAR and MNAR. The first example shows a typical regression model in biostatistics with real missing data. This is useful to show how a typical multiple linear regression can be used for multiple imputation. The second one is based on spatially correlated data to assess the performance of our proposal on a simulated study in which a spatially correlated covariate is missing. Note that the aim is not to provide a comprehensive analysis of the dataset with missing values but to illustrate the methods described in this paper.

All models have been fit with INLA and its associated `R` package `R-INLA`. The `R` code to reproduce the examples described here is available from a GitHub repository at https://github.com/becarioprecario/MIINLA_paper.

5.1. Imputation using linear models

The `nhanes2` dataset (**?**) in the `mice` `R` package (van Buuren and Groothuis-Oudshoorn, 2011) records data on 25 participants in the National Health and Nutrition Examination Survey (NHANES). Variables in the dataset include body mass index, cholesterol level, age group and hypertensive status. The dataset presents missing observations in body mass index, hypertensive status and cholesterol level.

We will use this dataset to build a model to explain cholesterol level on age group and body mass index, where this is imputed. The imputation model will be based on a linear regression on the age group. There are three age groups 20-39, 40-59 and 60+ years, and the first group will be set as the reference level.

It is worth noting that having missing values in the response variable (i.e., cholesterol level) is not a problem as the predictive distribution can be easily computed with INLA.

Hence, the output from fitting this model will include the posterior distribution of the imputed values as well as the predictive distribution for the missing responses.

The analysis model is the following:

$$chol_i = \beta_0 + \beta_1 age_i^{40-59} + \beta_2 age_i^{60+} + \beta_3 bmi_i + \varepsilon_i,\ i = 1, \ldots, 25$$

where $chol_i$ refers to the cholesterol level, bmi_i to the body mass index, age_i^{40-59} and age_i^{60+} are indicator variables of age for groups 40-59 and 60+, respectively, and ε_i is a Gaussian error term with zero mean and precision τ.

Note that the missing values of bmi_i are obtained from the imputation model based on linear regression discussed above using as predictors variables age_i^{40-59} and age_i^{60+}. The imputation model is specified as

$$bmi_i = \beta_{I0} + \beta_{I1} age_i^{40-59} + \beta_{I2} age_i^{60+} + \varepsilon_{Ii},\ i \in \mathcal{I}.$$

Here, $\mathcal{I}$ represents the set of indices of the observations with missing values of body mass index. Parameters β_{I0}, β_{I1}, β_{I2} represent the intercept and the covariate coefficients used in the imputation model, and ε_{Ii} is a Gaussian error with zero mean and precision τ_I. Note that all the parameters in the imputation model are mainly informed from the observed values of the body mass index and age, and their prior distributions. Because the imputation model is part of the joint model there is also feedback from all the other parts of the model when estimating the imputation model parameters and the imputed values of body mass index.

A logistic regression is used for the missingness mechanism of bmi_i under MAR or MNAR. For MAR we assumed an intercept plus the covariate of age group, while for MNAR we assumed an intercept plus the covariate of bmi_i (that includes the imputed values). For simplicity, the model with both covariates can be represented as

$$\begin{aligned} M_i &\sim Bernoulli(p_i),\ i = 1\ldots, 25 \\ logit(p_i) &= \gamma_0 + \gamma_1 age_i^{40-59} + \gamma_2 age_i^{60+} + \delta bmi_i \end{aligned} \tag{5}$$

where M_i is a missingness indicator for bmi_i (0 for observed and 1 for missing).

Finaly, the priors for the coefficients of the fixed effects are independent Normal distributions with zero mean and precision 0.001. For the precision parameters, a Gamma with parameters 0.01 and 0.01 is used to provide a vague prior. All parameters are considered to be independent a priori.

Note that the model for analysis and the imputation model are the same for the three missingness scenarios (i.e., MCAR, MAR and MNAR). However, the missingness models differ to include different terms to accomodate the different missingness mechanisms; see Table 1 to assess which terms are included in each missingness model.

Table 1. *Posterior mean (and standard deviation) of the parameters from the joint models in the* `nhanes2` *dataset.*

Model	Parameter	Missingness mechanism in the model		
		MCAR	MAR	MNAR
Analysis	β_0	-4.084 (1.209)	-4.233 (0.816)	-4.864 (1.247)
	β_1	1.145 (0.421)	1.154 (0.398)	1.229 (0.447)
	β_2	1.866 (0.541)	1.879 (0.501)	1.940 (0.580)
	β_3	0.111 (0.049)	0.145 (0.044)	0.156 (0.044)
	τ	2.219 (0.786)	2.568 (1.312)	2.620 (1.169)
Imputation	β_{I0}	31.195 (1.569)	30.046 (1.515)	30.401 (1.296)
	β_{I1}	-5.902 (1.985)	-5.204 (2.316)	-4.711 (1.742)
	β_{I2}	-7.395 (1.733)	-5.561 (2.372)	-6.153 (2.126)
	τ_I	0.058 (0.027)	0.073 (0.023)	0.096 (0.030)
Missingness	γ_0	–	-0.337 (0.585)	-4.633 (4.892)
	γ_1	–	1.879 (0.501)	–
	γ_2	–	-0.377 (1.044)	–
	δ	–	–	0.092 (0.167)

Table 1 also shows the different estimates for all the models considered. Regarding the Gaussian analysis model, it seems that all three covariates included in the model play a significant role when explaining cholesterol level. In addition, point estimates are very similar across different missingness mechanisms. In the imputation model, we also observe that point estimates are very similar across missingness mechanisms. Age also plays an important role when imputing the missing values of body mass index. Finally, the different models for the missingness mechanism are not directly comparable.

Under MAR, age^{40-59} helps to explain why some values of body mass index are missing, while under MNAR the missing values do not appear to depend on their actual values as the estimate of δ is close to zero. We have not included age under MNAR in the missingness sub-model because this covariate is already used when imputing the missing values of body mass index, which is included in the linear predictor of the missingness model.

Cholesterol level seems to increase with age. In addition, the imputation models point to that body mass index seems to decrease with age. Although this is counterintuitive, we believe that is due to the general pattern observed in the dataset, which contains data on 25 people and only 13 of them have a complete record (i.e., all the values for all the covariates have been observed so that there are no missing values in the covariates).

As a final remark, it is worth noting that fitting these models took a few seconds. Hence, the sensitivity analysis could include other models than the ones presented here. See, for example, Mason et al. (2012) for a general discussion and alternative models for the sensitivity analysis. Larger datasets may take longer to run, but INLA will be able to fit these models faster than typical MCMC algorithms.

5.1.1. Imputation of categorical covariates with missing values

As we have mentioned in the description, this dataset includes an indicator of hypertensive status of the subjects. This categorical covariate also contains several missing values. To illustrate how missing values in continuous and categorical covariates can be handled at the same time we fit a model in which body mass index and hypertensive status are included. The imputation of body mass index will be done within the joint model as previously described, but the imputation of hypertension will be done using a multiple imputation approach; this means that an imputation model will be fit for hypertension, values of hypertensive status sampled from this model and used to fill the gaps in the original dataset. This will provide a number of complete datasets to which the analysis model will be fit; then the results will be pooled to obtain final estimates using Bayesian model averaging with equal weights Gómez-Rubio, Bivand and Rue (2020).

The analysis model becomes:

$$chol_i = \beta_0 + \beta_1 age_i^{40-59} + \beta_2 age_i^{60+} + \beta_3 bmi_i + \beta_4 hyp_i + \varepsilon_i,\ i = 1, \ldots, 25.$$

For simplicity, the missingness mechanism will not be assessed now. This implies assuming MCAR, but we have already seen that the model estimates will be close to model fit under MAR and MNAR for the case of body mass index.

The imputation model for hypertensive status (hyp_i) will be a multinomial model fit using the multinomial-Poisson transformation (Baker, 1994). This will provide estimates of the posterior probabilities of being hypertensive given the age group, which will be used to impute the missing values according to the age group of the patient. These posterior probabilities are shown in Table 2. Note that in this particular case a logistic regression would have been enough, but we have preferred to use the multinomial-Poisson transformation because it is a more general approach for the case of more than two categories.

Table 2. *Posterior probabilities of being hypertensive for the different age groups.*

	Age group		
Hypertensive	20-39	40-59	60+
Yes	1.00	0.66	0.49
No	0.00	0.34	0.51

We have drawn 100 samples to fill in the missing values of the hypertensive status, so that 100 different completed datasets have been used to fit the model. The resulting models have been pooled to obtained the posterior marginals of the model parameters using Bayesian model averaging with equal weights (Gómez-Rubio et al., 2020). These are shown in Table 3.

Table 3. *Estimates of the model parameters using multiple imputation on body mass index and hypertensive status.*

Analysis model	
Parameter	Estimate
β_0	-4.981 (1.166)
β_1	1.208 (0.518)
β_2	1.985 (0.635)
β_3	0.134 (0.072)
β_4	0.027 (0.566)
τ	1.965 (0.994)
Imputation model for bmi_i	
β_{I0}	29.612 (1.474)
β_{I1}	-3.899 (2.114)
β_{I2}	-6.116 (2.337)
τ_I	0.092 (0.034)

As expected, the estimates of the coefficients of age are close to the ones in the previous models. The coefficient of hypertensive status is close to zero, which indicates no association between cholesterol level and hypertensive status. Furthermore, the imputation model for body mass index based on a linear regression on age provides similar estimates to the imputation models fit previously and with similar effects of age on body mass index.

5.2. Simulation study: imputation of correlated data

The second example that we present is a simulation study based on the North Carolina Sudden Infant Death Syndrome (SIDS) dataset. It records several variables, which include the number of sudden infant deaths per county in the period 1974-78 (O_i), the total number of births (N_i), as well as the number of non-white births (NW_i). The expected number of cases in each county (E_i) can be obtained using internal standardization, so that the standardized mortality ratio (SMR) can be computed as O_i/E_i. Furthermore, several authors (see, for example, Cressie, 2015) have described the strong spatial pattern in the data, in the relative risk (estimated using the SMR, for example) and its correlation with the proportion of non-white births.

The model of interest to be fit is simply a Poisson regression, as follows:

$$O_i \sim Po(\mu_i); \mu_i = E_i\theta_i,\ i = 1,\ldots,100,$$

$$\log(\theta_i) = \beta_0 + \beta_1\ nwp_i.$$

Here, covariate nwp_i is the logit of the proportion of non-white births (NW_i), so that it is not bounded, that has been re-centered and re-scaled. This derived covariate has still a strong spatial pattern and a high correlation with the SMR.

Figure 1 shows the SMR for the period 1974-78 and the transformed proportion of non-white births (nwp_i). The SMR shows some areas of high risk and a strong correlation with the proportion of non-white births. Hence, this covariate can be useful when building models to explain the spatial variation of SIDS in North Carolina.

