Spanish Journal of Statistics

VOLUME 2, NUMBER 1, 2020

INE

NIPO: 096-20-014-2
ISSN: 2695 - 9070
DOI: https://doi.org//10.37830/SJS.2020.1

Spanish Journal of Statistics

Published by the Spanish National Statistical Institute

Spanish Journal of Statistics

Volume 2, Number 1, 2020

Contents

Spanish Journal of Statistics
Vol. 2 No. 1, 2020, Pages 5 – 6
doi: https://doi.org/10.37830/SJS.2020.1.01

Editorial

Presentation of Volume 2, 1, 2020

José María Sarabia

Editor-in-Chief Spanish Journal of Statistics

Dear readers and dear statistical community:

It is my pleasure to present the second volume of the Spanish Journal of Statistics, corresponding to the year 2020. This volume includes four papers: two papers in the General Section and another two papers in the Official Statistics section.

The first paper is titled "Financial and Actuarial Properties of the Beta-Pareto as a Long-Tail Distribution" by Emilio Gómez-Déniz and Enrique Calderín-Ojeda. This paper presents several properties of the Beta-Pareto distribution, which might be extremely useful in Economics, and in Financial and Actuarial fields. These properties are mainly related to the analysis of the tail of the distribution that makes it a candidate model for fitting actuarial data with extreme observations.

The second article of the general section is titled: "The Gamma-Chen distribution: a new family of distributions with applications", by Lucas David R. Reis, Gauss M. Cordeiro and Maria do Carmo S. Lim. This paper presents the gamma-Chen distribution and derive some of its mathematical and statistical properties. The authors present empirical evidence that support that this new distribution is better than other relevant statistical distributions using several data sets.

The following two papers cover relevant aspects of Official Statistics. The first of them is: "Towards a modular end-to-end statistical production process with mobile network data", by David Salgado, Luis Sanguiao, Bogdan Oancea, Sandra Barragán, and Marian Necula. In the context of the European Statistical System (ESS), the authors introduce the so-called ESS Reference Methodological Framework for Mobile Network Data with the first modular and evolvable statistical process, which involves five different aspects: the geolocation of mobile devices; the deduplication of mobile devices; the statistical filtering to identify the target population; the aggregation into territorial units, and inference to the target population. The proposal methodology is illustrated with synthetic data generated from a network event data simulator developed for these purposes.

The second paper on Official Statistics is titled: "Commonly used methods for measuring output quality of multisource statistics", by Ton de Waal, Arnout van Delden and Sander Scholtus. The estimation of output quality based on sample surveys is well established. The paper presents results of the ESSnet project Quality of Multisource Statistics that studied methods to estimate output quality. The authors distinguish three main groups of methods: scoring methods, (re)sampling methods and methods based on parametric modeling. All of these methodologies are developed and

Published by the Spanish National Statistical Institute

discussed in detail within the paper.

I would like to conclude by thanking all the authors of this volume for choosing our journal as a means of disseminating their work.I am also extremely grateful to the editors and the reviewers of the papers for the work devoted to the journal, which is key to maintain a high scientific quality standard.

Spanish Journal of Statistics
Vol. 2 No. 1 2020, Pages 7–21
doi:https://doi.org/10.37830/SJS.2020.1.02

Regular Article

Financial and Actuarial Properties of the Beta-Pareto as a Long-Tail Distribution

Emilio Gómez-Déniz[1], Enrique Calderín-Ojeda[2]
[1]Department of Quantitative Methods and TIDES Institute - University of Las Palmas de Gran Canaria
emilio.gomez-deniz@ulpgc.es
[2]Centre for Actuarial Studies, Department of Economics - The University of Melbourne, Australia
enrique.calderin@unimelb.edu.au

Received: December 18, 2020. Accepted: March 1, 2021.

Abstract: Undoubtedly, the single parameter Pareto distribution is one of the most attractive distribution in statistics; a power-law probability distribution that is found in a large number of real-world situations inside and outside the field of economics. Furthermore, it is usually used as a basis for excess of loss quotations as it gives a pretty good description of the random behaviour of large losses. In this paper, we provide properties of the Beta-Pareto distribution which can be useful in Economics, and in Financial and Actuarial fields, mainly related to the analysis of the tail of the distribution that makes it a candidate model for fitting actuarial data with extreme observations. As empirical applications two well-known data sources considered in general insurance are used to account for the suitability of the model.

Keywords: insurance, Beta-Pareto distribution, Danish and Norweigian data; Pareto distribution, right tail

MSC: 62E10, 62F10, 62P05

1 Introduction

Probability distributions such as the exponential, Pareto, gamma, lognormal and Weibull are frequently used in survival analysis, engineering applications and, specifically, in actuarial statistics to model losses in insurance and finance. Besides, other parametric families, e.g. Pareto and lognormal distributions are particularly appropriate to describe data that include large losses (see Boland, 2007, p. IX). More precisely, the study of the right tail of the distribution is an important issue in order to not underestimate the size of large claims. This is for example the case of the suitability of the Pareto distribution to describe fire claim data (Rolski et al., 1999, p. 49). This is also common in defaulted loans in banking sector. It is needless to say that, due to the simple form of its survival function, the Pareto distribution is commonly used in these scenarios.

It is well-known that the classical Pareto distribution (for a detailed discussion of the Pareto distribution see Arnold, 1983) with scale parameter $\sigma > 0$ and shape parameter $\theta > 0$ with probability density function

$$g(x) = \frac{\theta \sigma^{\theta}}{x^{\theta+1}}, \quad x > \sigma > 0, \ \theta > 0 \tag{1}$$

and survival function

$$\bar{G}(x) = \left(\frac{\sigma}{x}\right)^{\theta}, \quad x > \sigma > 0, \ \theta > 0 \tag{2}$$

has been proved to be useful as predicting tools in different socioeconomic contexts such as income (Mandelbrot, 1960), insurance (for applications of the Pareto model in rating property excess-of-loss reinsurance, the Pareto distribution has been used by Boyd, 1988, Hesselager, 1993 and Brazauskas and Serfling, 2003, among others), city size (Rosen and Resnick, 1980) and also in other fields as queue service (Harris, 1968). A thorough review of the reinsurance issue can be viewed in Albrecher et al. (2017). Perhaps, one of the most important characteristics of the Pareto distribution is that it produces a better extrapolation from the observed data when pricing high excess layers, in situations where there is little or no experience. In this regard, its efficacy dealing with inflation in claims and with the effect of deductibles and excess-of-loss levels for reinsurance has been demonstrated. Henceforward, a continuous random variable that follows the Pareto distribution with pdf as in (1) will be denoted as $X \sim Par(\theta, \sigma)$. Surely, one of the advantages of working with this probability distribution is, similarly to the exponential case, the simple form of its survival function which allows us to easily derive interesting properties. For example, it is straightforward to observe that if $X \sim Par(\theta, \sigma)$, then $\tau X \sim Par(\theta, \tau\sigma)$, $\tau > 0$. This property is useful when dealing with proportional reinsurance and also with claims inflation. Furthermore, if $X > Z$ we have that $X - Z \sim Par(\theta, Z)$. That is, the excess of X over Z is also Pareto (see Boland, 2007, p. 39).

In the last decades, a lot of attempts have been made to achieve generalizations of the classical Pareto distributions. Many of these new models try to obtain better fits to empirical data related to city populations and insurance losses. Some of them are the Stoppa's generalized Pareto distribution (see Stoppa, 1990 and Kleiber and Kotz, 2003); the Beta-Pareto distribution due to Akinsete et al. (2008); the Pareto positive stable distribution provided by Sarabia and Prieto (2009) and the recently proposals of Gómez-Déniz and Calderín (2014), Gómez-Déniz and Calderín (2015) and Ghitany et al. (2018). In general insurance settings and also in city size, mainly seeking to better adjust the right tail of the distribution, the recently proposed composite models have also made use of the Pareto distribution in their formulation and, therefore, can be considered as generalizations of the latter distribution (see Scollnik, 2007, Calderín-Ojeda and Kwok, 2016 and Calderín-Ojeda, 2016).

In actuarial settings, the single parameter Pareto distribution has been largely considered against other probability distributions, not only for its nice properties, but also for its appropriateness to describe the claims size. When modeling losses, there is widely concern on the frequencies and sizes of large claims, in particular, the study of the right tail of the distribution. On this subject, the single parameter Pareto distribution gives a good description of the random behaviour of large losses. See, for instance Boyd (1988) and Brazauskas and Serfling (2003), among others.

