EDITORIAL: SPECIAL ISSUE ON BIOSTATISTICS AND BIOMETRY IN THE ERA OF DATA SCIENCE

Paulo Canas RODRIGUES[1]

Luiz Ricardo NAKAMURA[2]

Carlos Alberto de Bragança PEREIRA[3]

This special issue of the Brazilian Journal of Biometrics (BJB) contains 16 papers. The main topic of the special issue is "Biostatistics and Biometry in the Era of Data Science", and it is the result of a collaboration between the BJB and the Brazilian Region of the International Biometric Society (RBras) in the sequence of the cancelled 2020 RBras annual meeting due to the COVID-19 pandemic.

RBras is a scientific society founded in 1955, of a cultural character, non-profit, dedicated to Brazilian researchers who work with the mathematical and statistical aspects of agronomy, biological sciences and related areas. According to its statute, RBras seeks to stimulate the research activities of its members, encouraging and supporting scientific events with the following objectives: (i) to bring together researchers in the area of applied statistics, biometrics and biostatistics; (ii) to promote scientific events to discuss research topics in applied statistics; (iii) to publicize the methods developed in the area of statistics; and (iv) to offer courses on new topics in applied statistics. RBras belongs to The International Biometric Society, made up of several other regions and networks, and involving researchers in the area of Biometrics from all over the world.

The history of The Brazilian Journal of Biometrics began at the São Paulo State University "Júlio de Mesquita Filho" (UNESP), Jaboticabal Campus, where it was launched in 1983 under the name "Revista de Matemática e Estatística". In 2007, it was renamed "Revista Brasileira de Biometria" (RBB) with the sole purpose of promoting the development and application of statistical methods to problems in

[1]Universidade Federal da Bahia - UFBA, Departamento de Estatística, CEP: 40170-110, Salvador, BA, Brasil. E-mail: *paulocanas@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 de São Paulo - USP, Departamento de Estatística, CEP: 05508-010, São Paulo, SP, Brasil. E-mail: *cadebp@gmail.com*

different areas of biological sciences. As of January 2016, the journal started to be published by the Editora UFLA, from the Federal University of Lavras, MG, under the editorial responsibility of the Department of Statistics of the Federal University of Lavras. With more than thirty years of uninterrupted publication, currently the BJB is fully online, with a new issue every three months, without publication fees, indexed on several international databases, and all articles have an associated digital object identifier (DOI). Currently, the BJB is undergo some major changes aiming at having all articles published in English, with international high-profile editorial board members, and applications to impact factor indicators in JCR and Scopus on the way.

This special issue aims at helping in that transition. We were very pleased with the outstanding outcome in terms of the number and quality of the submitted papers, which exceeded all our expectations. There are 16 papers in this special issue that cover a number of the topics related to "'Biostatistics and Biometry in the Era of Data Science" and another relevant areas.

Nomelini et al. (2021) highlight the statistical process control charts that are used to illustrate the repeatability and reproducibility of the techniques. This work focused on an application for two sources of variation: patched aluminum beams and types of damage. It was shown that the control charts helped in the pre-processing step and in the detection of measurement errors.

Calcagnoto et al. (2021) deal with the identification and understanding of the causes that generate absenteeism from work, in a transport company from the north of state of Paraná, in order to prevent possible financial losses for the company. The authors use hierarchical cluster analysis to group the employees accordingly to the considered variables. Principal component analysis was used to access the signal of the contribution of the factors to the absence. Three main homogeneous clusters were obtained, and the potential factors that cause absenteeism in each cluster were identified.

De Pauli et al. (2021) propose a Bayesian adaptation of the multilayer perceptron artificial neural network, considering Markov chain Monte Carlo. They assumed a priori distribution for the model weights. A simulated data set was considered to access the convergence of the chains and to understand the learning process and the estimation. A real data application was presented, where a historical daily series of WTI oil price, between January 1, 2015 and December 16, 2019, was considered.

Luiz and Lima (2021) present the non-parametric Kolmogorov-Sminov (KS) test as an alternative to compare the effect of flooded irrigation management on methane emission througout the rice crop cycle. They verified that, while the classical analysis of variance was unable to detect differences between the irrigation management types in their application, the KS test showed to be an interesting and powerful tool to compare such characteristics over time.

Ióca and Zuanetti (2021) present several methodologies for variable selection, such as random forest, LASSO and the stepwise method, and apply and compare them to genetic data to select single nucleotide polymorphism (SNP) markers that characterize the presence or absence of a disease. The application and comparison was done by using the Genetic Analysis Workshop 17 database, that was built from simulated and real data for 697 unrelated individuals, 327 men and 370 women. The simulation included 24,487 SNPs divided in 22 chromosomes, and was made for a common and complex disease that has a prevalence of 30% in the population. The authors showed that the random forest and LASSO show similar prediction performance, but none of them correctly select the relevant SNPs, which might be due to the fact that most of the SNPs present very low variability in the considered sample.

