ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

REVSTAT
Statistical Journal

Special issue
«Celebrating the World Statistics Day»

Volume 18, No.5
October 2020

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Trimestral. - Continuação de : Revista de Estatística = ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726 ; e-ISSN 2183-0371

EDITORIAL

#StatsDay2020 — The WORLD STATISTICS DAY, 20 October 2020

This year we celebrate on 20 October 2020 the third World Statistics Day (WSD).

The first WSD was celebrated on 20 October 2010 with the topic "Service, professionalism, integrity celebrating the many contributions and achievements of official statistics". Later on, in June 2015, the United Nations (UN) General Assembly adopted a resolution, in which 20 October 2015 was designated as the second WSD, with the theme "Better data, better lives"; moreover, it was decided to celebrate the WSD every five years on 20 October.

The WSD 2020 was launched by the UN Statistical Commission at the 51st Session, and will be commemorated around the world under the theme "Connecting the world with data we can trust", giving head to the importance of trust on reliable data in official statistics.

On the other hand, 2020 will also be a year in which COVID-19 pandemic has become an unavoidable topic. More than ever, scientists from all branches put efforts together in order to make out relevant information as much as possible from data labelled as reliable. Statistics play here a main role, no doubt.

In this Special Issue, in Celebration of the WORLD STATISTICS DAY 20 October 2020, *REVSTAT* offers to its readers a first Invited Paper with discussion, on the pandemic issue. We are grateful to all the authors involved in the main paper and subsequent discussants. Special thanks are due to our Past Editor Antónia Amaral-Turkman, a major engine for guiding the invitation process.

The other seven articles present original contributions, with data applications mainly in life sciences (*two treatments using matched pairs, leukaemia, diabetic retinopathy, bladder cancer, tree mortality, maximum time of breastfeeding, level of a biomedical marker, lung cancer rates, diagnostic tests on epithelial ovarian cancer*) and official statistics (*food security and poverty*).

It is our pleasure to invite our readers to enjoy this Special Issue of *REVSTAT – Statistical Journal*.

Happy World Statistics Day!

October 20, 2020

Isabel Fraga Alves
Giovani Loiola da Silva

INDEX

Abstracted/indexed in: *Current Index to Statistics, DOAJ, Google Scholar, Journal Citation Reports/Science Edition, Mathematical Reviews, Science Citation Index Expanded®, SCOPUS* and *Zentralblatt für Mathematic.*

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 553–564

STATISTICS IN TIMES OF PANDEMICS: THE ROLE OF STATISTICAL AND EPIDEMIOLOGICAL METHODS DURING THE COVID-19 EMERGENCY

(*Invited Paper with Discussion*)

Authors: Baltazar Nunes
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, and Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Portugal
baltazar.nunes@insa.min-saude.pt

Constantino Caetano
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal
constantino.caetano@insa-min-saude.pt

Liliana Antunes
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal
liliana.antunes@insa-min-saude.pt

Carlos Dias
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, and Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Portugal
carlos.dias@insa-min-saude.pt

Received: September 2020 Revised: October 2020 Accepted: October 2020

Abstract:

- Statistical and epidemiological methods play an essential role in producing information for the public health decision process. They allow the collection, analysis, reporting and interpretation of data necessary to inform public health officials in the decision-making process, enabling the diagnosis of the situation, the selection of the most adequate measures and to monitor and evaluate their impact.

Key-Words:

- *epidemiology; SARS-CoV-2; COVID-19; epidemiological surveillance; mathematical modelling.*

AMS Subject Classification:

- 92B15, 92D30.

1. INTRODUCTION

After the identification of the first cases of COVID-19 in December 2019 in Wuhan, China, it took one month (30 January 2020) for the World Health Organization to declare the epidemic a Public Health Emergency of Global concern, and almost two months to be declared a pandemic (11 March 2020) ([11]).

The start of the rampant community transmission of the SARS-CoV-2 virus in Italy and Spain in early February ([11]), which meant that most new cases did not have an epidemiological link, set a decisive date for public health action in Portugal. The first case of COVID-19 was diagnosed in Portugal on the 2nd of March 2020, and it was only a few weeks before community spread of the virus started in Portugal, which brought urgency to the implementation of population-base public health interventions.

Although the infection and disease occurs at the individual level, the epidemic occurs at the population level. This means that the control and mitigation of the infection or disease occurrence cannot be only achieved through individual targeted measures. Population level intervention is crucial to control the spread of the infection and to mitigate the impact of the disease. Among these, one can list vaccination, social distancing, respiratory hygiene, closing of schools, working from home, closing of commerce, restaurants and bars, or even more severe lockdowns, which involve, in addition to the previous measures, stay at home policies.

Public health authorities need information almost on a real time basis to be able to track the infection evolution among the population, the potential effect of interventions and their impact. The information provided can range from descriptive analysis of the distribution of the number of new cases of infection in time and space, population or individual characteristics (age, sex, education, health status), to projections of the infection spread and impact according to different epidemiological scenarios, or control and mitigation strategies.

Throughout an epidemic it is usual for certain geographical locations to be affected differently, probably due to factors that promote the spread of infection, such as population density and connectivity, the inbound and outbound of infected individuals from the region, amongst others. Which means that nowadays, the access to accurate and detailed information is paramount to mount proper public health measures, which can be targeted during a certain period of time, in a certain geographical area or population group.

In the context of a pandemic there are three areas where statistical and epidemiological methods are crucial, namely: epidemiological surveillance, specific epidemiological observational studies aimed at describing or measuring epidemiological parameters and the development of mathematical models, which focus on simulating the disease transmission dynamics and assess impact scenarios under different control and mitigation measures.

2. EPIDEMIOLOGICAL SURVEILLANCE

The epidemiological surveillance is defined as the systematic collection, analysis, interpretation and reporting of data for public health action. During the COVID-19 public health emergency, most of the epidemiological surveillance data comes from the case noti-

fication of SARS-CoV-2 infection. Individuals with a laboratory diagnosis of SARS-CoV-2 infection, with or without the presentation of COVID-19 symptoms at diagnosis, are notified and registered in a database. In Portugal, data is collected through the SINAVE system (Sistema Nacional de Vigilância Epidemiológica) ([28]), where medical doctors or laboratories, authorized to perform the SARS-CoV-2 test, report new cases of infection or disease on a daily basis. After the case notification, public health officials perform an epidemiological enquire where data is collected on individual characteristics, disease characteristics (symptoms), probable routes of infection, date of symptom onset, laboratory diagnosis date and notification date, travel history, and close contacts.

Among the first questions that arise, and that can be answered with epidemiological surveillance data, are about the distribution of cases in the time scale, namely: what is the recent incidence of the disease and its trend? Are the number of new cases of the disease increasing, decreasing or stable?

A simple but essential way to study the course of the epidemic is by plotting the epidemic curve. This graph basically depicts the number of new cases by date of disease onset.

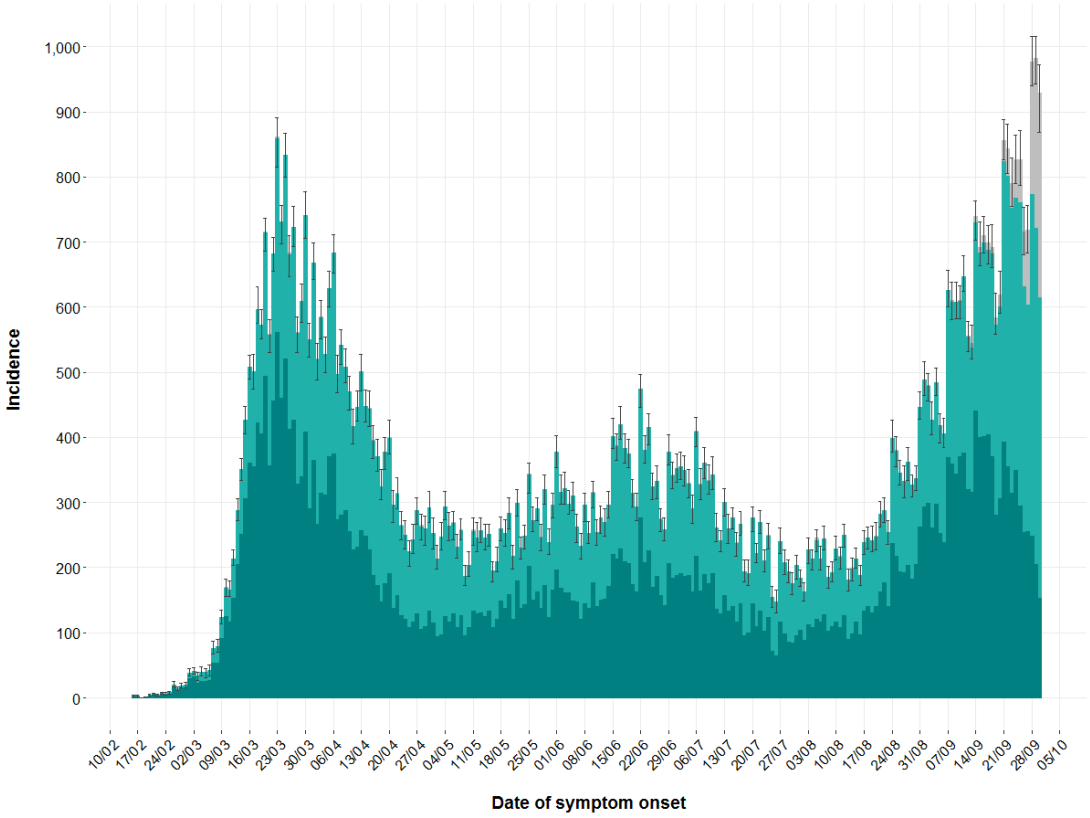


Figure 1: SARS-CoV-2 epidemic curve by date of symptom onset corrected for notification delay (Dark blue – Observed cases with date of symptom onset; Light blue – Observed cases with imputed date of symptom onset; Grey – Occurred but not yet reported cases (nowcasted)).

In actuality, the graph should be represented by date of infection, but for the majority of the cases this information is not available. The analysis of this figure is of utmost importance, given that allows the epidemiologist and public health officials to analyse the evolution of the

infection transmission, its trends, growing and decreasing phases, the impact of public health measures or changes in the epidemiology of the infection.

The classical presentation of an epidemic curve is divided in four phases, the establishment phase, with a sporadic number of new cases, the exponential growing phase, the peak and the decreasing phase. But in reality, mainly with pandemics like the one we are facing with SARS-CoV-2 virus, there is a diversity in the presentation of the epidemic curve. As Adam Kurcharski, from London School of Hygiene and Tropical Medicine, referred in his recent book, The Rules of Contagion ([22]), "If you have seen one pandemic, you have seen one pandemic.". In reality, the form of the epidemic curve can present several growing, decreasing and endemic phases, depending on the presence of several factors, including public health measures, population behaviour, population susceptibility, virus changes or climate effects.

The quality of the surveillance data used in the production and analysis of an epidemic curve in real time can be affected by several factors. Surveillance data is collected hastily, since the priority of medical personnel is to isolate the cases and quarantine their contacts, in order to be able to interrupt the chains of transmission. So, collecting perfect data in these circumstances is very difficult, and usually, data quality, completeness and timeliness are substandard during pandemic emergencies. This problem is further exacerbated during the epidemic growth and epidemic peak phases, which is when this information is mostly needed.

As mentioned above, in order to draw a proper and useful epidemic curve the date of disease onset for each case is essential. Unfortunately, during an epidemic this is one of the variables that suffers more from incompleteness. Moreover, there is the need to draw conclusions on the epidemic course for the last few days or weeks recorded. As can be easily understood, before an infected individual is accounted for in the surveillance system, several other events must occur, namely: disease onset, which is when the individual starts developing symptoms; medical visit, which is usually the time it takes an individual to seek health care after developing symptoms; diagnosis, when the infection has laboratory confirmation; and ultimately, notification, which is when the case is introduced in the surveillance system database.

Naturally, from disease onset to the notification of the case to the surveillance system, several days can go by. Not only due to the disease and reporting process described above but also due to other factors, such as the awareness of the population, preparedness of the surveillance system, overload of the professionals involved in the reporting process during the growth or peak phases, weekends when most medical sites are closed, personal and regional characteristics, among others ([15]).

The time between disease onset and case notification, precludes the monitoring of the epidemic in real time, translated in epidemic curves showing the last days/weeks always with fewer cases than the day before, suggesting a declining trend in incidence, when in fact, as more data is available, it might actually show it was increasing.

The notification delay varies across different surveillance systems because of the intrinsic nature of the reporting process implemented for each disease and country, and also due to changes in time. During the COVID-19 emergency, the time between disease onset to

notification delays have been described as 8 days on average, also 95% of the cases with onset of symptoms in one specific day take 15 days to be notified ([30, 12]). In Portugal, the median time between disease onset and notification decreased from 8 to 4 days throughout the course of the epidemic.

Here the statistical and epidemiological methods can be used to our advantage ([23]). Imputation methods can be used to estimate missing dates of disease onset, and nowcasting techniques can be applied to estimate the occurred-but-not-yet-reported cases in the last weeks of the epidemic curve ([23, 25, 19]). In order to be applied, both these methods need the distribution of the diagnosis and/or notification delay.

Several approaches have been proposed to impute date of symptom onset and to nowcast the recent days in public health setting ([15]). Nowcasting methods have also been developed in both frequentist and Bayesian framework ([19]).

There are several constraints that difficult the real time implementation of nowcast procedures such as the distribution of the delay from disease onset to notification not being invariant in time, place and individual characteristics. During the first weeks of the epidemic, the delay between disease onset and notification is usually longer, given that the first traced cases are usually detected at a later date. This means that several weeks of epidemic data are necessary in order to be able to estimate a complete distribution of the delay. So, for the first weeks of the epidemic, the delay distribution is truncated, and we need to recur to other sources of information, like historical data from other epidemics in order to estimate the notification delay distribution and nowcast the most recent days of the epidemic curve.

Besides the level of incidence and its distribution in time, public health officials usually need to know the level of transmission of the infection in the community. The reproduction number (R) of the infection is the measure used to monitor the transmission of infectious diseases. During the COVID-19 emergency it has been given a very important role in decision making, such as being used to implement or lift lockdown measures. This index measures the average number of secondary cases a typical infectious case gives origin to. When R is above one it means that the epidemic is in a growing phase, and each generation of infected individuals, gives rise to a new generation with a higher number of infected individuals. Otherwise if R is below one, the next generation of infected individuals is smaller than the previous one, which means that the size of the epidemic converges to 0, leading to its extinction. If R is equal to or close to 1, then the number of new infected individuals in each subsequent generation is approximately constant, which corresponds to an endemic or stable phase of disease spread.

It is important to estimate the reproduction number during the first phase of a pandemic, when all the population is susceptible and there are no control measures in place, in order to have a natural measure of the infection transmissibility in the population in study. This parameter is called R_0 ("R naught") and has a very crucial role in infectious disease epidemiology. It can be interpreted as the average number of infectious cases resulting from one infectious case, after its introduction in a completely susceptible population ([31]).

R_0 is used in mathematical modelling studies to set up the base scenario of an epidemic, allowing to find the level of public health measures needed to bring it below one. Or it can be used as reference to be compared with the effective reproduction number in order to evaluate

changes in the infection transmission, due to the implementation of public health measures or due to the natural course of the epidemic.

R_0 can be estimated from transmission chains data, which are usually difficult to obtain, measuring the distribution of the secondary cases each initial case gave origin to. It can also be estimated from the epidemic curve growth in the initial phase of the epidemic, with additional information on the generation time distribution of the infection ([31]).

The generation time of infection corresponds to the time elapsed between the times of infection of infector and infectee individuals. Given that it is difficult to collect data on infection dates, this distribution is substituted by the serial time distribution, which is given by the time interval between symptom onset of the infector and the symptom onset of the infectee. Regarding the COVID-19 infection, the serial time is being described with a mean from 3.1 to 7.5 according to the study published ([2]). This distribution plays also a very relevant role in the epidemic transmission, given that it represents the time elapsed between the disease onset in two generations of infected individuals. The shorter the serial time, the quicker the infection spread.

Besides the R_0, another reproduction number that has been recurrently used by several public health agencies and governments to monitor the SARS-CoV-2 transmission, is the time dependent effective reproduction number. Firstly proposed by Wallinga and Teunis in 2004 ([32]), to measure the average number of secondary cases originated by the incident infectious cases in time t, it has been showed to be an excellent tool to monitor transmission over time and evaluate in a real-time manner the impact of public health measures and changes in the infection epidemiology. Nevertheless, for surveillance proposes, where date-of-onset data is always incomplete in the last few recorded weeks, several researchers have recommended the use of Cori *et al.* method ([8]) to estimate the time dependent effective reproduction number. In this method the R_t represents the average number of secondary cases that originated the new cases observed at time t.

As stated by other researchers, today's challenge is how to include the imputation and nowcasting methods uncertainty into the estimate of R_t.

In Portugal the R_t has been estimated from the epidemic curve after imputation and nowcasting ([20]), using a method developed by Antunes *et al.* in 2014 ([3]) for daily mortality monitoring. The R_t has been estimated using the Cori *et al.* method ([8]). This analysis did not include a method for error propagation from the imputation and nowcasting to the estimate of the reproduction number.

3. SPECIAL STUDIES

Special studies are specifically designed to measure epidemiological parameters of infection or disease and to identify risk factors of infection or disease severity, including the need for critical care and case-fatality.

One of the most relevant studies that the WHO recommends, at the start of the emergency of a new infectious agent with pandemic potential, is the First Few X Cases study (FFX) ([29]). This study should be implemented as quickly as the first cases are identified,

and is aimed at estimating key epidemiological indicators that are essential both for surveillance and for modelling the transmission and impact of the disease. FFX study aims at describing the clinical presentation of the disease and the routes of infection, the secondary infection and clinical attack rate, the serial interval, proportion of asymptomatic infections, the basic reproduction number, the incubation period, and eventually preliminary estimates of infection and disease-severity ratios (case-severity and case-fatality ratios).

Early estimates of these parameters are essential to the design and selection of the more effective public health measures, however, during emergency situations, it might not be feasible to carry out such studies. During the COVID-19 emergency, Portugal included the FFX protocol in the list of studies to implement ([26]), but once the number of cases started to increase exponentially, its implementation was not possible. Since only a few countries were able to conduct this study ([5]), the majority of the estimates of key epidemiological parameters came from the secondary analysis of data from routine epidemiological surveillance systems.

According to the different phases of the pandemic, and with the increase in the number of cases, other studies can be implemented. Such as, cohort or case-control studies aimed at identifying risk or protective factors of infection, disease or disease severity ([26]). These studies are also crucial for the implementation of targeted public health measures, but also to feed the mathematical models that aim to produce impact scenarios in the health care sector.

Other special studies of high importance are population-based surveys. These are cross-sectional studies that aim to measure seroprevalence of antibodies against the infectious agent, or surveys focused on evaluating the knowledge about the infection and disease, the adoption of preventive measures (hand washing, social distancing or use of mask) or the impact of the universal public health measures, such as lockdown of specific areas or countries, on the socioeconomic indicators or even on mental health. A very specific survey that some countries have implemented before and during the COVID-19 emergency, aims at measuring the profile of the contacts between individuals, according to their age and setting (home, school, workplace and general community). These are very important to feed the mathematical models in order to include age heterogeneity in the contacts between individuals and the observed impact of social distancing measures on these contact matrices.

Finally, it is important to mention those studies dedicated at measuring the effectiveness or impact of the public health measures, vaccines or treatments. These are essential to adjust the set of intervention and treatments. During the COVID-19 emergency several studies were performed to evaluate the impact of the lockdown measures, using quasi-experimental designs like interrupted time series analysis. Moreover, complex randomized clinical trials aimed at identifying effective treatments against COVID-19 disease and their complications were also implemented.

In all these studies the role of the statistical methods is undeniable in the several phases of the study development. These encompass the design, field implementation and data validation, analysis, reporting and interpretation of results. One of the lessons learned during the COVID-19 emergency is that the development of these studies needs generic scientific protocols developed and approved during inter-pandemic phases, and most importantly, dedicated teams for their implementation.

4. MATHEMATICAL MODELS OF DISEASE TRANSMISSION

Epidemic models for the spread of infectious diseases date back to the beginning of the twentieth century. Pioneer mathematical techniques were proposed to describe the dynamics of disease transmission and these are key tools to lay out proper mitigation and suppression measures to deal with an epidemic ([13]). The susceptible-infected-recovered (SIR) models developed by Kermack and McKendrick (1927) were the first mathematical models developed to the transmission dynamics of infectious diseases. These compartmental models have been further extended and adapted to an assortment of different diseases ([31]). Nowadays, these models are seldom used in their original form, since they are too simplistic to account for the inherent heterogeneity of disease transmission.

The SIR and SEIR (Susceptible, Exposed, Infectious and Removed) models have been improved to account for several crucial factors, such as the heterogeneity of human contact ([16, 31]), geographical distribution ([4]) and disease susceptibility, as well as accounting for the underlying uncertainty in disease transmission, i.e. stochastic models ([35, 34]). These extended models have been developed for different goals:

- **a**) assessing the epidemic preparedness of health systems ([27, 7]), by taking into consideration the susceptibility and contact pattern of individuals, as well as the infectivity of the pathogen;
- **b**) evaluate the impact of intervention measures that aim to reduce mortality and healthcare demand during an epidemic ([13]);
- **c**) account for the geographical distribution of the population, i.e. metapopulation models ([4, 33]);
- **d**) study the seasonality of disease prevalence ([10]);
- **e**) evaluate the necessary herd immunity vaccination threshold ([17]).

Since the beginning of the SARS-CoV-2 epidemic, a number of different working groups have employed SIR-like deterministic and stochastic models to evaluate the spread of the disease. The authors in [33] used a deterministic SEIR model to nowcast and forecast the national spread of SARS-CoV-2 in China, by creating scenarios for the transmissibility reduction and mobility reduction associated with the measures implemented in Wuhan, China. A SEIR-type stochastic model was developed by the authors of [35] to estimate key latent epidemiological parameters and states, such as the proportion of asymptomatic individuals, and the strength of the contact tracing. The authors in [9] employed the use of a healthy-asymptomatic-sick-dead model to assess the relaxation of social-distancing measures in Germany, a similar approach was also employed by the authors in [24] for Spain.

One of the biggest challenges that epidemiologists and modellers face during this epidemic is to foresee the duration of different epidemic phases. The first phase consisted in the containment of imported cases and identification of transmission chains, after which, an exponential growth of incident cases is expected, hence, the next phase consisted in the introduction of suppression measures by enforcing strict social-distancing. Lately, several countries having been phasing-out social-distancing measures, which might result in higher disease transmission during next winter along with the transmission of other seasonal respira-

tory viruses. For each of these phases several measures are necessary to prevent, control and mitigate COVID-19's impact, which need to be evaluated and simulated with best available modelling tools.

Other very important discussion for which mathematical modelling has provided very important insights and discussion is the future of the pandemic and its end, such as the herd immunity threshold, with estimates ranging from around 10% to 70% ([14, 6, 1]). Furthermore, should we expect that SARS-CoV-2 will be present for several years and became a seasonal respiratory virus ([21]) together with influenza and other respiratory virus ([14, 6, 1])?

5. FINAL REMARKS

The emergence of SARS-CoV-2 and the COVID-19 pandemic was really a black swan event, although international and national institutions have been preparing for new influenza pandemic, in truth, the majority of the countries were not prepared for all the implications of this event.

Data and information about the epidemic course, the risk factors for infection and disease, the effectiveness of public health measures and treatments, and the future scenarios of the course of the pandemic are among the most wanted pieces of knowledge by all the sectors of society.

These can be obtained through the development of surveillance systems, special epidemiological studies and mathematical modelling, areas that must be developed together in comprehensive and timely manner. Data from surveillance systems is important to feed special studies and mathematical modelling. Special studies allow the estimation of relevant parameters to feed the mathematical models and to estimate transmission parameters in real time (R_t). The public health measures that are evaluated prospectively through mathematical modelled scenarios, can be retrospectively evaluated using surveillance systems data or special studies.

Epidemiologists, statisticians and mathematical modellers have been probably among the most needed professionals during this phase, due to the amounts of data and information that needs to be collected, analysed, reported and interpreted. Data and information is demanded by decision makers with higher levels of certainty and lowest timeliness.

On the side of those that must make decisions, the information and knowledge that public health officials and decision-makers (at the highest level) receive in real time to decide is overwhelming. During this emergency, decision makers needed to deal with complex epidemiological concepts, with high levels of uncertainty. This placed very high pressure in public institutions responsible for delivering the information needed for decision, but also on decision makers and public health authorities that are confronted with uncertain information and are asked to make correct decisions ([18]).

A reflection should be made about several of these issues in the light of the COVID-19 pandemic experience, from the point of view of data and information. These reflections should include the anticipation of data, studies and information needed during the different phases of

a pandemic; methods and models available for analysis in order to develop a public available data analysis toolkit and clear rules for data disclosure and availability. The collaboration between public health institutes and the science academies should also be enhanced for the development of these three areas. Finally, efforts should be allocated in the translation of data and information for decision makers in order to contribute to more informed decisions, with a focus on the population health and welfare.

REFERENCES

[1] Aguas, R.; Corder, R.M.; King, J.G.; Gonçalves, G.; Ferreira, U.M. and Gomes, M.G.M. (2020). Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics, *medRxiv*, https://doi.org/10.1101/2020.07.23.20160762.

[2] Ali, S.T.; Li, W.; Lau, E.H.Y.; Xu, X.-K.; Du, Z.; Wu, Y.; Leung, G. and Colwing, B.J. (2020). Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions, *Science*, **369**(6507), 1106–1109.

[3] Antunes, L.; Antunes, M. and Nunes, B. (2014). *Modelação do atraso na notificação dos óbitos no sistema de vigilância diária da mortalidade*, Master's thesis, FCUL, Lisbon, Portugal.

[4] Apolloni, A.; Poletto, C.; Ramasco, J.J.; Jensen, P. and Colizza, V. (2014). Metapopulation epidemic models with heterogeneous mixing and travel behaviour, *Theoretical Biology and Medical Modelling*, **11**(1), 3.

[5] Boddington, N. (2020). COVID-19 in Great Britain: epidemiological and clinical characteristics of the first few hundred (FF100) cases: a descriptive case series and case control analysis, *medRxiv*, https://doi.org/10.1101/2020.05.18.20086157.

[6] Britton, T.; Ball, F. and Trapman, P. (2020). A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2, *Science*, **360**(6505), 846–849.

[7] Chitnis, N.; Hyman, J.M. and Valle, S.Y.D. (2017). Mathematical models of contact patterns between age groups for predicting the spread of infectious diseases, *Mathematical Biosciences and Engineering*, **15**(5), 1475–1497.

[8] Cori, A.; Ferguson, N.; Fraser, C. and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics, *American Journal of Epidemiology*, **178**(9), 1505–1512.

[9] Donsimoni, J.R.; Glawion, R.; Plachter, B.; Weiser, C. and Wälde, K. (2020). Should contact bans be lifted in Germany? A quantitative prediction of its effects, *medRxiv*, https://doi.org/10.1101/2020.04.10.20060301.

[10] Dorélien, A.M.; Ballesteros, S. and Grenfell, B.T. (2013). Impact of birth seasonality on dynamics of acute immunizing infections in Sub-Saharan Africa, *PLoS ONE*, **8**, 10.

[11] ECDC: Event Background-COVID-19, https://www.ecdc.europa.eu/en/novel-coronavirus/event-background-2019.

[12] Estimation of the current development of the SARS-CoV-2 epidemic in Germany – nowcasting Epidemiologisches Bulletin. 17/2020 Robert Koch Institute. https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2020/Ausgaben/17_20.pdf.

[13] Ferguson, N.M.; Laydon, D.; Nedjati-Gilani, G.; Imai, N.; Ainslie, K.; Baguelin, M.; Bhatia, S.; Boonyasiri, A.; Cucunubá, Z.; Cuomo-Dannenburg, G.; Dighe, A.; Dorigatti, I.; Fu, H.; Gaythorpe, K.; Green, W.; Hamlet, A.; Hinsley, W.; Okell, L.C.; van Elsland, S.; Thompson, H.; Verity, R.; Volz, E.; Wang, H.; Wang, Y.; Walker, P.G.T.; Walters, C.; Winskill, P.; Whittaker, C.; Donnelly, C.A.;

Riley, S. and Ghani, A.C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand, *Imperial College of London*, Report 9, https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/.

[14] Gomes, M.G.M.; Corder, G.; King, J.; Langwig, K.E.; Souto-Maior, C.; Carneiro, J.; Gonçalves, G.; Penha-Gonçalves, C.; Ferreira, M.U. and Aguas, R. (2020). Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold, *medRxiv*, https://doi.org/10.1101/2020.04.27.20081893.

[15] Held, L.; Hens, N.; O'Neill, P. and Wallinga, J. (2019). *Handbook of Infectious Disease Data Analysis*, Chapman and Hall/CRC, New York.

[16] Hens, N. (2012). *Modeling Infectious Disease Parameters Based On Serological and Social Contact Data: A Modern Statistical Perspective*, Springer, New York.

[17] Hethcote, H.W. (2013). The mathematics of infectious diseases, *SIAM Review*, **42**(4), 599–653, https://doi.org/10.1137/s0036144500371907.

[18] Hilderink, H.B.M. (2020). The corona crisis and the need for public health foresight studies, *European Journal of Public Health*, **30**(4), 616.

[19] Höhle, M. and van der Heiden, M. (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, *Biometrics*, **70**, 993–1002.

[20] INSA: Covid-19 curva epidémica e parâmetros de transmissibilidade, http://www.insa.min-saude.pt/category/areas-de-atuacao/epidemiologia/covid-19-curva-epidemica-e-parametros-de-transmissibilidade/.

[21] Kissler, S.; Tedijanto, C.; Goldstein, E.; Grad, Y.H. and Lipsitch, M. (2020). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period, *Science*, **368**(6493), 860–868.

[22] Kurcharki, A. (2020). *The Rules of Contagion: Why Things Spread – and Why They Stop*, Profile Books Ltd, London.

[23] Lipsitch, M.; Finelli, L.; Heffernan, R.T.; Leung, G.M. andRedd, S.C. (2011). Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1, *Biosecur. Bioterror.*, **9**, 89–115.

[24] Lopez, L. and Rodo, X. (2020). The end of the social confinement in Spain and the COVID-19 re-emergence risk, *medRxiv*, https://doi.org/10.1101/2020.04.14.20064766.

[25] Nunes, B.; Natário I. and Carvalho, M.L. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models, *Stat. Med.*, **32**(15), 2643–2660.

[26] Plano Nacional de Preparação e Resposta à Doença por novo coronavírus (COVID-19), https://www.dgs.pt/documentos-e-publicacoes/plano-nacional-de-preparacao-e-resposta-para-a-doenca-por-novo-coronavirus-covid-19-pdf.aspx.

[27] Saunders-Hastings, P.; Hayes, B.Q.; Smith, R. and Krewski, D. (2017). National assessment of Canadian pandemic preparedness: employing InFluNet to identify high-risk areas for inter-wave vaccine distribution, *Infectious Disease Modelling*, **2**(3), 341–352.

[28] SINAVE: Sistema Nacional de Vigilância Epidemiológica, https://www.dgs.pt/servicos-on-line1/sinave-sistema-nacional-de-vigilancia-epidemiologica.aspx.

[29] The First Few X (FFX) Cases and contact investigation protocol for 2019-novel coronavirus (2019-nCoV) infection, https://www.who.int/publications/i/item/the-first-few-x-(ffx)-cases-and-contact-investigation-protocol-for-2019-novel-coronavirus-(2019-ncov)-infection.

[30] Tsang, T.K.; Wu, P.; Lin, Y.; Lau, E.H.Y.; Leung, G.M. and Cowling, B.J. (2020). Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study, *Lancet Public Health*, **5**, e289–e296.

[31] Vynnycky, E. and White, R.G. (2019). *An Introduction to Infectious Disease Modelling*, Oxford Univ. Press, Oxford.

[32] Walling, J. and Teunis, P. (2014). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures, *American Journal of Epidemiology*, **160**(6), 509–516.

[33] Wu, J.T.; Leung, K. and Leung, G.M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study, *The Lancet*, **395**(10225), 689–697.

[34] Yan, P. and Chowell, G. (2019). *Quantitative Methods for Investigating Infectious Disease Outbreaks*, Springer, Cham.

[35] Zhang, Y.; You, C.; Cai, Z.; Sun, J.; Hu, W. and Zhou, X.H. (2020). Prediction of the COVID-19 outbreak based on a realistic stochastic model, *medRxiv*, https://doi.org/10.1101/2020.03.10.20033803.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 565–574

ENDEMIC-EPIDEMIC FRAMEWORK USED IN COVID-19 MODELLING

(*Discussion on the paper by Nunes, Caetano, Antunes and Dias*)

Authors: M. Bekker-Nielsen Dunbar
– Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland
maria.dunbar@uzh.ch

L. Held
– Epidemiology, Biostatistics and Prevention Institute, and Center for Reproducible Science, University of Zurich, Switzerland
leonhard.held@uzh.ch

Abstract:

- Nunes *et al.* ([54]) provide an overview of mathematical models used to analyse epidemics and techniques for conducting studies to obtain parameter estimates for such models. They discuss the SEIR model which has been used in much coronavirus disease 2019 (COVID-19) analysis. Our discussion presents a modelling framework based in time series analysis developed for the analysis of infectious disease surveillance data, as well as our use of the framework in analysing COVID-19. We believe many of the purposes of modelling infectious disease outlined by Nunes *et al.* ([54]) as well as the benefits of mathematical modelling highlighted can also be found in the statistical modelling techniques we use in our work.

1. ENDEMIC-EPIDEMIC MODELLING FRAMEWORK

Multiple epidemic data sources provide valuable information on different aspects of an infectious disease outbreak ([19]). Indeed, a recent simulation study by Colón-González *et al.* [18] indicates that use of multiple data streams arising from such surveillance activities can be a useful approach to disease detection. However, it is pertinent that appropriate statistical techniques be used in analysing such data sources to incorporate the associated uncertainties to avoid introducing bias and artificial precision in estimates of disease outcomes, impacts of disease control interventions, and real-time predictions ([9]). One such statistical technique is the endemic-epidemic (EE) modelling framework. The EE framework is a multivariate time series model created for analysis of infectious disease surveillance data ([30]). The simplest EE model is a spatio-temporal multivariate time-series model of disease incidence from surveillance data. The model additively decomposes incidence into endemic and epidemic components. The endemic component covers exogenous factors such as seasonality, sociodemography, and population while the epidemic component is autoregressive and is driven by previous case counts ("infectiousness"), i.e. the force of infection. We will discuss this modelling approach and its applicability to the current COVID-19 setting.

1.1. Applications

The EE framework has been applied to a multitude of infectious diseases classified as various types, e.g. diseases with other reservoirs than just humans, vaccine-preventable diseases, and vector-borne diseases, showcasing its versatility. See Table 1 for an overview of diseases analysed using the method. COVID-19 and SFTS are both currently considered emerging infectious diseases, showcasing the EE framework's flexibility and ability to consider both novel and established diseases. Since its introduction, the EE framework has been extended to cover many different aspects of disease modelling and statistical analysis ([58, 57, 33, 47, 32, 48, 62]). Recent extensions include the possibility to estimate the serial interval distribution ([13]) and methodology to adjust for underreporting ([12]).

Table 1: Applications of the EE framework.

Disease	Reference(s)
Endemic porcine diseases	[3]
Leishmaniasis	[1, 52]
Dengue	[16, 75]
Invasive pneumococcal disease	[17]
Campylobacteriosis	[69]
Hand, foot and mouth disease	[6]
Measles	[30, 34, 50, 56]
Influenza	[58, 46, 47, 62]
Norovirus	[32, 31, 13]
Rotavirus	[13, 12]
Pertussis	[51]
Tuberculosis	[76]
Meningococcal disease	[58]
Severe fever with thrombocytopenia syndrome (SFTS)	[71]
Coronavirus disease 2019 (COVID-19)	see next section

NB: A regularly updated table of use cases is maintained by Sebastian Meyer at https://github.com/rforge/surveillance/blob/master/www/applications_EE.csv.

The EE framework is considered state-of-the-art and is often used used as a benchmark model for comparison in infectious disease modelling and probabilistic forecasting ([7, 61, 64]). In model construction using the EE framework, it is possible to incorporate dependencies such as the spatial movements of a population under study; the effects of human movement can be examined statistically using gravity models ([14]). Gravity models examine the flow from one subpopulation to another taking into account locations on mobility networks rather than geographical distance. Such models have been used to examine measles epidemics ([72, 38]) and influenza pandemics ([68]). This is but one example of where the EE framework works well with other available modelling options. Finally, the EE framework is implemented in a readily available software package ([50]), and its extensions are included in a wider ecosystem of packages within the same software ([11, 49, 10]).

1.2. Comparison with mathematical modelling approaches

To provide further synergy with established methods, the EE framework can be motivated from the discrete-time SIR compartmental model ([36, 6, 67]) and can be adjusted to include natural depletion of susceptibles ([66]) as well as to incorporate potential future pharmaceutical countermeasures and vaccines ([34]). Co-occurrences and co-infections as well as the existence of multi-strain pathogens provide added levels of complexity to disease transmission. Specifically for multiple strains of a disease, Wakefield *et al.* [67] outline models that can be used for outbreak detection in settings — including the EE framework. Additionally, bivariate analysis of different diseases can be conducted, see e.g. the example of influenza and meningococcal disease by Paul *et al.* [58]. For this reason we believe the EE framework should conceivably be able to include 2020 influenza season in the northern hemisphere in a model for COVID-19. This is a co-occurrence and co-infection scenario being raised by policy makers as we enter the final months of 2020. We conceptualise one situation where full synergy between mathematical and statistical modelling may not be possible: consider the Anderson–May equation for calculating the basic reproductive rate of sexually transmitted infections. Mathematically it might make sense to consider the interplay between the five or so parameters but once we examine case count data, we may only really be able to estimate one parameter in place of the five. If the effects are not inseparable, additional data may be required to estimate them.

2. COVID-19 CASE STUDIES

We feel the EE framework is particularly well-suited to being adapted to examine COVID-19. The EE framework was developed in surveillance situations of weekly case counts of established diseases, both in terms of biology as well as available information capturing infrastructure of their associated surveillance systems, meaning no new data gathering approaches were required. The EE framework offers increased flexibility and robustness compared to more standard epidemic models which may need to be constructed on a disease-by-disease basis. The framework allows us to incorporate available evidence at various levels of detail and examine intervention measures and other explanatory variables, e.g. meteorological ([6]) with all unknown parameters being estimated with likelihood techniques from the available data. Spatio-temporal spread can be captured by suitably parametrised power laws ([47]) and gravity models ([72]), and long term predictions can be produced ([32]). Importantly, the spatio-temporal formulation of the EE model can be extended to include age-dependent contact information ([48]), which is often considered a proxy of transmission events for respiratory disease such as COVID-19. The EE model has also been extended to include higher order lags in the epidemic component, allowing for the inclusion of infectiousness from the entire serial interval in the analysis of daily COVID-19 counts ([13]). This allows us to consider data at a finer temporal resolution than weekly, and analyse the near-real-time daily COVID-19 case information. We are aware that the EE framework has been used in the epidemiologic and economic studies of COVID-19 listed in Table 2.

Table 2: EE models focusing on COVID-19.

Author	Area of focus
Dickson *et al.* [20]	Italy
Giuliani *et al.* [28]	Italy
Alipour *et al.* [4]	Germany
Berlamann and Haustein [8]	Germany
Fritz and Kauermann [23]	Germany
Fronterre *et al.* [24]	England
Ssentongo *et al.* [63]	African continent
SUSPend	Switzerland and surrounding

2.1. Introduction of the SUSPend project

We have been using EE modelling approaches in the *SUSPend: Impact of Social distancing policies and Underreporting on the SPatio-temporal spread of COVID-19* project. This project is funded by the Swiss national science foundation's emergency support for research into coronaviruses as project number 196247. A description of the project can be found at https://data.snf.ch/covid-19/snsf/196247. Within the SUSPend project we are working on two subprojects, both concern the introduction of time-varying transmission weights in the model. The first subproject incorporates a contact matrix which changes over time. In particular, we are considering a synthetic contact matrix for Switzerland ([25]). The benefit to using the synthetic contact matrix for Switzerland rather than the single empirical one which exists ([35]) is that the sampling approach for the synthetic matrix is well-designed and the sample size is sufficiently large. The Swiss contact matrix considers contacts in various settings and we have adjusted these to reflect social distancing measures put in place, similar to other approaches seen in COVID-19 modelling ([70, 60, 55, 21]). The obvious alternative to adjusting contact matrices would be to consider instead contact surveys conducted during the COVID-19 outbreak as part of the EpiPose project (https://cordis.europa.eu/project/id/101003688), whose contact survey work has recently expanded to cover additional countries, including Switzerland [personal communication]. Such information has mainly been gathered in the United Kingdom, an island nation in northern Europe, which may be very different to landlocked Switzerland. Additional contact surveys conducted during the 2020 COVID-19 outbreak have been done by Feehan and Mahmud [22] in a north American setting and Latsuzbaia *et al.* [45] in a central European setting, indicating there is an increasing awareness that understanding the evolution of contacts established during an ongoing outbreak is useful for informing future outbreak modelling efforts.

The second subproject we are working on uses time-varying adjacency matrices and focuses on spatial spread of COVID-19. We create time-varying adjacency matrices for the seven Swiss NUTS-2 regions and their immediate neighbouring regions. These matrices are adjusted from baseline adjacencies based on mobility data gathered from smartphone users available at subregional level. Such mobility data has been used in studies of COVID-19 in multiple countries, including: China ([63, 44, 42, 27, 2]), Taiwan ([15]), Japan ([43]), Italy

([59, 26]), France ([26]), the United Kingdom ([39, 5, 26]), the United States of America ([65, 40, 37, 74, 41]), and Brazil, Chile, Bolivia, Colombia, and Peru ([74]). Our first project concerns the first half of 2020 while the timescale considered in the second project is longer. The EE framework is suitably flexible to allow us to incorporate additional information as it is found to be important. Thus, considerations nested in both policy making and biological can be included in the model as they are identified.

An issue common to the COVID-19 pandemic, and thus both of our subprojects, is the presence of underreporting and reporting delays in case data ([53]). Simple multiplication factors can be applied to address the former. However, such multiplication factors need to be time-dependent to incorporate increased testing capacities and changes in testing strategies observed in some countries. Multiplication factors may also vary across age groups, which is particularly relevant for the subproject with time-varying contact matrices as this has an age focus. The usefulness of incorporating delays in disease surveillance models has been shown ([12]). Nowcasting allows us to predict the true number of case counts based on available data and can be used to address reporting delays. Within compartmental modelling, nowcasting is often referred to as "real-time modelling". Nowcasting requires information both on test and reporting date on an individual basis. Unfortunately such information is rarely available in surveillance systems.

ACKNOWLEDGMENTS

We thank Lucas H. Kook for fruitful discussion on the Anderson–May equation. Our suggestions of use of mobility data made available by software companies is not an endorsement of the entities who gathered it. Country designation used in this manuscript are without prejudice to positions on status of the Republic of China (Taiwan) under the terms of international law.

REFERENCES

[1] Adegboye, O.A. and Adegboye, M. (2017). Spatially correlated time series and ecological niche analysis of cutaneous leishmaniasis in Afghanistan, *Int. J. Environ. Res. Public Health*, **14**(3), 309.

[2] Ainslie, K.E.C.; Walters, C.E.; Fu, H.; Bhatia, S.; Wang, H.; Xi, X.; Baguelin, M.; Bhatt, S.; Boonyasiri, A.; Boyd, O.; Cattarino, L.; Ciavarella, C.; Cucunuba, Z.; Cuomo-Dannenburg, G.; Dighe, A; Dorigatti, I.; van Elsland, S.L.; FitzJohn, R.; Gaythorpe, K; Ghani, A.C.; Green, W.; Hamlet, A.; Hinsley, W.; Imai, N.; Jorgensen, D.; Knock, E.; Laydon, D; Nedjati-Gilani, G.; Okell, L.C.; Siveroni, I.; Thompson, H.A.; Unwin, H.J.T.; Verity, R.; Vollmer, M.; Walker, P.G.T.; Wang, Y.; Watson, O.J.; Whittaker, C.; Winskill, P.; Donnelly, C.A.; Ferguson, N.M. and Riley, S. (2020). Evidence of initial success for China exiting COVID-19 social distancing policy after achieving containment, *Wellcome Open Res.*, **5**, 81.

[3] Alba-Casals, A.; Allue, E.; Tarancon, V.; Baliellas, J.; Novell, E.; Napp, S. and Fraile, L. (2020). Near real-time monitoring of clinical events detected in swine herds in Northeastern Spain, *Front. Vet. Sci.*, **7**, 68.

[4] Alipour, J.-V.; Fadinger, H. and Schymik, J. (2020). My home is my castle: the benefits of working from home during a pandemic crisis. Evidence from Germany, *ifo Working Papers*, **329**.

[5] Basellini, U.; Alburez-Gutierrez, D.; Del Fava, E.; Perrotta, D.; Bonetti, M.; Camarda, C.G. and Zagheni, E. (2020). Linking excess mortality to Google mobility data during the COVID-19 pandemic in England and Wales, *HAL archives-ouvertes*, https://hal.archives-ouvertes.fr/hal-02899654.

[6] Bauer, C. and Wakefield, J. (2018). Stratified space?time infectious disease modelling, with an application to hand, foot and mouth disease in China, *J. R. Stat. Soc. Ser. C Appl. Stat.*, **67**(5), 1379–1398.

[7] Bauer, C.; Wakefield, J.; Rue, H.; Self, S.; Feng, Z. and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data, *Stat. Med.*, **35**(11), 1848–1865.

[8] Berlamann, M. and Haustein, E. (2020). Right and yet wrong: a spatio-temporal evaluation of Germany?s COVID-19 containment policy, *CESifo Working Papers*, **8446**, https://www.cesifo.org/DocDL/cesifo1_wp8446.pdf.

[9] Birrell, P.J.; Wernisch, L.; Tom, B.D.M.; Held, L.; Roberts, G.O.; Pebody, R.G. and De Angelis, D. (2020). Efficient real-time monitoring of an emerging influenza pandemic: how feasible?, *Ann. Appl. Stat.*, **14**(1), 74–93.

[10] Bracher, J. (2019). hhh4underreporting. R package, https://github.com/jbracher/hhh4underreporting.

[11] Bracher, J. and Held, L. (2019). hhh4addon: extending the functionality of surveillance:hhh4. R package, https://github.com/jbracher/hhh4addon.

[12] Bracher, J. and Held, L. (2020). A marginal moment matching approach for fitting endemic-epidemic models to underreported disease surveillance counts, *Biometrics*, https://doi.org/10.1111/biom.13371.

[13] Bracher, J. and Held, L. (2020). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction, *Int. J. Forecast.*, https://doi.org/10.1016/j.ijforecast.2020.07.002.

[14] Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena, *Science*, **342**(6164), 1337–1342.

[15] Chang, M.-C.; Kahn, R.; Li, Y.-A.; Lee, C.-S.; Buckee, C.O. and Chang, H.-H. (2020). Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan, *medRxiv*, https://doi.org/10.1101/2020.04.07.20053439.

[16] Cheng, Q.; Lu, X.; Wu, J.T.; Liu, Z. and Huang, J. (2016). Analysis of heterogeneous dengue transmission in Guangdong in 2014 with multivariate time series model, *Sci. Rep.*, **6**, 33755.

[17] Chiavenna, C.; Presanis, A.M.; Charlett, A.; de Lusignan, S.; Ladhani, S.; Pebody, R.G. and De Angelis, D. (2019). Estimating age-stratified influenza-associated invasive pneumococcal disease in England: a time-series model based on population surveillance data, *PLOS Med*, **16**(6), 1–21.

[18] Colón-González, F.J.; Lake, I.R.; Morbey, R.A.; Elliot, A.J.; Pebody, R. and Smith, G.E. (2018). A methodological framework for the evaluation of syndromic surveillance systems: a case study of England, *BMC Public Health*, **18**, 544.

[19] De Angelis, D. and Presanis, A.M. (2020). *Analysing multiple epidemic data sources.* In "Handbook of Infectious Disease Data Analysis" (L. Held, N. Hens, P.D. O'Neill, and J. Wallinga, Eds.), Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 477–508.

[20] Dickson, M.M.; Espa, G.; Giuliani, D.; Santi, F. and Savadori, L. (2020). Assessing the effect of containment measures on the spatio-temporal dynamic of COVID-19 in Italy, *Nonlinear Dyn.*, https://doi.org/10.1007/s11071-020-05853-7.

[21] Di Domenico, L.; Pullano, G.; Sabbatini, C.E.; Boëlle, P.-Y. and Colizza, V. (2020). Impact of lockdown on COVID-19 epidemic in Île-de-France and possible exit strategies, *BMC Med.*, **18**, 240.

[22] Feehan, D. and Mahmud, A. (2020). Quantifying interpersonal contact in the United States during the spread of COVID-19: first results from the Berkeley Interpersonal Contact Study, *medRxiv*, https://doi.org/10.1101/2020.04.13.20064014.

[23] Fritz, C. and Kauermann, G. (2020). On the interplay of regional mobility, social connectedness, and the spread of COVID-19 in Germany, *arXiv*, https://arxiv.org/abs/2008.03013.

[24] Fronterre, C.; Read, J.M.; Rowlingson, B.; Bridgen, J.; Alderton, S.; Diggle, P.J. and Jewell, C.P. (2020). COVID-19 in England: spatial patterns and regional outbreaks, *medRxiv*, https://doi.org/10.1101/2020.05.15.20102715.

[25] Fumanelli, L.; Ajelli, M.; Manfredi, P.; Vespignani, A. and Merler, S. (2012). Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread, *PLOS Comput. Biol.*, **8**(9), 1–10.

[26] Galeazzi, A.; Cinelli, M.; Bonaccorsi, G.; Pierri, F.; Schmidt, A.L.; Scala, A.; Pammolli, F. and Quattrociocchi, W. (2020). Human mobility in response to COVID-19 in France, Italy and UK, *arXiv*, https://arxiv.org/abs/2005.06341.

[27] Gibbs, H.; Liu, Y.; Pearson, C.A.B.; Jarvis, C.I.; Grundy, C.; Quilty, B.J.; Diamond, C.; LSHTM CMMID COVID-19 working group and Eggo, R.M. (2020). Changing travel patterns in China during the early stages of the COVID-19 pandemic, *medRxiv*, https://doi.org/10.1101/2020.05.14.20101824.

[28] Giuliani, D.; Dickson, M.M.; Espa, G. and Santi, F. (2020). Modelling and predicting the spatio-temporal spread of coronavirus disease 2019 (COVID-19) in Italy, *BMC Infectious Diseases*, **20**, 700.

[29] Held, L.; Hofmann, M.; Höhle, M. and Schmid, V. (2006). A two-component model for counts of infectious diseases, *Biostatistics*, **7**, 422–437.

[30] Held, L.; Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts, *Stat. Model.*, **5**(3), 187–199.

[31] Held, L. and Meyer, S. (2020). *Forecasting based on surveillance data*. In "Handbook of Infectious Disease Data Analysis" (L. Held, N. Hens, P.D. O'Neill, and J. Wallinga, Eds.), Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 509–528.

[32] Held, L.; Meyer, S. and Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture, *Stat. Med.*, **36**(22), 3443–3460.

[33] Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data, *Biom. J.*, **54**(6), 824–843.

[34] Herzog, S.; Paul, M. and Held, L. (2011). Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data, *Epidemiol. Infect.*, **139**(4), 505–515.

[35] Hoang, T.; Coletti, P.; Melegaro, A.; Wallinga, J.; Grijalva, C.G.; Edmunds, J.W.; Beutels, P. and Hens, N. (2019). A systematic review of social contact surveys to inform transmission models of close-contact infections, *Epidemiology*, **30**(5), 723–736.

[36] Höhle, M. (2016). *Infectious disease modelling*. In "Handbook of Spatial Epidemiology" (A. Lawson, S. Banerjee, R. Haining, and L. Ugarte, Eds.), Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

[37] Huang, X.; Li, Z.; Jiang, Y.; Ye, X.; Deng, C.; Zhang, J. and Li, X. (2020). The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the U.S. during the COVID-19 pandemic, *medRxiv*, https://doi.org/10.1101/2020.07.31.20143016.

[38] Jandarov, R.; Haran, M.; Bjørnstad, O. and Grenfell, B. (2020). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease, *J. R. Stat. Soc. Ser. C Appl. Stat.*, **63**(3), 423–444.

[39] Jeffrey, B.; Walters, C.E.; Ainslie, K.E.C.; Eales, O.; Ciavarella, C.; Bhatia, S.; Hayes, S.; Baguelin, M.; Boonyasiri, A.; Brazeau, N.F.; Cuomo-Dannenburg, G.; FitzJohn, R.G.; Gaythorpe, K.; Green, W.; Imai, N.; Mellan, T.A.; Mishra, S.; Nouvellet, P.; Unwin, H.J.T.; Verity, R.; Vollmer, M.; Whittaker, C.; Ferguson, N.M.; Donnelly, C.A. and Riley, S. (2020). Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with COVID-19 social distancing interventions was high and geographically consistent across the UK, *Wellcome Open Res.*, **5**, 170.

[40] Kissler, S.; Kishore, N.; Prabhu, M.; Goffman, D.; Beilin, Y.; Landau, R.; Gyamfi-Bannerman, C.; Bateman, B.; Katz, D.; Gal, J.; Bianco, A.; Stone, J.; Larremore, D.; Buckee, C. and Grad, Y. (2020). Reductions in commuting mobility predict geographic differences in SARS-CoV-2 prevalence in New York City, http://nrs.harvard.edu/urn-3:HUL.InstRepos:42665370.

[41] Klein, B.; LaRock, T.; McCabe, S.; Torres, L.; Friedland, L.; Privitera, F.; Lake, B.; Kraemer, M.U.G.; Brownstein, J.S.; Lazer, D.; Eliassi-Rad, R.; Scarpino, S.V.; Vespignani, A. and Chinazzi, M. (2020). Reshaping a nation: Mobility, commuting, and contact patterns during the COVID-19 outbreak, https://www.mobs-lab.org/uploads/6/7/8/7/6787877/covid19mobility_report2.pdf.

[42] Kraemer, M.U.G.; Yang, C.-H.; Gutierrez, B.; Wu, C.-H.; Klein, B.; Pigott, D.M.; Open COVID-19 Data Working Group; du Plessis, L.; Faria, N.R.; Li, R.; Hanage, W.P.; Brownstein, J.S.; Layan, M.; Vespignani, A.; Tian, H.; Dye, C.; Pybus, O.G. and Scarpino, S.V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China, *Science*, **368**(6490), 493–497.

[43] Kurita, J.; Sugawara, T. and Ohkusa, Y. (2020). NTT Docomo and Apple mobility data compared as countermeasures against COVID-19 outbreak in Japan, *medRxiv*, https://doi.org/10.1101/2020.05.01.20087155.

[44] Lai, S.; Ruktanonchai, N.W.; Zhou, L.; Prosper, O.; Luo, W.; Floyd, J.R.; Wesolowski, A.; Santillana, M.; Zhang, C.; Du, X.; Yu, H. and Tatem, A.J. (2020). Effect of non-pharmaceutical interventions to contain COVID-19 in China, *Nature*, **585**(7825), 410–413.

[45] Latsuzbaia, A.; Herold, M.; Bertemes, J.-P. and Mossong, J. (2020). Evolving social contact patterns during the COVID-19 crisis in Luxembourg, *PLOS ONE*, **15**(8), 1–13.

[46] Lu, J. and Meyer, S. (2020). Forecasting flu activity in the United States: benchmarking an endemic-epidemic beta model, *Int. J. Environ. Res. Public Health*, **17**(4), 1381.

[47] Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread, *Ann. Appl. Stat.*, **8**, 1612–1639.

[48] Meyer, S. and Held, L. (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread, *Biostatistics*, **18**, 338–351.

[49] Meyer, S. and Held, L. (2017). hhh4contacts: Age-Structured Spatio-Temporal Models for Infectious Disease Counts. R package, https://github.com/cran/hhh4contacts.

[50] Meyer, S.; Held, L. and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance, *J. Stat. Softw.*, **77**(11), 1–55.

[51] Munro, A.; Smallman-Raynor, M. and Algar, A.C. (2020). Long-term changes in endemic threshold populations for pertussis in England and Wales: a spatiotemporal analysis of Lancashire and South Wales, 1940-69, *Soc. Sci. Med.*, 113295, https://doi.org/10.1016/j.socscimed.2020.113295.

[52] Nightingale, E.S.; Chapman, L.A.C.; Srikantiah, S.; Subramanian, S.; Purushothaman, J.; Bracher, J.; Cameron, M. and Medley, G. (2020). A spatio-temporal approach to short-term forecasting of visceral leishmaniasis diagnoses in India, *medRxiv*, https://doi.org/10.1101/19009258.

[53] Noufaily, A. (2020). *Underreporting and reporting delays*. In "Handbook of Infectious Disease Data Analysis" (L. Held, N. Hens, P.D. O'Neill, and J. Wallinga, Eds.), Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 437–454.

[54] Nunes, B.; Caetano, C.; Antunes, L. and Dias, C. (2020). Statistics in times of pandemics: the role of statistical and epidemiological methods during the COVID-19 emergency, *REVSTAT – Statistical Journal*, **18**(5), 553–564.

[55] Panovska-Griffiths, J.; Kerr, C.C.; Stuart, R.M.; Mistry, D.; Klein, D.J.; Viner, R.M. and Bonell, C. (2020). Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study, *Lancet Child Adolesc. Health*, https://doi.org/10.1016/S2352-4642(20)30250-9.

[56] Parpia, A.S.; Skrip, L.A.; Nsoesie, E.O.; Ngwa, M.C.; Abah Abah, A.S.; Galvani, A.P. and Ndeffo-Mbah, M.L. (2020). Spatio-temporal dynamics of measles outbreaks in Cameroon, *Ann. Epidemiol.*, **42**, 64–72.e3.

[57] Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts, *Stat. Med.*, **30**, 1118–1136.

[58] Paul, M.; Held, L. and Toschke, A.M. (2008). Multivariate modelling of infectious disease surveillance data, *Stat. Med.*, **27**(29), 6250–6267.

[59] Pepe, E.; Bajardi, P.; Gauvin, L.; Privitera, F.; Lake, B.; Cattuto, C. and Tizzoni, M. (2020). COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown, *Sci. Data*, **7**(1), 230.

[60] Prem, K.; Liu, Y.; Russell, T.; Kucharski, A.J.; Eggo, R.M.; Davies, N.; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group; Jit, M. and Klepac, P. (2020). The effect of control strategies that reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China, *Lancet Public Health*, **5**(5), 261–270.

[61] Ray, E.L.; Sakrejda, K.; Lauer, S.A.; Johansson, M.A. and Reich, N.G. (2017). Infectious disease prediction with kernel conditional density estimation, *Stat. Med.*, **36**(30), 4908–4929.

[62] Reeve, K. (2017). *Spatio-temporal forecasting and infectious disease count data*, Master's thesis, University of Zurich, https://www.math.uzh.ch/li/index.php?file&key1=54294.

[63] Ssentongo, P.; Fronterre, C.; Geronimo, A.; Greybush, S.J.; Mbabazi, P.K.; Muvawala, J.; Nahalamba, S.; Omadi, P.O.; Opar, B.T.; Sinnar, S.A.; Wang, Y.; Whalen, A.J.; Held, L.; Jewell, C.; Muwanguzi, A.J.B.; Greatrex, H.; Norton, M.M.; Diggle, P. and Schiff, S.J. (2020). *Tracking and predicting the African COVID-19 pandemic*, (in prep.).

[64] Stojanović, O.; Leugering, J.; Pipa, G.; Ghozzi, S. and Ullrich, A. (2019). A Bayesian Monte Carlo approach for predicting the spread of infectious diseases, *PLOS ONE*, **14**(12), 1–20.

[65] Unwin, H.; Mishra, S.; Bradley, V.C.; Gandy, A.; Vollmer, M.; Mellan, T.; Coupland, H.; Ainslie, K.; Whittaker, C.; Ish-Horowicz, J.; Filippi, S.; Xi, X.; Monod, M.; Ratmann, O.; Hutchinson, M.; Valka, F.; Zhu, H.; Hawryluk, I.; Milton, P.; Baguelin, M.; Boonyasiri, A.; Brazeau, N.; Cattarino, L.; Charles, G.; Cooper, L.; Cucunuba Perez, Z.; Cuomo-Dannenburg, G.; Djaafara, A.; Dorigatti, I.; Eales, O.; Eaton, J.; Van Elsland, S.; Fitzjohn, R.; Gaythorpe, K.; Green, W.; Hallett, T.; Hinsley, W.; Imai, N.; Jeffrey, B.; Knock, E.; Laydon, D.; Lees, J.; Nedjati Gilani, G.; Nouvellet, P.; Okell, L.; Ower, A.; Parag, K.; Siveroni, I.;

Thompson, H.; Verity, R.; Walker, P.; Walters, C.; Wang, Y.; Watson, O.; Whittles, L.; Ghani, A.; Ferguson, N.; Riley, S.; Donnelly, C.; Bhatt, S. and Flaxman, S. (2020). Report 23: State-level tracking of COVID-19 in the United States, https://doi.org/10.25561/79231.

[66] Wakefield, J. (2018). Spatio-temporal modeling of infectious disease data in a developing world setting — Presentation at the University of Montreal, http://www.crm.umontreal.ca/2018/Spatial18/pdf/Wakefield_slides.pdf.

[67] Wakefield, J.; Dong, T.Q. and Minin, V.N. (2020). *Spatio-temporal analysis of surveillance data*. In "Handbook of Infectious Disease Data Analysis" (L. Held, N. Hens, P.D. O'Neill, and J. Wallinga, Eds.), Chapman & Hall/CRC Handbooks of Modern Statistical Methods, 455–476.

[68] Walters, C.E.; Meslé, M.M.I. and Hall, I.M. (2018). Modelling the global spread of diseases: a review of current practice and capability, *Epidemics*, **25**, 1–8.

[69] Wei, W.; Schüpbach, G. and Held, L. (2015). Time-series analysis of Campylobacter incidence in Switzerland, *Epidemiol. Infect.*, **143**(9), 1982–1989.

[70] Willem, L.; Hoang, T.V.; Funk, S.; Coletti, P.; Beutels, P. and Hens, N. (2020). SOCRATES: An online tool leveraging a social contact data sharing initiative to assess mitigation strategies for COVID-19, *BMC Res. Notes*, **13**(1), 293.

[71] Wu, H.; Wu, C.; Lu, Q.; Ding, Z.; Xue, M. and Lin, J. (2020). Spatial-temporal characteristics of severe fever with thrombocytopenia syndrome and the relationship with meteorological factors from 2011 to 2018 in Zhejiang Province, China, *PLOS Negl. Trop. Dis.*, **14**(4), 1–17.

[72] Xia, Y.; Bjørnstad, O.N. and Grenfell, B.T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics, *Am. Nat.*, **164**(2), 267–281.

[73] Yuan, Z.; Xiao, Y.; Dai, Z.; Huang, J.; Zhang, Z. and Chen, Y. (2020). Modelling the effects of Wuhan's lockdown during COVID-19, China, *Bull. World Health Organ*, **98**, 484–494, https://doi.org/10.2471/BLT.20.254045.

[74] Zhu, D.; Mishra, S.R.; Han, X. and Santo, K. (2020). Social distancing in Latin America during the COVID-19 pandemic: an analysis using the Stringency Index and Google Community Mobility Reports, *J. Travel Med.*, https://doi.org/10.1093/jtm/taaa125.

[75] Zhu, G.; Xiao, J.; Liu, T.; Zhang, B.; Hao, Y. and Ma, W. (2019). Spatiotemporal analysis of the dengue outbreak in Guangdong Province, China, *BMC Infect. Dis.*, **19**, 493.

[76] Zuo, Z.; Wang, M.; Cui, H.; Wang, Y.; Wu, J.; Qi, J.; Pan, K.; Sui, D.; Liu, P. and Xu, A. (2020). Spatiotemporal characteristics and the epidemiology of tuberculosis in China from 2004 to 2017 by the nationwide surveillance system, *BMC Public Health*, **20**(1), 1284.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 575–576

A BRIEF APPRAISAL OF THE COVID-19 PANDEMIC IN PORTUGAL

(*Discussion on the paper by Nunes, Caetano, Antunes and Dias*)

Author: M. Gabriela M. Gomes
– University of Strathclyde, Glasgow, United Kingdom,
and University of Porto, Porto, Portugal
gabriela.gomes@strath.ac.uk

In this issue, Nunes *et al.* ([4]) review the prominent role embraced by statistical and epidemiological researchers in conducting societies through the COVID-19 pandemic. Basic epidemiological concepts, such as infection-fatality ratios (IFR), temporal infection reproduction numbers (R_t) and herd immunity thresholds (HIT), became part of every person vocabulary. Daily new cases, recoveries and deaths, have been diligently tracked and featured at the opening of prime-time news since the first case was confirmed in each country (February–March in most of Europe). Besides reviewing research conducted by themselves and others, Nunes *et al.* ([4]) provide a frank account of the challenges associated with conducting scientific research under such spotlight. Their paper should interest a very wide readership.

Comparing COVID-19 trajectories across countries and regions and appraising control strategies became topical in most social encounters, whether physical or virtual. Europe was the first major epicentre outside the source in China, and European countries quickly started to be classified into those with high death toll (such as Italy, UK, Belgium and Sweden) or low death toll (most prominently, Czech Republic) during the first wave in the spring 2020. Due to combinations of non-pharmaceutical interventions based on social distancing measures and naturally acquired immunity in populations, epidemics curbed throughout Europe and cases were brough to very low levels during the summer. By the end of the summer and into the autumn, Europe started to experience a second wave. Countries who were least affected in the spring (such as Czech Republic) are seeing steeper rises now, most plausibly due to having acquired less immunity. Portugal appears in neither of these extremes. Rates of infection and death were moderate throughout and the epidemic is under control. Although a final assessment is not possible until the pandemic is over, I expect the Portuguese strategy to rank among the most balanced. This would almost certainly not have been the case without the dedicated work of statisticians and epidemiologists.

The authors touch briefly on the role of mathematical modelling of the COVID-19 pandemic. Models were developed early in the pandemic to project epidemic trajectories in various countries ([5, 3]). Initial projections for Portugal suggested that, without mitigation, up to 70% of the population would be infected before cases started to decline (HIT), 85% would be infected by the time the epidemic ended and 1% would die as a result ([5]).