In this paper, we pay special attention to one generalization of the Pareto distribution, built from the scheme proposed by Jones (2004) and which was considered by Akinsete et al. (2008), the Beta-Pareto distribution. We will see that this distribution can be used as a basis for excess of loss quotations, and similarly to the Pareto distribution (see for instance, Rytgaard, 1990), providing a good description of the random behaviour of large losses.

In order to make the paper self-contained, some of the basic properties provided in Akinsete et al. (2008) are again reproduced here. Furthermore, new properties that are important in financial and actuarial applications are also provided. In particular, we give expressions for the limited expected values, integrated tail distribution and mean excess function, among others. Finally, the performance of the model is examined by using two well-known examples of real claims data in actuarial statistics.

The remainder of the paper is organized as follows. Basic background of the Beta-Pareto distribution is shown in Section 2. We pay special attention here to some of its more basic properties and the estimation of the parameters of the distribution by maximum likelihood method. Section 3 discusses properties related to the right tail of the distribution that are very relevant in the field of reinsurance. Two numerical applications are shown in Section 4 and conclusions are provided in the last Section.

2 Preliminaries

In an appealing paper Jones (2004) proposed a method to add more flexibility to a parent probability function by starting with a distribution function G (in that work author only considered symmetric distributions but the methodology is applicable to any distribution function) and generating the new one by adding two parameters in order to include skewness and vary the tail weight. The method is based on order statistics by using the classical Beta distribution. Specifically, for a probability density function $g(x)$ with distribution function $G(x)$ and survival function $\bar{G}(x) = 1-G(x)$, the author studied the family of probability distributions given by

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} g(x)\,[G(x)]^{\alpha-1}\left[\bar{G}(x)\right]^{\beta-1}, \quad \alpha>0,\ \beta>0, \tag{3}$$

where $B(\cdot,\cdot)$ is the Euler Beta function.

When (3) is applied to (2), the probability density function of the Beta-Pareto distribution studied in Akinsete et al. (2008) is obtained with analytical expression given by

$$f(x) = \frac{1}{B(\alpha,\beta)}\frac{\theta}{x}\left(\frac{\sigma}{x}\right)^{\alpha\theta}\left[1-\left(\frac{\sigma}{x}\right)^{\theta}\right]^{\beta-1}, \quad x>\sigma. \tag{4}$$

This distribution includes a wide range of curve shapes as illustrated by the density plots shown in Figure 1.

Some special cases of the distribution provided in (4) are given below:

- If $\alpha=\beta=1$ we get the classical Pareto distribution given in (1).
- The case $\beta=1$ reduces to a $Par(\alpha\theta,\sigma)$.
- The case $\alpha=1$ to the Stoppa distribution (see Stoppa, 1990 and Kleiber and Kotz, 2003).

Hereafter, a random variable X that follows the probability density function (4) will be denoted as $X\sim BP(\alpha,\beta,\theta,\sigma)$.

Simple computations show that the distribution is unimodal with modal value located at

$$x = \sigma\left[\frac{1+(\alpha+\beta-1)\theta}{1+\alpha\,\theta}\right]^{1/\theta}.$$

All moments of order $r>0$ exist and they are given by,

$$E(X^r) = \frac{\sigma\Gamma(\alpha+\beta)\Gamma(\alpha-r/\theta)}{\Gamma(\alpha)\Gamma(\alpha+\beta-r/\theta)}.$$

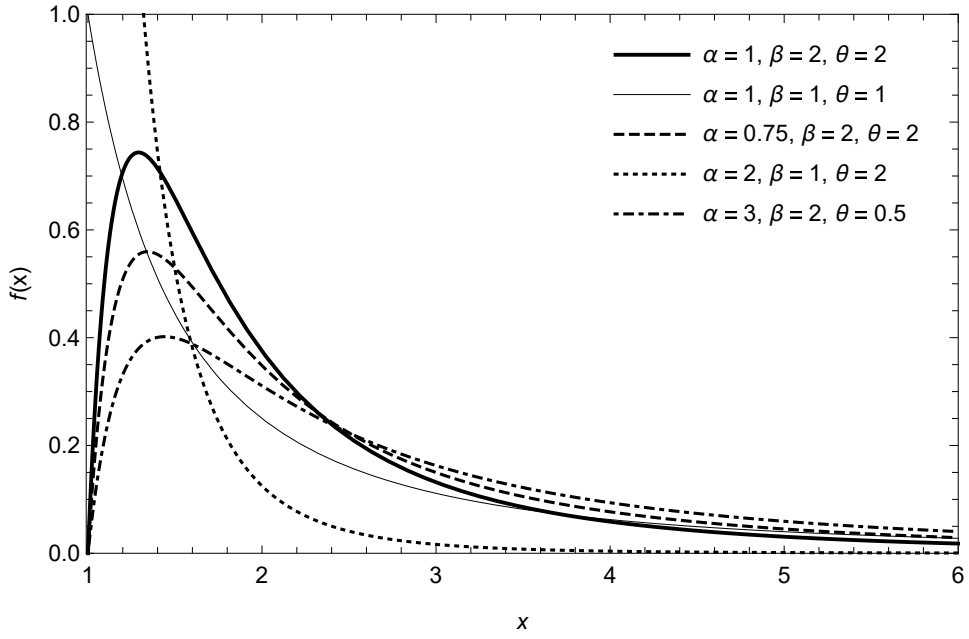

Figure 1: Graphs of the probability density function (4) for different values of parameter α, β and θ assuming in all the cases $\sigma = 1$.

In particular, the mean value is given by

$$E(X) = \frac{\sigma B(\alpha^*, \beta)}{B(\alpha, \beta)}, \quad \alpha > \frac{1}{\theta}, \tag{5}$$

where $\alpha^* = \alpha - 1/\theta$.

The variance is easily computed and it can be seen that the mean value increases with α and β (in this case when $\theta > 1$) and decreases with θ.

One of the advantage of this distribution is its simple form of its survival function, which is expressed in terms of the incomplete beta function ratio, a special function available in many statistical software and spreadsheet packages. That is, the survival function of the random variable following the probability distribution (4) results

$$\bar{F}(x) = I_{z(x)}(\alpha, \beta), \tag{6}$$

where $z(x) = (\sigma/x)^\theta$ and $I_u(\cdot,\cdot)$ represents the incomplete beta function ratio, given by

$$I_c(a, b) = \frac{1}{B(a, b)} \int_0^c t^{a-1}(1-t)^{b-1}\, dt.$$

Furthermore, the hazard rate function, $h(x) = f(x)/\bar{F}(x)$, has also a simple and closed-form expression.

Below, the hazard rate function has been plotted in Figure 2 for different values of the parameters α, β and θ and assuming again that $\sigma = 1$. It is observable that the hazard rate function is monotonically decreasing when $\beta \leq 1$. When $\beta > 1$ the hazard rate function has inverted-U shape.

Also, if $\beta < 1$ the distribution is log-convex, i.e. $(\log f(x))'' > 0$. Finally, closed-form expression for the entropy of the distribution can be viewed in Akinsete et al. (2008)

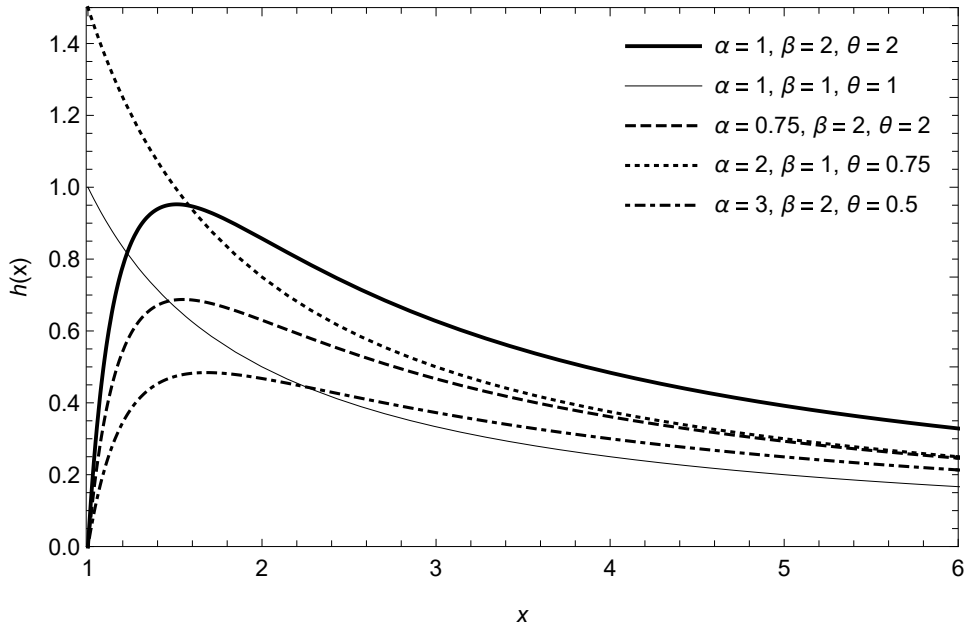

Figure 2: Graphs of the hazard rate function for different values of parameter α, β and θ, assuming again $\sigma = 1$.