The simulation study will remove 5%, 10%, 15%, 30% and 50% of the covariate values (i.e., proportion of non-white births) using MCAR and MNAR mechanisms. Note that MAR can be regarded as an extension to MCAR that considers other observed covariates in the linear predictor of the logistic regression in the imputation model. Although MAR may seem more reasonable, it is simply a matter of including other covariates in the linear predictor of the missingness model so it is computationally feasible but it adds little to the comparison. This is why we have not considered it.

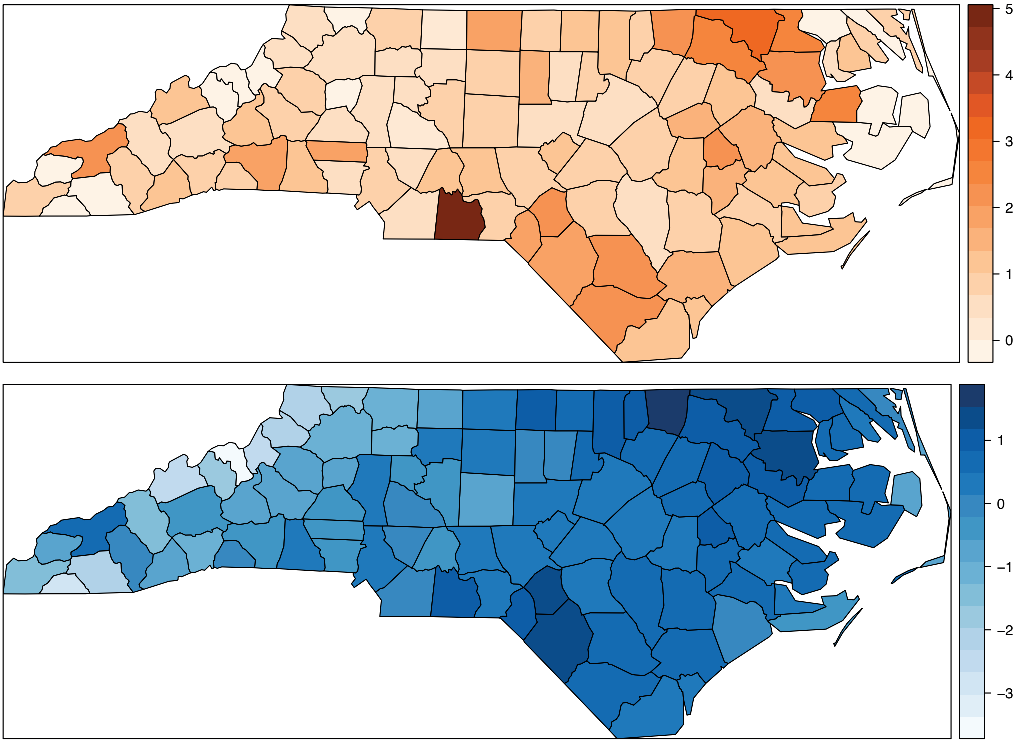


Figure 1. *Standardized mortality ratio (SMR, top) and proportion of non-white births (bottom) in North Carolina in the period 1974-78.*

The missing observations will be nested across the five scenarios, i.e., the observations removed in the 10% scenario will also be removed in the 15% scenario and so on. Furthermore, the probability of being missing under the MNAR mechanism p_i is

$$logit(p_i) = \alpha_M + 5x_i$$

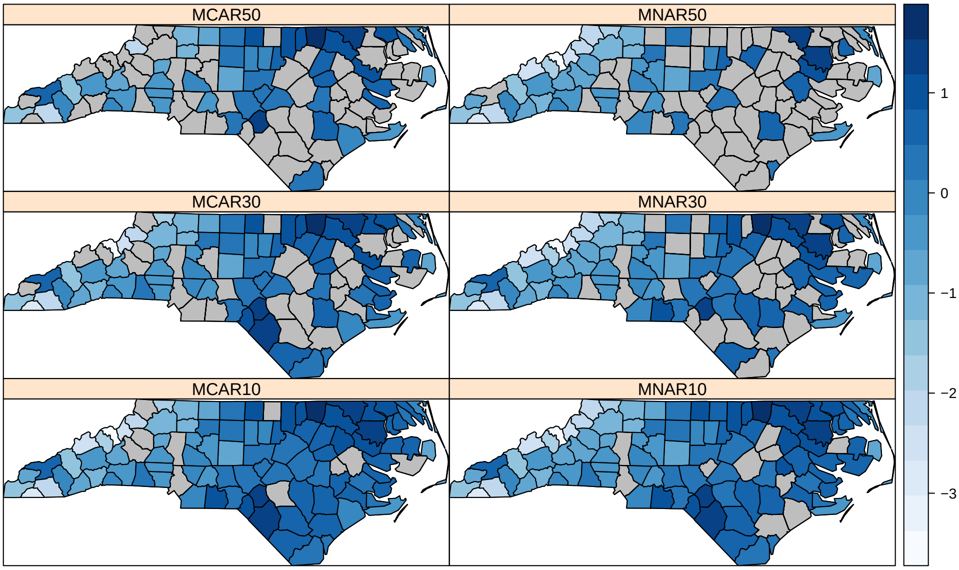


Figure 2. *Missing observations (in grey) of the proportion of non-white births.*

where α_M is set as the logit of 0.5 and x_i represents the value of the covariate with missing values.

This simulation is intended to compare mild to severe missingness under five different scenarios for MCAR and MNAR. Models will be fit assuming MCAR and MNAR missingness, so that we fit 20 models in total. Under MCAR, we only fit the analysis and imputation model. Under MNAR, in addition we will assess whether the joint approach including the missingness mechanism is able to capture the type of missingness.

Figure 2 shows the missing values of the proportion of non-white births for three of the scenarios considered in this simulation study. As it can be seen, when the percentage of missing values is 50% under MNAR missing values concentrate in the counties with high values of the covariate.

In addition, the imputation model proposed is based on the conditional autoregressive specification presented in Section 3.3, so that imputation is included within the main model. This imputation model will have the following parameters: τ is the precision of the CAR specification, ρ the spatial autocorrelation and α the mean value of the covariates.

Finally, a logistic regression on the missingness variable M_i (0 for observed and 1 for missing) is used to model the missingness mechanism (under MNAR):

$$\begin{aligned} M_i &\sim Bernoulli(p_i);\ i = 1, \ldots, 100 \\ logit(p_i) &= \gamma_0 + \gamma_1 nwp_i \end{aligned} \tag{6}$$

Note that the imputed values appear both in the Poisson regression and the sub-model on the missingness mechanism. Non-zero values of γ_1 indicate that the probability of being missing depends on the actual values.

Table 4 summarises the models fit to the data under MCAR. Here, an imputation sub-model for the covariate has been included but not a joint model for the missingness as under MCAR it is not necessary. In general, there are not large differences between the different models fit to the datasets regarding percentage of missing values and type of actual missingness. However, these differences become larger as the proportion of missing values increases, which was to be expected. These differences are noticeable for the case of 50% of missing values both under MCAR and MNAR.

The estimates of the imputation models are quite similar as well, across missingness type in the data and proportion of missing values. However, some differences are observed for 30% and 50% of missing values. In particular, the estimates of α differ.

Table 5 summarises the (joint) models fit to the data considering a MNAR scenario. This includes the model fit to the complete dataset, and the binomial sub-model in the joint model to assess the missingness mechanism. First of all, the posterior distribution of γ_1 helps to determine the missingness mechanism. Its posterior estimate is very close to zero under MCAR, while it is above zero under MNAR (but for the case of 5% of missing values). It is worth stating that it is possible to assess this now because these are simulated data and the true missingness mechanism is known.

Regarding the imputation model, the estimates are very similar across scenarios. Finally, the estimates of the parameters in the Poisson model are in general very close to the model fit to the full dataset.

It is worth noting that under MNAR with 50% of missing observations the point estimates of the parameters in the Poisson sub-model show the largest departure from the model fit to the full dataset. This is probably due to the fact that the imputation model is not able to fully recover the values of the covariates as missing values tend to have high values and there is not enough information in the observed values as to recover this pattern.

To sum up, imputation models behave as expected and provide a good performance in all cases. Most importantly, the joint model is able to identify between MCAR and MNAR situations as well as imputing the covariates and fit the model of interest to the data. Again, this is possible now because the missingness mechanism is known but in real applications we would propose different models and conduct a sensitivity analysis.

When the models fit under MCAR (Table 4) and under MNAR (Table 5) are compared, it should be mentioned that when data under MCAR are analysed both models produce very similar results because the missingness mechanism is, in fact, independent of the observed data. For the analysis of the data simulated under MNAR, differences can be observed because now the missingness mechanism depends on the covariate (including the imputed data) and the estimates of the parameters in the imputation sub-model are different.

Table 4. *Posterior mean (and standard deviation) of the model parameters under MCAR.*

		Model under MCAR						
		Poisson		Imputation			Missingness	
Missingness	% missing	β_0	β_1	τ	ρ	α	γ_0	γ_1
–	0	-0.141 (0.046)	0.524 (0.068)	–	–	–	–	–
MCAR	5	-0.126 (0.047)	0.518 (0.068)	2.129 (0.305)	0.977 (0.022)	-0.211 (0.162)	–	–
MCAR	10	-0.114 (0.047)	0.496 (0.069)	2.076 (0.301)	0.976 (0.024)	-0.215 (0.165)	–	–
MCAR	15	-0.120 (0.048)	0.504 (0.067)	1.915 (0.294)	0.973 (0.027)	-0.234 (0.175)	–	–
MCAR	30	-0.099 (0.049)	0.507 (0.065)	1.776 (0.295)	0.960 (0.039)	-0.175 (0.183)	–	–
MCAR	50	-0.077 (0.051)	0.518 (0.070)	2.461 (0.481)	0.957 (0.044)	0.034 (0.169)	–	–
MNAR	5	-0.131 (0.045)	0.506 (0.067)	2.040 (0.292)	0.977 (0.022)	-0.236 (0.166)	–	–
MNAR	10	-0.138 (0.048)	0.506 (0.068)	1.991 (0.288)	0.976 (0.023)	-0.220 (0.167)	–	–
MNAR	15	-0.110 (0.048)	0.495 (0.068)	1.966 (0.289)	0.976 (0.024)	-0.238 (0.170)	–	–
MNAR	30	-0.105 (0.050)	0.453 (0.070)	1.827 (0.291)	0.975 (0.025)	-0.342 (0.189)	–	–
MNAR	50	-0.064 (0.055)	0.419 (0.061)	1.421 (0.279)	0.964 (0.037)	-0.423 (0.226)	–	–

Table 5. *Posterior mean (and standard deviation) of the model parameters under MNAR.*