These results relied on the hypothesis that populations were homogeneously susceptible, which was recently refuted ([2]). According to models constructed to account for complete heterogeneity in susceptibility or exposure to infection, and conditional on the accuracy of currently available seroprevalence results, herd immunity is expected at much lower infection levels (around 10–20%) which I estimate to be happening this Autumn in Portugal. Continuing monitoring will inform the accuracy of these estimates but, if confirmed, these results imply that the COVID-19 pandemic is ending in Europe. Incident cases will continue but sustained epidemic growth will not be expected more than for other seasonal respiratory viruses. Given the disease fatality observed in the spring, protection of the most vulnerable is critical until a vaccine is available but the risk of complications is considered low for the majority of the population.

REFERENCES

[1] Aguas, R.; Corder, R.M.; King, J.G.; Gonçalves, G.; Ferreira, U.M. and Gomes, M.G.M. (2020). Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics, *medRxiv*, https://doi.org/10.1101/2020.07.23.20160762.

[2] Colombo, M.; Mellor, J.; Colhoun, H.M.; Gomes, M.G.M. and McKeigue, P.M. (2020). Trajectory of COVID-19 epidemic in Europe, *medRxiv*, https://doi.org/10.1101/2020.09.26.20202267.

[3] Flaxman, S.; Mishra, S.; Gandy, A.; Unwin, H.J.T.; Mellan, T.A.; Coupland, H.; Whittaker, C.; Zhu, H.; Berah, T.; Eaton, J.W.; Monod, M.; Imperial College COVID-19 Response Team; Ghani, A.C.; Donnelly, C.A.; Riley, S.; Vollmer, M.A.C.; Ferguson, N.M.; Okell, L.C. and Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, *Nature*, **584**, 257–261.

[4] Nunes, B.; Caetano, C.; Antunes, L. and Dias, C. (2020). Statistics in times of pandemics: the role of statistical and epidemiological methods during the COVID-19 emergency, *REVSTAT – Statistical Journal*, **18**(5), 553–564.

[5] Walker, P.G.T.; Whittaker, C.; Watson, O.; Baguelin, M.; Ainslie, K.E.C.; Bhatia, S.; Bhatt, S.; Boonyasiri, A.; Boyd, O.; Cattarino, L.; Cucunubá, Z.; Cuomo-Dannenburg, G.; Dighe, A.; Donnelly, C.D.; Dorigatti, I.; van Elsland, S.; FitzJohn, R.; Flaxman, S.; Fu, H.; Gaythorpe, K.; Geidelberg, L.; Grassly, N.; Green, W.; Hamlet, A.; Hauck, K.; Haw, D.; Hayes, S.; Hinsley, W.; Imai, N.; Jorgensen, D.; Knock, E.; Laydon, D.; Mishra, S.; Nedjati-Gilani, G.; Okell, L.C.; Riley, S.; Thompson, H.; Unwin, J.; Verity, R.; Vollmer, M.; Walters, C.; Wang, H.W.; Wang, Y.; Winskill, P.; Xi, X.; Ferguson, N.M. and Ghani, A.C. (2020). The global impact of COVID-19 and strategies for mitigation and suppression, *Imperial College COVID-19 Response Team*, Report 12, https://doi.org/10.25561/77735.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 577–579

REJOINDER

(*Discussion on the paper by Nunes, Caetano, Antunes and Dias*)

Authors: Baltazar Nunes
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, and Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Portugal
baltazar.nunes@insa.min-saude.pt

Constantino Caetano
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal
constantino.caetano@insa-min-saude.pt

Liliana Antunes
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal
liliana.antunes@insa-min-saude.pt

Carlos Dias
– Dep. Epidemiology, Instituto Nacional de Saúde Dr. Ricardo Jorge, and Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Portugal
carlos.dias@insa-min-saude.pt

The authors would like to thank the commentaries from Bekker-Nielsen Dunbar and Held ([2]) and from Gomes ([3]) to their article *Statistics in times of pandemics: the role of statistical and epidemiological methods during the COVID-19 emergency*, and for sharing their experience, knowledge and work in the response to the COVID-19 epidemic.

In their commentary, Bekker-Nielsen Dunbar and Held ([2]) introduced the time-series framework of endemic-epidemic models, showing the benefit of this approach to resolve and overcome many of the challenges identified during the analysis of surveillance data and epidemic dynamics modelling. They also share their experience using Switzerland's data and present the work developed under the SUSPend project, that aims at bring knowledge on essentials components of the epidemic transmission, for its control and mitigation.

The same is revealed in the Gomes ([3]) commentary that, based on new mathematical modelling approaches, suggests that by introducing different levels of heterogeneity in the immunity or susceptibility of the population, the size of the epidemics and its conclusion can change dramatically.

Considering the four main components of infectious diseases transmission, summarised in the acronym DOTS (Duration of infectious period, Opportunities of transmission, probability of Transmission upon contact and Susceptibility), as described by Adam Kucharski from the London School of Hygiene and Tropical Medicine ([4]), it is possible to verify that, Bekker-Nielsen Dunbar and Held ([2]) and Gomes ([3]), cover, in their works, the majority of

these components. The understanding of these four components is crucial for the epidemic control and mitigation.

Gomes ([3]) describe works developed with colleagues ([1]), on the role that the heterogeneity in population susceptibility and connectivity play on the SARS-CoV-2 epidemic dynamics. They propose that by using classical mathematical disease transmission models, that assume everyone is equally susceptible, one can overestimate the Herd Immunity Threshold (HIT). Their results show that by increasing the heterogeneity of the population susceptibility, the HIT could be as low as 10%. Which means that the COVID-19 epidemic could be resolved during the 2020–21 autumn–winter season. This is an important and hopeful hypothesis, however it cannot be used as the base model for planning and decision making during the public health emergency, where it is necessary to plan for the worst and hope for the best.

The works described by Bekker-Nielsen Dunbar and Held ([2]), namely in the project SUSPend, look more deeply into the role of the Opportunities of transmission, and indirectly into the component probability of Transmission upon contact, by introducing in their models population contact matrices to model COVID-19 disease dynamics. One important feature is the possibility of modelling the changes in population contacts as a function of time, by allowing the introduction of time-dependent contact matrices. Modelling these components and measuring their impact on the disease transmission is of utmost importance, given that a relevant part of Non Pharmaceutical Interventions (NPI), recommended by Public Health Authorities, are: incentivizing social distancing, use of mask and respiratory hygiene. This pandemic has shown to us the capability of NPI, and their ability to interrupted or reduce infection transmission, given that they can be implemented at levels that do not suspend the needed social interaction and the economy.

Finally, the last component of DOTS is the Duration of infectiousness that was not directly discussed in the commentaries of Bekker-Nielsen Dunbar and Held ([2]) and Gomes ([3]). This component relates to the number of infectious individuals present in the general population, that are not isolated or in a hospital, and their infectious period. The lower the number of infectious individuals and the shorter their infectious period, results in lower transmissibility of the infection. The reduction of the transmission in this component it is in part achievable by the early identification of infection cases, its quick isolation and quarantine of the closer contacts. A group of mathematical modellers from Universidade de Trás-dos-Montes, Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa and the Instituto Nacional de Saúde Dr. Ricardo Jorge are currently developing an age-structured SEIR model that accounts for different levels of contact tracing and symptomatic cases identification. This model is being developed under the project COVID-19 in-CTRL funded by Fundação para a Ciência e Tecnologia, under the program RESEARCH4COVID ([5]). The aim of this project is to evaluate the necessary coverage of SARS-CoV-2 symptomatic cases that need to be identified by public health teams, and the proportion of close contacts that need to be traced in order to reduce the effective reproduction number below 1.

As a final note, the authors would like to emphasize the important role that previous statistical and mathematical modelling of infectious diseases research had in the scientific response during the emergency of COVID-19. Previous developments, including open source R packages like the ones made available by R Epidemics Consortium (RECON) ([5]), or the ones described in the Bekker-Nielsen Dunbar and Held ([2]) commentary, like the surveillance R ([6, 7]) package, were of invaluable importance during all the phases of this pandemic.

In fact, it has shown the benefit of sharing knowledge and resources in the scope of enhancing the global response to current and future pandemics.

REFERENCES

[1] Aguas, R.; Corder, R.M.; King, J.G.; Gonçalves, G.; Ferreira, U.M. and Gomes, M.G.M. (2020). Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics, *medRxiv*, https://doi.org/10.1101/2020.07.23.20160762.

[2] Bekker-Nielsen Dunbar, M. and Held, L. (2020). Endemic-epidemic framework used in COVID-19 modelling (Discussion on the paper by Nunes, Caetano, Antunes and Dias), *REVSTAT – Statistical Journal*, **18**(5), 565–574.

[3] Gomes, M.G.M. (2020). A brief appraisal of the COVID-19 pandemic in Portugal (Discussion on the paper by Nunes, Caetano, Antunes and Dias), *REVSTAT – Statistical Journal*, **18**(5), 575–576.

[4] Kurcharki, A. (2020). *The Rules of Contagion: Why Things Spread – and Why They Stop*, Profile Books Ltd, London.

[5] Nunes, B. (coordinator) (2020). COVID-19 in-CTRL: Projeção do Impacte das medidas Não-farmacológicas de Controlo e mitigação da epidemia de COVID-19 em Tempo ReaL, *FCT project*, https://www.cma.fct.unl.pt/news/2020/07/projecao-do-impacte-das-medidas-nao-farmacologicas-de-controlo-e-mitigacao-da-epidemia-de-covid-19-e.

[6] RECON team (2020). R Epidemic Consortium, (visited 07/10/2020), https://www.repidemicsconsortium.org/.

[7] Salmon, M.; Schumacher, D. and Höhle, M. (2016). Monitoring count time series in R: aberration detection in public health surveillance, *Journal of Statistical Software*, **70**(10), 1–35.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 581–592

MATCHED PAIRS WITH BINARY OUTCOMES

Authors: Christiana Kartsonaki
– MRC Population Health Research Unit, Nuffield Department of Population Health, University of Oxford,
Oxford OX3 7LF, UK
christiana.kartsonaki@dph.ox.ac.uk

D.R. Cox
– Nuffield College,
Oxford OX1 1NF, UK
david.cox@nuffield.ox.ac.uk

Received: September 2018 Revised: May 2019 Accepted: June 2019

Abstract:

- Comparison of two treatments in matched pairs is a powerful general method for improving precision. When the outcome is binary the formulation in terms of logistic comparisons leads to an analysis in which concordant pairs, that is pairs in which both members show the same outcome, are discarded. The present paper discusses a number of conceptual aspects of this including a comparison with a linear in probabilities formulation, the relation between logistic parameters in different designs and in particular some new efficiency comparisons. Some emphasis is based on new relations between estimated effects derived from different formulations and on comparative calculations of asymptotic efficiency.

Key-Words:

- *asymptotic relative efficiency; comparison of distinct models; log odds; logistic regression.*

AMS Subject Classification:

- 62BO5:62F**.

1. INTRODUCTION

The comparison of two treatments, T_0 and T_1, using matched pairs of individuals is a simple and often effective way of improving precision and is the basis for many generalizations. When each outcome is binary, 0 or 1, say, there are four possible outcomes from a pair, (0, 0) and (1, 1), called concordant pairs, and (1, 0) and (0, 1), called discordant pairs. The analysis of such data has been extensively discussed, partly because of the broader implications for approaching formalized statistical inference; see, for example, the wide-ranging review of Agresti [1] (1990).

McNemar [10] (1947) suggested testing the null hypothesis of treatment equivalence by discarding the concordant pairs and testing the discordant pairs for an equal split between the two possibilities, using the binomial distribution with parameter 1/2. Cox [3, 4] (1958a, b) formalized this within a systematic approach to the analysis of binary data using an exponential family setting based on a linear logistic model. In the psychometric literature the problem is considered in this way as a special case of the Rasch model (Rasch [12], 1960).

One approach, possibly closer in spirit to McNemar's paper, is to treat the analysis as a simple significance test (Fisher [8, Chapter II], 1935) in which the strong null hypothesis is that the outcome on each individual is totally unaffected by the treatment allocations, taken to be by design independent randomization for each pair between the two possible assignments. With m pairs and two treatments there are thus 2^m possible configurations that might be observed, each with the same probability under the null hypothesis; McNemar's test follows from that. Here a stochastic model for the data is not needed; rather the stochastic element comes from the randomization in design. This is a powerful argument but limited in its implications because an estimation formulation attached to it is rather contrived. The extensive literature on the matched pairs and related issues has tended not to follow that route.

A general aspect that underlies the discussion of binary data goes back in particular to earlier differences between Karl Pearson and Yule (Pearson [11], 1907; Yule [13], 1903). The former treated the two binary variables in a simple 2×2 contingency table as derived from an underlying standardized bivariate Gaussian distribution whose correlation coefficient is to be estimated, whereas Yule considered the binary variables as such.

In the logistic formulation, let Y_{s0}, Y_{s1} be independent random variables representing the observations on the sth pair and suppose that for m pairs

$$P(Y_{s0} = i) = L_i(\alpha_s - \theta/2), \quad P(Y_{s1} = j) = L_j(\alpha_s + \theta/2), \quad i, j \in \{0, 1\}, \tag{1.1}$$

where $\alpha_s, \theta \in \mathbb{R}$ are unknown parameters, $L_1(x) = e^x/(1 + e^x)$ is the unit logistic function and $s = 1, ..., m$ and we write $L_0(x) = 1 - L_1(x)$. Interest is typically focused on θ whereas the α_s specify inter-pair differences supposed to be of no direct interest. Here i, j take values 0 and 1 and the parameter space is unconstrained.

It follows from the existence of complete sufficient statistics that if study of θ is to be made regarding the α_s as totally arbitrary nuisance parameters, then to achieve a procedure not formally depending on those parameters, inference is made conditionally on the pair totals,

and therefore confined to the discordant pairs. In the nonnull case this leads to confidence limits for θ based on the binomial distribution.

The discarding of concordant pairs has often been regarded with unease, especially if there are many such pairs, this superficially pointing to treatment equivalence. If, though, it is required to cover the possibility that many of the α_s are large in absolute value the rejection seems inevitable. If, however, implicit or explicit restrictions are placed on the variation of these parameters some information may be recoverable from the concordant pairs. For example, Lee [9] (2001) replaced the logistic model by a broadly equivalent Poisson model. Because of the richer reference set somewhat improved estimates involving the concordant pairs were obtained.

Another route is to replace the logistic function in the above formulation by some other function, for example the linear or Gaussian functions. Such a change might give a better fit or a more direct interpretation or accommodate several related studies more conformably. Aranda-Ordaz [2] (1981) studied a parametric family of transformations as a basis for choosing the best fitting model. Empirical discrimination betweeen different models typically requires extensive data.

In essentially discrete problems "exact" significance testing involves reference to a discrete distribution and hence for each data configuration to a limited set of achievable significance levels. There is a very extensive literature on how the discrete test can be augmented to achieve some pre-specified level, such as 0.05. For interpretative purposes such arbitrarily defined levels are irrelevant. Repetitive binary decision problems such as routine screening need to be treated as such.

2. OUTLINE ANALYSES OF TWO MODELS

For data from m independent pairs we write for pair s the likelihood contribution for outcome (i, j) as

$$L_i(\alpha_s - \theta/2)L_j(\alpha_s + \theta/2).$$

It follows that, if $\hat{\pi}_{ij}$ is the proportion of pairs with $Y_0 = i, Y_1 = j$ with corresponding probabilities π_{ij}, then

$$\pi_{ij} \;=\; \mathrm{Ave}_s L_i(\alpha_s - \theta/2)L_j(\alpha_s + \theta/2), \tag{2.1}$$

where s indexes the pairs and Ave_s is the average over the m pairs in the study or over a population of pairs from which the observed pairs have been randomly chosen. For each fixed θ the sufficient statistics for the α_s are the pair totals. Conditioning on these leaves no information in the concordant pairs and the contribution to the conditional log likelihood is thus $L_1(\theta)$ from each of the N_{01} pairs with outcome (0, 1) and $L_0(\theta)$ from each of N_{10} pairs (1, 0). It follows that $\hat{\theta} = \log(N_{01}/N_{10})$ and $L_1(\hat{\theta}) = N_{01}/(N_{01} + N_{10})$, so that, in particular, from the variance of a binomial distribution,

$$\mathrm{var}\{L_1(\hat{\theta})\} = L_0(\theta)L_1(\theta)/(m\pi_D), \tag{2.2}$$

and then by the formula for the asymptotic variance of nonlinear function

$$\text{var}(\hat{\theta}) = 1/\{mL_0(\theta)L_1(\theta)\pi_D\},$$

where π_D is the average probability that a pair is discordant so that $m\pi_D$ is the expected number of discordant pairs.

Suppose now that we replace the logistic formulation by the linear representation

$$P(Y_{s0} = 1) = 1/2 + \beta_s - \phi/2,\ P(Y_{s1} = 1) = 1/2 + \beta_s + \phi/2, \tag{2.3}$$

where β_s specifies the impact of the sth pair and ϕ gives the difference of probabilities between the two groups. The parameter space is constrained so that all probabilities are in [0, 1]. This places relatively complicated restrictions on the component parameters. If we write $\mu_\beta = \Sigma\beta_s/m,\ \ \sigma^2_\beta = \Sigma\beta_s^2/m - \mu^2_\beta$, then the four cell probabilities for the expected outcome proportions are in a symmetrized notation

$$\begin{aligned}
\pi_{00} &= 1/4 - \phi^2/4 - \mu_\beta + \gamma_\beta,\\
\pi_{01} &= 1/4 + \phi/2 + \phi^2/4 - \gamma_\beta,\\
\pi_{10} &= 1/4 - \phi/2 + \phi^2/4 - \gamma_\beta,\\
\pi_{11} &= 1/4 - \phi^2/4 + \mu_\beta + \gamma_\beta,
\end{aligned}$$

where $\gamma_\beta = \mu^2_\beta + \sigma^2_\beta$. Here ϕ specifies the inter-treatment differences and β_s characterizes the sth pair. Explicit characterization of the parameter space, that is the non-negativity of p_{ij} is not simple.

It follows that $\phi = \pi_{01} - \pi_{10}$ is estimated by

$$\hat{\phi} = \hat{\pi}_{01} - \hat{\pi}_{10} = (N_{01} - N_{10})/m. \tag{2.4}$$

The numerator is the sum of independent random variables taking the values $(-1, 0, 1)$ and it follows that

$$\text{var}(\hat{\phi}) = \frac{1 - \phi^2 - 4\gamma_\beta}{2m} = (\pi_D - \phi^2)/m. \tag{2.5}$$

This depends not only on the discordant pairs but, through the denominator, also on the total number of concordant pairs.

The variance component σ^2_β can be estimated through its equivalence to

$$1/4 - (\pi_{0.} - \pi_{1.})(\pi_{.0} - \pi_{.1})/4 - \pi_D/2.$$

In this discussion μ_β and σ^2_β are the mean and variance of the finite population of values of β_s. Alternatively if the β_s correspond to independent and identically distributed random variables and expectations are taken over their distribution the parameters μ_β and σ^2_β refer to that distribution.

3. SOME SIMPLE COMPARISONS

Comparison of logistic and linear in probability and indeed other models can be viewed in a number of distinct ways. From the viewpoint of formal statistical theory the logistic model has the major advantage of leading to a full exponential family form with the regression coefficients as canonical parameters (Cox [3], 1958a) and associated "exact" methods. Fully efficient estimation for the linear in probability models requires iterative calculation. However the use of ordinary least squares, treating the binary (0, 1) outcomes as if quantitative, has high efficiency so long as the probabilities are in a central range, say $(0.2, 0.8)$ (Cox and Wermuth [7], 1992). The direct subject-matter interpretation of differences in probabilities in terms of expected numbers of individuals affected is an advantage of the linear in probabilities model but the severe restrictions to specified regions of the parameter space are a major disadvantage of that formulation.

There is, however, a further general consideration applying to all issues connected with binary data and going back to the early work of Karl Pearson [11] (1907) and Yule [13] (1903) on the simpler 2×2 table. Pearson treated binary variables as formed from dichotomizing unobserved continuous variables having a bivariate normal distribution whose correlation is the focus of interest, whereas Yule treated binary variables directly in their own right. In many contexts the distinction is nugatory, although for quantal bioassay the former approach is directly relevant. Each study individual has a just critical dose above which, say, a lethal response is observed; each individual can be tested only once. Treating the unobserved critical dose levels as having a normal distribution, virtually indistinguishable from a continuous logistic distribution, is often reasonable; treating it as uniform, the implication of the linear in probabilities model, would typically not be.

The distinction between logistic and linear formulation disappears at the null hypothesis $\theta = \phi = 0$ and locally the parameter estimated in the linear in probability model is $\phi = \text{Ave}_s\{L_1(\alpha_s + \theta/2) - L_1(\alpha_s - \theta/2)\} = \theta \text{Ave}_s\{L_1'(\alpha_s)\}$, where $L_1'(\cdot)$ is the derivate of $L_1(\cdot)$, and this is approximately

$$\tilde{\phi} = \theta \text{Ave}_s\{L_1(\alpha_s)L_0(\alpha_s)\} = \theta\pi_D/2.$$

Here $\text{Ave}_s(b_s)$ is the unweighted average $\Sigma b_s/m$. The asymptotic relative efficiency of the linear and logistic procedures is thus given by the ratio $\text{var}(\tilde{\phi})/\text{var}(\hat{\phi})$ evaluated at the null hypothesis and this is one.

Both logistic and linear formulations have three free parameters and are therefore saturated families for the distribution over the four possible outcomes. The linear in probability model has for most purposes the more directly understandable interpretation, although if the proportions of, say, 1's are small, the interpretation of the logistic model in terms of proportional effects is attractive and the positivity constraints on the linear model are severe. Often the most appealing base for choosing between different formulations is stability of estimated effects across replicate sets of data, that is relative constancy of either θ or of ϕ, potentially favouring the logistic formulation.

Instead of matching in pairs it would be possible to randomize the allocation of individuals to the two groups, leading to a comparison of two binomially distributed random variables. We study the consequences of this in Section 8.

4. SOME APPROXIMATIONS

A number of aspects of the study of logistic models involve the evaluation of expectations typified in its simplest form by $E\{L_1(\mu + A)\}$, where A is a random variable of zero mean and variance σ^2. There are a number of approximations for small σ equivalent to order σ^2 but one that is likely to be better over a wider range of values. The simplest is based on Taylor expansion of $L_1(\mu + A)$ for small A and is

$$E\{L_1(\mu + A)\} = L_1(\mu) + \sigma^2 L_1''(\mu)/2, \tag{4.1}$$

where

$$L_1''(\mu) = L_1(\mu)L_0(\mu)\{L_0(\mu) - L_1(\mu)\}.$$

The second approximation is based on absorbing the correction term in (4.1) into the first by writing the approximation

$$L_1\left\{\mu + \sigma^2\{L_0(\mu) - L_1(\mu)\}/2\right\},$$

differing from (4.1) by terms of order $O(\sigma^4)$.

A third approximation is obtained less directly but is more stable for larger values of σ^2. We approximate the logistic function $L_1(x)$ by the standardized normal integral $\Phi(kx)$ for a suitable constant k; this gives a good approximation over a wide range of arguments. Then the expectation of interest is approximately $E\{\Phi(k\mu + kA)\}$ and if also A is normally distributed this expectation is itself a normal integral. On re-expressing this as a logistic function we obtain the third approximation

$$L_1\left\{\frac{\mu}{(1 + k^2\sigma^2)^{1/2}}\right\}.$$

Suitable values of k are suggested by Cox and Snell [6, p. 21–22] (1989); a compromise value over the central part of the range is $k = 0.607$. A major advantage of this third approximation is that, unlike the other two, it gives qualitatively sensible answers even for large values of σ^2.

To aid interpretation, suppose for instance that the probabilities varied with roughly 95% of values being between 0.6 and 0.9. Then the corresponding logistic function varies between 0.4 and 2.2 suggesting a σ of roughly 0.45. Then the correction factor $\sqrt{\{1 + (0.607^2 \cdot 0.45^2)\}}$ would be about 1.04, implying a quite modest adjustment.

For more detailed comparisons more explicit information about the probability that a pair is discordant is needed. We treat α_s as a random variable A, so that

$$\pi_D = E_A\{L_0(\mu + A - \theta/2)L_1(\mu + A + \theta/2) + L_1(\mu + A - \theta/2)L_0(\mu + A + \theta/2)\}. \tag{4.2}$$

This can be expanded in terms of σ by the methods outlined above. The complex details will not be given.

Table 1 shows π_D against θ, μ and σ. The calculated values of π_D were confirmed by simulation. The proportion of discordant pairs decreases rather slowly with σ.

Table 1: Probability of a pair being discordant, π_D, against θ, μ and σ.

		$\pi_D(0)$				
	θ	0	0.5	1	1.5	2
σ	μ					
	0	0.500	0.530	0.607	0.702	0.790
	0.5	0.470	0.500	0.578	0.676	0.769
	1	0.393	0.422	0.500	0.604	0.709
	1.5	0.298	0.324	0.396	0.500	0.615
		$\pi_D(\sigma)$ / $\pi_D(0)$				
0.5	0	0.94	0.94	0.95	0.96	0.97
	0.5	0.95	0.95	0.96	0.97	0.97
	1	0.98	0.98	0.97	0.97	0.98
	1.5	1.01	1.01	1.00	0.99	0.98
	2	1.05	1.04	1.03	1.01	1.00
1	0	0.75	0.76	0.80	0.85	0.90
	0.5	0.79	0.80	0.83	0.86	0.90
	1	0.91	0.90	0.90	0.89	0.90
	1.5	1.05	1.04	1.00	0.95	0.93
	2	1.19	1.17	1.11	1.04	0.98
1.5	0	0.44	0.47	0.56	0.66	0.76
	0.5	0.54	0.56	0.61	0.69	0.77
	1	0.80	0.79	0.76	0.76	0.78
	1.5	1.12	1.08	0.99	0.90	0.84
	2	1.42	1.37	1.25	1.10	0.96

Estimation of σ is in principle possible by first estimating θ and μ and then comparing the proportion of discordant observations with that to be expected in the homogenous case, $\sigma = 0$. Table 1 shows that it is only for rather large value of σ and even then for certain ranges of the other parameters that such estimation is likely to be effective.

5. UNCONDITIONAL ANALYSIS

Suppose that instead of pairing, individuals are randomized to two groups, 0 and 1, therefore with probabilities of success

$$P(Y_0 = 1) = E\left\{L_1(\mu + A - \theta/2)\right\}, \quad P(Y_1 = 1) = E\left\{L_1(\mu + A + \theta/2)\right\},$$

respectively. The resulting unconditional analysis uses all pairs.

Thus, for example, the probability of success for an individual in group 0 is approximately

$$\psi_0 \simeq L_1\left(\frac{\mu - \theta/2}{\sqrt{(1 + k^2\sigma^2)}}\right) \tag{5.1}$$

and that for an individual in group 1 is

$$\psi_1 \simeq L_1\left(\frac{\mu + \theta/2}{\sqrt{(1 + k^2\sigma^2)}}\right). \tag{5.2}$$

To estimate the marginal log odds ratio, we calculate

$$\text{logit}(\psi_1) - \text{logit}(\psi_0) = \frac{\theta}{\sqrt{(1+k^2\sigma^2)}}.$$

Thus the sample proportions can be used to obtain an unconditional estimate of θ

$$\hat{\theta}_U = \sqrt{(1+k^2\sigma^2)} \left\{ \log \frac{\hat{\psi}_1}{1-\hat{\psi}_1} - \log \frac{\hat{\psi}_0}{1-\hat{\psi}_0} \right\},$$

where in this discussion we shall treat σ as known or, more realistically, treated by sensitivity analysis.

The asymptotic variance of the estimate of the treatment effect θ in the unconditional analysis is then

$$\text{var}(\hat{\theta}_U) \simeq (1+k^2\sigma^2) \left\{ \frac{1}{m\psi_0(1-\psi_0)} + \frac{1}{n\psi_1(1-\psi_1)} \right\}, \tag{5.3}$$

which can be expressed in terms of the functions L_i. The parameter σ^2 might possibly be estimated from the proportion of discordant pairs, although the resulting precision is likely to be low.

Table 2 shows $\text{var}(\hat{\theta}_U)$ against θ, μ and σ. The variance of the estimate of θ from the unconditional analysis increases with μ. The relation between σ and the variance of the estimate of the treatment effect from the unconditional analysis is rather weak.

Table 2: $\text{var}(\hat{\theta}_U)$ against θ, μ and σ.

$\text{var}(\hat{\theta}_U)$										
		θ								
		0			2			4		
		μ			μ			μ		
		0	1	2	0	1	2	0	1	2
	0	0.020	0.025	0.048	0.025	0.034	0.068	0.048	0.068	0.152
σ	0.5	0.022	0.027	0.049	0.027	0.035	0.067	0.049	0.067	0.142
	1	0.027	0.033	0.053	0.033	0.040	0.068	0.053	0.068	0.125

6. COMPARISON OF THE EFFICIENCY OF THE CONDITIONAL AND UNCONDITIONAL ANALYSES

The variances of $\hat{\theta}_C$ and $\hat{\theta}_U$, the estimates from the conditional and unconditional analysis respectively, are next compared. The parameter θ is defined in terms of the conditional formulation so that naive estimates of the log odds ratio are not directly comparable. Of the values in Table 3, $\theta = 4$ corresponds to a quite extreme odds ratio.

Table 3: $\text{var}(\hat{\theta}_C)/\text{var}(\hat{\theta}_U)$ against θ, μ and σ.

$\text{var}(\hat{\theta}_C)/\text{var}(\hat{\theta}_U)$										
		θ								
		0			2			4		
		μ			μ			μ		
		0	1	2	0	1	2	0	1	2
	0	1	1	1	1.54	1.41	1.20	3.76	2.93	1.87
σ	0.5	0.98	0.96	0.93	1.52	1.39	1.18	3.77	3.03	2.00
	1	0.97	0.85	0.76	1.49	1.33	1.08	3.80	3.25	2.30

Table 3 shows the ratio of the variance of the estimate of θ_C to the variance of the estimate of θ_U against θ, μ and σ. The ratio $\text{var}(\hat{\theta}_C)/\text{var}(\hat{\theta}_U)$ is equal to one when $\theta = \sigma = 0$, that is for the null hypothesis with effectively random pairing. As to be expected from the matching, near $\theta = 0$ the gain from using the conditional estimate increases with σ. Especially for larger values of μ and θ, however, the unconditional estimate using the concordant pairs is to be preferred.

The values of the variances as a function of μ, σ and θ were checked by simulation. For θ_C there was good agreement and also for θ_U for small values of σ, but for large σ the calculated variance was larger than the simulated variance.

We now return to testing the hypothesis of no difference between the two groups. In comparing the conditional and unconditional analyses, it is important that the parameters used to specify departures from the null hypothesis have broadly comparable interpretations in the different formulations.

For the conditional analysis, described in Section 6, we take the test statistic to be $T_C = \log(N_{01}/N_{10})$ and in the discussion to follow of the unconditional analysis we take $T_U = \log\{(N_{\cdot 1}N_{0\cdot})/(N_{1\cdot}N_{\cdot 0})\}$.

Then T_C, interpreted as the logit difference between the two individuals in an arbitrary pair, has asymptotic expected value $E(T_C) = \theta$. At the null hypothesis we have that, asymptotically,

$$\text{var}(T_C) = \left\{ \frac{1}{\frac{1}{2}m_D} + \frac{1}{\frac{1}{2}m_D} \right\} = \frac{4}{m_D} = \frac{4}{m\pi_D},$$

where m is the number of pairs, m_D the number of discordant pairs and π_D the probability of a pair being discordant. The Pitman efficacy (Cox and Hinkley [5, p. 337–338], 1974) for testing the hypothesis that $\theta = 0$ is thus

$$\mathcal{E}_C = \frac{\{\partial E(T_C)/\partial\theta|_{\theta=0}\}^2}{m\text{var}(T_C)|_{\theta=0}} = \pi_D.$$

Under the null hypothesis, the probability of a pair being discordant, and hence also $\mathcal{E}_C$, is

$$\pi_D \simeq 2L_0(\mu)L_1(\mu)\left\{1 + \frac{1}{2}\sigma^2\left(1 - 6L_0(\mu)L_1(\mu)\right)\right\}. \tag{6.1}$$

In the unmatched analysis of Section 7 we have, for instance, that $P(Y_0 = 1) = E\{L_1(A - \theta)\}$. Such probabilities can be calculated approximately by using a Taylor expansion or by approximating $L_1(\cdot)$ by $\Phi(\cdot)$, the standard normal cumulative distribution function. For group 0 this gives

$$P(Y_0 = 1) \simeq L_1(\mu - \theta/2) + \frac{1}{2}\sigma^2 L_1(\mu - \theta/2)L_0(\mu - \theta/2)\{L_0(\mu - \theta/2) - L_1(\mu - \theta/2)\}.$$

For group 1 an analogous expression holds, with $\mu - \theta/2$ replaced by $\mu + \theta/2$. By a further approximation,

$$\text{logit}\{P(Y_0 = 1)\} \simeq \mu + \frac{1}{2}\sigma^2\{L_0(\mu - \theta/2) - L_1(\mu - \theta/2)\},$$

so that T_U, the log odds contrast in the unconditional analysis, has asymptotic expected value

$$E(T_U) = \frac{1}{4}\theta + \frac{1}{4}\sigma^2\{L_0(\mu + \theta/2) - L_1(\mu + \theta/2) - L_0(\mu - \theta/2) + L_1(\mu - \theta/2)\}.$$

Then

$$\frac{\partial E(T_U)}{\partial \theta} \simeq \frac{1}{2}\{1 - \sigma^2 L_0(\mu + \theta/2)L_1(\mu + \theta/2)\}$$

which under the null hypothesis is $\{1 - \sigma^2 L_0(\mu)L_1(\mu)\}/2$. The variance under the null hypothesis is that of the comparison of two independent logits, each based on m observations and thus is

$$\text{var}(T_U) = \frac{1}{2mL_0(\mu)L_1(\mu)}\left\{1 - \frac{1}{2}\sigma^2\left(L_0(\mu) - L_1(\mu)\right)^2\right\},$$

assuming σ^4 is negligible. Therefore the Pitman efficacy for T_U is after some simplification

$$\mathcal{E}_U \simeq \frac{L_0(\mu)L_1(\mu)}{2}\left\{1 + \frac{1}{2}\sigma^2\left(1 - 8L_0(\mu)L_1(\mu)\right)\right\}, \tag{6.2}$$

ignoring terms of order σ^4 and above.

Therefore to assess the relative efficiency for $\theta = 0$, we compare $\mathcal{E}_U$ and $\mathcal{R}_C$. Since in this special case $\mathcal{E}_U$ is smaller than $\mathcal{E}_C$, near the null hypothesis of zero treatment effect the matched design tends to be slightly more efficient than the unmatched one, as is confirmed by the comparison of the variances.

Often $L_0(\mu)L_1(\mu) \simeq 1/4$ and then

$$\mathcal{E}_C \simeq \frac{1}{2}\left(1 - \frac{1}{4}\sigma^2\right)$$

and for comparison

$$\mathcal{E}_U \simeq \frac{1}{2}\left(1 - \frac{1}{2}\sigma^2\right).$$

Thus for testing the hypothesis of no treatment effect the conditional analysis is asymptotically slightly better than the unconditional analysis, depending on the amount of variability between pairs.

7. DISCUSSION

The main qualitative aspects in this discussion, some with broader implications, are as follows. Most importantly, should conclusions be formulated in terms of differences of probabilities or as logistic differences or possibly on some other scale? For any specific set of data the choice is likely to be numerically unimportant if all probabilities are in the central range, say between 0.2 and 0.8 (Cox and Wermuth [7], 1992). The choice becomes important if several sets of data are considered together, when stability of contrasts across data sets, if achievable, is desirable. The direct interpretation of differences of probabilities in terms of the numbers of individuals notionally affected by a change in treatment is attractive but in general decreasingly so at the extremes of the scale, where the logistic comparisons, essentially log ratios at the two ends of the scale, become more appealing, especially so for case-control studies, where there are quite strong specific arguments for the use of logistic differences.

The second general issue applying to the logistic analysis of matched pairs is that the parameter of interest, a difference of log odds, is notionally specific to each pair. This implies, in particular, that it is not directly comparable with the same difference calculated from an unmatched randomized comparison of the same two treatments from the same population. The exception is when the variation between pairs is small. Otherwise some correction based on the inter-pair variability can be made, essentially using the relation between that variability and the proportion of discordant pairs, but such adjustments are likely to be quite fragile.

The third issue is that detailed comparison of the conclusions from different studies, some matched and some totally randomized, requires recognition that different ways of expressing the comparisons of interest by an unknown parameter may be involved.

Finally our detailed results show when the gain in sensitivity from matching is likely to be appreciable.

ACKNOWLEDGMENTS

We are grateful to the referee for constructive comments.

REFERENCES

[1] Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

[2] Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data, *Biometrika*, **68**, 357–363.

[3] Cox, D.R. (1958a). The regression analysis of binary sequences (with discussion), *J. R. Statist. Soc. B*, **20**, 215–232.

[4] Cox, D.R. (1958b). Two further applications of a model for binary regression, *Biometrika*, **45**, 562–565.

[5] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

[6] Cox, D.R. and Snell, E.J. (1989). *The Analysis of Binary Data*, Second Edition, Chapman and Hall, London.

[7] Cox, D.R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables, *Biometrika*, **79**, 441–461.

[8] Fisher, R.A. (1935). *Design of Experiments*, Oliver and Boyd, Edinburgh.

[9] Lee, Y. (2001). Can we recover information from concordant pairs in binary matched pairs? *J. Applied Statistics*, **28**, 239–246.

[10] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, **12**, 153–157.

[11] Pearson, K. (1907). Reply to certain criticisms of Mr. G.U. Yule, *Biometrika*, **5**, 470–476.

[12] Rasch, G. (1960). *Probability Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen.

[13] Yule, G.U. (1903). Notes on the theory of the association of attributes in statistics, *Biometrika*, **2**, 121–134.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 593–612

THE XGAMMA FAMILY: CENSORED REGRESSION MODELLING AND APPLICATIONS

Authors: Gauss M. Cordeiro
– Department of Statistics, Universidade Federal de Pernambuco, Recife, Brazil
gausscordeiro@gmail.com

Emrah Altun
– Department of Mathematics, Bartin University, Bartin, Turkey
emrahaltun@bartin.edu.tr

Mustafa Ç. Korkmaz
– Department of Measurement and Evaluation, Artvin Çoruh University, Artvin, Turkey
mcagatay@artvin.edu.tr

Rodrigo R. Pescim
– Department of Statistics, Universidade Estadual de Londrina, Londrina, Brazil
rrpescim@gmail.com

Ahmed Z. Afify
– Department of Statistics, Mathematics and Insurance, Benha University, Benha, Egypt
ahmed.afify@fcom.bu.edu.eg

Haitham M. Yousof
– Department of Statistics, Mathematics and Insurance, Benha University, Benha, Egypt
haitham.yousof@fcom.bu.edu.eg

Received: May 2019 Revised: December 2019 Accepted: December 2019

Abstract:

- In this paper, a new family of distributions with one extra shape parameter, called the xgamma-G, is proposed. comprehensive treatment of some of its mathematical properties including ordinary and incomplete moments and quantile and generating functions are derived. The unknown model parameters are estimated by the maximum likelihood method and the performance of the maximum likelihood estimators are assessed via two extensive simulation studies. Additionally, the log-location-scale regression model for censored data based on a special member of the family is introduced. The usefulness of the proposed models is illustrated utilizing three real data sets.

Key-Words:

- *censored data; maximum likelihood estimation; moment; regression model; xgamma distribution.*

AMS Subject Classification:

- 60E05, 62F10, 62N01.

1. INTRODUCTION

Statistical distributions are important tools to model the characteristics of data sets such as right or left skewness, bi-modality or multi-modality observed in different applied sciences such as engineering, medicine and finance and among others. The well-known distributions such as normal, Weibull, gamma, Lindley are extensively used because of theirs simple forms and identifiability properties. However, in the last decade, researchers have focused on the more complex and flexible distributions to increase the modeling ability of these distributions by adding one or more shape parameters. The well-known family of distributions can be cited as follows: Marshall-Olkin-G (Marshall and Olkin [19], 1997), beta-G (Eugene *et al.* [11], 2002), gamma-G (Zografos and Balakrishnan [33], 2009), type I half-logistic-G (Cordeiro *et al.* [8], 2016), Burr X-G (Yousof *et al.* [32], 2016), generalized transmuted-G (Nofal *et al.* [23], 2017) and exponentiated Weibull-H (Cordeiro *et al.* [7], 2017), among others.

Recently, Sen *et al.* [29] (2016) proposed and studied the xgamma (XG) distribution with cumulative distribution function (cdf) and probability density function (pdf) (for $\theta > 0$) given by

$$G(x;\theta) = 1 - \frac{1+\theta+\theta x+\frac{1}{2}\theta^2 x^2}{1+\theta}\exp(-\theta x),\ x>0 \tag{1.1}$$

and

$$g(x;\theta) = \frac{\theta^2}{1+\theta}\left(1+\frac{\theta}{2}x^2\right)\exp(-\theta x), \tag{1.2}$$

respectively. During the recent years, the xgamma distribution has been shown great interest by researchers. Altun and Hamedani [2] (2018) introduced a new bounded distribution using the transformation $Y = \exp(-X)$ as an alternative to the beta distribution based on the xgamma distribution. Biçer [4] (2019) introduced the transmuted-xgamma distribution and studied its statistically properties comprehensively. The another generalization of xgamma distribution was provided by Sen *et al.* [27] (2018a) on the basis of special mixture of exponential and gamma distributions. Sen *et al.* [28] (2018b) studied the parameter estimation of xgamma distribution under progressively type-II right censoring scheme by maximum likelihood and Bayesian estimation methods. Sen and Chandra [25] (2017) introduced the quasi-xgamma distribution by using the xgamma distribution as a baseline distribution. The weighted generalization of xgamma distribution, using $w(x) = x^r$ as a weighting function, was studied by Sen *et al.* [26] (2017).

In this paper, we introduce and study a new class of distributions called the *xgamma-G* (XG-G) family. The idea is to incorporate any distribution into a larger family through an application of the XG cdf. In fact, based on the T–X transform defined by Alzaatreh *et al.* [3] (2013) and the XG cdf, we construct the XG-G family. The some of its mathematical properties are provided comprehensively. The new family has flexible shapes to model various lifetime data sets. Additionally, its special models produce better fits than other well-known families.

To this end, we define the cdf of the XG-G family with one extra shape parameter $\theta > 0$ by

$$\begin{aligned} F(x;\theta,\boldsymbol{\xi}) &= \frac{\theta^2}{1+\theta}\int_0^{-\log \overline{G}(x;\boldsymbol{\xi})}\left(1+\frac{\theta}{2}t^2\right)\exp\left(-\theta t\right)\,dt \\ &= 1-\frac{1+\theta-\theta\log\overline{G}(x;\boldsymbol{\xi})+\frac{1}{2}\theta^2\left[\log\overline{G}(x;\boldsymbol{\xi})\right]^2}{1+\theta}\overline{G}(x;\boldsymbol{\xi})^{\theta}, \end{aligned} \tag{1.3}$$

where $\overline{G}(x;\boldsymbol{\xi}) = 1 - G(x;\boldsymbol{\xi})$ and $G\left(x;\boldsymbol{\xi}\right)$ is a baseline cdf with a parameter vector $\boldsymbol{\xi}$.

The pdf corresponding to (1.3) reduces to

$$f(x;\theta,\boldsymbol{\xi}) = \frac{\theta}{1+\theta}g(x;\boldsymbol{\xi})\overline{G}(x;\boldsymbol{\xi})^{\theta-1}\left\{\theta+\frac{1}{2}\theta^2\left[\log\overline{G}(x;\boldsymbol{\xi})\right]^2\right\}, \tag{1.4}$$

where $g(x;\boldsymbol{\xi}) = dG(x;\boldsymbol{\xi})/dx$. If the random variable (rv) T has the xgamma distribution (1), then $X = G^{-1}\left[1-\exp\left(-T\right)\right]$ follows the XG-G family (4). Henceforth, we denote by $X \sim$XG-G$(\theta,\boldsymbol{\xi})$ a rv having density (1.4). The hazard rate function (hrf) of X is given by

$$\tau\left(x;\theta,\boldsymbol{\xi}\right) = \frac{\theta r(x;\boldsymbol{\xi})\left\{\theta+\frac{1}{2}\theta^2\left[\log\overline{G}(x;\boldsymbol{\xi})\right]^2\right\}}{\left\{1+\theta-\theta\log\overline{G}(x;\boldsymbol{\xi})+\frac{1}{2}\theta^2\left[\log\overline{G}(x;\boldsymbol{\xi})\right]^2\right\}}.$$

The identifiability is an important property of the statistical distributions to satisfy the precise inference for the model parameters. The following theorem is given to prove the identifiability property of XG-G family.

Theorem 1.1. *The cdf* (1.3) *is identifiable.*

Proof: Assume that the baseline cdf $G\left(x;\boldsymbol{\xi}\right)$ is identifiable. The cdf (1.3) is identifiable once $F\left(x;\theta_1\right) = F\left(x;\theta_2\right)$ is valid if and only if $\theta_1 = \theta_2$. Using (1.3), we have

$$\begin{aligned} &F\left(x;\theta_1\right) = F\left(x;\theta_2\right) \\ &1-\left(1+\theta_1-\theta_1 A+\tfrac{1}{2}\theta_1^2A^2\right)\left(1+\theta_1\right)^{-1}\exp\left(A\theta_1\right) \\ &= 1-\left(1+\theta_2-\theta_2 A+\tfrac{1}{2}\theta_2^2A^2\right)\left(1+\theta_2\right)^{-1}\exp\left(A\theta_2\right) \end{aligned} \tag{1.5}$$

where $A = \log\bar{G}\left(x\right)$. (1.5) can be simplified as follows

$$\begin{aligned} &\left[\frac{\exp\left(A\theta_2\right)}{\left(1+\theta_2\right)}-\frac{\exp\left(A\theta_1\right)}{\left(1+\theta_1\right)}\right]+\left[\frac{\exp\left(A\theta_2\right)\theta_2}{\left(1+\theta_2\right)}-\frac{\exp\left(A\theta_1\right)\theta_1}{\left(1+\theta_1\right)}\right] \\ &-\left[\frac{\exp\left(A\theta_2\right)\theta_2 A}{\left(1+\theta_2\right)}-\frac{\exp\left(A\theta_1\right)\theta_1}{\left(1+\theta 1\right)}\right]+\left[\frac{\exp\left(A\theta_2\right)\theta_2^2A^2}{2\left(1+\theta_2\right)}-\frac{\exp\left(A\theta_1\right)\theta_1^2A^2}{2\left(1+\theta_1\right)}\right]=0 \end{aligned} \tag{1.6}$$

The expression (1.6) is equal to zero for all x only when the parameters $\theta_1 = \theta_2$. Since the parameter $\theta > 0$, it is concluded that the model is identifiable: $F\left(x;\theta_1\right) = F\left(x;\theta_2\right) \Leftrightarrow \theta_1 = \theta_2$. □

The purpose of the generation of the XG-G family is to provide new opportunities to model the different characteristics of the data sets such as left skewness, excess kurtosis and bathtub failure rate. The well-known distributions are insufficient to model these kinds of

data sets. The special members of the XG-G family can be used to model skewed and long-tailed data sets to improve the modeling accuracy of interested data set with only one extra shape parameter. Moreover, the proposed family is highly effective in modeling the censored lifetimes of individuals with some covariates in a location-scale regression framework.

The remaining part of the paper is organized as follows. In Section 2, three special cases of the XG-G family are given. In Section 3, a linear representation of the XG-G density is provided. The comprehensive mathematical properties of the XG-G density are obtained and reported in Section 4. Section 5 is devoted to the maximum likelihood estimation of the model parameters for uncensored and censored data. In Section 6, we present a new log-location-scale regression model based on the log XG-Weibull distribution. Section 7 deals with simulation studies to evaluate the maximum likelihood estimators of the parameters of proposed models. In Section 8, three applications to the real data sets are given to prove empirically the importance of XG-G family. Section 9 contains the concluding remarks of the study.

2. SOME SPECIAL XG-G MODELS

2.1. The XG-Lindley (XG-Li) model

Consider the cdf $G(x) = 1 - \frac{1+a+ax}{1+a}\exp(-ax)$ of the Li distribution with scale parameter $a > 0$. The XG-Li density (for $x > 0$) can be determined from (1.4). Some plots of the XG-Li density and hazard functions for selected parameter values are displayed in Figure 1. These plots reveal that the pdf of the XG-Li model can be reversed J-shape, right skewed or unimodal. The hrf can be unimodal or bathtub.

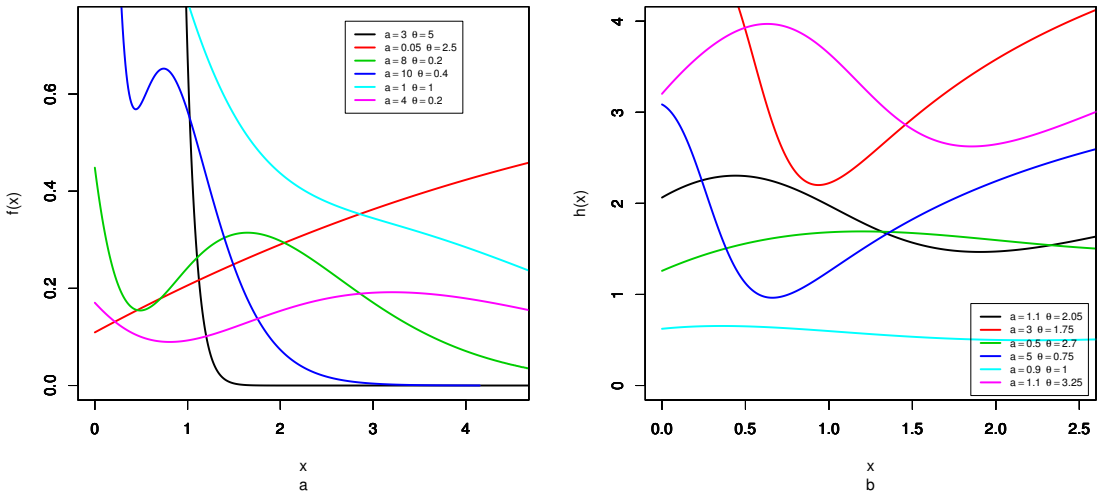


Figure 1: Plots of the XG-Li pdf (left) and hrf (right) for some parameter values.

2.2. The XG-Weibull (XG-W) model

Consider the cdf $G(x) = 1 - \exp[-(ax)^b]$ of the W distribution with scale $a > 0$ and shape $b > 0$. The pdf of the XG-W model (for $x > 0$) follows from (1.4). Some plots of the XG-W pdf and hrf for selected parameter values are displayed in Figure 2. Figure 2 reveals that the XG-W density can be concave down, left skewed or right skewed. The hrf of the XG-W model can be increasing, decreasing, bathtub or unimodal.

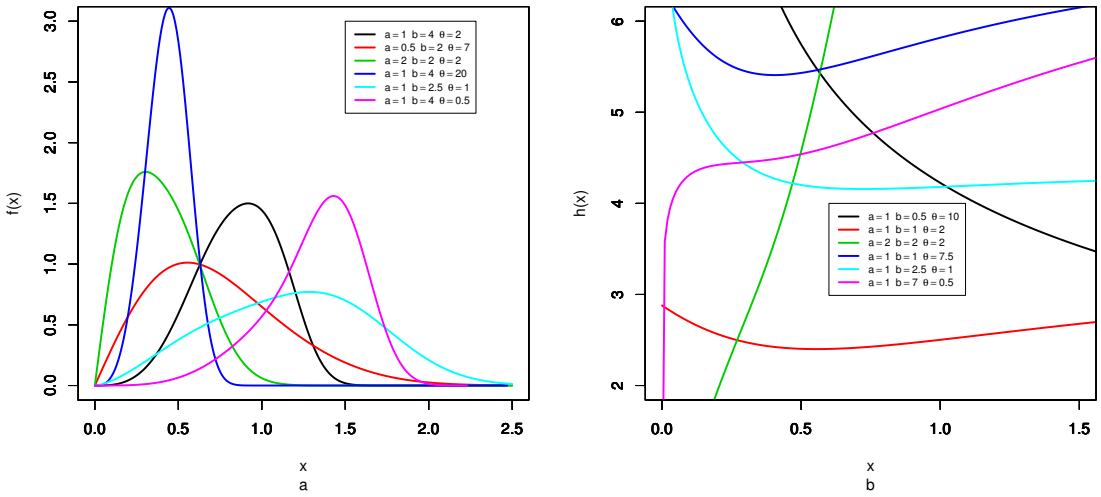


Figure 2: Plots of the XG-W pdf (left) and hrf (right) for some parameter values.

2.3. The XG-BurrXII (XG-BXII) model

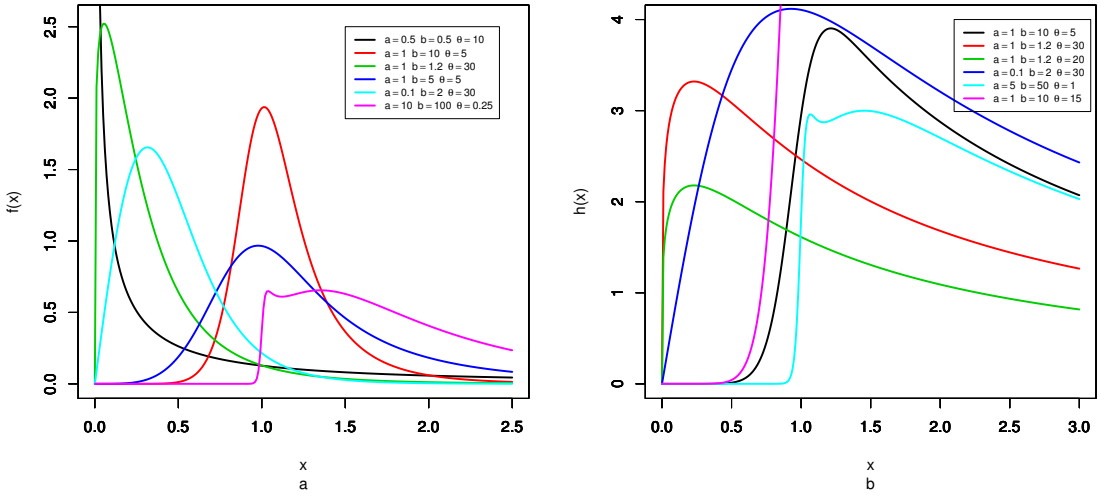


Figure 3: Plots of the XG-BXII pdf (left) and hrf (right) for some parameter values.

Consider the cdf $G(x) = 1 - (1 + x^a)^{-b}$ of the BXII distribution with parameters $a > 0$ and $b > 0$. The pdf of the XG-BXII model (for $x > 0$) can be obtained from (1.4). Some plots of the XG-BXII pdf and hrf for selected parameter values are displayed in Figure 3. These plots reveal that the pdf of the XG-BXII model can be reversed J-shape, concave down or right skewed. Its hrf can be increasing or unimodal.

3. USEFUL REPRESENTATION OF PDF AND CDF

The XG-G family density in (1.4) can be expressed as

$$f(x) = \frac{\theta^2 g(x)}{1+\theta}\overline{G}(x)^{\theta-1} + \frac{\theta^3 g(x)}{2(1+\theta)}\overline{G}(x)^{\theta-1}\underbrace{[\log \overline{G}(x)]}_{A}^2.$$

Consider

$$\log(1-z) = -\sum_{i=0}^{\infty}\frac{z^{i+1}}{i+1}, \ |z| < 1, \tag{3.1}$$

and the power series raised to a positive integer n (Gradshteyn and Ryzhik [14, Section 0.314], 2002)

$$\left(\sum_{j=0}^{\infty} a_j\, u^j\right)^n = \sum_{j=0}^{\infty} c_{n,j}\, u^j, \tag{3.2}$$

where the coefficients $c_{n,j}$ (for $j = 1, 2, ...$) can be easily determined from the recurrence equation

$$c_{n,j} = (j a_0)^{-1}\sum_{m=1}^{j}[m(n+1) - j]\, a_m\, c_{n,j-m} \quad \text{and} \quad c_{n,0} = a_0^n.$$

The coefficient $c_{n,j}$ can be calculated from $c_{n,0}, ..., c_{n,j-1}$ and hence from the quantities $a_0, ..., a_j$. For $|z| < 1$ and $b > 0$, the power series holds

$$(1-z)^{b-1} = \sum_{k=0}^{\infty}\frac{(-1)^k\,\Gamma(b)}{k!\,\Gamma(b-k)}\, z^k. \tag{3.3}$$

Applying (3.1) to the quantity A gives

$$f(x) = \frac{\theta^2 g(x)}{1+\theta}\overline{G}(x)^{\theta-1} + \frac{\theta^3 g(x)}{2(1+\theta)}\overline{G}(x)^{\theta-1}G(x)^2\underbrace{\left[\sum_{i=0}^{\infty}\frac{G(x)^i}{i+1}\right]^2}_{B}.$$

Next, the quantity B follows using (3.2) as

$$f(x) = \frac{\theta^2 g(x)}{1+\theta}\underbrace{\overline{G}(x)^{\theta-1}}_{C} + \frac{\theta^3 g(x)}{2(1+\theta)}\sum_{i=0}^{\infty} c_{2,i}\, G(x)^{i+2}\underbrace{\overline{G}(x)^{\theta-1}}_{C},$$

where $a_i = 1/(i+1)$.

Applying the power series (3.3) to the quantity C, we obtain

$$f(x) = \sum_{k=0}^{\infty} \left[b_k \ \pi_{k+1}(x) + \sum_{i=0}^{\infty} b_{i,k} \ \pi_{i+k+3}(x) \right], \tag{3.4}$$

where

$$b_k = \frac{(-1)^k \, \theta^2 \Gamma(\theta)}{(k+1)\,(1+\theta)\,\Gamma(\theta-k)}, \quad b_{i,k} = \frac{(-1)^k \, \theta^3 \Gamma(\theta) \ c_{2,i}}{2(1+\theta)(i+k+3)\,k!\,\Gamma(\theta-k)},$$

and $\pi_\alpha(x) = \alpha \, g(x) \, G(x)^{\alpha-1}$ is the exponentiated-G (Exp-G) density function with power parameter $\alpha > 0$. So, the density of X is a linear combination of Exp-G densities.

The properties of Exp-G distributions have been studied by many authors in recent years, see, for example, Mudholkar and Srivastava [20] (1993) and Mudholkar *et al.* [21] (1995) for exponentiated Weibull (EW), Gupta and Kundu [16] (1999) for exponentiated exponential and Nadarajah and Gupta [22] (2007) for exponentiated gamma, among others.

The cdf of X follows by integrating (3.4) as

$$F(x) = \sum_{k=0}^{\infty} \left[b_k \ \Pi_{k+1}(x) + \sum_{i=0}^{\infty} b_{i,k} \ \Pi_{i+k+3}(x) \right], \tag{3.5}$$

where $\Pi_\alpha(x) = G(x)^\alpha$ is the Exp-G cdf with power parameter α. Equations (3.4) and (3.5) are the main results of this section.

4. PROPERTIES

In this section, we investigate some mathematical properties of the XG-G family.

4.1. Quantile function

The quantile function (qf) of X can be determined by inverting $F(x) = u$ in (1.3). We require numerical methods to obtain the quantiles. For given u, we solve numerically for $z = z(u)$ in the equation

$$\left[1 + \theta - \theta \log(z) + 0.5\,\theta^2 \ \log^2(z)\right] \ z^\theta = (1+\theta)(1-u),$$

and then $x = Q(u) = G^{-1}(1-z)$ is a variate from the XG-G family (1.4).

4.2. Moments

Let Y_α be a rv having density $\pi_\alpha(x)$. The r-th ordinary moment of X, say μ'_r, follows from (3.4) as

$$\mu'_r = E(X^r) = \sum_{k=0}^{\infty} \left[b_k \ E\left(Y_{k+1}^r\right) + \sum_{i=0}^{\infty} b_{i,k} \ E\left(Y_{i+k+3}^r\right) \right], \tag{4.1}$$

where $E(Y_\alpha^r) = \alpha \int_{-\infty}^{\infty} x^r \; g(x)\, G(x)^{\alpha-1} \; dx$ can be evaluated numerically in terms of the baseline qf $Q_G(u) = G^{-1}(u)$ as $E(Y_\alpha^n) = \alpha \int_0^1 Q_G(u)^n \; u^{\alpha-1} du$. Setting $r = 1$ in (4.1) gives the mean of X. Table 1 lists the first three ordinary moments of XG-W distribution. The results given in this table show that when the parameter θ increases, the ordinary moments of XG-W decrease for fixed a and b parameters.

Table 1: Moments of XG-W distribution for several parameter values.

Parameters			$\mu_1^{'}$	$\mu_2^{'}$	$\mu_3^{'}$
θ	a	b			
2	2	2	1.619	3.333	7.990
2	2	1	0.809	0.833	0.999
2	2	0.5	0.405	0.208	0.125
2	1	0.5	0.417	0.333	0.375
2	0.5	0.5	0.667	2.125	14.035
1	0.5	0.5	3.500	48.000	1304.865
0.5	0.5	0.5	17.231	953.497	94367.230

4.3. Incomplete moments

The r-th incomplete moment of X is given by

$$m_r(y) = \int_{-\infty}^{y} x^r \, f(x) dx. \tag{4.2}$$

Using (3.4), the r-th incomplete moment of XG-G family is

$$m_r(y) = \sum_{k=0}^{\infty} \left[b_k \; m_{r,k+1}(y) + \sum_{i=0}^{\infty} b_{i,k} \; m_{r,i+k+3}(y) \right],$$

where $m_{r,\alpha}(y) = \int_0^{G(y)} Q_G^r(u)\, u^{\alpha-1} \; du$. The $m_{r,\alpha}(y)$ can be calculated numerically by using the software such as **Matlab**, **R**, **Mathematica** etc. The incomplete moments of the XG-W distribution are given in Table 2. As seen from the results given in Table 2, the incomplete moments of XG-W distribution increases for fixed a and b parameters when the parameter θ increases.

Table 2: Incomplete moments of XG-W distribution for several parameter values.

Parameters			$\mu_1(0.5)$	$\mu_1(1)$	$\mu_1(2)$
θ	a	b			
2	2	2	0.026	0.170	0.790
2	2	1	0.085	0.395	0.799
2	2	0.5	0.197	0.400	0.405
2	1	0.5	0.135	0.293	0.406
2	0.5	0.5	0.080	0.168	0.313
1	0.5	0.5	0.051	0.129	0.320
0.5	0.5	0.5	0.022	0.060	0.166

4.4. Moment generating function

The moment generating function (mgf) of X, say $M(t) = E(\mathrm{e}^{t\,X})$, is obtained from (3.4) as

$$M(t) = \sum_{k=0}^{\infty} \left[b_k \; M_{k+1}\,(t) + \sum_{i=0}^{\infty} b_{i,k} \; M_{i+k+3}\,(t) \right],$$

where $M_\alpha(t)$ is the generating function of Y_α given by

$$M_\alpha(t) = \alpha \int_{-\infty}^{\infty} \mathrm{e}^{t\,x} G(x) g(x)^{\alpha-1}\, dx = \alpha \int_0^1 \exp[t\, Q_G(u;\alpha)]\, u^{\alpha-1} du.$$

The last two integrals can be computed numerically for most parent distributions.

5. ESTIMATION

This section deals with the maximum likelihood estimation of the unknown model parameters.

5.1. Maximum likelihood estimation

Let $x_1, \cdots, x_n$ be a random sample from the XG-G models with a parameter vector $\mathbf{\Phi} = (\theta, \xi^\intercal)^\intercal$. The log-likelihood function is given by

$$\begin{aligned} \ell_n(\mathbf{\Phi}) \;=\; & n \log\theta - n \log(1+\theta) + \sum_{i=1}^{n} \log g(x_i;\xi) + (\theta-1) \sum_{i=1}^{n} \log \overline{G}(x_i;\xi) \\ & + \sum_{i=1}^{n} \log \left\{ \theta + \frac{1}{2}\theta^2 \left[\log \overline{G}(x_i;\xi) \right]^2 \right\}. \end{aligned}$$

Taking the partial derivatives of the log-likelihood function concerning the parameters, we obtain the score vectors. The simultaneous solution of these equations for zero gives the maximum likelihood estimate of $\mathbf{\Phi}$. Since it is not possible to obtain closed-form expressions of the maximum likelihood estimators of the parameters of XG-G family, direct maximization of the log-likelihood is needed. In this study, the **optim** function of **R** software is used to minimize the minus of the log-likelihood function which is equivalent to the maximization of log-likelihood.

5.2. Multi-censored maximum likelihood estimation

Censored data are often encountered in survival analysis and reliability studies. Here, the general case of multi-censored data is considered. Assume that m_0 subjects of m are failed at the times $x_1, \cdots, x_{m_0}$, m_1 subjects of m are failed in (s_{j-1}, s_j) interval where $j = 1, ..., m_1$ and m_2 subjects of m survived until a time r_j, $j = 1, ..., m_2$. Note that $m_0 + m_1 + m_2 = m$. The log-likelihood function for $\mathbf{\Phi}$ is

$$\begin{aligned}\ell_m(\mathbf{\Phi}) \;=\; & m_0 \log\theta - m_0 \log(1+\theta) + \sum_{i=1}^{m_0} \log g(x_i,\xi) \\ & + (\theta-1)\sum_{i=1}^{m_0} \log \overline{G}(x_i,\xi) + \sum_{i=1}^{m_0} \log\left\{\theta + \frac{1}{2}\theta^2 \left[\log \overline{G}(x_i,\xi)\right]^2\right\} \\ & + \sum_{i=1}^{m_2} \log\left\{\frac{1}{1+\theta}\left[1+\theta-\theta\log t_{r_i} + \frac{(\log t_{r_i})^2}{2\theta^{-2}}\right] t_{r_i}^{\theta}\right\} \\ & + \sum_{i=1}^{m_1} \log\left(\left\{1 - \frac{1}{1+\theta}\left[1+\theta-\theta\log t_{s_i} + \frac{(\log t_{s_i})^2}{2\theta^{-2}}\right] t_{s_i}^{\theta}\right\}\right. \\ & \left. - \left\{1 - \frac{1}{1+\theta}\left[1+\theta-\theta\log t_{s_{i-1}} + \frac{(\log t_{s_{i-1}})^2}{2\theta^{-2}}\right] t_{s_{i-1}}^{\theta}\right\}\right),\end{aligned}$$

where $t_{r_i} = \overline{G}(r_i,\xi)$, $t_{s_i} = \overline{G}(s_i,\xi)$, $t_{s_{i-1}} = \overline{G}(s_{i-1},\xi)$ and the normal equations are available before.