2.1 Transformations

Let

$$X = \sigma(1 - Z)^{-1/\theta},$$

then, it is easy to see that the random variable Z follows a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. This change of variable facilitates computations on properties of the BP distribution studied here.

2.2 Estimation

In this subsection, we show how to estimate the parameters of the distribution. For that reason, let us assume that $\{x_1, x_2, \ldots, x_n\}$ is a random sample selected from the distribution (4) and also assume that $\sigma = \min\{x_i\}$, $i = 1, \ldots, n$. By using the first three moments, numerical computation can be carried out to obtain the moment estimates of the distribution. Alternatively, by using the maximum likelihood method, the likelihood function is given by

$$\begin{aligned} \ell(\boldsymbol{\omega}; \tilde{x}) &= n[\log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) + \log\theta + \alpha\theta\log\sigma] \\ &\quad -\alpha\theta\sum_{i=1}^{n}\log x_i + (\beta-1)\sum_{i=1}^{n}\log[1 - z(x_i)]. \end{aligned}$$

The maximum likelihood estimates (MLEs) $\widehat{\boldsymbol{\omega}} = (\widehat{\alpha}, \widehat{\beta}, \widehat{\theta})$, of the parameters $\boldsymbol{\omega} = (\alpha, \beta, \theta)$ are obtained by solving the score equations

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \alpha} &= n[\psi(\alpha+\beta) - \psi(\alpha) + \theta \log \sigma] - \theta \sum_{i=1}^{n} \log x_i = 0, \quad (7)\\
\frac{\partial \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \beta} &= n[\psi(\alpha+\beta) - \psi(\beta)] + \sum_{i=1}^{n} \log[1 - z(x_i)] = 0, \quad (8)\\
\frac{\partial \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \theta} &= n\left(\frac{1}{\theta} + \alpha \log \sigma\right) - \alpha \sum_{i=1}^{n} \log x_i \\
&\quad + (\beta - 1) \sum_{i=1}^{n} \left(\frac{\sigma}{x_i}\right)^{\theta} \frac{\log(\sigma/x_i)}{1 - z(x_i)} = 0, \quad (9)
\end{aligned}$$

where $\psi(\cdot)$ gives the derivative of the digamma function (the logarithm of the gamma function).

Observe that from equation (7) we get

$$\theta = \frac{n[\psi(\alpha+\beta) - \psi(\alpha)]}{\sum_{i=1}^{n} \log x_i - n \log \sigma},$$

which can be plugged into equations (8) and (9) in order to derive system of equations which only depends on two parameters and that can be solved by a numerical method such as Newton-Raphson.

The second partial derivatives are as follows.

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \alpha^2} &= n[\psi_1(\alpha+\beta) - \psi_1(\alpha)],\\
\frac{\partial^2 \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \alpha \partial \beta} &= n\psi_1(\alpha+\beta),\\
\frac{\partial^2 \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \alpha \partial \theta} &= \log \theta - \sum_{i=1}^{n} \log x_i,\\
\frac{\partial^2 \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \beta^2} &= n[\psi_1(\alpha+\beta) - \psi_1(\beta)],\\
\frac{\partial^2 \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \beta \partial \theta} &= -\sum_{i=1}^{n} \left(\frac{\sigma}{x_i}\right)^{\theta} \frac{\log(\sigma/x_i)}{1 - z(x_i)},\\
\frac{\partial^2 \ell(\boldsymbol{\omega}; \tilde{x})}{\partial \theta^2} &= -\frac{n}{\theta^2} - (\beta - 1) \sum_{i=1}^{n} \frac{(\sigma/x_i)^{\theta} \log^2(\sigma/x_i)}{[1 - z(x_i)]^2}.
\end{aligned}$$

Once the parameters have been estimated the entries of the expected Fisher's information matrix, $\mathcal{I}(\widehat{\boldsymbol{\omega}})$, can be approximated by the following expressions

$$\begin{aligned}
\mathcal{I}_{11}(\widehat{\boldsymbol{\omega}}) &= -n\Big[\psi_1(\widehat{\alpha}+\widehat{\beta})-\psi_1(\widehat{\alpha})\Big],\\
\mathcal{I}_{12}(\widehat{\boldsymbol{\omega}}) &= -n\psi_1(\widehat{\alpha}+\widehat{\beta}),\\
\mathcal{I}_{13}(\widehat{\boldsymbol{\omega}}) &\approx -\log\widehat{\theta}-\sum_{i=1}^{n}\log x_i,\\
\mathcal{I}_{22}(\widehat{\boldsymbol{\omega}}) &= -n\Big[\psi_1(\widehat{\alpha}+\widehat{\beta})-\psi_1(\widehat{\beta})\Big],\\
\mathcal{I}_{23}(\widehat{\boldsymbol{\omega}}) &\approx -\sum_{i=1}^{n}\left(\frac{\sigma}{x_i}\right)^{\widehat{\theta}}\frac{\log(\sigma/x_i)}{1-(\sigma/x_i)^{\widehat{\theta}}},\\
\mathcal{I}_{33}(\widehat{\boldsymbol{\omega}}) &\approx \frac{n}{\widehat{\theta}^2}+(\widehat{\beta}-1)\sum_{i=1}^{n}\frac{(\sigma/x_i)^{\widehat{\theta}}\log^2(\sigma/x_i)}{\Big[1-(\sigma/x_i)^{\widehat{\theta}}\Big]^2}.
\end{aligned}$$

3 Tail of the distribution and related issues

As it was previously mentioned, a random variable with non-negative support, such as the classic Pareto distribution, is commonly used in insurance to model the amount of claims (losses). In this sense, the size of the distribution tail is of vital importance in actuarial and financial scnearios, if it is desired that the chosen model allows to capture amounts sufficiently far from the start of the distribution support, that is, extreme values. Consequently, the use of heavy right-tailed distributions such as the Pareto, lognormal and Weibull (with shape parameter smaller than 1) distributions, among others, have been employed to model losses in motor third-party liability insurance, fire insurance or catastrophe insurance.

3.1 Right tail of the BP distribution

It is already known that any probability distribution, that is specified through its cumulative distribution function $F(x)$ on the real line, is heavy right-tailed (see Rolski et al., 1999) if $\limsup_{x\to\infty}(-\log\bar{F}(x)/x)=0$. Observe that $-\log\bar{F}(x)$ is the hazard function of $F(x)$. Next result shows that the BP is a heavy tail distribution.

Proposition 1. *The cumulative distribution function $F(x)$ of the Beta-Pareto distribution is a heavy tail distribution.*

Proof. We have that

$$\begin{aligned}
\limsup_{x\to\infty}\frac{1}{x}\log\bar{F}(x) &= \limsup_{x\to\infty}\frac{1}{x}\log\left[\frac{1}{B(\alpha,\beta)}\int_0^{z(x)}t^{\alpha-1}(1-t)^{\beta-1}\right]\\
&= -\limsup_{x\to\infty}\frac{\theta\,[1-z(x)]^{\beta-1}}{\sigma B(\alpha,\beta)I_{z(x)}(\alpha,\beta)}\left(\frac{\sigma}{x}\right)^{\theta(\alpha+1)}\\
&= -\limsup_{x\to\infty}\frac{\theta}{xB(\alpha,\beta)}\left[\alpha+1-\frac{(\beta-1)\sigma^{\theta}}{1-z(x)}\right]=0,
\end{aligned}$$

where we have applied twice L'Hospital rule and the Fundamental Theorem of Calculus. Hence the result. □

Corollary 1. *It is verified that* $\limsup_{x\to\infty} e^{sx}\bar{F}(x) = \infty$, $x > \sigma$, $s > 0$.

Proof. This is a direct consequence of Proposition 1. □

Therefore, as a long-tailed distribution is also heavy right-tailed, the Beta-Pareto distribution introduced in this manuscript is heavy right-tailed.

An important issue in extreme value theory is the regular variation (see Bingham, 1987 and Konstantinides, 2018). This is, a fexible description of the variation of some function according to the polynomial form of the type $x^{-\delta}+o(x^{-\delta})$, $\delta > 0$. This concept is formalized in the following definition.