		Model under MNAR						
		Poisson		Imputation			Missingness	
Missingness	% missing	β_0	β_1	τ	ρ	α	γ_0	γ_1
–	0	-0.141 (0.046)	0.524 (0.068)	–	–	–	–	–
MCAR	5	-0.121 (0.047)	0.512 (0.068)	2.120 (0.305)	0.977 (0.022)	-0.217 (0.163)	-3.218 (0.565)	-0.514 (0.465)
MCAR	10	-0.111 (0.048)	0.494 (0.069)	2.073 (0.301)	0.975 (0.024)	-0.216 (0.165)	-2.271 (0.349)	-0.074 (0.392)
MCAR	15	-0.127 (0.049)	0.505 (0.067)	1.903 (0.293)	0.972 (0.027)	-0.218 (0.176)	-1.821 (0.309)	0.359 (0.396)
MCAR	30	-0.110 (0.050)	0.507 (0.065)	1.768 (0.294)	0.960 (0.039)	-0.141 (0.187)	-0.896 (0.232)	0.339 (0.309)
MCAR	50	-0.079 (0.054)	0.518 (0.070)	2.458 (0.480)	0.956 (0.044)	0.040 (0.176)	-0.014 (0.203)	0.038 (0.307)
MNAR	5	-0.132 (0.045)	0.502 (0.068)	2.046 (0.293)	0.976 (0.023)	-0.236 (0.165)	-3.286 (0.720)	0.810 (0.795)
MNAR	10	-0.153 (0.049)	0.486 (0.071)	1.964 (0.287)	0.977 (0.022)	-0.225 (0.170)	-2.947 (0.849)	1.661 (0.828)
MNAR	15	-0.133 (0.049)	0.481 (0.069)	1.928 (0.287)	0.977 (0.023)	-0.227 (0.173)	-2.225 (0.529)	1.306 (0.592)
MNAR	30	-0.152 (0.052)	0.423 (0.069)	1.688 (0.285)	0.976 (0.024)	-0.190 (0.200)	-1.385 (0.450)	1.477 (0.492)
MNAR	50	-0.172 (0.060)	0.380 (0.060)	1.230 (0.266)	0.969 (0.032)	-0.093 (0.253)	-0.303 (0.351)	1.576 (0.434)

Finally, we have included the posterior distributions of some imputed values of the covariate in Figure 3. In particular, we have considered the dataset with 50% missing values under MNAR and taken nine counties with missing values that have missing values also in the simulated data under MCAR. This produces a set of counties with a wide variety in the posterior marginals of the imputed values. The posterior marginals shown are for the imputation model under MCAR in Table 4 (dashed line) and the imputation model under MNAR in Table 5 (dotted line). The vertical solid line shows the actual value of the missing covariate. Furthermore, we have kept the same axes scale in all plots so that differences are appreciated better.

In general, both marginals are close in all cases. Under MNAR (dotted lines), the posterior mode seems to be closer to the actual value for most of the counties in the plot. This should not be surprising as this is the actual missingness mechanism in the data.

As the counties considered here are also present in the case in which the missingness mechanims is MCAR, it could be possible to check what happens between models that assumed MCAR and MNAR when the actual missingness is MCAR. In this case, the posterior marginals of the missing values (assuming MCAR and MNAR) look the same for each county because accounting for the missingness model does not affect the model estimates. This shows that handling imputation of missing values with INLA is an interesting way to conduct sensitivity analysis.

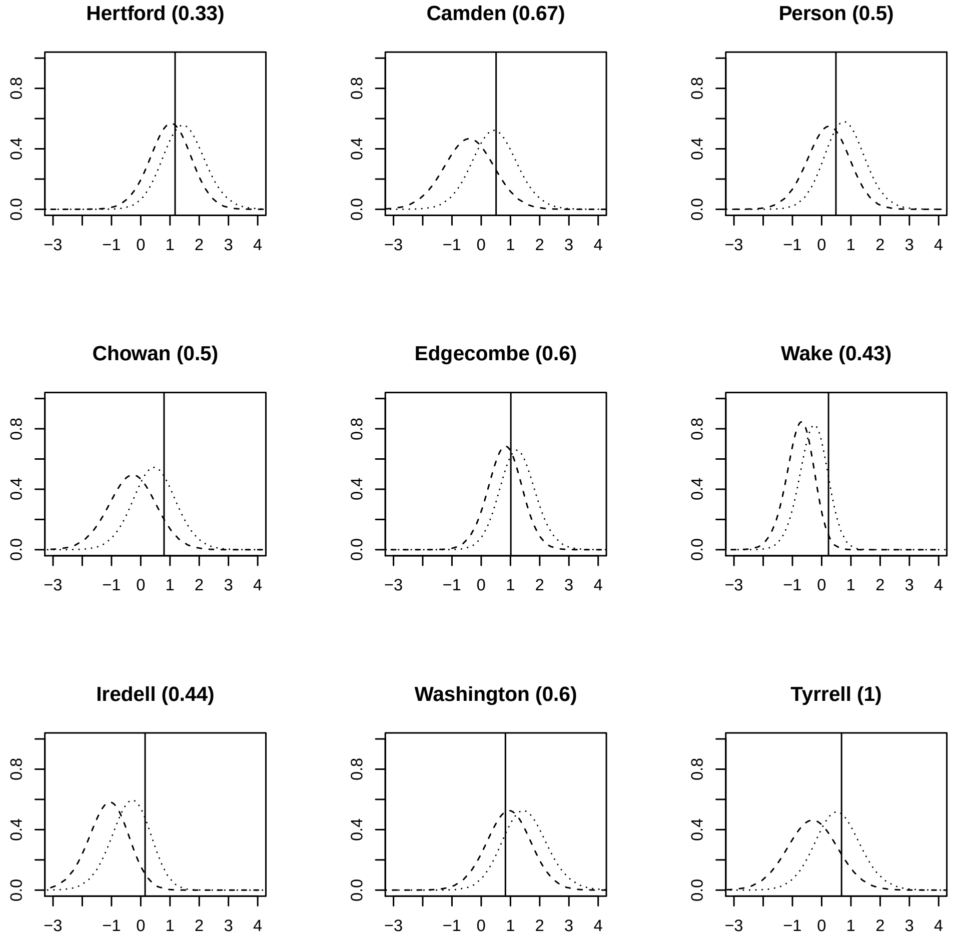


Figure 3. *Posterior marginal distributions of some of the imputed values for missingness of 50% under MNAR. The lines represent the actual value (solid vertical line), the posterior marginal from the MCAR model (dashed line) and the posterior marginal from the MNAR model (dotted line). The value between parenthesis corresponds to the proportion of missing values in the neighbour counties.*

6. Discussion

This paper shows how the general problem of dealing with missing observations in the covariates and performing multiple imputation under different missingness mechanisms can be recast within the framework of latent Gaussian Markov random field models. This has the main advantage that models expressed as latent GMRFs can be fit through INLA, making inference fast. Furthermore, this fills an important gap in the INLA methodology as now models with missing values in the covariates can be easily fit.

Imputation models for the covariates can also take many different forms when defined as GMRFs. In this work we have only considered a linear regression model and spatially correlated model for imputation, but other similar imputation models could be easily developed. For example, these could tackle missing observations in longitudinal data or time series. Furthermore, the methods proposed can be extended to consider imputation of more than one covariate at the same time by relying on multivariate Gaussian models.

The implementation of the multiple imputation models take the form of new latent effects for the `R-INLA` package and they are available within the `MIINLA` package for the `R` programming language. These new latent effects have been developed using the `rgeneric` framework for latent effects development within the `R-INLA` package. Nonetheless, this approach could be implemented in any other software packages for Bayesian inference.

Although we have focused on imputation of continuous covariates, missing values in categorical covariates can also be handled. However, as stated in the paper, this case does not fit within the paradigm of latent GMRF models easily. However, INLA can be used to propose an imputation model for the missing categorical data and to fit the model of interest to these imputed datasets. The fitted models can then be combined to account for the uncertainty of the imputed values in the estimation of the model parameters using Bayesian model averaging.

When the missing values of the categorical covariates index a latent effect the imputation of missing values becomes more complex. This is the case, for example, when random effects are estimated for different groups in the data using multilevel models. However, this scenario could also be handled using the multiple imputation methods described in this paper.

In addition to handling and imputing missing values, this new framework allows us to consider the missingness mechanism using a joint model fit within the INLA methodology. Hence, the analysis of data with missing observations can now be completely carried out within the INLA framework.

Sensitivity analysis on the missingness mechanism, required when it is not ignorable, can benefit from the the computational speed of the INLA method. First of all, models are fit faster than with typical MCMC methods, which helps to define the scenarios to test. Secondly, more scenarios can be tested as the time required to fit the models is reduced.

Acknowledgements

V. Gómez-Rubio has been supported by grants MTM2016-77501-P and PID2019-106 341GB-I00 from the Spanish Ministry of Economy and Competitiveness co-financed with FEDER funds, grant SBPLY/17/180501/000491 and SBPLY/21/180501/000241 funded by Consejería de Educación, Cultura y Deportes (JCCM, Spain) and FEDER. Marta Blangiardo acknowledges partial support through the grant R01HD092580 funded by the National Institute of Health and from the MRC Centre for Environment and Health, which is currently funded by the Medical Research Council (MR/S019669/1).

References

Baker, S. G. (1994). The multinomial-Poisson transformation. *The Statistician*, 43(4):495–504.

Barber, X., Conesa, D., Lladosa, S., and Lòpez-Quílez, A. (2016). Modelling the presence of disease under spatial misalignment using Bayesian latent Gaussian models. *Geospatial Health*, 11(1):1–10.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, Ltd, Chichester, UK.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, Boca Raton, FL.

Carpenter, J. R. and Kenward, M. G. (2012). *Multiple Imputation and its Application*. John Wiley & Sons, Ltd, Chichester, UK.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and inverse probability weighting for analyses with missing data. *Journal of the Royal Statistical Society: Series A*, 169:571–584.

Carpenter, J. R., Kenward, M. G., and White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random - a weighting approach. *Statistical Methods in Medical Research*, 16:259–275.

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons, Inc., Hoboken, NJ, revised edition.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press, New York, NY.

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W. V., Franco, O. H., and Lesaffre, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, 35(17):2955–2974.

Forlani, C., Bhatt, S., Cameletti, M., Krainski, E., and Blangiardo, M. (2020). A joint Bayesian space-time model to integrate spatially misaligned air pollution data in R-INLA. *Environmetrics*, 31(8):e2644.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/ Hierarchical Models*. Cambridge University Press, New York.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. Chapman & Hall/CRC, Boca Raton, FL.

Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2020). Bayesian model averaging with the integrated nested Laplace approximation. *Econometrics*, 8(2):23.

Gómez-Rubio, V. and Rue, H. (2018). Markov chain Monte Carlo with the integrated nested Laplace approximation. *Statistics and Computing*, 28(5):1033–1051.

Held, L. and Rue, H. (2010). Conditional and intrinsic autoregressions. In Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, chapter 13, pages 201–216. Chapman & Hall.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, 73(4):423–498.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Little, R. J. A. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons, Inc., Hoboken, NJ, 3rd edition.

Martino, S. and Riebler, A. (2019). Integrated nested Laplace approximations (INLA). arXiv:1907.01248 [stat.CO].

Mason, A., Richardson, S., Plewis, I., and Best, N. (2012). Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics*, 28(2):279–302.

Mason, A. J. (2009). *Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies*. Ph.D., Imperial College London.

