NAIVE STATISTICAL ANALYSES FOR COVID-19: APPLICATION TO DATA FROM BRAZIL AND ITALY

Carlos Alberto de Bragança PEREIRA[1]
Luiz Ricardo NAKAMURA[2]
Paulo Canas RODRIGUES[3]

■ ABSTRACT: This article is a direct consequence of the authors' desire to discuss the role of statistics in data analysis. The analysis of coronavirus (COVID-19) databases are used as to show simple, but powerful statistical frameworks. We do believe that models for assessing future trends in temporal data in general, and in cases and/or deaths of COVID-19, belongs to the area of (Bio)Statistics. Just as engineers use knowledge of physics, chemistry and often architecture, when constructing bridges, buildings and roads, statisticians use knowledge of mathematics, computer science and even physics for modelling, analysing, and forecasting in order to transform data into information. While the statistician's contribution is rarely acknowledged, everyone knows that a building is a work of an engineer. Nonetheless, nowadays statistics has been gaining the attention that it deserves due to the rise of big data and data science that was built on the foundations of statistics. This article shows that, even with only basic knowledge of statistics, one can adequately collaborate with the community in dealing with very important issues such as the COVID-19 numbers. In order to model and to obtain predictions we use well-known distributions to statisticians working on survival analysis: gamma, Weibull and log-normal distributions. We also make use of singular spectrum analysis, a simple non-parametric time series methodology, for an analogous purpose. Survival analysis is a research area widely used in Biostatistics and even in Reliability, while time series analysis is widely used across areas where the data is measured along the time.

■ KEYWORDS: COVID-19; Statistics in practice; Survival analysis; Time series analysis.

[1]Universidade de São Paulo - USP, Departamento de Estatística, CEP: 05508-010, São Paulo, SP, Brasil. E-mail: *cadebp@gmail.com*
[2]Universidade Federal de Santa Catarina - UFSC, Departamento de Informática e Estatística, CEP: 88040-900, Florianópolis, SC, Brasil. E-mail: *luiz.nakamura@ufsc.br*
[3]Universidade Federal da Bahia - UFBA, Departamento de Estatística, CEP: 40170-110, Salvador, BA, Brasil. E-mail: *paulocanas@gmail.com*

1 Introduction

Statistical concepts have been used since the beginning of civilization, where census were conducted by empires and the trade of commodities was recorded. On the 18th century, the term "statistics" was used in the context of official statistics, mostly in terms of demographic and economic data. In the current days, after many major developments in terms of methodology, but also in terms of computational power, statistical science is of utmost importance in most areas of science and society, including policy maker, health and environment. In the modern world, with the current generation and collection of immense amounts of data, from many areas and in many formats, the quantitative ability to analyse and to transform them into information and decision making, provided by statistical analyses, is essential. On the one hand, statistics uses important mathematical concepts without being part of mathematics. On the other hand, statisticians use important computer science concepts and programming without being part of computer science. Statistics, as a discipline, includes methodological and applied work, being the recent concept of "data science", in its majority, similar to applied statistics.

In this paper we present two simple statistical techniques (DEGROOT and SCHERVISH, 2011), part of two major statistical sub-areas: survival analysis and time series analysis. The use and usefulness of these techniques are illustrated with daily sequence of the frequency of COVID-19 deaths. Other important information such as contamination rates, ages of deaths, type of housing, etc. are left for future work because this study does not intend to be exhaustive, but illustrative.

The motivation for the type of modelling developed here comes from the techniques of survival analysis (biostatistics), failure analysis (reliability) of production systems, and time series analysis.

Let us imagine that the future purchase of lamps for an industry will be carried out by competition among sellers. Information is being collected on the lifetime of the lamps in use at the largest factory in this industry. The objective is to develop a model of probabilities that well describes the performance of the set of lamps. To measure the expected lifetime of a lamp, the statistician bases his analysis on a report containing the moments when each lamp came into operation, the moment when each of the burnt out lamps stopped working and the final moment of data collection with the number of lamps that were still working. Of course, the number of lamps under test is, in this case, known. In the statistician's view, the sampling units are the lamps and the variable to be studied is the lamps survival times.

The above example served as an analogy for the statistical analysis of the pandemic data that currently plagues us. The individual corresponds to a light bulb, the sampling unit, and the number of days, D, that an individual lived until death is the random variable under study, whose baseline is the moment of the first notified death. Information, in our case, is the day of death and not the moment when death occurred. Censorship is done when it is observed that the individual died within the 24 hours of the day of recording. An observation in the sample is the number of days from baseline up the death day of an individual. As in the case

of lamps, the observations are already ordered. For instance, the event observed on the 20th is the frequency of deaths that have already occurred until the 20th; Thus, it is observed the number of times, the frequency, that the event $\{D < 20\}$ occurred.

With these observations, our goal is to find a distribution function for variable D that adequately represents the behaviour of this variable. It is important to remember that our observations do not cover all the support for the possible values of D, but an observation of the tail of the distribution function. The difficulty then is to estimate the parameters of any function in a family of distribution functions. To draw the empirical tail of the distribution, divide the accumulated frequencies obtained in each day by the accumulated frequency of the last observational day. Thus, an empirical distribution function conditional to the data of the last observation day is obtained. We may call this function as the empirical conditional distribution.

Here, we use three families of distributions: gamma (WALCK, 2007), Weibull (RINNE, 2009), and log-normal (AITCHISON and BROWN, 1957). Our job is to find among members of a specific family of distributions the one whose conditional distribution function is the closest to the corresponding empirical conditional distribution function: conditional is on the last day being observed.

The measure of approximation between conditional distributions used in this article is Aitchison's compositional distance (AITCHISON, 1986). After obtaining the best conditional functions of the models, one may compare them and choose the one that has better adjustment for the empirical one.

As a second approach, the non-parametric time series technique singular spectrum analysis (SSA; GOLYANDINA et al. 2001; GOLYANDINA and ZHIGLJAVSKY, 2013; HASSANI and MAHMOUDVAND, 2018) is considered to fit the individual time series, and to illustrate how to estimate the maximum number of deaths. SSA is a simple technique that combines elements of time series analysis, matrix algebra and multivariate statistics. It allows to decompose a given time series into a set of components that can be interpreted as trend components, seasonal components or noise components.