6. THE LXG-W REGRESSION MODEL FOR CENSORED DATA

Let X be a rv having the XG-W density function. The rv $Y = \log(X)$ defines the *log-xgamma Weibull* (LXG-W) distribution. Let $a = \mathrm{e}^{-\mu}$ and $b = \sigma^{-1}$. Then, the pdf of Y (for $y \in \Re$) is given by

$$f(y) = \frac{\theta}{\sigma(1+\theta)} \exp\left[(1-\theta)\left(\frac{y-\mu}{\sigma}\right)\right]\left\{\theta + \frac{\theta^2}{2}\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]^2\right\}, \tag{6.1}$$

where $\mu \in \Re$, $\sigma > 0$ and $\theta > 0$. If Y is a rv having density function (6.1), we can write $Y \sim$ LXG-W(θ, μ, σ). For $\sigma = 1$, the LXG-W distribution reduces to the log-xgamma-exponential (LXG-E) distribution. The survival function (sf) corresponding to (6.1) is given by

$$S(y) = \frac{1}{1+\theta}\left[(1+\theta) + \left(\theta - \frac{\theta^2}{2}\right)\exp\left(\frac{y-\mu}{\sigma}\right)\right]\left\{\exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\theta}.$$

We define the standardized rv $Z = (Y - \mu)/\sigma$ with pdf (for $z \in \Re$) given by

$$f(z) = \frac{\theta}{(1+\theta)} \exp\left[(1-\theta)z\right] \left[\theta + \frac{\theta^2}{2} \exp\left(2z\right)\right]. \tag{6.2}$$

Regression models are widely used to model dependent variable with some covariates. The lifetimes of individuals are generally effected by some explanatory variables such as gender, age, alcohol abuse or smoking. To model these kind of data sets, we propose a new log-location-scale regression model based on the LXG-W density. Let yi be the response variable and $\mathbf{v}_i^T = (v_{i1}, ..., v_{ip})$ is the explanatory variable vector, we consider the following regression model

$$y_i = \mathbf{v}_i^T \beta + \sigma\, z_i,\ i = 1, ..., n. \tag{6.3}$$

where y_i follows the LXG-W density with unknown parameters $\mu_i \in \Re$, $\theta > 0$, and $\sigma > 0$. The location of y_i, μ_i, is modeled by using the identity link function, $\mu_i = \mathbf{v}_i^T \beta$. The vector $\mu = (\mu_1, ..., \mu_n)^T$ is defined as $\mu = \mathbf{V}\beta$, where $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_n)^T$ is a known model matrix.

Let the random sample $(y_1, \mathbf{v}_1), ..., (y_n, \mathbf{v}_n)$ are independent and the response variable is defined as $y_i = \min\{\log(t_i), \log(c_i)\}$. Assume that the lifetimes and censoring times are independent. F and C represent the sets of individuals for the log-lifetime and log-censoring, respectively. The log-likelihood function for the vector of parameters $\eta = (\beta^T, \theta, \sigma)^T$ is given by

$$\begin{aligned} l(\eta) \;=\; & r\, \log\left[\frac{\theta}{\sigma\,(1+\theta)}\right] + (1-\theta) \sum_{i \in F} \frac{y_i - \mathbf{v}_i^T \beta}{\sigma} \\ & + \sum_{i \in F} \log\left\{\theta + \frac{\theta^2}{2}\left[-\exp\left(\frac{y_i - \mathbf{v}_i^T \beta}{\sigma}\right)\right]^2\right\} \\ & + c\, \log\left(\frac{1}{1+\theta}\right) - \theta \sum_{i \in C} \exp\left(\frac{y_i - \mathbf{v}_i^T \beta}{\sigma}\right) \\ & + \sum_{i \in C} \log\left[(1+\theta) + \left(\theta - \frac{\theta^2}{2}\right) \exp\left(\frac{y_i - \mathbf{v}_i^T \beta}{\sigma}\right)\right], \end{aligned} \tag{6.4}$$

where r and c are the number of uncensored (failures) and censored observations. The parameter vector, η, of the LXG-W regression model is estimated by minimizing the minus of log-likelihood function, given in (6.4). To do this, the **optim** function of R software is used. The inverse of the observed information matrix is used to obtain corresponding standard errors and construct 95% asymptotic confidence intervals of the parameters. The observed information matrix is evaluated numerically at $\hat{\eta}$ by **hessian** function of R software.

7. SIMULATION STUDIES

In this section, three simulation studies are given to evaluate the finite sample performance of the parameters of proposed models.

7.1. Simulations of XG-W and XG-N distributions

Here, we perform two simulation studies using the XG-W and XG-normal (XG-N) distributions. To verify the performance of the MLEs for these distributions, we generate 1,000 samples of sizes 20, 50 and 100 from their qfs by inverting the cdfs. The simulation results are reported in Tables 3 and 4. These results reveal that the mean estimates become closer to the true parameter values when the sample size increases, whereas the standard errors of the estimates decrease.

The cdfs of the XG-W and XG-N distributions are given here for convenience

$$F(x) = 1 - \frac{1 + \theta + \theta\,(ax)^{\,b} + \frac{\theta^2}{2}(ax)^{2b}}{1+\theta}\exp\left[-\theta(ax)^{b}\right]$$

and

$$F(x) = 1 - \frac{1 + \theta - \theta\log\left[1 - \mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right] + \frac{1}{2}\theta^2\left\{\log\left[1 - \mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]\right\}^2}{1+\theta} \times \left[1 - \mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{\theta},$$

respectively, where $x\,, \mu \in \Re, \theta, \sigma > 0$.

Table 3: Empirical means and standard errors (in parentheses) for different values of the XG-W parameters.

Parameters		$n = 20$			$n = 50$			$n = 100$	
a, b, θ	$\widehat{a}$	$\widehat{b}$	$\widehat{\theta}$	$\widehat{a}$	$\widehat{b}$	$\widehat{\theta}$	$\widehat{a}$	$\widehat{b}$	$\widehat{\theta}$
5, 5, 5	5.3196	5.2006	4.7679	5.2329	4.8741	4.9432	5.0876	4.9086	5.0471
	(0.5927)	(0.9124)	(1.5578)	(0.5339)	(0.5449)	(1.5381)	(0.1677)	(0.3679)	(0.3703)
50,3,3	50.6839	3.0894	3.2746	49.5001	2.9022	3.1821	49.9190	2.9504	3.1318
	(1.9717)	(0.6324)	(5.0238)	(1.9357)	(0.3407)	(0.5332)	(1.8658)	(0.2525)	(0.4192)
3,3,50	3.3019	3.2521	50.0152	3.0971	3.0684	50.0124	3.0622	3.0433	49.9866
	(0.7011)	(0.6395)	(0.1291)	(0.4194)	(0.3396)	(0.1232)	(0.2885)	(0.2412)	(0.3244)
3,10,3	3.0714	10.3185	2.9319	3.0328	9.8262	3.0470	3.0203	9.8692	3.0153
	(0.1308)	(1.3464)	(0.6916)	(0.0526)	(0.6161)	(0.1897)	(0.0460)	(0.5184)	(0.1689)
50,10,50	51.0354	10.6994	50.1005	50.4395	10.2804	50.0442	50.3066	10.1571	50.0298
	(3.2573)	(2.0583)	(0.3469)	(2.0441)	(1.1738)	(0.2420)	(1.4197)	(0.7910)	(0.1551)
0.01,2,5	0.0107	1.9949	4.9998	0.0106	1.9961	4.9999	0.0105	1.9970	5.0001
	(0.0013)	(0.1175)	(0.0013)	(0.0008)	(0.0391)	(0.0003)	(0.0005)	(0.0264)	(0.0004)
1,1,1	0.9338	1.1141	1.2913	0.9526	1.0424	1.1839	1.0430	1.0327	1.0584
	(0.5201)	(0.2313)	(0.5837)	(0.3619)	(0.1177)	(0.4090)	(0.3041)	(0.1023)	(0.3542)
1,2,3	1.1736	2.0477	2.9188	1.1361	1.9665	3.0585	1.0643	1.9385	3.0133
	(0.4524)	(0.3743)	(0.7855)	(0.3183)	(0.2056)	(0.7672)	(0.1287)	(0.1538)	(0.3812)
2,2,2	2.2538	1.9168	2.4294	2.0430	1.9369	2.3574	1.9826	1.9603	2.3360
	(0.8596)	(0.3450)	(1.2688)	(0.5113)	(0.2314)	(0.7653)	(0.4837)	(0.1494)	(0.7579)
5,0.9,5	5.6190	0.9244	5.5789	5.2117	0.8744	5.2745	5.1839	0.9059	5.1761
	(0.8374)	(0.1629)	(1.0001)	(0.5833)	(0.0824)	(0.5583)	(0.4753)	(0.0792)	(0.4106)
0.025,0.9,1	0.0271	0.9142	1.0044	0.0254	0.9081	0.9965	0.0253	0.8999	0.9968
	(0.0131)	(0.1036)	(0.0730)	(0.0041)	(0.0767)	(0.0365)	(0.0040)	(0.0485)	(0.0540)

Table 4: Empirical means and standard errors (in parentheses) for different values of the XG-N parameters.

Parameters		$n = 20$			$n = 50$			$n = 100$	
θ, μ, σ	$\widehat{\theta}$	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\theta}$	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\theta}$	$\widehat{\mu}$	$\widehat{\sigma}$
5,0,1	4.9748	-0.0626	0.9531	5.0029	-0.0163	0.9829	5.0023	-0.0111	0.9918
	(0.7832)	(0.2480)	(0.1611)	(0.4507)	(0.1539)	(0.0968)	(0.2156)	(0.0966)	(0.0674)
1,0,1	0.9254	0.0846	0.9605	1.0915	0.0519	0.9791	1.0126	0.1369	1.0126
	(0.3459)	(0.6632)	(0.1990)	(0.5446)	(0.5433)	(0.1819)	(0.1570)	(0.4313)	(0.1570)
5,-1,1	4.8906	-1.0657	0.9416	5.0272	-1.0172	0.9850	4.9931	-1.0043	0.9922
	(0.7691)	(0.2425)	(0.1598)	(1.0933)	(0.1696)	(0.1023)	(0.1220)	(0.0951)	(0.0685)
5,-1,2	4.5540	-1.4028	1.8114	4.8770	-1.1455	1.9313	5.0085	-1.0687	1.9668
	(1.5270)	(0.7732)	(0.3485)	(1.1746)	(0.5544)	(0.2326)	(1.2758)	(0.5046)	(0.1903)
1,0,2	1.1759	0.2696	1.9480	1.0623	0.0886	1.9207	1.0115	0.0078	2.0198
	(0.6231))	(1.0898)	(0.4291)	(0.5966)	(1.0551)	(0.3179)	(0.3250)	(0.1924)	(0.2135)
5,0.25,0.5	4.9820	0.2220	0.4792	5.0115	0.2426	0.4914	4.9943	0.2457	0.4948
	(0.2254)	(0.0996)	(0.0719)	(0.2712)	(0.0716)	(0.0476)	(0.0751)	(0.0444)	(0.0340)
1,1,1	1.2206	1.1946	0.9452	1.1730	1.1117	0.9606	0.9647	1.0068	0.9741
	(0.5303)	(0.5923)	(0.1978)	(0.6568)	(0.6751)	(0.1619)	(0.1463)	(0.0056)	(0.1575)
50,5,5	50.0425	4.5608	4.8112	49.9274	4.7581	4.9854	49.9130	4.9392	4.9764
	(0.5333)	(1.6910)	(0.8088)	(0.9864)	(1.0880)	4.9854	49.9130	4.9392	4.9764
4,-50,10	4.2033	-50.6667	9.4225	4.0616	-50.5333	9.6902	4.0208	-49.8745	10.0017
	(1.4918)	(2.5538)	(1.4512)	(1.5078)	(2.4856)	(1.1384)	(1.0730)	(2.2228)	(0.8002)
0.9,0,0.01	0.8998	0.00042	0.009603	0.9001	0.0001	0.0098	0.9000	0.0000	0.0099
	(0.0000)	(0.0032)	(0.0014)	(0.0000)	(0.0019)	(0.0009)	(0.0000)	(0.0014)	(0.0006)
0.9,50,10	0.9594	50.2913	10.2145	0.9038	50.0927	9.8318	0.9012	50.0459	9.9480
	(0.3181)	(1.5910)	(1.3660)	(0.1473)	(1.1424)	(1.2016)	(0.1046)	(0.9813)	(0.7806)

7.2. Simulation of the LXG-W regression model

Table 5: Simulation results of LXG-W regression model.

Censoring rate=0.10		n=50			n=200			n=500	
Parameters	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
θ	2.4958	0.4958	0.7937	2.2815	0.2815	0.4029	2.0743	0.0743	0.1866
σ	0.5077	0.0077	0.0061	0.5090	0.0090	0.0028	0.5048	0.0048	0.0010
β_0	1.9266	-0.0734	0.4430	1.9563	-0.0437	0.2712	1.9870	-0.0130	0.1184
β_1	1.9992	-0.0008	0.0215	1.9997	-0.0003	0.0054	2.0010	0.0010	0.0020
Censoring rate=0.20		n=50			n=200			n=500	
Parameters	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
θ	2.3648	0.3648	0.4329	2.1059	0.1059	0.3109	2.0528	0.0528	0.0722
σ	0.5059	0.0059	0.0030	0.5120	0.0120	0.0008	0.5147	0.0147	0.0001
β_0	1.9669	-0.0331	0.2197	1.9776	-0.0224	0.0840	1.8440	-0.1560	0.0080
β_1	2.0047	0.0047	0.0135	2.0037	0.0037	0.0020	1.9994	-0.0006	0.0001
Censoring rate=0.30		n=50			n=200			n=500	
Parameters	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
θ	2.2911	0.2911	0.6217	2.1508	0.1508	0.1559	1.9377	-0.0623	0.0262
σ	0.5011	0.0011	0.0018	0.5077	0.0077	0.0026	0.5084	0.0084	0.0014
β_0	2.0236	0.0236	0.1682	2.0620	0.0620	0.3145	1.9769	-0.0231	0.1693
β_1	2.0030	0.0030	0.0100	1.9962	-0.0038	0.0061	2.0009	0.0009	0.0027

The simulation study is given to evaluate the MLEs of the parameters of LXG-W regression model. The three censoring rates (10%, 20%, 30%) and sample sizes ($n = 50, 200, 500$) are used. The simulation replication is $N = 1,000$. The lifetimes are generated by using the quantile function of the LXG-W distribution. The following parameter vector is used: $(\theta = 2, \sigma = 0.5, \beta_0 = 2, \beta_1 = 2)$. For each generated sample sizes, the biases, average of estimates (AEs) and MSEs are calculated. The simulation results are reported in Table 5. As seen from the results, the estimated biases and MSEs are near the desired value, zero. Moreover, the estimated AEs are closer the nominal values which indicates that the estimates are stable. The similar results can be also obtained for different parameter vector.

8. DATA ANALYSIS

In this section, we provide three applications to real data to illustrate the importance and flexibility of the XG-W, XG-N and LXG-W distributions. The Akaike Information Criteria (AIC), Bayesian information criterion (BIC) and Kolmogorov-Smirnov (K-S) statistic are used to compare the fitted distributions. All computations are performed using the **maxLik** routine in the R software.

8.1. Application 1: Glass fibres data

The first data set represents the strength of 1.5 cm glass fibres measured at National physical laboratory, England (Smith and Naylor [30], 1987). These data have been analyzed by Korkmaz and Genç [18] (2017). We shall compare the fits of the XG-W, Kumaraswamy-Weibull (Kw-W) (Cordeiro and de Castro [9], 2011), beta-Weibull (BW) (Famoye *et al.* [12], 2005), Lindley-Weibull (LW) (Cakmakyapan and Ozel [6], 2016), EW (Mudholkar and Srivastava [20], 1993) and odd log-logistic-Weibull (OLL-W) (Gleaton and Lynch [13], 2010; da Cruz *et al.* [10], 2016) distributions to the glass fibres data. The cdfs of the Kw-W, BW, LW, EW and OLL-W models (for $x > 0$) are given by

$$
\begin{aligned}
F(x) &= 1 - \left(1 - \left\{1 - \exp\left[-(ax)^b\right]\right\}^{\gamma}\right)^{\eta}, \\
F(x) &= \frac{1}{B(\gamma,\eta)} B\left(1 - \exp\left[-(ax)^b\right], \gamma, \eta\right), \\
F(x) &= 1 - \exp\left[-\theta (ax)^{\beta}\right]\left[1 + \frac{\theta}{\theta+1}(ax)^{\beta}\right], \\
F(x) &= \left\{1 - \exp\left[-(\alpha x)^{\beta}\right]\right\}^{\theta}, \\
F(x) &= \frac{\left\{1 - \exp\left[-(ax)^b\right]\right\}^{\theta}}{\left\{1 - \exp\left[-(ax)^b\right]\right\}^{\theta} + \exp\left[-\theta (ax)^b\right]},
\end{aligned}
$$

respectively, where $B(\gamma, \eta)$ is the complete beta function and the parameters of the above densities are all positive real numbers. The MLEs (and their corresponding standard errors in parentheses) of the parameters, AIC, BIC and K-S statistics for the above fitted models are displayed in Table 6. The values in this table indicate that the XG-W model provides a

better fit than the other fitted models because proposed model has the smallest values of the AIC, BIC and K-S statistics and has the largest p-value of the K-S statistic.

Table 6: The MLEs (standard errors in parentheses), AIC, BIC and K-S (with p-values in {·}) statistics for glass fibres data.

Model	$\widehat{\gamma}$	$\widehat{\eta}$	$\widehat{\theta}$	$\widehat{a}$	$\widehat{b}$	AIC	BIC	K-S
XG-W	-	-	0.4392	0.8952	4.6911	31.4342	37.8636	0.1210
	-	-	(0.0421)	(0.0250)	(0.1330)			{0.3151}
Kw-W	0.7910	112.6514	-	0.2702	7.2790	38.3943	46.9669	0.1524
	(0.1088)	(15.9962)	-	(0.0226)	(0.5671)			{0.1073}
BW	0.6207	120.6149	-	0.3051	7.7653	37.1752	45.7477	0.1455
	(0.0947)	(0.3733)	-	(0.0088)	(0.1522)			{0.1388}
LW	-	-	117.0336	0.2698	5.7804	36.4135	42.8429	0.1522
	-	-	(9.4031)	(0.0215)	(0.5800)			{0.1078}
EW	-	-	0.6713	0.5821	7.2841	35.3510	41.7804	0.1462
	-	-	(0.2876)	(0.0332)	(2.0252)			{0.1351}
OLL-W	-	-	0.9438	0.6159	6.0252	36.3736	42.8030	0.1537
	-	-	(0.2655)	(0.0163)	(1.3273)			{0.1018}

8.2. Application 2: Leukemia data

The second data set represents the lifetimes in days of 40 patients suffering from leukemia from one of the Ministry of Health Hospitals in Saudi Arabia (Abouammoh *et al.* [1], 1994). The data have been analyzed by Sarhan *et al.* [24] (2013). We compare the XG-N distribution with the Kumaraswamy-normal (Kw-N) (Cordeiro and de Castro [9], 2011), power-normal (PN) (Gupta and Gupta [15], 2008), logistic-normal (L-N) (Tahir *et al.* [31], 2016) and odd log-logistic-normal (OLL-N) (Braga *et al.* [5], 2016) distributions. The cdfs of the Kw-N, PN, L-N, and OLL-N models are given by

$$
\begin{aligned}
F(x) &= 1-\left\{1-\left[\mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{\gamma}\right\}^{\eta},\\
F(x) &= \left[\mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{\theta},\\
F(x) &= \left\{1+\left[1-\mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{-\theta}\right\}^{-1},\\
F(x) &= \frac{\left[\mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{\theta}}{\left[\mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{\theta}+\left[1-\mathbf{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right]^{\theta}},
\end{aligned}
$$

respectively, where $x, \mu \in \Re, \gamma, \eta, \sigma > 0$ and $\mathbf{\Phi}(\cdot)$ is the cdf of the standard normal distribution.

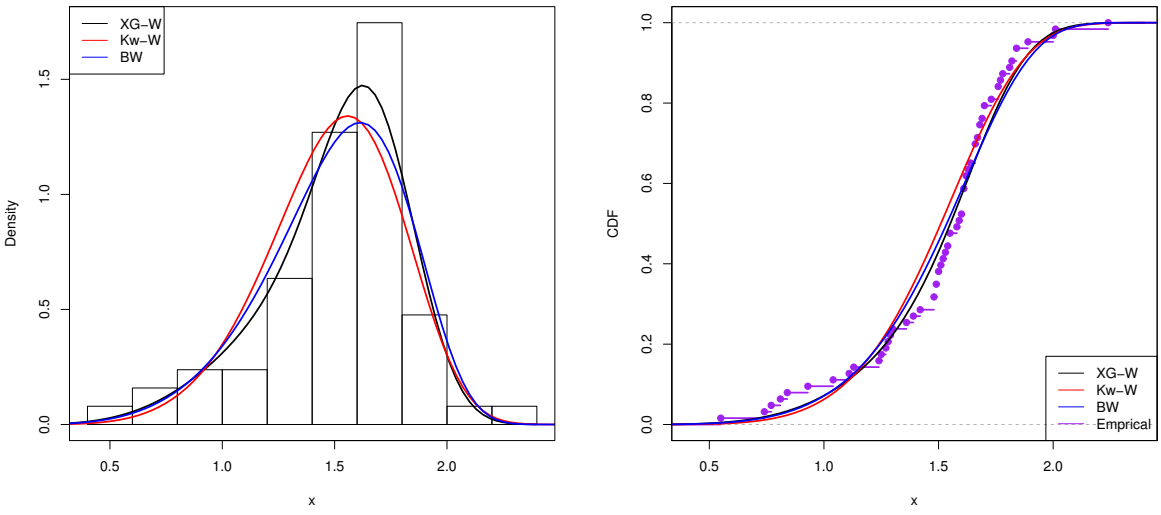
Table 7 lists the MLEs (and their standard errors) of the parameters and the K-S statistic for the fitted models. The figures in this table reveal that the XG-N distribution has the smallest values of the AIC, BIC and K-S statistics and has the largest p-value of the

K-S statistic. Therefore, we can conclude that the XG-N distribution could be chosen as the most adequate model for this data set.

Table 7: The MLEs (standard errors in parentheses) AIC, BIC and K-S (with p-values in $\{\cdot\}$) statistic for leukemia data.

Model	$\widehat{\gamma}$	$\widehat{\eta}$	$\widehat{\theta}$	$\widehat{\mu}$	$\widehat{\sigma}$	AIC	BIC	K-S
XG-N	-	-	0.6892	662.6324	609.7157	609.7157	614.7824	0.0825
	-	-	(0.0717)	(0.9018)	(1.5488)			{0.9484}
Kw-N	0.8320	0.2217	-	614.6680	294.0911	616.3179	623.0734	0.1314
	(1.0023)	(0.2217)	-	(0.5583)	(1.7209)			{0.4942}
PN	-	-	5.5078	189.8173	776.1012	615.2070	620.2736	0.1196
	-	-	(0.8745)	(4.2845)	(0.8745)			{0.6163}
L-N	-	-	4.4833	719.6505	1329.4302	617.0567	622.1234	0.1022
	-	-	(0.5873)	(0.5873)	(4.1943)			{0.7976}
OLL-N	-	-	36.6070	1169.5520	16331.9508	614.8475	619.9142	0.0869
	-	-	(5.9316)	(4.1943)	(7.2647)			{0.9228}

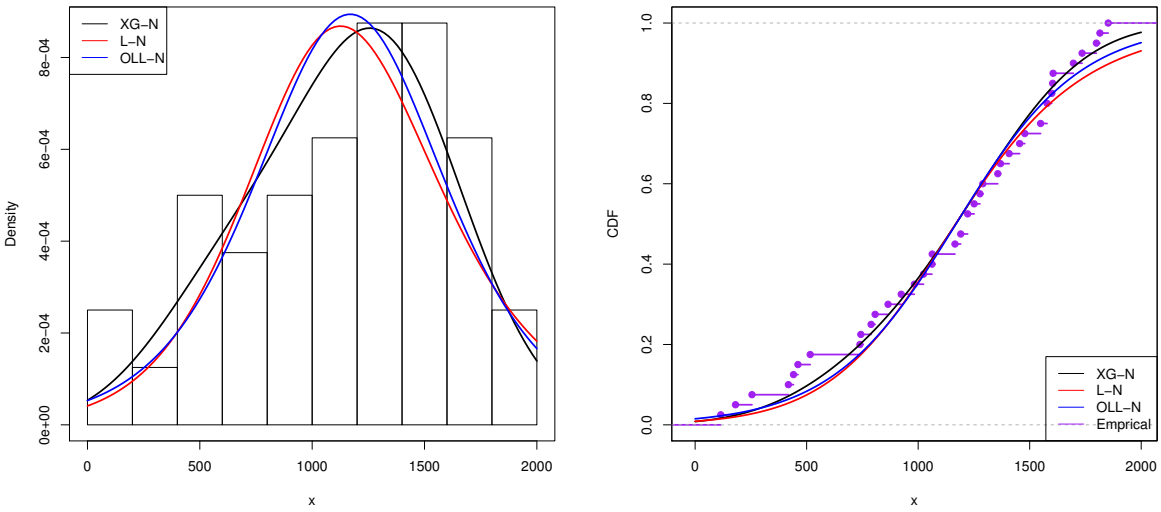
The histogram of both data sets and the estimated pdfs and cdfs of the XG-W and XG-N models and their competitive models are displayed in Figures 4 and 5, respectively. It is clear from these plots that the XG-W and XG-N models provide the best fits to both data sets.



(a) Fitted pdfs for data set I.

(b) Fitted cdfs for data set I.

Figure 4: Plots of the estimated pdfs and cdfs of the XG-W distribution and other competitive models.



(a) Fitted pdfs for data set II.

(b) Fitted cdfs for data set II.

Figure 5: Plots of the estimated pdfs and cdfs of the XG-N distribution and other competitive models.

8.3. Application 3: Diabetic retinopathy study

We consider a data set analyzed by Huster *et al.* [17] (1989) which represents patients with diabetic retinopathy in both eyes and 20/100 or better visual acuity for both eyes were eligible for the study. The patients were followed for two consecutively completed 4 month follow-ups and the endpoint was the occurrence of visual acuity less than 5/200. We choose only the treatment time. A 50% sample of the high-risk patients defined by diabetic retinopathy criteria was taken for the data set (n =197) and the percentage of censored observations was 72.4%. The variables involved in the study are: t_i – failure time for the treatment (in min); censoring indicator (0 = censoring, 1 = lifetime observed); x_{i1} – age (0 = patient is an adult diabetic, 1 = patient is a juvenile diabetic). The below regression structure is fitted by LXG-W regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \sigma\, z_i,$$

where the rv Y_i has the LXG-W distribution (6.1) for $i = 1, ..., 197$. The statistical software **R** is used to estimate the unknown model parameters by MLE approach. The **optim** function of **R** software is used to minimize the minus of log-likelihood function, given in (6.4). The initial values of the parameters are taken from the fitted LXG-E regression model (with $\sigma = 1$). The MLEs of the parameters of LXG-W regression model (approximate standard errors and p-values in parentheses) are: $\hat{\theta} = 1.7187\,(1.8739)$, $\hat{\sigma} = 1.2085\,(0.1518)$, $\hat{\beta}_0 = 4.2902\,(1.9308)\,(0.0068)$ and $\hat{\beta}_1 = 0.6474\,(0.3755)\,(0.0215)$. The explanatory variable x_1 is found statistically significant at the 5% significance level. In order to assess the validity of the fitted regression model, the estimated survival functions of the LXG-W regression model and empirical one are displayed in Figure 6. As seen from this figure, the LXG-W regression model provides substantial fit to these data.

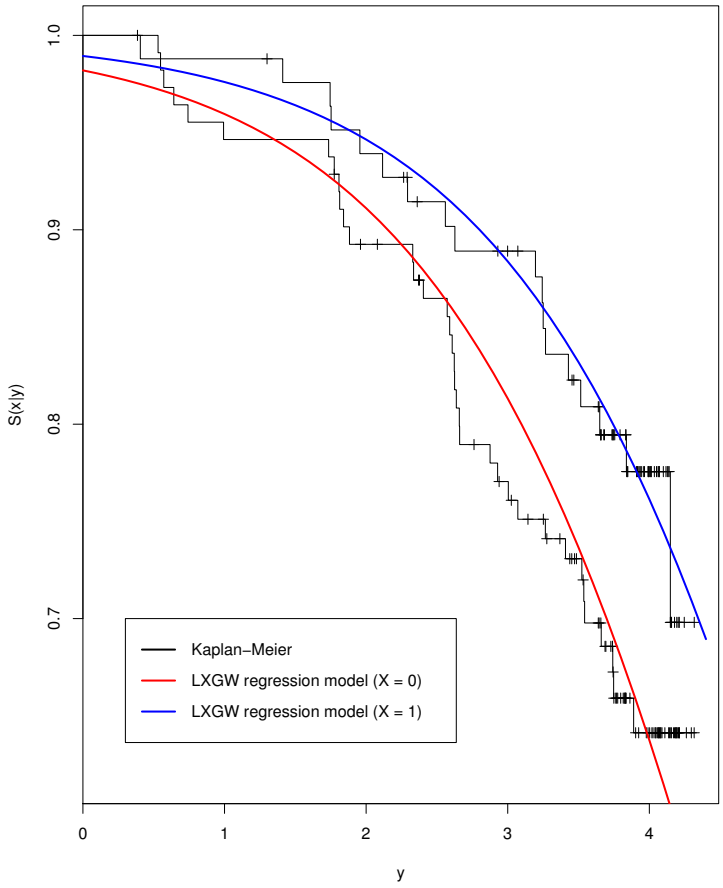


Figure 6: Estimated survival function by fitting the LXG-W regression model and the empirical survival for each level of the diabetic retinopathy study.

9. CONCLUSIONS

A new class of distributions called the xgamma-G family with one extra positive parameter is introduced and studied. We provide some mathematical properties of the new family including ordinary and incomplete moments, quantile and generating functions and mean deviations. The maximum likelihood method is used for estimating the model parameters. We assess of the performance of the maximum likelihood estimators in terms of biases and mean squared errors by means of two simulation studies. We also introduced a new linear regression model based on the logarithm of the xgamma random variable for uncensored and censored data. We prove that the special models of the proposed family provide consistently better fits than other competitive models by means of three real data sets.

REFERENCES

[1] Abouammoh, A.M.; Abdulghani, S.A. and Qamber, I.S. (1994). On partial orderings and testing of new better than renewal used classes, *Reliability Engineering & System Safety*, **43**, 37–41.

[2] Altun, E. and Hamedani, G.G. (2018). The log-xgamma distribution with inference and application, *Journal de la Société Française de Statistique*, **159**, 40–55.

[3] Alzaatreh, A.; Lee, C. and Famoye, F. (2013). A new method for generating families of continuous distributions, *Metron*, **71**, 63–79.

[4] Biçer, H.D. (2019). Properties and inference for a new class of XGamma distributions with an application, *Mathematical Sciences*, **4**, 335–346.

[5] Braga, A.S.; Cordeiro, G.M.; Ortega, E.M.M. and da Cruz, J.N. (2016). The odd log–logistic normal distribution: theory and applications in analysis of experiments, *Journal of Statistical Theory and Practice*, **10**, 311–335.

[6] Cakmakyapan, S. and Ozel, G. (2016). The Lindley family of distributions: properties and applications, *Hacettepe Journal of Mathematics and Statistics*, **46**, 1113–1137.

[7] Cordeiro, G.M.; Afify, A.Z.; Yousof, H.M.; Pescim, R.R. and Aryal, G.R. (2017). The exponentiated Weibull-H family of distributions: theory and applications, *Mediterranean Journal of Mathematics*, **14**, 155.

[8] Cordeiro, G.M.; Alizadeh, M. and Diniz Marinho, P.R. (2016). The type I half-logistic family of distributions, *J. Stat. Comput. Simul.*, **86**, 707–728.

[9] Cordeiro, G.M. and de Castro, M. (2011). A new family of generalized distributions, *J. Stat. Comput. Simul.*, **81**, 883–898.

[10] da Cruz, J.N.; Ortega, E.M.M. and Cordeiro, G.M. (2016). The log-odd log-logistic Weibull regression model: modelling, estimation, influence diagnostics and residual analysis, *J. Stat. Comput. Simul.*, **86**, 1516–1538.

[11] Eugene, N.; Lee, C. and Famoye, F. (2002). Beta-normal distribution and its applications, *Commun. Stat. Theory Methods*, **31**, 497–512.

[12] Famoye, F.; Lee, C. and Olumolade, O. (2005). The beta-Weibull distribution, *Journal of Statistical Theory and Applications*, **4**, 121–136.

[13] Gleaton, J.U. and Lynch (2010). Extended generalized log-logistic families of lifetime distributions with an application, *Journal of Probability and Statistical Science*, **8**, 1–17.

[14] Gradshteyn, I.S. and Ryzhik, I.M. (2002). *Table of Integrals, Series, and Products*, Academic Press, San Diego, CA.

[15] Gupta, R.D. and Gupta. R.C. (2008). Analyzing skewed data by power normal model, *Test*, **17**, 197–210.

[16] Gupta, R.D. and Kundu, D. (1999). Generalized Exponential distributions, *Australian and New Zealand Journal of Statistics*, **41**, 173–188.

[17] Huster, W.J.; Brookmeyer, R. and Self, S.G. (1989). Modelling paired survival data with covariates, *Biometrics*, **45**, 145–156.

[18] Korkmaz, M.Ç. and Genç, A.I. (2017). A new generalized two-sided class of distributions with an emphasis on two-sided generalized normal distribution, *Comm. Stat. Simulation and Computation*, **46**, 1441–1460.

[19] Marshall, A.W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families, *Biometrika*, **84**, 641–652.

[20] Mudholkar, G.S. and Srivastava, D.K. (1993). Exponentiated Weibull family for analysing bathtub failure rate data, *IEEE Transactions on Reliability*, **42**, 299–302.

[21] Mudholkar, G.S.; Srivastava, D.K. and Freimer, M. (1995). The exponentiated Weibull family: a reanalysis of the bus-motor-failure data, *Technometrics*, **37**, 436–445.

[22] Nadarajah, S. and Gupta, A.K. (2007). The exponentiated gamma distribution with application to drought data, *Calcutta Statistical Association Bulletin*, **59**, 29–54.

[23] Nofal, Z.M.; Afify, A.Z.; Yousof, H.M. and Cordeiro, G.M. (2017). The generalized transmuted-G family of distributions, *Commun. Stat. Theory Methods*, **46**, 4119–4136.

[24] Sarhan, A.M.; Ahmad, A.A. and Alasbahi, A.I. (2013). Exponentiated generalized linear exponential distribution, *Applied Mathematical Modelling*, **37**, 2838–2849.

[25] Sen, S. and Chandra, S.S.N. (2017). The quasi xgamma distribution with application in bladder cancer data, *Journal of Data Science*, **15**, 61–76.

[26] Sen, S.; Chandra, N. and Maiti, S.S. (2017). The weighted xgamma distribution: properties and application, *Journal of Reliability and Statistical Studies*, **10**, 43–58.

[27] Sen, S.; Chandra, N. and Maiti, S.S. (2018a). On properties and applications of a two-parameter xgamma distribution, *Journal of Statistical Theory and Applications*, **4**, 674–685.

[28] Sen, S.; Chandra, N. and Maiti, S.S. (2018b). Survival estimation in xgamma distribution under progressively type-II right censored scheme, *Model Assisted Statistics and Applications*, **2**, 107–121.

[29] Sen, S.; Maiti, S.S. and Chandra, N. (2016). The xgamma distribution: statistical properties and application, *Journal of Modern Applied Statistical Methods*, **15**, 774–788.

[30] Smith, R.L. and Naylor, J.C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution, *Appl. Statist.*, **36**, 358–369.

[31] Tahir, M.H.; Cordeiro, G.M.; Alzaatreh, A.; Mansoor, M. and Zubair, M. (2016). The logistic-X family of distributions and its applications, *Comm. Stat. Theory and Methods*, **45**, 7326–7349.

[32] Yousof, H.M.; Afify, A.Z.; Hamedani, G.G. and Aryal, G. (2016). The Burr X generator of distributions for lifetime data, *Journal of Statistical Theory and Applications*, **16**, 1–19.

[33] Zografos, K. and Balakrishnan, N. (2009). On families of beta and generalized gamma-generated distributions and associated inference, *Statistical Methodology*, **6**, 344–362.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 613–635

THE FAY–HERRIOT MODEL IN SMALL AREA ESTIMATION: EM ALGORITHM AND APPLICATION TO OFFICIAL DATA

Authors: José Luis Ávila-Valdez
– Department of Mathematics, Universidad Popular Autónoma del Estado de Puebla, Mexico
joseluis.avila@upaep.mx

Mauricio Huerta
– School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile
mauricio.huerta.a@gmail.com

Víctor Leiva
– School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile
victorleivasanchez@gmail.com www.victorleiva.cl

Marco Riquelme
– Institute of Statistics, Universidad de Valparaíso, Chile
mriquelmealamos@gmail.com

Leonardo Trujillo
– Department of Statistics, Universidad Nacional de Colombia, Colombia
ltrujilloo@unal.edu.co

Received: July 2017 Revised: May 2018 Accepted: July 2018

Abstract:

- Standard methods of variance component estimation used in the Fay-Herriot model for small areas can produce problems of inadmissible values (negative or zero) for these variances. This implies that the empirical best linear unbiased predictor of a small area mean does not take into account the variance of the random effect of the corresponding area, reducing it to a regression estimator. In this paper, we propose an approach based on the expectation-maximization (EM) algorithm to solve the problem of inadmissibility. As stated in the theory of variance component estimation, we confirm through Monte Carlo simulations that the EM algorithm always produces strictly positive variance component estimates. In addition, we compare the performance of the proposed approach with two recently proposed methods in terms of relative bias, mean square error and mean square predictor error. We illustrate our approach with official data related to food security and poverty collected in Mexico, showing their potential applications.

Key-Words:

- *empirical best linear unbiased predictor; food security and poverty; Monte Carlo simulation; R software; random effects; variance components.*

AMS Subject Classification:

- 62D05, 62F99, 62J99.

1. INTRODUCTION

Surveys are intended not only to estimate population target parameters, but also to estimate characteristics for a variety of subpopulations commonly known as domains or areas. An area is considered as small if the sample domain is not sufficiently large to have a direct estimate of the area parameter with adequate precision. Then, the goal of the small area estimation is to produce reliable estimates of subpopulation target parameters for areas with small samples or even where the area is not sampled at all; see Pfeffermann [30] (2013).

Currently, the small area estimation methodology is playing an important role in both public and private sectors. Different government agencies around the world, for example, the Bureau of Labor Statistics and Census Bureau in the United States (US), Ministry of Social Development of Chile, National Administrative Department of Statistics in Colombia, National Council for the Evaluation of Social Development Policy in Mexico and Office of National Statistics in the United Kingdom (UK) are adopting such a methodology. This is due to the need for reliable estimates on parameters of interest in specific areas or domains; see Rao and Molina [35] (2015).

Because of the wide acceptance about small area estimation in recent years, several models have been developed, applied and studied. Pfeffermann [30] (2013), Rao [34] (2003) and Rao and Molina [35] (2015) reviewed the advances in this methodology from its beginnings to the present. Small area estimation methodology can be divided into two parts (Lohr [24, pp. 518-522], 1999): (i) design-based techniques (for example, direct, synthetic and composite estimators) and (ii) model-based techniques (for example, area-level models and unit-level models); see Coelho and Pereira [6] (2011), Pereira and Coelho [29] (2012) and Rueda *et al.* [36] (2018). On the one hand, in design-based techniques, the existence of a model is not recognized. Implicit models are sometimes proposed as an assisting tool, linking a number of small areas according to administrative or census records, which is considered as auxiliary data. Then, even when the model is misspecified, design-based properties can hold; see Lehtonen and Veijanen [22] (2009). On the other hand, model-based techniques rely on explicit super-population models (Datta [7], 2009) and include area-level models, relating each small area characteristic to auxiliary data that are available for each area. Area-level modeling is often described by the popular Fay-Herriot (FH) model (Fay and Herriot [13], 1979), which has been widely used in small area estimation. Li and Lahiri [23] (2010) emphasized that the main reasons for its widespread usage include: (i) its simplicity, (ii) its ability to protect the confidentiality of microdata, and (iii) its ability to produce design-consistent estimators. Other advantages of the FH model are that it takes into account the sampling design (level 1 model) and only requires area auxiliary variables that, in general, are more easily available in practice than unit auxiliary variables. Applications of the FH model have been extensive, mainly in the study of poverty and other related socio-demographic variables. For a reference in the context of big data sources in small area estimation through the FH model, see Marchetti *et al.* [25] (2015). A recent application of the FH model for poverty mapping in Chile can be found in Casas-Cordero *et al.* [2] (2016). Also, model-based techniques include unit-level models relating the unit values of the response variable to auxiliary variables for each individual in the survey; see Coelho and Casimiro [5] (2008). A well-known model proposed by Battese *et al.* [1] (1988) is a particular example of a unit-level model, corresponding to a nested regression model. Area and unit model-based

techniques in small area estimation are presented by Jiang and Lahiri [20] (2006), Datta [7] (2009) and Datta and Ghosh [8] (2012), among others. Linear mixed models have played a crucial role in model-based techniques. Note also that these techniques can be based on either Bayesian or frequentist methods. In this paper, we consider a frequentist model-based technique employing the FH model under a non-informative sampling design. For informative sampling, see Pfeffermann and Sverchkov [31] (2007).

A problem detected in small areas, using the FH model, is that the standard methods utilized for variance component estimation may produce a negative or zero value. For more details about these methods, see Fay and Herriot [13] (1979), Prasad and Rao [33] (1990) for moment estimation (PR method), and Datta and Lahiri [9] (2000) for maximum likelihood (ML) and residual or restricted ML (RML) estimation. Note that the empirical best linear unbiased predictor (EBLUP) of a small area mean does not take into account the variance of the random effect for the corresponding area, reducing it to a regression estimator. Li and Lahiri [23] (2010) and Yoshimori and Lahiri [42] (2014) solved this problem adjusting the associated likelihood function.

The expectation-maximization (EM) algorithm is a popular iterative approach to estimate parameters by the ML method in models with incomplete data (unobserved or missing). This algorithm is used in many applications of mixed models, because there the unobserved data occur naturally. A comprehensive account of the EM algorithm is found in Laird and Were [21] (1982), van Dyk [39] (2000) and McLachlan and Krishnan [26] (2008). Some advantages of the EM algorithm are the following: (i) it is more stable than other algorithms, due to its property of monotone convergence (Laird and Were [21], 1982); (ii) it is more robust to starting values than other algorithms (Demidenko [11], 2013); and (iii) it generates positive definite matrix estimates if the starting matrix is positive definite (Thompson and Meyer [38], 1986; Searle *et al.* [37], 2006; Demidenko [11], 2013; El-Leithy *et al.* [12], 2016). An important feature related to (iii), stated by Searle *et al.* [37, pp. 297-298] (2006), is that the iterations will always remain in the parameter space, since the ML estimation is performed for the complete data.

The main objectives of this research are: (i) to review the estimation methods proposed at date on the topic; (ii) to propose an alternative approach for avoiding a negative or zero value in the variance component estimates, using the EM algorithm in both ML (MLEM) and RML (RMLEM) methods; (iii) to evaluate the proposed approach by Monte Carlo (MC) simulations; and (iv) to illustrate potential applications of our approach with official data related to food security and poverty. The proposed approach is compared to the methods presented in Li and Lahiri [23] (2010) and Yoshimori and Lahiri [42] (2014).

The outline of this paper is as follows. Section 2 introduces background to the FH model, the EBLUP of a small area mean, and a measure of its uncertainty. In addition, in this section, some variance component estimation methods are reviewed, highlighting their advantages and shortcomings. In Section 3, we propose an approach based on the EM algorithm to get positive values for the variance component estimates. In Section 4, the results of an MC simulation are presented to assess the performance of the proposed approach, comparing it to two alternative methods recently introduced. In Section 5, we apply our approach to estimate the small area means of monthly per capita expenditure in a food security and poverty study conducted in Mexico; see CIESIN [4] (2005). Conclusions and future research are discussed in Section 6.

2. THE FAY-HERRIOT MODEL

2.1. Formulation

Suppose that there are m small areas labeled as $i = 1, ..., m$. Assuming a $p \times 1$ vector of observed values $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^\top$ for auxiliary variables is available for each area i, Fay and Herriot [13] (1979) proposed their model to improve the direct estimator $\widehat{\theta}_i$ used to compute the true small area mean θ_i, consisting of the following two levels:

- Level 1 (sampling model): $\widehat{\theta}_i|\theta_i \overset{\text{IND}}{\sim} \text{N}(\theta_i, \psi_i)$,
- Level 2 (linking model): $\theta_i \overset{\text{IND}}{\sim} \text{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, $\quad i = 1, ..., m$,

where "IND" stands for "independent", ψ_i is the known variance of the sampling error, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is a vector of unknown regression coefficients to be estimated, and σ^2 is the unknown variance of the area-specific random effect to be estimated. Level 1 accounts for the sampling variability of the survey estimates $\widehat{\theta}_i$ of θ_i, whereas Level 2 links θ_i to the vector of p known area-specific auxiliary variables; see Jiang and Lahiri [20] (2006) and Li and Lahiri [23] (2010). Then, the FH model can be written as

$$\widehat{\theta}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + b_i + \varepsilon_i, \quad i = 1, ..., m, \tag{2.1}$$

where $b_i \overset{\text{IID}}{\sim} \text{N}(0, \sigma^2)$ are independent and identically distributed (IID) area-specific random effects with unknown σ^2 to be estimated from the data, and $\varepsilon_i \overset{\text{IND}}{\sim} \text{N}(0, \psi_i)$ represent sampling errors with known variances ψ_i. Although in this paper we are considering ψ_i as known, in practical cases when the variances ψ_i are not available, Fay and Herriot [13] (1979) employed generalized variance functions (Wolter [41, Chapter 7], 2007) to estimate them. In addition, it is assumed that b_i and ε_i are independent.

2.2. Estimation of a small area mean

We are interested in estimating or predicting the small area mean $\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + b_i$, as well as in obtaining a measurement of uncertainty associated with that prediction. Under the model given in (2.1), the best predictor (BP) of θ_i, which minimizes the mean squared error (MSE), can be expressed by a weighted average of the direct estimator $\widehat{\theta}_i$ and the regression-synthetic estimator $\boldsymbol{x}_i^\top \boldsymbol{\beta}$, being it defined as

$$\widehat{\theta}_i^{\text{BP}} = (1 - B_i)\widehat{\theta}_i + B_i \, \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad i = 1, ..., m, \tag{2.2}$$

with the weight $0 < B_i < 1$ defined as $B_i = \psi_i/(\sigma^2 + \psi_i)$. Note that $(1 - B_i)$ is a function of the variance ratio σ^2/ψ_i and measures the uncertainty when estimating θ_i relative to the total variance $\sigma^2 + \psi_i$; see Rao and Molina [35] (2015). In addition, the parameter σ^2 is a measure of homogeneity of the areas after accounting for the auxiliary variables $\boldsymbol{x}_i$. If σ^2 is known, $\boldsymbol{\beta}$ can be obtained by the standard weighted least squares estimator $\overline{\boldsymbol{\beta}}$; see Mert [27] (2015). Hence, by replacing it in (2.2), one gets the best linear unbiased predictor (BLUP)

of θ_i expressed as

$$\widehat{\theta}_i^{\,\mathrm{BLUP}} = (1 - B_i)\widehat{\theta}_i + B_i\, \boldsymbol{x}_i^\top \overline{\boldsymbol{\beta}}, \quad i = 1, ..., m, \tag{2.3}$$

where

$$\overline{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{m} \boldsymbol{x}_i \widehat{\theta}_i/(\sigma^2 + \psi_i)}{\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top/(\sigma^2 + \psi_i)}.$$

The BLUP of θ_i given in (2.3) depends on σ^2, which is unknown in practical applications. Replacing σ^2 in (2.3) with a general estimator, that for now we denote by $\widehat{\sigma}^{\,2}$ (see details in Section 2.3), we obtain the EBLUP of θ_i as

$$\widehat{\theta}_i^{\,\mathrm{EBLUP}} = (1 - \widehat{B}_i)\widehat{\theta}_i + \widehat{B}_i\, \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}, \tag{2.4}$$

where $\widehat{B}_i$ and $\widetilde{\boldsymbol{\beta}}$ are the estimators of B_i and $\boldsymbol{\beta}$ when σ^2 is replaced with $\widehat{\sigma}^{\,2}$ in (2.2) and (2.3), respectively. Note that the model given in (2.1) can be rewritten in matrix terms as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{I}_m \boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{2.5}$$

where $\boldsymbol{Y} = (Y_1, ..., Y_m)^\top$, with $Y_i = \widehat{\theta}_i$, $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_m)^\top$ is of full rank, $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, $\boldsymbol{\beta}$ is defined as above, $\boldsymbol{b} = (b_1, ..., b_m)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_m)^\top$. In addition, as mentioned in scalar terms, $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ are independently distributed with $\boldsymbol{b} \sim \mathrm{N}_m(\boldsymbol{0}, \boldsymbol{G})$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}_m(\boldsymbol{0}, \boldsymbol{S})$, for $\boldsymbol{G} = \sigma^2 \boldsymbol{I}_m$ and $\boldsymbol{S} = \mathrm{diag}\{\psi_1, ..., \psi_m\}$. The model defined in (2.5) is a particular case of a more general linear mixed model (Datta *et al.* [10], 2005) with its variance-covariance matrix taking the form $\boldsymbol{V} = \boldsymbol{G} + \boldsymbol{S}$.

2.3. Estimation of σ^2

Note that the EBLUP given in (2.4) depends on the way how σ^2 is estimated. Different methods have been proposed in the literature to estimate σ^2; see Fay and Herriot [13] (1979) and Prasad and Rao [33] (1990). In those cases when the estimate of σ^2 takes a negative value, Prasad and Rao [33] (1990) suggested to truncate the negative estimate at zero. They also showed that the probability of having a negative estimate goes to zero as $m \to \infty$; see Datta [7] (2009). As an alternative, the ML method has been widely used in small area estimation; see Jiang and Lahiri [20] (2006) and Rao and Molina [35] (2015). It was employed by Datta and Lahiri [9] (2000) in the context of the FH model, in whose case the log-likelihood function takes the form

$$\ell_{\mathrm{ML}}(\sigma^2, \boldsymbol{\beta}; \boldsymbol{Y}) = c - \frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.6}$$

where c is a constant that is independent of σ^2. By differentiating (2.6) with respect to $\boldsymbol{\beta}$ and σ^2, we have

$$\frac{\partial \ell_{\mathrm{ML}}(\sigma^2, \boldsymbol{\beta}; \boldsymbol{Y})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}\boldsymbol{\beta}, \tag{2.7}$$

$$\frac{\partial \ell_{\mathrm{ML}}(\sigma^2, \boldsymbol{\beta}; \boldsymbol{Y})}{\partial \sigma^2} = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}). \tag{2.8}$$

Thus, equating (2.7) and (2.8) to zero, and solving them simultaneously with respect to σ^2 and $\boldsymbol{\beta}$, we obtain the ML estimators of σ^2, denoted by $\widehat{\sigma}^2_{\mathrm{ML}}$, and of $\boldsymbol{\beta}$ given in (2.3). If we replace $\boldsymbol{\beta}$ by $\widetilde{\boldsymbol{\beta}}$ in (2.6), we have the corresponding profile log-likelihood (PML) function expressed as

$$\ell_{\mathrm{PML}}(\sigma^2;\boldsymbol{Y}) = c - \frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}\boldsymbol{Y}^\top \boldsymbol{P}\boldsymbol{Y}, \tag{2.9}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}$. Equating (2.9) to zero and solving it with respect to σ^2, we have an estimator of σ^2 identical to that obtained through (2.6) by means of the ML method; see Jiang [19] (2007). Consequently, the associated estimate computed with the PML method is not analyzed here. Datta and Lahiri [9] (2000) obtained both the asymptotic variance and bias of $\widehat{\sigma}^2_{\mathrm{ML}}$, given respectively by

$$\begin{aligned} \overline{V}[\widehat{\sigma}^2_{\mathrm{ML}}] &= \frac{2}{\mathrm{tr}(\boldsymbol{V}^{-2})} + o(m^{-1}), \\ \mathrm{Bias}[\widehat{\sigma}^2_{\mathrm{ML}}] &= \frac{\mathrm{tr}(\boldsymbol{P} - \boldsymbol{V}^{-1})}{\mathrm{tr}(\boldsymbol{V}^{-2})} + o(m^{-1}). \end{aligned} \tag{2.10}$$

Note that the ML estimates tend to underestimate the variance components and then the RML estimation is preferred; see Pinheiro and Bates [32] (2004). A feature of the RML method is that, when estimating variance components, it takes into account the degrees of freedom involved in estimating the fixed effects, which is not considered by the ML method; see Searle *et al.* [37] (2006). Several alternative derivations of the RML method have been presented in the literature; see Harville [18] (1977), Jiang [19] (2007) and references therein.

Verbyla [40] (1990) proposed an approach which divides the likelihood function into two independent parts, one related to the fixed effect ($\boldsymbol{Y}_1 = \boldsymbol{L}_1^\top\boldsymbol{Y}$) and the another part related to the residual contrasts $\boldsymbol{Y}_2 = \boldsymbol{L}_2^\top\boldsymbol{Y}$, where $\boldsymbol{L} = [\boldsymbol{L}_1 \;\; \boldsymbol{L}_2]$ is a non-singular matrix, with $\boldsymbol{Y}$ given in (2.5) and $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ being $m \times p$ and $m \times (m-p)$ matrices, respectively, both of full column rank, which are chosen to satisfy $\boldsymbol{L}_1^\top\boldsymbol{X} = \boldsymbol{I}_p$ and $\boldsymbol{L}_2^\top\boldsymbol{X} = \boldsymbol{0}$. Therefore, $\boldsymbol{Y}$ is transformed as

$$\boldsymbol{L}^\top\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{L}_1^\top\boldsymbol{Y} \\ \boldsymbol{L}_2^\top\boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} \sim \mathrm{N}_m\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{L}_1^\top\boldsymbol{V}\boldsymbol{L}_1 & \boldsymbol{L}_1^\top\boldsymbol{V}\boldsymbol{L}_2 \\ \boldsymbol{L}_2^\top\boldsymbol{V}\boldsymbol{L}_1 & \boldsymbol{L}_2^\top\boldsymbol{V}\;\boldsymbol{L}_2 \end{bmatrix} \right).$$

The probability density function (PDF) of $\boldsymbol{L}^\top\boldsymbol{Y}$ can be expressed as the product of the conditional PDF of $\boldsymbol{Y}_1$ given $\boldsymbol{Y}_2$ and the marginal PDF of $\boldsymbol{Y}_2$. Hence, the log-likelihood function of $\boldsymbol{L}^\top\boldsymbol{Y}$ is $\ell(\boldsymbol{\beta},\sigma^2;\boldsymbol{L}^\top\boldsymbol{Y}) = \ell(\boldsymbol{\beta},\sigma^2;\boldsymbol{Y}_1|\boldsymbol{Y}_2) + \ell(\sigma^2;\boldsymbol{Y}_2)$. Since $\boldsymbol{Y}_1$ is a $p \times 1$ vector and $\ell(\sigma^2;\boldsymbol{Y}_2)$ is not a function of $\boldsymbol{\beta}$, the fixed effects are estimated from $\ell(\boldsymbol{\beta},\sigma^2;\boldsymbol{Y}_1|\boldsymbol{Y}_2)$. Once $\boldsymbol{\beta}$ has been estimated, there is no information left for estimating σ^2 and $\ell(\sigma^2;\boldsymbol{Y}_2)$ is used for estimating σ^2. The function $\ell(\sigma^2;\boldsymbol{Y}_2)$ is known as residual or restricted log-likelihood function, from which the RML estimator is obtained. Then, under the FH model defined in (2.5), it is expressed as

$$\ell_{\mathrm{RML}}(\sigma^2;\boldsymbol{Y}_2) = c - \frac{1}{2}\log(|\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}|) - \frac{1}{2}\log(|\boldsymbol{V}|) - \frac{1}{2}\boldsymbol{Y}^\top\boldsymbol{P}\boldsymbol{Y}. \tag{2.11}$$

Thus, the RML estimator of σ^2, $\widehat{\sigma}^2_{\mathrm{RML}}$ namely, is generated as a solution of the equation

$$\frac{\partial \ell_{\mathrm{RML}}(\sigma^2;\boldsymbol{Y}_2)}{\partial \sigma^2} = 0.$$

Datta and Lahiri [9] (2000) showed that $\overline{V}[\widehat{\sigma}^2_{\mathrm{RML}}]$ is identical to $\overline{V}[\widehat{\sigma}^2_{\mathrm{ML}}]$ given in (2.10), whereas $\widehat{\sigma}^2_{\mathrm{RML}}$ is asymptotically unbiased for σ^2. All above estimators hold the following properties: (i) they are $m^{\frac{1}{2}}$-consistent, that is, $\widehat{\sigma}^{\,2} - \sigma^2 = O(m^{\frac{1}{2}})$; (ii) they are even functions of $\boldsymbol{Y}$ and hence $\widehat{\sigma}^{\,2}(-\boldsymbol{Y}) = \widehat{\sigma}^{\,2}(\boldsymbol{Y})$; and (iii) they are invariant functions under translation and so $\widehat{\sigma}^{\,2}(\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{d}) = \widehat{\sigma}^{\,2}(\boldsymbol{Y})$, for any $\boldsymbol{d} \in \mathbb{R}^p$ and for all $\boldsymbol{Y}$; see Datta *et al.* [10] (2005). In contrast to these properties (i)-(iii), the FH, ML, PR and RML methods can provide non-admissible negative or zero values for the estimates of $\widehat{\sigma}^{\,2}$, especially when the number of small areas is low; see Li and Lahiri [23] (2010) and Yoshimori and Lahiri [42] (2014). However, as happens in practice with any of these methods, the estimate $\widehat{\sigma}^{\,2} = \max(\widehat{\sigma}^2_{\mathrm{M}}, 0)$ is used, where "M" indicates the FH, PR, ML or RML method. Then, if $\widehat{\sigma}^{\,2} = 0$ (when the Level 2 model is perfect), the EBLUP in (2.4) reduces to the simple regression-synthetic estimator (since $\widehat{B}_i = 1$), which typically has an overshrinking problem. Thus, as mentioned by Li and Lahiri [23] (2010), this situation is unrealistic, because Level 2 model cannot be perfect and $\widehat{\sigma}^{\,2}$ should be always greater than zero. To solve the problem of a negative or zero value for the variance component estimate, various methods have been proposed. Li and Lahiri [23] (2010) adjusted the ML (LML) method defining a product of σ^2 and a standard likelihood function, introducing the adjusted log-likelihood function $\ell_{\mathrm{LML}}(\sigma^2; \boldsymbol{Y}) = \ell(\sigma^2; \boldsymbol{Y}) + \log(\sigma^2)$, where $\ell(\sigma^2; \boldsymbol{Y})$ may be chosen from (2.9) or (2.11). Its maximization produces the LML and Li-Lahiri RML (LRML) estimators of σ^2, denoted as $\widehat{\sigma}^2_{\mathrm{LML}}$ and $\widehat{\sigma}^2_{\mathrm{LRML}}$, respectively. Both $\widehat{\sigma}^2_{\mathrm{LML}}$ and $\widehat{\sigma}^2_{\mathrm{LRML}}$ are strictly positive, even for small m. Li and Lahiri [23] (2010) showed that their asymptotic variances are as given in (2.10). In addition, the corresponding biases are expressed as

$$\begin{aligned}
\mathrm{Bias}[\widehat{\sigma}^2_{\mathrm{LML}}] &= \frac{\mathrm{tr}(\boldsymbol{P} - \boldsymbol{V}^{-1}) + 2/\sigma^2}{\mathrm{tr}(\boldsymbol{V}^{-2})} + o(m^{-1}), \\
\mathrm{Bias}[\widehat{\sigma}^2_{\mathrm{LRML}}] &= \frac{2/\sigma^2}{\mathrm{tr}(\boldsymbol{V}^{-2})} + o(m^{-1}).
\end{aligned}$$

Yoshimori and Lahiri [42] (2014) proposed other adjusted ML method, with adjusted likelihood function defined as the product of a function $h(\sigma^2)$ and a standard likelihood function. In this case, the adjusted log-likelihood function is defined as

$$\ell_{\mathrm{YML}}(\sigma^2; \boldsymbol{Y}) = \ell(\sigma^2; \boldsymbol{Y}) + \log(h(\sigma^2)), \tag{2.12}$$

where $\ell(\sigma^2; \boldsymbol{Y})$ expressed in (2.12) can be chosen from (2.9) or (2.11), and $h(\sigma^2) = (\tan^{-1}(\mathrm{tr} - (\boldsymbol{I}_m - \boldsymbol{B})))^{\frac{1}{m}}$, with $\boldsymbol{B} = \mathrm{diag}\{B_1, ..., B_m\}$ and B_i as defined in (2.2). Thus, the Yoshimori-Lahiri ML (YML) and Yoshimori-Lahiri RML (YRML) estimators of σ^2, denoted by $\widehat{\sigma}^2_{\mathrm{YML}}$ and $\widehat{\sigma}^2_{\mathrm{YRML}}$, respectively, are obtained by maximizing (2.12) with respect to σ^2. Both $\widehat{\sigma}^2_{\mathrm{YML}}$ and $\widehat{\sigma}^2_{\mathrm{YRML}}$ are also strictly positive, even for small m. Yoshimori and Lahiri [42] (2014) showed that their asymptotic variances are identical as in (2.10). In addition, we have that

$$\mathrm{Bias}[\widehat{\sigma}^2_{\mathrm{YML}}] = \frac{\mathrm{tr}(\boldsymbol{P} - \boldsymbol{V}^{-1})}{\mathrm{tr}(\boldsymbol{V}^{-2})} + o(m^{-1})$$

and $\widehat{\sigma}^2_{\mathrm{YRML}}$ is asymptotically unbiased for σ^2.

2.4. Uncertainty of the EBLUP

A measure of uncertainty of the EBLUP of θ_i given in (2.4) is obtained by its mean squared predicted error (MSPE), also known as MSE or predicted mean squared error (Rao

and Molina [35, Section 5.2, p. 119], 2015), defined by

$$\text{MSPE}[\widehat{\theta}_i^{\text{EBLUP}}] = \text{E}[\widehat{\theta}_i^{\text{EBLUP}} - \theta_i]^2, \tag{2.13}$$

which, under certain regularity conditions, can be decomposed as (Datta *et al.* [10], 2005)

$$\text{MSPE}[\widehat{\theta}_i^{\text{EBLUP}}] = g_{1i}(\sigma^2) + g_{2i}(\sigma^2) + \text{E}[\widehat{\theta}_i^{\text{EBLUP}} - \widehat{\theta}_i^{\text{BLUP}}]^2, \tag{2.14}$$

where

$$g_{1i}(\sigma^2) = \frac{\sigma^2\psi_i}{\sigma^2+\psi_i}, \quad g_{2i}(\sigma^2) = \frac{\psi_i^2}{(\sigma^2+\psi_i)^2}\boldsymbol{x}_i^\top\left(\sum_{j=1}^{m}\frac{1}{\sigma^2+\psi_i}\boldsymbol{x}_j\boldsymbol{x}_j^\top\right)^{-1}\boldsymbol{x}_i. \tag{2.15}$$

The term $g_{1i}(\sigma^2)$ is of order $O(1)$, which captures the uncertainty of the BP given in (2.2), whereas the term $g_{2i}(\sigma^2)$ is of order $O(m^{-1})$, capturing the uncertainty due to the estimation of $\boldsymbol{\beta}$. The last term in (2.14) considers the uncertainty due to the estimation of σ^2. Ignoring this term seriously underestimates the MSPE. However, there is no closed-form expression available for it, but an approximation of order $O(m^{-1})$ can be expressed by (Li and Lahiri [23], 2010)

$$\text{E}[\widehat{\theta}_i^{\text{EBLUP}} - \widehat{\theta}_i^{\text{BLUP}}]^2 = g_{3i}(\sigma^2) + o(m^{-1}),$$

where

$$g_{3i}(\sigma^2) = \frac{\psi_i^2}{(\sigma^2+\psi_i)^3}\overline{V}[\widehat{\sigma}^2]. \tag{2.16}$$

Therefore, a second-order approximation to $\text{MSPE}[\widehat{\theta}_i^{\text{EBLUP}}]$ in (2.13) or (2.14), under certain regularity conditions, is defined as

$$\text{MSPE}[\widehat{\theta}_i^{\text{EBLUP}}] = g_{1i}(\sigma^2) + g_{2i}(\sigma^2) + g_{3i}(\sigma^2) + o(m^{-1}). \tag{2.17}$$

It is noteworthy that both terms $g_{1i}(\sigma^2)$ and $g_{2i}(\sigma^2)$ given in (2.17) do not depend on the estimation method for σ^2 or B_i, but σ^2 affects the term $g_{3i}(\sigma^2)$ through $\overline{V}[\widehat{\sigma}^2]$. For the FH model, Datta *et al.* [10] (2005) and Datta [7] (2009) showed that the term $g_{3i}(\sigma^2)$ is the smallest in the ML and RML methods, but it is the largest in the PR and FH methods.

Note that $\text{MSPE}[\widehat{\theta}_i^{\text{EBLUP}}]$ defined in (2.17) depends on σ^2, which is unknown and hence cannot be used to asses the uncertainty of the EBLUP for a certain data set. Then, it is of interest to obtain a second-order unbiased estimator of $\text{MSPE}(\widehat{\theta}_i^{\text{EBLUP}})$, denoted as $\widehat{\text{MSPE}}[\widehat{\theta}_i^{\text{EBLUP}}]$, which must satisfy

$$\text{E}[\widehat{\text{MSPE}}[\widehat{\theta}_i^{\text{EBLUP}})] - \text{MSPE}[\widehat{\theta}_i^{\text{EBLUP}}] = o(m^{-1}).$$

Datta and Lahiri [9] (2000) derived a standard second-order unbiased approximation to the MSPE of the EBLUP, which is valid for all estimation methods of σ^2 discussed in this paper, and given by

$$\widehat{\text{MSPE}}[\widehat{\theta}_i^{\text{EBLUP}}] = g_{1i}(\widehat{\sigma}^2) + g_{2i}(\widehat{\sigma}^2) + 2g_{3i}(\widehat{\sigma}^2) - \widehat{B}_i^2\widehat{\text{Bias}}[\widehat{\sigma}^2], \tag{2.18}$$

where $g_{1i}(\widehat{\sigma}^2)$, $g_{2i}(\widehat{\sigma}^2)$ and $g_{3i}(\widehat{\sigma}^2)$ are defined in (2.15) and (2.16), respectively, when σ^2 is replaced by $\widehat{\sigma}^2$ and $\widehat{\text{Bias}}[\widehat{\sigma}^2]$ is a second-order unbiased estimator of $\text{Bias}[\widehat{\sigma}^2]$. It is important to note that a disadvantage of the method proposed by Li and Lahiri [23] (2010) for estimating σ^2 is that it can yield a negative value for the corresponding MSPE; see Yoshimori and Lahiri [42] (2014).