Definition 1. *A distribution function (measurable function) is called regular varying at infinity with index* $-\delta$ *if it holds*

$$\lim_{x\to\infty} \frac{\bar{F}(\tau x)}{\bar{F}(x)} = \tau^{-\delta},$$

where $\tau > 0$ *and the parameter* $\delta \geq 0$ *is called the tail index.*

Next theorem establishes that the survival function given in (6) is a regular variation Lebesgue measure.

Proposition 2. *The survival function given in (6) is a survival function with regularly varying tails.*

Proof. Let us firstly consider the survival function given in (6). Then, after applying L'Hospital rule and Fundamental Theorem of Calculus we get

$$\begin{aligned}\limsup_{x\to\infty} \frac{\bar{F}(\tau x)}{\bar{F}(x)} &= \limsup_{x\to\infty} \frac{\int_0^{z(\tau x)} t^{\alpha-1}(1-t)^{\beta-1}\,dt}{\int_0^{z(x)} t^{\alpha-1}(1-t)^{\beta-1}\,dt} \\ &= \limsup_{x\to\infty} \frac{t(\theta/\sigma)(\sigma/(\tau x))^{\theta(\alpha+1)}[1-z(\tau x)]^{\beta-1}}{(\theta/\sigma)(\sigma/x)^{\theta(\alpha+1)}[1-z(x)]^{\beta-1}} = \tau^{-\theta\,\alpha},\end{aligned}$$

and taking into account that $\theta, \alpha > 0$ the result follows. □

An immediate consequence of the previous result is the following (see Jessen and Mikosch, 2006).

Corollary 2. *If* $X, X_1, \ldots, X_n$ *are iid random variables with common survival function given by (6) and* $S_n = \sum_{i=1}^n X_i$, $n \geq 1$, *then*

$$\Pr(S_n > x) \sim \Pr(X > x) \quad as \quad x \to \infty.$$

Thus, if $X, X_1, \ldots, X_n$ are iid random variables with common survival function given by (6) and $S_n = \sum_{i=1}^n X_i$, $n \geq 1$, then

$$\Pr(S_n > x) \sim \Pr(X > x) \quad \text{as} \quad x \to \infty.$$

Therefore, if $P_n = \max_{i=1,\ldots,n} X_i$, $n \geq 1$, we have that

$$\Pr(S_n > x) \sim n\Pr(X > x) \sim \Pr(P_n > x).$$

This means that for large x the event $\{S_n > x\}$ is due to the event $\{P_n > x\}$. Therefore, exceedances of high thresholds by the sum S_n are due to the exceedance of this threshold by the largest value in the sample.

The integrated tail distribution or equilibrium distribution (see for example Yang, 2004), given by

$$F_I(x) = \frac{1}{E(X)} \int_\sigma^x \bar{F}(y)\, dy.$$

is an important concept that often appears in insurance and many other applied probability models. For the BP distribution studied in this work, the integrated tail distribution can be written as a closed-form expression as it is given in the following Proposition.

Proposition 3. *Let X be a random variable that follows the probability density function given in (4). Then, the integrated tail distribution of this random variable is given by*

$$F_I(x) = \frac{B(\alpha,\beta)}{B(\alpha^*,\beta)} \left[\frac{x}{\sigma} I_{z(x)}(\alpha,\beta) - I_1(\alpha,\beta) \right] + I_1(\alpha^*,\beta) - I_{z(x)}(\alpha^*,\beta). \tag{10}$$

Proof. First, we make the change of variable $u = (\sigma/y)^\theta$ by obtaining that

$$\int_\sigma^x \bar{F}(y)\, dy = -\frac{\sigma}{\theta} \int_1^{z(y)} u^{-1-1/\theta} I_u(\alpha,\beta)\, du$$

Now, using the indefinite integration of power functions of the beta incomplete ratio function given by[1]

$$\int u^{r-1} I_u(s,t)\, du = \frac{u^r}{r} I_u(s,t) - \frac{\Gamma(s+t)\Gamma(s+r)}{r\Gamma(s)\Gamma(s+t+r)} I_u(s+r,t)$$

and by using (5) and Fundamental Theorem of Calculus, we get the result after some computations. □

3.2 Actuarial tools

The surplus process of an insurance portfolio is defined as the wealth obtained by the premium payments minus the reimbursements made at the times of claims. When this process becomes negative (if ever), we say that ruin has occurred. Let $\{U(t)\}_{t\geq 0}$ be a classical continuous time surplus process, the surplus process at time t given the initial surplus $u = U(0)$, the dynamic of $\{U(t)\}_{t\geq 0}$ is given by

$$U(t) = u + c\,t - S(t),$$

where $S(t) = \sum_{i=1}^{N(t)} X_i$ is the aggregate claim amount up to time t and $S(t) = 0$ if $N(t) = 0$. Here, $u \geq 0$ is the insurer's initial risk surplus at $t = 0$ and $c = (1+\theta)\alpha\mu$ is the insurer's rate of premium income per unit time with loading factor $\rho \geq 0$. Here the random variables $\{X_i\}$ are independent and identically distributed random variables with $E(X_i) = \mu$.

Under the classical model of ruin theory (Yang, 2004) and assuming a positive security loading, ρ, for the claim size distributions with regularly varying tails it is known that by using (10), an approximation of the probability of ultimate ruin,

$$\psi(u) = \Pr[U(t) < 0 \text{ for some } t > 0 | U(0) = u].$$

[1] See The Wolfram functions site (`https://functions.wolfram.com`)

can be obtained. This asymptotic approximation of the ruin function is given by

$$\psi(u) \sim \frac{1}{\rho}\bar{F}_I(u), \quad u \to \infty,$$

where $\bar{F}_I(u) = 1 - F_I(u)$.

On the other hand, let the random variable X represent either a policy limit or reinsurance deductible (from an insurer's perspective); then the limited expected value function L of X with cdf $F(x)$, is defined by

$$L(x) = E[\min(X, x)] = \int_{\sigma}^{x} y\, dF(y) + x\bar{F}(x), \tag{11}$$

which is the expectation of the cdf $F(x)$ truncated at this point. In other words, it represents the expected amount per claim retained by the insured on a policy with a fixed amount deductible of x.

A variant of this last expression is given by

$$E[\min(N, \max(0, X - M))] = \int_{M}^{M+N} (x - M) f(x)\, dx + N\bar{F}(M + N), \tag{12}$$

which represents the expected cost per claim to the reinsurance layer when the losses excess of $M > \sigma$ subject to a maximum of $N > M$.

The following result, concerning to the classical Beta distribution, is useful to derive the Propositions which will be given later in order to calculate the limited expected value function for the Beta-Pareto distribution.

Proposition 4. *Let $h(y)$ the probability density function of a classical Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. Then, it is verified that,*

$$\int_0^s (1-y)^r h(y)\, dy = \frac{1}{B(\alpha,\beta)} I_s(\alpha, \beta + r), \tag{13}$$

$$\int_s^{s+t} (1-y)^r h(y)\, dy = \frac{1}{B(\alpha,\beta)} [I_{s+t}(\alpha, \beta + r) - I_s(\alpha, \beta + r)]. \tag{14}$$

Proof. It is straightforward. □

Proposition 5. *Let X be a random variable denoting the individual claim size taking values only for individual claims greater than $M > \sigma$. Let us also assumed that X follows the probability density function (4). Then the expected cost per claim of the reinsurance layer when the losses excess of $M > \sigma$ is given by*

$$L(x) = \frac{\sigma B(\alpha^*, \beta)}{B(\alpha, \beta)} \left[1 - I_{z(M)}(\alpha^*, \beta)\right] + M I_{z(M)}(\alpha, \beta), \quad \alpha > \frac{1}{\theta}. \tag{15}$$

Proof. By taking (11), making the change of variable $u = 1 - z(y)$ and using (13) we get the result after some algebra. □

Proposition 6. *Let X be a random variable denoting the individual claim size taking values only for individual claims greater than $M > \sigma$. Let us also assumed that X follows the probability density function*

(4). Then the expected cost per claim of the reinsurance layer when the losses excess of $M > \sigma$ subject to a maximum of $N > M$ is given by

$$\begin{aligned} L(x) &= \frac{\sigma}{B(\alpha,\beta)}\Big\{B(\alpha^*,1+\beta)\Big[I_{z(M)}(\alpha^*,1+\beta)-I_{z(M+N)}(\alpha^*,1+\beta)\Big] \\ &\quad +B(1+\alpha^*,\beta)\Big[I_{z(M)}(1+\alpha^*,\beta)-I_{z(M+N)}(1+\alpha^*,\beta)\Big]\Big\} \\ &\quad +(M+N)I_{z(M+N)}(\alpha,\beta)-MI_{z(M)}(\alpha,\beta), \quad \alpha>\frac{1}{\theta}. \end{aligned}$$

Proof. The proof is similar to that in Propostion 5 but using now (12) and (14). □

3.3 Mean excess function

The failure rate of the integrated tail distribution, which is given by $\gamma_I(x) = \bar{F}(x)/\int_x^\infty \bar{F}(y)\,dy$ is also obtained in closed-form. Furthermore, the reciprocal of $\gamma_I(x)$ is the mean residual life that can be easily derived. For a claim amount random variable X, the mean excess function or mean residual life function is the expected payment per claim for a policy with a fixed amount deductible of $x > 0$, where claims with amounts less than or equal to x are completely ignored. Then,

$$e(x) = E(X-x|X>x) = \frac{1}{\bar{F}(x)}\int_x^\infty \bar{F}(u)\,du. \tag{16}$$

Next result gives the mean excess function of the BP distribution in a closed-form expression.