Ribeiro et al. (2021) propose a new covariance structure with an exponential correlation function for bivariate random fields, called the simpler exponential covariance (SEC) model. A simulation study is presented to illustrate some features of the proposed model. The potentiality of the model is illustrated by means of a real weather data set from Brazil, where even though the SEC model has a simple structure, it can perform as good as the known bivariate separable Matérn model and the bivariate Matérn model with constraints.

From the point of view of applied statistics, Cardoso et al. (2021) present a study on traffic accidents with coalition. The objective was to assist technicians in the transport department of the city of Londrina in ordering the different types of vehicle coalition, since these accidents are responsible for more than 50% of the total accidents and among these 60% were with fatal victims.

Silva et al. (2021a) start their paper highlighting the great expansion of electronic sports (eSports) in the last few years. Then, they consider the well-known League of Legends (LoL) game, comparing the performance of the four top jungler champions (characters in the game) through their win rate, performing an analysis of variance followed by the Tukey's test, identifying a specific champion who outstands all others.

Silva et al. (2021b) consider the nonlinear Logistic, Gompertz and von Bertalanffy regression models to describe the growth in wood volume of *Eucalyptus urophylla x Eucalyptus grandis* hybrids in three forest site categories. The results show that for all site categories, the Gompertz model with the addition of autoregressive parameters AR(1) is the most appropriate model to describe the growth in wood volume of *Eucalyptus urophylla x Eucalyptus grandis* hybrids.

Batista et al. (2021) present an univariate and multivariate temporal analysis of the series of returns of Petrobrás and Ibovespa in two periods, the fist between 2005 and 2015, and the second between 2015 and 2019, the prior and posterior period of the beginning of the corruption crisis at Petrobras. The data was modelled with an ARIMA model, and its square residuals were modelled with a GARCH model. The authors concluded that, in the analysed lag period, there is little influence of

Petrobras shares in the Bovespa index.

Pereira et al. (2021) discuss the role of statisticians in the field of data analysis, and analyse Brazilian (per state) and Italian deaths by COVID-19, between March 17, 2020 and August 20, 2020, and between February 20, 2020 and August 20, 2020, respectively. They consider survival analysis and singular spectrum analysis for model fit and model forecasting.

Araujo et al. (2021) aim at identifying, by spectral analysis, the groups of patterns of points in space. They considered that the selected points were generated in a stochastic manner. Emphasis was given to the relationship between the structure of the periodograms and the type of stochastic process that generates the points. They also shown that the spectral analysis methods developed can identify the patterns of three-point processes and improve the characterization of the spatial points patterns.

Oliveira et al. (2021) suggest the use of discriminant analysis to define functions to separate different groups of *Coffea canephora*, namely varietal groups (Conilon and Robusta) and hybrids obtained from these varieties, establishing the most relevant phenotypic traits in these functions in order to classify future new individuals. Results show that the most important features are vegetative vigor, evaluation of the incidence of coffee rust, incidence of cercosporiosis, plant height and diameter of the canopy projection.

Santos and Bazán (2021) consider a Bayesian alternative to estimate the parameters of the Rasch Poisson count models, which are commonly used to identify individual latent traits, through the integrated nested Laplace approximation (INLA). Furthermore, they propose the use of the well known randomized quantile residuals in such models. Finally, they present a real data application, where they show the importance of performing proper residual analysis to identify potential problems of the fitted model.

Neisse et al. (2021) deal with chronic fatigue, that is related to generalized pain - fibromyalgia for instance - and often causes depression and absenteeism. Since LDL, evil cholesterol, and triglycerides are possible indicators of the level of generalized pain, this study, in the search for efficient indicators, made use of three regression methods: Lasso, Elastic-net and Stepwise. Stepwise selects the independent variables that effectively influence the levels of pain and help to define the regression model. Lasso and Elastic net show the level of influence of LDL. The elastic-net model suggests in addition an effect of Potassium. This paper illustrates how important Statistics is for medical doctors and physiotherapists.