The data used in the paper corresponds to the daily sequence of the frequency of COVID-19 deaths in Brazil and Italy. For Brazil, besides the total number of daily deaths, we also consider the daily number of deaths for each of the 26 states plus the federal district.

The rest of this paper is organised as follows. Section 2 presents a short description of the data, the concepts of probability necessary to adjust the models for the survival analysis and a brief description of SSA. Section 3 presents the results for both survival analysis and time series analysis. On the one hand it illustrates the survival analysis based methodology using the COVID-19 data from Italy, chosen for having a complete database with the entire first phase of contamination and with the important inflection points already observed, and from São Paulo. On the other hand the time series based methodology is illustrated with the data from all 26 Brazil states plus the federal district. The paper ends in Section 4, where some

final considerations are given.

2 Materials and methods

2.1 The data

The data used in the paper includes the official daily number of deaths by COVID-19 in Brazil, as a whole and for each of its 26 states and the federal district, obtained from `https://github.com/wcota/covid19br/blob/master/cases-brazil-states.csv`. The first observation was considered to be the day of the first death, on March 17, 2020, and the last observation was considered to be August 20, 2020, with a total of 157 observations. Figure 1 shows the number of daily deaths, per state and globally. Moreover, we also consider the official daily number of deaths by COVID-19 in Italy from the day of first death, February 20, up to August 20, 2020, with a total number of 182 observations, obtained from `https://www.worldometers.info/coronavirus/country/italy`.

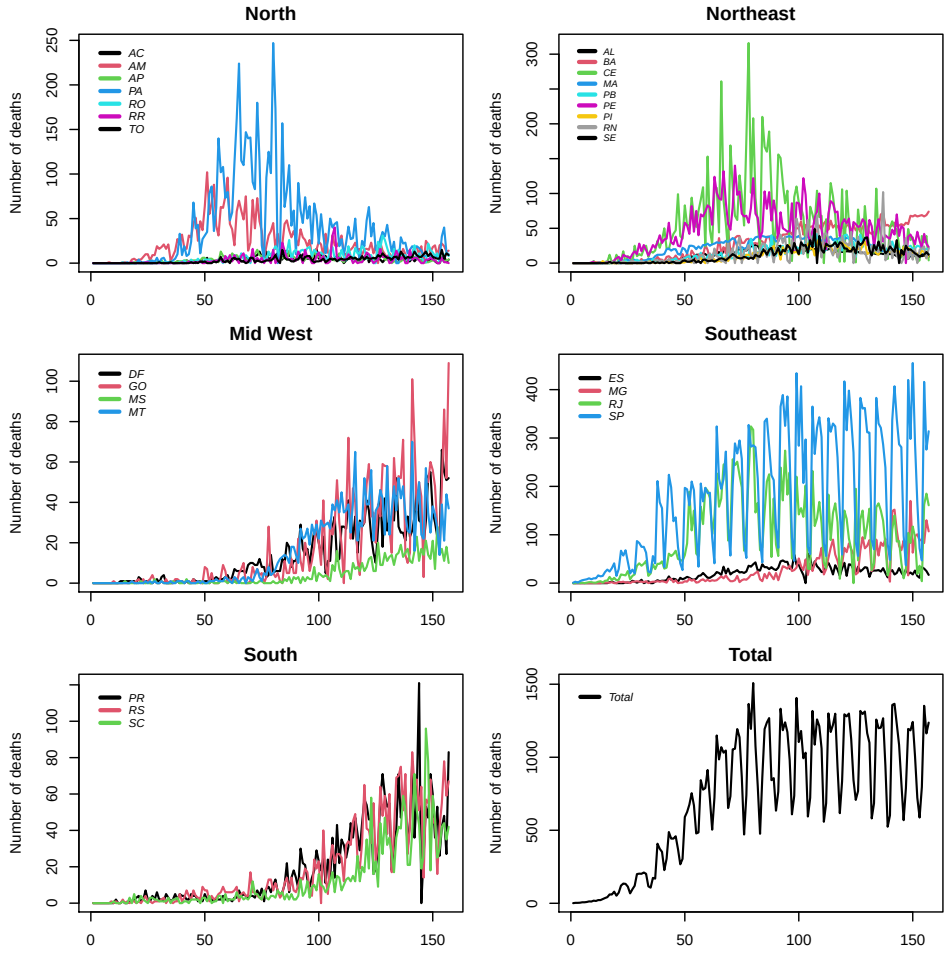

Figure 1 - Number of daily deaths by COVID-19 in Brazil, between March 17, 2020 and August 16, 2020. The top-left plot includes the states in the North region – AC: Acre, AP: Amapá, AM: Amazonas, PA: Pará, RO:Rondônia, RR: Roraima, and TO: Tocantins. The top-right plot includes the states in the Northeast region – AL: Alagoas, BA: Bahia, CE; Ceará, MA: Maranhão, PB: Paraíba, PE: Pernanbuco, PI: Piauí, RN: Rio Grande do Norte, and SE: Sergipe. The mid-left plot includes the states in the Mid West region – DF: Distrito Federal, GO: Goiás, MS: Mato Grosso do Sul, and MT: Mato Grosso. The mid-right plot includes the states in the Southeast region – ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, and SP: São Paulo. The bottom-left plot includes the states in the South region – PR: Paraná, RS: Rio Grande do Sul, and SC: Santa Catarina. The bottom-right plot includes the overall Brazil.

2.2 Survival analysis

The data analysed in this paper are the daily frequency of deaths – the frequency d_i of deaths at the i^{th} day and the accumulated frequency of deaths, D_i, up to the day. The data consists of non-negative integer numbers. Our first step is to define a conditional empirical distribution function (CEDF). The second step is to look for a statistical model whose conditional distribution function (MCDF), adjusts well the CEDF.

Suppose that we are at the m^{th} day of observation. As explained above, for every $i \leq j$, $D_j = \sum_{i=1}^{j} d_i$ and now the CEDF of D_i given D_j can be defined as $Em(D_i|D_j) = D_i/Dj$.