Proposition 7. *The mean excess function of the BP distribution is given by*

$$e(x) = \frac{\sigma^2 B(\alpha^*,\beta)I_{z(x)}(\alpha^*,\beta)}{B(\alpha,\beta)I_{z(x)}(\alpha,\beta)} - x. \tag{17}$$

Proof. Using the expression

$$e(x) = \frac{E(X)-L(x)}{\bar{F}(x)},$$

which relates the mean excess function given in (16) with the limited expected value function (see Hogg and Klugman, 1984, p. 59), the result follows by using and (5), (6), (15) and a some little algebra. □

Figure 3 shows the mean residual life function (16) for special cases of parameters. It can be seen that this function can be increasing, decreasing, unimodal or anti-unimodal.

4 Numerical application

Two well-known datasets in the actuarial literature will be used here to analyze hoe the BP distribution works. The first dataset deals with large losses in a fire insurance portfolio in Denmark. These dataset include 2157 losses over 1 million Danish Krone in the years 1980-1990. A detailed statistical analysis of this set of data can be seen in McNeil (1997) in Albrecher et al. (2017) and also in Embrechts et al. (1997). It can be found in the *R* package `CASdatasets` collected at *Copenhagen Reinsurance*. The second dataset is `norfire` comprises 9181 fire losses over the period 1972 to 1992 from an unknown Norwegian company. A priority of 500 thousands of Norwegian Krone (NKR) (if

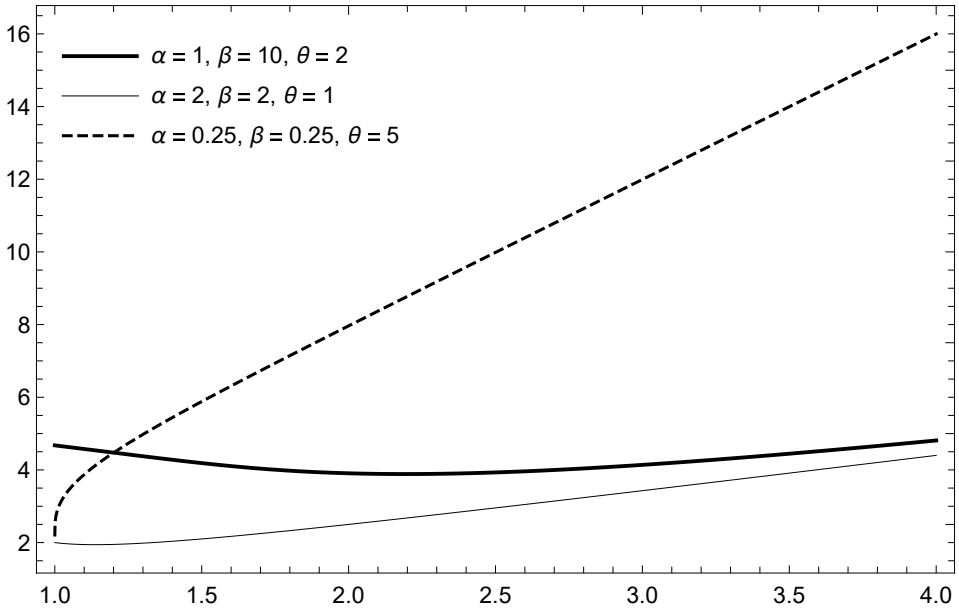

Figure 3: Mean residual life function of BP distribution for selected values of parameters when $\sigma = 1$.

this amount is exceeded, the reinsurer becomes liable to pay) was applied to obtain this dataset. This set of data is is also available in the *R* package `CASdatasets`.

Below in Table 1, parameter estimates and their corresponding *p*-values together with the negative value of the maximum likelihood function (NLL) evaluated at the maximum likelihood estimates for the two datasets considered are shown. Also the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) test statistics are displayed. As judged by the corresponding *p*-values, the BP distribution is not rejected for neither of the tests for the Danish dataset. However, for the Norwegian set of data, the BP distribution is rejected at the 5% significance level.

Danish dataset						
BP	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\theta}$	NLL	KS	AD
$\sigma = 0.999$	32.113	1.157	0.046	3341.895	0.027	0.00051
	(< 0.001)	(< 0.001)	(< 0.001)		(0.398)	(0.545)
Norwegian dataset						
BP	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\theta}$	NLL	KS	AD
$\sigma = 490$	77.090	1.549	0.020	20979.635	0.039	0.0011
	(< 0.001)	(< 0.001)	(< 0.001)		0.038	0.057

Table 1: Parameter estimates and their *p*-values (in brackets), negative of the maximum of the log likelihood function, Kolmogorov-Smirnov and Anderson-Darling test for the BP distribution.

These results are confirmed in Figure 4 where the empirical and theoretical cdf are plotted. It is observable that for the Danish dataset (left panel) the theoretical model adheres closer to the empirical data than for the Norwegian dataset.

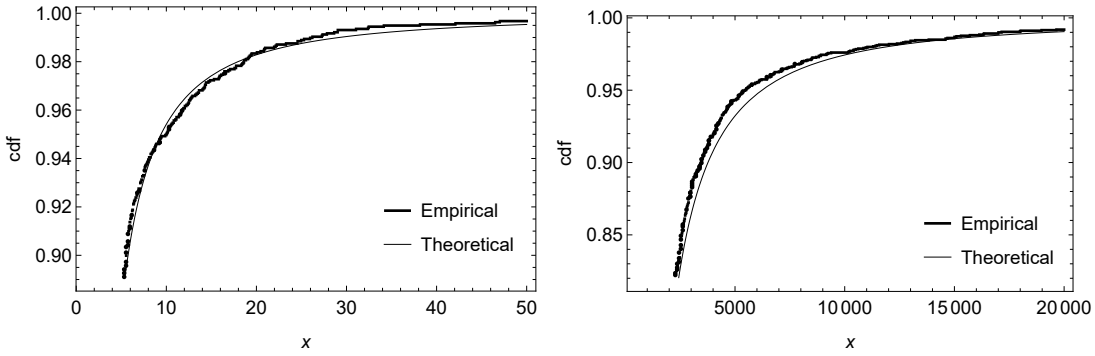

Figure 4: Empirical (thick line) and fitted cumulative distribution function for the Danish (left) and Norwegian (right) datasets.

In Figure 5, the limited expected value for the two sets of data have been plotted. It can be seen that when the policy limit x increases the theoretical model overestimates the empirical values for the Danish dataset. The converse occurs for the other set of data.

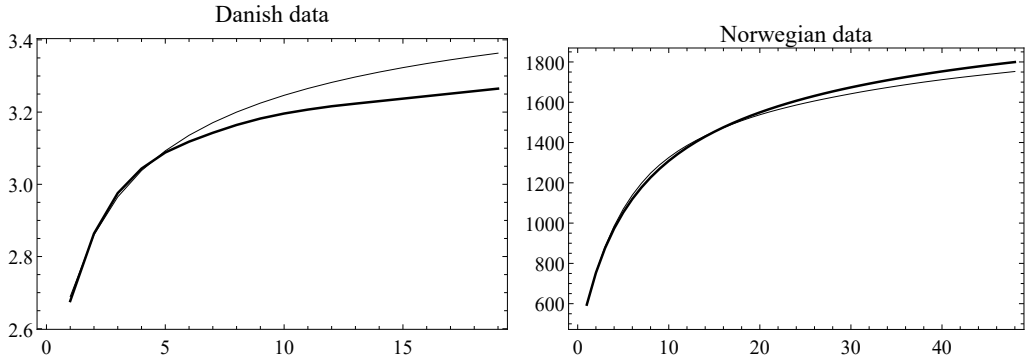

Figure 5: Empirical (thick line) and fitted limited expected values for the Danish (left) and Norwegian (right) datasets.