Nakagawa, S. (2015). Missing data: mechanisms, methods, and messages. In Fox, G. A., Negrete-Yankelevich, S., and Sosa, V. J., editors, *Ecological Statistics: Comtemporary Theory and Practice*, chapter 4, pages 81–105. Oxford University Press, Oxford.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons., Hoboken, NJ.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Boca Raton, FL.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 2(71):1–35.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4(1):395–421.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Trivellore, R. (2015). *Missing Data Analysis in Practice*. Chapman & Hall/CRC, Boca Raton, FL.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). `mice`: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

DOI: 10.2436/20.8080.02.125

Alternate-wrapped circular distributions

Savitri Joshi[1] and R. N. Rattihalli[2]

Abstract

To generate a circular distribution, we use the alternate-wrapping technique (unlike the usual wrapping), by wrapping in the alternate directions, after each single-wrapping. The resulting distribution is called alternate-wrapped distribution. Some general properties and distinctions between the two wrapping schemes are indicated. As an illustration, alternate-wrapped-exponential distribution and alternate-wrapped-normal distribution are considered. The moment and maximum likelihood estimator of the parameters of alternative-wrapped-exponential distribution are obtained and their performance is evaluated using simulation. Maximum likelihood estimators are obtained for the parameters of the alternate-wrapped-normal distribution and simulation study is conducted, and this distribution is used to analyse a real-life data set and is compared with the wrapped normal distribution.

MSC: *62H11, 62P12.*

Keywords: *Akaike information criterion, Bayesian information criterion, circular distribution, exponential distribution, trigonometric moments, wrapped normal distribution.*

1. Introduction

In many real-life situations, characteristics of interest are not linear. For instance, wind directions, the direction of migration of birds, time of occurrence of an event in a day etcetera. These cannot be measured on a linear scale and are circular. If X is a univariate real-valued random variable then $\theta = X(\text{mod}2\pi)$ is called a wrapped circular random variable. The density function of θ is $g_w(\theta) = \sum_{m=-\infty}^{\infty} f(2m\pi + \theta), \quad 0 \leqslant \theta < 2\pi$. For illustrative examples of circular random variables and models, one may refer to Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001). Here, the density function

[1] Corresponding author - Savitri Joshi. Department of Applied Sciences, Indian Institute of Information Technology Allahabad, India. Address: 5015, CC-3, Indian Institute of Information Technology Allahabad, India-211015. Email: savitri.joshi@iiita.ac.in.

[2] 27, Sarnaik Mal, Samrat Nagar, Kolhapur, India. Email: ranganath.rattihalli@gmail.com
Received: March 2022
Accepted: December 2022

of θ is obtained by wrapping the density of X around a unit radius circular cylinder in the anti-clockwise direction, so that $x = 0$ matches with $\theta = 0$. In the literature, many wrapped distributions have been developed: to mention a few, Jammalamadaka and Kozubowski (2004) have discussed some properties of wrapped exponential (WE) and wrapped Laplace distributions. Sarma, Rao, and Girija (2011) discussed the characteristic function of wrapped log-normal and wrapped Weibull distributions. Roy and Adnan (2012) have proposed a wrapped weighted exponential distribution. Adnan and Roy (2014) have introduced wrapped variance gamma distribution. Joshi and Jose (2018) have discussed wrapped Lindley distribution. Yilmaz and Bicer (2018) have introduced a new version of wrapped exponential distribution namely transmuted exponential distribution and have discussed its properties.

In this paper, we propose a new wrapping technique called "alternate-wrapping" and the generated distribution is called an alternate-wrapped distribution. To begin with, consider a density function $f(x)$ corresponding to a non-negative random variable X. Start wrapping the density around the unit radius circular cylinder in the anti-clockwise direction, by setting 0 of X with angle 0. After the first single wrapping in the anti-clockwise direction, the next single wrapping is done in the other direction, and so on. However, for the usual wrapped densities, the wrapping is continued in the same direction. The resulting alternate-wrapped density function has a period 2π. To define the density function on the entire real line, we extend the density function with periodicity 2π. In usual wrapping, for $0 < \delta < 2\pi$, the value of the density f at $\delta,\ 2\pi + \delta,\ 4\pi + \delta, \ldots$ and so on contribute to the wrapped density at $\theta = \delta$. But in alternate-wrapping, the values of the density f at $\delta, 4\pi - \delta, 4\pi + \delta, 8\pi - \delta, \ldots$ and so on contribute to the alternate wrapped density at δ. If f is an arbitrary density function then the density on the support $[0, \infty)$ is wrapped as described above and alternate wrapping of the density on the negative part starts from 0 with the first wrapping of the density (towards $-\infty$) in the clockwise direction and the next wrapping in the anti-clockwise direction and so on.

The rest of the paper is organized as follows. In Section 2, we define alternate-wrapping technique in detail and give some of its properties. In Section 3, alternate-wrapped-exponential (AWE) distribution is considered and some of its properties are given. In Section 4, the alternate-wrapped-normal (AWN) distribution is discussed. In Section 5, a data set of 506 cases of onset of lymphatic leukemia, reported in different months in the UK, is analysed using AWN and wrapped-normal (WN) distributions. Lastly, Section 6 concludes the findings.

2. Alternate-wrapping technique

Let X have a continuous distribution with density function $f(x)$. The total contribution at θ $(0 \leq \theta < 2\pi)$, by the density on the domain $\{x \geq 0\}$ is

$$f^{+}(\theta) = f(\theta) + f(4\pi - \theta) + f(4\pi + \theta) + f(8\pi - \theta) + f(8\pi + \theta) + \ldots.$$

The total contribution at θ by the density on the domain $\{x<0\}$ is

$$f^{-}(\theta)=f(-2\pi+\theta)+f(-2\pi-\theta)+f(-6\pi+\theta)+f(-6\pi-\theta)+\ldots.$$

Thus, under alternate-wrapping the total contribution at θ by the wrapping density is $f^{+}(\theta)+f^{-}(\theta)$.

Definition A circular random variable θ is said to have an *alternate-wrapped* distribution, if the density function is given by

$$g_{aw}(\theta)=f(\theta)+\sum_{m=1}^{\infty}\left(f((-1)^{m}2m\pi+\theta)+f((-1)^{m}2m\pi-\theta)\right),\theta\in[0,2\pi). \qquad (1)$$

For a density with positive support, the alternate-wrapped density is given by

$$g_{aw}(\theta)=f(\theta)+\sum_{m=1}^{\infty}\left(f(4m\pi-\theta)+f(4m\pi+\theta)\right),\quad \theta\in[0,2\pi). \qquad (2)$$

The alternate-wrapping of f for $x>0$ can be viewed as the usual wrapping of its modified version h (say). Under alternate-wrapping of f, for $x>0$, anti-clockwise wrapping is over intervals $(4m\pi,(4m+2)\pi)$ for $m=0,1,2,\ldots,$ and clockwise over the other intervals. To make it anti-clockwise over $(0,\infty)$, we modify f over the intervals $((2m+1)2\pi,(2m+2)2\pi)$, to be $h(x)=f((8m+6)\pi-x)$.
Hence for $x>0$ and $m=0,1,2,\ldots,$ let

$$h(x)=\begin{cases} f(x), & 4m\pi<x\leq(4m+2)\pi \\ f((8m+6)\pi-x), & (4m+2)\pi<x\leq(4(m+1))\pi. \end{cases} \qquad (3)$$

Similarly, for $x\leq 0$, to have clockwise wrapping through out, we need to have modification on the intervals $-(4m+4)\pi<x\leq-(4m+2))\pi$. The resulting function for $x\leq 0$ and $m=0,1,2,\ldots$ is given by

$$h(x)=\begin{cases} f(x), & -(4m+2)\pi<x\leq-4m\pi \\ f(-(8m+6)\pi-x), & -(4m+4)\pi<x\leq-(4m+2))\pi. \end{cases} \qquad (4)$$

Hence, we have the following.

Property 2.1. *If $g_{w}^{(f)}(\theta),g_{aw}^{(f)}(\theta)$ are respectively wrapped and alternate-wrapped density functions generated from f, then*

$$g_{aw}^{(f)}(\theta)=g_{w}^{(h)}(\theta),$$

where h is as defined in (3) *and* (4).

Property 2.2. *Let f and h be the probability density functions of X and $-X$ respectively. Then*

$$g_{aw}^{(h)}(\theta) = g_{aw}^{(f)}(2\pi - \theta).$$

Property 2.3. *The alternate-wrapped density g_{aw} can be written as a mixture of the usual wrapped densities;*

$$g_{aw}(\theta) = p g_{w}^{(f_1)}(\theta) + (1-p) g_{w}^{(f_2)}(\theta), \quad 0 \le p \le 1,$$

where $g_w^{(f_i)}(\theta)$ is the usual wrapped density obtained by wrapping linear density $f_i(x)$, $i = 1, 2$, respectively, as defined in (17) *in the Appendix.*

The proofs of the above properties are given in the Appendix.

Remark 2.1. *If X is symmetric about* 0*, then, g_{aw} is symmetric about* 0 *(or π.)*

Remark 2.2. *In alternate-wrapping, the density function $g(\theta)$ is not necessarily continuous at $\theta = 0$. However, the distribution function is continuous and satisfies the properties of a distribution function.*

Remark 2.3. *For an arbitrary density $f(x)$, if the density on the support $[0, \infty)$ is wrapped first in the clockwise direction and next in the anti-clockwise direction alternatively and the density on the support $(-\infty, 0)$ in first anti-clockwise and next in clockwise direction alternatively, then a new wrapped circular density, say g_{aw}^{-} is obtained. This g_{aw}^{-} is the same as alternate-wrapped density g_{aw} obtained by wrapping $f(-x)$.*

Remark 2.4. *Characteristic Functions: The characteristic functions of the linear distribution and its usual wrapped distribution remain the same. But, in general, the characteristic function of the alternate-wrapped distribution is not equal to that of the linear distribution. However, this is true if the support of the density of X is a subset of $(-2\pi, 2\pi)$, as in this case the alternate-wrapped and the usual wrapped densities are the same.*

Distinction between usual and alternate-wrapping: Let X have the density function $f(x)$ and θ_w be a usual wrapped circular random variable with density function $g_w(\theta)$, obtained by wrapping the density f. Then, we have

$$g_w(\theta) = \sum_{k=-\infty}^{\infty} f(k2\pi + \theta),$$

and the distribution function of θ_w is given by

$$G_w(\alpha) = P(\theta_w \le \alpha) = \int_0^{\alpha} \left(\sum_{k=-\infty}^{\infty} f(k2\pi + \theta) \right) d\theta = \sum_{k=-\infty}^{\infty} \left(\int_0^{\alpha} f(k2\pi + \theta) d\theta \right),$$
$$0 \le \alpha \le 2\pi.$$

Consider the transformation

$$Y = \sum_{k=-\infty}^{\infty} (X - k2\pi) I_{k2\pi \le X < (k+1)2\pi}. \tag{5}$$

Then, the distribution function of Y is given by

$$H(\alpha) = P(Y \le \alpha) = P\left(k(2\pi) \le X < k(2\pi) + \alpha; \text{ for some } k = \cdots, -2, -1, 0, 1, 2, \cdots\right).$$

This implies $H(\alpha) = \sum_{k=-\infty}^{\infty} \int_{k(2\pi)}^{k(2\pi)+\alpha} f(x)dx = \sum_{k=-\infty}^{\infty} \int_0^{\alpha} f(k2\pi + t)dt = P(\theta_w \le \alpha)$. Hence, Y and θ_w are identically distributed. Hence, to generate an observation on θ_w, we generate X and obtain Y by using (5).