In order to find the member of the family such that $F(i|j)$ is close to $Em(D_i|D_j)$ we use the Aitchison distance (AITCHISON, 1986), $A(F; Em)$, to choose the F with smaller A, building a mesh in the parametric space to find the optimal point. It is noteworthy here that the finer the mesh, the better the result, but unfortunately at a much higher cost. The Aitchison compositional distance can be understood as a standard deviation of the natural log of the ratio of the corresponding elements of two vectors. For more details about the Aitchison's distance, please see Appendix A.

Once the optimal F for the current observation day is found, this probability distribution is used to cover the entire sample space and make predictions associated with its probabilities. Note that for each new data coming from the following observation day, all work must be redone, and the fitted model must produce a new set of forecasts with additional information being incorporated.

In this analysis we considered the gamma, Weibull, and log-normal families, which cumulative distribution functions are given by

- Gamma distibution

$$F(y|\alpha, \beta) = \frac{\gamma(\alpha, y/\beta)}{\Gamma(\alpha)},$$

  where $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter and $\gamma(\alpha, y/\beta) = \int_0^{y/\beta} t^{\alpha-1} e^{-t} dt$ is the lower incomplete gamma function.

- Weibull distribution

$$F(y|\lambda, k) = 1 - \exp\left[-\left(\frac{y}{\lambda}\right)^k\right],$$

  where $\lambda > 0$ and $k > 0$ are the scale and shape parameters, respectively.

- Log-normal distribution

$$F(y|\mu, \sigma^2) = \Phi\left(\frac{\log y - \mu}{\sigma}\right),$$

  where $-\infty < \mu < \infty$ is a location parameter, $\sigma > 0$ is a scale parameter and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

2.3 Singular spectrum analysis

Techniques for time series analysis are a natural choice for data modelling and forecasting when the measurements are conducted across time. Despite the many possible approaches and models, in this paper we will consider a non-parametric time series technique called singular spectrum analysis (SSA), which incorporates elements of classical time series analysis, matrix algebra, and multivariate statistics. SSA allows the decomposition of the original univariate time series into a set of components that can be interpreted as trend, seasonal and cycle components and noise (BROOMHEAD and KING, 1986; GOLYANDINA et al., 2001; GOLYANDINA and ZHIGLJAVSKY, 2013; RODRIGUES et al., 2020), and consists of two stages: decomposition and reconstruction, with two steps each. A short description of the SSA technique is given below, and mode details can be found in, e.g. Golyandina et al. (2001), Golyandina and Zhigljavsky (2013), Hassani and Mahmoudvand (2019) and Rodrigues and Mahmoudvand (2018).

2.3.1 First stage: Decomposition

1st step: Embedding. Let $y_1, \ldots, y_N$ be a time series of length N. Considering a window length L the result of this step is a $L \times K$ matrix $\mathbf{Y} = [Y_1 : \ldots : Y_K]$, where $K = N - L + 1$ and $Y_i = (y_i, \ldots, y_{i+L-1})^T,\ 1 \leq i \leq K$.

2nd step: Singular value decomposition (SVD). In this step, the matrix $\mathbf{Y}$ will be decomposed using SVD as $\mathbf{Y} = \mathbf{Y}_1 + \cdots + \mathbf{Y}_L$, where $\mathbf{Y}_i = \sqrt{\lambda_i} U_i V_i^T$, $\mathbf{Y}_i = \mathbf{0}$ when $\lambda_i = 0$, and $V_i = \mathbf{Y}^T U_i / \sqrt{\lambda_i}$ with $\lambda_1, \ldots, \lambda_L$, the eigenvalues of $\mathbf{S} = \mathbf{Y}\mathbf{Y}^T$ and $U_1, \ldots, U_L$, the corresponding eigenvectors.

2.3.2 Second stage: Reconstruction

3rd step: Grouping. The grouping step corresponds to splitting the elementary matrices into m disjunct subsets $I_1, \ldots, I_m$, and summing the matrices within each group. In our application we will focus in $m = 2$, i.e. only two groups. $I_1 = \{1, \ldots, r\}$ and $I_2 = \{r+1, \ldots, L\}$ are associated with the signal and noise components, respectively.

4th step: Diagonal averaging. This step transforms each matrix $\mathbf{Y}_{I_j}$ into a new series of length N. Using diagonal averaging we have that $\mathbf{Y} = \widetilde{\mathbf{Y}}_{I_1} + \cdots + \widetilde{\mathbf{Y}}_{I_m}$, where $\widetilde{\mathbf{Y}}_{I_j}$ is the Hankelized form of $\mathbf{Y}_{I_j},\ j = 1, \ldots, m$. Considering $\widetilde{y}_{m,n}^{(I_j)}$ the $(m,n)^{th}$ entry of the estimated matrix $\widetilde{\mathbf{Y}}_{I_j}$ and denoting by $\{\widetilde{y}_{j_1}, \ldots, \widetilde{y}_{j_N}\}$ the reconstructed components in the matrix $\widetilde{\mathbf{Y}}_{I_j},\ j = 1, \ldots, m$, applying diagonal averaging follows that

$$\widetilde{y}_{j_l} = \begin{cases} \frac{1}{j_\ell - 1} \sum_{n=1}^{j_\ell - 1} \widetilde{y}_{n,j_\ell - n}^{(I_j)} & 2 \leq j_\ell \leq L-1, \\ \frac{1}{L} \sum_{n=1}^{L} \widetilde{y}_{n,j_\ell - n}^{(I_j)} & L \leq j_\ell \leq K+1, \\ \frac{1}{K+L-j_\ell+1} \sum_{n=n-K}^{L} \widetilde{y}_{n,j_\ell - n}^{(I_j)} & K+2 \leq j_\ell \leq K+L. \end{cases}$$

In this paper, we will illustrate how to make use of SSA to reconstruct the trend of COVID-19 daily deaths and, when possible, to predict the peak of the curve. For that, we will group together the trend components obtained in the third step above.

In this analysis we decided not to focus on model forecasting for a given number of steps ahead. However, the discussion of out-of-sample forecasting in the context of SSA can be found elsewhere, e.g. Danilov (1997), Golyandina et al. (2001), Mahmoudvand et al. (2017), Mahmoudvand & Rodrigues (2018), Rodrigues et al. (2020).