In Table 2, the tail value at risk (TVaR) (or first order tail moment), for different security levels has been calculated for the BP distribution. This risk measure describes the expected loss given that the loss exceeds the security level (quantile). These values have been calculated directly from the data. Empirical values have also been obtained. For the different risk levels it is discernible that the BP distribution overestimate the empirical TVaR values for the three security levels considered and the two sets of data.

5 Conclusions

In this work, the Beta-Pareto distribution, a generalization of the Pareto distribution that was introduced in the statistical literature not long time ago, has been extended and applied in financial and actuarial settings. In addition, several interesting properties related with the right-tail of the distribution were provided including the integrated tail distribution and the limited expected values among others. These properties, which had not been revealed until now, make the Beta-Pareto

	Risk Level α		
Risk Level	0.90	0.95	0.99
Danish dataset			
Empirical	15.637	24.305	61.376
BP	18.556	33.079	85.719
Norwegian dataset			
Empirical	9936.597	15635.295	42475.200
BP	11423.301	17576.914	53270.056

Table 2: Tail Value at Risk for different risk levels.

distribution a plausible alternative for applications in these fields. Additionally, its usefulness has been proven in its good performance against some well-known datasets usually considered in general insurance, improving the performance of other traditionally-used loss models.

Acknowledgments

The authors would like to express their gratitude to the Editor J.M. Sarabia for his kind invitation to write this paper and also for his valuable comments which have improved the content of this manuscript. EGD would also like to acknowledge partial financial support received from Ministerio de Economía, Industria y Competitividad. Agencia Estatal de Investigación (project ECO2017-85577-P).

References

Akinsete, A., F. Famoye, and C. Lee (2008). The beta-Pareto distribution. *Statistics 42*(6), 547–563.

Albrecher, H., J. Beirlant, and J.L. Teugels (2017). *Reinsurance: Actuarial and Statistical Aspects*. Wiley.

Arnold, B.C. (1983). *Pareto Distributions*. International Cooperative Publishing House, Silver Spring, MD.

Bingham, N.H. (1987). *Regular Variation*. Cambridge University Press, Cambridge.

Boland, P.J. (2007). *Statistical and Probabilistic Methods in Actuarial Science*. Chapman & Hall.

Boyd, A.V. (1988). Fitting the truncated Pareto distribution to loss distributions. *Journal of the Staple Inn Actuarial Society 31*, 151–158.

Brazauskas, V. and R. Serfling (2003). Favorable estimator for fitting Pareto models: a study using goodness-of-fit measures with actual data. *ASTIN Bulletin 33*(2), 365–381.

Calderín-Ojeda, E. (2016). The distribution of all French communes: A composite parametric approach. *Physica A 450*, 385–394.

Calderín-Ojeda, E. and C.F. Kwok (2016). Modeling claims data with composite stoppa models. *Scandinavian Actuarial Journal 2016*(9), 817–836.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997). *Modelling Extremal Events for Insurance and Finance*. Springer Verlag.

Ghitany, M.E., E. Gómez-Déniz, and S. Nadarajah (2018). A new generalization of the Pareto distribution and its application to insurance data. *Journal of Risk and Financial Management (Special Issue: Extreme Values and Financial Risk) 11*(10), 1–14.

Gómez-Déniz, E. and E. Calderín (2014). A suitable alternative to the Pareto distribution. *Hacettepe Journal of Mathematics and Statistics 43*(5), 843–860.

Gómez-Déniz, E. and E. Calderín (2015). On the use of the Pareto ArcTan distribution for describing city size in Australia and New Zealand. *Physica A 436*, 821–832.

Harris, C.M. (1968). The Pareto distribution as a queue service discipline. *Operations Research 16*(2), 307–313.

Hesselager, O. (1993). A class of conjugate priors with applications to excess-of-loss reinsurance. *ASTIN Bulletin 23*(1), 77–93.

Hogg, R. and S.A. Klugman (1984). *Loss Distributions*. John Wiley and Sons, New York.

Jessen, A.H. and T. Mikosch (2006). Regularly varying functions. *PUBLICATIONS DE L'INSTITUT MATHÉMATIQUE. Nouvelle série 80*(94), 171–192.

Jones, M.C. (2004). Families of distributions arising from distributions of order statistics. *Test 13*, 1–43.

Kleiber, C. and S. Kotz (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons, Inc.

Konstantinides, D.G. (2018). *Risk Theory. A Heavy Tail Approach*. World Scientific Publishing.

Mandelbrot, B. (1960). The Pareto-Lévy law and the distribution of income. *International Economic Review 1*(2), 79–106.

McNeil, A.J. (1997). Estimating the tails of loss severity distribution using extreme value theory. *ASTIN Bulletin 27*(1), 117–137.

Rolski, T., H. Schmidli, V. Schmidt, and J. Teugel (1999). *Stochastic Processes for Insurance and Finance*. John Wiley & Sons.

Rosen, K.T. and M. Resnick (1980). The size distribution of cities: An examination of the Pareto law and primacy. *Journal of Urban Economics 8*(2), 165–186.

Rytgaard, M. (1990). Estimation in the Pareto distribution. *ASTIN Bulletin 20*(2), 201–216.

Sarabia, J.M. and F. Prieto (2009). The Pareto-positive stable distribution: A new descriptive model for city size data. *Physica A 388*, 4179–4191.

Scollnik, D. (2007). On composite lognormal-Pareto models. *Scandinavian Actuarial Journal 2007*(1), 20–33.

Stoppa, G. (1990). Proprietà campionarie di un nuovo modello Pareto generalizzato. *XXXV Riunione Scientifica della Società Italiana di Statistica, Padova: Cedam*, 137–144.

Yang, H. (2004). Crámer-Lundberg asymptotics. In *Encyclopedia of Actuarial Science*, pp. 1–6. Wiley.

Spanish Journal of Statistics
Vol. 2 No. 1 2020, Pages 23–40
doi:https://doi.org/10.37830/SJS.2020.1.03

Regular Article

The Gamma-Chen distribution: a new family of distributions with applications

Lucas David R. Reis[1], Gauss M. Cordeiro[2], Maria do Carmo S. Lima[3]
[1]Federal University of Pernambuco, econ.lucasdavid@gmail.com
[2]Federal University of Pernambuco, gausscordeiro@gmail.com
[3]Federal University of Pernambuco, maria@de.ufpe.br

Received: July 21, 2020. Accepted: February 15, 2021.

Abstract: The generalized gamma-generated family adds one shape parameter to a baseline distribution. We define the gamma-Chen and derive some of its mathematical properties. Its hazard rate may have increasing, decreasing, bathtub and unimodal shapes due to the extra parameter, which portrays a positive point of the proposed model. We perform Monte Carlo simulations to prove that the asymptotic properties of the maximum likelihood estimators hold. We show empirically that the new distribution is better than ten others known distributions using engineering-related data sets.

Keywords: Bathtub, Chen distribution, gamma-Chen distribution, Maximum likelihood, Simulation study

MSC: 33B15, 33C15, 62E20, 62P30, 65C05

1 Introduction

In the area of survival analysis and new distributions, much is said about proposing families and, consequently, distributions, which model fatigue data sets, failure time of electronic components, etc., which constitute engineering data.

Several types of data sets have been used for new distributions from medical and different branches of engineering and industry. However, Brazilian data sets are seldom used in international statistical papers. In this context, this work focuses on two applications in engineering area from Brazil. In addition, given that many workers used data collected over 20 years, we adopt more recent data sets.

The Weibull and Birnbaum-Saunders models are among the most widely distributions taken for baseline for several generators in different areas of engineering. The main goal here is to propose a new distribution that is as flexible as, or more than, the aforementioned, and that fits recent real engineering data. We believe that this purpose is valid and innovative.

One of the most used methods in the construction of new lifetime distributions is based on well-established generators by adding shape(s) parameters to parent models. This method is adopted in

this paper. The proposed distribution is interesting for lifetime data analysis as a further option, where some known distributions do not fit well.