Now, on the other hand, let θ_{aw} be an alternate-wrapped circular random variable with density function $g_{aw}(\theta)$, obtained by alternate-wrapping of the density function $f(x)$. Then, we have $g_{aw}(\theta) = \sum_{k=-\infty}^{\infty} f(2k(2\pi) + \theta) + \sum_{k=-\infty}^{\infty} f((2k-1)(2\pi) + (2\pi - \theta))$, which implies

$$g_{aw}(\theta) = \sum_{k=-\infty}^{\infty} \left(f(2k(2\pi) + \theta) + f((2k)(2\pi) - \theta)\right).$$

The distribution function of θ_{aw} is given by

$$P(\theta_{aw} \le \alpha) = \int_0^{\alpha} \left(\sum_{k=-\infty}^{\infty} \left(f(2k(2\pi) + \theta) + f((2k)(2\pi) - \theta)\right) \right) d\theta,$$

which gives

$$P(\theta_{aw} \le \alpha) = \sum_{k=-\infty}^{\infty} \int_{4k\pi - \alpha}^{4k\pi + \alpha} f(x)dx.$$

Consider the transformation

$$Z = \sum_{k=-\infty}^{\infty} |X - 4k\pi| I_{\{(2k-1)2\pi \le X < (2k+1)2\pi\}}. \tag{6}$$

The distribution function of Z is given by

$$M(\alpha) = P(Z \le \alpha) = P\left(|X - 4k\pi| \le \alpha; \text{ for some } k = \cdots, -2, -1, 0, 1, 2, \cdots\right),$$

which implies

$$M(\alpha) = \sum_{k=-\infty}^{\infty} \int_{4k\pi - \alpha}^{4k\pi + \alpha} f(x)dx = P(\theta_{aw} \le \alpha).$$

Thus, θ_{aw} and Z are identically distributed. Thus, to generate observations on θ_{aw}, we generate X and use the transformation Z as given in (6). The distinctions between the usual wrapped variable Y and the alternate-wrapped variable Z can also be viewed in Figure 1.

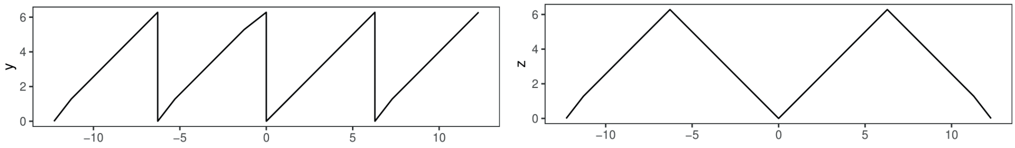


Figure 1. *Periodical behavior of usual and alternate-wrapped circular random variable Y and Z respectively.*

Remark 2.5. *Change of scale: Let X be a random variable with probability density function* $f(x)$*. Let* $U = cX, c > 0$*. For* $0 < \alpha \leq 2\pi$*, we have* $G^U_{aw}(\alpha) = P(U \leq \alpha) = \sum_{k=-\infty}^{\infty} \int_{4k\pi-\alpha}^{4k\pi+\alpha} c^{-1} f(u/c)du$*,*
which implies, $G^U_{aw}(\alpha) = \sum_{k=-\infty}^{\infty} \left(F\left(\frac{4k\pi+\alpha}{c}\right) - F\left(\frac{4k\pi-\alpha}{c}\right)\right)$*, where,* $F(.)$ *is the distribution function of* X*.*

3. Alternate-wrapped exponential (AWE) distribution

Let X follow the exponential distribution, then, we have

$$f(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, x > 0.$$

A random variable θ is said to have an AWE distribution if, the circular density function of θ is given by (2)

$$g_{aw}(\theta) = \lambda e^{-\lambda\theta} + \sum_{m=1}^{\infty} \left(\lambda e^{-\lambda(4m\pi+\theta)} + \lambda e^{-\lambda(4m\pi-\theta)}\right), \ \theta \in [0, 2\pi).$$

That is

$$g_{aw}(\theta) = \frac{\lambda}{(1-e^{-4\pi\lambda})}\left(e^{-\lambda\theta} + e^{-\lambda(4\pi-\theta)}\right), \quad \lambda > 0, \theta\varepsilon[0, 2\pi). \tag{7}$$

For $\lambda = 1$ the usual, circular (Rao et al. (2013)), and 3D circular representations of (7) are given in Figure 2.

It is easy to verify (7) is a mixture of two WE circular densities, wherein the mixing proportion is a function of λ.

Remark 3.1. $g_{aw}(\theta) = g_1(\theta)\rho(\lambda) + g_2(\theta)(1-\rho(\lambda))$ *where* $g_1(\theta) = \frac{\lambda e^{-\lambda\theta}}{1-e^{-2\pi\lambda}}$*,* $g_2(\theta) = \frac{\lambda e^{-\lambda(2\pi-\theta)}}{1-e^{-2\pi\lambda}}$ *and* $\rho(\lambda) = \frac{e^{2\pi\lambda}}{1+e^{2\pi\lambda}}$*.*

Note that, $g_1(\theta)$ is the usual wrapped exponential density, see Jammalamadaka and Kozubowski (2004), and $g_2(\theta) = g_1(2\pi - \theta)$.

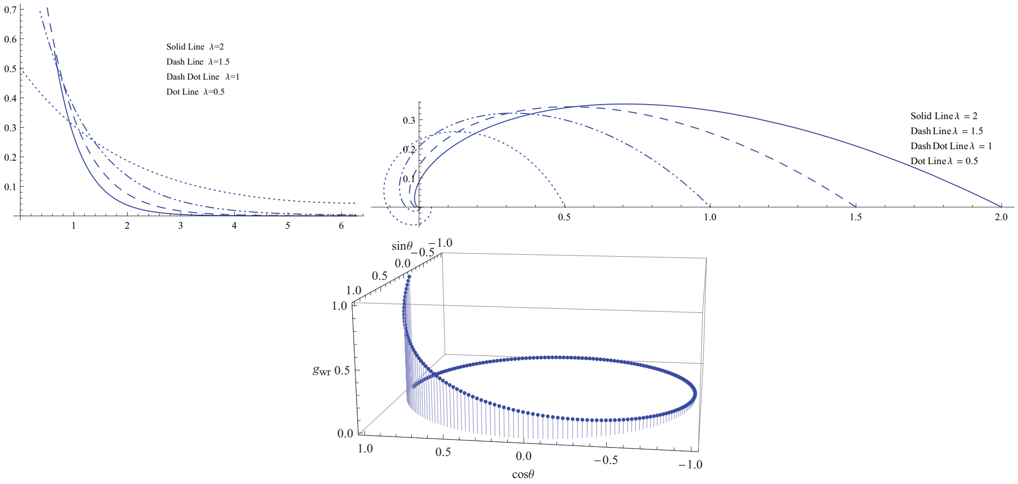


Figure 2. *Usual, circular, and 3D representations of the AWE density.*

3.1. Some properties of the AWE distribution

We obtain trigonometric moments, characteristic function, central trigonometric moments, see Mardia and Jupp (2000), and some other constants for the AWE distribution. All these expressions have been obtained by using Mathematica version 9. However, the same can also be obtained using Remark 3.1.

Trigonometric moments: To obtain the trigonometric moments, we first obtain the following.

$$\alpha_p = E(\cos p\theta) = \frac{\lambda(\lambda + p\operatorname{csch}(2\pi\lambda)\sin(2\pi p))}{\lambda^2 + p^2}, \tag{8}$$

and

$$\beta_p = E(\sin p\theta) = \frac{\lambda p\operatorname{csch}(2\pi\lambda)(\cosh(2\pi\lambda) - \cos(2\pi p))}{\lambda^2 + p^2}.$$

The p^{th} trigonometric moments of θ, denoted by ϕ_p, is the value of the characteristic function at an integer p and can be expressed in terms of α_p and β_p as follows

$$\phi_p = \alpha_p + i\beta_p = \rho_p e^{i\mu_p},$$

where $\rho_p = \sqrt{\alpha_p^2 + \beta_p^2}$ and $\mu_p = \arctan^*\left(\frac{\beta_p}{\alpha_p}\right)$, where the two-argument operator arctan* is as defined in (1.3.5) (Jammalamadaka and SenGupta (2001)).

The mean resultant length ρ is given by $\rho = \rho_1 = \sqrt{\frac{\lambda^2(\lambda^2+\tanh^2(\pi\lambda))}{(\lambda^2+1)^2}}$.

Characteristic function: Since, θ is a periodic circular random variable with period 2π, the characteristic function (corresponding to density (7)) is given by

$$\Phi_p = E(e^{ip\theta}) = E(e^{ip(2\pi+\theta)}) = E(\cos p\theta + i\sin p\theta),$$

which implies $\Phi_p = \frac{\lambda\left(\left(e^{4\pi\lambda}-1\right)\lambda+ip\left(e^{4\pi\lambda}-2e^{2\pi(\lambda+ip)}+1\right)\right)}{\left(e^{4\pi\lambda}-1\right)(\lambda^2+p^2)}, \quad p = 0, \pm 1, \pm 2, \cdots.$

Mean direction: The mean direction is $\mu = \mu_1 = \tan^{-1}\left(\frac{\tanh(\pi\lambda)}{\lambda}\right)$.

The circular variance: The circular variance is

$$V_0 = 1-\rho = 1-\sqrt{\frac{\lambda^2\left(\lambda^2+\tanh^2(\pi\lambda)\right)}{\left(\lambda^2+1\right)^2}},$$

where, $\rho = \rho_1$.