3. EM ALGORITHM IN THE ML ESTIMATION OF σ^2

3.1. The EM algorithm

Let $\boldsymbol{Y}_o$ be the random vector corresponding to the observed data $\boldsymbol{y}_o$, and $\boldsymbol{\theta}$ the parameter of interest corresponding to a $d \times 1$ vector with parameter space Θ. The vector $\boldsymbol{y}_o$ is viewed as being incomplete and is regarded as an observable function of the complete data. The random vector $\boldsymbol{Y}_c = (\boldsymbol{Y}_o^\top, \boldsymbol{U}^\top)^\top$ corresponds to the complete-data vector $\boldsymbol{y}_c = (\boldsymbol{y}_o^\top, \boldsymbol{u}^\top)^\top$, where $\boldsymbol{U}$ is the random vector associated with $\boldsymbol{u}$, the vector of unobserved or missing data. Let $\ell(\boldsymbol{\theta}|\boldsymbol{y}_o)$ be the log-likelihood function for $\boldsymbol{\theta}$ based on observed data. The EM algorithm approaches the problem of solving the incomplete-data likelihood equation $\partial\ell(\boldsymbol{\theta}|\boldsymbol{y}_o)/\partial\boldsymbol{\theta} = 0$ indirectly by proceeding in an iterative form in terms of the complete-data log-likelihood function, $\ell(\boldsymbol{\theta}|\boldsymbol{y}_c)$. As it is unobservable, it is replaced by its conditional expectation given $\boldsymbol{Y}_o = \boldsymbol{y}_o$, using a current estimate of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^{(0)}$ be a starting value for $\boldsymbol{\theta}$. Then, on the first iteration, the E-step of the EM algorithm requires the calculation of

$$Q \equiv Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) = \mathrm{E}[\ell(\boldsymbol{\theta}|\boldsymbol{Y}_c)|\boldsymbol{Y}_o, \boldsymbol{\theta}^{(0)}], \tag{3.1}$$

whereas its M-step needs the maximization of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ with respect to $\boldsymbol{\theta}$ over the parameter space $\boldsymbol{\Theta}$. Hence, we choose $\boldsymbol{\theta}^{(1)}$ such that $Q(\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}^{(0)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$, for all $\boldsymbol{\theta} \in \Theta$. The E-step and M-step must be iterated until reaching convergence, for example, when $|\ell(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{Y}_o) - \ell(\boldsymbol{\theta}^{(r)}|\boldsymbol{Y}_o| < 10^{-5}$, where $\widehat{\boldsymbol{\theta}}^{\,(r+1)}$ is the current ML estimate of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}^{\,(r)}$ is its previous estimate; see McLachlan and Krishnan [26, pp. 18-20] (2008). Thus, the $(r+1)$-th iteration of the EM algorithm consists of an E-step followed by an M-step described as:

E-step: Given $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{(r)}$, compute $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)}) = \mathrm{E}[\ell(\boldsymbol{\theta}|\boldsymbol{Y}_c)|\boldsymbol{Y}_o, \widehat{\boldsymbol{\theta}}^{(r)}]$.

M-step: Find $\widehat{\boldsymbol{\theta}}^{(r+1)}$ maximizing $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)})$ such that $Q(\boldsymbol{\theta}^{(r+1)}|\widehat{\boldsymbol{\theta}}^{(r)}) \geq Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(r)})$, for all $\boldsymbol{\theta} \in \Theta$.

3.2. The EM algorithm in the ML method for small area estimation

To solve the problem of negative or zero values when estimating the strictly positive variance components mentioned in Section 2.3, we propose to use the EM algorithm. Then, we derive the MLEM and RMLEM approaches. Let $\boldsymbol{Y}_o = \boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{b}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)^\top$. From (2.5), we have that

$$\boldsymbol{Y}_c = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{b} \end{pmatrix} \sim \mathrm{N}_{2m}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{V} & \sigma^2\boldsymbol{I}_m \\ \sigma^2\boldsymbol{I}_m & \sigma^2\boldsymbol{I}_m \end{bmatrix} \right),$$

with $\boldsymbol{V} = \boldsymbol{G} + \boldsymbol{S}$ given below (2.5). Then, the distribution of $\boldsymbol{b}$ conditional on $\boldsymbol{Y}$ is $\boldsymbol{b}|\boldsymbol{Y} = \boldsymbol{y} \sim \mathrm{N}_m(\sigma^2\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \sigma^2(\boldsymbol{I}_m - \sigma^2\boldsymbol{V}^{-1}))$. Thus, the log-likelihood function for $\boldsymbol{\theta}$ based on $\boldsymbol{y}_c$ can be expressed as $\ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}_c) = \ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}|\boldsymbol{b}) + \ell(\sigma^2; \boldsymbol{b})$. Hence, we have that

$$\ell(\beta, \sigma^2; \boldsymbol{y}_c) = c - \frac{1}{2}\log(|\boldsymbol{S}|) - \frac{1}{2}\boldsymbol{\varepsilon}^\top\boldsymbol{S}^{-1}\boldsymbol{\varepsilon} - \frac{1}{2}\log(|\sigma^2\boldsymbol{I}_m|) - \frac{1}{2\sigma^2}\boldsymbol{b}^\top\boldsymbol{b}, \tag{3.2}$$

where $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b}$ and c is a constant that is independent of σ^2.

Let $Q_1 \equiv Q_1(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}})$. By eliminating the constant term in (3.2) and according to (3.1), we obtain $Q_1 = \mathrm{E}[\ell(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{Y}_c) | \boldsymbol{Y}, \boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}]$ as

(3.3)
$$Q_1 = -\frac{1}{2}\log(|\boldsymbol{S}|) - \frac{1}{2}\mathrm{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{S}^{-1}\boldsymbol{\varepsilon} | \boldsymbol{Y}, \boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}] - \frac{1}{2}\log(|\sigma^2 \boldsymbol{I}_m|) - \frac{1}{2\sigma^2}\mathrm{E}[\boldsymbol{b}^\top \boldsymbol{b} | \boldsymbol{Y}, \boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}].$$

After some algebraic steps, we get

$$\mathrm{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{S}^{-1}\boldsymbol{\varepsilon} | \boldsymbol{Y}, \boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}] = \mathrm{tr}(\boldsymbol{S}^{-1}(\widetilde{\boldsymbol{\varepsilon}}_1 \widetilde{\boldsymbol{\varepsilon}}_1^\top + \mathrm{Var}[\boldsymbol{b}_1])), \tag{3.4}$$

with $\sigma^{2^{(0)}}$ and $\boldsymbol{\beta}^{(0)}$ being starting values for σ^2 and $\boldsymbol{\beta}$, respectively, where $\widetilde{\boldsymbol{\varepsilon}}_1 = \mathrm{E}[\boldsymbol{\varepsilon} | \boldsymbol{Y}, \boldsymbol{\beta}^{(0)} \sigma^{2^{(0)}}]$ $= \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \widetilde{\boldsymbol{b}}_1$, and

$$\widetilde{\boldsymbol{b}}_1 = \left(\frac{1}{\sigma^{2^{(0)}}}\boldsymbol{I}_m + \boldsymbol{S}^{-1}\right)^{-1} \boldsymbol{S}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(0)}), \quad \mathrm{Var}[\boldsymbol{b}_1] = \left(\frac{1}{\sigma^{2^{(0)}}}\boldsymbol{I}_m + \boldsymbol{S}^{-1}\right)^{-1}.$$

In addition, we have that

$$\mathrm{E}[\boldsymbol{b}^\top \boldsymbol{b} | \boldsymbol{Y}, \boldsymbol{\beta}^{(0)}, \sigma^{2^{(0)}}] = \mathrm{tr}(\widetilde{\boldsymbol{b}}_1 \widetilde{\boldsymbol{b}}_1^\top + \mathrm{Var}[\boldsymbol{b}_1]), \tag{3.5}$$

so that substituting (3.4) and (3.5) in (3.3), we obtain

$$Q_1 = -\frac{1}{2}\log(|\boldsymbol{S}|) - \frac{1}{2}\widetilde{\boldsymbol{\varepsilon}}_1^\top \boldsymbol{S}^{-1}\widetilde{\boldsymbol{\varepsilon}}_1 - \frac{1}{2}\log(|\sigma^2 \boldsymbol{I}_m|) - \frac{1}{2\sigma^2}\widetilde{\boldsymbol{b}}_1^\top \widetilde{\boldsymbol{b}}_1 - \frac{1}{2}\mathrm{tr}\left(\left(\frac{1}{\sigma^2}\boldsymbol{I}_m + \boldsymbol{S}^{-1}\right)\mathrm{Var}[\boldsymbol{b}_1]\right).$$

Maximizing Q_1 with respect to $\boldsymbol{\beta}$ and σ^2, we get

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{S}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{S}^{-1}(\boldsymbol{Y} - \widetilde{\boldsymbol{b}}_1), \quad \widehat{\sigma^2}^{(1)} = \frac{1}{m}\left(\widetilde{\boldsymbol{b}}_1^\top \widetilde{\boldsymbol{b}}_1 + \mathrm{tr}\left(\frac{1}{\sigma^{2^{(0)}}}\boldsymbol{I}_m + \boldsymbol{S}^{-1}\right)^{-1}\right).$$

Thus, the first main result of this study based on the EM algorithm, for the ML method used in the FH model, is described as follows:

Step 0. Set $r = 0$ and choose starting values $\boldsymbol{\beta}^{(0)}$ and $\sigma^{2^{(0)}}$.

Step 1. For $r \geq 0$, calculate

$$\widetilde{\boldsymbol{b}}_1^{(r+1)} = \left(\frac{1}{\sigma^{2^{(r)}}}\boldsymbol{I}_m + \boldsymbol{S}^{-1}\right)^{-1} \boldsymbol{S}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(r)}).$$

Step 2. For $r \geq 0$, compute

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{(r+1)} &= (\boldsymbol{X}^\top \boldsymbol{S}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{S}^{-1}(\boldsymbol{Y} - \widetilde{\boldsymbol{b}}_1^{(r+1)}), \\ \widehat{\sigma^2}^{(r+1)} &= \frac{1}{m}\left(\widetilde{\boldsymbol{b}}_1^{(r+1)\top} \widetilde{\boldsymbol{b}}_1^{(r+1)} + \mathrm{tr}\left(\frac{1}{\widehat{\sigma^2}^{(r)}}\boldsymbol{I}_m + \boldsymbol{S}^{-1}\right)^{-1}\right). \end{aligned}$$

Step 3. Iterate Steps 1 and 2 from $r = 1$ until reaching convergence when the difference in absolute value between the iterations $(r+1)$-th and r-th is less than a small preset precision value (for example10^{-5}).

The EM algorithm generates positive definite matrix estimates in Step 2, if the starting matrix is positive definite according to Thompson and Meyer [38] (1986), Searle *et al.* [37] (2006), Demidenko [11] (2013) and El-Leithy *et al.* [12] (2016) in the context of mixed models.

3.3. EM algorithm in the RML method for small area estimation

The joint distribution of $\boldsymbol{b}$ and $\boldsymbol{Y}_2$ (defined in Section 2.3) is given by

$$\begin{pmatrix} \boldsymbol{Y}_2 \\ \boldsymbol{b} \end{pmatrix} \sim \mathrm{N}_{2m-p}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{L}_2^\top \boldsymbol{V} \boldsymbol{L}_2 & \sigma^2 \boldsymbol{L}_2^\top \\ \sigma^2 \boldsymbol{L}_2 & \sigma^2 \boldsymbol{I}_m \end{bmatrix} \right).$$

From results of the multivariate normal distribution and after some matrix operations, we have that the distribution of $\boldsymbol{b}$ conditional on $\boldsymbol{Y}_2$ is $\boldsymbol{b}|\boldsymbol{Y}_2 = \boldsymbol{y}_2 \sim \mathrm{N}_m(\sigma^2 \boldsymbol{P} \boldsymbol{y}_2,\, \sigma^2(\boldsymbol{I}_m - \sigma^2 \boldsymbol{P}))$. Since we propose to use the EM algorithm to estimate σ^2 with the RML method, we rewrite the log-likelihood function for $\boldsymbol{Y}_c$ in (3.2) as

$$\ell(\sigma^2; \boldsymbol{Y}_c) = c - \frac{1}{2}\log(|\boldsymbol{S}|) - \frac{1}{2}\boldsymbol{\varepsilon}^\top \boldsymbol{S}^{-1}\boldsymbol{\varepsilon} - \frac{1}{2}\log(|\sigma^2 \boldsymbol{I}_m|) - \frac{1}{2\sigma^2}\boldsymbol{b}^\top \boldsymbol{b}. \tag{3.6}$$

Then, we maximize it conditional on $\boldsymbol{Y}_2$.

Let $Q_2 \equiv Q_2(\sigma^2|\sigma^{2^{(0)}})$. By eliminating the constant term in (3.6) and according to (3.1), we have that $Q_2 = \mathrm{E}[\ell(\sigma^2|\boldsymbol{Y}_c)|\boldsymbol{Y}_2, \sigma^{2^{(0)}}]$ is such that

$$Q_2 = -\frac{1}{2}\log(|\boldsymbol{S}|) - \frac{1}{2}\mathrm{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{S}^{-1}\boldsymbol{\varepsilon}|\boldsymbol{Y}_2, \sigma^{2^{(0)}}] - \frac{1}{2}\log(|\sigma^2 \boldsymbol{I}_m|) - \frac{1}{2\sigma^2}\mathrm{E}[\boldsymbol{b}^\top \boldsymbol{b}|\boldsymbol{Y}_2, \sigma^{2^{(0)}}]. \tag{3.7}$$

After some algebraic steps, we obtain

$$\mathrm{E}[\boldsymbol{\varepsilon}^\top \boldsymbol{S}^{-1}\boldsymbol{\varepsilon}|\boldsymbol{Y}_2, \sigma^{2^{(0)}}] = \mathrm{tr}(\boldsymbol{S}^{-1}(\widetilde{\boldsymbol{\varepsilon}}_2\, \widetilde{\boldsymbol{\varepsilon}}_2^\top + \mathrm{Var}[\boldsymbol{b}_2])), \tag{3.8}$$

where $\mathrm{Var}[\boldsymbol{b}_2] = \sigma^{2^{(0)}}(\boldsymbol{I}_m - \sigma^{2^{(0)}}\boldsymbol{P}^{(0)})$, $\widetilde{\boldsymbol{\varepsilon}}_2 = \mathrm{E}[\boldsymbol{\varepsilon}|\boldsymbol{Y}_2, \sigma^{2^{(0)}}] = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \widetilde{\boldsymbol{b}}_2$, with $\widetilde{\boldsymbol{b}}_2 = \sigma^{2^{(0)}}\boldsymbol{P}^{(0)}\boldsymbol{Y}$ and $\boldsymbol{P}^{(0)}$ being a starting value for $\boldsymbol{P}$. In addition, we have that

$$\mathrm{E}[\boldsymbol{b}^\top \boldsymbol{b}|\boldsymbol{Y}_2, \sigma^{2^{(0)}}] = \mathrm{tr}(\widetilde{\boldsymbol{b}}_2\, \widetilde{\boldsymbol{b}}_2^\top + \mathrm{Var}[\boldsymbol{b}_2]), \tag{3.9}$$

so that substituting (3.8) and (3.9) in (3.7), it conducts to

$$Q_2 = -\frac{1}{2}\log(|\boldsymbol{S}| - \frac{1}{2}\widetilde{\boldsymbol{\varepsilon}}_2^\top \boldsymbol{S}^{-1}\widetilde{\boldsymbol{\varepsilon}}_2 - \frac{1}{2}\log(|\sigma^2 \boldsymbol{I}_m|) - \frac{1}{2\sigma^2}\widetilde{\boldsymbol{b}}_2^\top \widetilde{\boldsymbol{b}}_2 - \frac{1}{2}\mathrm{tr}((\sigma^2 \boldsymbol{I}_m + \boldsymbol{S}^{-1})\,\mathrm{Var}[\boldsymbol{b}_2]).$$

Maximizing Q_2 with respect to σ^2, we obtain $\sigma^{2^{(1)}} = (1/m)(\widetilde{\boldsymbol{b}}_2^\top\, \widetilde{\boldsymbol{b}}_2 + \mathrm{tr}(\sigma^{2^{(0)}}(\boldsymbol{I}_m - \sigma^{2^{(0)}}\boldsymbol{P}^{(0)})))$. Thus, the second main result of this study based on the EM algorithm, for the RML method in the FH model, is summarized as follows:

Step 0. Set $s = 0$, and choose a starting value $\sigma^{2^{(0)}}$.

Step 1. For $s \geq 0$, calculate $\widetilde{\boldsymbol{b}}_2^{(s+1)} = \widehat{\sigma^2}^{(s)}\boldsymbol{P}^{(s)}\boldsymbol{Y}$.

Step 2. For $s \geq 0$, compute $\widehat{\sigma^2}^{(s+1)} = (1/m)(\widetilde{\boldsymbol{b}}_2^{(s+1)^\top}\, \widetilde{\boldsymbol{b}}_2^{(s+1)} + \mathrm{tr}(\widehat{\sigma^2}^{(s)}(\boldsymbol{I}_m - \widehat{\sigma^2}^{(s)}\boldsymbol{P}^{(s)})))$.

Step 3. Iterate Steps 1 and 2 from $r = 1$ until reaching convergence when the difference in absolute value between the iterations $(r+1)$-th and r-th is less a small preset precision value (for example10^{-5}).

4. MONTE CARLO SIMULATION STUDY

4.1. Scenario of the simulation

We present a multi-sample MC simulation study to compare the performance of the approaches proposed in this paper (MLEM and RMLEM) as alternative solutions to the problem of a negative or zero value in the estimate of σ^2. The multi-sample simulation is a common practice in MC procedures whenever we do not have an easy way to estimate measures of dispersion of a statistic, like for the MSE; see Fishman [15] (1973), Figueiredo and Gomes [14] (2004) and Gomes *et al.* [16, 17] (2011, 2016) for details about multi-sample MC simulation. The idea is reasonably simple: in a multi-sample simulation of size $R \times T$, instead of generating a sample of very large size of observed values of a statistic, $N_{\text{sim}} = R \times T$ say, we collect T observations of the statistic on each of the R independent replications of the experiment. The value of T also needs to be large enough to reduce the bias, and eventually provide asymptotic normality. Then, we take the average of the corresponding R estimates as overall estimate of the parameter of interest, where each estimate is computed from T runs. Thus, under very broad conditions, the overall estimator converges to normality as R increases. Moreover, we may estimate the standard error (SE) of this overall estimator, even if R is small. For small values of R, and whenever we may guarantee the asymptotic normality of the estimator for the parameter of interest, we may use the t-student distribution with $R-1$ degrees of freedom to approximate its true distribution. The performance of the approaches proposed in this paper is compared to the LML, LRML, YML and YRML methods, according to their percentage relative bias (PRB) and MSE, as well as the MSPE of the EBLUP estimator. We follow the same scenario used in Yoshimori and Lahiri [42] (2014) to do an effective comparison in relation to that work. Specifically, we consider the FH model defined in (2.1) with a common mean $\mu = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. As the MSE is invariant under translation, we set $\mu = 0$ without loss of generality. However, to account for the uncertainty in the estimation of the common mean that arises in practice, we treat the mean as unknown. We generate $R = 20$ independent MC replications with $T = 500$ runs ($N_{\text{sim}} = 20 \times 500 = 10,000$) of $\{Y_i,\, i = 1, ..., m\}$ using the FH model: $Y_i = b_i + \varepsilon_i$, where b_i and ε_i are independent with $b_i \overset{\text{IID}}{\sim} \text{N}(0, \sigma^2)$ and $\varepsilon_i \overset{\text{IND}}{\sim} \text{N}(0, \psi_i)$. We analyze both balanced (equal sampling variances ψ_i) and unbalanced (unequal sampling variances ψ_i) cases for different values of m. To examine the effect of the number of small areas on the performance of several estimators, we use values of $m \in \{15, 30, 45\}$. In the balanced case, we consider each of the combinations of m and ψ_i, where $\psi_i \in \{0.05, 0.1, 1, 10, 20\}$, fixing $\sigma^2 = 1$. We also examine the effect of σ^2/ψ_i on the performance of the estimators as in Yoshimori and Lahiri [42] (2014). In the unbalanced case, we also fix $\sigma^2 = 1$ and assume the following three patterns of sampling variances as in Yoshimori and Lahiri [42] (2014): (i) Pattern A, $\psi_i \in \{0.1, 0.4, 0.5, 0.6, 4.0\}$, where almost all (but one) of the sampling variances are smaller than σ^2; (ii) Pattern B, $\psi_i \in \{1.5, 2.0, 2.5, 3.0, 3.5\}$, where all sampling variances are slightly greater than σ^2; and (iii) Pattern C, $\psi_i \in \{2, 4, 5, 6, 20\}$, where not only are all sampling variances greater than σ^2, but one is much greater than σ^2, representing a case for extremely small area. Pattern A was also used by Datta and Lahiri [9] (2000) and Datta *et al.* [10] (2005), and Pattern C by Chen and Lahiri [3] (2008) in their simulation studies. In each pattern, we consider five groups (g) of small areas, each with three, six or nine small areas according to $m = 15$, $m = 30$ or

$m = 45$, such that the sampling variances ψ_i are the same within a given group. For example, in Pattern A for $m = 15$, we simulate three small areas for each case with sampling variances $\psi_i = 0.1, 0.4, 0.5, 0.6$ and 4.0. Similarly the other patterns of sampling variances and m were simulated.

4.2. Behavior of $\widehat{\sigma}^2$

The empirical probabilities of obtaining a zero estimate of σ^2 by different methods for balanced and unbalanced cases are reported in Tables 1 (balanced case) and 2 (unbalanced case). In both cases, the MLEM/RMLEM approaches and the LML/LRML/YML/YRML methods produce strictly positive estimates of σ^2. As mentioned in Yoshimori and Lahiri [42] (2014), only the ML and RML methods could yield negative or zero estimate of σ^2. For the balanced case and the ML/RML methods, the probability of getting negative or zero estimate increases as σ^2/ψ_i decreases in both methods, being slightly smaller in the RML method.

Table 1: Percentage of negative or zero estimate of σ^2 for the indicated m, variance ratio and method.

m	σ^2/ψ_i	ML	RML	LML	LRML	YML	YRML	MLEM	RMLEM
15	0.05	57.57	50.50	0	0	0	0	0	0
	0.1	52.28	44.89	0	0	0	0	0	0
	1	8.49	6.49	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0
	20	0	0	0	0	0	0	0	0
30	0.05	51.14	46.09	0	0	0	0	0	0
	0.1	43.91	38.97	0	0	0	0	0	0
	1	1.42	1.09	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0
	20	0	0	0	0	0	0	0	0
45	0.05	48.18	44.25	0	0	0	0	0	0
	0.1	39.68	36.02	0	0	0	0	0	0
	1	0.26	0.18	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0
	20	0	0	0	0	0	0	0	0

Table 2: Percentage of negative or zero estimate of σ^2 for the indicated m, pattern and method.

m	Pattern	ML	RML	LML	LRML	YML	YRML	MLEM	RMLEM
15	A	0.92	0.37	0	0	0	0	0	0
	B	27.33	21.65	0	0	0	0	0	0
	C	35.37	27.04	0	0	0	0	0	0
30	A	0.06	0.03	0	0	0	0	0	0
	B	13.87	11.29	0	0	0	0	0	0
	C	28.21	23.39	0	0	0	0	0	0
45	A	0	0	0	0	0	0	0	0
	B	7.50	6.24	0	0	0	0	0	0
	C	21.40	18.21	0	0	0	0	0	0

As m increases, this probability decreases and is very similar in both methods. In the unbalanced case, Pattern C (having an extreme value in the sampling variance) yields the largest percentages of negative or zero variance component estimates. Similarly to the balanced case, as m increases, this probability decreases, being smaller for the RML method than the ML method. In the remainder of this section, we consider only the performance of those methods mentioned above that produce strictly positive variance, because these methods solve the problem of inadmissibility presented in this paper (we do not further comparisons for the ML and RML methods). An aspect to be evaluated for assessing the performance of different estimators of σ^2 is their bias. We use the PRB of a given estimator of σ^2, $\widehat{\sigma}^{\,2}$ say, defined in our simulation study as the sample mean $\widehat{\mathrm{PRB}}[\widehat{\sigma}^{\,2}] = (1/R)\sum_{r=1}^{R}(\widehat{\mathrm{PRB}}_r[\widehat{\sigma}^{\,2})]$ of the PRB calculated on the $R = 20$ replications, with $\widehat{\mathrm{PRB}}_r[\widehat{\sigma}^{\,2}] = (1/T)\sum_{t=1}^{T}((\widehat{\sigma}^{\,2^{(t)}} - \sigma^2)/\sigma^2) \times 100$, where $\widehat{\sigma}^{\,2^{(r)}}$ denotes an estimate of σ^2 for the t-th instance in the r-th replication, with an associated SE defined as $((1/(R-1))\sum_{r=1}^{R}(\widehat{\mathrm{PRB}}_r[\widehat{\sigma}^{\,2}] - \widehat{\mathrm{PRB}}[\widehat{\sigma}^{\,2})]^2)^{1/2}$. The PRBs of estimators for σ^2 are presented in Tables 3 and 4 for balanced and unbalanced cases, respectively.

Table 3: PRB and its corresponding SE (in parentheses) of estimators of σ^2 for the indicated m, variance ratio and method.

m	σ^2/ψ_i	LML		LRML		YML		YRML		MLEM		RMLEM	
15	0.05	1094.36	(24.35)	1272.41	(27.71)	224.55	(18.36)	309.57	(21.42)	171.28	(19.42)	262.68	(22.69)
	0.1	541.57	(13.50)	636.53	(15.31)	94.98	(10.38)	143.26	(12.07)	69.39	(10.79)	121.47	(12.60)
	1	44.33	(2.96)	62.77	(3.25)	-11.38	(2.68)	1.38	(2.90)	-12.14	(2.71)	0.81	(2.93)
	10	10.61	(2.06)	20.65	(2.23)	-6.94	(1.79)	0.42	(1.92)	-6.95	(1.79)	0.42	(1.92)
	20	9.34	(1.64)	18.86	(1.78)	-6.61	(1.42)	0.42	(1.53)	-6.61	(1.42)	0.42	(1.53)
30	0.05	672.31	(13.29)	737.93	(14.26)	144.30	(11.8)	187.01	(12.62)	124.84	(12.02)	175.96	(12.76)
	0.1	322.88	(9.24)	358.37	(9.88)	53.59	(9.12)	78.31	(9.93)	44.51	(9.43)	73.99	(10.29)
	1	21.76	(2.17)	29.46	(2.26)	-6.19	(2.16)	0.44	(2.24)	-6.24	(2.16)	0.51	(2.24)
	10	4.89	(1.22)	9.14	(1.27)	-3.51	(1.14)	0.16	(1.18)	-3.51	(1.14)	0.17	(1.18)
	20	4.19	(1.26)	8.23	(1.31)	-3.44	(1.18)	0.06	(1.22)	-3.44	(1.18)	0.06	(1.22)
45	0.05	515.15	(17.7)	554.32	(18.67)	111.86	(16.27)	140.66	(17.42)	101.51	(16.38)	139.52	(17.67)
	0.1	243.74	(6.9)	265.20	(7.21)	39.55	(7.13)	56.53	(7.48)	34.93	(7.31)	57.02	(7.65)
	1	13.55	(1.98)	18.40	(2.04)	-4.94	(1.99)	-0.51	(2.03)	-4.95	(1.99)	-0.39	(2.04)
	10	3.08	(0.83)	5.78	(0.85)	-2.43	(0.79)	0.01	(0.81)	- 2.43	(0.79)	0.02	(0.81)
	20	2.70	(0.67)	5.26	(0.68)	-2.32	(0.64)	0.01	(0.65)	-2.32	(0.64)	0.01	(0.65)

Table 4: PRB and its corresponding SE (in parentheses) of estimators of σ^2 for the indicated m, pattern and method.

m	Pattern	LML		LRML		YML		YRML		MLEM		RMLEM	
15	A	26.01	(3.07)	42.66	(3.46)	-9.20	(2.43)	1.99	(2.7)	-9.28	(2.42)	1.97	(2.7)
	B	114.71	(5.43)	146.20	(6.02)	-3.76	(4.90)	15.63	(5.46)	-8.00	(5.09)	12.30	(5.65)
	C	258.95	(7.35)	323.53	(8.43)	24.72	(5.83)	58.94	(6.81)	14.73	(5.86)	48.72	(7.19)
30	A	10.19	(1.33)	16.72	(1.39)	-5.91	(1.24)	-0.51	(1.28)	-5.91	(1.24)	-0.47	(1.28)
	B	58.85	(4.26)	71.31	(4.49)	-6.07	(4.24)	4.24	(4.46)	-7.03	(4.30)	4.02	(4.53)
	C	123.79	(6.31)	145.68	(6.79)	0.97	(5.56)	16.90	(6.04)	-2.10	(5.63)	13.38	(6.20)
45	A	6.47	(1.43)	10.52	(1.47)	-4.00	(1.36)	-0.42	(1.39)	-4.00	(1.36)	-0.38	(1.39)
	B	39.56	(2.73)	47.34	(2.81)	-6.07	(2.87)	1.05	(2.93)	-6.40	(2.87)	1.38	(2.93)
	C	87.02	(4.51)	100.23	(4.73)	-1.16	(4.49)	9.85	(4.71)	-2.68	(4.52)	7.83	(4.95)

In the balanced case, when $\sigma^2/\psi_i < 1$, all methods widely overestimate σ^2, with the best performance in those methods based on the ML method than those based on the RML method. In this case, the performance of the MLEM approach is better having the smallest PRBs. When $\sigma^2/\psi_i \geq 1$, the RMLEM approach and the YRML method are the best

options being them very similar. Also, the MLEM approach and the YML method always underestimate σ^2. For the unbalanced case, the performance of the RMLEM approach and the YRML method are very similar and have the smallest PRBs in the following cases: (i) for Pattern A and all m (with the PRBs being always smallest for the proposed RMLEM approach) and (ii) for Pattern B, when $m = 30$ and $m = 45$. The performance of the MLEM approach is better for Pattern C when $m = 15$ than in the other cases. For all remaining situations (Pattern B with $m = 15$ and Pattern C with $m = 30, 45$), the smallest PRBs correspond to the YML method. In both balanced and unbalanced cases, as m increases, the performance of all estimators improves. We define the empirical percentage MSE of an estimator $\widehat{\sigma}^2$ of σ^2 based in our simulation study as (Yoshimori and Lahiri [42], 2014) $\widehat{\text{MSE}}_r[\widehat{\sigma}^2] = (1/T)\sum_{t=1}^{T}\left(\widehat{\sigma}^{2^{(t)}} - \sigma^2\right)^2 \times 100$, with $1 \le r \le R$, for the t-th instance in the r-th replication, where the overall MSE is then the sample mean $\widehat{\text{MSE}}[\widehat{\sigma}^2] = (1/R)\sum_{t=1}^{T}\widehat{\text{MSE}}_r[\widehat{\sigma}^2]$, with an associated SE given by $((1/(R-1))\sum_{r=1}^{R}(\widehat{\text{MSE}}_r[\widehat{\sigma}^{\,2}] - \widehat{\text{MSE}}[\widehat{\sigma}^{\,2})]^2)^{1/2}$. The empirical percentage MSEs of different estimators of σ^2 are shown in Tables 5 (balanced case) and 6 (unbalanced case). In Table 5, the performance of the MLEM approach and the YML method are better than the other ones, with a performance much better when σ^2/ψ_i is small, for all m. In other cases, when σ^2/ψ_i is large, all methods have a similar performance, particularly for $m = 45$, but the performance of the MLEM approach and the YML method are still slightly better than other methods. In the unbalanced case, again the MLEM approach and the YML method have a better performance than the other methods, followed by the RMLEM approach and the YRML method, for all patterns and values of m.

Table 5: Percentage MSE and its corresponding SE (in parentheses) of estimators of σ^2 for the indicated m, variance ratio and method.

m	σ^2/ψ_i	LML		LRML		YML		YRML		MLEM		RMLEM	
15	0.05	15530.4	(815.1)	20780.8	(1043.4)	2522.0	(354.8)	3726.0	(460.9)	2507.6	(352.9)	3722.4	(459.3)
	0.1	4031.4	(223.7)	5457.1	(287.1)	770.5	(98.7)	1114.5	(127.4)	789.7	(98.5)	1133.4	(127.0)
	1	74.5	(6.1)	105.5	(8.0)	46.2	(2.9)	52.9	(3.7)	47.4	(2.8)	53.9	(3.7)
	10	21.1	(1.6)	27.7	(2.1)	15.5	(1.0)	17.2	(1.2)	15.5	(1.0)	17.2	(1.2)
	20	20.0	(1.7)	26.0/	(2.2)	14.8	(1.1)	16.5	(1.4)	14.8	(1.1)	16.5	(1.4)
30	0.05	5940.1	(268.4)	7084.0	(308.8)	1332.7	(147.8)	1711.6	(174.0)	1348.0	(147.9)	1746.4	(174.0)
	0.1	1489.0	(90.5)	1795.1	(104.6)	414.3	(48.1)	517.9	(57.2)	426.2	(47.9)	530.9	(56.9)
	1	31.3	(2.0)	37.5	(2.4)	26.1	(1.6)	27.6	(1.7)	26.1	(1.6)	27.6	(1.7)
	10	9.3	(0.9)	10.6	(1.0)	8.1	(0.7)	8.5	(0.7)	8.1	(0.7)	8.5	(0.7)
	20	8.3	(0.7)	9.4	(0.8)	7.2	(0.5)	7.5	(0.6)	7.2	(0.5)	7.5	(0.6)
45	0.05	3545.6	(250.6)	4063.1	(279.6)	935.9	(138.9)	1133.6	(158.9)	947.8	(138.7)	1163.8	(158.5)
	0.1	888.1	(58.1)	1027.1	(64.3)	307.1	(36.6)	361.1	(41.3)	314.0	(36.6)	368.6	(41.3)
	1	19.2	(1.6)	21.7	(1.8)	17.6	(1.3)	18.1	(1.4)	17.6	(1.3)	18.1	(1.4)
	10	5.9	(0.3)	6.5	(0.4)	5.4	(0.3)	5.6	(0.3)	5.4	(0.3)	5.6	(0.3)
	20	5.2	(0.4)	5.7	(0.4)	4.7	(0.3)	4.9	(0.3)	4.7	(0.3)	4.9	(0.3)

Note that, although the results of the MSE in the estimation of the variance component in our simulation study are similar under the YML and MLEM methods, we observe that the estimation by means of the EM algorithm is slightly more accurate (with smaller SEs) than the estimation under the YML and YRML methods, such as occurs when comparing the YRML and RMLEM methods.

Table 6: Percentage MSE and its corresponding SE (in parentheses) of estimators of σ^2 for the indicated m, pattern and method.

m	Pattern	LML		LRML		YML		YRML		MLEM		RMLEM	
15	A	46.6	(5.8)	66.5	(8.1)	29.4	(2.7)	33.5	(3.7)	29.5	(2.7)	33.5	(3.7)
3	B	261.9	(22.0)	375.9	(28.9)	98.3	(8.6)	124.7	(11.8)	104.8	(8.4)	130.0	(11.5)
	C	990.7	(58.4)	1466.0	(78.5)	207.4	(26.6)	303.4	(34.3)	219.8	(26.8)	315.8	(34.4)
30	A	18.3	(1.7)	21.5	(2.0)	15.1	(1.2)	15.9	(1.4)	15.1	(1.2)	15.9	(1.4)
	B	94.2	(8.9)	117.2	(10.4)	59.9	(3.8)	66.3	(4.8)	61.5	(3.7)	67.2	(4.8)
	C	282.7	(28.5)	361.3	(34.1)	112.5	(13.8)	134.5	(16.9)	117.3	(13.8)	140.6	(16.9)
45	A	11.0	(0.7)	12.3	(0.8)	9.7	(0.5)	10.1	(0.6)	9.7	(0.5)	10.1	(0.6)
	B	54.0	(5.9)	63.6	(6.6)	42.4	(4.0)	44.8	(4.5)	42.9	(4.0)	44.9	(4.6)
	C	162.1	(13.9)	195.5	(15.8)	86.7	(7.9)	97.1	(9.2)	89.1	(7.9)	101.1	(9.2)

4.3. Estimator of the MSPE

We study the performance of the estimators of the MSPE of the EBLUP using different estimators of σ^2. Let $\theta_i^{(t)}$ and $\widehat{\theta}_i^{\mathrm{EBLUP}(t)}$ be the empirical mean and EBLUP corresponding to several estimators of σ^2 of the area i for the t-th simulation in the r-th replication, respectively, for $i = 1, ..., m,\ t = 1, ..., T$ and $r = 1, ..., R$, with $T = 500$, $R = 20$ and $N_{\mathrm{sim}} = 10,000$, as mentioned. The empirical MSPE of the EBLUP estimator in the area i for the r-th replication is given by $\widehat{\mathrm{MSPE}}_i^{(r)} = (1/T)\sum_{t=1}^{T}(\widehat{\theta}_i^{\mathrm{EBLUP}(t)} - \widehat{\theta}_i^{(t)})^2$. Then, the overall MSPE is $\widehat{\mathrm{MSPE}}_i = (1/R)\sum_{r=1}^{R}\widehat{\mathrm{MSPE}}_i^{(r)}$, with an associated SE given by $(\sum_{r=1}^{R}(\widehat{\mathrm{MSPE}}_i^{r} - \widehat{\mathrm{MSPE}}_i)^2/(R-1))^{1/2}$. Tables 7 (balanced case) and 8 (unbalanced case) report the empirical MSPEs of EBLUPs for different estimation methods of σ^2. In the balanced case, when σ^2/ψ_i is small, both the MLEM approach and the YML method present a performance slightly better than the RMLEM approach and the YRML method, and much better than the LML and LRML methods. This behavior is more pronounced when m is small. If $\sigma^2/\psi_i \geq 1$, all methods provide similar results, and for a fixed σ^2/ψ_i, the performance does not change substantially as m increases. For the unbalanced case, all methods have a similar performance.

Table 7: Empirical MSPE and its corresponding SE (in parentheses) of EBLUPs of $\widehat{\theta}_i$ for the indicated m, ratio and method.

m	σ^2/ψ_i	LML		LRML		YML		YRML		MLEM		RMLEM	
15	0.05	4.75	(0.15)	5.22	(0.16)	2.86	(0.11)	3.08	(0.12)	2.86	(0.11)	3.08	(0.12)
	0.1	2.68	(0.08)	2.90	(0.08)	1.85	(0.06)	1.96	(0.07)	1.87	(0.06)	1.97	(0.07)
	1	0.58	(0.01)	0.59	(0.01)	0.60	(0.01)	0.59	(0.01)	0.60	(0.01)	0.59	(0.01)
	10	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)
	20	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)
30	0.05	2.95	(0.06)	3.13	(0.07)	1.99	(0.05)	2.08	(0.05)	2.00	(0.05)	2.09	(0.05)
	0.1	1.81	(0.05)	1.90	(0.06)	1.43	(0.04)	1.47	(0.04)	1.44	(0.04)	1.48	(0.04)
	1	0.54	(0.01)	0.55	(0.01)	0.55	(0.01)	0.55	(0.01)	0.55	(0.01)	0.55	(0.01)
	10	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)
	20	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)
45	0.05	2.31	(0.07)	2.40	(0.08)	1.68	(0.06)	1.73	(0.06)	1.69	(0.06)	1.74	(0.06)
	0.1	1.53	(0.03)	1.57	(0.03)	1.29	(0.02)	1.31	(0.02)	1.29	(0.02)	1.32	(0.02)
	1	0.53	(0.01)	0.53	(0.01)	0.54	(0.01)	0.53	(0.01)	0.54	(0.01)	0.53	(0.01)
	10	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)
	20	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)	0.05	(<0.01)

Table 8: Empirical MSPE and its corresponding SE (in parentheses) of EBLUP of $\widehat{\theta}_i$ for indicated m, pattern group and method.

m	Pattern	g	LML		LRML		YML		YRML		MLEM		RMLEM	
15	A	1	0.92	(0.03)	0.94	(0.03)	0.91	(0.03)	0.92	(0.03)	0.92	(0.03)	0.92	(0.03)
		2	0.41	(0.01)	0.41	(0.01)	0.42	(0.01)	0.42	(0.01)	0.43	(0.01)	0.42	(0.01)
		3	0.36	(0.01)	0.36	(0.01)	0.37	(0.01)	0.37	(0.01)	0.37	(0.01)	0.37	(0.01)
		4	0.31	(0.01)	0.31	(0.01)	0.32	(0.01)	0.32	(0.01)	0.32	(0.01)	0.32	(0.01)
		5	0.09	(<0.01)	0.09	(<0.01)	0.10	(<0.01)	0.10	(<0.01)	0.10	(<0.01)	0.10	(<0.01)
	B	1	1.10	(0.04)	1.15	(0.04)	1.02	(0.04)	1.03	(0.04)	1.03	(0.04)	1.04	(0.05)
		2	1.07	(0.04)	1.11	(0.04)	1.00	(0.03)	1.01	(0.03)	1.02	(0.03)	1.02	(0.03)
		3	1.01	(0.04)	1.05	(0.04)	0.95	(0.04)	0.96	(0.04)	0.96	(0.04)	0.97	(0.04)
		4	0.93	(0.03)	0.96	(0.04)	0.88	(0.03)	0.89	(0.03)	0.90	(0.03)	0.90	(0.03)
		5	0.80	(0.04)	0.82	(0.04)	0.77	(0.04)	0.77	(0.04)	0.79	(0.04)	0.79	(0.04)
	C	1	1.70	(0.06)	1.86	(0.07)	1.39	(0.04)	1.43	(0.04)	1.40	(0.04)	1.44	(0.04)
		2	1.68	(0.08)	1.84	(0.09)	1.32	(0.05)	1.37	(0.06)	1.34	(0.05)	1.39	(0.06)
		3	1.61	(0.10)	1.75	(0.10)	1.27	(0.07)	1.32	(0.08)	1.29	(0.07)	1.34	(0.08)
		4	1.50	(0.07)	1.61	(0.07)	1.21	(0.06)	1.25	(0.06)	1.23	(0.06)	1.27	(0.07)
		5	1.12	(0.04)	1.18	(0.04)	0.96	(0.04)	0.98	(0.04)	0.99	(0.04)	1.01	(0.04)
30	A	1	0.86	(0.02)	0.86	(0.03)	0.86	(0.03)	0.86	(0.03)	0.86	(0.03)	0.86	(0.03)
		2	0.40	(0.01)	0.40	(0.01)	0.40	(0.01)	0.40	(0.01)	0.40	(0.01)	0.40	(0.01)
		3	0.35	(0.01)	0.35	(0.01)	0.35	(0.01)	0.35	(0.01)	0.35	(0.01)	0.35	(0.01)
		4	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)
		5	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)
	B	1	0.93	(0.03)	0.94	(0.03)	0.92	(0.03)	0.92	(0.03)	0.92	(0.03)	0.92	(0.03)
		2	0.90	(0.02)	0.91	(0.02)	0.89	(0.02)	0.89	(0.02)	0.89	(0.02)	0.89	(0.02)
		3	0.85	(0.02)	0.86	(0.02)	0.85	(0.02)	0.85	(0.02)	0.85	(0.02)	0.85	(0.02)
		4	0.80	(0.02)	0.81	(0.02)	0.80	(0.02)	0.80	(0.02)	0.81	(0.02)	0.80	(0.02)
		5	0.70	(0.01)	0.71	(0.01)	0.71	(0.01)	0.71	(0.01)	0.72	(0.01)	0.71	(0.01)
	C	1	1.24	(0.04)	1.27	(0.05)	1.17	(0.04)	1.18	(0.04)	1.18	(0.04)	1.18	(0.04)
		2	1.20	(0.03)	1.24	(0.03)	1.11	(0.03)	1.12	(0.03)	1.11	(0.03)	1.13	(0.03)
		3	1.16	(0.03)	1.20	(0.03)	1.08	(0.03)	1.09	(0.03)	1.09	(0.03)	1.10	(0.04)
		4	1.12	(0.03)	1.16	(0.04)	1.03	(0.03)	1.04	(0.03)	1.04	(0.03)	1.05	(0.03)
		5	0.91	(0.02)	0.93	(0.02)	0.87	(0.02)	0.87	(0.02)	0.88	(0.02)	0.89	(0.02)
45	A	1	0.83	(0.02)	0.83	(0.02)	0.83	(0.02)	0.83	(0.02)	0.83	(0.02)	0.83	(0.02)
		2	0.39	(0.01)	0.39	(0.01)	0.39	(0.01)	0.39	(0.01)	0.39	(0.01)	0.39	(0.01)
		3	0.34	(0.01)	0.34	(0.01)	0.34	(0.01)	0.34	(0.01)	0.34	(0.01)	0.34	(0.01)
		4	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)	0.30	(0.01)
		5	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)	0.09	(<0.01)
	B	1	0.88	(0.02)	0.88	(0.02)	0.88	(0.02)	0.88	(0.02)	0.88	(0.02)	0.88	(0.02)
		2	0.85	(0.02)	0.86	(0.02)	0.86	(0.02)	0.86	(0.02)	0.86	(0.02)	0.86	(0.02)
		3	0.80	(0.02)	0.81	(0.02)	0.81	(0.02)	0.81	(0.02)	0.81	(0.02)	0.81	(0.02)
		4	0.75	(0.01)	0.75	(0.01)	0.76	(0.02)	0.76	(0.02)	0.76	(0.02)	0.76	(0.01)
		5	0.67	(0.02)	0.68	(0.02)	0.69	(0.02)	0.69	(0.02)	0.69	(0.02)	0.69	(0.02)
	C	1	1.13	(0.04)	1.14	(0.04)	1.10	(0.03)	1.11	(0.03)	1.10	(0.03)	1.11	(0.03)
		2	1.07	(0.03)	1.09	(0.03)	1.03	(0.03)	1.03	(0.03)	1.03	(0.03)	1.04	(0.03)
		3	1.05	(0.03)	1.07	(0.03)	1.01	(0.03)	1.02	(0.03)	1.02	(0.03)	1.02	(0.03)
		4	1.01	(0.03)	1.03	(0.03)	0.97	(0.03)	0.98	(0.03)	0.98	(0.03)	0.99	(0.03)
		5	0.83	(0.02)	0.84	(0.02)	0.82	(0.02)	0.82	(0.02)	0.83	(0.02)	0.83	(0.02)

5. APPLICATION

We illustrate the results of this study with real-world data taken from http://dx.doi.org/10.7927/H4FF3Q9B; see CIESIN [4] (2005). The small areas in our application correspond to the 32 states of Mexico. Here, m_i is the number of municipalities in the state i that were used for direct estimation within each state, which ranges from 5 to 578 municipalities.

We are interested in modeling the "average monthly per capita food expenditure for rural household in 2000 (AMPCFERH)". The direct estimator $\widehat{\theta}_i$ is available. We use three auxiliary variables: (i) the size of the illiterate population aged 15 years and above (X_1); (ii) the percentage of the population living in rural areas (X_2); and (iii) the fraction of rural households below the food poverty line (X_3). Our small area model is given by

$$\widehat{\theta}_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + b_i + \varepsilon_i, \quad i = 1, ..., 32,$$

where $b_i \overset{\text{IID}}{\sim} \text{N}(0, \sigma^2)$ are area-specific random effects with unknown σ^2; $\varepsilon_i \overset{\text{IND}}{\sim} \text{N}(0, \psi_i)$ represent the sampling errors in the area i with known variance ψ_i, which was calculated from the variance of the AMPCFERH within each municipality. Note that $\psi_i \neq \psi_j$, for $i \neq j$, and hence we are in the unbalanced case. We estimate σ^2 with the LML, LRML, YML, YRML, MLEM and RMLEM methods obtaining the following values: $\widehat{\sigma}^2_{\text{LML}} = 105147.20$, $\widehat{\sigma}^2_{\text{LRML}} = 123306.20$, $\widehat{\sigma}^2_{\text{YML}} = 93504.80$, $\widehat{\sigma}^2_{\text{YLML}} = 108737.40$, $\widehat{\sigma}^2_{\text{MLEM}} = 93503.32$ and $\widehat{\sigma}^2_{\text{RMLEM}} = 108735.6$. The results of the EBLUP of $\widehat{\theta}_i$ generated under different estimators of σ^2 are shown in Table 9.

Table 9: EBLUP of $\widehat{\theta}_i$ with estimates of σ^2 for the indicated Mexican state, m_i and method.

State	m_i	$\widehat{\theta}_i$	LML	LRML	YML	YRML	MLEM	RMLEM
Aguascalientes (Ags)	11	939.35	918.37 ± 21.27	921.38 ± 18.11	915.63 ± 24.81	918.88 ± 21.21	883.70 ± 164.85	921.38 ± 18.11
Baja California (BC)	5	1169.42	1126.60 ± 36.11	1132.35 ± 31.93	1121.75 ± 39.89	1127.78 ± 35.52	1061.41 ± 327.17	1132.35 ± 31.93
Baja California Sur (BCS)	5	1356.74	995.54 ± 182.87	1035.31 ± 166.05	963.68 ± 200.21	1003.73 ± 183.40	912.80 ± 420.93	1035.31 ± 166.05
Campeche (Camp)	11	460.49	479.69 ± 18.28	477.29 ± 15.77	481.72 ± 20.75	479.23 ± 17.97	504.41 ± 148.49	477.29 ± 15.77
Chiapas (Chis)	119	325.82	340.69 ± 15.38	338.74 ± 13.19	342.42 ± 17.79	340.35 ± 15.30	357.94 ± 103.04	338.74 ± 13.19
Chihuahua (Chih)	67	962.03	871.35 ± 94.37	879.37 ± 86.44	864.92 ± 101.19	872.73 ± 93.51	844.74 ± 166.78	879.37 ± 86.44
Coahuila (Coah)	38	879.43	856.88 ± 41.54	860.06 ± 35.87	853.96 ± 47.65	857.35 ± 41.47	834.94 ± 133.54	860.06 ± 35.87
Colima (Col)	10	803.55	789.86 ± 20.16	791.98 ± 17.31	787.95 ± 23.02	790.23 ± 19.89	752.55 ± 189.81	791.98 ± 17.31
Distrito Federal (DF)	7	1526.98	409.99 ± 335.33	459.56 ± 342.75	374.67 ± 332.40	418.91 ± 339.88	355.75 ± 364.64	459.56 ± 342.75
Durango (Dgo)	39	754.39	788.84 ± 47.27	784.37 ± 42.01	792.78 ± 52.38	788.10 ± 46.84	812.29 ± 139.89	784.37 ± 42.01
Guanajuato (Gto)	46	597.01	593.20 ± 14.86	593.98 ± 12.64	592.41 ± 17.36	593.32 ± 14.77	576.37 ± 92.82	593.98 ± 12.64
Guerrero (Gro)	76	605.11	550.19 ± 60.47	557.22 ± 53.54	544.14 ± 66.68	551.49 ± 59.49	495.19 ± 302.22	557.22 ± 53.54
Hidalgo (Hgo)	85	641.63	630.10 ± 43.78	631.72 ± 38.25	628.68 ± 48.94	630.41 ± 43.06	611.97 ± 127.91	631.72 ± 38.25
Jalisco (Jal)	124	808.46	657.37 ± 112.62	664.11 ± 110.76	652.36 ± 114.25	658.54 ± 112.67	639.18 ± 131.28	664.11 ± 110.76
México (Méx)	123	1036.92	622.15 ± 167.89	628.90 ± 167.92	617.30 ± 168.26	623.35 ± 168.21	597.08 ± 180.24	628.90 ± 167.92
Michoacán (Mich)	113	577.81	581.91 ± 17.52	581.69 ± 15.00	582.02 ± 20.23	581.85 ± 17.33	577.63 ± 55.60	581.69 ± 15.00
Morelos (Mor)	33	926.05	787.61 ± 107.11	804.31 ± 96.10	773.48 ± 118.51	790.61 ± 106.99	719.25 ± 310.25	804.31 ± 96.10
Nayarit (Nay)	20	677.62	668.68 ± 34.96	670.25 ± 30.90	667.21 ± 38.93	668.92 ± 34.66	656.15 ± 78.34	670.25 ± 30.90
Nuevo León (NL)	50	1232.20	859.37 ± 132.73	883.75 ± 131.00	841.01 ± 135.47	863.66 ± 134.46	814.11 ± 203.67	883.75 ± 131.00
Oaxaca (Oax)	578	412.72	576.55 ± 100.21	559.66 ± 92.17	590.51 ± 109.23	573.70 ± 101.17	618.84 ± 199.55	559.66 ± 92.17
Puebla (Pue)	223	423.22	462.18 ± 25.78	457.26 ± 22.49	466.51 ± 29.59	461.34 ± 25.91	491.48 ± 139.54	457.26 ± 22.49
Querétaro (Qro)	18	641.84	641.50 ± 18.55	641.64 ± 16.00	641.37 ± 21.14	641.53 ± 18.32	634.05 ± 73.27	641.64 ± 16.00
Quintana Roo (QR)	8	632.76	589.84 ± 73.18	596.10 ± 64.57	584.20 ± 81.90	590.85 ± 72.60	550.42 ± 205.72	596.10 ± 64.57
San Luis Potosí (SLP)	58	504.47	536.93 ± 24.67	532.63 ± 21.44	540.76 ± 28.47	536.20 ± 24.78	567.95 ± 146.40	532.63 ± 21.44
Sinaloa (Sin)	18	927.84	826.22 ± 66.88	838.54 ± 59.69	815.67 ± 74.98	828.40 ± 67.41	765.16 ± 278.01	838.54 ± 59.69
Sonora (Son)	72	989.80	954.36 ± 51.44	958.16 ± 45.92	951.17 ± 56.45	955.01 ± 50.76	937.69 ± 123.42	958.16 ± 45.92
Tabasco (Tab)	17	541.45	554.30 ± 10.05	552.53 ± 8.67	555.89 ± 11.56	553.98 ± 10.00	586.06 ± 161.98	552.53 ± 8.67
Tamaulipas (Tamps)	41	793.39	798.74 ± 33.53	798.14 ± 29.30	799.22 ± 37.98	798.61 ± 33.23	800.82 ± 84.05	798.14 ± 29.30
Tlaxcala (Tlax)	51	777.61	742.60 ± 66.61	747.53 ± 58.27	738.16 ± 75.27	743.37 ± 66.16	715.98 ± 157.75	747.53 ± 58.27
Veracruz (Ver)	216	515.10	515.72 ± 59.07	516.32 ± 52.16	515.10 ± 65.23	515.83 ± 58.07	501.88 ± 126.33	516.32 ± 52.16
Yucatán (Yuc)	100	344.22	369.75 ± 23.10	366.72 ± 20.06	372.30 ± 25.99	369.16 ± 22.72	392.23 ± 136.02	366.72 ± 20.06
Zacatecas (Zac)	57	842.43	836.09 ± 15.82	836.93 ± 13.70	835.33 ± 18.11	836.23 ± 15.62	826.72 ± 73.82	836.93 ± 13.70

Following to Figueiredo and Gomes [14] (2004), we have generated 10000 bootstrap samples of size 32 (the same as the number of states) to calculate the SEs. With these SEs, we build the corresponding bootstrap confidence intervals ($\text{BCI}_{95\%}$) for each EBLUP, using a confidence level of 95%. For the EBLUP, these intervals are obtained as

$$\text{BCI}_{95\%}(\widehat{\theta}_i^{\text{EBLUP}}) = \left[\overline{\widehat{\theta}}_i^{\text{EBLUP}(B)} \pm z_{1-\alpha/2} \text{SD}\left(\widehat{\theta}_i^{\text{EBLUP}(B)}\right) \right], \tag{5.1}$$

where $\overline{\widehat{\theta}}_i^{\text{EBLUP}(B)}$ and $\text{SD}(\widehat{\theta}_i^{\text{EBLUP}(B)})$ are the mean and standard deviation bootstrap, respectively, and $z_{1-\alpha/2}$ is the $(1-\alpha/2) \times 100$-th percentile of the standard normal distribution. The results for BCIs based on (5.1) with the Mexican data are resented in Table 9. Figure 1 shows the estimated AMPCFERH, where the states are colored according to a classification into quartiles for the values of the estimates by the RMLEM approach. Note that, in general, the states with smaller AMPCFERH are mainly concentrated in the southwest part of the country, except for Quintana Roo, while the states with larger AMPCFERH are located at the north part. A measure of uncertainty of the EBLUP of $\widehat{\theta}_i$ is given in Table 10.

Figure 1: Estimated average monthly per capita food expenditure for rural household in 2000 with the RMLEM approach for Mexican data.

We compute the MSPE and its corresponding $\text{BCI}_{95\%}$, similarly as for the EBLUP given in (5.1), using (2.18) for each of the estimation methods. We highlight two aspects: (i) the values obtained by the MLEM and RMLEM approaches are similar to the corresponding values of the YML and RYML methods, and moreover, its variability presents the same results; and (ii) note that the LML and LRML approaches show lower MSPE indicators, but, as mentioned, these methods can provide a negative value for the MSPE, making their results underestimated and unreliable. As in Molina *et al.* [28] (2014), we calculate the coefficients of variation (CVs), and its correspondig $\text{BCI}_{95\%}$ in terms of the MSPE estimates as $\text{CV}[\widehat{\theta}_i^{\text{EBLUP}}] = ((\text{MSPE}(\widehat{\theta}_i^{\text{EBLUP}}))^{1/2}/\widehat{\theta}_i^{\text{EBLUP}}) \times 100$ for each bootstrap sample, in order to analyze the gain in efficiency of the $\widehat{\theta}_i^{\text{EBLUP}}$ in comparison with direct estimates. Table 11 displays these CVs, from which the two following aspects can be mentioned: (i) there is a clear overall gain in precision when the EBLUP of $\widehat{\theta}_i$ is obtained with the YML/YRML/MLEM/RMLEM methods, if σ^2 is estimated, since in almost all cases the CVs are less than the CVs of the direct estimator; and (ii) in general, this gain in precision has a greater effect

in the RMLEM method, and moreover, the variability obtained by the bootstrapp resampling shows that it also has less variability in comparison to the rest of the methods.

Table 10: MSPE of EBLUP of $\widehat{\theta}_i$ with estimates of σ^2 for the indicated Mexican state, m_i and method.

State	m_i	$\widehat{\theta}_i$	LML	LRML	YML	YRML	MLEM	RMLEM
Aguascalientes	11	–	7062.57 ± 88.31	7072.20 ± 84.91	7065.37 ± 87.95	7063.48 ± 87.76	7065.36 ± 87.95	7063.47 ± 87.76
Baja California	5	–	8533.92 ± 113.74	8537.96 ± 110.40	8545.70 ± 112.33	8533.08 ± 113.05	8545.69 ± 112.33	8533.07 ± 113.05
Baja California Sur	5	–	38291.19 ± 3898.23	38779.60 ± 3658.90	38101.79 ± 4049.52	38373.13 ± 3851.71	38101.59 ± 4049.52	38372.91 ± 3851.71
Campeche	11	–	6002.27 ± 58.62	6005.50 ± 56.77	6007.47 ± 57.86	6002.09 ± 58.32	6007.46 ± 57.86	6002.09 ± 58.32
Chiapas	119	–	4727.16 ± 36.06	4728.90 ± 34.86	4730.72 ± 35.61	4726.98 ± 35.83	4730.72 ± 35.61	4726.98 ± 35.83
Chihuahua	67	–	43346.23 ± 3745.72	43585.34 ± 3567.53	43361.26 ± 3843.81	43372.80 ± 3706.31	43361.08 ± 3843.81	43372.60 ± 3706.31
Coahuila	38	–	13307.16 ± 322.52	13340.25 ± 309.35	13314.31 ± 323.26	13310.50 ± 320.07	13314.29 ± 323.26	13310.48 ± 320.07
Colima	10	–	4821.23 ± 42.43	4824.96 ± 40.71	4823.34 ± 42.17	4821.47 ± 42.13	4823.34 ± 42.17	4821.47 ± 42.13
Distrito Federal	7	–	111400.65 ± 32054.55	113227.77 ± 30823.28	110394.68 ± 33003.71	111738.73 ± 31720.72	110393.75 ± 33003.71	111737.70 ± 31720.72
Durango	39	–	17327.26 ± 542.56	17353.87 ± 522.28	17358.39 ± 543.24	17326.94 ± 538.59	17358.36 ± 543.24	17326.91 ± 538.59
Guanajuato	46	–	6224.48 ± 74.76	6232.15 ± 71.48	6226.63 ± 74.66	6225.22 ± 74.12	6226.63 ± 74.66	6225.22 ± 74.12
Guerrero	76	–	18966.35 ± 589.66	18872.63 ± 556.05	19098.20 ± 606.77	18939.14 ± 579.86	19098.18 ± 606.77	18939.12 ± 579.86
Hidalgo	85	–	15972.62 ± 491.00	16015.04 ± 469.76	15984.87 ± 494.07	15976.51 ± 486.51	15984.84 ± 494.07	15976.48 ± 486.51
Jalisco	124	–	95196.34 ± 29228.59	97750.38 ± 28174.99	93689.46 ± 30058.90	95683.09 ± 28945.16	93688.45 ± 30058.90	95681.96 ± 28945.16
México	123	–	171437.83 ± 54229.90	174617.81 ± 52689.00	169339.63 ± 55318.79	172075.09 ± 53758.50	169338.20 ± 55318.79	172073.49 ± 53758.50
Michoacán	113	–	9280.13 ± 175.17	9298.77 ± 167.71	9282.71 ± 175.54	9282.20 ± 173.85	9282.70 ± 175.54	9282.19 ± 173.85
Morelos	33	–	26572.95 ± 1626.86	26759.52 ± 1546.85	26534.16 ± 1661.29	26600.49 ± 1611.46	26534.08 ± 1661.29	26600.39 ± 1611.46
Nayarit	20	–	22314.46 ± 1123.78	22447.65 ± 1070.15	22291.05 ± 1142.82	22333.72 ± 1113.91	22290.98 ± 1142.82	22333.66 ± 1113.91
Nuevo León	50	–	68062.42 ± 12681.30	69275.95 ± 12132.42	67486.95 ± 13078.86	68276.16 ± 12554.36	67486.42 ± 13078.86	68275.58 ± 12554.36
Oaxaca	578	–	35112.56 ± 2740.44	35374.18 ± 2611.48	35068.08 ± 2802.07	35149.58 ± 2712.87	35067.93 ± 2802.07	35149.43 ± 2712.87
Puebla	223	–	10943.82 ± 240.24	10967.14 ± 230.28	10948.83 ± 240.76	10946.20 ± 238.56	10948.81 ± 240.76	10946.18 ± 238.56
Querétaro	18	–	8204.86 ± 134.72	8217.40 ± 129.18	8208.76 ± 134.56	8206.02 ± 133.81	8208.75 ± 134.56	8206.01 ± 133.81
Quintana Roo	8	–	19958.04 ± 797.45	20018.77 ± 765.16	19976.57 ± 803.66	19963.41 ± 791.20	19976.53 ± 803.66	19963.36 ± 791.20
San Luis Potosí	58	–	9360.25 ± 174.23	9377.92 ± 166.87	9363.88 ± 174.46	9362.08 ± 173.01	9363.87 ± 174.46	9362.07 ± 173.01
Sinaloa	18	–	21443.10 ± 972.04	21545.23 ± 928.59	21438.25 ± 984.28	21456.33 ± 963.25	21438.19 ± 984.28	21456.27 ± 963.25
Sonora	72	–	23983.14 ± 1109.54	24042.02 ± 1065.55	24024.91 ± 1119.54	23985.42 ± 1100.99	24024.85 ± 1119.54	23985.35 ± 1100.99
Tabasco	17	–	3501.87 ± 22.68	3503.64 ± 21.93	3503.21 ± 22.36	3501.95 ± 22.57	3503.21 ± 22.36	3501.95 ± 22.57
Tamaulipas	41	–	16459.47 ± 547.96	16507.93 ± 525.90	16469.43 ± 550.71	16464.38 ± 544.25	16469.40 ± 550.71	16464.34 ± 544.25
Tlaxcala	51	–	20426.38 ± 867.03	20518.08 ± 828.69	20423.87 ± 876.94	20438.07 ± 859.53	20423.82 ± 876.94	20438.02 ± 859.53
Veracruz	216	–	22062.93 ± 993.54	22146.37 ± 950.78	22075.62 ± 1005.02	22071.75 ± 984.50	22075.56 ± 1005.02	22071.69 ± 984.50
Yucatán	100	–	7949.13 ± 112.31	7949.75 ± 109.16	7961.83 ± 110.59	7947.76 ± 111.89	7961.82 ± 110.59	7947.75 ± 111.89
Zacatecas	57	–	8189.27 ± 131.47	8198.87 ± 126.70	8195.48 ± 130.75	8189.79 ± 130.78	8195.47 ± 130.75	8189.78 ± 130.78

Table 11: CVs of direct estimator and EBLUP of $\widehat{\theta}_i$ with estimates of σ^2 for the indicated Mexican state, m_i and method.