The probability density function (pdf) and cumulative distribution function (cdf) of the gamma-G family (Zografos and Balakrishnan, 2009) for a baseline G are

$$f_{\mathrm{GG}}(x;a,\boldsymbol{\eta}) = \frac{1}{\Gamma(a)}\{-\log[1-G(x;\boldsymbol{\eta})]\}^{a-1} g(x;\boldsymbol{\eta}) \tag{1}$$

and

$$F_{\mathrm{GG}}(x;a,\boldsymbol{\eta}) = \frac{\gamma(a,-\log[1-G(x;\boldsymbol{\eta})])}{\Gamma(a)} = \frac{1}{\Gamma(a)}\int_0^{-\log[1-G(x;\boldsymbol{\eta})]} t^{a-1}\mathrm{e}^{-t}dt,$$

respectively, where $a > 0$, $\boldsymbol{\eta}$ is the q-parameter vector of the baseline distribution, $g(x;\boldsymbol{\eta}) = dG(x;\boldsymbol{\eta})/dx$, $\Gamma(a) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}dt$ is the gamma function and $\gamma(a,z) = \int_0^z t^{a-1}\mathrm{e}^{-t}dt$ denotes the lower incomplete gamma function. This family is flexibilized by the shape parameter a and the support of $f_{\mathrm{GG}}(x)$ is the same of $g(x)$.

Many papers adopt the gamma-G family in order to fit several types of data sets. The great advantage in choosing this family over that one proposed by Torabi and Hedesh (2012) is the reduction in the problems of parameter estimation, since the proposal by Zografos and Balakrishnan (2009) has one parameter less than the other one. There are many works involving this family and its structural properties (Nadarajah et al., 2015). Table 1 lists some of its special models and associated data sets. In all cases, the gamma-G provides better fits when compared with another well-know distributions including beta-G models.

Table 1: Some gamma-G models

Model	Authors (year)	Application
Gamma-Birnbaum-Saunders	Cordeiro et al. (2016)	Failure and fatigue
Gamma-Normal	Lima (2015)	Agronomy and levels nicotine
Gamma-Lindley	Lima (2015)	Reliability and SAR images
Gamma-Nadarajah-Haghighi	Lima (2015)	Failure and fatigue
Gamma-Extended Weibull	Lima (2015)	Failure and fatigue
Gamma-Pareto	Alzaatreh et al. (2012)	River flood rates, fatigue and frequencies for Tribolium Confusum Strain
Gamma-Exponentiated Weibull	Pinho et al. (2012)	Daily minimum wind speed
Gamma exponentiated exponential-Weibull distribution	Pogány and Saboor (2016)	Remission times and fatigue

Major topics studied in the sections are as follows. In Section 2, we define the gamma-Chen (GC) model. In Section 3, we obtain some of its properties. In Section 4, we examine the accuracy of the maximum likelihood estimators (MLEs). The superiority of the GC model in relation to ten known distributions (including the well-known exponentiated Weibull model) is proved by means of two engineering data sets in Section 5. These competitors were chosen based on previous works in engineering data management (focus of the applications of this work). In Section 6, we conclude the paper.

2 Proposed model

In survival analysis it is very common to look for new distributions that have great versatility in the hazard rate function (hrf). The most common forms of hrfs are bathtub and unimodal. Chen (2000) proposed a two-parametric distribution that accommodates increasing and bathtub hrf forms, thus showing the great flexibility of this distribution.

Recently, some extensions of the Chen distribution (Chen, 2000) have appeared in the literature. Dey et al. (2017) proposed the exponentiated-Chen (exp-Chen) and showed that this distribution also has unimodal hrf. Among others extensions, we can mention Kumaraswamy-exponentiated-Chen (Khan et al., 2018) distribution, Weibull-Chen (Tarvirdizade and Ahmadpour, 2019) distribution, modified Weibull extension (Xie et al., 2002) and odd Chen-G family (Anzagra et al., 2020).

The cdf and pdf of the random variable $Y \sim \text{Chen}(\lambda, \beta)$ are

$$G(y;\lambda,\beta) = 1 - \mathrm{e}^{\lambda(1-\mathrm{e}^{y^\beta})}, \quad y > 0$$

and

$$g(y;\lambda,\beta) = \lambda\beta y^{\beta-1}\mathrm{e}^{y^\beta+\lambda(1-\mathrm{e}^{y^\beta})}, \tag{2}$$

respectively, where $\lambda > 0$ is a scale parameter and $\beta > 0$ is a shape parameter.

The GC density is determined from (1) and the last two equations

$$f_{\text{GC}}(x;a,\lambda,\beta) = \frac{\lambda^a\beta}{\Gamma(a)}\, x^{\beta-1}\,(\mathrm{e}^{x^\beta}-1)^{a-1}\,\mathrm{e}^{x^\beta+\lambda(1-\mathrm{e}^{x^\beta})}, \quad x > 0. \tag{3}$$

For $a = 1$, we have the Chen density. Henceforth, $X \sim \text{GC}(a,\lambda,\beta)$ denotes a random variable with pdf (3). The three-parameter GC distribution has no problem of identifiability. The Chen distribution is clearly identifiable since different parameter vectors imply different cumulative distributions. So, the GC is also identifiable.

A simple motivation for the GC density follows from Zografos and Balakrishnan (2009). If $Y_{(1)} < \cdots < Y_{(p)} < \cdots$ are upper record values arising from a sequence of Chen independent and identically random variables $Y_1, Y_2, \cdots$, then the order statistic $Y_{(p)}$ has the GC density with $a = p$. So, the density of X can approximate the density of the pth order statistic of the Chen(λ,β) distribution by taking p as the greatest integer less than or equal to a. So, the GC distribution is generated by Chen record value densities. This explicitly means that the GC distribution is a direct record-Chen analog.

A practical relevance and applicability of the GC distribution is for the lifetime system with n independent components which function if and only if at least k of the components function is a "k out of n" system. For such a system, k is less than n, and it includes some parallel, fail-safe and series systems all as special cases for $k = 1, k = n-1$ and $k = n$, respectively. Suppose $Y_1, \cdots, Y_n$ denote the lifetimes of n components having the Chen distribution of a system, where k is assumed unknown and n is very large. Then, the lifetime of a k-out-of-n system consisting of these components can be represented by the order statistic $Y_{(n-k+1)}$, which can be modeled by the GC distribution to estimate a and then k.

Figure 1(a) displays plots of the density of X for some parameter values, which show that it accommodates several forms. By combining different values of β and a provide great flexibility for the GC density. In fact, this density can be symmetric, left-skewed or right-skewed, and the parameter a has significant effects on both skewness and kurtosis.

The cdf and hrf of X are

$$F_{\text{GC}}(x;a,\lambda,\beta) = \frac{\gamma(a,-\lambda(1-\mathrm{e}^{x^\beta}))}{\Gamma(a)},$$

and

$$\tau_{\text{GC}}(x;a,\lambda,\beta) = \frac{\lambda^a \beta\, x^{\beta-1}\,(\mathrm{e}^{x^\beta}-1)^{a-1}\;\mathrm{e}^{x^\beta+\lambda(1-\mathrm{e}^{x^\beta})}}{\Gamma(a)-\gamma(a,-\lambda(1-\mathrm{e}^{x^\beta}))},$$

respectively. Note that for $a=\beta=1$ the shape of the hrf is independent of λ. Chen (2000) showed that the Chen hrf can only be increasing ($\beta \geq 1$) and bathtub ($\beta < 1$). However, the hrf of X can be increasing, decreasing, unimodal and bathtub-shaped as shown in Figure 1(b). Further, the bathtub shape can be obtained even when $\beta > 1$. This fact reveals that the hrf of X gains more flexibility with the extra parameter a since it can take the most four common forms for applications to real data: increasing for any positive value of β, bathtub-shaped, unimodal, and also decreasing, which shows that it has great flexibility due to the parameter a (see Figure 1(b)).

Following the idea in Qian (2012), we can to determine the parameter ranges for the density shapes. Setting $z = \exp(x^\beta)$, we obtain from (3)

$$r(z) = f([\log z]^{1/\beta}) = \frac{\lambda^a \beta \log z}{\Gamma(a)}(\log z)^{-1/\beta}(z-1)^{a-1}\exp[\log z + \lambda(1-z)].$$

Applying logarithms of both sides of the previous equation,

$$\begin{aligned}\log r(z) =& a\log\lambda + \log\beta + \log(\log z) - \log\Gamma(a) - \frac{1}{\beta}\log(\log z)\\ &+ (a-1)\log(z-1) + \log z + \lambda(1-z).\end{aligned}$$

By taking derivatives of both sides of the last equation, we have

$$\begin{aligned}\frac{r'(z)}{r(z)} =& \frac{1}{z\log z} - \frac{1}{\beta z\log z} + \frac{a-1}{z-1} + \frac{1}{z} - \lambda\\ =& \frac{\beta(z-1)-(z-1)+\beta z(a-1)\log z+\beta(z-1)\log z-\lambda\beta z(z-1)\log z}{\beta z(z-1)\log z}.\end{aligned}$$

If $s(z)$ is the numerator of the right side of this equation, we can write

$$r'(z) = \frac{r(z)s(z)}{\beta z(z-1)\log z}.$$

Hence, $r'(z)$ and $s(z)$ have the same signs, since $r(z) > 0$ and $\beta z(z-1)\log z > 0$ for $z > 1$. The condition $z > 1$ holds since $x > 0$. In this case, $x = \log(z)^{1/\beta} \Longleftrightarrow z > 1$.