The circular standard deviation: is

$$\sigma_0 = \sqrt{-2\log(1-V_0)} = \sqrt{2}\sqrt{-\log\left(\sqrt{\frac{\lambda^2\left(\lambda^2+\tanh^2(\pi\lambda)\right)}{\left(\lambda^2+1\right)^2}}\right)}$$

Central trigonometric moments: The central trigonometric moments are given by

$$\bar{\phi}_p = E(e^{ip(\theta-\mu)}) = \bar{\alpha}_p + i\bar{\beta}_p,$$

where, μ is the mean direction. This implies

$$\bar{\phi}_p = \frac{\lambda\left(\left(e^{4\pi\lambda}-1\right)\lambda+ip\left(e^{4\pi\lambda}-2e^{2\pi(\lambda+ip)}+1\right)\right)e^{-i\mu p}}{\left(e^{4\pi\lambda}-1\right)\left(\lambda^2+p^2\right)}.$$

Since, $\bar{\phi}_p = E(e^{ip(\theta-\mu)}) = E(\cos p(\theta-\mu)+i\sin p(\theta-\mu))$, therefore,

$$\begin{aligned} E(\cos p(\theta-\mu)) &= \bar{\alpha}_p \\ &= \frac{\lambda\left(2e^{2\pi\lambda}p\sin((2\pi-\mu)p)+\left(e^{4\pi\lambda}+1\right)p\sin(\mu p)+\left(e^{4\pi\lambda}-1\right)\lambda\cos(\mu p)\right)}{\left(e^{4\pi\lambda}-1\right)\left(\lambda^2+p^2\right)}, \end{aligned}$$

and

$$\begin{aligned} E(\sin p(\theta-\mu)) &= \bar{\beta}_p \\ &= \frac{\lambda\left(\left(e^{4\pi\lambda}+1\right)p\cos(\mu p)-2e^{2\pi\lambda}(\lambda\sinh(2\pi\lambda)\sin(\mu p)+p\cos((2\pi-\mu)p))\right)}{\left(e^{4\pi\lambda}-1\right)\left(\lambda^2+p^2\right)}. \end{aligned}$$

The coefficient of skewness, $\zeta_1^0 = \frac{\bar{\beta}_2}{V_0^{3/2}}$, is given by

$$\zeta_1^0 = \frac{\lambda\left(2\left(e^{2\pi\lambda}-1\right)\cos(2\mu)-\left(e^{2\pi\lambda}+1\right)\lambda\sin(2\mu)\right)}{\left(e^{2\pi\lambda}+1\right)\left(\lambda^2+4\right)\left(1-\sqrt{\frac{\lambda^2\left(\lambda^2+\tanh^2(\pi\lambda)\right)}{\left(\lambda^2+1\right)^2}}\right)^{3/2}}.$$

The coefficient of kurtosis, $\zeta_2^0 = \frac{\bar{\alpha}_2-(1-V_0)^4}{V_0^2}$, is given by

$$\zeta_2^0 = \frac{\frac{\lambda(\lambda\cos(2\mu)+2\tanh(\pi\lambda)\sin(2\mu))}{\lambda^2+4} - \frac{\lambda^4\left(\lambda^2+\tanh^2(\pi\lambda)\right)^2}{(\lambda^2+1)^4}}{\left(\sqrt{\frac{\lambda^2\left(\lambda^2+\tanh^2(\pi\lambda)\right)}{(\lambda^2+1)^2}}-1\right)^2}.$$

3.2. Estimation of parameter

In this section, we obtain the MLE and the moment estimator of the parameter λ of AWE distribution.

MLE for AWE

Here we obtain the MLE of λ, the parameter of the AWE distribution. Let $\theta_1, \theta_2, \cdots, \theta_n$ be a random sample from the model (7). Then, the log-likelihood function is given by $\log L = n\log\lambda - n\log(1-e^{-4\pi\lambda}) + \sum_{i=1}^{n}\log\left(e^{-\lambda\theta_i} + e^{-\lambda(4\pi-\theta_i)}\right)$. The MLE of λ is obtained by solving the likelihood equation given by

$$\frac{\partial \log L}{\partial\lambda} = \frac{n}{\lambda} + \frac{4n\pi e^{-4\pi\lambda}}{1-e^{-4\pi\lambda}} - \sum_{i=1}^{n}\frac{4\pi e^{-\lambda(4\pi-\theta_i)} - \theta_i e^{-4\pi\lambda}}{\left(e^{-\lambda\theta_i}+e^{-\lambda(4\pi-\theta_i)}\right)} = 0. \tag{9}$$

Since (9) can not be solved analytically, we use a numerical method to obtain the MLE of λ. For this, we use the maxLik package in R which maximizes a function by using Newton-Raphson algorithm.

Moment estimator for AWE

Here we obtain the moment estimator for the parameter λ. Let $\theta_1, \cdots, \theta_n$ be a random sample from AWE distribution. To obtain the moment estimator, we equate $E(\cos\theta)$ to the mean of $\cos\theta_i's$. From (8) with $p = 1$, we will have $\frac{\lambda^2}{\lambda^2+1} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i = \bar{C}$ and hence the moment estimator is

$$\tilde{\lambda} = \sqrt{\frac{\bar{C}}{1-\bar{C}}}. \tag{10}$$

3.3. Simulation study

In this section, to evaluate the performance of the estimators, we carry out a simulation study. For large samples, consistency, asymptotic unbiasedness and asymptotic normality of MLE and moment estimator follow from the large sample theory. It is also well supported by the simulation study conducted. Here, we generate 10000 samples for finite sample sizes, 20 and 50 from the density (7) for different values of λ. Here instead of generating θ using (7), we generate Z as defined in (6). To obtain the MLE, we use the maxLik package in R since (9) can not be solved analytically. To obtain the moment estimators we use (10). Based on the simulation study, performances of estimators for small and moderate sample sizes are evaluated and results are reported in Tables 1 and 2.

Table 1. *Average values of bias, mean square error (MSE) and variance (var) for $\hat{\lambda}$ based on 10000 samples.*

n	$\lambda = 0.5$			$\lambda = 1.0$		
	bias($\hat{\lambda}$)	MSE($\hat{\lambda}$)	var($\hat{\lambda}$)	bias($\hat{\lambda}$)	MSE($\hat{\lambda}$)	var($\hat{\lambda}$)
20	0.017117	0.020537	0.019378	0.046972	0.060530	0.058539
50	0.010271	0.007256	0.007245	0.018571	0.024428	0.021548
	$\lambda = 2.0$			$\lambda = 4.0$		
20	0.118710	0.271947	0.237354	0.204775	1.042776	0.933834
50	0.040330	0.089551	0.085022	0.075890	0.351541	0.339197

Table 2. *Average values of bias, mean square error (MSE), and variance (var) for $\tilde{\lambda}$ based on 10000 samples.*

n	$\lambda = 0.5$			$\lambda = 1.0$		
	bias($\tilde{\lambda}$)	MSE($\tilde{\lambda}$)	var($\tilde{\lambda}$)	bias($\tilde{\lambda}$)	MSE($\tilde{\lambda}$)	var($\tilde{\lambda}$)
20	0.043820	0.051156	0.049241	0.038183	0.086479	0.085029
50	0.002439	0.023788	0.023785	0.015761	0.030834	0.030588
	$\lambda = 2.0$			$\lambda = 4.0$		
20	0.131001	0.309610	0.292478	0.347391	1.338005	1.217446
50	0.048435	0.097680	0.095343	0.127815	0.416024	0.399727

Based on the results obtained from Table 1 and 2 the following are some observations.
(i) The values of bias, MSE, and variance decrease as the sample size increases.
(ii) The bias, MSE, and variance values increase with the increase in the value of λ and decrease with the increase in sample size.
(iii) The MSE and variance of the MLE is lesser than that of the moment estimator for all values of λ.
(iv) The bias of MLE is higher than the moment estimator for $\lambda < 1$ and is lesser otherwise.
The box plots of the estimators and kernel density estimators of the densities have been plotted. Additionally, the histograms have been drawn for the estimated values and normal distribution has been fitted to those histograms which validate that the estimators are asymptotically normal. All these graphs and plots have been provided as supplementary material.

4. Alternate-wrapped normal (AWN) distribution

Let X have $N(\mu,\sigma^2)$, then the density function of X is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2};-\infty<x<\infty,-\infty<\mu<\infty,\sigma^2>0. \tag{11}$$

From (1), the alternate-wrapped density function corresponding to (11) is given by

$$\begin{aligned}g_{aw}(\theta)=\frac{1}{\sigma\sqrt{2\pi}}&\left[e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(-2\pi+\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(-2\pi-\theta-\mu)^2}\right.\\&\left.+\sum_{m=1}^{\infty}\left(e^{-\frac{1}{2\sigma^2}(4m\pi\pm\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(-(4m+2)\pi\pm\theta-\mu)^2}\right)\right].\end{aligned} \tag{12}$$

$$\begin{aligned}g_{aw}(\theta)=\frac{1}{\sigma\sqrt{2\pi}}&\left[\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)+\exp\left(-\frac{(-2\pi-\theta-\mu)^2}{2\sigma^2}\right)+\exp\left(-\frac{(-2\pi+\theta-\mu)^2}{2\sigma^2}\right)\right.\\&+\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{16\pi^2m^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{8\pi m(\theta-\mu)}{2\sigma^2}\right)\\&+\exp\left(-\frac{(-\theta+\mu+2\pi)^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{16\pi^2m^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{8\pi m(-\theta+\mu+2\pi)}{2\sigma^2}\right)\\&+\exp\left(-\frac{(\theta+\mu)^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{16\pi^2m^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{-8\pi m(\theta+\mu)}{2\sigma^2}\right)\\&\left.+\exp\left(-\frac{(\theta+\mu+2\pi)^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{16\pi^2m^2}{2\sigma^2}\right)\sum_{m=1}^{\infty}\exp\left(-\frac{8\pi m(\theta+\mu+2\pi)}{2\sigma^2}\right)\right].\end{aligned}$$

$$\begin{aligned}g_{aw}(\theta)=\frac{1}{2\sigma\sqrt{2\pi}}&\left[2\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)+2\exp\left(-\frac{(-2\pi-\theta-\mu)^2}{2\sigma^2}\right)\right.\\&+2\exp\left(-\frac{(-2\pi+\theta-\mu)^2}{2\sigma^2}\right)+\sum_{m=1}^{\infty}\exp\left(-\frac{16\pi^2m^2}{2\sigma^2}\right)\left[\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)\right.\\&\left(\frac{1}{\exp\left(-\frac{2\pi(\theta-\mu)}{\sigma^2}\right)-1}\right)+\exp\left(-\frac{(\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(-\frac{2\pi(\theta+\mu)}{\sigma^2}\right)-1}\right)\\&+\exp\left(-\frac{(2\pi+\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(\frac{2\pi(2\pi+\theta+\mu)}{\sigma^2}\right)-1}\right)\\&\left.\left.+\exp\left(-\frac{(2\pi-\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(\frac{2\pi(2\pi-\theta+\mu)}{\sigma^2}\right)-1}\right)\right]\right].\end{aligned} \tag{13}$$

Since, $\vartheta_3\left(0, e^{-\frac{8\pi^2}{\sigma^2}}\right) = 1 + 2\sum_{n=1}^{\infty}\exp\left(-8\pi^2/\sigma^2\right)^{n^2}$ is the elliptic theta function (please refer EllipticTheta 2022), (13) can also be written as

$$\begin{aligned}
g_{aw}(\theta) = \frac{1}{2\sigma\sqrt{2\pi}}\Bigg[&2\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right) + 2\exp\left(-\frac{(-2\pi-\theta-\mu)^2}{2\sigma^2}\right)\\
&+ 2\exp\left(-\frac{(-2\pi+\theta-\mu)^2}{2\sigma^2}\right) + \left(\vartheta_3\left(0, e^{-\frac{8\pi^2}{\sigma^2}}\right) - 1\right)\Bigg[\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)\\
&\left(\frac{1}{\exp\left(-\frac{2\pi(\theta-\mu)}{\sigma^2}\right) - 1}\right) + \exp\left(-\frac{(\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(-\frac{2\pi(\theta+\mu)}{\sigma^2}\right) - 1}\right)\\
&+ \exp\left(-\frac{(2\pi+\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(\frac{2\pi(2\pi+\theta+\mu)}{\sigma^2}\right) - 1}\right)\\
&+ \exp\left(-\frac{(2\pi-\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(\frac{2\pi(2\pi-\theta+\mu)}{\sigma^2}\right) - 1}\right)\Bigg]\Bigg].
\end{aligned}$$