State	m_i	$\widehat{\theta}_i$	LML	LRML	YML	YRML	MLEM	RMLEM
Aguascalientes	11	9.21	9.15 ± 0.21	9.13 ± 0.18	9.18 ± 0.25	9.15 ± 0.21	9.51 ± 0.24	9.12 ± 0.18
Baja California	5	8.14	8.20 ± 0.27	8.16 ± 0.24	8.24 ± 0.30	8.19 ± 0.27	8.71 ± 0.29	8.16 ± 0.24
Baja California Sur	5	17.58	19.66 ± 4.56	19.02 ± 3.76	20.26 ± 5.56	19.52 ± 4.52	21.38 ± 4.45	18.92 ± 3.77
Campeche	11	17.20	16.15 ± 0.60	16.24 ± 0.53	16.09 ± 0.68	16.17 ± 0.59	15.37 ± 0.62	16.23 ± 0.53
Chiapas	119	21.47	20.18 ± 0.24	20.30 ± 0.78	20.09 ± 1.01	20.20 ± 0.88	19.22 ± 0.90	20.30 ± 0.78
Chihuahua	67	25.74	23.89 ± 3.06	23.74 ± 2.74	24.08 ± 3.34	23.86 ± 3.02	24.65 ± 2.77	23.68 ± 2.76
Coahuila	38	13.85	13.46 ± 0.68	13.43 ± 0.58	13.51 ± 0.80	13.46 ± 0.68	13.82 ± 0.72	13.41 ± 0.59
Colima	10	8.81	8.79 ± 0.23	8.77 ± 0.20	8.81 ± 0.26	8.79 ± 0.23	9.23 ± 0.25	8.77 ± 0.20
Distrito Federal	7	39.79	81.41 ± 80.28	73.22 ± 72.47	88.68 ± 83.59	79.80 ± 77.55	93.40 ± 79.40	72.74 ± 70.28
Durango	39	18.60	16.69 ± 1.01	16.79 ± 0.91	16.62 ± 1.11	16.70 ± 1.00	16.22 ± 0.98	16.78 ± 0.91
Guanajuato	46	13.55	13.30 ± 0.35	13.29 ± 0.30	13.32 ± 0.41	13.30 ± 0.35	13.69 ± 0.39	13.28 ± 0.30
Guerrero	76	23.69	25.03 ± 3.17	24.65 ± 2.71	25.40 ± 3.60	24.95 ± 3.09	27.91 ± 2.97	24.70 ± 2.72
Hidalgo	85	21.00	20.06 ± 1.55	20.03 ± 1.35	20.11 ± 1.75	20.05 ± 1.53	20.66 ± 1.51	20.01 ± 1.35
Jalisco	124	77.60	46.94 ± 11.61	47.08 ± 11.20	46.92 ± 11.97	46.97 ± 11.56	47.89 ± 9.14	46.58 ± 11.29
México	123	130.27	66.55 ± 24.00	66.44 ± 23.43	66.66 ± 24.51	66.55 ± 23.95	68.92 ± 17.91	65.96 ± 23.48
Michoacán	113	17.33	16.55 ± 0.53	16.58 ± 0.46	16.55 ± 0.60	16.56 ± 0.52	16.68 ± 0.56	16.56 ± 0.46
Morelos	33	19.88	20.70 ± 3.34	20.34 ± 2.84	21.06 ± 3.94	20.63 ± 3.33	22.65 ± 3.23	20.28 ± 2.85
Nayarit	20	24.39	22.34 ± 1.35	22.35 ± 1.20	22.38 ± 1.49	22.34 ± 1.34	22.75 ± 1.30	22.30 ± 1.21
Nuevo León	50	30.82	30.36 ± 5.98	29.78 ± 5.61	30.89 ± 6.36	30.25 ± 5.99	31.91 ± 5.06	29.57 ± 5.65
Oaxaca	578	53.03	32.50 ± 5.71	33.61 ± 5.56	31.71 ± 5.90	32.68 ± 5.77	30.26 ± 4.71	33.50 ± 5.57
Puebla	223	25.85	22.63 ± 1.28	22.90 ± 1.14	22.43 ± 1.42	22.68 ± 1.28	21.29 ± 1.24	22.88 ± 1.14
Querétaro	18	14.59	14.12 ± 0.43	14.13 ± 0.37	14.13 ± 0.48	14.12 ± 0.42	14.29 ± 0.45	14.12 ± 0.37
Quintana Roo	8	24.17	23.95 ± 3.47	23.74 ± 2.95	24.19 ± 4.10	23.91 ± 3.46	25.68 ± 3.35	23.70 ± 2.96
San Luis Potosí	58	19.93	18.02 ± 0.81	18.18 ± 0.71	17.89 ± 0.92	18.05 ± 0.81	17.04 ± 0.82	18.17 ± 0.72
Sinaloa	18	17.31	17.72 ± 1.66	17.50 ± 1.44	17.95 ± 1.91	17.68 ± 1.65	19.14 ± 1.63	17.47 ± 1.45
Sonora	72	17.13	16.23 ± 0.96	16.18 ± 0.86	16.30 ± 1.06	16.22 ± 0.95	16.53 ± 0.94	16.16 ± 0.87
Tabasco	17	11.08	10.68 ± 0.19	10.71 ± 0.17	10.65 ± 0.22	10.68 ± 0.19	10.10 ± 0.21	10.71 ± 0.17
Tamaulipas	41	17.29	16.06 ± 0.70	16.10 ± 0.62	16.06 ± 0.78	16.07 ± 0.69	16.03 ± 0.70	16.08 ± 0.62
Tlaxcala	51	20.06	19.25 ± 1.96	19.16 ± 1.68	19.36 ± 2.27	19.23 ± 1.95	19.96 ± 1.93	19.12 ± 1.69
Veracruz	216	31.55	28.80 ± 3.66	28.82 ± 3.21	28.84 ± 4.08	28.80 ± 3.59	29.60 ± 3.33	28.77 ± 3.22
Yucatán	100	26.60	24.11 ± 1.46	24.31 ± 1.29	23.97 ± 1.62	24.15 ± 1.44	22.75 ± 1.40	24.31 ± 1.29
Zacatecas	57	11.09	10.82 ± 0.22	10.82 ± 0.19	10.84 ± 0.25	10.82 ± 0.22	10.95 ± 0.24	10.81 ± 0.20

6. CONCLUSIONS AND FUTURE RESEARCH

One of the advantages of using a methodology based on small area estimation is that through auxiliary data we can improve direct estimates of a parameter of interest in small areas. Standard methods of variance component estimation used in the FH model for small areas produce a negative or zero estimate for these variances, with severe implications. In such a context, we proposed alternative approaches to those available in the literature, based on the EM algorithm, for estimating the variance of the random effects in the FH model, when estimating small area means. We showed through a simulation study that the EM algorithm is a good alternative to compute the ML estimate of the variance components, ensuring its strictly positive value. We compared the performance of our approaches with two recently proposed methods by means of statistical indicators. In general, the MLEM and RMLEM approaches performed well and similarly to the YML and YRML methods proposed by Yoshimori and Lahiri [42] (2014), but better than the LML and LRML methods proposed by Li and Lahiri [23] (2010). The proposed approaches have the advantage of working directly with the likelihood function without having to adjust it. A shortcoming of the LML and LRML methods in comparison to the approaches proposed here is that they can yield a negative value for the MSPE. Also, although the results of the MSE in the estimation of the variance component are similar under the YML and MLEM methods, note that the estimation with the EM algorithm is slightly more accurate in terms of SEs than the estimation with the YML and YRML methods, such as occurs when comparing the YRML and RMLEM methods. In an application from the real-world, we confirmed that small area estimation through the FH model helped to improve the direct estimates of the average monthly per capita food expenditure for Mexican rural households in 2000 according to three auxiliary variables. A possible future study can be conducted to compare the YML and YRML methods to their analogous based on the EM algorithm.

ACKNOWLEDGMENTS

The authors thank the editors and reviewers for their constructive comments on an earlier version of this manuscript which resulted in this improved version. This research was supported partially by fellowship "Becas-Conicyt" (J.L. Ávila-Valdez) and grant "Fondecyt 1160868" (V. Leiva), both from the Chilean Council for Scientific and Technology Research.

REFERENCES

[1] Battese, G.E.; Harter, R.M. and Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data, *Journal of the American Statistical Association*, **83**, 28–36.

[2] [Casas-Cordero et al., 2016]casasep:16 Casas-Cordero, C.; Encina, J. and Lahiri, P. (2016). *Poverty mapping for the Chilean comunas*. In "Analysis of Poverty Data by Small Area Estimation" (M. Pratesi, Ed.), volume 20, pages 379–404, Wiley, Chichester, UK.

[3] Chen, S. and Lahiri, P. (2008). On mean squared prediction error estimation in small area estimation problems, *Communications in Statistics: Theory and Methods*, **37**, 1792–1798.

[4] CIESIN (2005). *Poverty mapping project: poverty and food security case studies*, Technical Report http://dx.doi.org/10.7927/H4FF3Q9B (accessed at 26-Nov-2015), Center for International Earth Science Information Network (CIESIN), Columbia University and Centro Internacional de Tropical Agriculture, New York, US.

[5] Coelho, P.S. and Casimiro, F. (2008). Post Enumeration Survey of the 2001 Portuguese population and housing censuses, *REVSTAT – Statistical Journal*, **6**, 231–252.

[6] Coelho, P.S. and Pereira, L.N. (2011). A spatial unit level model for small area estimation, *REVSTAT – Statistical Journal*, **9**, 155–180.

[7] Datta, G.S. (2009). *Model-based approach of small area estimation*. In "Handbook of Statistics. Sample Surveys: Inference and Analysis" (D. Pfeffermann and C.R. Rao, Eds.), volume 29B, pages 251–288, Elsevier, Oxford, UK.

[8] Datta, G.S. and Ghosh, M. (2012). Small area shrinkage estimation, *Statistical Science*, **27**, 95–114.

[9] Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, **10**, 613–627.

[10] Datta, G.S.; Rao, J.N.K. and Smith, D.D. (2005). On measuring the variability of small area estimators under a basic area level model, *Biometrika*, **92**, 183–196.

[11] Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*, Wiley, New Jersey, US.

[12] El-Leithy, H.A.; Wahed, Z.A.A. and Abdallah, M.S. (2016). On non-negative estimation of variance components in mixed linear models, *Journal of Advanced Research*, **7**, 59–68.

[13] Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association*, **74**, 269–277.

[14] Figueiredo, F. and Gomes, M.I. (2004). The total median in statistical quality control, *Applied Stochastic Model in Business and Industry*, **20**, 339–353.

[15] Fishman, G.S. (1973). *Concepts and Methods in Discrete Event Digital Simulation*, Wiley, New York, US.

[16] Gomes, M.I.; Henriques-Rodrigues, L. and Miranda, M. (2011). Reduced-bias location-invariant extreme value index estimation: a simulation study, *Communications in Statistics: Simulation and Computation*, **40**, 424–447.

[17] Gomes, M.I.; Henriques-Rodrigues, L. and Miranda, M. (2016). Mean-of-order-p location-invariant extreme value index estimation, *REVSTAT – Statistical Journal*, **14**, 273–296.

[18] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320–340.

[19] Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, California, US.

[20] Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation, *TEST*, **15**, 1–96.

[21] Laird, N.M. and Were, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, **38**, 963–974.

[22] Lehtonen, R. and Veijanen, A. (2009). *Design-based methods of estimation for domains and small areas*, In "Handbook of Statistics. Sample Surveys: Inference and Analysis" (D. Pefeffermann and C.R. Rao, Eds.), volume 29B, pages 219–249, Elsevier, Oxford, UK.

[23] Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems, *Journal of Multivariate Analysis*, **101**, 882–892.

[24] Lohr, S.L. (1999). *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove, CA, US.

[25] Marchetti, S.; Giusti, C.; Pratesi, M.; Salvati, N.; Giannotti, F.; Pedreschi, D.; Rinzivillo, R.; Pappalardo, L. and Gabrielli, L. (2015). Small area model-based estimators using big data sources, *Journal of Official Statistics*, **31**, 263–281.

[26] McLachlan, G.J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*, Wiley, New Jersey, US.

[27] Mert, Y. (2015). Generalized least squares and weighted least squares estimation methods for distributional parameters, *REVSTAT – Statistical Journal*, **13**, 263–282.

[28] Molina, I.; Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach, *The Annals of Applied Statistics*, **8**, 852–885.

[29] Pereira, L.N. and Coelho, P.S. (2012). Small area estimation using a spatio-temporal linear mixed model, *REVSTAT – Statistical Journal*, **10**, 285–308.

[30] Pfeffermann, D. (2013). New important developments in small area estimation, *Statistical Science*, **28**, 40–68.

[31] Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas, *Journal of the American Statistical Association*, **102**, 1427–1439.

[32] Pinheiro, J.C. and Bates, D.M. (2004). *Mixed-Effects Models in S and S-Plus*, Springer, New York, US.

[33] Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association*, **85**, 163–171.

[34] Rao, J.N.K. (2003). *Small Area Estimation*, Wiley, New Yersey, US.

[35] Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*, Wiley, New Yersey, US.

[36] Rueda, M.M.; Arcos, A.; Molina, D. and Trujillo, M. (2018). Model-assisted and model-calibrated estimation for class frequencies with ordinal outcomes, *REVSTAT – Statistical Journal*, **16**, 323–348.

[37] Searle, S.R.; Casella, G. and McCulloch, C.E. (2006). *Variance Components*, Wiley, New York, US.

[38] Thompson, R. and Meyer, K. (1986). Estimation of variance components: what is missing in the EM algorithm, *Journal of Statistical Computation and Simulation*, **24**, 215–230.

[39] van Dyk, D.A. (2000). Fitting mixed-effects models using efficient EM-type algorithms, *Journal of Computational and Graphical Statistics*, **9**, 78–98.

[40] Verbyla, A.P. (1990). A conditional derivation of residual maximum likelihood, *Australian Journal of Statistics*, **32**, 227–230.

[41] Wolter, K.M. (2007). *Introduction to Variance Estimation*, Springer-Verlag, New York, US.

[42] Yoshimori, M. and Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model, *Journal of Multivariate Analysis*, **124**, 281–294.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 637–660

A UNIFICATION OF FAMILIES OF BIRNBAUM–SAUNDERS DISTRIBUTIONS WITH APPLICATIONS

Authors: Guillermo Martínez-Flórez
– Departamento de Matemáticas y Estadística, Facultad de Ciencias Básicas, Universidad de Córdoba, Córdoba, Colombia
guillermomartinez@correo.unicordoba.edu.co

Heleno Bolfarine
– Departamento de Estatítica, IME, Universidade de São Paulo, São Paulo, Brazil
hbolfar@ime.usp.br

Yolanda M. Gómez
– Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile
yolanda.gomez@uda.cl

Héctor W. Gómez
– Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile
hector.gomez@uantof.cl

Received: March 2019 Accepted: April 2019

Abstract:

- This paper considers an extension for the skew-elliptical Birnbaum–Saunders model by considering the power-normal model. Some properties of this family are studied and it is shown, in particular, that the range of asymmetry and kurtosis surpasses that of the ordinary skew-normal and power-normal models. Estimation is dealt with by using the maximum likelihood approach. Observed and expected information matrices are derived and it is shown to be nonsingular at the vicinity of symmetry. The applications illustrate the better performance of the new distribution when compared with other recently proposed alternative models.

Key-Words:

- *elliptical Birnbaum–Saunders distribution; maximum likelihood; power-normal distribution.*

1. INTRODUCTION

Vilca-Labra and Leiva-Sánchez ([30]) extended the ordinary Birnbaum–Saunders (BS) distribution by considering the generalized Birnbaum–Saunders skew-elliptical distribution which is based on replacing the normal distribution by the elliptical family of distributions of which the normal distribution is a special case. Such general family of distributions is very successful in dealing with data sets with high degrees of asymmetry and kurtosis.

In this paper, we consider an extension of the generalized BS (GBS) model proposed in Díaz-García and Leiva-Sánchez ([9]) to the case of elliptical distributions. A comprehensive review of the GBS model can be found in Sanhueza *et al.* ([29]). Another important feature of this distribution is related to robustness with respect to parameter estimation which was studied in Barros *et al.* ([4]). The generalized Birnbaum–Saunders skew-elliptical distribution represents an important extension of the ordinary BS distribution to the case of symmetrical and asymmetrical distributions, which can be appropriate for applications in life data and material fatigue data.

The family of elliptical distributions has proved to be an important alternative to the normal distribution. The distributions in this family are symmetric and include distributions with greater and smaller kurtosis than the normal distribution. The normal distribution is an important member of the family. The elliptical family of distributions has been studied by many authors including Fang and Zang ([12]), Fang *et al.* ([11]), Gupta and Varga ([13]), Arellano-Valle and Bolfarine ([2]), among others.

A random variable X is distributed according to the elliptical distribution with location parameter ξ and scale parameter η if its pdf can be written as

$$f(x) = \frac{c}{\eta} g\left(\left(\frac{x-\xi}{\eta}\right)^2\right), \tag{1.1}$$

for some nonnegative function $g(u)$, $u > 0$, such that $\int_0^\infty u^{-\frac{1}{2}} g(u)du = 1/c$, where c is a normalizing constant. The function $g(\cdot)$ is known as the density generator function. If X is elliptically distributed with location-scale parameters ξ and η and generator function g, denoted $X \sim EC(\xi, \eta; g)$. If $\xi = 0$ and $\eta = 1$, then X has spherical distribution, denoted as $X \sim EC(0, 1; g)$. Properties of this family can be studied in Kelker ([15]), Cambanis *et al.* ([5]), Fang *et al.* ([11]), Arellano-Valle and Bolfarine ([2]) and Gupta and Varga ([13]) among others. Particular cases of the $X \sim EC(0, 1; g)$ distribution are the Pearson type VII distribution, the type Kotz distribution, the Student-t (t_ν) distribution, the Cauchy distribution and the normal distribution, among others.

Díaz-García and Leiva-Sánchez ([9]) present the GBS distribution, by assuming that

$$Z = \frac{1}{\gamma}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right) \sim EC(0, 1; g).$$

where $\gamma > 0$ is the shape parameter and $\beta > 0$ is the scale parameter and the distribution median. Then, from

$$T = \frac{\beta}{4}\left[\gamma Z + \sqrt{\gamma^2 Z^2 + 4}\right]^2,$$

the GBS distribution follows, which we denote by $T \sim GBS(\gamma, \beta; g)$. The pdf for the random variable T is given by

$$f_{GBS}(t) = cg\left(\frac{1}{\gamma^2}\left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right)\frac{t^{-3/2}(t+\beta)}{2\gamma\beta^{1/2}}, \qquad t > 0, \tag{1.2}$$

where c is a normalizing constant and g is the generator function. Moreover, letting

$$a_t(\gamma, \beta) = a_t = \frac{1}{\gamma}\left(\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right), \tag{1.3}$$

it follows that

$$A_t(\gamma, \beta) = \frac{d}{dt}a_t(\gamma, \beta) = \frac{t^{-3/2}(t+\beta)}{2\gamma\beta^{1/2}},$$

so that (1.2) can be written as

$$f_{GBS}(t) = f(a_t(\gamma, \beta))A_t(\gamma, \beta),$$

where f is given in (1.1).

An extension of the elliptical model to the asymmetric case was given in Vilca-Labra and Leiva-Sánchez ([30]), where it is defined the standard elliptical asymmetric or skew-elliptical (SE) model as

$$f_Y(y; \lambda) = 2f(y)F(\lambda y); \quad y, \lambda \in \mathbb{R},$$

where f is given in (1.1), F is its respective cumulative distribution function (cdf) and λ is an asymmetry parameter. We use the notation $Y \sim SE(0, 1; g, \lambda)$. The cumulative distribution function for this model is given by

$$F_Y(y) = 2\int_{-\infty}^{y} f(t)F(\lambda t)dt. \tag{1.4}$$

A particular case of model (1.4) is the skew-normal (SN) distribution (see Azzalini, ([3])) with $f = \phi$ and $F = \Phi$ with pdf and cdf given, respectively, by

$$\phi_{SN}(y) = 2\phi(y)\Phi(\lambda y), \quad y \in \mathbb{R}, \tag{1.5}$$

$$\Phi_{SN}(y) = \Phi(y) - 2T(y; \lambda), \quad y \in \mathbb{R},$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of $N(0, 1)$ (the standard normal distribution), respectively and $T(\cdot; \cdot)$ is Owen's ([25]) function.

Extensions of the BS model to elliptical distributions were studied in Vilca-Labra and Leiva-Sánchez ([30]), namely, skew-elliptical Birnbaum–Saunders (SEBS) distribution. Model construction is based on the condition that

$$Z = \frac{1}{\gamma}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right) \sim SE(0, 1; g, \lambda).$$

We use the notation $SEBS(\gamma, \beta; g, \lambda)$. The case of model SEBS based on SN distribution, we denote $SNBS(\gamma, \beta, \lambda)$. Additional references on the BS distribution can be found in the recent book by Leiva ([18]).

An alternative asymmetric distribution is studied in Durrans ([10]), by introducing the fractional order statistical model, with pdf given by

$$\varphi_H(z;\alpha) = \alpha h(z)\{H(z)\}^{\alpha-1}, \qquad z \in \mathbb{R}, \tag{1.6}$$

where H is an absolutely continuous cumulative distribution function with pdf h and $\alpha > 0$ is a parameter that controls the distributional shape. The case $H = \Phi$ is called the power-normal (PN) distribution, with pdf given by

$$\varphi_\Phi(z;\alpha) = \alpha\phi(z)\{\Phi(z)\}^{\alpha-1}, \qquad z \in \mathbb{R},$$

denoted $Z \sim PN(\alpha)$. This model is an alternative to adjust data with asymmetry and kurtosis above (or below) the expected for the normal distribution.

In this paper we extend the SEBS model considered in Vilca-Labra and Leiva-Sánchez ([30]), using the fractionary order statistical model of Durrans ([10]). This generalization leads to a more flexible model in what concerns asymmetry and kurtosis, that the SEBS model, given that those models are special cases (hence also the ordinary BS model). It than can used for fitting fatigue data as well as life data.

The paper is organized as follows. Section 2 is devoted to study extensions of the GBS elliptical model by using the fractionary order statistical model in Durrans ([10]). Some properties of this family are studied and it is shown, in particular, that the range of asymmetry and kurtosis surpasses that of the ordinary skew-normal and power-normal models. Maximum likelihood estimation for the model proposed is implemented in Section 3. Observed and expected information matrices are derived and it is shown to be nonsingular at the vicinity of symmetry. Results of three real data application is presented in Section 4. The main conclusion is that the model proposed offers a viable alternative to others considered in the literature.

2. POWER SKEW-ELLIPTICAL BIRNBAUM–SAUNDERS DISTRIBUTIONS

We start by extending the model (1.6) assuming that the pdf h it is as follows

$$h(y;\lambda) = 2f(y)F(\lambda y); \quad y, \lambda \in \mathbb{R}, \tag{2.1}$$

where f is given in (1.1), F is its respective cumulative distribution function and λ is an asymmetry parameter. We call it the power skew-elliptical(PSE) model with pdf given by

$$\varphi_{PSE}(z;\lambda,\alpha) = \alpha h(z;\lambda)\{H(z;\lambda)\}^{\alpha-1}, \quad z \in \mathbb{R}. \tag{2.2}$$

We use the notation $Z \sim PSE(0,1;g,\lambda,\alpha)$.

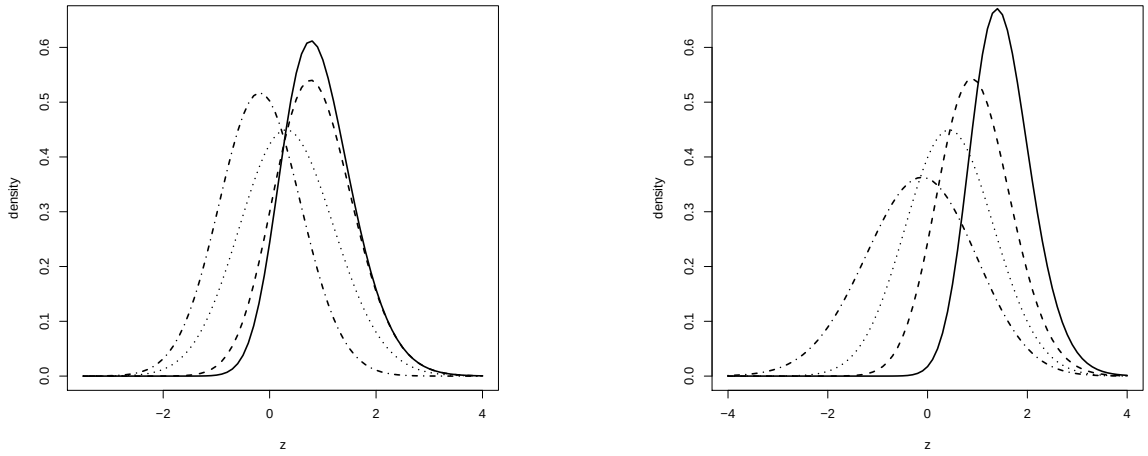
Moments of the random variable Z have no closed form, but under a variable change the r-th moment of the random variable Z can be written as

$$E(Z^r) = \alpha \int_0^1 [H^{-1}(z;\lambda)]^r z^{\alpha-1} dx,$$

where H^{-1} is the inverse of the function H.

If the pdf h follows model (1.5), then, we have the power skew-normal (PSN) model with parameters λ and α introduced in Martínez-Flórez *et al.* ([23]). This model we denote by $PSN(\lambda, \alpha)$.

Special cases of model PSN occur with $\alpha = 1$, so that the skew-normal model $\phi_{SN}(x)$, follows. On the other hand, with $\lambda = 0$ the model with pdf $\varphi_{\Phi}(x)$, that is, Durrans generalized normal model follows. The ordinary standard normal model is also a special case which follows by taking $\alpha = 1$ and $\lambda = 0$, that is, $\varphi_{PSN}(x; 0, 1) = \phi(x)$. Notice from Figure 1 (a) and (b) below that α and λ affect both, distribution asymmetry and kurtosis and hence the model proposed seems more flexible than the models by Azzalini ([3]) and Durrans ([10]).



(a) (b)

Figure 1: PSN model. (a) $\alpha = 1.5$ and λ =-0.75 (dotted dashed line), 0 (dotted line), 1 (dashed line) and 1.75 (solid line), (b) $\lambda =$ 0.70 and $\alpha = 0.50$ (dotted-dashed line), 1.0 (dotted line), 2.0 (dashed line) and 5.0 (solid line).

For some values of λ and $\alpha \in [0.1, 100]$, asymmetry and kurtosis coefficients namely $\sqrt{\beta_1}$ and β_2, for $Z \sim PSN(\lambda, \alpha)$, are in the intervals [-1.4676,0.9953) and [1.4672,5.4386] respectively, see Martínez-Flórez *et al.* ([23]). Such intervals contain the corresponding intervals for the skew-normal distribution, given by (-0.9953,0.9953) and [3,3.8692) respectively, and for the PN model, given by [-0.6115,0.9007] and [1.7170,4.3556], respectively, see Pewsey *et al.* ([26]). This illustrates the fact that the exponentiated skew-normal family contains models with greater (and smaller) asymmetry than both skew-normal (Azzalini, ([3])) and the power-normal (generalized normal) model (Durrans, ([10])). It then encompasses a family of distributions with more of both, platykurtic and leptokurtic, distributions. This illustrates the fact that the PSE model can be more flexible, respective to asymmetry and kurtosis, than the models characterized by density functions f_Y and φ_H.

We consider now an extension of the BS model to the case of exponentiated skew

elliptical distributions. Assuming that

$$Z = \frac{1}{\gamma}\left(\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}}\right) \sim PSE(0, 1; g, \lambda, \alpha),$$

it follow that Z is distributed according to model (2.2). Therefore, trough a simple variable change, it can be shown that the random variable

$$T = \frac{\beta}{4}\left[\gamma Z + \sqrt{\gamma^2 Z^2 + 4}\right]^2, \tag{2.3}$$

is distributed according to the power skew-elliptical Birnbaum–Saunders (PSEBS) distribution, denoted by $T \sim PSEBS(\gamma, \beta; g, \lambda, \alpha)$.

The pdf for random variable (2.3) is given by

$$\varphi_{PSEBS}(t; \gamma, \beta, \lambda, \alpha) = \alpha h(a_t(\gamma, \beta); \lambda)\{H(a_t(\gamma, \beta); \lambda)\}^{\alpha-1} A_t(\gamma, \beta), \quad t \in \mathbb{R}^+. \tag{2.4}$$

This model provides then a generalization for the model introduced by Díaz-García and Leiva-Sánchez ([9]) and Vilca-Labra and Leiva-Sánchez ([30]). Notice that for $\alpha = 1$, the SEBS model (Vilca-Labra and Leiva-Sánchez ([30])) is obtained and for $\lambda = 0$ and $\alpha = 1$ we obtain the GBS model (Díaz-García and Leiva-Sánchez ([9])). The case $\lambda = 0$ constitutes an extension for the BS model since it contains the ordinary BS model. This model has been studied in Martínez-Flórez *et al.* ([22]), supposing that $Z \sim PN(\alpha)$ and is called the power normal Birnbaum–Saunders (PNBS) model, denoted $PNBS(\gamma, \beta, \alpha)$ for the case of the normal distribution. Some properties and moments of the PSEBS distribution represented by the random variable T in (2.3) are presented next. Properties are similar to the ones derived for the SEBS distribution by Vilca-Labra and Leiva-Sánchez ([30]), for T with $Z \sim SE(0, 1; g, \lambda)$.

Theorem 2.1. *Let $T \sim PSEBS(\gamma, \beta; g, \lambda, \alpha)$. Then,*

1. $bT \sim PSEBS(\gamma, b\beta; g, \lambda, \alpha)$, $b > 0$ *and*

2. $T^{-1} \sim PSEBS(\gamma, \beta^{-1}; g, -\lambda, \alpha)$.

Proof: **1**. Let $T \sim PSEBS(\gamma, \beta; g, \lambda, \alpha)$ and $Y = bT$ for $b > 0$ so that $T = \frac{Y}{b}$, where the Jacobian is $J = \frac{1}{b}$. Moreover, since $a_t(\gamma, \beta) = a_{y/b}(\gamma, \beta) = a_y(\gamma, b\beta)$ and $|J|\frac{d}{dt}a_t(\gamma, \beta) = |J|\frac{d}{dt}a_{y/b}(\gamma, \beta) = \frac{d}{dy}a_y(\gamma, b\beta) = A_y(\gamma, b\beta)$, so that, from the above transformations we have

$$\begin{aligned} f_Y(y) &= \alpha h(a_{y/b}(\gamma, \beta); \lambda)\left\{H(a_{y/b}(\gamma, \beta); \lambda)\right\}^{\alpha-1} \frac{d}{dt}a_{y/b}(\gamma, \beta)|J| \\ &= \alpha h(a_y(\gamma, b\beta); \lambda)\left\{H(a_y(\gamma, b\beta); \lambda)\right\}^{\alpha-1} A_y(\gamma, b\beta), \end{aligned}$$

so that $Y = bT \sim PSEBS(\gamma, b\beta; g, \lambda, \alpha)$.

2. Let $T \sim PSEBS(\gamma, \beta; g, \lambda, \alpha)$ and $Y = T^{-1}$ then $T = Y^{-1}$ the jacobian of the transformation is $J = Y^{-2}$. Moreover, $a_t(\gamma, \beta) = a_{y^{-1}}(\gamma, \beta) = -a_y(\gamma, \beta^{-1})$ and $|J|\frac{d}{dt}a_t(\gamma, \beta) = |J|\frac{d}{dt}a_{y^{-1}}(\gamma, \beta) = \frac{d}{dy}a_y(\gamma, \beta^{-1}) = A_y(\gamma, \beta^{-1})$.

Then, $h(a_t(\gamma,\beta);\lambda) = h(a_{y^{-1}}(\gamma,\beta);\lambda) = h(a_y(\gamma,\beta);-\lambda)$ and

$$\begin{aligned} H(a_t(\gamma,\beta);\lambda) &= H(-a_y(\gamma,\beta^{-1});\lambda) \\ &= \int_{-\infty}^{-a_y(\gamma,\beta^{-1})} 2cg(x^2)F(\lambda x)dx \\ &= \int_0^y 2cg(a_x(\gamma,\beta^{-1})^2)F(-\lambda a_x(\gamma,\beta^{-1}))\frac{d}{dx}a_x(\gamma,\beta^{-1})dx \\ &= \int_{-\infty}^{a_y(\gamma,\beta^{-1})} h(x;-\lambda)dx \\ &= H(a_y(\gamma,\beta^{-1});-\lambda). \end{aligned}$$

Using the above transformations, we have that

$$\begin{aligned} f_Y(y) &= \alpha h(a_{y^{-1}}(\gamma,\beta);\lambda)\left\{H(a_{y^{-1}}(\gamma,\beta);\lambda)\right\}^{\alpha-1}\frac{d}{dt}a_{y^{-1}}(\gamma,\beta)|J| \\ &= \alpha h(a_y(\gamma,\beta^{-1});-\lambda)\left\{H(a_y(\gamma,\beta^{-1});-\lambda)\right\}^{\alpha-1}A_y(\gamma,\beta^{-1}) \end{aligned}$$

then we conclude that $Y = T^{-1} \sim PSEBS(\gamma,\beta^{-1};g,-\lambda,\alpha)$. □

Theorem 2.2. *Let $T \sim PSEBS(\gamma,\beta;g,\lambda,\alpha)$, $\mathbb{H}_T$ its cumulative distribution function and H the distribution function of $Z \sim PSE(0,1;g,\lambda,\alpha)$. Then,*

$$\mathbb{H}_T(t,\gamma,\beta;g,\lambda,\alpha) = \{H(a_t(\gamma,\beta);\lambda)\}^\alpha.$$

Proof: Let $a_x(\lambda,\beta)$, as defined above, so that

$$\begin{aligned} \mathbb{H}_T(t,\gamma,\beta;g\lambda,\alpha) &= \int_0^t \alpha h(a_x(\gamma,\beta);\lambda)\left\{H(a_x(\gamma,\beta);\lambda)\right\}^{\alpha-1}A_x(\gamma,\beta)dx \\ &= \int_0^t \alpha h(a_x(\gamma,\beta);\lambda)\left\{H(a_x(\gamma,\beta);\lambda)\right\}^{\alpha-1}\frac{d}{dx}a_x(\gamma,\beta)dx \\ &= \int_{-\infty}^{a_t(\gamma,\beta)} \alpha h(x;\lambda)\left\{H(x;\lambda)\right\}^{\alpha-1}dx \\ &= \mathbb{F}_Z(a_t(\gamma,\beta);\lambda,\alpha). \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathbb{F}_Z(a_t(\gamma,\beta);\lambda,\alpha) &= \int_{-\infty}^{a_t(\gamma,\beta)} \alpha h(x;\lambda)\left\{H(x;\lambda)\right\}^{\alpha-1}dx \\ &= \int_{-\infty}^{a_t(\gamma,\beta)} \frac{d}{dx}\left\{H(x;\lambda)\right\}^{\alpha}dx \\ &= \{H(a_t(\gamma,\beta);\lambda)\}^\alpha, \end{aligned}$$

concluding the proof. □

Theorem 2.3. *The p-th percentile of the $PSEBS(\gamma,\beta;g,\lambda,\alpha)$, $t_p = \mathbb{H}^{-1}(p,\gamma,\beta;g\lambda,\alpha)$, is given by:*

$$t_p = \beta\left[\frac{\lambda}{2}z_p + \sqrt{\left(\frac{\lambda}{2}z_p\right)^2+1}\right]^2,$$

where z_p is the p-th percentile of the distribution of $PSE(0,1;g,\lambda,\alpha)$, given by $z_p = H^{-1}(p^{1/\alpha};\lambda)$.

Proof: For $p \in (0,1)$ as in Theorem 2.2, it follows that $p = \{H(a_t(\gamma,\beta);\lambda)\}^\alpha$ so that $a_T(\gamma,\beta) = Z_p = H^{-1}(p^{1/\alpha};\lambda) \sim PSE(0,1:g,\lambda,\alpha)$ where H^{-1} is the inverse of H. Therefore, result follows from (2.3). □

Theorem 2.4. *The survivor function, cumulative risk function, risk and inverted risk functions for model PSEBS are given, respectively, by:*

$$S(t) = 1 - \{H(a_t(\gamma,\beta);\lambda)\}^\alpha, \quad M(t) = -\log[S(t)],$$

$$r(t) = \alpha r_{SEBS}(t)\frac{\{H(a_t(\gamma,\beta);\lambda)\}^{\alpha-1} - \{H(a_t(\gamma,\beta);\lambda)\}^\alpha}{1 - \{H(a_t(\gamma,\beta);\lambda)\}^\alpha} \text{ and } R(t) = \alpha R_{SEBS}(t),$$

where $r_{SEBS}(t)$ and $R_{SEBS}(t)$ denote the risk and inverted risk for the skew-elliptical BS model.

Proof: Result follows directly from the definitions of survival function risk and inverse risk using the result in Theorem 2.2. □

From Theorem 2.4 we can conclude that the inverse risk rate is proportional to the risk rate for the SEBS distribution. Hence, the intervals where $R(t)$ is decreasing or increasing, are the same as the intervals where $R_{SEBS}(t)$ is decreasing or increasing.

The following two Theorem discuss the existence and the r-th moment of a random variable $T \sim PSEBS(\gamma,\beta;g,\lambda,\alpha)$.

Theorem 2.5. *Let $T \sim PSEBS(\gamma,\beta;g,\lambda,\alpha)$ and $Z \sim PSE(0,1;g,\lambda,\alpha)$. Hence, $E(T^r)$ exists if and only if,*

$$\mathbb{E}\left[\left(\frac{\gamma Z}{2}\right)^{k+l}\left(\left(\frac{\gamma Z}{2}\right)+1\right)^{\frac{k-l}{2}}\right] \tag{2.5}$$

exists $k = 1,2,...,r$ with $l = 0,1,...,k$.

Proof: Taking $Z \sim PSE(0,1;g,\lambda,\alpha)$ it follows that

$$\begin{aligned}\mathbb{E}\left\{\left[\frac{T}{\beta}\right]^n\right\} &= \mathbb{E}\left\{\left[\frac{\gamma}{2}Z + \sqrt{\left(\frac{\gamma}{2}Z\right)^2+1}\right]^2\right\}^n \\ &= \mathbb{E}\left\{\left[1 + \left\{\frac{\gamma^2}{2}Z^2 + \gamma Z\sqrt{\left(\frac{\gamma}{2}Z\right)^2+1}\right\}\right]^n\right\}.\end{aligned}$$

Therefore, using the binomial expansion, we have

$$\mathbb{E}\left\{\left[\frac{T}{\beta}\right]^n\right\} = \sum_{k=0}^{n}\binom{n}{k}\mathbb{E}\left\{\left[\frac{\gamma^2}{2}Z^2 + \gamma Z\sqrt{\left(\frac{\gamma}{2}Z\right)^2+1}\right]^k\right\}$$

and doing another binomial expansion, we obtain

$$\mathbb{E}\left\{\left[\frac{T}{\beta}\right]^n\right\} = \sum_{k=0}^{n}\binom{n}{k}\sum_{l=0}^{k}\binom{k}{l}2^k\mathbb{E}\left\{\left[\left(\frac{\gamma}{2}Z\right)^{k+l}\left[\left(\frac{\gamma}{2}Z\right)^2+1\right]^{\frac{k-l}{2}}\right]\right\},$$

so that $\mathbb{E}\left\{\left[\frac{T}{\beta}\right]^n\right\}$ exists if, and only if, $\mathbb{E}\left\{\left[\left(\frac{\gamma}{2}Z\right)^{k+l}\left[\left(\frac{\gamma}{2}Z\right)^2+1\right]^{\frac{k-l}{2}}\right]\right\}$ exists, for $k = 0, 1, ..., n$ and $l = 0, 1, ..., k$. □

Theorem 2.6. *Let* $T \sim PSEBS(\gamma, \beta; g, \lambda, \alpha)$ *and* $Z \sim PSE(0, 1; g, \lambda, \alpha)$. *If* $\mathbb{E}[Z^r]$ *exists for* $r = 1, 2, ...,$ *then*

$$\begin{aligned}\mu_r = \mathbb{E}(T^r) = \beta^r \sum_{[0\le k\le r/2]}\binom{r}{2k}\left(\frac{1}{2}\right)^{2k}\sum_{j=0}^{2k}\binom{2k}{j}\mathbb{E}[(\gamma Z)^{4k-j}(\gamma^2Z^2+4)^{j/2}]\\ + \beta^r \sum_{[0\le k<r/2]}\binom{r}{2k+1}\left(\frac{1}{2}\right)^{2k+1}\sum_{j=0}^{2k+1}\binom{2k+1}{j}\mathbb{E}[(\gamma Z)^{4k+2-j}(\gamma^2Z^2+4)^{j/2}]\end{aligned}$$

where $[\cdot]$ *corresponds to the sum of the integer part of the subscripts.*

Corollary 2.1. *For* $r = 1, 2$ *we have that*

$$\mathbb{E}(T) = \frac{\beta}{2}[2+\gamma^2\nu_2+\gamma\kappa_1] \quad \text{and} \quad \mathbb{E}(T^2) = \frac{\beta^2}{2}[2+4\gamma^2\nu_2+\gamma^4\nu_4+2\gamma\kappa_1+\gamma^3\kappa_3],$$

where $\nu_k = \mathbb{E}[Z^k]$ *and* $\kappa_k = \mathbb{E}\left[Z^k\left(\gamma^2Z^2+4\right)^{1/2}\right]$. *Then, the variance is given by*

$$Var(T) = \mathbb{E}(T^2) - \mathbb{E}^2(T) = \frac{\gamma^2\beta^2}{4}[4\nu_2 - \kappa_1^2 + 2\gamma\kappa_3 - 2\gamma\nu_2\kappa_1 - \gamma^2\nu_2^2 + 2\gamma^2\nu_4].$$

The central moments, $\acute{\mu}_r = \mathbb{E}(T - \mathbb{E}(T))^r$, for $r = 2, 3, 4$ can be obtained using $\acute{\mu}_2 = \mu_2 - \mu_1^2$, $\acute{\mu}_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3$ and $\acute{\mu}_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$. Hence, variation coefficient, asymmetry and kurtosis can be obtained by using:

$$CV = \frac{\sqrt{\sigma_T^2}}{\mu_1}, \qquad \sqrt{\beta_1} = \frac{\acute{\mu}_3}{[\acute{\mu}_2]^{3/2}} \qquad \text{and} \qquad \beta_2 = \frac{\acute{\mu}_4}{[\acute{\mu}_2]^2}.$$

2.1. Power skew-normal Birnbaum–Saunders distribution

The power skew-normal Birnbaum–Saunders distribution is obtained by taking $H = \Phi_{SN}$ (and $h = \phi_{SN}$) in (2.4) and is denoted by PSNBS. It follows then that the density function is given by

$$\varphi_{PSNBS}(t; \gamma, \beta; \phi, \lambda, \alpha) = \alpha\phi_{SN}(a_t(\gamma, \beta))\{\Phi_{SN}(a_t(\gamma, \beta))\}^{\alpha-1}A_t(\gamma, \beta),$$

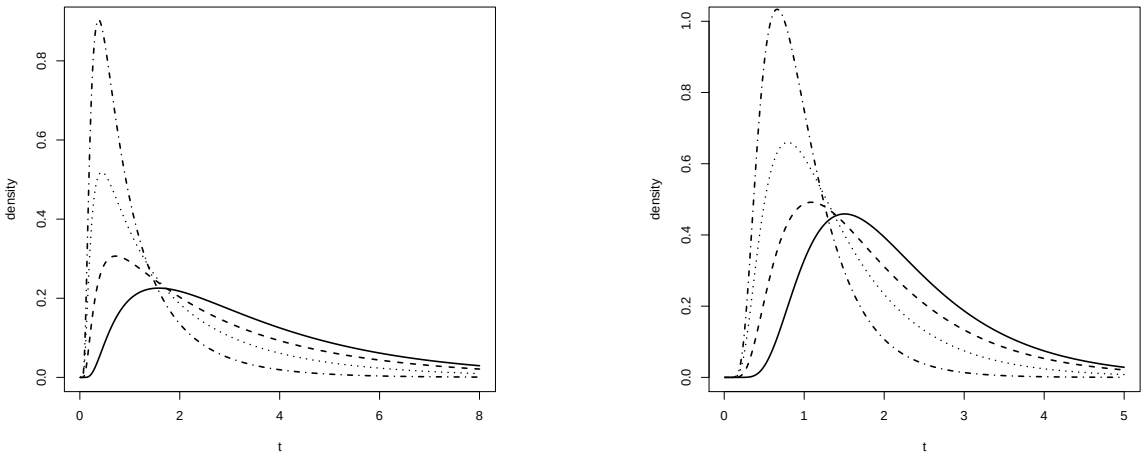
with a_t given in (1.3). Notice that the ordinary BS is a special case which follows by taking $F = \Phi$, $\lambda = 0$ and $\alpha = 1$. If $\alpha = 1$, the asymmetric BS model studied in Vilca-Labra and

Leiva-Sánchez ([30]) is derived and for $\lambda = 0$, we obtain the power-normal BS model studied in Martínez-Flórez *et al.* ([22]). Moreover, some properties of the BS distribution holds for the PSNBS distribution.

The cumulative distribution function for this model is given by

$$\mathbb{H}_{PSNBS}(t, \gamma, \beta; \lambda, \alpha) = \{\Phi(a_t(\gamma, \beta)) - 2T(a_t(\gamma, \beta); \lambda)\}^{\alpha}, \quad t > 0,$$

Figures 2 and 3 depicts the behavior of the PSNBS distribution for those values of α and λ.



(a) (b)

Figure 2: Plots for density function $\varphi_T(t; \gamma, \beta, \lambda, \alpha)$. (a) $(\gamma, \beta, \lambda, \alpha) =$ (0.75,1,-1,1.75) (dashed and dotted lines), $(\gamma, \beta, \lambda, \alpha) =$ (0.75,1,-0.25,1.75) (dotted line), $(\gamma, \beta, \lambda, \alpha) =$ (0.75,1,0.25,1.75) (dashed line) and $(\gamma, \beta, \lambda, \alpha) =$ (0.75,1,1,1.75) (solid line). (b)$(\gamma, \beta, \lambda, \alpha) =$ (1.25,1,-1,1.75) (dashed and dotted lines), $(\gamma, \beta, \lambda, \alpha) =$ (1.25,1,-0.25,1.75) (dotted line), $(\gamma, \beta, \lambda, \alpha) =$ (1.25,1,0.25,1.75) (dashed line) and $(\gamma, \beta, \lambda, \alpha) =$ (1.25,1,1,1.75) (solid line).

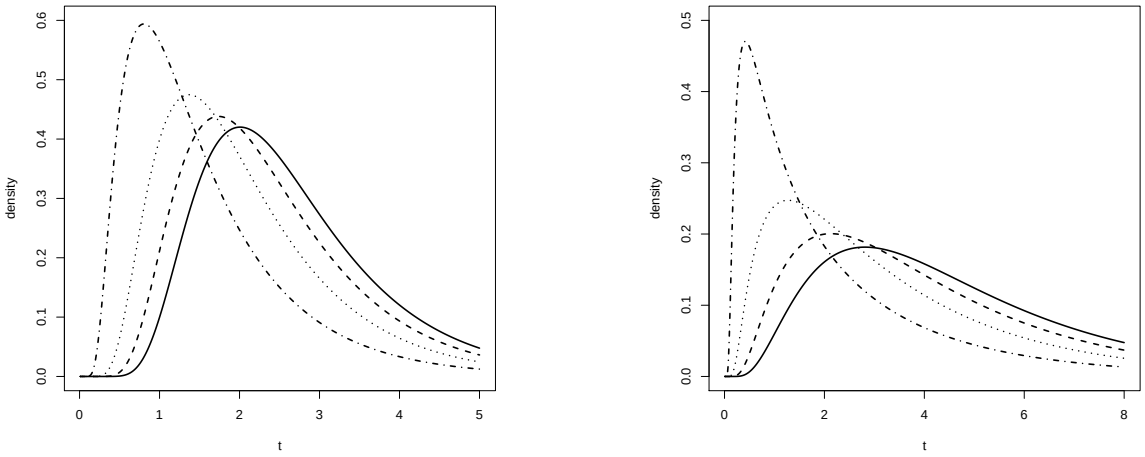
From Theorem 2.4, the survivor function, risk and inverted risk functions for model PSNBS are given, respectively, by

$$S(t) = 1 - \{\Phi_{SN}(a_t(\gamma, \beta))\}^{\alpha}, \quad M(t) = -\log[S(t)], \tag{2.6}$$

$$r(t) = \alpha r_{SNBS}(t) \frac{\{\Phi_{SN}(a_t(\gamma, \beta))\}^{\alpha-1} - \{\Phi_{SN}(a_t(\gamma, \beta))\}^{\alpha}}{1 - \{\Phi_{SN}(a_t(\gamma, \beta))\}^{\alpha}} \text{ and } R(t) = \alpha R_{SNBS}(t),$$

where $r_{SNBS}(t)$ and $R_{SNBS}(t)$ respectively denote the risk and inverted risk of the skew-normal Birnbaum–Saunders.

The following Theorem shows that for $t \to \infty$ the limit of the risk function of the PSNBS model coincides with the limit to infinity for the risk function of the SNBS model, result found by Leiva *et al.* ([20]).



a) b)

Figure 3: Plots for density function $\varphi_T(t;\gamma,\beta,\lambda,\alpha)$. a) $(\gamma,\beta,\lambda,\alpha)$ = (0.75,1,1,0.75) (dashed and dotted lines), $(\gamma,\beta,\lambda,\alpha)$ =(0.75,1,1,1.5) (dotted line), $(\gamma,\beta,\lambda,\alpha)$ =(0.75,1,1,2.25) (dashed line) and $(\gamma,\beta,\lambda,\alpha)$ =(0.75,1,1,3) (solid line). b)$(\gamma,\beta,\lambda,\alpha)$ = (1.25,1,1,0.75) (dashed and dotted lines), $(\gamma,\beta,\lambda,\alpha)$ = (1.25,1,1,1.5) (dotted line), $(\gamma,\beta,\lambda,\alpha)$ = (1.25,1,1,2.25) (dashed line) and $(\gamma,\beta,\lambda,\alpha)$ = (1.25,1,1,3) (solid line).

Theorem 2.7.

$$\lim_{t\to\infty} r(t) = (1+\lambda^2)(2\gamma^2\beta)^{-1}.$$

Proof: Rewritting the risk function in the form

$$r(t) = \alpha r_{SNBS}(t)\{\Phi_{SN}(a_t(\gamma,\beta))\}^{\alpha-1}\frac{1-\Phi_{SN}(a_t(\gamma,\beta))}{1-\{\Phi_{SN}(a_t(\gamma,\beta))\}^{\alpha}},$$

and using L'Hôpital rule, we obtain

$$\lim_{t\to\infty}\frac{1-\Phi_{SN}(a_t(\gamma,\beta))}{1-\{\Phi_{SN}(a_t(\gamma,\beta))\}^{\alpha}} = \lim_{t\to\infty}\frac{-\phi_{SN}(a_t(\gamma,\beta))A_t(\gamma,\beta)}{-\alpha\{\Phi_{SN}(a_t(\gamma,\beta))\}^{\alpha-1}\phi_{SN}(a_t(\gamma,\beta))A_t(\gamma,\beta)} = \frac{1}{\alpha},$$

where $A_t(\gamma,\beta) = \frac{d}{dt}a_t(\gamma,\beta)$.

Therefore,

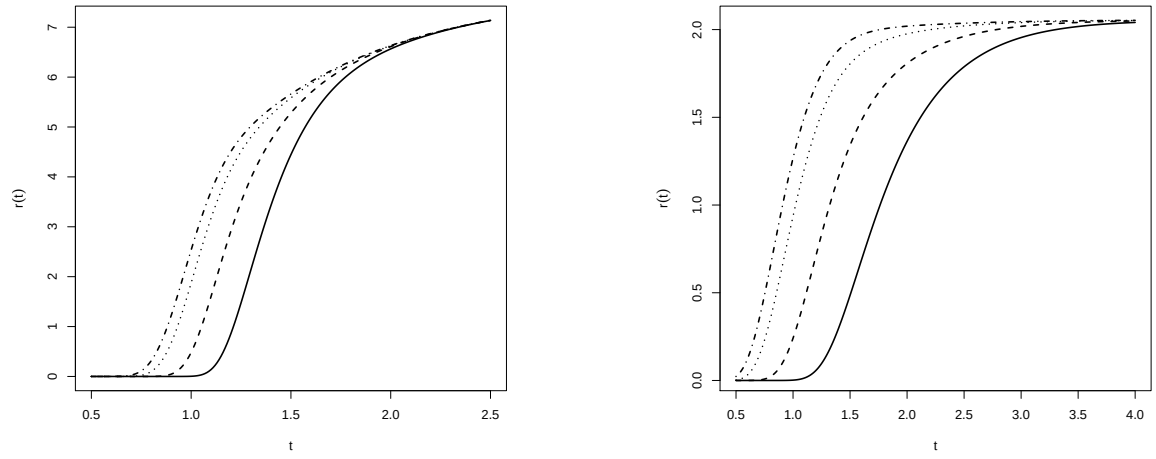
$$\lim_{t\to\infty} r(t) = \alpha\lim_{t\to\infty} r_{SNBS}(t)\frac{1}{\alpha} = \lim_{t\to\infty} r_{SNBS}(t) = (1+\lambda^2)(2\gamma^2\beta)^{-1}$$

where

$$\lim_{t\to\infty} r_{SNBS}(t) = (1+\lambda^2)(2\gamma^2\beta)^{-1},$$

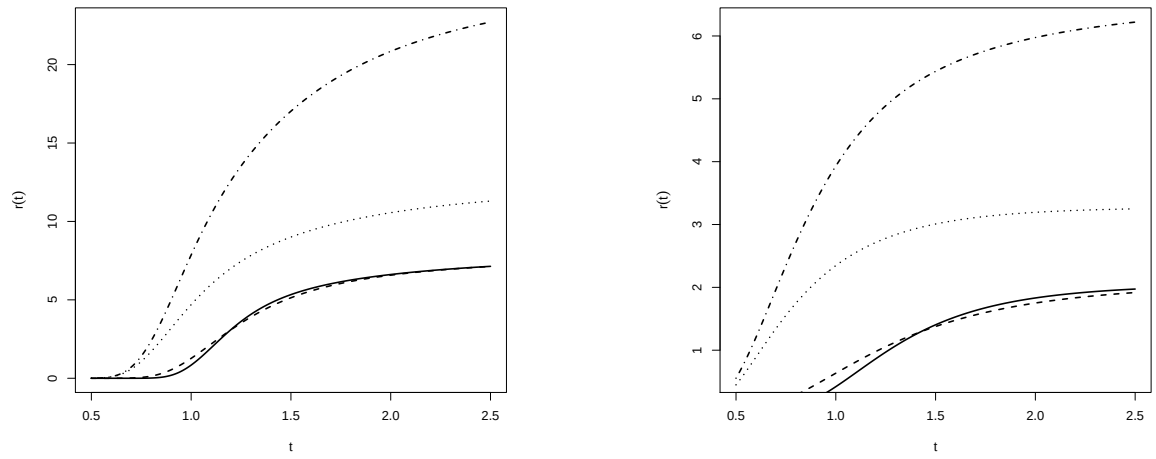
as shown in Leiva *et al.* ([20]). □

Figures 4 and 5 reveals the fact that the risk function is a non decreasing (and unimodal) function of t, but an increasing function of parameter α. Moreover, $r(t)$ is a non decreasing function for parameter γ.



(a) (b)

Figure 4: Function $r(t)$, for (a) $\gamma = 0.25$, $\beta = 1.0$, $\lambda = 2$ and $\alpha = 0.75$ (dashed and dotted line), α =1 (dotted line), α =2 (dashed line) and α =5 (solid line). (b)$\gamma = 0.5$, $\beta = 1.0$, $\lambda = 2$ and $\alpha = 0.75$ (dashed and dotted line), α =1 (dotted line), α =2 (dashed line) and α =5 (solid line).



(a) (b)

Figure 5: Function $r(t)$, for (a) $\gamma = 0.25$, $\beta = 1.0$, $\alpha = 1.75$ and λ =-1.5 (dashed and dotted lines), λ =-0.75 (dotted line), λ =0.75 (dashed line) and λ =1.5 (solid line). (b)$\gamma = 0.5$, $\beta = 1.0$, $\alpha = 1.75$ and λ = -1.5 (dashed and dotted lines), λ =-0.75 (dotted line), λ =0.75 (dashed line) and λ =1.5 (solid line).

2.2. Inference for the PSNBS model

We present in this section the score functions and the observed and expected information matrices for the parameter $\theta = (\gamma, \beta, \lambda, \alpha)$. Given a random sample of size n, $\mathbf{t} = (t_1, ..., t_n)'$, from the distribution $PSNBS(\gamma, \beta, \lambda, \alpha)$, the log-likelihood function for $\theta = (\gamma, \beta, \lambda, \alpha)'$ can be written as

$$\begin{aligned}\ell(\theta; \mathbf{t}) \;=\; & n\left[\log(\alpha) - \log(\gamma) - \frac{1}{2}\log(\beta)\right] + \sum_{i=1}^{n}\log(t_i+\beta) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) \\ & - \frac{1}{2\gamma^2}\sum_{i=1}^{n}\left[\frac{t_i}{\beta} + \frac{\beta}{t_i} - 2\right] + \sum_{i=1}^{n}\log(\Phi(\lambda a_{t_i})) + (\alpha-1)\sum_{i=1}^{n}\log(\Phi_{SN}(a_{t_i})).\end{aligned} \tag{2.7}$$

The maximum likelihood (ML) estimators are obtained by maximizing the log-likelihood function given in (2.7). The score function, defined as the derivative of the likelihood function with respect to model parameters is denoted by $U(\theta) = (U(\gamma), U(\beta), U(\lambda), U(\alpha))'$ so that the score equations follow by equating the scores to zero, leading to the following equations

$$U(\gamma) = -\frac{1}{\gamma}\sum_{i=1}^{n}\left[1 - a_{t_i}^2 + a_{t_i}\left[\lambda w_i + (\alpha-1)w_{1i}\right]\right] = 0,$$

$$U(\beta) = -\frac{n}{2\beta} + \sum_{i=1}^{n}\frac{1}{\beta+t_i} - \frac{1}{2\gamma^2}\sum_{i=1}^{n}\left[\frac{1}{t_i} - \frac{t_i}{\beta^2}\right] - \frac{1}{2\gamma\beta^{\frac{3}{2}}}\sum_{i=1}^{n}\frac{t_i+\beta}{t_i^{\frac{1}{2}}}\left[\lambda w_i + (\alpha-1)w_{1i}\right] = 0,$$

$$U(\lambda) = \sum_{i=1}^{n} a_{t_i}\frac{\phi(\lambda a_{t_i})}{\Phi(\lambda a_{t_i})} - \sqrt{\frac{2}{\pi}}\frac{(\alpha-1)}{1+\lambda^2}\sum_{i=1}^{n} w_{2i} = 0, \qquad U(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n} u_i = 0,$$

where $u_i = \log\{\Phi_{SN}(a_{t_i})\}$,

$$w_i = \frac{\phi(\lambda a_{t_i})}{\Phi(\lambda a_{t_i})}, \qquad w_{1i} = \frac{\phi_{SN}(a_{t_i})}{\Phi_{SN}(a_{t_i})},$$

and

$$w_{2i} = \frac{\phi\left(\sqrt{1+\lambda^2}a_{t_i}\right)}{\Phi_{SN}(a_{t_i})}, \qquad i = 1, ..., n.$$

Numerical approaches are required for solving the above system of equations.

The elements of the observed information matrix are the negative of the second partial derivatives of the likelihood function with respect to the model parameters evaluated at the ML estimators. We use the notation $j_{\gamma\gamma}$, $j_{\beta\gamma}$, $j_{\lambda\gamma}$, $j_{\alpha\gamma}$, ..., $j_{\alpha\lambda}$, $j_{\alpha\alpha}$ so that, after extensive algebraic manipulations,

$$\begin{aligned}j_{\gamma\gamma} = & -\frac{n}{\gamma^2} + \frac{3}{\gamma^2}\sum_{i=1}^{n} a_{t_i}^2 + \frac{\lambda}{\gamma^2}\sum_{i=1}^{n} a_{t_i}^2 w_i\left[\lambda^2 a_{t_i} + \lambda w_i - 2\right] - \sqrt{\frac{2}{\pi}}\frac{\lambda(\alpha-1)}{\gamma^2}\sum_{i=1}^{n} a_{t_i}^2 w_{2i} \\ & - \frac{(\alpha-1)}{\gamma^2}\sum_{i=1}^{n} a_{t_i} w_{1i}\left[2 + a_{t_i}^2 - a_{t_i} w_{1i}\right].\end{aligned}$$

$$j_{\beta\gamma} = \frac{1}{\gamma^3}\sum_{i=1}^{n}\left[\frac{t_i}{\beta^2} - \frac{1}{t_i}\right] - \frac{\lambda}{2\gamma^2\beta^{3/2}}\sum_{i=1}^{n}\frac{t_i+\beta}{\sqrt{t_i}}w_i\left[1-\lambda a_{t_i}(\lambda a_{t_i}+w_i)\right]$$
$$- \frac{\alpha-1}{2\gamma^2\beta^{3/2}}\sum_{i=1}^{n}\frac{t_i+\beta}{\sqrt{t_i}}\left[\sqrt{\frac{2}{\pi}}\lambda a_{t_i}w_{2i} + w_{1i}(1+a_{t_i}^2 - a_{t_i}w_{1i})\right],$$

$$j_{\lambda\gamma} = \frac{1}{\gamma}\sum_{i=1}^{n}a_{t_i}w_i\left[1-\lambda a_{t_i}w_i(\lambda a_{t_i}+w_i)\right] + \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{\gamma}\sum_{i=1}^{n}a_{t_i}w_{2i}\left[a_{t_i}+\frac{1}{1+\lambda^2}w_{1i}\right],$$

$$j_{\beta\beta} = -\frac{n}{2\beta^2} + \sum_{i=1}^{n}\frac{1}{(t_i+\beta)^2} + \frac{1}{\gamma^2\beta^3}\sum_{i=1}^{n}t_i - \frac{1}{2\gamma\beta^{5/2}}\sum_{i=1}^{n}\frac{3t_i+\beta}{\sqrt{t_i}}[\lambda w_i + (\alpha-1)w_{1i}]$$
$$+ \frac{1}{4\gamma^2\beta^3}\sum_{i=1}^{n}\frac{(t_i+\beta)^2}{t_i}\left[\lambda^2 w_i\left(\frac{\lambda(t_i-\beta)}{\gamma\beta^{1/2}t_i^{1/2}} + w_i\right)\right.$$
$$\left. + (\alpha-1)\left(\frac{t_i-\beta}{\gamma\beta^{1/2}t_i^{1/2}}w_{1i} + w_{1i}^2 - \sqrt{\frac{2}{\pi}}\lambda w_{2i}\right)\right],$$

$$j_{\lambda\beta} = \frac{1}{2\gamma\beta^{3/2}}\sum_{i=1}^{n}\frac{t_i+\beta}{\sqrt{t_i}}w_i\left[1-\lambda a_{t_i}w_i(\lambda a_{t_i}+w_i)\right]$$
$$+ \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{2\gamma\beta^{3/2}}\sum_{i=1}^{n}\frac{t_i+\beta}{\sqrt{t_i}}w_{2i}\left[a_{t_i}+\frac{1}{1+\lambda^2}w_{1i}\right],$$

$$j_{\alpha\gamma} = \frac{1}{\gamma}\sum_{i=1}^{n}a_{t_i}w_{1i}, \qquad j_{\alpha\beta} = \frac{1}{2\gamma\beta^{3/2}}\sum_{i=1}^{n}\frac{t_i+\beta}{\sqrt{t_i}}w_{1i},$$

$$j_{\lambda\lambda} = \sum_{i=1}^{n}a_{t_i}^2 w_i(\lambda a_{t_i}+w_i) - \sqrt{\frac{2}{\pi}}\frac{2\lambda(\alpha-1)}{(1+\lambda^2)^2}\sum_{i=1}^{n}w_{2i}$$
$$+ \sqrt{\frac{2}{\pi}}\frac{\alpha-1}{1+\lambda^2}\sum_{i=1}^{n}w_{2i}\left[-\lambda a_{t_i}^2 + \sqrt{\frac{2}{\pi}}\frac{1}{1+\lambda^2}w_{2i}\right],$$

$$j_{\alpha\lambda} = \sqrt{\frac{2}{\pi}}\frac{1}{1+\lambda^2}\sum_{i=1}^{n}w_{2i}, \qquad j_{\alpha\alpha} = \frac{n}{\alpha^2}.$$

The elements of the Fisher information matrix are n^{-1} times the expected values of the elements of the matrix of second derivatives of the log-likelihood function.

Considering now $\lambda = 0$ and $\alpha = 1$ and using the approximation in Cribari-Neto and Branco ([8]), we can write the expected Fisher information matrix as

$$I_F(\theta) = \begin{pmatrix} \frac{1}{\gamma^2} & 0 & 0 & \frac{1}{4\gamma}\frac{\pi^2}{\sqrt{8+\pi^2}} \\ 0 & \frac{\sqrt{2\pi}+\gamma p(\gamma)}{\sqrt{2\pi}\gamma^2\beta^2} & A_1(\gamma,\beta) & A_2(\gamma,\beta) \\ 0 & A_1(\gamma,\beta) & \frac{2}{\pi} & \sqrt{\frac{1}{2}} \\ \frac{1}{4\gamma}\frac{\pi^2}{\sqrt{8+\pi^2}} & A_2(\gamma,\beta) & \sqrt{\frac{1}{2}} & 1 \end{pmatrix},$$

where $p(\gamma) = \gamma\sqrt{\frac{2}{\pi}} - \frac{\pi \exp(\frac{2}{\gamma^2})}{2} erfc(\frac{2}{\gamma})$, with $erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-t^2)dt$ being the error function (see Prudnikov *et al.* ([27])), $A_1(\gamma, \beta) = \sqrt{\frac{2}{\pi}} \frac{1}{4\gamma^2\beta^2} \int_0^\infty \left(1 + \frac{\beta}{t}\right) \phi(a_t)dt$ and $A_2(\gamma, \beta) = \sqrt{\frac{2}{\pi}} \frac{1}{4\gamma^2\beta^2} \int_0^\infty \left(1 + \frac{\beta}{t}\right) \phi(2\sqrt{2}a_t/\pi)\Phi(-a_t)dt$.

The 2x2 superior submatrix of $I(\theta)$ is the Fisher information matrix for the ordinary BS distribution, as can be seen in Lemonte *et al.* ([21]). It can be verified that the columns (lines) of the matrix $I_F(\theta)$ are linearly independent and hence, it is invertible. Hence, for large samples, the MLE $\hat{\theta}$ of θ is asymptotically normal, that is,

$$\hat{\theta} \overset{A}{\to} N_4(\theta, I_F(\theta)^{-1}),$$

resulting that the asymptotic variance of the ML estimators $\hat{\theta}$ is the inverse of $I_F(\theta)$, which we denote by $\Sigma_{\hat{\theta}} = I_F(\theta)^{-1}$.

Approximation $N_4(\theta, \Sigma_{\hat{\theta}})$ can be used to construct confidence intervals for θ_r, which are given by $\hat{\theta}_r \mp z_{1-\rho/2}\sqrt{\hat{\sigma}(\hat{\theta}_r)}$, where $\hat{\sigma}(\cdot)$ corresponds to the r-th diagonal element of the matrix $\Sigma_{\hat{\theta}}$ and $z_{1-\rho/2}$ denotes $100(1-\rho/2)$-quantile of the standard normal distribution.
On the other hand, in presence of right-censoring we can adopt the following scheme. Assuming that for each individual the failure time is independent of the censoring time (say, Y_i and C_i for $i = 1, ..., n$ respectively). The observed times are given by $T_i = \min(Y_i, C_i)$ and the failure indicator is denoted as $\delta_i = I(Y_i \leq C_i)$. Given a sample of observed times and failure indicators $(t_1, \delta_1), (t_2, \delta_2), ..., (t_n, \delta_n)$ and under the additional assumption of non-informative censoring, i.e., the distribution of failure times (Y_i) don't provide information about the censoring times (C_i) and viceversa (see Lagakos ([16])), the log-likelihood function for θ is given

$$l(\theta; \mathbf{t}) = \sum_{i=1}^{n} \left[\delta_i \log \varphi_{PSNBS}(t_i; \gamma, \beta; \phi, \lambda, \alpha) + (1 - \delta_i) \log S((t_i; \gamma, \beta; \phi, \lambda, \alpha))\right]. \tag{2.8}$$

For $\delta_i = 1$, $i = 1, ..., n$, equation (2.8) is reduced to (2.7). Finally, inference based on (2.8) can be performed in a similar manner as was done in the uncensored case, as described above.