3 Results and discussion

3.1 Survival analysis

In a sequel we illustrate our survival analysis approach by performing the steps described in Section 2.2 in the COVID-19 data, between the date of the first occurance and August 20, 2020. We use data from the state of São Paulo and Italy since they present comparable areas and population size.

The Italian case

First, we consider the three classes of distributions, searching for the best parameter's values in each of them. Table 1 displays the best set of parameter estimates for each of the distribution with their respectively Aitchison's distances obtained against the CEDF. We shall highlight here that for precision adjustment we start the comparison between the distributions from the day where the number of deaths was larger than nine (February 25, 2020) and ending on the August 20, 2020.

Table 1 - The Italy case: parameter estimates and Aitchison's distance (empirical vs model)

Distribution	Aitchison's distance
Gamma($\alpha = 4.9$; $\beta = 11$)	0.24
Weibull($\lambda = 2.6$; $k = 59$)	0.42
Log-normal($\mu = 49$; $\sigma = 1.56$)	0.20

According to the Aitchison's distance values displayed in Table 1, we choose the log-normal distribution to represent the COVID-19 data. Figure 2(a) presents the

two conditional distributions, empirical and log-normal with mean 49 and standard deviation 1.56.

(a) (b)

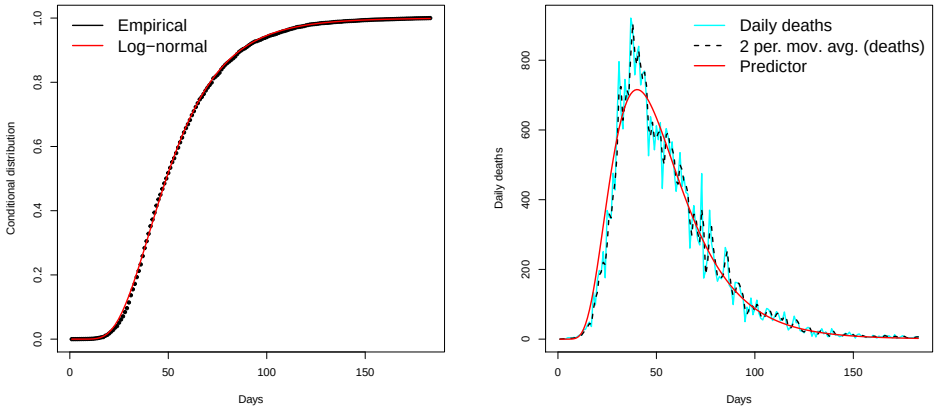

Figure 2 - The Italian case: (a) conditional (August 20, 2020) distributions: empirical vs log-normal distributions; and (b) daily deaths, moving average of order 2 and predictions from log-normal model.

In order to assess the modelling produced by log-normal predictive function, Figure 2(b) shows the predictions, the moving average function of order 2, and daily death observed values. The small variability around the predictive function may indicate that the Italian protocol for the population quarantine should have adopted in other countries or regions. There are countries where there have been a great deal of disagreement between the subregion's protocols.

It is important to observe that with a standard moving average adjustment there is no probability associated to perform interesting predictions. For instance, at the August 20 we may predict the proportion of deaths we have up to that moment. Since these observed data were from the hundred and eighty-second day, we can use the fitted log-normal distribution to compute F(182) which is equal to 0.998415326 (99.84%), leading us to imagine that Italy is almost at the ending of the pandemic first phase. Also, the mode of the density is 717 in March 31. The highest number of deaths was 921 in March 27 with the prediction of having 27% of the total deaths due to the pandemics. Finally, we can predict 35,474 total deaths at December 31, 2020.

The São Paulo case

Once again, we consider the data up to August 20, 2020, but starting in March 16, 2020 (the day before of first death). As we will see, the situation (behaviour)

here is quite different from the previous case. Based on the Aitchison's distance values, as can be seen in Table 2, the best family to describe the COVID-19 data from São Paulo was the Weibull distribution.

Table 2 - The São Paulo case: parameter estimates and Aitchison's distance (empirical vs model)

Distribution	Aitchison's distance
Gamma($\alpha = 4.97$; $\beta = 25$)	0.67
Weibull($\lambda = 2.1$; $k = 196$)	0.21
Log-normal($\mu = 190$; $\sigma = 2.1$)	0.56

To illustrate the quality of the adjustment, Figure 3(a) shows the conformity among the fitted Weibull conditional distribution and the CEDF. In Panel (b), we can see the high variability of daily deaths in the state of São Paulo, which can make predictability to be difficult. Further, we are able to see the differences between the Weibull density and the moving average function of order 7. Nevertheless, Panel (c) shows that the estimated Weibull model could be used for predictions.

We can observed that on August 20, the total death toll was 27,905, which, based on the Weibull estimate, represented 46.61% of the forecast total number of deaths at the end of the pandemic. The adjusted Weibull model predicts that on December 31, there would be 53,518 deaths, corresponding to 89.74% of the total estimated deaths, which would be 58,572. The final number of deaths, and, consequently, the end of the pandemic, was predict to occur on October 7, 2021, by the Weibull model.