Remark 4.1. *If X follows $N(0,1)$, then $g_w(\theta) = g_{aw}(\theta)$.*

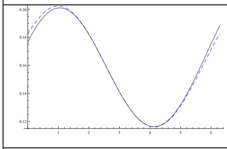
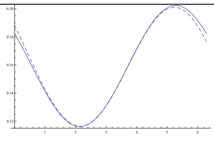
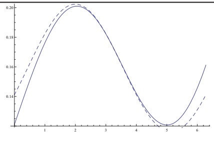
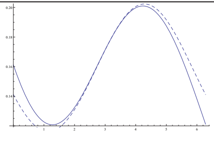
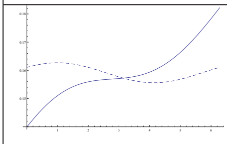
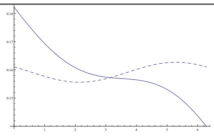
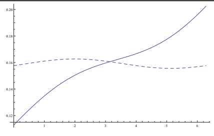
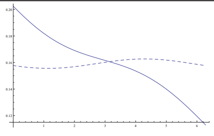
Property 4.1. *Let $g_{aw}(\theta,\mu)$ be the AWN density corresponding to $N(\mu,\sigma^2)$, then*

$$g_{aw}(\theta,\mu) = g_{aw}(2\pi-\theta,-\mu), \quad 0<\theta<2\pi.$$

The proof of this property is given in the Appendix.

The AWN and WN densities are plotted for different values of μ and σ^2 in Table 3.

Table 3. *AWN (solid line) and WN (dashed line) densities for different values of μ and σ^2.*

$\mu=1$, $\sigma^2=4$	$\mu=-1$, $\sigma^2=4$	$\mu=2$, $\sigma^2=4$	$\mu=-2$, $\sigma^2=4$
			
$\mu=1$, $\sigma^2=9$	$\mu=-1$, $\sigma^2=9$	$\mu=2$, $\sigma^2=9$	$\mu=-2$, $\sigma^2=9$
			

4.1. Estimation of parameters

In this section, we obtain the MLEs of μ and σ, the parameters of the AWN distribution. Let $\theta_1, \theta_2, \cdots, \theta_n$ be a random sample from the model (13). Then, the log-likelihood function is given by

$$\begin{aligned}\log L = & -\frac{n}{2}\log 2\pi - n\log\sigma - n\log 2 + \sum_{i=1}^{n}\log\Bigg[2\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right) \\ & + 2\exp\left(-\frac{(-2\pi-\theta-\mu)^2}{2\sigma^2}\right) + 2\exp\left(-\frac{(-2\pi+\theta-\mu)^2}{2\sigma^2}\right) + \sum_{m=1}^{\infty}\exp\left(-\frac{16\pi^2m^2}{2\sigma^2}\right) \\ & \Bigg[\exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(-\frac{2\pi(\theta-\mu)}{\sigma^2}\right)-1}\right) + \exp\left(-\frac{(\theta+\mu)^2}{2\sigma^2}\right) \\ & \left(\frac{1}{\exp\left(-\frac{2\pi(\theta+\mu)}{\sigma^2}\right)-1}\right) + \exp\left(-\frac{(2\pi+\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(\frac{2\pi(2\pi+\theta+\mu)}{\sigma^2}\right)-1}\right) \\ & + \exp\left(-\frac{(2\pi-\theta+\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\exp\left(\frac{2\pi(2\pi-\theta+\mu)}{\sigma^2}\right)-1}\right)\Bigg]\Bigg]. \end{aligned} \tag{14}$$

To obtain the MLEs of μ and σ, we maximize the log-likelihood (14) numerically by using the maxLik package in *R*. For the computational purpose, we use the first three terms of the series $\sum_{m=1}^{\infty}\exp\left(-\frac{8\pi^2m^2}{\sigma^2}\right)$ yielding accuracy to at least five decimal places.

4.2. Simulation study

The simulation study is conducted by using different values of μ and σ for the sample sizes of 20 and 50 with 10000 replications. The AWN random variables are generated using (6). The MLEs of μ and σ are obtained by maximizing (14) using the maxLik package in R. The simulation results are reported as supplementary material. The following observations are made based on the simulation results.

1. For a given value of μ, the bias, MSE, and variance of $\hat{\mu}$ and $\hat{\sigma}$ increase as σ increases.

2. The bias, MSE, and variance of $\hat{\mu}$ and $\hat{\sigma}$ decrease with the increase in sample size.

For large samples, consistency, asymptotic unbiasedness, and asymptotic normality of MLEs follow from the large sample theory.

5. Data analysis

In this section, we fit the AWN and WN distributions to the data set of 506 cases of onset of lymphatic leukemia, reported in different months in the UK during 1946-1960 (Mardia and Jupp (2000)). For analysis purposes, the months are transformed into angles by assigning 30° sector to every month. The data are grouped, therefore all observations recorded every month are assigned to the midpoint of the interval, for example for the month of January, February, and March, the observations are assigned to the corresponding angles 15°, 45°, and 75°, respectively. By using the maxLik package in R we obtain, the MLEs (standard error) for both parameters of the AWN distribution as $\hat{\mu} = -2.3925$ (0.2010) and $\hat{\sigma} = 2.0465$ (0.1096). For the AWN model, μ and σ are the parameters and μ need not correspond to the mean direction of the corresponding AWN model. We also obtain the MLEs with standard errors for the parameters of WN distribution by using again the maxLik package in R: the values of the estimators are $\hat{\mu} = -2.7640$ (0.3181) and $\hat{\sigma} = 2.1440$ (0.1411). We apply the chi-square goodness of fit tests to both the AWN and WN models, by making six classes of the given data. We obtain the chi-square statistic values as 1.89 and 1.69 for the AWN and WN models respectively. The p-values of the statistics for both AWN and WN models are 0.8641 and 0.8901, respectively, which indicates that WN fits marginally better than AWN. Calculations of the chi-square statistics are given in supplementary material.

To evaluate the performance of the estimators under the AWN and WN models, we obtain AIC and BIC values. The AIC and BIC values for the AWN model are 1854.46 and 1862.91, which are very close to those of the WN model, 1853.31 and 1861.76, respectively. Based on the AIC and BIC values, we observe that the WN performs marginally better than the AWN, whereas based on the standard errors, the AWN performs marginally better than the WN.

Thus, one may conclude that for the considered data set, overall both the models perform almost the same.

6. Conclusion

In this paper, the concept of a novel wrapping technique called the alternate-wrapping technique has been introduced to generate circular models. Though the alternate-wrapped distributions are unable to retain some of the properties such as continuity at zero and the simplicity of obtaining characteristic functions, as are in the usual wrapping, they have some interesting properties, for example being expressible as a mixture of two usual wrapped distributions and others as indicated in the manuscript. The class of alternate-wrapped distributions widens the scope for research in circular models and data analysis. To enhance the class of circular distributions and for the circular data analysis, one can generate new alternate-wrapped versions of different distributions.

References

Adnan, M. A. S. and Roy, S. (2014). Wrapped variance gamma distribution with an application to wind direction. Journal of Environmental Statistics, 6(2), 1-10.

EllipticTheta 2022, accessed on 09 August 2022, https://reference.wolfram.com/language/ref/EllipticTheta.html.

Jammalamadaka, S. R. and SenGupta, A. (2001). Topics in Circular Statistics, World Scientific, New York.

Jammalamadaka, S. R. and Kozubowski, T. J. (2004). New families of wrapped distributions for modeling skew circular data. Communications in Statistics-Theory and Methods, 33(9), 2059-2074.

Joshi, S. and Jose, K. K. (2018). Wrapped Lindley Distribution. Communication in Statistics: Theory and Methods, 47(5): 1013-1021.

Mardia, K. V. and Jupp, P. E. (2000). Directional Statistics, 2nd edition, Wiley, New York.

Roy, S. and Adnan, M. A. S. (2012). Wrapped weighted exponential distributions. Statistics and Probability Letters, 82, 77-83.

Sarma, R., Rao, A. V. D. and Girija, S. V. S. (2011). On characteristic functions of the wrapped lognormal and the wrapped Weibull distributions. Journal of Statistical Computation and Simulation, 81(5), 579-589.

Yilmaz, A. and Bicer, C. (2018). A new wrapped exponential distribution. Mathematical Sciences, 12, 285-293.

Appendix

Property 2.2 Let f and h be the probability density functions of X and $-X$ respectively. Then

$$g_{aw}^{(h)}(\theta) = g_{aw}^{(f)}(2\pi - \theta).$$

Proof Let the density of X be $f(x)$. Now, if $Y = -X$ then, the density of Y can be given as $h(y) = f(-x)$. The total contribution at θ by the density h on the domain $\{y \geq 0\}$ is given by $h^+(\theta) = h(\theta) + h(4\pi - \theta) + h(4\pi + \theta) + h(8\pi - \theta) + h(8\pi + \theta) + \ldots$, which implies $h^+(\theta) = f(-\theta) + f(-4\pi + \theta) + f(-4\pi - \theta) + f(-8\pi + \theta) + f(-8\pi - \theta) + \ldots$. Hence,

$$\begin{aligned} h^+(2\pi - \theta) &= f(-2\pi + \theta) + f(-2\pi - \theta) + f(6\pi + \theta) + f(-6\pi - \theta) + f(-10\pi - \theta) + \ldots \\ &= f^-(\theta). \end{aligned} \tag{15}$$

The total contribution at θ by the density h on the domain $\{y < 0\}$ is $h^-(\theta) = h(-2\pi + \theta) + h(-2\pi - \theta) + h(-6\pi + \theta) + h(-6\pi - \theta) + h(-10\pi + \theta) + h(-10\pi - \theta) + \ldots$, which implies $h^-(\theta) = f(2\pi - \theta) + f(2\pi + \theta) + f(6\pi - \theta) + f(6\pi + \theta) + f(10\pi - \theta) + h(10\pi + \theta) + \ldots$. Hence,

$$\begin{aligned} h^-(2\pi - \theta) &= f(\theta) + f(4\pi - \theta) + f(4\pi + \theta) + f(8\pi - \theta) + f(8\pi + \theta) + f(12\pi - \theta) + \ldots \\ &= f^+(\theta). \end{aligned} \tag{16}$$

Thus, from (15) and (16), we have $h^{+}(2\pi-\theta)+h^{-}(2\pi-\theta)=f^{+}(\theta)+f^{-}(\theta)$. That is, $g_{aw}^{(h)}(\theta)=g_{aw}^{(f)}(2\pi-\theta)$ Hence the proof.