3. RELATIONSHIP AMONG DISTRIBUTIONS OF THE FAMILY PSEBS

The pdf for the PSEBS model with t_ν distribution (denoted PSTBS) is given by:

$$\varphi_{PSTBS}(t; \xi) = \frac{\alpha\Gamma(\frac{\nu+1}{2})}{(\nu\pi)^{1/2}\Gamma(\frac{\nu}{2})} \left[1 + \frac{a_t^2}{\nu}\right]^{-\frac{\nu+1}{2}} F_{st}(\lambda a_t)\{H_{st}(a_t; \lambda)\}^{\alpha-1} A_t(\gamma, \beta), \tag{3.1}$$

where $\xi = (\gamma, \beta; \lambda, \alpha, \nu)$ and ν representing degrees of freedom and F_{st} is the cdf of the t_ν distribution (see Johnson *et al.* ([14])) and H_{st} is the cdf of the skew-t$_\nu$ distribution. The power skew-Cauchy Birnbaum–Saunders (PSCBS) model follows from pdf (3.1) by taking $\nu = 1$. Note that in the particular case that $\lambda = 0$ and $\alpha = 1$, the PSTBS coincides with the Birnbaum–Saunders-t$_\nu$ (BST) distribution studied in Díaz-García and Leiva-Sánchez ([9]) and

for $\lambda = 0$ is obtained the Power Birnbaum–Saunders Student-t distribution studied in ([24]). Moreover, for $\alpha = 1$, we obtain the skew-t$_\nu$-Birnbaum–Saunders (STBS) model, studied in Vilca-Labra and Leiva-Sánchez ([30]). The relationships among some of those models are presented in Figure 6.

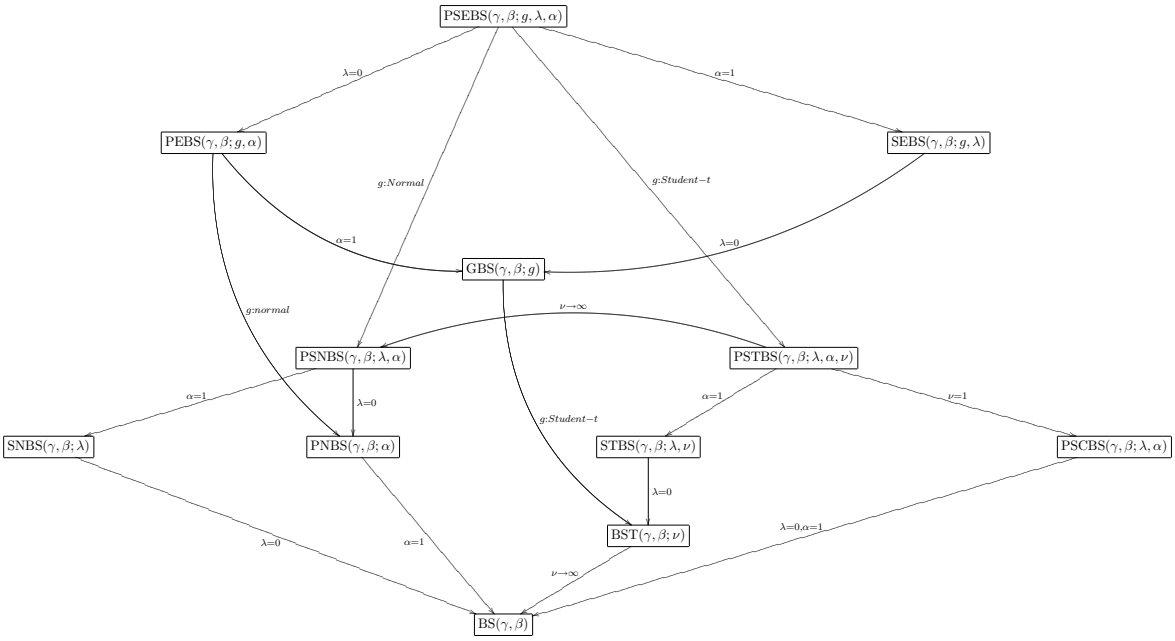


Figure 6: Relationship among distributions of the family PSEBS.

The density generator of the normal, Cauchy, t_ν, generalized t_ν, type I logistic, type II logistic and power exponential are, respectively, given by $g(u) = (2\pi)^{-1/2}\exp(-u/2)$, $g(u) = \{\pi(1+u)\}^{-1}$, $g(u) = \nu^{\nu/2}B(1/2,\nu/2)^{-1}(\nu+u)^{-(\nu+1)/2}$, where $\nu > 0$ and $B(\cdot,\cdot)$ is the beta function, $g(u) = s^{r/2}B(1/2,r/2)^{-1}(s+u)^{-(r+1)/2}$ $(s,r>0)$, $g(u) = c\exp(-u)(1+\exp(-u))^{-2}$, where $c \approx 1.484300029$ is the normalizing constant obtained from $\int_0^\infty u^{-1/2}g(u)du = 1$, $g(u) = \exp(-\sqrt{u})(1+\exp(-\sqrt{u}))^{-2}$ and $g(u) = c(k)\exp(-\frac{1}{2}u^{1/(1+k)})$, $-1 < k \leq 1$, where $c(k) = \Gamma(1+(k+1)/2)2^{1+(1+k)/2}$.

4. APPLICATIONS

In this section, it is shown that the model discussed in the previous sections can give good feedback to understand relations between variables in applied problems. The first application considers the remission times (in months) of the bladder cancer patients. The second application presented is based on certain features of the trees in a forestry area, and the last applications is a censured data.

4.1. Application I

We consider an uncensored data set corresponding to remission times (in months) of a random sample of 128 bladder cancer patients. These data were previously studied by Lee and Wang ([17]). Bladder cancer is a disease in which abnormal cells multiply without control in the bladder. The most common type of bladder cancer recapitulates the normal histology of the urothelium and is known as transitional cell carcinoma.

Descriptive statistics results are summarized in Table 1, where $\sqrt{b_1}$ and b_2 are sample asymmetry and sample kurtosis coefficients, respectively. There is indication of high kurtosis in this data set, which suggest that PSNBS model can be more appropriate than BS model. ML estimators were computed by maximizing log-likelihood using function "optim" in R Core Team ([28]). Table 2 shows the fitting of the BS, SNBS, PNBS and PSNBS models (standard error are in parenthesis). To compare the fitting of these models, we use Akaike ([1]) criterion, namely

$$AIC = -2\ell(\cdot; \mathbf{t}) + 2k,$$

we consider also the AICC (corrected Akaike information criterion), namely

$$AICC = AIC + \frac{2k(k+1)}{n-(k+1)},$$

where k is the number of parameters in the model. According to this criterion the model that best fits the data is the one with the lowest AIC or AICC. We also apply the formal goodness-of-fit tests in order to verify which distribution fits better to these data. We consider the Cramér-von Mises (W^*), Anderson-Darling (A^*) statistics, Kolmolgorov- Smirnov(K-S) test statistics and p-value. The statistics W^* and A^* are described in detail in Chen and Balakrishnan ([6]). In general, the smaller the values of the statistics W^* and A^*, the better the fit to the data.

Table 1: Descriptive statistics for the data set.

n	$\bar{t}$	s^2	$\sqrt{b_1}$	b_2
128	4.1293	9.3660	3.2480	15.1950

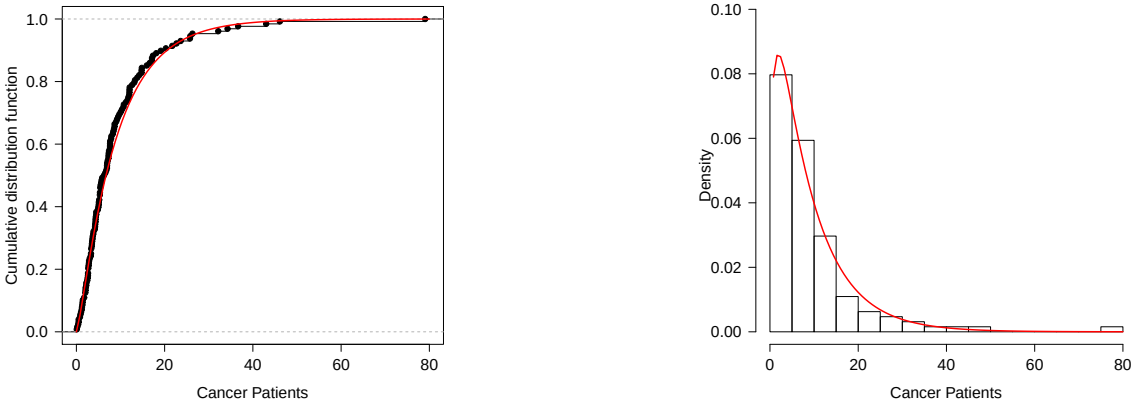
Table 2: ML estimates for BS, PNBS, SNBS and PSNBS models.

Parameters	γ	β	α	λ
BS	1.3740(0.0862)	4.5711(0.4461)	–	–
PNBS	3.2915(0.2856)	0.4227(0.6321)	5.1830(0.2051)	–
SNBS	2.3350(0.4131)	1.3566(0.3849)	–	1.9050(1.1294)
PSNBS	5.3315(3.0351)	0.1764(0.2060)	2.3024(0.4235)	2.5762(3.7211)

The values of these statistics for all models are given in Table 3. As expected, the values of AIC, AICC, W^*, A^*, K-S and p-value indicates better fit for the PSNBS model over the SNBS, PNBS and BS models. Figure 7 shows graphs for PSNBS model (a) empirical cdf (b) histogram and Figure 8 (a) and (b) shows the qq-plot for the models with better fit.

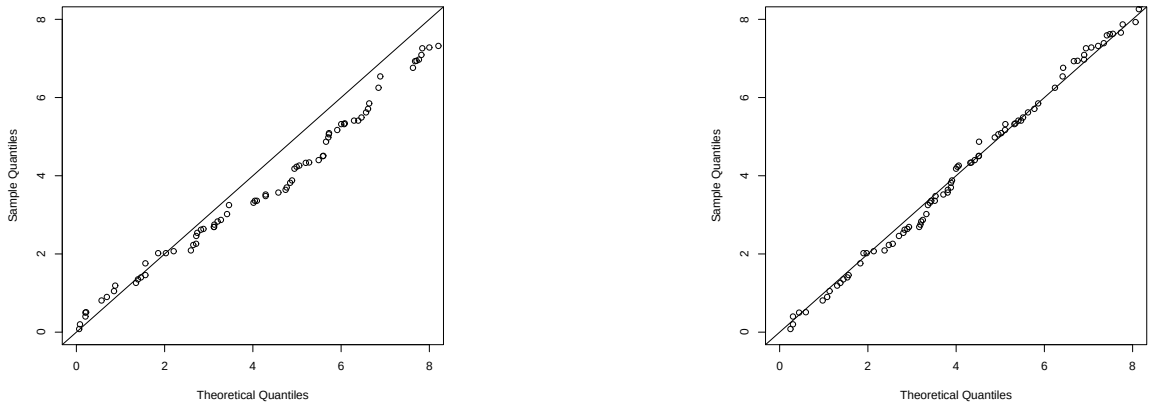
Table 3: AIC, AICC, W^*, A^*, K-S and p-value for the remission times of bladder cancer data for BS, PNBS, SNBS and PSNBS models.

	$\ell(\theta)$	AIC	AICC	W^*	A^*	K-S	P-value
BS	430.0420	864.0836	864.1898	0.4136	2.5615	0.1689	0.0013
PNBS	413.0645	832.1290	832.3433	0.1196	0.7219	0.0694	0.5680
SNBS	418.8570	843.7140	843.9283	0.1667	1.0930	0.1214	0.0459
PSNBS	411.8310	831.6620	832.0224	0.0829	0.5073	0.0623	0.7037



(a) (b)

Figure 7: Graphs for PSNBS model (a) empirical cdf (b) histogram.



(c) (d)

Figure 8: (a) qq-plot PNBS and (b) qq-plot PSNBS.

Note that the PSNBS model provides better fit to the data set analyzed. Therefore, the PSNBS model fits better than the other models, although it has one more parameter.

4.2. Application II

A major problem with forest areas is tree mortality due to various factors that can be seen as caused by stress through a phenomenon similar to material fatigue. In this context, two problems of great interest are tree mortality and the distribution of the diameter at the breast height (DBH). It has been observed that the BS distribution has a failure rate that can capture such features. As seen above, the ordinary BS is a particular case of the PSEBS distribution, so that the PSEBS is more flexible to explain skewness and kurtosis excess. Thus, we apply this distribution to explain the behavior of the variable DHB (in cm) in explaining forest mortality of Gray Birch (Betula populifolia Marshall) of a perennial with an average height of ten meters. The data basis consists of 160 trees and are available in Leiva *et al.* ([19]). Descriptive statistics results are summarized in Table 4. There is indication of high kurtosis in this data set, that suggest a more flexible model than the BS model, such as the PSTBS model. For this reason we implement the BS, BST, STBS and PSTBS models.

Table 4: Descriptive statistics for the data set.

n	$\bar{t}$	s^2	$\sqrt{b_1}$	b_2
160	14.5387	13.0510	2.8893	13.9716

Table 5 reports the estimates of the degrees of freedom, ν, for each model based on the t_ν distribution, which are obtained by maximizing the profile log-likelihood function. ML estimates (standard errors in parenthesis) are presented in Table 6.

Table 5: Estimation of ν for the BST, STBS and PSTBS models by maximizing the log-likelihood function.

	Log-likelihood	Log-likelihood	Log-likelihood
ν	BST	STBS	PSTBS
1	-406.4265	-402.8126	-390.2868
2	-392.7834	-387.5216	-383.0684
3	-389.9824	-383.4061	-381.0513
4	**−389.4381**	-381.8612	-380.0609
5	-389.5679	-381.1933	-379.6883
6	-389.9238	-380.8925	-379.4448
7	-390.3497	-380.8505	-379.4001
8	-390.7852	**−380.7285**	-379.0779
9	-391.2060	-381.0066	**−378.8113**
10	-391.6025	-383.8818	-379.3470

Table 6: ML estimates for BS, BST ($\nu = 4$), STBS ($\nu = 8$) and PSTBS ($\nu = 9$) models.

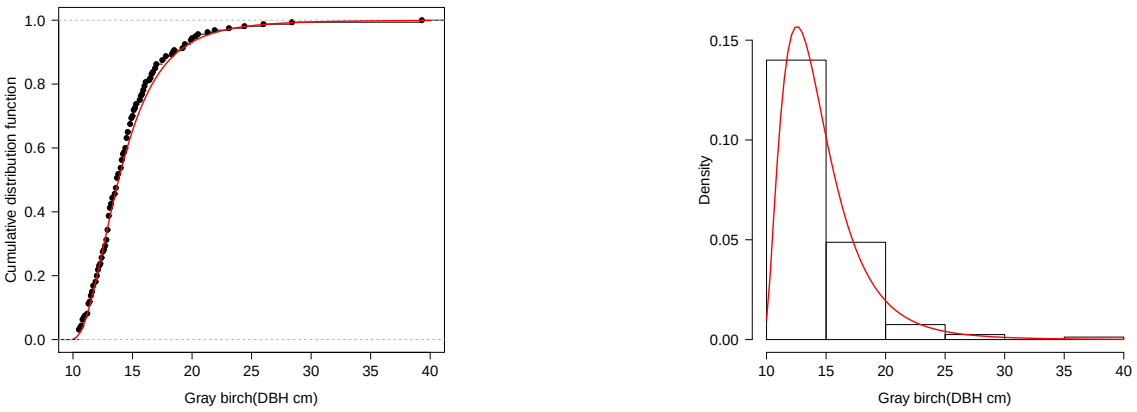
	γ	β	α	λ
BS	0.2083(0.0116)	14.2302(0.2331)	–	–
BST	0.151(0.074)	13.818(0.014)	–	–
STBS	0.2653(0.103)	11.346(0.025)	–	3.325(1.174)
PSTBS	0.2796(0.1135)	9.8844(0.1244)	2.3178(0.9654)	7.7185(11.4417)

According to the AIC and AICC criteria, W^*, A^*, K-S and p-value indicates better fit for the PSNBS model over the other models. See Table 7.

Table 7: AIC, AICC, W^*, A^*, K-S and p-value for the remission times of Gray birch data for BS, BST_4, STBS_8 and PSTBS_9 models.

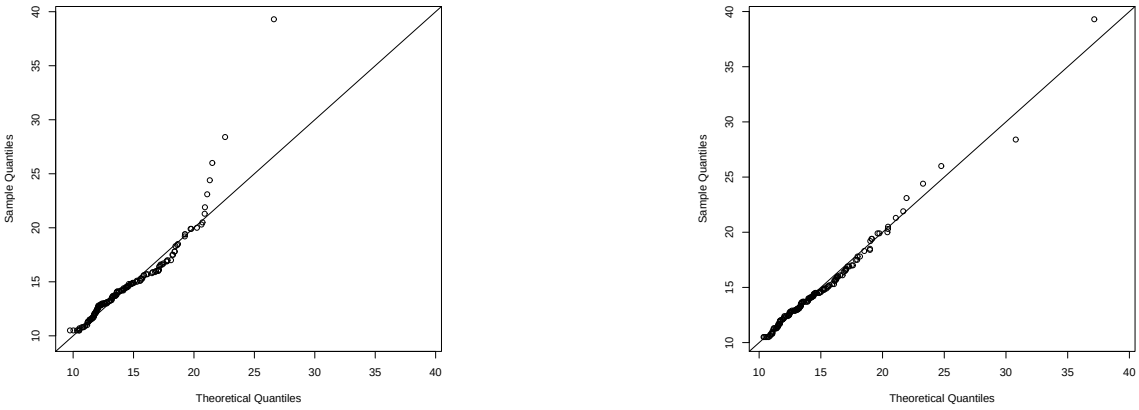
	$\ell(\theta)$	AIC	AICC	W^*	A^*	K-S	P-value
BS	399.7764	803.5528	803.6590	0.4396	2.7084	0.1066	0.0526
BST	389.4381	782.8762	782.9526	0.16602	1.1472	0.0707	0.4004
STBS	380.7285	767.4569	767.6108	0.04515	0.3166	0.0535	0.7506
PSTBS	378.8113	764.355	764.6131	0.040	0.2966	0.047	0.8614

Figure 9 shows graphs for PSTBS_9 model (a) empirical cdf (b) histogram and Figure 10 (a) and (b) shows the qq-plot for the models with better fit.



(a) (b)

Figure 9: Graphs for PSTBS_9 model (a) empirical cdf and (b) histogram.



(a) (b)

Figure 10: (a) qq-plot STBS_8 and (b) qq-plot PSTBS_9.

4.3. Application III (censored data)

The World Health Organization recommends breastfeeding exclusive for babies until 4 and 6 months. For this reason, an study from Universidade Federal de Minas Gerais UFMG main breastfeeding practice, as well as the possible factors of risk for an early weaning. The study consists of 150 mothers with children under 2 years of age. The response variable was the maximum time of breastfeeding, i.e., the time counted from birth to the weaning. More details on this data set can be found in Colosimo and Giolo ([7]). The values of the ML estimates for the BS, SNBS and PSNBS statistics for all models are given in Table 8. As expected, the values of AIC better fit for the PSNBS over other models, and the Figure 11 we can see that most babies stop having exclusive breastfeeding after 7 or 8 months.

Table 8: ML estimates for BS, SNBS and PSNBS models and AIC criteria.

	γ	β	α	λ	$\ell(\theta)$	AIC
BS	2.362 (0.268)	6.696(1.372)	–	–	-243.545	491.090
SNBS	5.380(1.538)	0.591(0.277)	–	4.015 (1.712)	-230.047	466.094
PSNBS	6.441(2.794)	0.252(0.227)	1.489(0.299)	3.761(2.287)	-228.357	464.715

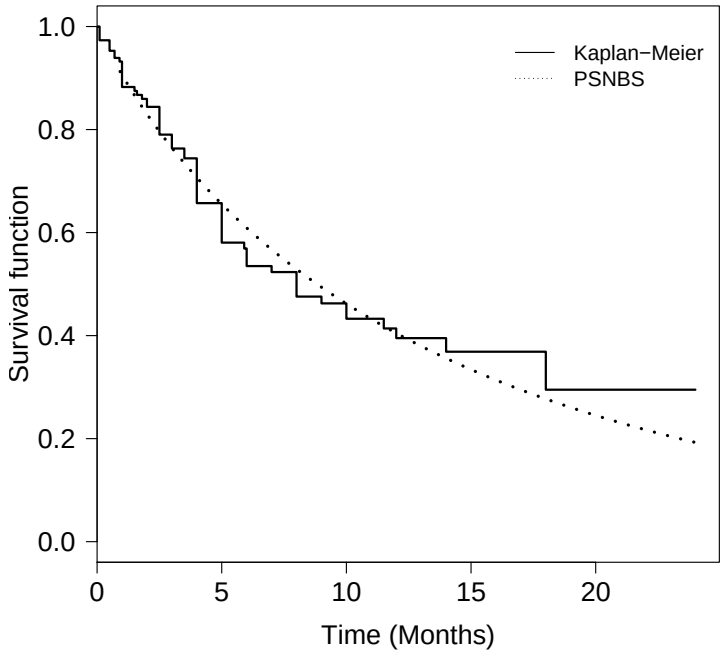


Figure 11: Estimated survival function for weaning study data under PSNBS model.

5. FINAL COMMENTS

This Paper proposes a flexible asymmetric BS distribution which contains previous ones as special cases and is able to surpass traditional models in terms of wider ranges of asymmetry and kurtosis. It is also shown that it is able to perform well in real applications, outperforming potential rival models. Maximum likelihood estimation is implemented and Fisher and observed information matrices are derived. It is shown that both are nonsingular. Some more features of this family of distributions are:

- The PSEBS model contains, as special cases, the SEBS model proposed by Vilca-Labra and Leiva-Sánchez ([30]) and the PEBS model proposed by Martínez-Flórez *et al.* ([22]).
- The proposed model it has a closed expression and presents more flexible asymmetry and kurtosis coefficients than PEBS and SEBS models.
- Some properties of the BS distribution were extended for the PSEBS model.
- The moments of the PSEBS family are finite.
- In the three applications it is shown that the PSEBS model fit better than the other models. This is confirmed by the different criteria used.

ACKNOWLEDGMENTS

The research of H.W. Gómez was supported by SEMILLERO UA-2020 (Chile). The research of H. Bolfarine was supported by CNPq and Fapesp (Brasil). We also acknowledge the valuable suggestions from the referees.

REFERENCES

[1] Akaike, H. (1974). A new look at statistical model identification, *IEEE Transaction on Automatic Control*, **19**, 716–723.

[2] Arellano-Valle, R.B. and Bolfarine, H. (1995). On some characterizations of the t-distribution, *Statistics & Probability Letters*, **25**, 79–85.

[3] Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

[4] Barros, M.; Paula, G.A. and Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics, *Lifetime Data Analysis*, **14**, 316–332.

[5] Cambanis, S.; Huang, S. and Simons, G. (1981). On the theory of elliptically contoured distributions, *J. Multivar. Anal.*, **11**, 365–385.

[6] Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test, *Journal of Quality Technology*, **27**, 154–161.

[7] Colosimo, E. and Giolo, S. (2006). *Análise de sobrevivencia aplicada*, ABE-Projeto Fisher.

[8] Cribari-Neto, E. and Branco, M. (2003). *Bayesian reference analysis for binomial calibration problem*, RT MAE 2003-12: IME-USP.

[9] Díaz-García, J.A. and Leiva-Sánchez, V. (2005). A new family of life distributions based on the elliptically contoured distributions, *J. Statist. Plann. Inference*, **128**, 445–457.

[10] Durrans, S.R. (1992). Distributions of fractional order statistics in hydrology, *Water Resources Research*, **28**, 1649–1655.

[11] Fang, K.T.; Kotz, S. and Ng, K.W. (1990). *Symmetric Multivariate and Related Distribution*, Chapman and Hall, London.

[12] Fang, K.T. and Zhang, Y.T. (1990). *Generalized Multivariate Analysis*, Sciences Press, Beijing, Springer-Verlag, Berlin.

[13] Gupta, A.K. and Varga, T. (1993). *Elliptically Contoured Models in Statistics*, Kluwer Academic Publishers, Boston.

[14] Johnson, S.; Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, (2nd ed.), Wiley, New York.

[15] Kelker, D. (1970). Distribution theory of spherical distributions and a location scale parameter generalization, *JSankhya (Ser. A)*, **32**, 419–430.

[16] Lagakos, S.W. (1979). A stochastic model for censored-survival data in the presence of an auxiliary variable, *Biometrics*, **35**, 139–156.

[17] Lee, E.T. and Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis*, (3rd ed.), Wiley, New York.

[18] Leiva, V. (2016). *The Birnbaum–Saunders Distribution*, Academic Press, New York.

[19] Leiva, V.; Ponce, M.; Marchant, C. and Bustos, O. (2012). Fatigue statistical distributions useful for modeling diameter and mortality of trees, *Revista Colombiana de Estadística*, **35**, 349–370.

[20] Leiva, V.; Vilca, F.; Balakrishnan, N. and Sanhueza, A. (2010). A skewed sinh-normal distribution and its properties and application to air pollution, *Communications in Statistics Theory and Methods*, **39**, 426–443.

[21] Lemonte, A.; Cribari-Neto, F. and Vasconcellos, K. (2007). Improved statistical inference for the two-parameter Birnbaum–Saunders distribution, *Computational Statistics and Data Analysis*, **51**, 4656–4681.

[22] Martínez-Flórez, G.; Bolfarine, H. and Gómez, H.W. (2014a). An alpha-power extension for the Birnbaum–Saunders distribution, *Statistics*, **48**(4), 896–912.

[23] Martínez-Flórez, G.; Bolfarine, H. and Gómez, H.W. (2014b). Skew-normal alpha-power model, *Statistics*, **48**(6), 1414–1428.

[24] Moreno-Arenas, G.; Martínez-Flórez, G. and Bolfarine, H. (2017). Skew-normal alpha-power model, *Revista Integración. Temas Mat.*, **35**(1), 51–70.

[25] Owen, D.B. (1956). Tables for computing bi-variate normal probabilities, *Annals Mathematical Statistics*, **27**, 1075–1090.

[26] Pewsey, A.; Gómez, H.W. and Bolfarine, H. (2012). Likelihood-based inference for power distributions, *TEST*, **21**(4), 775–789.

[27] Prudnikov, A.P.; Brychkov, Y.A. and Marichev, O.I. (1990). *Integrals and Series*, vols. 1, 2 and 3, Gordon and Breach Science Publishers, Amsterdam.

[28] R Development Core Team R. (2016). *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[29] Sanhueza, A.; Leiva, V. and Balakrishnan, N. (2008). The generalized Birnbaum–Saunders distribution and its theory, methodology and application, *Communications in Statistics-Theory and Methods*, **37**, 645–670.

[30] Vilca-Labra, F. and Leiva-Sánchez, V. (2006). A new fatigue life model based on the family of skew-elliptical distributions, *Communications in Statistics-Theory and Methods*, **35**, 229–244.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 661–672

MODELLING IRREGULARLY SPACED TIME SERIES UNDER PREFERENTIAL SAMPLING

Authors: Andreia Monteiro
– University of Minho & CIDMA, Portugal
andreiaforte50@gmail.com

Raquel Menezes
– University of Minho, Portugal
rmenezes@math.uminho.pt

Maria Eduarda Silva
– Faculty of Economics, University of Porto & CIDMA, Portugal
mesilva@fep.up.pt

Received: April 2018 Revised: September 2018 Accepted: October 2018

Abstract:

- Irregularly spaced time series are commonly encountered in the analysis of time series. A particular case is that in which the collection procedure over time depends also on the observed values. In such situations, there is stochastic dependence between the process being modeled and the times at which the observations are made. Ignoring this dependence can lead to biased estimates and misleading inferences. In this paper, we introduce the concept of preferential sampling in the temporal dimension and we propose a model to make inference and prediction. The methodology is illustrated using artificial data as well a real data set.

Key-Words:

- *preferential sampling; time series; continuous time autoregressive process; SPDE.*

AMS Subject Classification:

- 62M10, 62M20.

1. INTRODUCTION

Analysis of experimental data that have been observed at different points in time leads to specific problems in statistical modeling and inference. In traditional time series the main emphasis is on the case when a continuous variable is measured at discrete equispaced time points, [22]. There is an extensive body of literature on analyzing equally spaced time series data, see for example [3] and [6]. However, unevenly spaced (also called unequally or irregularly spaced) time series data naturally occurs in many scientific domains. Natural disasters such as earthquakes, floods, or volcanic eruptions typically occur at irregular time intervals. In observational astronomy, for example, measurements of properties such as the spectra of celestial objects are taken at irregularly spaced times determined by seasonal, weather conditions, and availability of observation time slots. In clinical trials (or more generally, longitudinal studies), a patient's state of health may be observed only at irregular time intervals, and different patients are usually observed at different points in time.

It must be noted that sometimes equally spaced time series are treated as irregularly spaced time series, namely time series with missing observations and multivariate data sets that consist of time series with different frequencies, even if the observations of each time series are reported at regular intervals. One of the first to treat evenly sampled gene expression time series with missing values as unevenly sampled data is [19].

There are few methods available in the literature for the analysis of irregularly spaced series. Some authors, such us [10], [12], [2] and [5] have suggested an embedding into continuous diffusion processes, with the aim of using the well established tools for univariate autoregressive moving average (ARMA) processes.

Observations with irregularly spaced sampling times are much harder to work with, partly because the established and efficient algorithms developed for equally spaced sampling times are no longer applicable [15]. A common approach to perform parametric estimation is to construct a log-likelihood function in terms of the unknown parameter [4]. When the sampling times are considered deterministic, the traditional approach is to build the classical Gaussian log-likelihood function. However, because the inversion of the covariance matrix has to be performed, numerical evaluation of this Gaussian log-likelihood function is in general very expensive [14]. One way to overcome this computational effort is to regulate the sampling scheme, using some form of interpolation, and consider it as being equally spaced. Under the assumption of equally spaced sampling times, the Gaussian log-likelihood function can be approximated, at least for a sufficiently large sample, by the Whittle log-likelihood function [24]. This approach has been successfully applied to irregularity caused by missing values, [16]. While, it may be reasonable to use this methodology, to deal with the minor irregularities in sampling times caused by missing values, the interpolation procedure will typically change the dynamic of the underlying process, leading to biased estimates for the parameters [9]. Moreover, there is little understanding of which particular interpolation method is the most appropriate on a given data set. Alternatively, a convenient continuous time domain dynamic model may be assumed for the underlying continuous time stationary process such as the Continuous time ARMA (CARMA) model. The application of Kalman recursion techniques to the parametric estimation of CARMA processes is reviewed in [22]. Additionally, [13] estimate the parameters of an irregularly sampled CARMA process using a Bayesian framework.

A particular case of irregularly spaced data is that in which the collection procedure along time depends also, for practical constraints, on the observed values. For example, a certain health indicator for an individual may be measured at different time points and with different frequencies depending on his health state. In a completely different setting, the times of occurrence of transactions in the financial markets depend largely on the value of the underlying asset. In environmental monitoring applications, or in the context of smart cities if it is decided to monitor more frequently when a value considered critical to human health is exceeded. Therefore, additional information on the phenomena under study is obtained from the frequency or time occurrence of the observations. In such situations, there is stochastic dependence between the process being modeled and times of the observations, which may be coined as temporal preferential sampling following [8] in the context of spatial statistics.

In this work, we propose a model-based approach to analyze a time series observed under preferential sampling. The suggested framework considers the observed time points as the realization of a time point process stochastically dependent on an underlying latent process (e.g. an individual health indicator or the underlying asset). This latent process is assumed as Gaussian without loss of generality.

The paper is organized as follows. Section 2 describes our proposed model for preferential sampling in time dimension, namely to make inference and prediction. In Section 3 we describe the Monte Carlo Maximum Likelihood Estimation. In section 4 we conduct a numerical illustration, in an artificial data set, to analyze the quality of the proposed model. We then show the application of the previously described methodology to a real data set related to monitoring the level of a biomedical marker, after a cancer patient undergoes a bone marrow transplant. Section 5 is devoted to make some concluding remarks.

2. A MODEL FOR PREFERENTIAL SAMPLING

In time series, data are obtained by sampling a phenomenon $S(t) : t > 0$ at a discrete set of times t_i, $i = 1, ..., n$. Admiting the possibility that the sampling design may be stochastic, $T = (t_1, ..., t_n)$ denotes a stochastic process of observation times. In many situations, $S(t)$ cannot be measured without error, hence, if $Y(t_i)$ denotes the measured value at time t_i, a model for the data takes the form:

$$Y(t) = \mu + S(t) + N(0, \tau^2),\ t > 0 \tag{2.1}$$

where μ is a constant mean effect and $S(\cdot)$ is a stationary Gaussian process with $E[S(t)] = 0$. An equivalent formulation is that conditional on $S(\cdot)$, the $Y(t_i)$ are mutually independent, normally distributed with mean $\mu + S(t_i)$ and common variance τ^2.

We consider $S(\cdot)$ as a continuous time autoregressive process of order 1, CAR(1), that satisfies the differential equation $dS(t) + \alpha_0 S(t)dt = dW(t)$ where, α_0 is the autoregressive coefficient, $S(\cdot)$ is asymptotically stationary if an only if $\alpha_0 > 0$ and $W(t)$ is a Brownian motion with variance parameter σ_w^2. For notation simplification let us denote $Y_i = Y(t_i)$. Then $Y = (Y_1, ..., Y_n)$ is multivariate Gaussian with mean $\mu\mathbf{1}$ and covariance matrix $\Sigma_Y = \frac{\sigma_w^2}{2\alpha_0} R_y(\alpha_0) + \tau^2 I_n$, where $\mathbf{1}$ is a n-length vector of ones, I_n is the $n \times n$ identity matrix and

$R_y(\alpha_0)$ has elements $r_{ij} = \rho(|t_i - t_j|; \alpha_0)$ defined by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \exp(-\alpha_0 |h|) \tag{2.2}$$

being $\gamma(\cdot)$ the covariance function.

Admitting that the sampling times are stochastic, a complete model needs to specify the joint distribution of S, T and Y. Considering the stochastic dependence between S and T, the model to deal with preferential sampling is defined through $[S, T, Y]$ written as:

$$[S][T|S][Y|S(T)] \tag{2.3}$$

where $[\cdot]$ means "the distribution of", $S = \{S(t) : t > 0\}$, $T = (t_1, ..., t_n)$ and $S(T)$ represents $\{S(t_1), ..., S(t_n)\}$.

We define a specific class of models through the additional assumptions: conditional on S, T is an inhomogeneous Poisson process with intensity $\lambda(t) = \exp\{a + \beta S(t)\}$ and unconditionally T is a log-Gaussian Cox process. The log-Gaussian Cox process is a flexible class of point pattern models that allows conditioning the sampling times to the variable of interest. β is the parameter that controls the degree of preferentiality, for example, $\beta = 2$ corresponds to a situation when the sampling times are concentrated, predominantly, near the maximum of the observed values and $\beta = 0$ corresponds to the situation of an homogeneous, non-preferential, sampling. Conditional on S and T, Y is a set of mutually independent Gaussian variates with τ^2 being the measurement error variance.

The predicted value of $S(\cdot)$ at an unsampled time point $t_{n_i} < t_0 < t_{n_j}$, $S(t_0|T)$, is given by $S(t_0|T) = \mathrm{E}\big[S(t_0)|Y(T)$. Considering that the process CAR(1) is Markovian, [6, p.358] shows that the conditional mean of $S(t_0)$ given $Y(T)$ is

$$\begin{aligned} S(t_0|T) &= \mathrm{E}\big[S(t_0)|Y(T)\big] \\ &= \exp\left(-\alpha_0(t_0 - t_{n_i})\right) Y(T) + \mu\left(1 - \exp\left(-\alpha_0(t_0 - t_{n_i})\right)\right). \end{aligned} \tag{2.4}$$

The variance of the prediction is

$$\sigma^2(t_0) = Var\big[S(t_0)|Y(T)\big] = \frac{\sigma_w^2}{2\alpha_0}\left(1 - \exp\left(-2\alpha_0(t_0 - t_{n_i})\right)\right). \tag{2.5}$$

3. MONTE CARLO MAXIMUM LIKELIHOOD ESTIMATION

We consider a discretization of the S process with N points and a partition of S into $S = \{S_0, S_1\}$, where S_0 denotes the values of S at each of n times $t_i \in T$, and S_1 are the values of S at the remaining $(N - n)$.

The likelihood function for data T and Y can be expressed as

$$L(\boldsymbol{\theta}) = [T, Y] = \int_S [T, Y, S] dS = \int_S [S][T, Y|S] dS = \int_S [S][T|S][Y|T, S] dS \tag{3.1}$$

where $\boldsymbol{\theta} = (\mu, \sigma_w, \alpha_0, \tau, \beta)$ represents all the model parameters.
An algebraic simplification of $[Y|T,S]$ is $[Y|S_0]$ so, we can rewrite the integral as

$$L(\boldsymbol{\theta}) = \int_S [S][T|S][Y|S_0]\frac{[S|Y]}{[S|Y]}dS. \tag{3.2}$$

Considering that $[S] = [S_1, S_0] = [S_1|S_0][S_0]$ and replacing the term $[S|Y]$ in the denominator of expression (3.2) by $[S|Y] = [S_0, S_1|Y] = [S_1|S_0, Y][S_0|Y] = [S_1|S_0][S_0|Y]$, equation (3.2) becomes

$$\begin{aligned} L(\theta) &= \int_S [S_1|S_0][S_0][T|S][Y|S_0]\frac{[S|Y]}{[S_1|S_0][S_0|Y]}dS \\ &= \int_S [T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0][S|Y]dS \\ &= E_{S|Y}\left[[T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0]\right]. \end{aligned} \tag{3.3}$$

Taking into account that the above conditional expectation can be approximated by Monte Carlo, MLE's are obtained by maximizing the Monte Carlo likelihood

$$L_{MC}(\boldsymbol{\theta}) = m^{-1}\sum_{j=1}^{m}[T|S_j]\frac{[Y|S_{0j}]}{[S_{0j}|Y]}[S_{0j}] \tag{3.4}$$

where S_j are assumed as realizations of the distribution of S conditional on Y. S_{0j} denotes the values of S_j restricted to the n observed time points. We may notice that j takes a value from 1 to m, the total number of Monte Carlo replicates. With this purpose, we use a technique known as conditioning by kriging [18] and we use the following construction. The new sample $S_j = U + \Sigma_S A^T\left(A\Sigma_S A^T + \tau^2 I_n\right)^{-1}(V - AU)$ where A is the $n \times N$ matrix whose ith row consists of $N-1$ 0s and a single 1 to identify the position of t_i within $T = (t_1, ..., t_n)$; $U = \Sigma_S^{1/2}u \sim MVN(0, \Sigma_S)$ with $u \sim N(0,1)$ and $\Sigma_S^{1/2}$ is obtained from the Cholesky decomposition and $V \sim MVN(y, \Sigma_Y)$. Then S_j has the required multivariate Gaussian distribution of S given $Y = y$. In practice, we use antithetic pairs of realizations to reduce Monte Carlo variance [8].

$T|S_j$ in (3.4) is an inhomogeneous Poisson process with intensity

$$\lambda(t) = \exp\{a + \beta S_j(t)\}. \tag{3.5}$$

For computational reasons, we work with logarithm and thus,

$$\log([T|S_j]) = \sum_{i=1}^{n}(a + \beta S_j(t_i)) - n\log\left(\int_0^T \exp(a + \beta S_j(t))dt\right). \tag{3.6}$$

As the S_j replicate is not known in $[0,T]$ domain, we can not calculate the integral presented in expression (3.6), so, we approximate the integral using the composed trapezium formula for unequally spaced data.

$[S_{0j}]$ in (3.4) is multivariate Gaussian with mean 0 and covariance matrix $\Sigma_{S_{0j}} = \frac{\sigma_w^2}{2\alpha_0}R_{S_{0j}}(\alpha_0)$, where $R_{S_{0j}}(\alpha_0)$ is the $n \times n$ correlation matrix with elements $r_{ij} = \rho\left(|t_i - t_j|; \alpha_0\right)$ defined by (2.2).

$[S_{0j}|Y]$ in (3.4) is multivariate Gaussian with mean $\mu_{S_{0j}|Y} = \Sigma_{S_{0j}}\Sigma_Y^{-1}(y - \mu\mathbf{1})$ and covariance matrix $\Sigma_{S_{0j}|Y} = \Sigma_{S_{0j}} - \Sigma_{S_{0j}}\Sigma_Y^{-1}\Sigma_{S_{0j}}^t$. For more details about conditional distribution see for e.g. [1].

Obtained the Maximum Likelihood Estimates (MLE's), we can plug them into (2.4) and (2.5), treating them as known. We are in position of doing the so-called plug-in predictions.

4. NUMERICAL ILLUSTRATION

In this section we document the performance of the model with time series simulated under preferential and non preferential (irregular and regular sampling) scenarios. The simulation allows control the degree of preferentiality. In addition, we apply our modeling procedure to a time series related to the biomedical marker level of platelet after a cancer patient undergoes a bone marrow transplant. Taken together, these examples suggest that our model is effective at detecting potential preferential sampling situations, estimating an adequate model and obtaining predictions for the process. We compare the results from our model with the traditional Kalman filter approach to irregularly spaced data (cts package [23]). We begin by describing the procedure to simulate a time series under preferential sampling.

4.1. Artificial data

To generate a time series under preferential sampling we first generate a realization of S from model (2.1) with $\alpha_0 = 0.2$ and $\sigma_w^2 = 1$, discretized in 400 equally spaced time points. These values correspond to $Var[S(.)] = \sigma^2 = \frac{\sigma_w^2}{2\alpha_0} = (1.581)^2$ and $\phi = \frac{1}{\alpha_0} = 5$, being the latter related to the lag beyond which there is no correlation for practical purposes. To generate Y from model (2.1), we consider $\mu = 0$ and $\tau = 0.1$, conducting three separate sampling procedures over the realization of S:

- preferential sampling: conditional on the values of S, we obtain $n = 70$ sampling times T following an inhomogeneous Poisson process with intensity function defined in (3.5) and $\beta = 2$;
- irregular sampling: we obtain $n = 70$ sampling times T from (3.5) and with $\beta = 0$, illustrating the situation without preferential sampling;
- regular sampling: we obtain $n = 70$ sampling times with equidistant observations.

To illustrate the results of these sampling schemes, we represent in Figure 1 a realization of the process S (gray line) and the three resulting data sets. We have 70 sampling times (black points), considering $\beta = 2$ in the process intensity function, in which the preferential nature of the sampling process results in sample times falling predominantly near the maxima. For 70 sampling times (white points), we consider $\beta = 0$, the situation without preferential sampling and with irregularly sampling points. For the remaining 70 points (star points), we have the situation of regular spaced sampling times.

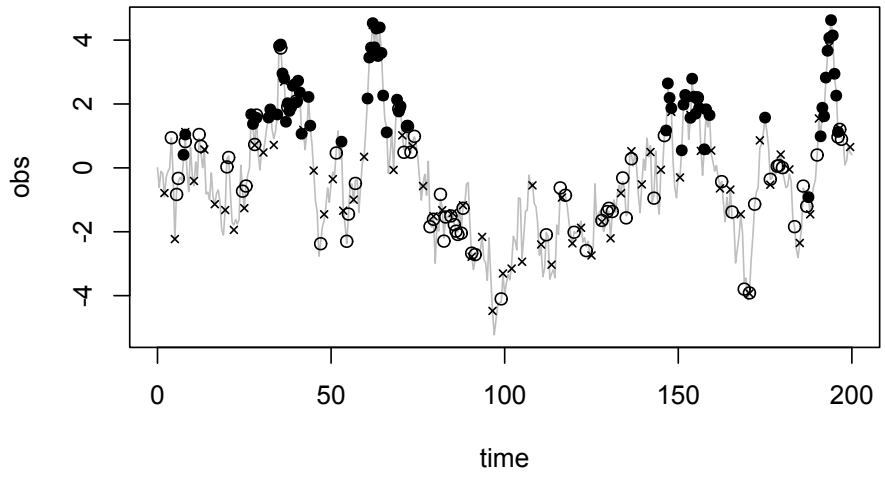


Figure 1: Sample times with preferential sampling nature (black points), without preferential sampling and irregularly spaced time points (white points), regular spaced time points (star points) and underlying process S (gray line).

The parameters μ, σ, ϕ, τ and β are the target of estimation. The estimates are obtained under (3.4), henceforward denoted by MCMLE's and from the Kalman filter, denoted by MLE's. For the maximization of our Monte Carlo log-likelihood function we considered a total of grid points $N = 400$ and a total number of MÇ replicates $m = 1000$. Mean and standard errors for the estimates obtained from 250 independent simulated samples are summarized in Table 1.

Table 1: Maximum likelihood estimates, under PS model (MCMLE's) and by cts package (MLE's), mean (standard errors) obtained from a total of 250 independent samples.

		PS Data set ($\beta = 2$)		Irregularly Sampling ($\beta = 0$)		Regular sampling	
	True	PS model	CTS	PS Model	CTS	PS Model	CTS
$\widehat{\mu}$	**0**	**0.13** (0.18)	**0.38** (0.31)	**0.04** (0.12)	**0.26** (0.34)	**0.02** (0.22)	**0.71** (0.62)
$\widehat{\sigma}$	**1.58**	**1.53** (0.21)	**0.99** (0.18)	**1.64** (0.11)	**1.52** (0.21)	**1.60** (0.13)	**1.45** (0.24)
$\widehat{\phi}$	**5**	**5.71** (1.01)	**3.17** (2.55)	**5.20** (0.48)	**5.52** (1.96)	**5.12** (0.89)	**6.78** (2.93)
$\widehat{\tau}$	**0.1**	**0.12** (0.04)	**0.27** (0.13)	**0.11** (0.01)	**0.30** (0.18)	**0.11** (0.02)	**0.55** (0.28)
$\widehat{\beta}$	**2 or 0**	**1.76** (0.39)		**0.00** (0.07)		**0.00** (0.02)	

Analysing Table 1 we conclude that the model for Temporal Preferential Sampling presents estimates for the parameters less biased, even when the preferability degree is null, with regular and irregularly sampling.

To analyse the impact of ignoring preferential sampling on the quality of predictions, we conducted a second simulation study. We simulated 250 realizations of S and for each we constructed a preferential sampling data set. Then, the proposed MCMLE's and the MLE's from the Kalman filter approach were obtained and plugged-in equation (2.4) to predict $S(t)$ at 50 equally spaced time points. These together with the corresponding standard errors, in (2.5), allowed us to calculate prediction 95% confidence intervals and estimate their coverage.

Figure 2 represents one simulation of $S(t)$ (black line), the corresponding preferential sampling data (black points) and the predictions acquired from MCMLE's (white points) and MLE's (gray points). MLE's which do not take into account the preferential character of the data lead to predictions with larger bias (overestimation of the observations) and smaller variance than that of MCMLE's. In fact, in the overall simulation results confidence intervals from MCMLE's present an estimated coverage of 88% while the MLE's provide an estimated coverage of just 73%. Thus, the proposed model leads to estimates that are less biased but with larger variance, reflecting the uncertainty associated with the observations.

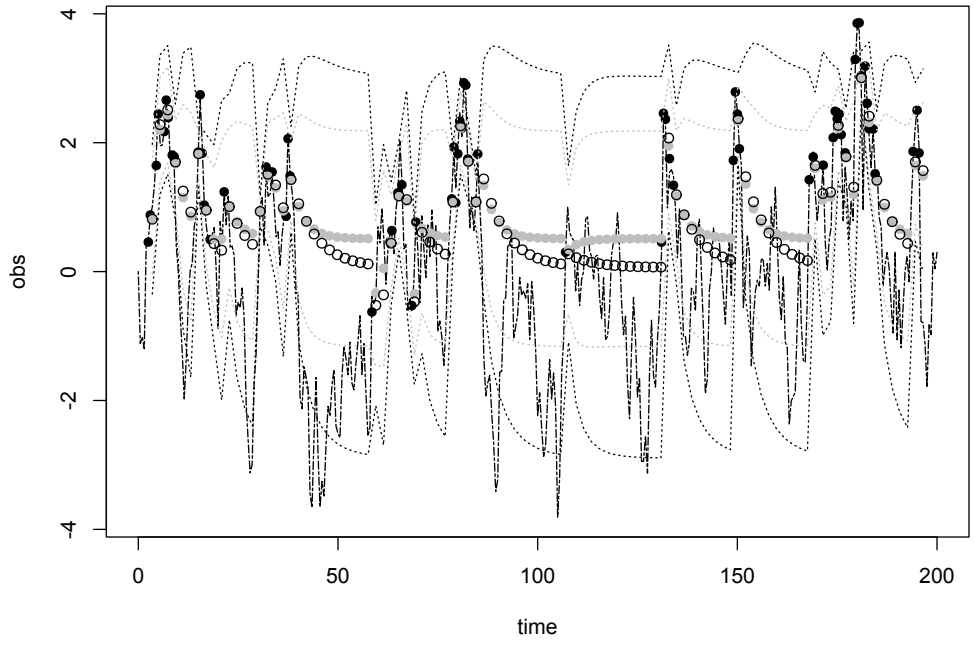


Figure 2: Predictions acquired from MCMLE's (white points) and MLE's (gray points), dashed line are confidence bands, black points are the preferential sampling data and black line is the underlying process S.

Further studies with β taking non-integer and negative values (sampling times are concentrated, predominantly, near the minima of the observed values) lead to similar conclusions.

4.2. Biomedical marker

We consider the problem of monitoring the level of a biomedical marker, platelet, after a cancer patient undergoes a bone marrow transplant. The data in Figure 3, studied in [20] as missing data problem, are 91 measurements made different days on variable log(platelet) [PLT]. In the first 35 days the data were observed daily and then irregularly, once the indicator began to show better results. According to [11], "Platelet count at about 100 days post transplant has previously been shown to be a good indicator of subsequent long term survival". This data is available in the package astsa [21] with the name of "blood".

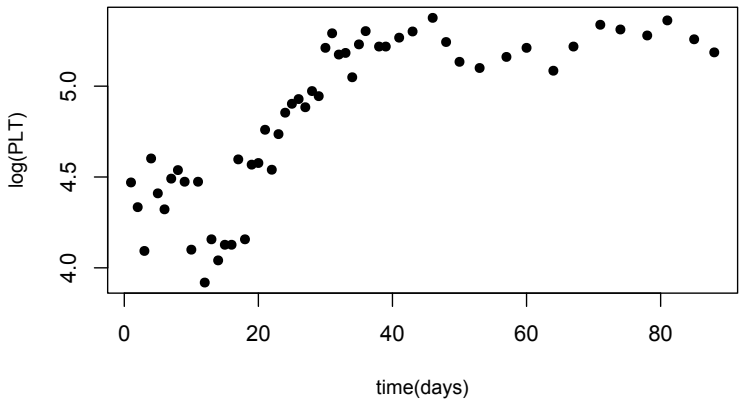


Figure 3: Measurements of the log(platelet) [PLT].

The MCMLE's for model parameters are: $\hat{\mu} = 1.99$, $\hat{\phi} = 66.18$, $\hat{\sigma} = 0.72$, $\hat{\tau} = 0.11$ and $\hat{\beta} = -2.01$. The estimated value for β with its negative sign indicates that the data was, in fact, observed under a preferential framework whereby the patient is observed more frequently when the biomarker shows lower values. Predictions of the biomarker within the period of observations are obtained plugging-in the estimated parameters in equations (2.4) and (2.5). Figure 4 top panel shows the 95% prediction intervals for (log of) the biomarker while the bottom panel represents the 95% prediction intervals obtained from the MLE's from the Kalman filter approach, with $\hat{\mu} = 1.57$, $\hat{\phi} = 53.94$, $\hat{\sigma} = 0.42$ and $\hat{\tau} = 0.13$. As expected in view of the simulation results, the predictions obtained from MCMLE present larger variance reflecting the uncertainty associated with the preferential data under analysis.

This kind of study is important, for example, to analyse when a new measurement of the patient's health indicator should be taken.

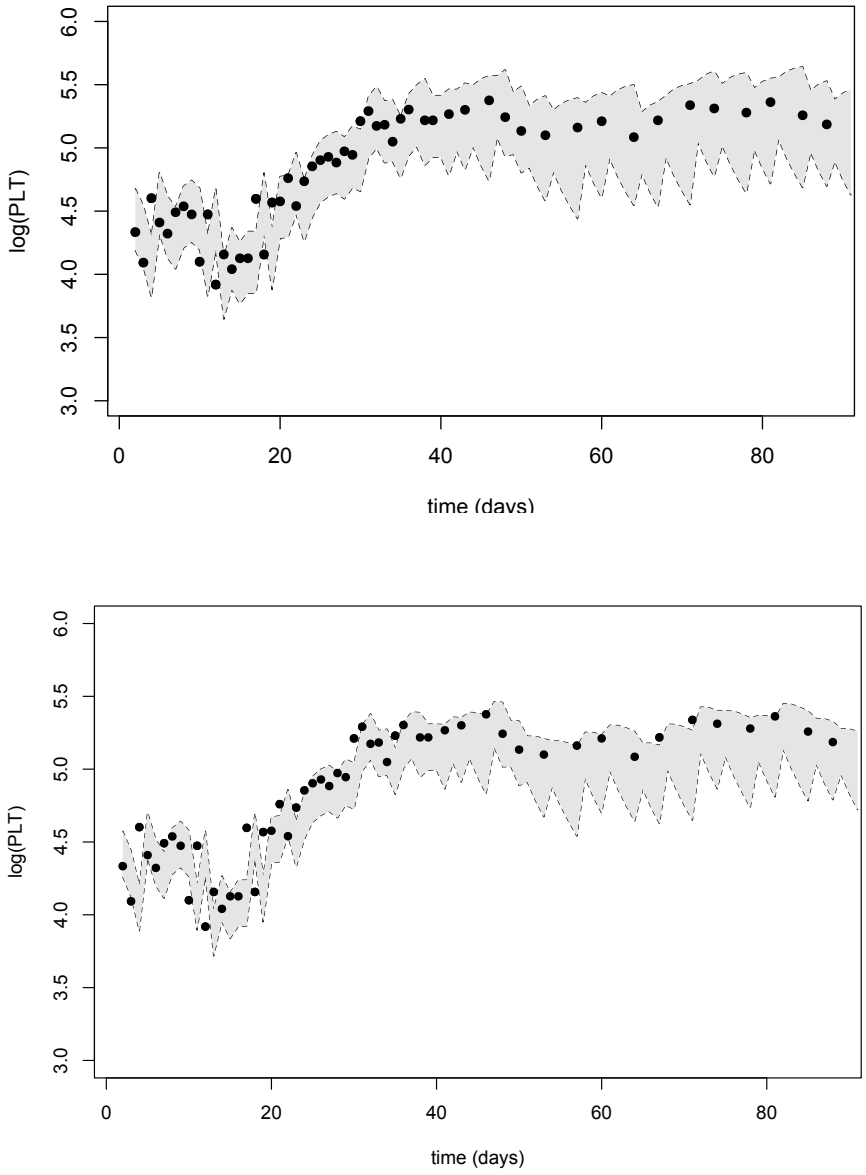


Figure 4: Prediction 95% confidence intervals using predictions acquired from MCMLE's (top) and MLE's (bottom).

5. CONCLUDING REMARKS AND FUTURE WORK

We propose, in this work, a methodology to deal with irregularly spaced time series but also a methodology that takes into account the frequency or time occurrence of the observations. The proposed model not only provides good estimates for model parameters but also reveals quite satisfactory results for prediction. A key aspect of this methodology is that it provides a tool, for example in the context of clinical trials, supporting a better knowledge of the underlying stochastic process, goal of study.

In their work, [7] affirm that the use of a single parameter in (3.5) to capture both the strength of the preferentiality and the amount of non-uniformity in sampling locations is somewhat inflexible. Alternatively, a more flexible and computational more efficient class of models, based on the proposal of [17], is discussed. These authors suggest an extension to the model proposed by [8], by adding a second Gaussian process and use of stochastic partial differential equation models. For future investigation we intend to adapt those suggestions to the time dimension.

ACKNOWLEDGMENTS

The authors acknowledge Foundation FCT (Fundação para a Ciência e Tecnologia) for funding through Individual Scholarship PhD PD/BD/ 105743/2014, Centre of Mathematics of Minho University and Center for Research & Development in Mathematics and Applications of Aveiro University within project UID/MAT/04106/2019.

REFERENCES

[1] Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

[2] Belcher, J.; Hampton, J. and Wilson, G.T. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data, *Journal of the Royal Statistical Society Series B (Methodological)*, **56**, 141–155.

[3] Box, G.E.; Jenkins, G.M.; Reinsel, G.C. and Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons.

[4] Brockwell, P.J. (2001). Continuous-time ARMA processes, *Handbook of Statistics*, **19**, 249–276.

[5] Brockwell, P.J. (2009). Levy-driven continuous-time ARMA processes, *Handbook of Financial Time Series*, 457–480.

[6] Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting* (second ed.), Springer.

[7] Diggle, P.J. and Giorgi, E. (2016). Preferential sampling of exposure levels, *Hanbook of Environmental and Ecological Statisics*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

[8] Diggle, P.J.; Menezes, R. and Su, T.l. (2010). Geostatistical inference under preferential sampling, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 191–232.

[9] Erdogan, E.; Ma, S.; Beygelzimer, A. and Rish, I. (2005). *Statistical models for unequally spaced time series*. In: "Proceedings of the 2005 SIAM International Conference on Data Mining", 626–630.

[10] Jones, R.H. (1981). Fitting a continuous time autoregression to discrete data, *Applied Time Series Analysis II*, Elsevier, 651–682.

[11] Jones, R.H. (1984). *Fitting multivariate models to unequally spaced data*. In "Time Series Analysis of Irregularly Observed Data" (E. Parzen, Ed.), Lecture Notes in Statistics, vol. 25, Springer, New York, pp. 158–188.

[12] Jones, R.H. (1985). Time series analysis with unequally spaced data, *Handbook of Statistics*, **5**, 157–177.

[13] Kelly, B.C.; Becker, A.C.; Sobolewska, M.; Siemiginowska, A. and Uttley, P. (2014). Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets, *The Astrophysical Journal*, **788**, 1–33.

[14] Lange, K. (2010). *Numerical Analysis for Statisticians*, Springer Science & Business Media.

[15] Li, Z. (2014). *Methods for irregularly sampled continuous time processes*, Ph.D. Thesis, UCL (University College of London).

[16] Little, R.J. and Rubin, D.B. (2014). *Statistical Analysis with Missing Data*, John Wiley & Sons.

[17] Pati, D.; Reich, B.J. and Dunson, D.B. (2011). Bayesian geostatistical modelling with informative sampling locations, *Biometrika*, **98**(1), 35–48.

[18] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, CRC press.

[19] Ruf, T. (1999). The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series, *Biological Rhythm Research*, **30**(2), 178–201.

[20] Shumway, R.H. and Stoffer, D.S. (2017). *Time Series Analysis and Its Applications: With R Examples*, Edition 4, Springer, New York.

[21] Stoffer, D.S. (2017). astsa: Applied Statistical Time Series Analysis, *R Package Version*, **1**, https://CRAN.R-project.org/package=astsa.

[22] Tomasson, H. (2015). Some computational aspects of Gaussian CARMA modelling, *Statistics and Computing*, **25**(2), 375–387.

[23] Wang, Z. (2013). cts: an R package for continuous time autoregressive models via kalman filter, *Journal of Statistical Software*, **53**(5), 1–19, http://www.jstatsoft.org/v53/i05/.

[24] Whittle, P. (1961). Gaussian estimation in stationary time series, *Bull. Internat. Statist. Inst.*, **39**, 105–129.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 673–695

ON A SUM AND DIFFERENCE OF TWO LINDLEY DISTRIBUTIONS: THEORY AND APPLICATIONS

Authors: Christophe Chesneau
– Université de Caen, LMNO, Campus II, Science 3, Caen, 14032, France
christophe.chesneau@unicaen.fr

Lishamol Tomy
– Deva Matha College, Department of Statistics, Kuravilangad, Kerala, 686633, India
lishatomy@gmail.com

Jiju Gillariose
– St. Thomas College, Department of Statistics, Pala, Kerala, 686574, India
jijugillariose@yahoo.com

Received: November 2019 Accepted: January 2020

Abstract:

- This paper investigates theoretical and practical aspects of two basic random variables constructed from Lindley distribution. The first one is defined as the sum of two independent random variables following the Lindley distribution (with the same parameter) and the second one is defined as the difference of two independent random variables following the Lindley distribution (with the same parameter). Then, statistical inference is performed. In both the cases, we assess the performance of the maximum likelihood estimators using simulation studies. The usefulness of the corresponding models are proved using goodness-of-fit tests based on different real datasets.

Key-Words:

- *convolution; data analysis; Lindley distribution; maximum likelihood estimation; moment estimator.*

AMS Subject Classification:

- 60E05, 62E15, 62F10.

1. INTRODUCTION

Statistical distributions have been widely applied over the past decades for modeling data in several areas. In fact, the statistics literature is filled with hundreds or thousands of continuous univariate distributions. Among them, the exponential distribution is perhaps the most widely applied statistical distribution in various fields, mainly because of the simplicity of its mathematical quantities like moments, moment generating function, etc. However, under some comparison criteria, it was shown that the Lindley distribution is a reliable alternative to the exponential distribution in modeling lifetime data. The Lindley distribution has a cumulative density function (cdf) of the form

$$F_*(x) = 1 - \left(1 + \frac{\theta}{1+\theta}x\right) e^{-\theta x}, \quad x, \theta > 0.$$

The corresponding probability density function (pdf) is given by

$$f_*(x) = \frac{\theta^2}{1+\theta}(1+x)e^{-\theta x}, \quad x, \theta > 0. \tag{1.1}$$

As indicated by its name, this distribution was introduced by [14, 15] to illustrate a difference between fiducial distribution and posterior distribution. In the recent years, the Lindley distribution is mainly used for studying stress-strength reliability modeling. It finds applications in various areas such as engineering, demography, reliability, medicine and biology. Its detailed properties can be found in [6], [10], [3], [1], [24], [27], etc.

In the last decades, its different generalizations have been emerged in distribution theory and applications. In particular, the reader is refereed to the three parameters-Lindley distribution [31], generalized Poisson-Lindley distribution [16], generalized Lindley distribution [22], Marshall-Olkin Lindley distribution [32], power Lindley distribution [5], two-parameter Lindley distribution [25], quasi Lindley distribution [26], transmuted Lindley distribution [18], transmuted Lindley-geometric distribution [19], beta-Lindley distribution [20] and discrete Harris extended Lindley distribution [29], among others. Moreover, a latest version of the Lindley distribution, called modified Lindley distribution, is given by [2]. Further, Lindley distribution and its generalizations have been studied extensively by [30].

In this note, we consider two independent random variables following the Lindley distribution with appropriate parameter and study the convolutions (sum and difference) of their distributions. In addition, we investigate applications and structural properties of the new models. In fact, this is a pioneering work in investigating comprehensively the applications and properties of exact distributions of the sum and difference of Lindley random variables. The article is outlined as follows: Section 2 deals with a detailed study of sum of two independent Lindley distributions. Section 3 presents difference of two independent Lindley distributions. Finally, Section 4 offers some concluding remarks.

2. ON THE SUM OF TWO INDEPENDENT LINDLEY DISTRIBUTION

This section is devoted to the sum of two independent random variables following the Lindley distribution with pdf given by equation (1.1), including its main theoretical properties and modeling.

2.1. Definition

We consider the pdf given by

$$f(x) = \frac{\theta^4}{(1+\theta)^2} x \left(\frac{x^2}{6} + x + 1 \right) e^{-\theta x}, \quad x, \theta > 0. \tag{2.1}$$

The feature of this distribution is the following: let X and Y be two independent random variables following the Lindley distribution with parameter θ. Then, the random variable $Z = X + Y$ has the pdf given by (2.1). This result is a particular case of [8, Theorem 2]. A crystal clear proof is given below. Since X and Y are independent, the pdf of Z is given by the following convolution product: for $x > 0$,

$$\begin{aligned} f(x) &= \int_{-\infty}^{+\infty} f_*(x-t) f_*(t) dt = \int_0^x \frac{\theta^2}{1+\theta}(1+x-t) e^{-\theta(x-t)} \frac{\theta^2}{1+\theta}(1+t) e^{-\theta t} dt \\ &= \frac{\theta^4}{(1+\theta)^2} e^{-\theta x} \int_0^x (1+x-t)(1+t) dt = \frac{\theta^4}{(1+\theta)^2} x \left(\frac{x^2}{6} + x + 1 \right) e^{-\theta x}. \end{aligned}$$

For the purpose of this study, the corresponding distribution is called the 2S-Lindley distribution (2S for Sum of 2 random variables). To the best of our knowledge, there is no work on the theoretical and practical aspect of this distribution, which motivates a part of this study.

As a first approach, some possible shapes of the pdf of the 2S-Lindley distribution are shown in Figure 1.

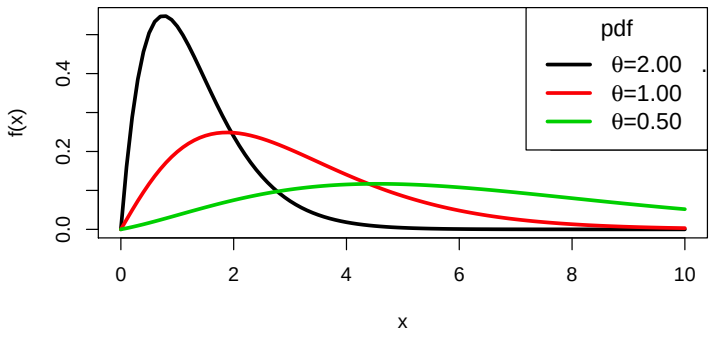


Figure 1: Plots of the pdf of the 2S-Lindley distribution for different values of θ.