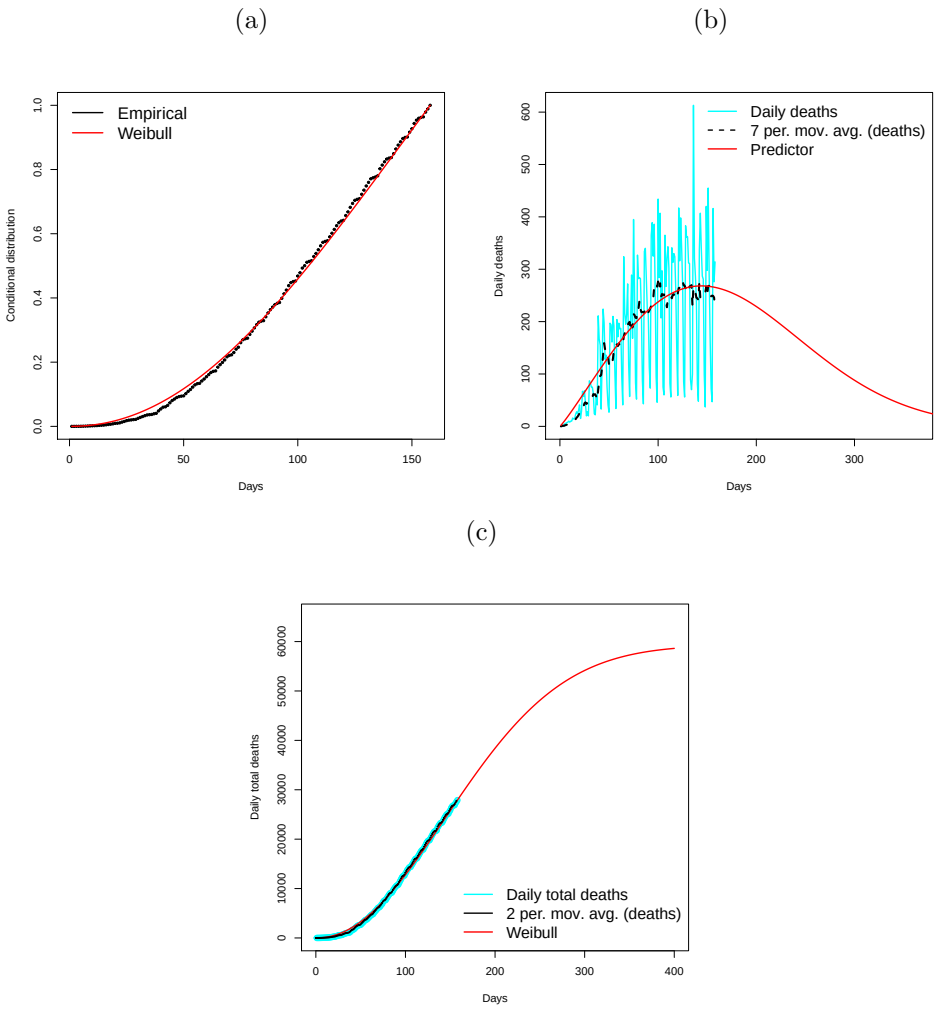

Figure 3 - The São Paulo case: (a) conditional (August 20, 2020) distributions: empirical vs Weibull distributions; (b) daily deaths, moving average of order 7 and predictions from log-normal model; and (c) Daily total deaths prediction vs empirical distribution function.

3.2 Singular spectrum analysis

In this analysis we considered the univariate time series for each Brazilian state, federal district and overall number of deaths by COVID-19, between the date of the first occurrence in each of the time series and August 20, 2020. We considered a window length L proportional to seven and close to the middle of the time series. Table 3 shows the date of the first death, the number of observations, the window length, the number of components used for reconstruction, and the set of indexes associated to the trend components, for each time series under consideration. Other approaches to study Brazilian COVID-19 data can be found elsewhere, e.g. Amaral et al. (2020).

Table 3 - Date of the first death (and the number of days since the first death in Brazil), number of observations (N), window length (L), and number of components used for reconstruction (r), set of trend components among the first r components, and the date where the peak was observed (and the number of days since the first death in Brazil), for each univariate time series

Region	State	1st death	N	L	r	Trend	Peak
North	AC: Acre	06/04/2020 (21)	137	63	7	$\{1,2,3\}$	06/06/2020 (62)
	AM: Amazonas	24/03/2020 (8)	150	70	6	$\{1,2,3,6\}$	14/05/2020 (52)
	AP: Amapá	04/04/2020 (19)	139	63	12	$\{1,2\}$	30/05/2020 (57)
	PA: Pará	01/04/2020 (16)	142	70	9	$\{1,2,3\}$	23/05/2020 (53)
	RO:Rondônia	30/03/2020 (14)	144	70	10	$\{1,2,7\}$	20/07/2020 (113)
	RR: Roraima	03/04/2020 (18)	140	70	12	$\{1,4\}$	01/07/2020 (90)
	TO: Tocantins	15/04/2020 (30)	128	63	6	$\{1,2\}$	20/08/2020 (128)
Northeast	AL: Alagoas	31/03/2020 (15)	143	70	7	$\{1,2,3\}$	14/06/2020 (76)
	BA: Bahia	29/03/2020 (13)	145	70	8	$\{1,2\}$	20/08/2020 (145)
	CE; Ceará	26/03/2020 (10)	148	70	13	$\{1,2,7\}$	02/06/2020 (69)
	MA: Maranhão	29/03/2020 (13)	145	70	8	$\{1,2\}$	22/06/2020 (86)
	PB: Paraíba	31/03/2020 (15)	143	70	10	$\{1,2\}$	16/07/2020 (108)
	PE: Pernanbuco	25/03/2020 (9)	149	70	10	$\{1,2,5\}$	23/05/2020 (60)
	PI: Piauí	27/03/2020 (11)	147	70	6	$\{1,2\}$	12/07/2020 (108)
	RN: Rio Grande do Norte	28/03/2020 (12)	146	70	7	$\{1,2,7\}$	05/07/2020 (100)
	SE: Sergipe	02/04/2020 (17)	141	70	9	$\{1,2\}$	12/07/2020 (102)
Mid West	DF: Distrito Federal	29/03/2020 (13)	145	70	6	$\{1,2\}$	20/08/2020 (145)
	GO: Goiás	26/03/2020 (10)	148	70	6	$\{1,6\}$	20/08/2020 (148)
	MS: Mato Grosso do Sul	31/03/2020 (15)	143	70	8	$\{1,2\}$	15/08/2020 (138)
	MT: Mato Grosso	03/04/2020 (18)	140	70	8	$\{1,2\}$	21/07/2020 (110)
Southeast	ES: Espírito Santo	02/04/2020 (17)	141	70	6	$\{1,2,7\}$	23/06/2020 (83)
	MG: Minas Gerais	30/03/2020 (14)	144	70	6	$\{1,6\}$	19/08/2020 (143)
	RJ: Rio de Janeiro	19/03/2020 (3)	155	77	13	$\{1,2,5\}$	29/05/2020 (72)
	SP: São Paulo	17/03/2020 (1)	157	77	8	$\{1,6\}$	20/08/2020 (157)
South	PR: Paraná	27/03/2020 (11)	147	70	9	$\{1,4\}$	18/08/2020 (145)
	RS: Rio Grande do Sul	25/03/2020 (9)	149	70	9	$\{1,8\}$	19/08/2020 (148)
	SC: Santa Catarina	26/03/2020 (10)	148	70	11	$\{1,6\}$	20/08/2020 (148)
Total	Brazil	17/03/2020 (1)	157	77	9	$\{1,2,7\}$	23/07/2020 (129)