Saraiva et al. (2021) deals with COVID-19 growth modelling, where the authors discuss a possible modelling strategy to try to describe and illustrate the type of growth of the group of infected and / or dead. In addition, it is of interest for health workers to prepare the environment for the next day. A combination of models was used here so that the types of growth at each stage of the process can

be monitored. The piecewise exponential estimator (PEXE)can be understood as the continuous alternative of the renowned Kaplan-Meier estimator, the latter has a discrete distribution as an estimator of the survival function. Although the work estimates the distribution function, PEXE motivated the presented solution. The appendix of the article discusses the distributions used and how they are used in composition.

It is our understanding that this Special Issue contributes to increasing knowledge in the fields of biostatistics, biometry and data science, by fostering discussions in real data problems.

We would like to finalise this editorial by thanking the authors for their interesting work and the reviewers for their important feedback that helped improving the quality of this special issue. We extend our acknowledgement to the Editors and Associate Editors that have been contributing for the development of the Brazilian Journal of Biometrics over the years.

References

ARAUJO, E. S. B.; SCALON, J. D.; BATISTA, L. S. Exploratory spectral analysis in three-dimensional spatial point patterns. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.177-192, 2021.

BATISTA, M. E. O., GOMES, R. S., GONÇALVES, L. R. Univariate and multivariate analysis of the Bovespa and Petrobras indices between 2005-2015. *Rev. Bras. Biom.* Lavras, v.39, n.1, p.139-157, 2021.

CALCAGNOTO, L. R.; SANTANA, T. V. F.; PESCIM, R. R. Classification and identification of the causes of absenteeism in a public transport company using cluster analysis and principal components techniques. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.25-44, 2021.

CARDOSO, M. G.; MARTINS, R. M.; STURION, L. A statistical study of traffic accidents attended by the Londrina-PR corporation in 2019. *Rev. Bras. Biom.* Lavras, v.39, n.1, p.103-113, 2021.

DE PAULI, S. T. Z.; KLEYNA, M.; BONAT, W. H. Multilayer perceptron artificial neural networks: an approach for learning through the bayesian framework. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.45-59, 2021.

IÓCA, M. P.; ZUANETTI, D. A. Selection of snp markers: analyzing gaw17 data using different methodologies. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.71-88, 2021.

LUIZ, A. J. B., LIMA, M. A. Application of the kolmogorov-smirnov test to compare greenhouse gas emissions over time. *Rev. Bras. Biom.* Lavras, v.39, n.1, p.60-70, 2021.

NEISSE, A. C.; OLIVEIRA, F. L. P.; OLIVEIRA, A. C. S.; CRUZ, F. R. B.; NETO, R. M. N. Chronic fatigue syndrome and its relation with absenteeism: elastic-net and stepwise applied to biochemical and anthropometric clinical measurements. *Rev. Bras. Biom.* Lavras, v.39, n.1, p.221-239, 2021.

NOMELINI, Q. S. S., SILVA, J. W., GALLO, C. A., FINZI NETO, R. M., MOURA JUNIOR, J. R. V. Statistical process control (spc) of damage metrics in the impedance-based structural health monitoring. *Rev. Bras. Biom.* Lavras, v.39, n.1, p.7-24, 2021.

OLIVEIRA, G. F., MIRANDA, T. L. R., NASCIMENTO, A. C. C., NASCIMENTO, M., CAIXETA, E. T., SILVA, L. F., ALKIMIM, E. R., SILVA, F. L. Discrimination of varietal groups and hybrids of coffea canephora species using multivariate analysis. *Rev. Bras. Biom.* Lavras, v.39, n.1, p.194-205, 2021.

PEREIRA, C. A. B.; NAKAMURA, L. R.; RODRIGUES, P. C. Naive statistical analyses for covid-19: application to data from Brazil and Italy. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.158-176, 2021.

RIBEIRO, A. M.; RIBEIRO JÚNIOR, P. J.; BONAT, W. H. Comparison of exponential covariance functions for bivariate geostatistical data. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.89-102, 2021.

SANTOS, N. C. A; BAZÁN, J. L. Residual analysis in rasch poisson counts models. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.206-220, 2021.

SARAIVA, E. F.; SAUER, L.; PEREIRA, B. B.;PEREIRA, C. A. B. A piecewise growth model for modeling the accumulated number of covid-19 cases in the city of Campo Grande. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.240-265, 2021.

SILVA, V. F.; SILVA, E. M.; LIMA, K. P.; SCALON, J . D. Performance of jungler's champions in the game league of legends. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.114-121, 2021a.

SILVA, W. S.; FERNANDES, F. A.; MUNIZ, F. R.; MUNIZ, J. A.; FERNANDES, T. J. Eucalyptus grandis x eucalyptus urophylla growth curve in different site classifications, considering residual autocorrelation. *Rev. Bras. Biom.*, Lavras, v.39, n.1, p.122-138, 2021b.