2.2. Probability functions

First of all, after some algebraic manipulations, the cdf of the 2S-Lindley distribution is given by

$$F(x) = 1 - \frac{1}{6(1+\theta)^2}\left[\theta^3 x(x^2+6x+6) + 3\theta^2(x^2+4x+2) + 6\theta(x+2)+6\right]e^{-\theta x},$$
$$x > 0.$$

The corresponding survival function (sf) is given by

$$\begin{aligned} S(x) &= 1 - F(x) \\ &= \frac{1}{6(1+\theta)^2}\left[\theta^3 x(x^2+6x+6) + 3\theta^2(x^2+4x+2) + 6\theta(x+2)+6\right]e^{-\theta x}, \\ & x > 0. \end{aligned}$$

The corresponding hazard rate function (hrf) is given by

$$h(x) = \frac{f(x)}{S(x)} = \frac{\theta^4 x(x^2+6x+6)}{\theta^3 x(x^2+6x+6) + 3\theta^2(x^2+4x+2) + 6\theta(x+2)+6}, \quad x > 0.$$

Also, the corresponding cumulative hazard rate function is given by

$$\begin{aligned} \Omega(x) &= -\log[S(x)] \\ &= \log(6) + 2\log(1+\theta) + \theta x \\ &\quad - \log\left[\theta^3 x(x^2+6x+6) + 3\theta^2(x^2+4x+2) + 6\theta(x+2)+6\right], \quad x > 0. \end{aligned}$$

The corresponding quantile function (qf), say $Q(u)$, can be obtained by solving the following equation: $F(Q(u)) = Q(F(u))$, $u \in (0,1)$. It can not be presented analytically but can be determined numerically for a given θ. Also, shapes of the hrf of the 2S-Lindley distribution are shown in Figure 2.

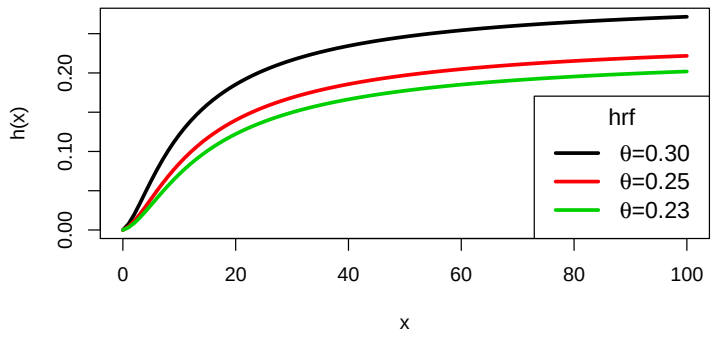


Figure 2: Plots of the hrf of the 2S-Lindley distribution for different values of θ.

2.3. Moments

The (ordinary) moments of the 2S-Lindley distribution are expressed in the following result.

Proposition 2.1. *Let $r \in \mathbb{N}$ and Z a random variable following the 2S-Lindley distribution with parameter θ. Then, the r-th moment of Z is given by*

$$\mu_r^* = E(Z^r) = \frac{1}{6\theta^r}\frac{1}{(1+\theta)^2}(r+1)!\left[6\theta^2 + 6\theta(r+2) + r^2 + 5r + 6\right].$$

Proof: Let us introduce the gamma function defined by $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$, $x > 0$. By using the pdf of Z given by (2.1), we have

$$\begin{aligned}
\mu_r^* = E(Z^r) &= \int_{-\infty}^{+\infty} x^r f(x)dx \\
&= \frac{\theta^4}{(1+\theta)^2}\left[\frac{1}{6}\int_0^{+\infty} x^{r+3}e^{-\theta x}dx + \int_0^{+\infty} x^{r+2}e^{-\theta x}dx + \int_0^{+\infty} x^{r+1}e^{-\theta x}dx\right] \\
&= \frac{\theta^4}{(1+\theta)^2}\left[\frac{1}{6}\frac{1}{\theta^{r+4}}\Gamma(r+4) + \frac{1}{\theta^{r+3}}\Gamma(r+3) + \frac{1}{\theta^{r+2}}\Gamma(r+2)\right] \\
&= \frac{1}{6\theta^r}\frac{1}{(1+\theta)^2}(r+1)!\left[6\theta^2 + 6\theta(r+2) + r^2 + 5r + 6\right].
\end{aligned}$$

This ends the proof of Proposition 2.1. □

An alternative proof of Proposition 2.1 using the Lindley distribution as baseline is given below. Let us recall that, for any $r \in \mathbb{N}$ and a random variable X following the Lindley distribution with parameter θ, the r-th moment of X is given by

$$\mu_r' = E(X^r) = \frac{r!(\theta + r + 1)}{\theta^r(1+\theta)}.$$

Therefore, by $Z = X + Y$ and the binomial formula, the r-th moment of Z is given by

$$\begin{aligned}
\mu_r^* = E((X+Y)^r) &= \sum_{k=0}^{r}\binom{r}{k}\mu_{r-k}'\mu_k' \\
&= \sum_{k=0}^{r}\binom{r}{k}\frac{(r-k)!(\theta + r - k + 1)}{\theta^{r-k}(1+\theta)}\frac{k!(\theta + k + 1)}{\theta^k(1+\theta)} \\
&= \frac{1}{\theta^r}\frac{1}{(1+\theta)^2}r!\sum_{k=0}^{r}(\theta + r - k + 1)(\theta + k + 1) \\
&= \frac{1}{6\theta^r}\frac{1}{(1+\theta)^2}(r+1)!\left[6\theta^2 + 6\theta(r+2) + r^2 + 5r + 6\right].
\end{aligned}$$

Also, owing to Proposition 2.1, we have

$$\mu_1^* = \frac{2(\theta+2)}{\theta(1+\theta)}, \quad \mu_2^* = \frac{6\theta(\theta+4)+20}{\theta^2(1+\theta)^2}, \quad \mu_3^* = \frac{24(\theta^2+5\theta+5)}{\theta^3(1+\theta)^2}$$

and

$$\mu_4^* = \frac{120(\theta^2+6\theta+7)}{\theta^4(1+\theta)^2}.$$

In particular, the mean of Z is given by $\mu = \mu_1^*$ and the variance of Z is given by

$$\sigma^2 = \mu_2^* - \mu^2 = \frac{2(\theta^2+4\theta+2)}{\theta^2(1+\theta)^2}.$$

Other important quantities can be defined via the moments as, for instance, the skewness and kurtosis coefficients of Z, respectively given by

$$\sqrt{\beta_1} = \frac{1}{\sigma^3}E\left[(Z-\mu)^3\right] = \frac{1}{\sigma^3}\sum_{k=0}^{3}\binom{3}{k}(-1)^{3-k}\mu_k^*\mu^{3-k}$$

and

$$\beta_2 = \frac{1}{\sigma^4}E\left[(Z-\mu)^4\right] = \frac{1}{\sigma^4}\sum_{k=0}^{4}\binom{4}{k}(-1)^{4-k}\mu_k^*\mu^{4-k}.$$

Table 1 indicates numerical values for the quantities above, i.e., μ_1^*, μ_2^*, μ_3^*, μ_4^*, σ^2, $\sqrt{\beta_1}$ and β_2, for selected values for θ.

Table 1: Numerical values of some measures of the 2S-Lindley distribution for selected values of parameter θ.

θ	μ_1^*	μ_2^*	μ_3^*	μ_4^*	σ^2	$\sqrt{\beta_1}$	β_2
0.002	1998.004	4992018	14970071832	5.23803e+13	999998	14.9442	4.5000
0.02	198.0392	49217.61	14707036	5132929642	9998.0787	14.4597	4.500
0.2	18.3333	434.7222	12583.33	429166.7	98.6111	10.8801	4.5420
0.1	38.1818	1856.198	109289.3	7547107	398.3471	12.6300	4.5124
1	3.0000	12.5000	66.0000	420.0000	3.5000	4.3525	4.8980
2	1.3333	2.5556	6.3333	19.1666	0.7778	0.4860	5.2347
5	0.4667	0.3222	0.2933	0.3305	0.1044	8.2294	5.6713
10	0.2182	0.0711	0.0307	0.0166	0.0235	22.0892	5.8688
20	0.1048	0.0164	0.0035	0.0009	0.0054	50.0478	5.9566
100	0.0202	0.0006	2.4715e-05	1.2477e-06	0.0002	275.9059	5.9978

2.4. Incomplete moments

A result on the incomplete moments for the 2S-Lindley distribution is given below.

Proposition 2.2. *Let r be a positive integer and Z a random variable following the 2S-Lindley distribution with parameter θ. Let us introduce the lower gamma function defined by $\gamma(x,y)=\int_0^y t^{x-1}e^{-t}dt$, $x>0$ and $y\geq 0$. Then, the r-th incomplete moment of Z is given by*

$$\mu_r^*(t)=E(Z^r\mathbf{1}_{\{Z\leq t\}})=\frac{1}{(1+\theta)^2\theta^r}\left[\frac{1}{6}\gamma(r+4,\theta t)+\theta\gamma(r+3,\theta t)+\theta^2\gamma(r+2,\theta t)\right].$$

Proof: By using the pdf of Z given by (2.1), we have

$$\begin{aligned}\mu_r^*(t)&=\int_{-\infty}^{t}x^r f(x)dx\\&=\frac{\theta^4}{(1+\theta)^2}\left[\frac{1}{6}\int_0^t x^{r+3}e^{-\theta x}dx+\int_0^t x^{r+2}e^{-\theta x}dx+\int_0^t x^{r+1}e^{-\theta x}dx\right]\\&=\frac{\theta^4}{(1+\theta)^2}\left[\frac{1}{6}\frac{1}{\theta^{r+4}}\gamma(r+4,\theta t)+\frac{1}{\theta^{r+3}}\gamma(r+3,\theta t)+\frac{1}{\theta^{r+2}}\gamma(r+2,\theta t)\right]\\&=\frac{1}{(1+\theta)^2\theta^r}\left[\frac{1}{6}\gamma(r+4,\theta t)+\theta\gamma(r+3,\theta t)+\theta^2\gamma(r+2,\theta t)\right].\end{aligned}$$

This ends the proof of Proposition 2.2. □

The incomplete mean given by $\mu_1^*(t)$ deserves a particular focus, because it allows to express several important quantities, as the mean deviation of Z about the mean given by $\delta_1=E(|Z-\mu|)=2\mu F(\mu)-2\mu_1^*(\mu)$, the mean residual life of Z given by $m_*(t)=E(Z-t\mid Z>t)=[1-\mu_1^*(t)]/[1-F(t)]-t$ and the mean waiting time of Z given by $M_*(t)=E(t-Z\mid Z<t)=t-\mu_1^*(t)/F(t)$, among others.

2.5. Characteristic function

The characteristic function of the 2S-Lindley distribution is provided in the following result.

Proposition 2.3. *Let Z be a random variable following the 2S-Lindley distribution with parameter θ. Then, the characteristic function of Z is given by*

$$\varphi(t)=\frac{\theta^4(\theta-it+1)^2}{(1+\theta)^2(\theta-it)^4},\quad t\in\mathbb{R}.$$

Proof: Let us recall that, for any $t\in\mathbb{R}$ and a random variable X following the Lindley distribution with parameter θ, the characteristic function of X is given by

$$\varphi_*(t)=E(e^{itX})=\frac{\theta^2(\theta-it+1)}{(1+\theta)(\theta-it)^2}.$$

Hence, using the representation $Z=X+Y$ with X and Y independent and identically distributed, the characteristic function for Z is given by

$$\varphi(t)=[\varphi_*(t)]^2=\frac{\theta^4(\theta-it+1)^2}{(1+\theta)^2(\theta-it)^4}.$$

This ends the proof of Proposition 2.3. □

2.6. Stochastic ordering

A result on stochastic ordering related to the 2S-Lindley distribution is now presented. Before that, some basics are recalled. Let X_1 and X_2 be two random variables having pdfs given by $f_1(x)$ and $f_2(x)$, respectively. Then, X_1 is said to be smaller than X_2 in the likelihood ratio order, denoted by $X_1 \leq_{lr} X_2$, if $f_1(x)/f_2(x)$ is decreasing in x. This property has important consequence in terms of distribution comparisons. We refer to [23] for the technical details.

Proposition 2.4. *Let X_1 be random variable following the 2S-Lindley distribution with parameter θ_1 and X_2 be a random variable following the 2S-Lindley distribution with parameter θ_2. Then, if $\theta_1 \geq \theta_2$, we have $X_1 \leq_{lr} X_2$.*

Proof: Let $f_1(x)$ and $f_2(x)$ be the pdfs of X_1 and X_2 given by (2.1) with $\theta = \theta_1$ and $\theta = \theta_2$, respectively. Then, for $x > 0$, we have

$$\frac{f_1(x)}{f_2(x)} = \frac{\theta_1^4(1+\theta_2)^2}{\theta_2^4(1+\theta_1)^2} e^{-(\theta_1-\theta_2)x},$$

which is clearly decreasing if $\theta_1 \geq \theta_2$, implying the desired result. This ends the proof of Proposition 2.4. □

2.7. Extreme order statistics

Let us consider a random sample $X_1, ..., X_n$ of size n from the 2S-Lindley distribution with parameter θ. Let $X_{1:n} = \min(X_1, ..., X_n)$ be the sample minima and $X_{n:n} = \max(X_1, ..., X_n)$ be the sample maxima. Then, we have the following limit results:

$$\lim_{t\to 0} \frac{F(xt)}{F(t)} =$$
$$\lim_{t\to 0} \frac{1 - \frac{1}{6(1+\theta)^2}\left[\theta^3 xt(x^2t^2+6xt+6) + 3\theta^2(x^2t^2+4xt+2) + 6\theta(xt+2)+6\right] e^{-\theta xt}}{1 - \frac{1}{6(1+\theta)^2}\left[\theta^3 t(t^2+6t+6) + 3\theta^2(t^2+4t+2) + 6\theta(t+2)+6\right] e^{-\theta t}}$$
$$= x.$$

Thus, [13, Theorem 1.6.2] ensures the existence of a_n and b_n such that

$$\lim_{n\to\infty} P(a_n(X_{1:n} - b_n) \leq x) = 1 - e^{-x}.$$

We recognize the cdf of the exponential distribution with parameter 1, showing that $a_n(X_{1:n} - b_n)$ can be approximated by this distribution.

Moreover, we have

$$\lim_{t\to+\infty} \frac{1 - F(x+t)}{1 - F(t)} =$$
$$\lim_{t\to+\infty} \frac{\theta^3(x+t)((x+t)^2+6(x+t)+6) + 3\theta^2((x+t)^2+4(x+t)+2) + 6\theta((x+t)+2)+6}{\theta^3 t(t^2+6t+6) + 3\theta^2(t^2+4t+2) + 6\theta(t+2)+6} e^{-\theta x}$$
$$= e^{-\theta x}.$$

Thus, [13, Theorem 1.6.2] ensures the existence of a_n and b_n such that

$$\lim_{n\to\infty} P(a_n(X_{n:n} - b_n) \leq x) = \exp\left(-e^{-\theta x}\right).$$

We recognize the cdf of the Gumbel distribution with parameters 1 and $1/\theta$, showing that $a_n(X_{n:n} - b_n)$ can be approximated by this distribution.

The form of the norming constants can also be determined using [13, Corollary 1.6.3].

2.8. Maximum likelihood estimator

Let $x_1, ..., x_n$ be n observations of a random variable Z following the 2S-Lindley distribution with parameter θ. Then, the likelihood and log-likelihood functions are, respectively, defined by

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \frac{\theta^{4n}}{(1+\theta)^{2n}} \left[\prod_{i=1}^{n} x_i\right] \left[\prod_{i=1}^{n} \left(\frac{x_i^2}{6} + x_i + 1\right)\right] e^{-\theta \sum_{i=1}^{n} x_i}$$

and

$$\ell(\theta) = \log[L(\theta)] = 4n\log(\theta) - 2n\log(1+\theta) - \theta\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \log(x_i) + \sum_{i=1}^{n} \log\left(\frac{x_i^2}{6} + x_i + 1\right).$$

The maximum likelihood estimator (MLE) of θ, denoted by $\widehat{\theta}$, is defined by the θ maximizing $L(\theta)$ or $\ell(\theta)$. Thus, it can be obtained by solving the following equation: $\partial\ell(\theta)/\partial\theta = 0$, i.e.,

$$\frac{4n}{\theta} - \frac{2n}{1+\theta} - \sum_{i=1}^{n} x_i = 0.$$

After some algebra, we have

$$\widehat{\theta} = \frac{-\sum_{i=1}^{n} x_i + \sqrt{\left(2n - \sum_{i=1}^{n} x_i\right)^2 + 16n\sum_{i=1}^{n} x_i} + 2n}{2\sum_{i=1}^{n} x_i}.$$

Hence, $\widehat{\theta}$ has a simple expression. As any MLE, it enjoys desirable properties of convergence, guarantied by the well-established theory of the maximum likelihood method.

2.9. Simulation study

In this section, we present some simulation results to examine the finite sample behavior of the MLE proposed in previous section in the case of the 2S-Lindley distribution.

The simulation study is repeated for $N = 1000$ iterations each with sample size $n = 25$, 50, 150 and 300 from the 2S-Lindley distribution. The 2S-Lindley random number generation was performed using the sum of *rlindley()* function from **LindleyR** package [17] and the parameters are estimated by using the method of MLE by using the package nlm in R. The evaluation of the assessment is based on two quantities such as the bias and the mean squared errors (MSE), as follows:

1) bias of the simulated N estimates of R:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{R}_i - R),$$

2) mean square error of the simulated N estimates of R:

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{R}_i - R)^2,$$

where R is the true value of parameters θ. The results of our simulation study are summarized in Table 2. Based on the table, notice that the MLEs are close to the true parameter values for the current sample sizes, which means that the maximum likelihood method can be used effectively for estimating θ. Also, we can see that the bias and MSE of the MLEs converge to zero when the sample size is increased, as expected.

Table 2: Bias and MSE of $\hat{\theta}$ for the 2S-Lindley distribution.

	$\theta = 0.3$		$\theta = 0.5$		$\theta = 1.0$		$\theta = 1.2$	
n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.0031	0.0009	0.0055	0.0027	0.0166	0.0124	0.0145	0.0178
50	0.0023	0.0004	0.0032	0.0013	0.0075	0.0055	0.0077	0.0079
100	0.0011	0.0002	0.0010	0.0007	0.0053	0.0029	0.0027	0.0041
200	6.7e-05	0.0001	0.0006	0.0003	0.0034	0.0015	0.0027	0.0022
300	4.6e-05	7.7e-05	5.1e-05	0.0002	0.0014	0.0010	0.0020	0.0014

2.10. Applications

Here, we use four data sets to illustrate the power of the proposed 2S-Lindley distribution. We compare the proposed distribution with the Lindley and exponential distributions. The first real data set corresponds to arose in tests on endurance of deep groove ball bearings from [12] on the number of million revolutions before failure for each of the 23 ball bearings in the life tests. The data are given below:

17.88 28.92 33.00 41.52 42.12 45.60 48.80 51.84 51.96 54.12 55.56 67.80 68.44 68.64 68.88 84.12 93.12 98.64 105.12 105.84 127.92 128.04 173.40 .

The second real data set is from [28]. It represents the strength of 1.5cm glass fibers measured at the National Physical Laboratory, England. The data are given below:

0.55 0.93 1.25 1.36 1.49 1.52 1.58 1.61 1.64 1.68 1.73 1.81 2.00 0.74 1.04 1.27 1.39 1.49 1.53 1.59 1.61 1.66 1.68 1.76 1.82 2.01 0.77 1.11 1.28 1.42 1.50 1.54 1.60 1.62 1.66 1.69 1.76 1.84 2.24 0.81 1.13 1.29 1.48 1.50 1.55 1.61 1.62 1.66 1.70 1.77 1.84 0.84 1.24 1.30 1.48 1.51 1.55 1.61 1.63 1.67 1.70 1.78 1.89

The third real data set is reported by [7]. It demonstrates the lifetime's data relating to relief times (in minutes) of 20 patients receiving an analgesic. The data are given below:

1.1 1.4 1.3 1.7 1.9 1.8 1.6 2.2 1.7 2.7 4.1 1.8 1.5 1.2 1.4 3 1.7 2.3 1.6 2

The fourth data set is taken from [4]. it gives the strength data of glass of the aircraft window. The data are given below:

18.83 20.8 21.657 23.03 23.23 24.05 24.321 25.5 25.52 25.8 26.69 26.77 26.78 27.05 27.67 29.9 31.11 33.2 33.73 33.76 33.89 34.76 35.75 35.91 36.98 37.08 37.09 39.58 44.045 45.29 45.381.

For comparing the goodness of fit of the models, we found the unknown parameters (by the maximum likelihood method), standard error (SE), −log likelihood (−logL), AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), corrected Akaike Information Criterion (AICc) and Kolmogorov-Smirnov (K-S) statistic, given by

$$-\text{LogL} = -\log(L), \quad \text{AIC} = -2\text{LogL} + 2k, \quad \text{BIC} = -2\text{LogL} + k\log(n),$$

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1},$$

and

$$\text{K-S} = \max\{|F(x_i) - \hat{F}(x_i)|, |F(x_i) - \hat{F}(x_{i-1})|\},$$

where L is the maximum value of the corresponding likelihood function, k is the number of parameters, n is the sample size, $F(x_i)$ denote the value of the cdf of the candidate distribution at x_i and $\hat{F}(x_i)$ denote the value of the empirical distribution function at x_i.

Table 3, Table 4 Table 5 and Table 6 summarize the results of the fitted 2S-Lindley, Lindley and exponential distributions for the four considered data sets.

Table 3: Estimated values, −logL, AIC, BIC, AICc and K-S statistics for the first data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc	K-S
2S-Lindley	0.0570 (0.0028)	113.0799	228.1598	229.2953	228.2254	0.10606
Lindley	0.0273 (0.0040)	115.7356	233.4713	234.6068	233.66	0.19299
exponential	0.0138 (0.0029)	121.4365	244.8731	246.0086	245.06	0.30677

Table 4: Estimated values, −logL, AIC, BIC, AICc and K-S statistics for the second data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc	K-S
2S-Lindley	1.8011 (0.1274)	62.2742	126.5484	128.6916	126.614	0.32852
Lindley	0.9961 (0.0948)	81.27844	164.5569	166.7	164.6225	0.38643
exponential	0.6636 (0.0836)	88.83032	179.6606	181.8038	179.7262	0.418

Table 5: Estimated values, −logL, AIC, BIC, AICc and K-S statistics for the third data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc	K-S
2S-Lindley	1.4775 (0.1822)	24.8511	51.70225	52.69799	51.92447	0.29271
Lindley	0.8161 (0.1361)	30.24955	62.4991	63.49483	62.72132	0.43951
exponential	0.5263 (0.1179)	32.83708	67.67416	68.66989	67.89638	0.43951

Table 6: Estimated values, −logL, AIC, BIC, AICc and K-S statistics for the fourth data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc	K-S
2S-Lindley	0.1227 (0.011)	117.8023	237.6046	239.0386	237.7425	0.26915
Lindley	0.0630 (0.008)	126.9942	255.9884	257.4224	256.1263	0.36548
exponential	0.0324 (0.0058)	137.2644	276.5289	277.9629	276.6668	0.4586

From these tables, it is obvious that the smallest −logL, AIC, BIC, AICc and K-S statistic are acquired for the 2S-Lindley distribution. In summary, we can conclude that the 2S-Lindley model can be adequate for modeling these data.

3. ON THE DIFFERENCE OF TWO INDEPENDENT LINDLEY DISTRIBUTION

This section now focuses on the properties of the difference of two independent random variables following the Lindley distribution with the same parameter.

3.1. Definition

We now consider the pdf given by

$$f(x) = \frac{\theta}{4(1+\theta)^2}\left[\theta(2\theta+1)|x| + 2\theta^2 + 2\theta + 1\right] e^{-\theta|x|}, \quad x \in \mathbb{R}, \theta > 0. \tag{3.1}$$

The feature of this pdf is described in the result below.

Proposition 3.1. *Let X and Y be two independent random variables both following the Lindley distribution with parameter θ. Then, the random variable $Z = X - Y$ has the pdf given by (3.1).*

Proof: First of all, since the support of X and Y is $(0, +\infty)$, the support of Z is $\mathbb{R}$. Now, let us notice that the cdf and pdf of $-Y$ are, respectively, given by

$$F_{**}(x) = \left[1 - \frac{\theta}{1+\theta}x\right] e^{\theta x}, \quad f_{**}(x) = \frac{\theta^2}{1+\theta}(1-x)e^{\theta x}, \quad x < 0.$$

Since X and $-Y$ are independent, the pdf of Z is given by the convolution product:

$$\begin{aligned}
f(x) &= (f_* \star f_{**})(x) = \int_{-\infty}^{+\infty} f_*(x-t) f_{**}(t)dt \\
&= \int_{-\infty}^{\inf(x,0)} \frac{\theta^2}{1+\theta}[1+(x-t)]e^{-\theta(x-t)} \frac{\theta^2}{1+\theta}(1-t)e^{\theta t}dt \\
&= \frac{\theta^4}{(1+\theta)^2} e^{-\theta x}\left\{\int_{-\infty}^{\inf(x,0)} (1-t)^2 e^{2\theta t}dt + x\int_{-\infty}^{\inf(x,0)} (1-t)e^{2\theta t}dt\right\} \\
&= \frac{\theta}{4(1+\theta)^2} e^{-\theta[x-2\inf(x,0)]}\left[2\theta^2 \inf(x,0)^2 - 2\theta^2 \inf(x,0)x - 4\theta^2 \inf(x,0) + 2\theta^2 x\right. \\
&\quad \left. + \; 2\theta^2 - 2\theta \inf(x,0) + \theta x + 2\theta + 1\right].
\end{aligned}$$

When $x \geq 0$, we have $\inf(x,0) = 0$ implying that

$$f(x) = \frac{\theta}{4(1+\theta)^2}\left[\theta(2\theta+1)x + 2\theta^2 + 2\theta + 1\right] e^{-\theta x}.$$

When $x < 0$, we have $\inf(x,0) = x$, implying that

$$f(x) = \frac{\theta}{4(1+\theta)^2}\left[-\theta(2\theta+1)x + 2\theta^2 + 2\theta + 1\right] e^{\theta x}.$$

By putting the above results together, we obtain the desired result. This ends the proof of Proposition 3.1. □

For the purpose of this study, the corresponding distribution is called the 2D-Lindley distribution (2D for Difference of 2 random variables). To the best of our knowledge, there is no work on the theoretical and practical aspect of this distribution, which motivates a part of this study. Figure 3 shows the behavior the pdf of the 2D-Lindley distribution for selected values of parameter θ.

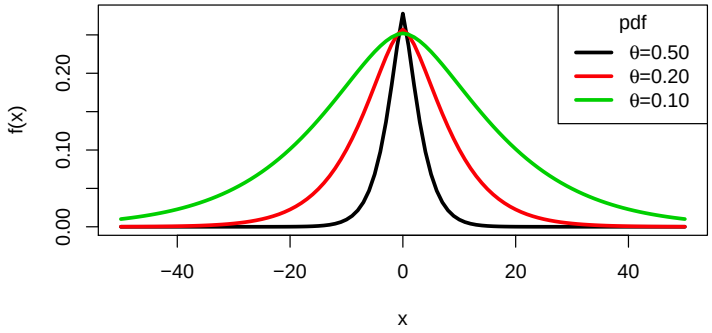


Figure 3: The pdf of the 2D-Lindley distribution for different values of θ.

3.2. Probability functions

The cdf of the 2D-Lindley distribution is presented in the proposition below.

Proposition 3.2. *The cdf of the 2D-Lindley distribution is given by*

$$F(x) = \begin{cases} \dfrac{1}{4(1+\theta)^2}\left[-\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{\theta x} & \text{if } x < 0, \\ 1 - \dfrac{1}{4(1+\theta)^2}\left[\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{-\theta x} & \text{if } x \geq 0. \end{cases} \tag{3.2}$$

Proof: For $x < 0$, by using (3.1), we have

$$\begin{aligned} F(x) = P(Z \leq x) &= \int_{-\infty}^{x} f(t)dt \\ &= \frac{\theta}{4(1+\theta)^2}\left[-\theta(2\theta+1)\int_{-\infty}^{x} te^{\theta t}dt + (2\theta^2+2\theta+1)\int_{-\infty}^{x} e^{\theta t}dt\right] \\ &= \frac{1}{4(1+\theta)^2}\left[-\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{\theta x}. \end{aligned}$$

Since the distribution of Z is symmetric around 0, for $x \geq 0$, we have

$$F(x) = 1 - F(-x) = 1 - \frac{1}{4(1+\theta)^2}\left[\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{-\theta x}.$$

We obtain the desired result by putting the above equalities together. This completes the proof of Proposition 3.2. □

By using Proposition 3.2, the corresponding survival function is given by

$$S(x) = \begin{cases} 1 - \dfrac{1}{4(1+\theta)^2}\left[-\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{\theta x} & \text{if } x < 0, \\ \dfrac{1}{4(1+\theta)^2}\left[\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{-\theta x} & \text{if } x \geq 0. \end{cases}$$

The corresponding hrf is given by

$$h(x) = \begin{cases} \dfrac{\theta\left[\theta(2\theta+1)|x| + 2\theta^2 + 2\theta + 1\right]}{4(1+\theta)^2 e^{-\theta x} + \theta(2\theta+1)x - 2(1+\theta)^2} & \text{if } x < 0, \\ \dfrac{\theta\left[\theta(2\theta+1)x + 2\theta^2 + 2\theta + 1\right]}{\theta(2\theta+1)x + 2(1+\theta)^2} & \text{if } x \geq 0. \end{cases}$$

Also, the corresponding chrf is given by

$$\Omega(x) = \begin{cases} -\log\left[1 - \dfrac{1}{4(1+\theta)^2}\left[-\theta(2\theta+1)x + 2(1+\theta)^2\right] e^{\theta x}\right] & \text{if } x < 0, \\ \log(4) + 2\log(1+\theta) + \theta x - \log\left[\theta(2\theta+1)x + 2(1+\theta)^2\right] & \text{if } x \geq 0. \end{cases}$$

The corresponding qf, say $Q(u)$, can be obtained by solving the following equation: $F(Q(u)) = Q(F(u))$, $u \in (0,1)$. It can not be presented analytically but can be determined numerically for a given θ. Further, Figure 4 depicts the behavior the hrf of the 2D-Lindley distribution for selected values of parameter θ.

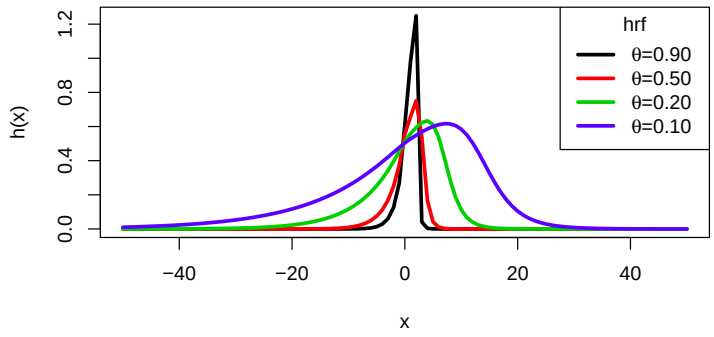


Figure 4: The hrf of the 2D-Lindley distribution for different values of θ.

3.3. Mixture

The 2D-Lindley distribution can be viewed as a particular mixture of distributions, as described below.

Proposition 3.3. *Let U, V and W be three random variables following the Laplace distribution with parameter θ and A a random variable following the Bernoulli distribution with parameter $\theta^2/(1+\theta)^2$, all these random variables are independent. Let Z be a random variable following the 2D-Lindley distribution with parameter θ. Then, we have the following stochastic representation:*

$$Z \overset{(d)}{=} AU + (1-A)(V+W).$$

Proof: It is enough to remark that we can write $f(x)$ given by (3.1) as

$$\begin{aligned} f(x) &= \frac{\theta^2}{(1+\theta)^2}\left[\frac{\theta}{2}e^{-\theta|x|}\right] + \frac{1+2\theta}{(1+\theta)^2}\left[\frac{\theta}{4}(1+\theta|x|)e^{-\theta|x|}\right] \\ &= pf_1(x) + (1-p)f_2(x), \end{aligned}$$

where

$$p = \frac{\theta^2}{(1+\theta)^2}, \quad f_1(x) = \frac{\theta}{2}e^{-\theta|x|}, \quad f_2(x) = \frac{\theta}{4}(1+\theta|x|)e^{-\theta|x|}.$$

One can notice that $f_1(x)$ is the pdf of the Laplace distribution with parameter θ and $f_2(x)$ is the pdf of the sum of two independent random variables both following the Laplace distribution with parameter θ as common distribution, see, [9, Section 2.3]. This ends the proof of Proposition 3.3. □

3.4. Moments

The moments of the 2D-Lindley distribution are described below.

Proposition 3.4. *Let $r \in \mathbb{N}$ and Z a random variable following the 2D-Lindley distribution with parameter θ. Then, the r-th moment of Z is given by*

$$\mu_r^* = E(Z^r) = [1+(-1)^r]\frac{1}{2\theta^r}r!\left[1+\frac{1+2\theta}{2(1+\theta)^2}r\right].$$

Proof: Since the distribution of Z is symmetric around 0 and the integral is well defined, for any $m \in \mathbb{N}$, we have $\mu_{2m+1}^* = 0$. By the use of the gamma function, for any $m \in \mathbb{N}$, we have

$$\begin{aligned} \mu_{2m}^* = E(Z^{2m}) &= \int_{-\infty}^{+\infty} x^{2m} f(x)dx \\ &= \frac{\theta^2}{(1+\theta)^2}\int_{-\infty}^{+\infty} x^{2m}\frac{\theta}{2}e^{-\theta|x|}dx + \frac{1+2\theta}{(1+\theta)^2}\int_{-\infty}^{+\infty} x^{2m}\frac{\theta}{4}(1+\theta|x|)e^{-\theta|x|}dx \\ &= \frac{\theta^2}{(1+\theta)^2}\frac{1}{\theta^{2m}}\Gamma(2m+1) + \frac{1+2\theta}{(1+\theta)^2}\frac{1}{2}\frac{1}{\theta^{2m}}\left[\Gamma(2m+1)+\Gamma(2m+2)\right] \\ &= \frac{1}{\theta^{2m}}(2m)!\left[1+\frac{1+2\theta}{(1+\theta)^2}m\right]. \end{aligned}$$

By distinguishing the odd and even integer, we prove the desired result, ending the proof of Proposition 3.4. □

Owing to Proposition 3.4, we have

$$\mu_1^* = 0, \quad \mu_2^* = \frac{2(\theta^2+4\theta+2)}{\theta^2(1+\theta)^2}, \quad \mu_3^* = 0, \quad \mu_4^* = \frac{24[\theta(\theta+6)+3]}{\theta^4(1+\theta)^2}.$$

In particular, the mean of Z is given by $\mu = 0$ and the variance of Z is given by

$$\sigma^2 = \mu_2^* = \frac{2(\theta^2+4\theta+2)}{\theta^2(1+\theta)^2}.$$

Without surprise, the variance of the 2S and 2D Lindley distributions are the same.

The skewness of Z is equal to 0 and the kurtosis of Z is given by

$$\beta_2 = \frac{1}{\sigma^4} E\left[(Z-\mu)^4\right] = \frac{6(\theta^2+6\theta+3)(1+\theta)^2}{(\theta^2+4\theta+2)^2}.$$

Table 7 indicates numerical values for the quantities above, that is, μ_2^*, μ_4^*, σ^2 and β_2, for selected values for θ.

Table 7: Numerical values of some measures of the 2D-Lindley distribution for selected values of parameter θ.

θ	μ_2^*	μ_4^*	σ^2	β_2
0.02	9998.078	449884660	9998.078	4.5006
0.01	39998.04	7199529458	39998.04	4.5001
0.1	398.3471	716033.1	398.3471	4.5124
1	3.500	60.00	3.5000	4.8980
2	0.7778	3.1667	0.7778	5.2347
5	0.1044	0.0619	0.1044	5.6713
10	0.0235	0.0032	0.0235	5.8688
20	0.0055	0.0002	0.0055	5.9565
100	0.0002	2.494e-07	0.0002	5.9978

3.5. Characteristic function

The characteristic function of the 2D-Lindley distribution is presented below.

Proposition 3.5. *Let Z be a random variable following the 2D-Lindley distribution with parameter θ. Then, the characteristic function of Z is given by*

$$\varphi(t) = \frac{\theta^4[(1+\theta)^2+t^2]}{(1+\theta)^2(\theta^2+t^2)^2}, \quad t \in \mathbb{R}.$$

Proof: Let us recall that, for any $t \in \mathbb{R}$ and a random variable X following the Lindley distribution with parameter θ, the characteristic function of X is given by

$$\varphi_*(t) = E(e^{itX}) = \frac{\theta^2(\theta - it + 1)}{(1+\theta)(\theta - it)^2}.$$

Hence, using the representation $Z = X - Y$ with X and Y independent and identically distributed, the characteristic function for Z is given by

$$\varphi(t) = \varphi_*(t)\varphi_*(-t) = \frac{\theta^2(\theta - it + 1)}{(1+\theta)(\theta - it)^2} \times \frac{\theta^2(\theta + it + 1)}{(1+\theta)(\theta + it)^2} = \frac{\theta^4[(1+\theta)^2+t^2]}{(1+\theta)^2(\theta^2+t^2)^2}.$$

This ends the proof of Proposition 3.5. □

Let us mention that can prove Proposition 3.3 by using the characteristic function. It is enough to observe that we can write $\varphi(t)$ as

$$\varphi(t) = \frac{\theta^2}{(1+\theta)^2}\frac{\theta^2}{\theta^2+t^2} + \left(1 - \frac{\theta^2}{(1+\theta)^2}\right)\left[\frac{\theta^2}{\theta^2+t^2}\right]^2,$$

which is exactly the characteristic function of $AU + (1-A)(V+W)$, implying the desired result.

3.6. Maximum likelihood estimator

Let $x_1, ..., x_n$ be n observations of a random variable Z following the 2D-Lindley distribution with parameter θ. Then, the likelihood and log-likelihood functions are, respectively, defined by

$$L(\theta) = \prod_{i=1}^{n} f(x_i) = \frac{\theta^n}{4^n(1+\theta)^{2n}}\left\{\prod_{i=1}^{n}\left[\theta(2\theta+1)|x_i| + 2\theta^2 + 2\theta + 1\right]\right\} e^{-\theta\sum_{i=1}^{n}|x_i|}$$

and

$$\begin{aligned}\ell(\theta) &= \log[L(\theta)] \\ &= n\log(\theta) - n\log(4) - 2n\log(1+\theta) - \theta\sum_{i=1}^{n}|x_i| \\ &+ \sum_{i=1}^{n}\log\left[\theta(2\theta+1)|x_i| + 2\theta^2 + 2\theta + 1\right].\end{aligned}$$

The MLE of θ can be obtained by solving the following equation: $\partial\ell(\theta)/\partial\theta = 0$, i.e.,

$$\frac{n}{\theta} - \frac{2n}{1+\theta} - \sum_{i=1}^{n}|x_i| + \sum_{i=1}^{n}\frac{(4\theta+1)|x_i| + 4\theta + 2}{\theta(2\theta+1)|x_i| + 2\theta^2 + 2\theta + 1} = 0.$$

This equation can not be solved analytically. However, some numerical algorithm allows to approach the solution in a precise way.

3.7. Simulation study

In this section, the simulation study is repeated for $N = 1000$ iterations from the 2D-Lindley distribution. For each replication, a random sample of size $n = 25$, 50, 100, 200 and 300 is drawn from the 2D-Lindley distribution. The 2D-Lindley random number generation was performed using difference of *rlindley()* function from **LindleyR** package [17] and the parameters are estimated by using the method of MLE by using the package nlm in R. The initial values of parameter are $\theta = 0.3, 0.5, 1.0$ and 1.5. The bias and MSE are presented in Table 8. From the table, we can observe that the bias and MSE of the MLEs converge to zero when the sample size is increased. This shows that the estimates are precise and accurate, hence, consistent and (asymptotically) unbiased.

Table 8: Bias and MSE of $\hat{\theta}$ for the 2D-Lindley distribution.

	$\theta = 0.3$		$\theta = 0.5$		$\theta = 1$		$\theta = 1.2$	
n	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
25	0.0066	0.0029	0.0055	0.0107	0.0288	0.0291	0.0307	0.0358
50	0.0048	0.0012	0.0090	0.0042	0.0133	0.0154	0.0154	0.0156
100	0.0020	0.0007	0.0027	0.0020	0.0057	0.0070	0.0074	0.0077
200	0.0012	0.0003	0.0010	0.0010	0.0017	0.0037	0.0040	0.0040
300	0.0008	0.0002	0.0007	0.0006	0.0016	0.0024	0.0022	0.0022

3.8. Applications

In this section, we analyze three data sets in order to illustrate the good performance of the 2D-Lindley distribution to compare with the Laplace and normal distributions, both with parameters standardly denoted by μ and σ. Here, we consider an extended form of the 2D-Lindley distribution by adding the location parameter μ in the pdf of the 2D-Lindley distribution. Thus, the related pdf is given by

$$f(x) = \frac{\theta}{4(1+\theta)^2}\left[\theta(2\theta+1)|x-\mu| + 2\theta^2 + 2\theta + 1\right] e^{-\theta|x-\mu|} \quad x, \mu \in \mathbb{R}, \theta > 0.$$

3.8.1. Comparison with the Laplace distribution

The first two data sets correspond to the age of the propellant and the tensile strength of kraft paper, respectively, reported in [21]. The data of the first set are given below:

15.5 23.75 8.0 17.0 5.5 19.0 24.0 2.5 7.5 11.0 13.0 3.75 25.0 9.75 22.0 18.0 6.0 12.5 2.0 21.5

The data of the second set are given below:

6.3 11.1 20.0 24.0 26.1 30.0 33.8 34.0 38.1 39.9 42.0 46.1 53.1 52.0 52.5 48.0 42.8 27.8 21.9

The third data set representing lung cancer rates data for 44 US states is given by www.calvin.edu/stob/data/cigs.csv. The data are given below:

17.05 19.8 15.98 22.07 22.83 24.55 27.27 23.57 13.58 22.8 20.3 16.59 16.84 17.71 25.45 20.94 26.48 22.04 22.72 14.2 15.6 20.98 19.5 16.7 23.03 25.95 14.59 25.02 12.12 21.89 19.45 12.11 23.68 17.45 14.11 17.6 20.74 12.01 21.22 20.34 20.55 15.53 15.92 25.88.

Table 9, Table 10 and Table 11 list the values of estimate, −logL, AIC, BIC and AICc, for the considered data sets.

Table 9: Estimated values, −logL, AIC, BIC and AICc for the first data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc
2D-Lindley	$\hat{\theta}$ **= 0.2335 (0.0447)**	**70.13335**	**144.2667**	**146.2582**	**144.9726**
	$\hat{\mu}$ **= 13.0205 (2.0049)**				
Laplace	$\hat{\mu}$ = 12.8350 (0.60)	71.33741	146.6748	148.6663	147.3807
	$\hat{\sigma}$ = 6.512916 (0.002)				

Table 10: Estimated values, −logL, AIC, BIC and AICc for the second data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc
2D-Lindley	$\hat{\theta}$ **= 0.1355 (0.0268)**	**77.22743**	**146.6748**	**148.6663**	**147.3807**
	$\hat{\mu}$ **= 34.7542 (3.4269)**				
Laplace	$\hat{\mu}$ = 34.00 (0.1062)	78.13456	160.2691	162.158	161.0191
	$\hat{\sigma}$ = 11.2337 (2.5791)				

Table 11: Estimated values, −logL, AIC, BIC and AICc for the third data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc
2D-Lindley	$\hat{\theta}$ **= 0.4182 (0.0536)**	**128.1709**	**260.3419**	**263.9103**	**260.6496**
	$\hat{\mu}$ **= 19.9190 (0.7894)**				
Laplace	$\hat{\mu}$ =20.3182 (0.27)	129.9786	263.9572	267.5255	264.2649
	$\hat{\sigma}$ = 3.5289 (0.03)				

From the tables, it may be noticed that the proposed 2D-Lindley model present the smallest values of the −logL, AIC, BIC and AICc and hence should be chosen as the best model for these datasets.

3.8.2. Comparison with the Normal distribution

Here we consider the data set artificially created from the standard Laplace distribution (with parameters 0 and 1) and truncated at the second decimal places which has been studied by [11]. The fourth data are given below:

-1.28 0.36 -1.29 -0.80 0.28 -0.06 -1.53 0.28 -0.54 0.17 0.59 6.22 2.41 0.33 -1.51 0.25 2.33 2.81 -0.92 2.12 -1.01 1.35 -0.37 -0.39 -4.39 -2.39 0.97 -0.58 -2.24 -0.05.

Table 12 shows the values of estimate, −logL, AIC, BIC and AICc, for the data set above.

Table 12: Estimated values, −logL, AIC, BIC and AICc for the fourth data set.

Distribution	Estimates (SE)	−logL	AIC	BIC	AICc
2D-Lindley	$\hat{\theta}$ **=1.0299 (0.1593)**	**59.4520**	**122.9040**	**125.7064**	**123.3484**
	$\hat{\mu}$ **= 0.0295 (0.4476)**				
Normal	$\hat{\mu}$ = 0.1228 (0.3445)	61.61703	127.2341	130.0365	127.6785
	$\hat{\sigma}$ = 1.8870 (0.2436)				

From the Table 12, we can see that the 2D-Lindley model present the smallest values of the −logL, AIC, BIC and AICc, which confirm the suitability behavior of the 2D-Lindley distribution.

4. CONCLUDING REMARKS

In this paper, we have derived single representations for the exact distribution of the sum and difference of independent Lindley random variables. We referred to the distributions of sum and difference of two independent Lindley random variables as the 2S-Lindley and 2D-Lindley distributions, respectively. Statistical properties such as moments, incomplete moments, characteristic function, stochastic ordering and extreme order statistics of the 2S-Lindley distribution have been provided. At the same time, a comprehensive study of statistical properties of the 2D-Lindley distribution also has been discussed. The model parameters are estimated by maximum likelihood method for both cases. From simulation studies, the performance of the maximum likelihood estimators has been assessed. The new models provide consistently better fit than some classical models available in the literature. In conclusion, proposed model with their attracting properties should have a promising future in distribution theory.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful comments and suggestions on the manuscript.

REFERENCES

[1] Al-Mutairi, D.K.; Ghitany, M.E. and Kundu, D. (2013). Inferences on stress-strength reliability from Lindley distributions, *Communications in Statistics – Theory and Methods*, **42**(8), 1443–1463.

[2] Chesneau, C.; Tomy, L. and Gillariose, J. (2019). *A new modified Lindley distribution with properties and applications*, preprint.

[3] Deniz, E. and Ojeda, E. (2011). The discrete Lindley distribution: properties and application, *Journal of Statistical Computation and Simulation*, **81**(11), 1405–1416.

[4] Fuller, E.J.; Frieman, S. and Quinn, J. (1994). Fracture mechanics approach to the design of glass aircraft windows: a case study, *SPIE Proceedings*, 419–430.

[5] Ghitany, M.E.; Al-Mutairi, D.K.; Balakrishnan, N. and Al-Enezi, L.J. (2013). Power lindley distribution and associated inference, *Computational Statistics and Data Analysis*, **64**, 20–33.

[6] Ghitany, M.E.; Atieh, B. and Nadarajah, S. (2008). Lindley distribution and its applications, *Mathematical Computation and Simulation*, **78**(4), 493–506.

[7] Gross, A.J. and Clark, V.A. (1975). *Survival Distributions: Reliability Applications in the Biometrical Sciences*, John Wiley, New York, USA.

[8] Hassan, M.K. (2014). On the convolution of Lindley distribution, *Columbia International Publishing Contemporary Mathematics and Statistics*, **2**(1), 47–54.

[9] Kotz, S.; Kozubowski, T.J. and Podgorski, K. (2001). *The Laplace Distribution and Generalizations. A Revisit with Applications to Communications, Economics, Engineering and Finance*, Birkhäuser, Boston.

[10] Krishna, H. and Kumar, K. (2011). Reliability estimation in Lindley distribution with progressively type II right censored sample, *Mathematics and Computers in Simulation*, **82**(2), 281–294.

[11] Kundu, D. (2004). *Discriminating between the Normal and the Laplace distributions*, Report, Department of Mathematics, Indian Institute of Technology Kanpur, India.

[12] Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York, USA.

[13] Leadbetter, M.R.; Lindgren, G. and Rootzén H. (1987). *Extremes and Related Properties of Random Sequences and Processes*, New York, Springer Verlag.

[14] Lindley, D.V. (1958). Fiducial distributions and Bayes theorem, *Journal of the Royal Statistical Society, A*, **20**, 102–107.

[15] Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part II: Inference*, Cambridge University Press, New York.

[16] Mahmoudi, E. and Zakerzadeh, H. (2010). Generalized Poisson-Lindley distribution, *Communications in Statistics – Theory and Methods*, **39**(10), 1785–1798.

[17] Mazucheli, J.; Fernandes, L.B. and de Oliveira, R.P. (2016). *LindleyR: The Lindley Distribution and Its Modifications*, R package version 1.1.0, https://CRAN.R-project.org/package=LindleyR.

[18] Merovci, F. (2013). Transmuted Lindley distribution, *International Journal of Open Problems in Computer Science and Mathematics*, **6**, 63–72.

[19] Merovci, F. and Elbatal, I. (2014). Transmuted Lindley-geometric distribution and its applications, *Journal of Statistics Applications and Probability*, **3**, 77–91.

[20] Merovci, F. and Sharma, V.K. (2014). The beta-Lindley distribution: properties and applications, *Journal of Applied Mathematics*, **2014**, 1–10.

[21] Montgomery, D.C.; Peck, E.A. and Vining, G.G. (2015). *Introduction to Linear Regression Analysis*, John Wiley and Sons.

[22] Nadarajah, S.; Bakouch, H.S. and Tahmasbi, R. (2011). A generalized Lindley distributionn, *Sankhya B – Applied and Interdisciplinary Statistics*, **73**, 331–359.

[23] Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*, Wiley, New York.

[24] Shanker, R.; Hagos, F. and Sujatha, S. (2015). On modeling of Lifetimes data using exponential and Lindley distributions, *Biometrics and Biostatistics International Journal*, **2**(5), 1–9.

[25] Shanker, R. and Mishra, A. (2013a). A two-parameter Lindley distribution, *Statistics in Transition New Series*, **14**(1), 45–56.

[26] Shanker, R. and Mishra, A. (2013b). A quasi Lindley distribution, *African Journal of Mathematics and Computer Science Research*, **6**(4), 64–71.

[27] Sharma, V.K.; Singh, S.; Singh, U. and Agiwal, V. (2015). The inverse Lindley distribution: a stress-strength reliability model with applications to head and neck cancer data, *Journal of Industrial and Production Engineering*, **32**(3), 162–173.

[28] Smith, R.L. and Naylor, J.C. (1987). A comparison of Maximum likelihood and Bayesian estimators for the three parameter Weibull distribution, *Journal of the Royal Statistical Society*, **36**, 358–369.

[29] Thomas, S.P.; Jose, K.K. and Tomy, L. (2019). *Discrete Harris extended Lindley distribution and applications*, preprint.

[30] Tomy, L. (2018). A retrospective study on Lindley distribution, *Biometrics and Biostatistics International Journal*, **7**(3), 163–169.

[31] Zakerzadeh, H. and Dolati, A. (2009). Generalized Lindley distribution, *Journal of Mathematical Extension*, **3**(2), 13–25.

[32] Zakerzadeh, H. and Mahmoudi, E. (2012). A new two parameter lifetime distribution: model and properties, *arXiv:1204.4248 [stat.CO]*, http://arxiv.org/abs/1204.4248.

REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 697–720

NONPARAMETRIC ESTIMATION OF ROC SURFACES UNDER VERIFICATION BIAS

Authors: Khanh To Duc
– Department of Statistical Sciences, University of Padova, Via C. Battisti, 241-243, 35121 Padova, Italy
toduc@stat.unipd.it

Monica Chiogna
– Department of Statistical Sciences, University of Padova, Via C. Battisti, 241-243, 35121 Padova, Italy
monica@stat.unipd.it

Gianfranco Adimari
– Department of Statistical Sciences, University of Padova, Via C. Battisti, 241-243, 35121 Padova, Italy
adimari@stat.unipd.it

Received: February 2017 Revised: April 2018 Accepted: August 2018

Abstract:

- Verification bias is a well known problem that can affect the statistical evaluation of the predictive ability of a diagnostic test when the true disease status is unknown for some of the patients under study. In this paper, we deal with the assessment of continuous diagnostic tests when an (ordinal) three-class disease status is considered and propose a fully nonparametric verification bias-corrected estimator of the ROC surface based on nearest-neighbor imputation. Consistency and asymptotic normality of the proposed estimator are proved under the missing at random assumption, and its finite sample behavior is investigated by means of Monte Carlo experiments. Variance estimation is also discussed and an illustrative example is presented.

Key-Words:

- *diagnostic tests; missing at random; true class fractions; nearest neighbor imputation.*

AMS Subject Classification:

- 62C99, 62P10.

1. INTRODUCTION

The assessment of diagnostic tests is an important issue in modern medicine. In a two-class problem, i.e. when the disease status has two categories (e.g., "healthy" and "diseased"), for a diagnostic test T that yields a continuous measure, the receiver operating characteristic (ROC) curve is a popular tool for displaying the ability of the test to distinguish between the classes. Assuming, without loss of generality, that higher test values indicate a higher likelihood of disease, the ROC curve is defined as the set of points $\{(1-\mathrm{TNR}(c), \mathrm{TPR}(c)), c \in (-\infty, \infty)\}$ in the unit square, where c is a cut point value, $\mathrm{TPR}(c) = \Pr(T \geq c \mid \text{subject is diseased})$ is the true positive rate at c and $\mathrm{TNR}(c) = \Pr(T < c \mid \text{subject is non–diseased})$ is the true negative rate at c. The shape of the ROC curve allows to evaluate the ability of the test. For example, a ROC curve equal to a straight line joining points $(0,0)$ and $(1,1)$ represents a diagnostic test which is the random guess. A commonly used summary measure of the overall performance of the test is the area under ROC curve (AUC). Under correct ordering, values of AUC range from 0.5, suggesting that the test is no better than chance alone, to 1.0, which indicates a perfect test. See, for example, [13] and [17] as general references.

In some medical studies, the disease status often involves three classes; see, for example, [5], [6] and [11]. In such situations, quantities used to evaluate the accuracy of tests are the true class fractions (TCF's). These quantities are defined as generalizations of TPR and TNR. For a given pair of cut points (c_1, c_2) such that $c_1 < c_2$, the true class fractions TCF's of the continuous test T at (c_1, c_2) are

$$\begin{aligned}
\mathrm{TCF}_1(c_1) &= \Pr(T < c_1 | \text{class } 1) = 1 - \Pr(T \geq c_1 | \text{class } 1), \\
\mathrm{TCF}_2(c_1, c_2) &= \Pr(c_1 \leq T < c_2 | \text{class } 2) \\
&= \Pr(T \geq c_1 | \text{class } 2) - \Pr(T \geq c_2 | \text{class } 2), \\
\mathrm{TCF}_3(c_2) &= \Pr(T \geq c_2 | \text{class } 3) = \Pr(T \geq c_2 | \text{class } 3).
\end{aligned}$$

The plot of (TCF_1, TCF_2, TCF_3) at various values of the pair (c_1, c_2) produces the ROC surface, a generalization of the ROC curve to the unit cube (see [11],[10],[15]). The ROC surface is the region defined by the triangle with vertices $(0,0,1), (0,1,0)$, and $(1,0,0)$ if the three TCF's are identical for every pair (c_1, c_2). In this case, we say that the diagnostic test is, again, the random guess. The ROC surface of an effective test lies in the unit cube above such region. A summary measure of the overall diagnostic accuracy of the test under consideration is the volume under the ROC surface (VUS), which can be seen as a generalization of the AUC. For correctly ordered categories, values of VUS vary from 1/6 to 1, ranging from bad to perfect diagnostic tests.

The application of a diagnostic test in the clinical practice requires a preliminary rigorous statistical assessment of its performance. Clearly, the true ROC curve (or surface) of the test under assessment and its AUC (or VUS) are unknown, so that the statistical evaluation relies on suitable inferential procedures, typically based on measurements collected on a sample of patients. The assessment requires to ascertain the true disease status of the patients in the sample, a verification that it is generally done by employing the most accurate available test, the so-called gold standard (GS) test. Some times, however, the GS test is too expensive, or too invasive, or both to be used on large samples, so that only a subset

of patients undergoes disease verification. It happens that statistical evaluations based only on data from subjects with verified disease status are typically biased, an effect known as verification bias.

Correcting for verification bias is a well known issue of medical statistics. Various methods have been developed to deal with the problem, most of which refer to the two-class case and assume that the true disease status, if missing, is missing at random (MAR, see [9]). We recall, among others, papers [1], [2], [3], [7], [14] and [17]. In particular, for continuous tests, [3] proposes four types of partially parametric estimators of TPR and TNR under the MAR assumption, i.e., full imputation (FI), mean score imputation (MSI), inverse probability weighting (IPW) and semiparametric efficient (SPE, also known as doubly robust DR) estimators. [1] and [2], instead, propose a fully nonparametric approach for ROC curve and AUC estimation, respectively.

The issue of correcting for verification bias in ROC surface analysis is very scarcely considered in the literature. To the best of our knowledge, only [5] and [16] discuss the issue. [5] proposes a maximum likelihood approach for estimation of the ROC surface and corresponding VUS for ordinal diagnostic tests, whereas [16] extends methods in [3] to the estimation of ROC surfaces of continuous diagnostic tests. It is worth noting that FI, MSI, IPW and SPE estimators in [16] are partially parametric estimators and their use requires the specification of parametric regression models for the probability of a subject being correctly classified with respect to the disease state, or the probability of a subject being verified (i.e., tested by GS), or both. As a consequence, a wrong specification of such parametric models negatively affects the behavior of the estimators, that no longer are consistent.

To avoid problems due to model misspecification, in this paper we propose a fully nonparametric approach to estimate TCF_1, TCF_2 and TCF_3 in the presence of verification bias, for continuous diagnostic tests. The proposed approach is based on a nearest-neighbor (NN) imputation of the missing data and extends an idea developed in [1]. Consistency and asymptotic normality of the estimators derived from the proposed method are studied. In addition, estimation of their variance is also discussed. Usefulness of our proposal and advantages in comparison with partially parametric estimators is assessed with the aid of some simulation experiments. An illustrative example is also given.

The rest of paper is organized as follows. In Section 2, we review partially parametric methods for correcting for verification bias in case of continuous tests. The proposed nonparametric method for (pointwise) estimating ROC surfaces and the related asymptotic results are presented in Section 3. In Section 4, we discuss variance-covariance estimation and in Section 5 we give some simulation results. An application is illustrated in Section 6. Finally, conclusions are drawn in Section 7. Some technical details and other simulation results are available in a Supplementary Material, downloadable at http://paduaresearch.cab.unipd.it/11221/ .

2. PARTIALLY PARAMETRIC ESTIMATORS OF ROC SURFACES

Consider a study with n subjects, for whom the result of a continuous diagnostic test T is available. For each subject, $\mathcal{D}$ denotes the true disease status, that can possibly be unknown. Hereafter, we will describe the true disease status as a trinomial random vector $\mathcal{D} = (D_1, D_2, D_3)$. D_k is a binary variable that takes 1 if the subject belongs to class k, $k = 1, 2, 3$ and 0 otherwise. Here, class 1, class 2 and class 3 can be referred, for example, as "non-diseased", "intermediate" and "diseased", and are assumed to be ordered. Further, let V be a binary verification status for a subject, such that $V = 1$ if he/she is undergoes the GS test, and $V = 0$ otherwise. In practice, some information, other than the results from the test T, can be obtained for each patient. Let A be the covariate vector for the patients, that may be associated both with $\mathcal{D}$ and V. We are interested in estimating the ROC surface of T, and hence the true class fractions $\mathrm{TCF}_1(c_1) = \Pr(T_i < c_1 | D_{1i} = 1)$, $\mathrm{TCF}_2(c_1, c_2) = \Pr(c_1 \leq T_i < c_2 | D_{2i} = 1)$ and $\mathrm{TCF}_3(c_2) = \Pr(T_i \geq c_2 | D_{3i} = 1)$, for fixed constants c_1, c_2, with $c_1 < c_2$.

When all patients have their disease status verified by a GS, i.e., $V_i = 1$ for all $i = 1, ..., n$, for any pair of cut points (c_1, c_2), the true class fractions $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ can be easily estimated by

$$\widehat{\mathrm{TCF}}_1(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) D_{1i}}{\sum_{i=1}^{n} D_{1i}},$$

$$\widehat{\mathrm{TCF}}_2(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) D_{2i}}{\sum_{i=1}^{n} D_{2i}},$$

$$\widehat{\mathrm{TCF}}_3(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) D_{3i}}{\sum_{i=1}^{n} D_{3i}},$$

where $\mathrm{I}(\cdot)$ is the indicator function. It is straightforward to show that the above estimators are consistent. However, they cannot be employed in case of incomplete data, i.e. when $V_i = 0$ for some $i = 1, ..., n$.

When only some subjects are selected to undergo the GS test, we need to make an assumption about the selection mechanism. We assume that the verification status V and the disease status $\mathcal{D}$ are mutually independent given the test result T and covariate A. This means that $\Pr(V|T, A) = \Pr(V|\mathcal{D}, T, A)$ or equivalently $\Pr(\mathcal{D}|T, A) = \Pr(\mathcal{D}|V, T, A)$. Such assumption is a special case of the missing at random (MAR) assumption (see [9]).

Under MAR assumption, verification bias-corrected estimation of the true class fractions is discussed in [16], where (partially) parametric estimators, based on four different approaches, are given. In particular, full imputation (FI) estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$

and $\mathrm{TCF}_3(c_2)$ are defined as

$$\widehat{\mathrm{TCF}}_{1,\mathrm{FI}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1)\hat{\rho}_{1i}}{\sum_{i=1}^{n} \hat{\rho}_{1i}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{FI}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2)\hat{\rho}_{2i}}{\sum_{i=1}^{n} \hat{\rho}_{2i}}, \tag{2.1}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{FI}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2)\hat{\rho}_{3i}}{\sum_{i=1}^{n} \hat{\rho}_{3i}}.$$

This method requires a parametric model (e.g. multinomial logistic regression model) to obtain the estimates $\hat{\rho}_{ki}$ of $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$, using only data from verified subjects. Differently, the mean score imputation (MSI) approach only uses the estimates $\hat{\rho}_{ki}$ for the missing values of disease status D_{ki}. Hence, MSI estimators are

$$\widehat{\mathrm{TCF}}_{1,\mathrm{MSI}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1)\left[V_i D_{1i} + (1 - V_i)\hat{\rho}_{1i}\right]}{\sum_{i=1}^{n} \left[V_i D_{1i} + (1 - V_i)\hat{\rho}_{1i}\right]},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{MSI}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2)\left[V_i D_{2i} + (1 - V_i)\hat{\rho}_{2i}\right]}{\sum_{i=1}^{n} \left[V_i D_{2i} + (1 - V_i)\hat{\rho}_{2i}\right]}, \tag{2.2}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{MSI}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2)\left[V_i D_{3i} + (1 - V_i)\hat{\rho}_{3i}\right]}{\sum_{i=1}^{n} \left[V_i D_{3i} + (1 - V_i)\hat{\rho}_{3i}\right]}.$$

The inverse probability weighting (IPW) approach weights each verified subject by the inverse of the probability that the subject is selected for verification. Thus, $\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ are estimated by

$$\widehat{\mathrm{TCF}}_{1,\mathrm{IPW}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) V_i \hat{\pi}_i^{-1} D_{1i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{1i}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{IPW}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) V_i \hat{\pi}_i^{-1} D_{2i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{2i}}, \tag{2.3}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{IPW}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) V_i \hat{\pi}_i^{-1} D_{3i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{3i}},$$

where $\hat{\pi}_i$ is an estimate of the conditional verification probabilities $\pi_i = \Pr(V_i = 1|T_i, A_i)$. Finally, the semiparametric efficient (SPE) estimators are

$$\widehat{\mathrm{TCF}}_{1,\mathrm{SPE}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) \left\{ \frac{V_i D_{1i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{1i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_{1i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{1i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{SPE}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) \left\{ \frac{V_i D_{2i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{2i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_{2i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{2i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}, \tag{2.4}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{SPE}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) \left\{ \frac{V_i D_{3i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{3i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_{3i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{3i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}.$$

Estimators (2.1)-(2.4) represent an extension to the three-classes problem of the estimators proposed in [3]. SPE estimators are also known to be doubly robust estimators, in the sense that they are consistent if either the ρ_{ki}'s or the π_i's are estimated consistently. However, SPE estimates could fall outside the interval $(0, 1)$. This happens because the quantities $V_i D_{ki} \hat{\pi}_i^{-1} - \hat{\rho}_{ki}(V_i - \hat{\pi}_i)\hat{\pi}_i^{-1}$ can be negative.

3. NONPARAMETRIC ESTIMATORS

3.1. The proposed method

All the verification bias-corrected estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ revised in the previous section belong to the class of (partially) parametric estimators, i.e., they need regression models to estimate $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$ and/or $\pi_i = \Pr(V_i = 1|T_i, A_i)$. In what follows, we propose a fully nonparametric approach to the estimation of $\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$. Our approach is based on the K-nearest neighbor (KNN) imputation method. Hereafter, we shall assume that A is a continuous random variable.

Recall that the true disease status is a trinomial random vector $\mathcal{D} = (D_1, D_2, D_3)$ such that D_k is a n Bernoulli trials with success probability $\theta_k = \Pr(D_k = 1)$. Note that $\theta_1 + \theta_2 + \theta_3 = 1$. Since parameters θ_k are the means of the random variables D_k, we can use the KNN estimation procedure discussed in [12] to obtain nonparametric estimates $\hat{\theta}_{k,\mathrm{KNN}}$. More precisely, we define

$$\hat{\theta}_{k,\mathrm{KNN}} = \frac{1}{n} \sum_{i=1}^{n} \left[V_i D_{ki} + (1 - V_i)\hat{\rho}_{ki,K} \right], \qquad K \in \{1, 2, 3, ...\},$$

where $\hat{\rho}_{ki,K} = \frac{1}{K} \sum_{l=1}^{K} D_{ki(l)}$, and $\left\{ (T_{i(l)}, A_{i(l)}, D_{ki(l)}) : V_{i(l)} = 1, l = 1, ..., K \right\}$ is a set of K observed data triplets and $(T_{i(l)}, A_{i(l)})$ denotes the l-th nearest neighbor to (T_i, A_i) among all (T, A)'s corresponding to verified patients, i.e., patients with $V = 1$.