Based on the SSA, for each time series, we created three groups of components: (i) the components associated with the trend; (ii) the components associated with the seasonality; and (iii) the components associated to the noise. Figures 4 and 5 show the sum of all components associated to the trend (7th columns of Table 3)

and to the seasonality ($\{1, \ldots, r\}$ except $\{Trend\}$ in Table 3), for all time series, respectively. To complement and summarise the results in Figure 4, the last column of Table 3 shows the date where the peak was observed and the number of days since the first death in a given state. We shall highlight here that the real peak for the time series that show an increasing trend (e.g. DF, GO, MG; Figure 4) was, most likely, not observed yet. For constant trends (e.g. SP, TOTAL; Figure 4) the peak is of difficult determination but it is a great indication that those number have reached a "plateau".

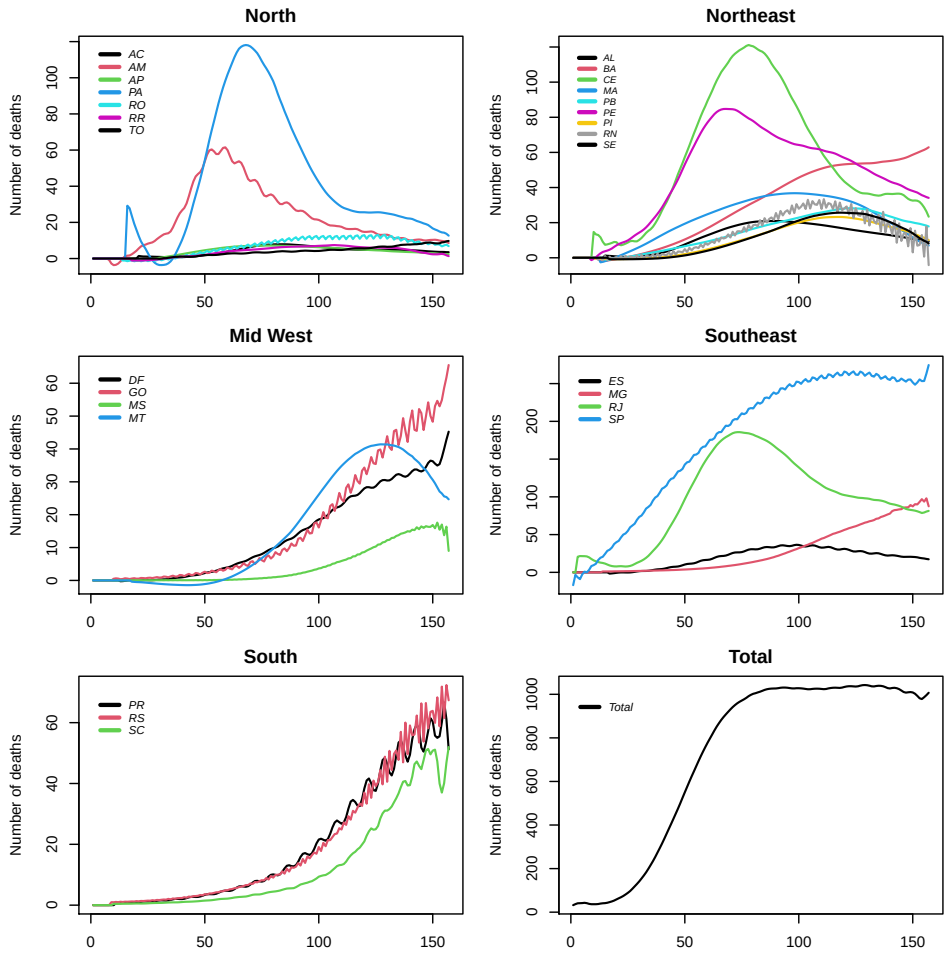

Figure 4 - Trend components based on the SSA. The top-left plot includes the states in the North region – AC: Acre, AP: Amapá, AM: Amazonas, PA: Pará, RO:Rondônia, RR: Roraima, and TO: Tocantins. The top-right plot includes the states in the Northeast region – AL: Alagoas, BA: Bahia, CE; Ceará, MA: Maranhão, PB: Paraíba, PE: Pernanbuco, PI: Piauí, RN: Rio Grande do Norte, and SE: Sergipe. The mid-left plot includes the states in the Mid West region – DF: Distrito Federal, GO: Goiás, MS: Mato Grosso do Sul, and MT: Mato Grosso. The mid-right plot includes the states in the Southeast region – ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, and SP: São Paulo. The bottom-left plot includes the states in the South region – PR: Paraná, RS: Rio Grande do Sul, and SC: Santa Catarina. The bottom-right plot includes the overall Brazil.