Property 2.3 The alternate-wrapped density g_{aw} can be written as a mixture of the usual wrapped densities;

$$g_{aw}(\theta)=pg_{w}^{(f_1)}(\theta)+(1-p)g_{w}^{(f_2)}(\theta), \quad 0\leq p\leq 1,$$

where $g_{w}^{(f_i)}(\theta)$ is the usual wrapped density obtained by wrapping linear density $f_i(x)$, $i=1,2$, respectively, as defined in (17).

Proof In the usual wrapping the entire density on positive support is wrapped in the anti-clockwise direction, starting from 0, and the entire density on negative support is wrapped in the clockwise direction, starting from 0. But in alternate-wrapping, the density on the positive support is alternatively wrapped, it is anti-clockwise wrapping for the intervals $(4\pi r, 4\pi r+2\pi)$ for $r=0,1,2,\ldots$ and on the remaining intervals it is clockwise wrapping. Similarly, in alternate-wrapping, the density on the negative support is alternatively wrapped, it is clockwise wrapping on the intervals $(-4\pi r-2\pi,-4\pi r)$ for $r=0,1,2,\ldots$, and on the remaining intervals it is anti-clockwise wrapping.
Let

$$A=\left\{\bigcup_{r=0}^{\infty}(4\pi r, 4\pi r+2\pi]\right\}\bigcup\left\{\bigcup_{r=0}^{\infty}(-4\pi r-2\pi,-4\pi r]\right\}.$$

That is, on the set A piece-wise wrapping directions for usual and alternate-wrapping are the same. On $R-A$, define a function f^*, by considering the reflection of the original function on each interval in $R-A$ of length 2π. On the interval $(k2\pi,\ (k+1)2\pi]$ for $k=1,3,5,\cdots$, we change the function by taking its reflection on $(2k+1)\pi$, the mid-point of the interval. Whereas, on the interval $(-(k+1)2\pi,\ -k2\pi]$ for $k=1,3,5,\cdots$, we change the function by taking its reflection on $-(2k+1)\pi$, the mid-point of the interval. The graphical representation of the set A together with the function f^* is given in Figure 4.

Let

$$f^*(t)=\begin{cases} f((2k+1)\pi-t) & \text{for}(k2\pi\leq t\leq (k+1)2\pi], k=1,3,5,\cdots \\ f(-(2k+1)\pi-t) & \text{for}(-(k+1)2\pi\leq t\leq -k2\pi], k=1,3,5,\cdots. \end{cases}$$

Let $p=\int_A f(x)dx$, if $p=1$ then the support for f will be a subset of A and in this case the alternate-wrapped and the usual wrapped densities will be the same. Let $0<p<1$, we note that

$$f_1(x)=\frac{f(x)I_{\{x\in A\}}}{p}, \quad f_2(x)=\frac{f^*(x)(1-I_{\{x\in A\}})}{1-p} \tag{17}$$

are density functions.

Let $g_{i,w}(\theta)$ be usual wrapped density corresponding to linear density $f_i(x)$, $i=1,2$, respectively. Then, the alternate-wrapped density $g_{aw}(\theta)$ can be written as

$$g_{aw}(\theta) = p g_w^{(f_1)}(\theta) + (1-p) g_w^{(f_2)}(\theta), \quad 0 \le p \le 1,$$

We know that for an integrable function $h(x)$ over a finite interval (a,b), we have $\int_a^b h(x)dx = \int_a^b h^*(x)dx$, where $h^*(x) = h(a+b-x)$ is the reflection of the function $h(x)$ about $(a+b)/2$. Hence the result.

If $p=0$, then we will have $g_{aw}(\theta) = g_w^{(f_2)}(\theta)$.

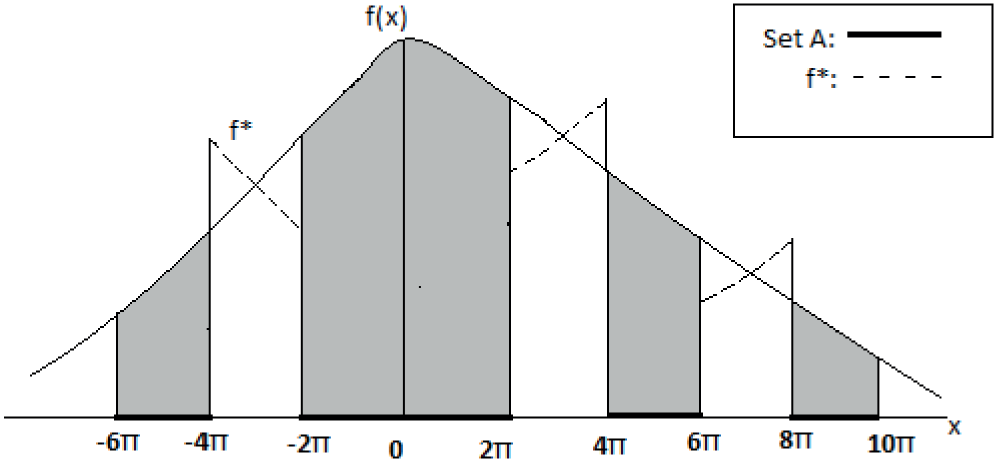


Figure 3. *Representation of the set A and the function f^*.*

Property 4.1 Let $g_{aw}(\theta,\mu)$ be the AWN density corresponding to $N(\mu,\sigma^2)$, then

$$g_{aw}(\theta,\mu) = g_{aw}(2\pi-\theta,-\mu), \quad 0<\theta<2\pi.$$

Proof Using (12), we can write

$$\begin{aligned} g_{aw}(2\pi-\theta,-\mu) = \frac{1}{\sigma\sqrt{2\pi}}\Bigg[& e^{-\frac{1}{2\sigma^2}(2\pi-\theta+\mu)^2} + e^{-\frac{1}{2\sigma^2}(-2\pi+2\pi-\theta+\mu)^2} + e^{-\frac{1}{2\sigma^2}(-2\pi-2\pi+\theta+\mu)^2} \\ & + \sum_{m=1}^{\infty} e^{-\frac{1}{2\sigma^2}(4m\pi+2\pi-\theta+\mu)^2} + \sum_{m=1}^{\infty} e^{-\frac{1}{2\sigma^2}(4m\pi-2\pi+\theta+\mu)^2} \\ & + \sum_{m=1}^{\infty} e^{-\frac{1}{2\sigma^2}(-(4m+2)\pi+2\pi-\theta+\mu)^2} + \sum_{m=1}^{\infty} e^{-\frac{1}{2\sigma^2}(-(4m+2)\pi-2\pi+\theta+\mu)^2}\Bigg], \end{aligned}$$

which implies

$$\begin{aligned}g_{aw}(2\pi-\theta,-\mu)=\frac{1}{\sigma\sqrt{2\pi}}\Bigg[&e^{-\frac{1}{2\sigma^2}(-2\pi+\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(4\pi-\theta-\mu)^2}\\&+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}(-(4m+2)\pi+\theta-\mu)^2}+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}((4m-2)\pi+\theta+\mu)^2}\\&+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}(4m\pi+\theta-\mu)^2}+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}((4m+4)\pi-\theta-\mu)^2}\Bigg].\end{aligned}$$

This gives

$$\begin{aligned}g_{aw}(2\pi-\theta,-\mu)=\frac{1}{\sigma\sqrt{2\pi}}\Bigg[&e^{-\frac{1}{2\sigma^2}(-2\pi+\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2}+e^{-\frac{1}{2\sigma^2}(-2\pi-\theta-\mu)^2}\\&+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}(-(4m+2)\pi+\theta-\mu)^2}+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}(-(4m+2)\pi-\theta-\mu)^2}\\&+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}(4m\pi+\theta-\mu)^2}+\sum_{m=1}^{\infty}e^{-\frac{1}{2\sigma^2}(4m\pi-\theta-\mu)^2}\Bigg]=g_{aw}(\theta,\mu).\end{aligned}$$

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics. Furthermore, once a paper is accepted it must not be published elsewhere in the same or similar form.

SORT is an **Open Access** journal which **does not** charge publication **fees**.

Articles should be preferably of an applied nature and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board.

Submission of papers must be in electronic form only at our **RACO** (Revistes Catalanes en Accés Obert) submission site. Initial submission of the paper should be a single document in **PDF** format, including all **figures and tables** embedded in the main text body. **Supplementary material** may be submitted by the authors at the time of submission of a paper by uploading it with the main paper at our RACO submission site. **New authors**: please register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

The article should be prepared in **double-spaced** format, using a **12-point** typeface. SORT strongly recommends the use of its LaTeX template.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75–100 words) followed by the keywords and MSC2010 Classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**. Otherwise the article may be returned for language improvement before entering the review process. The article must also use inclusive language, that is, language that avoids the use of certain expressions or words that might be considered to exclude particular groups of people, especially gender-specific words, such as "man", "mankind", and masculine pronouns, the use of which might be considered to exclude women. The article should also inform about whether the original data of the research takes gender into account, in order to allow the identification of possible differences.

Bibliographic references within the text must follow one of these formats, depending on the way they are cited: author surname followed by the year of publication in parentheses [e.g., Mahalanobis (1936) or Rao (1982b)])]; or author surname and year in parentheses, without comma [e.g. (Mahalanobis 1936) or (Rao 1982b) or (Mahalanobis 1936, Rao 1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

□ Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.

□ Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.

□ Chapter in book: Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.

□ Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be used only when absolutely necessary. They should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Papers should not normally exceed about **25 pages** of the **PDF** format (**40 pages** of the format provided by the SORT **LaTeX** template) including all figures, tables and references. Authors should consider transferring content such as long tables and supporting methodological details to the online supplementary material on the journal's web site, particularly if the paper is long.

Once the article has positively passed the first review round, the executive editor assigned with the evaluation of the paper will send comments and suggestions to the authors to improve the paper. At this stage, the executive editor will ask the authors to submit a revised version of the paper using the SORT **LaTeX** template.

Once the article has been accepted, the journal editorial office will **contact the authors** with further instructions about this final version, asking for the source files.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75-100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

The articles are licensed under a Creative Commons Atribution-NonCommercial-NoDerivs 3.0 Spain License.

You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work), you may not use the work for commercial purposes and you may not alter, transform, or build upon the work.

Published articles represent the author's opinions; the journal SORT-Statistics and Operations Research Transactions does not necessarily agree with the opinions expressed in the published articles.

SORT Statistics and Operations Research Transactions
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58 - 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name ______________________________

____

Organisation ______________________________

____

Street Address ______________________________

____

Zip/Postal code ______________ City ______________

State/Country ______________ Tel. ______________

Fax ______________ NIF/VAT Registration Number ______________

E-mail ______________________________

Date ______________

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)***) from now on

Annual subscription rates:

– Spain: €42 (4 % VAT included)
– Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:
sort@idescat.cat