Let $\beta_{jk} = \Pr(T \geq c_j, D_k = 1)$, with $j \in \{1, 2\}$, $k \in \{1, 2, 3\}$ and $k \geq j$. Then, we can define the KNN estimates of β_{jk} as

$$\hat{\beta}_{jk,\mathrm{KNN}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) \left[V_i D_{ki} + (1 - V_i) \hat{\rho}_{ki,K} \right].$$

It follows that the KNN imputation estimators for TCF_k are

$$\begin{aligned}
\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) &= 1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i < c_1) \left[V_i D_{1i} + (1 - V_i) \hat{\rho}_{1i,K} \right]}{\sum_{i=1}^{n} \left[V_i D_{1i} + (1 - V_i) \hat{\rho}_{1i,K} \right]}, \\
\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1, c_2) &= \frac{\hat{\beta}_{12} - \hat{\beta}_{22}}{\hat{\theta}_2} \\
&= \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) \left[V_i D_{2i} + (1 - V_i) \hat{\rho}_{2i,K} \right]}{\sum_{i=1}^{n} \left[V_i D_{2i} + (1 - V_i) \hat{\rho}_{2i,K} \right]}, \qquad (3.1) \\
\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) &= \frac{\hat{\beta}_{23}}{\hat{\theta}_3} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) \left[V_i D_{3i} + (1 - V_i) \hat{\rho}_{3i,K} \right]}{\sum_{i=1}^{n} \left[V_i D_{3i} + (1 - V_i) \hat{\rho}_{3i,K} \right]}.
\end{aligned}$$

Note that KNN estimators (3.1) can be seen as nonparametric versions of the MSI estimators (2.2).

3.2. Asymptotic distribution

Let $\rho_k(t, a) = \Pr(D_k = 1 | T = t, A = a)$ and $\pi(t, a) = \Pr(V = 1 | T = t, A = a)$. The KNN imputation estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ are consistent and asymptotically normal. In fact, we have the following theorems.

Theorem 3.1. *Assume the functions $\rho_k(t, a)$ and $\pi(t, a)$ are finite and first-order differentiable. Moreover, assume that the expectation of $1/\pi(T, A)$ exists. Then, for a fixed pair of cut points (c_1, c_2) such that $c_1 < c_2$, the KNN imputation estimators $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1, c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2)$ are consistent.*

Proof: Since the disease status D_k is a Bernoulli random variable, its second-order moment, $\mathbb{E}(D_k^2)$, is finite. According to the first assumption, we can show that the conditional variance of D_k given T and A, $\mathbb{V}\mathrm{ar}(D_k | T = t, A = a)$, is equal to $\rho_k(t, a) \left[1 - \rho_k(t, a) \right]$, which is clearly finite. Thus, by an application of Theorem 1 in [12], the KNN imputation estimators $\hat{\theta}_{k,\mathrm{KNN}}$ are consistent.

Now, observe that, for $j \in \{1,2\}$, $k \in \{1,2,3\}$ and $k \geq j$,

$$\begin{aligned}
\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)\left[V_i D_{ki} + (1-V_i)\rho_{ki}\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)(\hat{\rho}_{ki,K} - \rho_{ki}) - \beta_{jk} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) V_i \left[D_{ki} - \rho_{ki}\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk}\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)(\hat{\rho}_{ki,K} - \rho_{ki}) \\
&= S_{jk} + R_{jk} + T_{jk}.
\end{aligned}$$

Here, the quantities R_{jk}, S_{jk} and T_{jk} are similar to the quantities R, S and T in the proof of Theorem 2.1 in [4] and of Theorem 1 in [12]. Thus, we have that

$$\begin{aligned}
\sqrt{n}R_{jk} &\xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\right]\right), \\
\sqrt{n}S_{jk} &\xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\pi(T,A)\delta_{jk}^2(T,A)\right]\right),
\end{aligned}$$

where $\delta_{jk}^2(T,A)$ is the conditional variance of $\mathrm{I}(T \geq c_j, D_k = 1)$ given T, A. From proof of Theorem 1 in [12], we also get $T_{jk} = W_{jk} + o_p(n^{-1/2})$, where

$$W_{jk} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)\left[\frac{1}{K}\sum_{l=1}^{K}\left(V_{i(l)}D_{ki(l)} - \rho_{ki(l)}\right)\right],$$

with $\mathbb{E}(W_{jk}) = 0$, $\sqrt{n}W_{jk} \xrightarrow{d} \mathcal{N}\left(0, \sigma_{W_{jk}}^2\right)$, and

$$\sigma_{W_{jk}}^2 = \frac{1}{K}\mathbb{E}\left[(1-\pi(T,A))\delta_{jk}^2(T,A)\right] + \mathbb{E}\left[\frac{(1-\pi(T,A))^2\delta_{jk}^2(T,A)}{\pi(T,A)}\right]. \tag{3.2}$$

This leads to the consistency of $\hat{\beta}_{jk,\mathrm{KNN}}$, i.e, $\hat{\beta}_{jk,\mathrm{KNN}} \xrightarrow{p} \beta_{jk}$. It follows that $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) = 1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1}$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1,c_2) = \frac{\hat{\beta}_{12}-\hat{\beta}_{22}}{\hat{\theta}_2}$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) = \frac{\hat{\beta}_{23}}{\hat{\theta}_3}$ are consistent. □

Theorem 3.2. *Assume that the conditions in Theorem 3.1 hold. We get*

$$\sqrt{n}\left[\begin{pmatrix} \widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) \\ \widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1,c_2) \\ \widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) \end{pmatrix} - \begin{pmatrix} \mathrm{TCF}_1(c_1) \\ \mathrm{TCF}_2(c_1,c_2) \\ \mathrm{TCF}_3(c_2) \end{pmatrix}\right] \xrightarrow{d} \mathcal{N}(0,\Xi), \tag{3.3}$$

where Ξ is a suitable matrix.

Proof: From proof of Theorem 3.1, we have

$$\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk} = S_{jk} + R_{jk} + W_{jk} + o_p(n^{-1/2}),$$

$\sqrt{n}R_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\right]\right)$, $\sqrt{n}S_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\pi(T,A)\delta_{jk}^2(T,A)\right]\right)$ and $\sqrt{n}W_{jk} \xrightarrow{d} \mathcal{N}(0, \sigma_{W_{jk}}^2)$. Moreover, arguments in the proof of Theorem 2.1 in [4] and of Theorem 1 in [12],

allows to state that W_{jk} asymptotically behaves as a sample mean, S_{jk}, R_{jk} and W_{jk} are jointly asymptotically normal, and $\sqrt{n}(\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{jk})$, with $\sigma^2_{jk} = \left[\beta_{jk}\,(1-\beta_{jk}) + \omega^2_{jk}\right]$ and

$$\begin{aligned}\omega^2_{jk} \;=\; & \left(1+\frac{1}{K}\right)\mathbb{E}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))\right] \\ & + \mathbb{E}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\frac{(1-\rho_k(T,A))(1-\pi(T,A))^2}{\pi(T,A)}\right].\end{aligned} \tag{3.4}$$

Finally, a direct application of Theorem 1 in [12] gives that $\sqrt{n}(\hat{\theta}_{k,\mathrm{KNN}} - \theta_k)$ converges to a normal random variable with mean 0 and variance $\sigma^2_k = \left[\theta_k(1-\theta_k) + \omega^2_k\right]$, where

$$\begin{aligned}\omega^2_k \;=\; & \left(1+\frac{1}{K}\right)\mathbb{E}\left[\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))\right] \\ & + \mathbb{E}\left[\frac{\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))^2}{\pi(T,A)}\right].\end{aligned} \tag{3.5}$$

Since $\sqrt{n}\big(\hat{\theta}_{1,\mathrm{KNN}}, \hat{\theta}_{2,\mathrm{KNN}}, \hat{\beta}_{11,\mathrm{KNN}}, \hat{\beta}_{12,\mathrm{KNN}}, \hat{\beta}_{22,\mathrm{KNN}}, \hat{\beta}_{23,\mathrm{KNN}}\big)^\top$ is asymptotically normally distributed with mean $(\theta_1, \theta_2, \beta_{11}, \beta_{12}, \beta_{22}, \beta_{23})^\top$ and suitable covariance matrix Ξ^*, result (3.3) follows by applying the multivariate delta method to

$$h(\hat{\theta}_1, \hat{\theta}_2, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{22}, \hat{\beta}_{23}) = \left(1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1}, \frac{(\hat{\beta}_{12} - \hat{\beta}_{22})}{\hat{\theta}_2}, \frac{\hat{\beta}_{23}}{(1-\hat{\theta}_1-\hat{\theta}_2)}\right). \qquad \square$$

Let us denote elements in the asymptotic covariance matrix Ξ as follows

$$\Xi = \begin{pmatrix} \xi_1^2 & \xi_{12} & \xi_{13} \\ \xi_{12} & \xi_2^2 & \xi_{23} \\ \xi_{13} & \xi_{23} & \xi_3^2 \end{pmatrix}.$$

Recall that, from proof of Theorem 3.2, $\sigma^2_k = \left[\theta_k(1-\theta_k) + \omega^2_k\right]$ and $\sigma^2_{jk} = \beta_{jk}(1-\beta_{jk}) + \omega^2_{jk}$, where ω^2_k and ω^2_{jk} are given in (3.5) and (3.4), respectively. In Section S1, Supplementary Material, we show that

$$\begin{aligned}
\xi_1^2 \;=\; & \frac{\beta_{11}^2}{\theta_1^4}\sigma_1^2 + \frac{\sigma_{11}^2}{\theta_1^2} - \frac{\beta_{11}}{\theta_1^3}(\sigma_1^2 + \sigma_{11}^2 - \zeta_{11}^2), \\
\xi_2^2 \;=\; & \sigma_2^2\frac{(\beta_{12}-\beta_{22})^2}{\theta_2^4} + \frac{\lambda^2}{\theta_2^2} - \frac{\beta_{12}-\beta_{22}}{\theta_2^3}(\sigma_{12}^2 - \sigma_{22}^2 - \zeta_{12}^2 + \zeta_{22}^2), \\
\xi_3^2 \;=\; & \frac{\beta_{23}^2\sigma_3^2}{\theta_3^4} + \frac{\sigma_{23}^2}{\theta_3^2} - \frac{\beta_{23}}{\theta_3^3}(\sigma_3^2 + \sigma_{23}^2 - \zeta_{23}^2), \\
\xi_{12} \;=\; & \frac{1}{\theta_1\theta_2}\left[\psi_{1212}^2 + \beta_{11}(\beta_{12}-\beta_{22})\right] - \frac{\beta_{11}}{\theta_1^2\theta_2}\left[\psi_{1212}^2 + \theta_1(\beta_{12}-\beta_{22})\right] \\
& - \frac{\beta_{12}-\beta_{22}}{\theta_2^2\theta_1}\left(\frac{\beta_{11}}{\theta_1}\sigma_{12}^* + \psi_{112}^2 + \theta_2\beta_{11}\right), \\
\xi_{13} \;=\; & \frac{1}{\theta_3}\left[-\frac{\beta_{11}}{\theta_1^2}(\psi_{213}^2 + \theta_1\beta_{23}) + \frac{\psi_{213}^2 + \beta_{11}\beta_{23}}{\theta_1}\right] + \frac{\beta_{23}}{\theta_1\theta_3^2} \\
& \times \left[\frac{\beta_{11}}{\theta_1}\left(\sigma_1^2 + \sigma_{12}^*\right) - \psi_{113}^2 - \theta_3\beta_{11}\right], \\
\xi_{23} \;=\; & \frac{1}{\theta_2\theta_3}\left[-\beta_{23}(\beta_{12}-\beta_{22}) + \frac{\beta_{12}-\beta_{22}}{\theta_2}(\psi_{223}^2 + \theta_2\beta_{23})\right] \\
& + \frac{\beta_{23}}{\theta_2\theta_3^2}\left[\psi_{1223}^2 + \theta_3(\beta_{12}-\beta_{22}) - \frac{\beta_{12}-\beta_{22}}{\theta_2}(\sigma_2^2 + \sigma_{12}^*)\right],
\end{aligned} \tag{3.6}$$

where $\zeta_{jk}^2 = \gamma_{jk}(1-\gamma_{jk}) + \eta_{jk}^2$, $\lambda^2 = (\beta_{12} - \beta_{22})[1 - (\beta_{12} - \beta_{22})] + \omega_{12}^2 - \omega_{22}^2$, $\sigma_{12}^* = -(\theta_1\theta_2 + \psi_{12}^2)$, with $\gamma_{jk} = \Pr(T < c_j, D_k = 1)$ and

$$\begin{aligned}\eta_{jk}^2 &= \frac{K+1}{K}\mathbb{E}\left[\mathrm{I}(T < c_j)\rho_k(T,A)\{1-\rho_k(T,A)\}\{1-\pi(T,A)\}\right] \\ &\quad + \mathbb{E}\left[\mathrm{I}(T < c_j)\rho_k(T,A)\frac{\{1-\rho_k(T,A)\}\{1-\pi(T,A)\}^2}{\pi(T,A)}\right],\end{aligned}$$

$$\begin{aligned}\psi_{12}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\rho_1(T,A)\rho_2(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{1212}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \le T < c_2)\rho_1(T,A)\rho_2(T,A)\right\} \\ &\quad + \mathbb{E}\left\{[1-\pi(T,A)]^2\mathrm{I}(c_1 \le T < c_2)\frac{\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{112}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_2(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{213}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_2)\rho_1(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_2)\rho_1(T,A)\rho_3(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{113}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_3(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{223}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_2)\rho_2(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_2)\rho_2(T,A)\rho_3(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{1223}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \le T < c_2)\rho_2(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{[1-\pi(T,A)]^2\mathrm{I}(c_1 \le T < c_2)\frac{\rho_2(T,A)\rho_3(T,A)}{\pi(T,A)}\right\}.\end{aligned}$$

Therefore, from (3.6), the elements of Ξ depend, among others, on quantities as ω_k^2, ω_{jk}^2, γ_{jk}, η_{jk}^2, ψ_{1212}^2, ψ_{112}^2, ψ_{213}^2, ψ_{12}^2, ψ_{113}^2, ψ_{223}^2 and ψ_{1223}^2. As a consequence, to obtain consistent estimates of the asymptotic variances and covariances, we ultimately need to estimate these quantities.

3.3. Choice of K and of the distance measure

The proposed method is based on nearest-neighbor imputation, which requires the choice of a value for K as well as a distance measure.

In practice, the selection of a suitable distance is typically dictated by features of the data and possible subjective evaluations; thus, a general indication about an adequate choice is difficult to express. In many cases, the simple Euclidean distance may be appropriate. Other times, the researcher may wish to consider specific characteristics of data at hand, and then make a different choice. For example, the diagnostic test result T and the auxiliary covariate A could be heterogeneous with respect to their variances (in particular when the variables are measured on different scales). In this case, the choice of the Mahalanobis distance may be suitable. A further discussion on this topic in the context of medical studies can be found in [8]. Therein, we refer the reader to results relative to numerical datasets.

As for the choice of the size of the neighborhood, [12] argue that nearest-neighbor imputation with a small value of K typically yields negligible bias of the estimators, but a large variance; the opposite happens with a large value of K. The authors suggest that the choice of $K \in \{1, 2\}$ is generally adequate when the aim is to estimate a mean. A similar comment is also raised by [1] and [2], i.e., a small value of K, within the range 1–3, may be a good choice to estimate ROC curves and AUC. However, the authors stress that, in general, the choice of K may depend on the dimension of the feature space, and propose to use cross–validation to find K. Specifically, the authors indicate that a suitable value of the size of neighbor could be found by

$$K^* = \arg\min_K \frac{1}{n_{ver}} \left\| D - \hat{\rho}_K \right\|_1 ,$$

where D is a binary disease status, $\|\cdot\|_1$ denotes L_1 norm for vector and n_{ver} is the number of verified subjects. The formula above can be generalized to our three–class case. In fact, when the disease status has q categories ($q \geq 3$), the difference between $\mathcal{D}$ and $\hat{\rho}_K$ is a $n_{ver} \times (q-1)$ matrix. In such situation, the selection rule could be

$$K^* = \arg\min_K \frac{1}{n_{ver}(q-1)} \left\| \mathcal{D} - \hat{\rho}_K \right\|_{1,1} , \tag{3.7}$$

where $\|\mathcal{A}\|_{1,1}$ denotes $L_{1,1}$ norm of matrix $\mathcal{A}$, i.e.,

$$\|\mathcal{A}\|_{1,1} = \sum_{j=1}^{q-1} \left(\sum_{i=1}^{n_{ver}} |a_{ij}| \right) .$$

4. VARIANCE-COVARIANCE ESTIMATION

Consider first the problem of estimating the variances of $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1, c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2)$. In a nonparametric framework, quantities as $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2 in Section 3.2 can be estimated by their empirical counterparts, using also the plug–in method. Here, we consider an approach that uses a nearest-neighbor rule to estimate the functions $\rho_k(T, A)$

and the propensity score $\pi(T, A)$, that appear in the expressions of $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2. In particular, for the conditional probabilities of disease, we can use KNN estimates $\tilde{\rho}_{ki} = \hat{\rho}_{ki,\bar{K}}$, where the integer $\bar{K}$ must be greater than one to avoid estimates equal to zero. For the conditional probabilities of verification, we can resort to the KNN procedure proposed in [1], which considers the estimates

$$\tilde{\pi}_i = \frac{1}{K_i^*} \sum_{l=1}^{K_i^*} V_{i(l)},$$

where $\left\{(T_{i(l)}, A_{i(l)}, V_{i(l)}) : l = 1, ..., K_i^*\right\}$ is a set of K_i^* observed triplets and $(T_{i(l)}, A_{i(l)})$ denotes the l-th nearest neighbor to (T_i, A_i) among all (T, A)'s. When V_i equals 0, K_i^* is set equal to the rank of the first verified nearest neighbor to the unit i, i.e., K_i^* is such that $V_{i(K_i^*)} = 1$ and $V_i = V_{i(1)} = V_{i(2)} = ... = V_{i(K_i^*-1)} = 0$. In case of $V_i = 1$, K_i^* is such that $V_i = V_{i(1)} = V_{i(2)} = ... = V_{i(K_i^*-1)} = 1$, and $V_{i(K_i^*)} = 0$, i.e., K_i^* is set equal to the rank of the first non–verified nearest neighbor to the unit i. Such a procedure automatically avoids zero values for the $\tilde{\pi}_i$'s.

Then, based on the $\tilde{\rho}_{ki}$'s and $\tilde{\pi}_i$'s, we obtain the estimates

$$\begin{aligned}
\hat{\omega}_k^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right) + \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i}, \\
\hat{\omega}_{jk}^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}(T_i \geq c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i}, \\
\hat{\eta}_{jk}^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \mathrm{I}(T_i < c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}(T_i < c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i},
\end{aligned}$$

from which, along with $\hat{\theta}_{k,\mathrm{KNN}}$, $\hat{\beta}_{jk,\mathrm{KNN}}$ and

$$\hat{\gamma}_{jk,\mathrm{KNN}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(T_i < c_j) \left[V_i D_{ki} + (1 - V_i) \hat{\rho}_{ki,K}\right],$$

one derives the estimates of the variances of the proposed KNN imputation estimators.

To obtain estimates of covariances, we need to estimate also the quantities ψ_{1212}^2, ψ_{112}^2, ψ_{213}^2, ψ_{12}^2, ψ_{113}^2, ψ_{223}^2 and ψ_{1223}^2. However, estimates of such quantities are similar to those given above for $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2. For example,

$$\begin{aligned}
\hat{\psi}_{1212}^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) \tilde{\rho}_{1i} \tilde{\rho}_{2i} \left(1 - \tilde{\pi}_i\right) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}(c_1 \leq T_i < c_2) \tilde{\rho}_{1i} \tilde{\rho}_{2i} \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i}.
\end{aligned}$$

Of course, there are other possible approaches to obtain variance and covariance estimates. For instance, one could resort to a standard bootstrap procedure.

5. SIMULATION STUDY

In this section, the ability of KNN method to estimate TCF_1, TCF_2 and TCF_3 is evaluated by using Monte Carlo experiments. We also compare the proposed method with partially parametric approaches, namely, FI, MSI, IPW and SPE approaches. As already mentioned, partially parametric bias-corrected estimators of TCF_1, TCF_2 and TCF_3 require parametric regression models to estimate $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$, or $\pi_i = \Pr(V_i = 1|T_i, A_i)$, or both. A wrong specification of such models may affect the estimators. Therefore, in the simulation study we consider two scenarios: in the parametric estimation process,

(**i**) the disease model and the verification model are both correctly specified;

(**ii**) the disease model and the verification model are both misspecified.

In both scenarios, we execute 5000 Monte Carlo runs at each setting; we set three sample sizes, i.e., 250, 500 and 1000 in scenario (i) and a sample size of 1000 in scenario (ii).

We consider KNN estimators based on the Euclidean distance, with $K = 1$ and $K = 3$. This in light of the discussion in Section 3.4 and some results of a preliminary simulation study presented in Section S5, Supplementary Material. In such preliminary study, we compared the behavior of the KNN estimators for several choices of the distance measure (Euclidean, Manhattan, Canberra and Mahalanobis) and the size of the neighborhood ($K = 1, 3, 5, 10, 20$).

5.1. Correctly specified parametric models

The true disease is generated by a trinomial random vector (D_1, D_2, D_3), such that D_k is a Bernoulli random variable with success probability θ_k, $k = 1, 2, 3$. We set $\theta_1 = 0.4, \theta_2 = 0.35$ and $\theta_3 = 0.25$. The continuous test result T and a covariate A are generated from the following conditional models

$$T, A|D_k \sim \mathcal{N}_2\left(\mu_k, \Sigma\right), \qquad k = 1, 2, 3,$$

where $\mu_k = (2k, k)^\top$ and

$$\Sigma = \begin{pmatrix} \sigma^2_{T|D} & \sigma_{T,A|D} \\ \sigma_{T,A|D} & \sigma^2_{A|D} \end{pmatrix}.$$

We consider three different values for Σ, specifically

$$\begin{pmatrix} 1.75 & 0.1 \\ 0.1 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 5.5 & 3 \\ 3 & 2.5 \end{pmatrix},$$

giving rise to a correlation between T and A equal to $0.36, 0.69$ and 0.84, respectively. The verification status V is generated by the following model

$$\mathrm{logit}\left\{\Pr(V = 1|T, A)\right\} = \delta_0 + \delta_1 T + \delta_2 A,$$

where we fix $\delta_0 = 0.5, \delta_1 = -0.3$ and $\delta_2 = 0.75$. This choice corresponds to a verification rate of about 0.65. We consider six pairs of cut points (c_1, c_2), i.e., $(2, 4)$, $(2, 5)$, $(2, 7)$, $(4, 5)$, $(4, 7)$

and $(5, 7)$. Since the conditional distribution of T given D_k is the normal distribution, the true parameters values are

$$\begin{aligned}
\mathrm{TCF}_1(c_1) &= \Phi\left(\frac{c_1 - 2}{\sigma_{T|D}}\right), \\
\mathrm{TCF}_2(c_1, c_2) &= \Phi\left(\frac{c_2 - 4}{\sigma_{T|D}}\right) - \Phi\left(\frac{c_1 - 4}{\sigma_{T|D}}\right), \\
\mathrm{TCF}_3(c_2) &= 1 - \Phi\left(\frac{c_2 - 6}{\sigma_{T|D}}\right),
\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal random variable.

In this set–up, FI, MSI, IPW and SPE estimators are computed under correct working models for both the disease and the verification processes. Therefore, the conditional verification probabilities π_i are estimated from a logistic model for V given T and A with logit link. Under our data–generating process, the true conditional disease model is a multinomial logistic model

$$\Pr(D_k = 1|T, A) = \frac{\exp\left(\tau_{0k} + \tau_{1k}T + \tau_{2k}A\right)}{1 + \exp\left(\tau_{01} + \tau_{11}T + \tau_{21}A\right) + \exp\left(\tau_{02} + \tau_{12}T + \tau_{22}A\right)}$$

for suitable $\tau_{0k}, \tau_{1k}, \tau_{2k}$, where $k = 1, 2$.

Tables 1–3 show Monte Carlo means and standard deviations of the estimators for the three true class factions. Results concern the estimators FI, MSI, IPW, SPE, and the KNN estimator with $K = 1$ and $K = 3$ computed using the Euclidean distance. Also, the estimated standard deviations are shown in the tables. The estimates are obtained by using asymptotic results. To estimate standard deviations of KNN estimators, we use the KNN procedure discussed in Section 4, with $\bar{K} = 2$. Each table refers to a chosen value for Σ. The sample size is 250. The results for sample sizes 500 and 1000 are presented in Section S2 of Supplementary Material.

As expected, the parametric approaches work well when both models for $\rho_k(t, a)$ and $\pi(t, a)$ are correctly specified. FI and MSI estimators seem to be the most efficient ones, whereas the IPW approach seems to provide less powerful estimators, in general. The new proposals (1NN and 3NN estimators) yield also good results, comparable, in terms of bias and standard deviation, to those of the parametric competitors. Moreover, estimators 1NN and 3NN seem to achieve similar performances, and the results about estimated standard deviations of KNN estimators seem to show the effectiveness of the procedure discussed in Section 4.

Finally, some results of simulation experiments performed to explore the effect of a multidimensional vector of auxiliary covariates are given in Section S3, Supplementary Material. A vector A of dimension 3 is employed. The results in Table 7, Supplementary Material, show that KNN estimators still behave satisfactorily.

Table 1: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when $n = 250$ and the first value of Σ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut points $= (2, 4)$				
True	**0.5000**	**0.4347**	**0.9347**						
FI	0.5005	0.4348	0.9344	0.0537	0.0484	0.0269	0.0440	0.0398	0.0500
MSI	0.5005	0.4346	0.9342	0.0550	0.0547	0.0320	0.0465	0.0475	0.0536
IPW	0.4998	0.4349	0.9341	0.0722	0.0727	0.0372	0.0688	0.0702	0.0420
SPE	0.5010	0.4346	0.9344	0.0628	0.0659	0.0364	0.0857	0.0637	0.0363
1NN	**0.4989**	**0.4334**	**0.9331**	0.0592	0.0665	0.0387	0.0555	0.0626	0.0382
3NN	**0.4975**	**0.4325**	**0.9322**	0.0567	0.0617	0.0364	0.0545	0.0608	0.0372
					cut points $= (2, 5)$				
True	**0.5000**	**0.7099**	**0.7752**						
FI	0.5005	0.7111	0.7761	0.0537	0.0461	0.0534	0.0440	0.0400	0.0583
MSI	0.5005	0.7104	0.7756	0.0550	0.0511	0.0566	0.0465	0.0467	0.0626
IPW	0.4998	0.7108	0.7750	0.0722	0.0701	0.0663	0.0688	0.0667	0.0713
SPE	0.5010	0.7106	0.7762	0.0628	0.0619	0.0627	0.0857	0.0604	0.0611
1NN	**0.4989**	**0.7068**	**0.7738**	0.0592	0.0627	0.0652	0.0555	0.0591	0.0625
3NN	**0.4975**	**0.7038**	**0.7714**	0.0567	0.0576	0.0615	0.0545	0.0574	0.0610
					cut points $= (2, 7)$				
True	**0.5000**	**0.9230**	**0.2248**						
FI	0.5005	0.9229	0.2240	0.0537	0.0236	0.0522	0.0440	0.0309	0.0428
MSI	0.5005	0.9231	0.2243	0.0550	0.0285	0.0531	0.0465	0.0353	0.0443
IPW	0.4998	0.9238	0.2222	0.0722	0.0374	0.0765	0.0688	0.0360	0.0728
SPE	0.5010	0.9236	0.2250	0.0628	0.0362	0.0578	0.0857	0.0348	0.0573
1NN	**0.4989**	**0.9201**	**0.2233**	0.0592	0.0372	0.0577	0.0555	0.0366	0.0570
3NN	**0.4975**	**0.9177**	**0.2216**	0.0567	0.0340	0.0558	0.0545	0.0355	0.0563
					cut points $= (4, 5)$				
True	**0.9347**	**0.2752**	**0.7752**						
FI	0.9347	0.2763	0.7761	0.0245	0.0412	0.0534	0.0179	0.0336	0.0583
MSI	0.9348	0.2758	0.7756	0.0271	0.0471	0.0566	0.0220	0.0404	0.0626
IPW	0.9350	0.2758	0.7750	0.0421	0.0693	0.0663	0.0391	0.0651	0.0713
SPE	0.9353	0.2761	0.7762	0.0386	0.0590	0.0627	0.0377	0.0568	0.0611
1NN	**0.9322**	**0.2734**	**0.7738**	0.0374	0.0572	0.0652	0.0342	0.0553	0.0625
3NN	**0.9303**	**0.2712**	**0.7714**	0.0328	0.0526	0.0615	0.0332	0.0538	0.0610
					cut points $= (4, 7)$				
True	**0.9347**	**0.4883**	**0.2248**						
FI	0.9347	0.4881	0.2240	0.0245	0.0541	0.0522	0.0179	0.0444	0.0428
MSI	0.9348	0.4885	0.2243	0.0271	0.0576	0.0531	0.0220	0.0495	0.0443
IPW	0.9350	0.4889	0.2222	0.0421	0.0741	0.0765	0.0391	0.0713	0.0728
SPE	0.9353	0.4890	0.2250	0.0386	0.0674	0.0578	0.0377	0.0646	0.0573
1NN	**0.9322**	**0.4867**	**0.2233**	0.0374	0.0680	0.0577	0.0342	0.0633	0.0570
3NN	**0.9303**	**0.4852**	**0.2216**	0.0328	0.0630	0.0558	0.0332	0.0615	0.0563
					cut points $= (5, 7)$				
True	**0.9883**	**0.2132**	**0.2248**						
FI	0.9879	0.2118	0.2240	0.0075	0.0435	0.0522	0.0055	0.0336	0.0428
MSI	0.9882	0.2127	0.2243	0.0096	0.0467	0.0531	0.0084	0.0388	0.0443
IPW	0.9887	0.2130	0.2222	0.0193	0.0653	0.0765	0.0177	0.0618	0.0728
SPE	0.9888	0.2130	0.2250	0.0191	0.0571	0.0578	0.0184	0.0554	0.0573
1NN	**0.9868**	**0.2133**	**0.2233**	0.0177	0.0567	0.0577	0.0172	0.0532	0.0570
3NN	**0.9860**	**0.2139**	**0.2216**	0.0151	0.0519	0.0558	0.0168	0.0516	0.0563

Table 2: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when $n = 250$ and the second value of Σ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut points $= (2, 4)$				
True	**0.5000**	**0.3970**	**0.8970**						
FI	0.4999	0.3974	0.8973	0.0503	0.0421	0.0362	0.0432	0.0352	0.0466
MSI	0.5000	0.3975	0.8971	0.0521	0.0497	0.0416	0.0461	0.0451	0.0515
IPW	0.4989	0.3990	0.8971	0.0663	0.0685	0.0534	0.0647	0.0681	0.0530
SPE	0.5004	0.3980	0.8976	0.0570	0.0619	0.0516	0.0563	0.0620	0.0493
1NN	**0.4982**	**0.3953**	**0.8976**	0.0587	0.0642	0.0537	0.0561	0.0618	0.0487
3NN	**0.4960**	**0.3933**	**0.8970**	0.0556	0.0595	0.0494	0.0548	0.0600	0.0472
					cut points $= (2, 5)$				
True	**0.5000**	**0.6335**	**0.7365**						
FI	0.4999	0.6337	0.7395	0.0503	0.0436	0.0583	0.0432	0.0379	0.0554
MSI	0.5000	0.6330	0.7385	0.0521	0.0508	0.0613	0.0461	0.0469	0.0612
IPW	0.4989	0.6335	0.7386	0.0663	0.0676	0.0728	0.0647	0.0663	0.0745
SPE	0.5004	0.6333	0.7390	0.0570	0.0622	0.0682	0.0563	0.0612	0.0673
1NN	**0.4982**	**0.6304**	**0.7400**	0.0587	0.0645	0.0721	0.0561	0.0615	0.0672
3NN	**0.4960**	**0.6283**	**0.7396**	0.0556	0.0600	0.0670	0.0548	0.0597	0.0654
					cut points $= (2, 7)$				
True	**0.5000**	**0.8682**	**0.2635**						
FI	0.4999	0.8676	0.2655	0.0503	0.0316	0.0560	0.0432	0.0294	0.0478
MSI	0.5000	0.8678	0.2660	0.0521	0.0374	0.0583	0.0461	0.0364	0.0512
IPW	0.4989	0.8682	0.2669	0.0663	0.0507	0.0698	0.0647	0.0484	0.0692
SPE	0.5004	0.8681	0.2663	0.0570	0.0476	0.0608	0.0563	0.0459	0.0600
1NN	**0.4982**	**0.8672**	**0.2672**	0.0587	0.0495	0.0629	0.0561	0.0458	0.0609
3NN	**0.4960**	**0.8657**	**0.2671**	0.0556	0.0452	0.0610	0.0548	0.0442	0.0601
					cut points $= (4, 5)$				
True	**0.8970**	**0.2365**	**0.7365**						
FI	0.8980	0.2363	0.7395	0.0284	0.0367	0.0583	0.0239	0.0301	0.0554
MSI	0.8976	0.2356	0.7385	0.0318	0.0437	0.0613	0.0292	0.0386	0.0612
IPW	0.8975	0.2345	0.7386	0.0377	0.0594	0.0728	0.0373	0.0578	0.0745
SPE	0.8974	0.2353	0.7390	0.0364	0.0529	0.0682	0.0361	0.0522	0.0673
1NN	**0.8958**	**0.2352**	**0.7400**	0.0388	0.0540	0.0721	0.0373	0.0524	0.0672
3NN	**0.8946**	**0.2350**	**0.7396**	0.0362	0.0502	0.0670	0.0361	0.0510	0.0654
					cut points $= (4, 7)$				
True	**0.8970**	**0.4711**	**0.2635**						
FI	0.8980	0.4703	0.2655	0.0284	0.0512	0.0560	0.0239	0.0413	0.0478
MSI	0.8976	0.4703	0.2660	0.0318	0.0561	0.0583	0.0292	0.0490	0.0512
IPW	0.8975	0.4692	0.2669	0.0377	0.0693	0.0698	0.0373	0.0679	0.0692
SPE	0.8974	0.4701	0.2663	0.0364	0.0638	0.0608	0.0361	0.0629	0.0600
1NN	**0.8958**	**0.4719**	**0.2672**	0.0388	0.0666	0.0629	0.0373	0.0630	0.0609
3NN	**0.8946**	**0.4724**	**0.2671**	0.0362	0.0627	0.0610	0.0361	0.0611	0.0601
					cut points $= (5, 7)$				
True	**0.9711**	**0.2347**	**0.2635**						
FI	0.9710	0.2339	0.2655	0.0124	0.0407	0.0560	0.0104	0.0336	0.0478
MSI	0.9709	0.2348	0.2660	0.0166	0.0461	0.0583	0.0156	0.0412	0.0512
IPW	0.9709	0.2347	0.2669	0.0204	0.0568	0.0698	0.0202	0.0562	0.0692
SPE	0.9709	0.2348	0.2663	0.0202	0.0531	0.0608	0.0199	0.0524	0.0600
1NN	**0.9701**	**0.2368**	**0.2672**	0.0217	0.0549	0.0629	0.0213	0.0533	0.0609
3NN	**0.9695**	**0.2375**	**0.2671**	0.0200	0.0519	0.0610	0.0206	0.0517	0.0601

Table 3: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when $n = 250$ and the third value of Σ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
cut points $= (2, 4)$									
True	**0.5000**	**0.3031**	**0.8031**						
FI	0.5009	0.3031	0.8047	0.0488	0.0344	0.0495	0.0418	0.0284	0.0467
MSI	0.5005	0.3032	0.8045	0.0515	0.0448	0.0544	0.0460	0.0410	0.0542
IPW	0.5015	0.3030	0.8043	0.0624	0.0632	0.0649	0.0618	0.0620	0.0640
SPE	0.5007	0.3034	0.8043	0.0565	0.0576	0.0628	0.0564	0.0574	0.0614
1NN	**0.4997**	**0.3021**	**0.8047**	0.0592	0.0602	0.0682	0.0571	0.0584	0.0621
3NN	**0.4984**	**0.3018**	**0.8043**	0.0561	0.0565	0.0632	0.0556	0.0566	0.0601
cut points $= (2, 5)$									
True	**0.5000**	**0.4682**	**0.6651**						
FI	0.5009	0.4692	0.6668	0.0488	0.0384	0.0616	0.0418	0.0323	0.0536
MSI	0.5005	0.4687	0.6666	0.0515	0.0495	0.0658	0.0460	0.0455	0.0610
IPW	0.5015	0.4681	0.6670	0.0624	0.0671	0.0753	0.0618	0.0670	0.0743
SPE	0.5007	0.4690	0.6665	0.0565	0.0624	0.0721	0.0564	0.0622	0.0704
1NN	**0.4997**	**0.4676**	**0.6668**	0.0592	0.0661	0.0780	0.0571	0.0634	0.0717
3NN	**0.4984**	**0.4670**	**0.6666**	0.0561	0.0619	0.0729	0.0556	0.0614	0.0695
cut points $= (2, 7)$									
True	**0.5000**	**0.7027**	**0.3349**						
FI	0.5009	0.7030	0.3358	0.0488	0.0375	0.0595	0.0418	0.0318	0.0501
MSI	0.5005	0.7027	0.3360	0.0515	0.0474	0.0637	0.0460	0.0435	0.0563
IPW	0.5015	0.7026	0.3366	0.0624	0.0625	0.0730	0.0618	0.0618	0.0716
SPE	0.5007	0.7032	0.3362	0.0565	0.0591	0.0677	0.0564	0.0583	0.0657
1NN	**0.4997**	**0.7024**	**0.3366**	0.0592	0.0633	0.0712	0.0571	0.0592	0.0675
3NN	**0.4984**	**0.7016**	**0.3362**	0.0561	0.0590	0.0680	0.0556	0.0572	0.0660
cut points $= (4, 5)$									
True	**0.8031**	**0.1651**	**0.6651**						
FI	0.8042	0.1660	0.6668	0.0383	0.0277	0.0616	0.0323	0.0231	0.0536
MSI	0.8037	0.1655	0.6666	0.0415	0.0372	0.0658	0.0380	0.0333	0.0610
IPW	0.8039	0.1651	0.6670	0.0473	0.0503	0.0753	0.0473	0.0493	0.0743
SPE	0.8036	0.1655	0.6665	0.0456	0.0465	0.0721	0.0458	0.0455	0.0704
1NN	**0.8032**	**0.1655**	**0.6668**	0.0487	0.0481	0.0780	0.0472	0.0466	0.0717
3NN	**0.8020**	**0.1651**	**0.6666**	0.0460	0.0450	0.0729	0.0457	0.0451	0.0695
cut points $= (4, 7)$									
True	**0.8031**	**0.3996**	**0.3349**						
FI	0.8042	0.3999	0.3358	0.0383	0.0426	0.0595	0.0323	0.0349	0.0501
MSI	0.8037	0.3995	0.3360	0.0415	0.0522	0.0637	0.0380	0.0463	0.0563
IPW	0.8039	0.3996	0.3366	0.0473	0.0658	0.0730	0.0473	0.0645	0.0716
SPE	0.8036	0.3998	0.3362	0.0456	0.0618	0.0677	0.0458	0.0606	0.0657
1NN	**0.8032**	**0.4003**	**0.3366**	0.0487	0.0660	0.0712	0.0472	0.0619	0.0675
3NN	**0.8020**	**0.3998**	**0.3362**	0.0460	0.0617	0.0680	0.0457	0.0600	0.0660
cut points $= (5, 7)$									
True	**0.8996**	**0.2345**	**0.3349**						
FI	0.9003	0.2338	0.3358	0.0266	0.0351	0.0595	0.0224	0.0292	0.0501
MSI	0.9004	0.2340	0.3360	0.0308	0.0443	0.0637	0.0285	0.0398	0.0563
IPW	0.9005	0.2345	0.3366	0.0355	0.0555	0.0730	0.0353	0.0550	0.0716
SPE	0.9004	0.2342	0.3362	0.0349	0.0523	0.0677	0.0346	0.0517	0.0657
1NN	**0.9000**	**0.2348**	**0.3366**	0.0373	0.0556	0.0712	0.0361	0.0531	0.0675
3NN	**0.8992**	**0.2346**	**0.3362**	0.0349	0.0520	0.0680	0.0349	0.0515	0.0660

5.2. Misspecified models

We start from two independent random variables $Z_1 \sim \mathcal{N}(0, 0.5)$ and $Z_2 \sim \mathcal{N}(0, 0.5)$. The true conditional disease is generated by a trinomial random vector (D_1, D_2, D_3) such that

$$D_1 = \begin{cases} 1 \text{ if } Z_1 + Z_2 \leq h_1 \\ 0 \text{ otherwise} \end{cases}, \qquad D_2 = \begin{cases} 1 \text{ if } h_1 < Z_1 + Z_2 \leq h_2 \\ 0 \text{ otherwise} \end{cases},$$

and

$$D_3 = \begin{cases} 1 \text{ if } Z_1 + Z_2 > h_2 \\ 0 \text{ otherwise} \end{cases}.$$

Here, h_1 and h_2 are two thresholds. We choose h_1 and h_2 to make $\theta_1 = 0.4$ and $\theta_3 = 0.25$. The continuous test results T and the covariate A are generated to be related to $\mathcal{D}$ through Z_1 and Z_2. More precisely,

$$T = \alpha(Z_1 + Z_2) + \varepsilon_1, \qquad A = Z_1 + Z_2 + \varepsilon_2,$$

where ε_1 and ε_2 are two independent normal random variables with mean 0 and the common variance 0.25. We choose $\alpha = 0.5$. The verification status V is simulated by the following logistic model

$$\text{logit}\left\{\Pr(V = 1|T, A)\right\} = -1.5 - 0.35T - 1.5A.$$

Under this model, the verification rate is roughly 0.276. For the cut-point, we consider six pairs (c_1, c_2), i.e., $(-1.0, -0.5)$, $(-1.0, 0.7)$, $(-1.0, 1.3)$, $(-0.5, 0.7)$, $(-0.5, 1.3)$ and $(0.7, 1.3)$. Within this set–up, we determine the true values of TCF's as follows:

$$\begin{aligned}
\text{TCF}_1(c_1) &= \frac{1}{\Phi(h_1)} \int_{-\infty}^{h_1} \Phi\left(\frac{c_1 - \alpha z}{\sqrt{0.25}}\right) \phi(z)\mathrm{d}z, \\
\text{TCF}_2(c_1, c_2) &= \frac{1}{\Phi(h_2) - \Phi(h_1)} \int_{h_1}^{h_2} \left[\Phi\left(\frac{c_2 - \alpha z}{\sqrt{0.25}}\right) - \Phi\left(\frac{c_1 - \alpha z}{\sqrt{0.25}}\right)\right] \phi(z)\mathrm{d}z, \\
\text{TCF}_3(c_2) &= 1 - \frac{1}{1 - \Phi(h_2)} \int_{h_2}^{\infty} \Phi\left(\frac{c_2 - \alpha z}{\sqrt{0.25}}\right) \phi(z)\mathrm{d}z,
\end{aligned}$$

where $\phi(\cdot)$ denotes the density function of the standard normal random variable.

The aim in this scenario is to compare FI, MSI, IPW, SPE and KNN estimators when both the estimates for $\hat{\pi}_i$ and $\hat{\rho}_{ki}$ in the parametric approach are inconsistent. Therefore, $\hat{\rho}_{ki}$ is obtained from a multinomial logistic regression model with $\mathcal{D} = (D_1, D_2, D_3)$ as the response and T as predictor. To estimate π_i, we use a generalized linear model for V given T and $A^{2/3}$ with logit link. Clearly, the two fitted models are misspecified. The KNN estimators are obtained by using $K = 1$ and $K = 3$ and the Euclidean distance. Again, we use $\bar{K} = 2$ in the KNN procedure to estimate standard deviations of KNN estimators. As a large sample size is required to guarantee that FI, MSI, IPW, SPE and KNN estimators reach a substantial stability, we set $n = 1000$. For KNN estimators, results based on smaller sample sizes are reported in Section S4, Supplementary Material.

Table 4 presents Monte Carlo means and standard deviations (across 5000 replications) for the estimators of the true class fractions, TCF_1, TCF_2 and TCF_3. The table also gives the means of the estimated standard deviations (of the estimators), based on the asymptotic theory.

Table 4: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when both models for $\rho_k(t,a)$ and $\pi(t,a)$ are misspecified and the sample size $n = 1000$. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
cut points $= (-1.0, -0.5)$									
True	**0.1812**	**0.1070**	**0.9817**						
FI	0.1290	0.0588	0.9888	0.0153	0.0133	0.0118	0.0170	0.0126	0.0423
MSI	0.1299	0.0592	0.9895	0.0154	0.0153	0.0131	0.0171	0.0144	0.0427
IPW	0.1231	0.0576	0.9889	0.0178	0.0211	0.0208	0.0174	0.0201	0.2878
SPE	0.1407	0.0649	0.9877	0.0173	0.0216	0.0231	0.0171	0.0207	0.0125
1NN	**0.1809**	**0.1036**	**0.9817**	0.0224	0.0304	0.0255	0.0210	0.0257	0.0180
3NN	**0.1795**	**0.0991**	**0.9814**	0.0214	0.0258	0.0197	0.0207	0.0240	0.0190
cut points $= (-1.0, 0.7)$									
True	**0.1812**	**0.8609**	**0.4469**						
FI	0.1290	0.7399	0.5850	0.0153	0.0447	0.1002	0.0170	0.0403	0.0919
MSI	0.1299	0.7423	0.5841	0.0154	0.0453	0.1008	0.0171	0.0408	0.0926
IPW	0.1231	0.7690	0.5004	0.0178	0.0902	0.2049	0.0174	0.0824	0.1844
SPE	0.1407	0.7635	0.5350	0.0173	0.0702	0.2682	0.0171	0.0646	0.2171
1NN	**0.1809**	**0.8452**	**0.4406**	0.0224	0.0622	0.1114	0.0210	0.0503	0.0895
3NN	**0.1795**	**0.8285**	**0.4339**	0.0214	0.0521	0.0882	0.0207	0.0479	0.0929
cut points $= (-1.0, 1.3)$									
True	**0.1812**	**0.9732**	**0.1171**						
FI	0.1290	0.9499	0.1900	0.0153	0.0179	0.0550	0.0170	0.0203	0.0440
MSI	0.1299	0.9516	0.1902	0.0154	0.0184	0.0552	0.0171	0.0206	0.0442
IPW	0.1231	0.9645	0.1294	0.0178	0.0519	0.1795	0.0174	0.0268	0.0898
SPE	0.1407	0.9567	0.1760	0.0173	0.0425	0.3383	0.0171	0.0311	0.2127
1NN	**0.1809**	**0.9656**	**0.1124**	0.0224	0.0218	0.0448	0.0210	0.0272	0.0544
3NN	**0.1795**	**0.9604**	**0.1086**	0.0214	0.0172	0.0338	0.0207	0.0262	0.0567
cut points $= (-0.5, 0.7)$									
True	**0.4796**	**0.7539**	**0.4469**						
FI	0.3715	0.6811	0.5850	0.0270	0.0400	0.1002	0.0244	0.0353	0.0919
MSI	0.3723	0.6831	0.5841	0.0271	0.0409	0.1008	0.0246	0.0361	0.0926
IPW	0.3547	0.7114	0.5004	0.0325	0.0883	0.2049	0.0321	0.0815	0.1844
SPE	0.3949	0.6986	0.5350	0.0318	0.0687	0.2682	0.0312	0.0637	0.2171
1NN	**0.4783**	**0.7416**	**0.4406**	0.0361	0.0610	0.1114	0.0310	0.0526	0.0895
3NN	**0.4756**	**0.7294**	**0.4339**	0.0341	0.0499	0.0882	0.0303	0.0500	0.0929
cut points $= (-0.5, 1.3)$									
True	**0.4796**	**0.8661**	**0.1171**						
FI	0.3715	0.8910	0.1900	0.0270	0.0202	0.0550	0.0244	0.0218	0.0440
MSI	0.3723	0.8924	0.1902	0.0271	0.0211	0.0552	0.0246	0.0226	0.0442
IPW	0.3547	0.9068	0.1294	0.0325	0.0535	0.1795	0.0321	0.0384	0.0898
SPE	0.3949	0.8918	0.1760	0.0318	0.0451	0.3383	0.0312	0.0368	0.2127
1NN	**0.4783**	**0.8620**	**0.1124**	0.0361	0.0349	0.0448	0.0310	0.0373	0.0544
3NN	**0.4756**	**0.8613**	**0.1086**	0.0341	0.0285	0.0338	0.0303	0.0355	0.0567
cut points $= (0.7, 1.3)$									
True	**0.9836**	**0.1122**	**0.1171**						
FI	0.9618	0.2099	0.1900	0.0122	0.0317	0.0550	0.0114	0.0263	0.0440
MSI	0.9613	0.2093	0.1902	0.0125	0.0320	0.0552	0.0116	0.0265	0.0442
IPW	0.9548	0.1955	0.1294	0.0339	0.0831	0.1795	0.0278	0.0764	0.0898
SPE	0.9582	0.1932	0.1760	0.0332	0.0618	0.3383	0.0290	0.0577	0.2127
1NN	**0.9821**	**0.1204**	**0.1124**	0.0144	0.0494	0.0448	0.0109	0.0449	0.0544
3NN	**0.9804**	**0.1319**	**0.1086**	0.0138	0.0404	0.0338	0.0108	0.0429	0.0567

The table clearly shows limitations of the (partially) parametric approaches in case of misspecified models for $\Pr(D_k = 1|T, A)$ and $\Pr(V = 1|T, A)$. More precisely, in term of bias, the FI, MSI, IPW and SPE approaches perform almost always poorly, with high distortion in almost all cases. As we mentioned in Section 2, the SPE estimators could fall outside the interval $(0, 1)$. In our simulations, in the worst case, the estimator $\widehat{\mathrm{TCF}}_{3,\mathrm{SPE}}(-1.0, -0.5)$ gives rise to 20% of the values greater than 1. Moreover, the Monte Carlo standard deviations shown in the table indicate that the SPE approach might yield unstable estimates. Finally, the misspecification also has a clear effect on the estimated standard deviations of the estimators. On the other side, the estimators 1NN and 3NN seem to perform well in terms of both bias and standard deviation. In fact, KNN estimators yield estimated values that are near to the true values. In addition, we observe that the estimator 3NN has larger bias than 1NN, but with slightly less variance.

6. AN ILLUSTRATION

We use data on epithelial ovarian cancer (EOC) extracted from the Pre-PLCO Phase II Dataset from the SPORE/Early Detection Network/Prostate, Lung, Colon and Ovarian Cancer Ovarian Validation Study. [1]

As in [16], we consider the following three classes of EOC, i.e., benign disease, early stage (I and II) and late stage (III and IV) cancer, and 12 of the 59 available biomarkers, i.e. CA125, CA153, CA72–4, Kallikrein 6 (KLK6), HE4, Chitinase (YKL40) and immune costimulatory protein–B7H4 (DD–0110), Insulin–like growth factor 2 (IGF2), Soluble mesothelin-related protein (SMRP), Spondin–2 (DD–P108), Decoy Receptor 3 (DcR3; DD–C248) and Macrophage inhibitory cytokine 1 (DD–X065). In addition, age of patients is also considered.

After cleaning for missing data, we are left 134 patients with benign disease, 67 early stage samples and 77 late stage samples. As a preliminary step of our analysis we ranked the 12 markers according to value of VUS, estimated on the complete data. The observed ordering, consistent with medical knowledge, led us to select CA125 as the test T to be used to illustrate our method.

To mimic verification bias, a subset of the complete dataset is constructed using the test T and a vector $A = (A_1, A_2)$ of two covariates, namely the marker CA153 (A_1) and age (A_2). Reasons for using CA153 as a covariate come from the medical literature that suggests that the concomitant measurement of CA153 with CA125 could be advantageous in the pre-operative discrimination of benign and malignant ovarian tumors. In this subset, T and A are known for all samples (patients), but the true status (benign, early stage or late stage) is available only for some samples, that we select according to the following mechanism. We select all samples having a value for T, A_1 and A_2 above their respective medians, i.e. 0.87, 0.30 and 45; as for the others, we apply the following selection process

$$\Pr(V = 1|T, A) = 0.05 + 0.35\mathrm{I}(T > 0.87) + 0.25\mathrm{I}(A_1 > 0.30) + 0.35\mathrm{I}(A_2 > 45),$$

leading to a marginal probability of selection equal to 0.634.

[1] The study protocol and data are publicly available at the address: https://edrn.nci.nih.gov/protocols/119-spore-edrn-pre-plco-ovarian-phase-ii-validation.

Since the test T and the covariates A_1, A_2 are heterogeneous with respect to their variances, the Mahalanobis distance is used for KNN estimators. Based on the discussion in Section 3.4, we use the selection rule (3.7) to find the size K of the neighborhood. This leads to the choice of $K = 1$ for our data. In addition, we also employ $K = 3$ for the sake of comparison with 1NN result, and produce the estimate of the ROC surface based on full data (Full estimate), displayed in Figure 1.

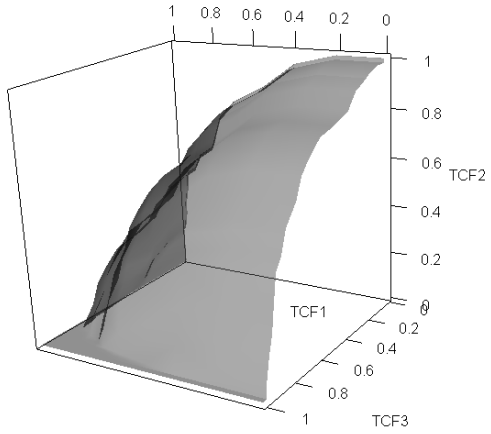
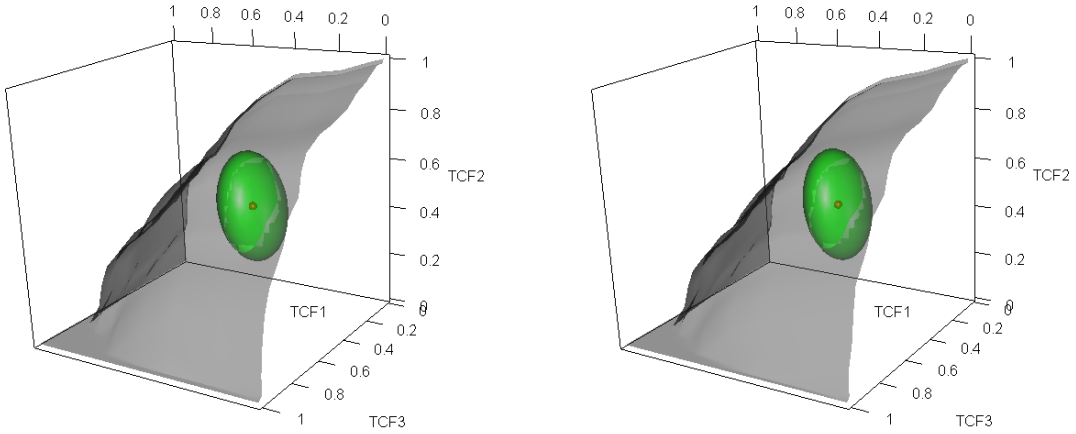


Figure 1: Estimated ROC surface for CA125, based on full data.

Figure 2 shows the 1NN and 3NN estimated ROC surfaces for the test T (CA125).



(a) 1NN (b) 3NN

Figure 2: Bias–corrected estimated ROC surfaces for CA125, based on incomplete data.

In this figure, we also give the 95% ellipsoidal confidence regions (green color) for (TCF_1, $\mathrm{TCF}_2, \mathrm{TCF}_3$) at cut points $(-0.56, 2.31)$. These regions are built using the asymptotic normality of the estimators. Compared with the Full estimate, KNN bias-corrected method

proposed in the paper appears to well behave, yielding reasonable estimates of the ROC surface with incomplete data. A closer inspection to the behavior at some chosen points can be taken by looking at Table 5.

Table 5: Comparison between Full and KNN estimates of the true class fractions for CA125, for some values of c_1 and c_2.

	Full			1NN			3NN		
(c_1, c_2)	TCF_1	TCF_2	TCF_3	TCF_1	TCF_2	TCF_3	TCF_1	TCF_2	TCF_3
(0, 0.5)	0.500	0.104	0.922	0.516	0.171	0.938	0.497	0.170	0.933
(0, 1)	0.500	0.254	0.883	0.516	0.271	0.838	0.497	0.275	0.858
(0, 2.6)	0.500	0.567	0.688	0.516	0.557	0.663	0.497	0.550	0.667
(0, 3)	0.500	0.612	0.623	0.516	0.614	0.612	0.497	0.605	0.617
(0, 4)	0.500	0.731	0.325	0.516	0.714	0.312	0.497	0.710	0.317
(0.4, 0.5)	0.694	0.030	0.922	0.688	0.043	0.938	0.670	0.040	0.933
(0.4, 1)	0.694	0.179	0.883	0.688	0.143	0.838	0.670	0.145	0.858
(0.4, 2.6)	0.694	0.493	0.688	0.688	0.429	0.663	0.670	0.420	0.667
(0.4, 3)	0.694	0.537	0.623	0.688	0.486	0.612	0.670	0.475	0.617
(0.4, 4)	0.694	0.657	0.325	0.688	0.586	0.312	0.670	0.580	0.317
(1, 2.6)	0.813	0.313	0.688	0.789	0.286	0.663	0.787	0.275	0.667
(1, 3)	0.813	0.358	0.623	0.789	0.343	0.612	0.787	0.330	0.617
(1, 4)	0.813	0.478	0.325	0.789	0.443	0.312	0.787	0.435	0.317
(2, 2.6)	0.955	0.149	0.688	0.945	0.143	0.663	0.942	0.130	0.667
(2, 3)	0.955	0.194	0.623	0.945	0.200	0.612	0.942	0.185	0.617
(2, 4)	0.955	0.313	0.325	0.945	0.300	0.312	0.942	0.290	0.317
(3.5, 4)	0.993	0.045	0.325	0.992	0.043	0.312	0.990	0.045	0.317

7. CONCLUSIONS

A general suitable strategy for reducing the effects of model misspecification in statistical inference is to resort on fully nonparametric methods. This paper proposes a nonparametric estimator of the ROC surface of a continuous diagnostic test. The estimator is based on nearest-neighbor imputation and works under MAR assumption. It represents an alternative to (partially) parametric estimators discussed in [16]. Our simulation results and the presented illustrative example show usefulness of the proposal.

Generally speaking, performances of our estimator depend on various intrinsic factors, and on some user-defined choices. Among intrinsic factors, we mention the unknown values of parameters TCF_1, TCF_2 and TCF_3 to be estimated, the rate of verified units in the sample at hand, and the nature of the unknown processes generating the observations. In particular, extreme values of the true class fractions, i.e. values close to 0 or 1, are difficult to estimate in an accurate way, especially when sample data are characterized by a low verification rate, which limits the amount of information available. On the basis of discussions in Section 3.3 (and in the last part of this section) and of simulation results in Section 5 (and in Supplementary Material), we offer some recommendations for tackling the user-defined choices. More precisely, we recommend: (a) to use the Euclidean distance, as the first choice,

and the Mahalanobis distance in case of heterogeneity among variables; (b) to keep small, from 1 to 3, say, the number of neighbors K. Our simulation results show satisfactory performances of the KNN estimator of the ROC surface when about 70 verified observations are present in the sample.

As in [1], a simple extension of our estimator, that could be used when categorical auxiliary variables are also available, is possible. Without loss of generality, we suppose that a single factor C, with m levels, is observed together with T and A. We also assume that C may be associated with both $\mathcal{D}$ and V. In this case, the sample can be divided into m strata, i.e. m groups of units sharing the same level of C. Then, for example, if the MAR assumption and first-order differentiability of the functions $\rho_k(t,a)$ and $\pi(t,a)$ hold in each stratum, a consistent and asymptotically normally distributed estimator of TCF_1 is

$$\widehat{\mathrm{TCF}}^{S}_{1,\mathrm{KNN}}(c_1) = \frac{1}{n}\sum_{j=1}^{m} n_j \widehat{\mathrm{TCF}}^{cond}_{1j,\mathrm{KNN}}(c_1),$$

where n_j denotes the size of the j-th stratum and the quantity $\widehat{\mathrm{TCF}}^{cond}_{1j,\mathrm{KNN}}(c_1)$ denotes the KNN estimator of the conditional TCF_1, i.e., the KNN estimator in (3.1) obtained from the patients in the j-th stratum. Of course, we must assume that, for every j, ratios n_j/n have finite and nonzero limits as n goes to infinity.

In our approach, the KNN method is used to estimate the probabilities $\rho_k(t,a)$ for non–verified subjects. A referee pointed out that KNN estimators might suffer from boundary effects, i.e., increases in bias when estimates are computed near the boundary of the support of the covariates. Indeed, near the boundaries, any smoothing method is less accurate, as fewer observations can be averaged, so that bias of estimators can be affected. In contrast to other nonparametric regression methods, however, KNN estimators always involve the same number of observations. Boundary effects, therefore, act on neighborhoods' sizes more than on the number of observations involved in the local fitting. For this reason, a prominent source of bias of KNN estimators is the shape at the boundary of the functions to be estimated. Steeper functions are more likely associated to a larger bias, an aspect pointing to small values of K as good choices to limit boundary effects. Moreover, it is worth noting that in the domain of our interest, i.e., evaluation of diagnostic tests, is hard to deal with test and covariate values close to the boundary of their support. More likely, one faces sparsity of data in some regions of the features space and, therefore, one has to deal with situations in which, for a fixed sample size, information brought by data on those regions is structurally low. This aspect also impacts on the neighborhoods' sizes, and probably amounts to a primary source of bias in our application contest. This remark is supported by results of some simulations that we carried out to evaluate possible bias due to boundary effects and/or sparsity of data (see Section S5, Supplementary Material). Overall, simulation results seem to show that the bias, when present, is driven more by sparsity of data issues than by boundary effects and that KNN estimators have their poorest performances on largest values of K, regardless of the position of points in the domain.

ACKNOWLEDGMENTS

This work has been supported by the grant number BIRD169208 from University of Padova, Italy. We also acknowledge the valuable suggestions from the Associated Editor and two anonymous Referees, who greatly contributed to improve presentation of the contents.

REFERENCES

[1] Adimari, G. and Chiogna, M. (2015). Nearest-neighbor estimation for ROC analysis under verification bias, *The International Journal of Biostatistics*, **11**, 109–124.

[2] Adimari, G. and Chiogna, M. (2017). Nonparametric verification bias-corrected inference for the area under the ROC curve of a continuous-scale diagnostic test, *Statistics and Its Interface*, **10**, 629–641.

[3] Alonzo, T.A. and Pepe, M.S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 173–290.

[4] Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random, *Journal of the American Statistical Association*, **89**, 81–87.

[5] Chi, Y.Y. and Zhou, X.H. (2008). Receiver operating characteristic surfaces in the presence of verification bias, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**, 1–23.

[6] Dreiseitl, S.; Ohno-Machado, L. and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**(3), 323–331.

[7] He, H. and McDermott, M.P. (2012). A robust method using propensity score stratification for correcting verification bias for binary tests, *Biostatistics*, **13**, 32–47.

[8] Hu, L.-Y.; Huang, M.-W.; Ke, S.-W. and Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets, *SpringerPlus*, **5**(1), 1304.

[9] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Wiley, New York.

[10] Nakas, C.T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems, *REVSTAT – Statistical Journal*, **12**, 43–65.

[11] Nakas, C.T. and Yiannoutsos, C.Y. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

[12] Ning, J. and Cheng, P.E. (2012). A comparison study of nonparametric imputation methods, *Statistics and Computing*, **22**, 273–285.

[13] Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

[14] Rotnitzky, A.; Faraggi, D. and Schisterman, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias, *Journal of the American Statistical Association*, **101**, 1276–1288.

[15] Scurfield, B.K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

[16] To Duc, K.; Chiogna, M. and Adimari, G. (2016). Bias-corrected methods for estimating the receiver operating characteristic surface of continuous diagnostic tests, *Electronic Journal of Statistics*, **10**(2), 3063–3113.

[17] Zhou, X.H.; Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, Wiley and Sons, New York.

REVSTAT – STATISTICAL JOURNAL

Background

Statistics Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of a scientific statistical journal called Revista de Estatística. The original language used in this publication was Portuguese and the idea behind it was to publish it, three times a year, containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided that the publication should also include papers in English. This step was taken to achieve a broader dissemination, and to encourage foreign contributors to submit their work for publication.

At the time, the Editorial Board was mainly comprised of Portuguese university professors. It is now comprised of international university faculties and this has been the first step aimed at changing the character of Revista de Estatística from a national to an international scientific journal.

We have also initiated a policy of publishing special volumes that may be thematic highlighting areas of interest or associated with scientific events in Statistics. For example, in 2001, a special issue of Revista de Estatística was published containing three volumes of extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

In 2003, the name of the Journal has been changed to REVSTAT - STATISTICAL JOURNAL, now fully published in English, with a prestigious international editorial board, aiming to become a reference scientific journal that promotes the dissemination of relevant research results in Statistics.

The editorial policy of REVSTAT Statistical Journal is mainly placed on the originality and importance of the research.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external.

The only working language allowed is English.

Abstract and Indexing Services

The REVSTAT is covered by the following abstracting/indexing services:

- Current Index to Statistics
- Google Scholar
- Mathematical Reviews® (MathSciNet®)
- Science Citation Index Expanded
- Zentralblatt für Mathematic
- Scimago Journal & Country Rank
- Scopus

Instructions to Authors

Articles must be written in English and will be submitted according to the following guidelines:

The corresponding author sends the manuscript in PDF format to the Executive Editor (revstat@ine.pt) with the Subject "New Submission to REVSTAT"; a MS#REVSTAT reference will be assigned later.

Optionally, in a mail cover letter, authors are welcome to suggest one of the Editors or Associate Editors, whose opinion may be considered suitable to be taken into account.

The submitted manuscript should be original and not have been previously published nor about to be published elsewhere in any form or language, avoiding concerns about self-plagiarism'.

Content published in this journal is peer-reviewed (Single Blind).

All research articles will be refereed by at least two researchers, including one from the Editorial Board unless the submitted manuscript is judged unsuitable for REVSTAT or does not contain substantial methodological novelty, in which case is desk rejected.

Manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and with less than 25 pages. Figures (minimum of 300dpi) will be reproduced online in colours, if produced this way; however, authors should take into account that the printed version is always in black and grey tones.

The first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s) and a summary of fewer than one hundred words, followed by a maximum of six keywords and the AMS 2000 subject classification.

Authors are encouraged to submit articles using LaTeX, in the REVSTAT style, which is available at the LaTeX2e MACROS webpage.

References about the format and other useful information on the submission are available in the LaTeX2e Templates page.

Acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text.

Supplementary files (in REVSTAT style) may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Any contact with REVSTAT must always contain the assigned REVSTAT reference number.

Accepted papers

Authors of accepted papers are requested to provide the LaTex files to the Secretary of the REVSTAT revstat@ine.pt. The authors should also mention if figure files were included, and submit electronic figures separately in .gif, .jpg, .png or .pdf format. Figures must be a minimum of 300dpi.

Copyright and reprints

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information, namely through the Statistics Portugal website (http://www.ine.pt).

After assigning copyright, authors may use their own material in other publications provided that REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

Editorial Board