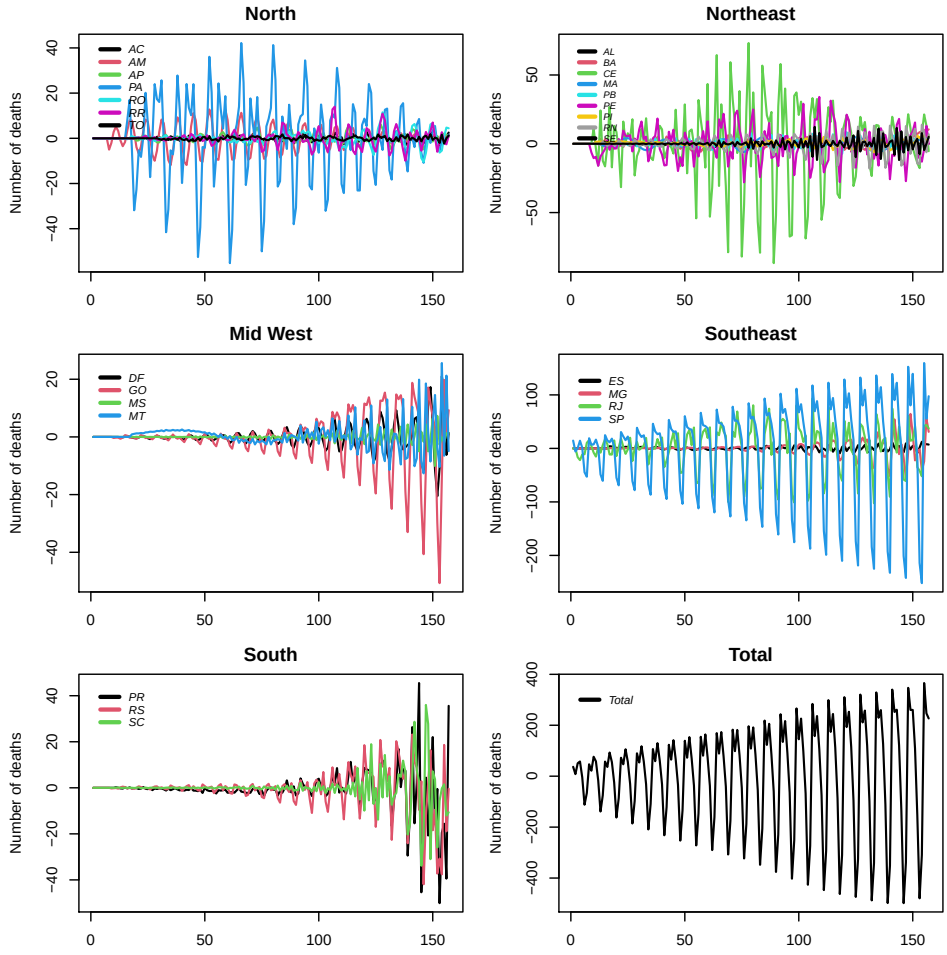

Figure 5 - Seasonal components based on the SSA. The top-left plot includes the states in the North region – AC: Acre, AP: Amapá, AM: Amazonas, PA: Pará, RO:Rondônia, RR: Roraima, and TO: Tocantins. The top-right plot includes the states in the Northeast region – AL: Alagoas, BA: Bahia, CE; Ceará, MA: Maranhão, PB: Paraíba, PE: Pernanbuco, PI: Piauí, RN: Rio Grande do Norte, and SE: Sergipe. The mid-left plot includes the states in the Mid West region – DF: Distrito Federal, GO: Goiás, MS: Mato Grosso do Sul, and MT: Mato Grosso. The mid-right plot includes the states in the Southeast region – ES: Espírito Santo, MG: Minas Gerais, RJ: Rio de Janeiro, and SP: São Paulo. The bottom-left plot includes the states in the South region – PR: Paraná, RS: Rio Grande do Sul, and SC: Santa Catarina. The bottom-right plot includes the overall Brazil.

4 Final considerations

The motivation that the authors had to produce this article comes from the desire to show the social utility of the field of statistics. They could not miss the opportunity to analyse data related to the pandemic that is currently attacking all of us. The reader should remember that the intense and essential statistical work that has been done by statisticians to collaborate with the authorities at crucial moments in decision making has been mentioned seldomly. The credit is given oftentimes to mathematicians, a common error that confuses the statistical work with the mathematical work. Statistics is not a sub-area of mathematics, as some mathematicians like to advertise; statistical foundations are far removed from mathematical thinking. As engineers and physicists, among other types of professionals, statisticians also make use of tools developed by mathematicians, without being one of its sub-areas.

As as example of a direct connection between science fiction and reality, Hari Seldon, a fictional character of the famous Isaac Asimov's Foundation trilogy, portrayed as the creator of a revolutionary science capable of predicting the future of humans, is referred as a mathematician, although Isaac Asimov knew that mathematicians do not do that kind of work with mathematical tools, e.g.

> "Without a doubt the greatest contributions of this personality were in the field of psychohistory. When Seldon started his saga, this field was little more than a set of vague axioms: he transformed it into a profound statistical science." - Asimov's Encyclopedia Galactica, `https://en.wikipedia.org/wiki/Encyclopedia_Galactica`

Having explained our motivation, we can say that the statistical methods described in this paper, and applied to COVID-19 data were chosen to illustrate to the reader with basic statistical training (last year of undergraduate or beginning of graduate studies) that simple methods can provide powerful insights.

With respect to the survival analysis, we can say that it is efficient in the sense of making "good" predictions. We must attract the reader's attention to the fact that the forecasts are based on the data collected until August 20, 2020. However, with the data of the following day, the adjustments, and predictions change. That is, adding relevant information to the database requires us to adjust the model for this new database. Of course, predictions change every time a daily observation is included into the database. We must repeat the adjustment every time new observations are added to the database.

Recall the case of Italy that on August 20, 2020, the total number of deaths was 35,418. The log-normal model used in this paper informs us that, of the estimated total deaths, 99.84% have already occurred. Thus, the total number of deaths on December 31 would be 35,474 with 99.9986% of the deaths occurring until that day. The total number of deaths in Italy is predicted to be 35,475. The same study now considering the state of São Paulo. We observed on August 20 that the total death toll was 27,905, which in the Weibull estimate represented 46.61% of the total death forecast at the end of the pandemic. The adjusted Weibull model predicts

that on December 31, there would be 53,518 deaths, corresponding to 89.74% of the total estimated deaths, which would be 58,572. Using the Weibull model, this final number of deaths was predicted to occur on October 7, 2021.

With respect to the time series analysis, it was possible to decompose each time series in trend, seasonal and noise components. Then, by analysing the overall trend component of each time series, it was possible to identify the peak in the number of deaths for some Brazilian states.

We should always keep in mind that any forecast using probabilistic models has a zero probability of being the exact observed value. What we want, as statisticians, is for the error to be within the allowable, considering losses and/or gains, functions of decision theory.

PEREIRA, C.A.B.; NAKAMURA, L.R.; RODRIGUES, P.C. Análise estatística elementar para COVID-19: Aplicação em dados do Brasil e da Itália. *Rev. Bras. Biom.,* Lavras, v.39, n.1, p.158-176, 2021.

- *RESUMO: Este artigo é uma consequência direta do desejo dos autores de discutir o papel da estatística na análise de dados. Para ilustrar metodologias estatísticas simples, robustas e poderosas, analisamos bancos de dados de coronavírus (COVID-19) obtidos nos seguintes links,* `https://github.com/wcota/covid19br/blob/master/cases-brazil-states.csv` *e* `https://www.worldometers.info/coronavirus/country/italy/`*: dados brasileiros e italianos, respectivamente. Acreditamos que os modelos de avaliação das tendências futuras dos dados temporais em geral e dos casos e/ou óbitos de COVID-19 pertencem à área da (Bio)Estatística. Assim como os engenheiros usam o conhecimento da física, química e muitas vezes da arquitetura, ao construir pontes, edifícios e estradas, os estatísticos usam o conhecimento da matemática, ciência da computação e até física para modelar, analisar e prever, com o objetivo de transformar dados em informação. Embora a contribuição do estatístico raramente seja reconhecida, todos sabem que um edifício é obra de um engenheiro. No entanto, atualmente a estatística vem ganhando a atenção que merece devido ao crescimento de big data e de ciência de dados que foi construída nos fundamentos da estatística. Este artigo mostra que, mesmo com apenas conhecimentos básicos de estatística, pode-se colaborar adequadamente com a comunidade no tratamento de questões muito importantes na sociedade, como os números do COVID-19. A fim de modelar e obter previsões, usamos distribuições bem conhecidas para estatísticos que trabalham com análises de sobrevivência: gamma, Weibull e log-normal. Também fazemos uso da análise espectral singular, uma metodologia não-paramétrica simples usada na análise de séries temporais, para uma finalidade análoga. A análise de sobrevivência é uma área de pesquisa amplamente utilizada em Bioestatística e até mesmo em Confiabilidade, enquanto a análise de séries temporais é amplamente utilizada em áreas onde os dados são medidos ao longo do tempo.*

- *PALAVRAS-CHAVE: COVID-19; Estatísticas na prática; Análise de sobrevivência; Análise de séries temporais.*

References

AITCHISON, J. The statistical analysis of compositional data. Chapman and Hall, 1986, 416p.

AITCHISON, J.; BROWN, J. A. C. The lognormal distribution, Cambridge, England: Cambridge University Press, 1957, 176p.

AMARAL, M. T. R., CONCEIÇÃO, K. S.; ANDRADE, M. G.; PADOVANI, C. R.Modelo de curva de crescimento generalizado para Covid-19 nos estados brasileiros. *Rev. Bras. Biom.*, Lavras, v.38, p.125-146, 2020.

BROOMHEAD, D. S.; KING, G. P. Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, v.20, p.217-236, 1986.

DANILOV, D. Principal components in time series forecast. *Journal of Computational and Graphical Statistics*, v.6, p.112-121, 1997.

DEGROOT, M. H.; SCHERVISH, M. J. Probability and Statistics, 4th ed., Boston, MA, USA: Pearson Education, Inc, 2011, 912p.

GOLYANDINA, N.; NEKRUKIN, V.; ZHIGLJAVSKY, A. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & HallCRC, 2001. 320p.

GOLYANDINA, N.; ZHIGLJAVSKY, A. *Singular Spectrum Analysis for Time Series*. Springer Science and Business Media, 2013. 128p.

HASSANI, H.; MAHMOUDVAND, R. *Singular Spectrum Analysis Using R*. Palgrave Advanced Texts in Econometrics, 2018. 149p.

MAHMOUDVAND, R.; KONSTANTINIDES, D.; RODRIGUES, P. C. Forecasting mortality rate by multivariate singular spectrum analysis. *Applied Stochastic Models in Business and Industry*, v.33, p.717-732, 2017.

MAHMOUDVAND, R.; RODRIGUES, P. C. A new parsimonious recurrent forecasting model in singular spectrum analysis. *Journal of Forecasting*, v.37, p.191-200, 2018.

RINNE, H. The Weibull distribution: A handbook, Boca Raton, FL, USA: Chapman & Hall/CRC, 2009, 761p.

RODRIGUES, P. C.; MAHMOUDVAND, R. The benefits of multivariate singular spectrum analysis over the univariate version. *Journal of the Franklin Institute*, v.355, p.544-564, 2018.

RODRIGUES, P. C.; PIMENTEL, J.; MESSALA, P.; KAZEMI, M. Decomposition and forecasting of mutual investment funds using singular spectral analysis. *Entropy*, v.22, n.1, e22010083, 2020.

WALCK, C. Hand-book on statistical distributions for experimentalists, Stockholm, Sweden: University of Stockholm, 2007. 202p.

Received on 04.09.2020.

Approved after revised on 28.01.2021.

Appendix A

The Aitchison's distance (AITCHISON, 1986) between two positive n-component vectors, $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})$ and $\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n})$, is defined as the standard deviation of the n-component vector with the natural logarithm of the ratio between the components of $\mathbf{X}$ and $\mathbf{Y}$. Let $\mathbf{X}$ and $\mathbf{Y}$ two positive n-components real vectors. Define the vector composed by the natural logarithm of the ratios of the components of $\mathbf{X}$ and $\mathbf{Y}$:

$$Ln\left(\frac{\mathbf{X}}{\mathbf{Y}}\right) = \left(Ln\left(\frac{\mathbf{x_1}}{\mathbf{y_1}}\right), \ldots, Ln\left(\frac{\mathbf{x_n}}{\mathbf{y_n}}\right)\right) \tag{1}$$

Consider now the mean and the standard deviation of the components of $Ln\left(\frac{\mathbf{X}}{\mathbf{Y}}\right)$,

$$E = \frac{1}{n}\sum_{i=1}^{n} Ln\left(\frac{\mathbf{x_i}}{\mathbf{y_i}}\right) \quad \text{and} \quad \alpha = \sqrt{\sum_{i=1}^{n} \frac{1}{n}\left(Ln\left(\frac{\mathbf{x_i}}{\mathbf{y_i}}\right) - E\right)^2} \tag{2}$$

The Aitchison's compositional distance between $\mathbf{X}$ and $\mathbf{Y}$ is the defined as $AD(\mathbf{X}, \mathbf{Y}) = \alpha$. Note that the compositional distance is independent of the sum of the components. For simplicity one can normalize the vectors by dividing the components of the vectors by their sums without changing the value of α.