Note that in Region I (Figure 2(a)), $s(z)$ takes positive values first and then negative values, which indicates the unimodal property of the density. In Region II (Figure 2(b)), $s(z)$ has only negative values, thus indicating decreasing shape. So, Figures 2(a) and 2(b) reveal that the pdf is unimodal for $a \geq 1$ and that it is decreasing for $a \in (0,1)$, respectively, as noted in Figure 1(a).

3 Properties

It is not possible to obtain some mathematical properties of the GC distribution in closed form, that is, according to known mathematical functions. Then, we determine these quantities from the weighted linear combination for its density function given in Theorem 2 below.

For a given cdf $G(z;\boldsymbol{\eta})$ with q-parameter vector $\boldsymbol{\eta}$, the cdf and pdf of the exponentiated-G (exp-G) random variable Z_a with power parameter $a > 0$, say $Z_a \sim \text{exp-G}(a,\boldsymbol{\eta})$, are

$$H(z;a,\boldsymbol{\eta}) = G(z;\boldsymbol{\eta})^a \quad \text{and} \quad h(z;a,\boldsymbol{\eta}) = a\,g(z;\boldsymbol{\eta})G(z;\boldsymbol{\eta})^{a-1},$$

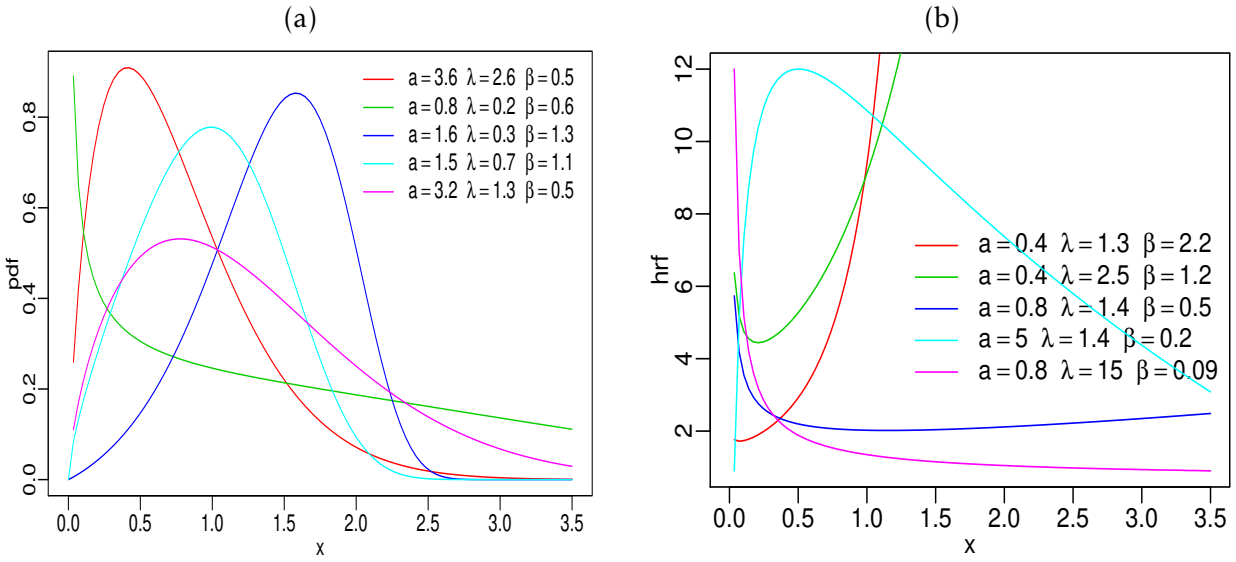

Figure 1: Plots of the pdf (a) and hrf (b) of the GC distribution.

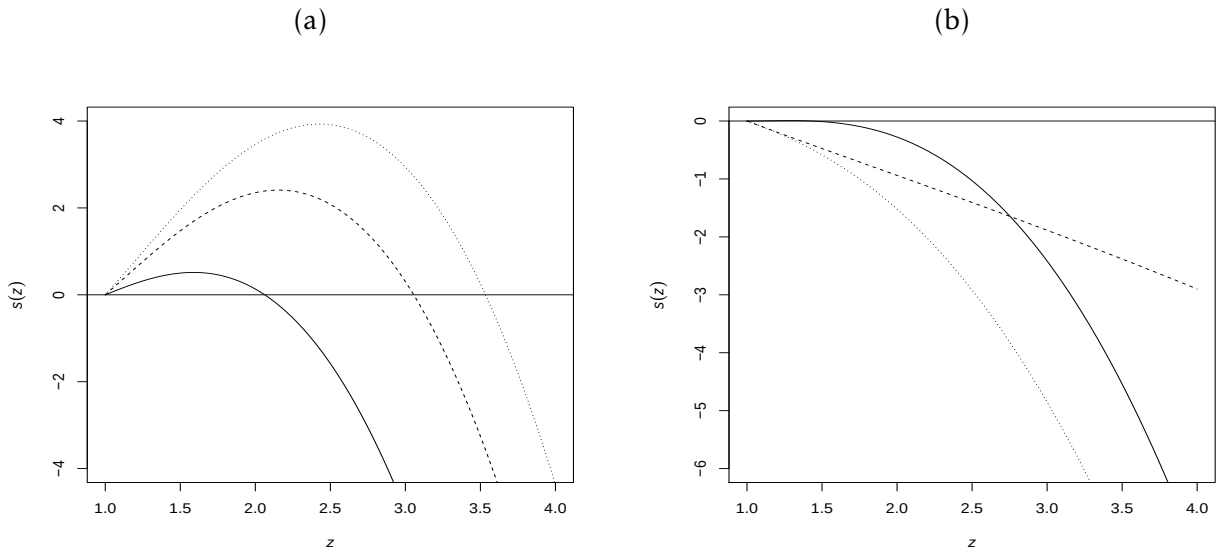

Figure 2: Regions for the density shapes. (a) Region I: $a \geq 1$ and (b) Region II: $a \in (0,1)$.

respectively, where $g(z;\boldsymbol{\eta}) = dG(z;\boldsymbol{\eta})/dz$.

The gamma-G cdf can be expressed as (Castellares and Lemonte, 2015)

$$F_{\text{GG}}(x;a,\boldsymbol{\eta}) = \sum_{k=0}^{\infty} \frac{\varphi_k(a)}{(a+k)} H(x;(a+k),\boldsymbol{\eta}), \tag{4}$$

where $\varphi_0(a) = \frac{1}{\Gamma(a)}$, $\varphi_k(a) = \frac{(a-1)}{\Gamma(a)}\,\psi_{k-1}(k+a-2)$ $(k \geq 1)$, $\psi_{k-1}(\cdot)$ are the Stirling polynomials

$$\psi_{k-1}(w) = \frac{(-1)^{k-1}}{(k+1)!}\left[T_k^{k-1} - \frac{(w+2)}{(k+2)}T_k^{k-2} + \frac{(w+2)(w+3)}{(k+2)(k+3)}T_k^{k-3} - \cdots + (-1)^{k-1}\frac{(w+2)(w+3)\cdots(w+k)}{(k+2)(k+3)\cdots(2k)}T_k^0\right],$$

$T_0^0 = 1$, $T_{k+1}^0 = 1 \times 3 \times \ldots \times (2k+1)$, $T_{k+1}^k = 1$ and T_k^m are positive integers determined from

$$T_{k+1}^m = (2k+1-m)T_k^m + (k-m+1)T_k^{m-1}.$$

The function $H(x;(a+k),\boldsymbol{\eta})$ denotes the cdf of Z_{a+k}. Thus, we can obtain the properties of the gamma-G model from those of the exp-G class.

Theorem 1. *Let Y be a random variable having density (2). Then, the exp-Chen(a,λ,β) density can be expressed as*

$$h(y;a,\lambda,\beta) = \sum_{m=1}^{\infty} p_m\, g(y;m\lambda,\beta),$$

where $p_m = p_m(a) = (-1)^{m+1}\binom{a}{m}$ and $g(y;m\lambda,\beta)$ is the Chen density with scale $m\lambda$ and shape β.

Proof. For $|x| < 1$ and any real $a \neq 0$, the power series

$$(1-x)^a = \sum_{m=0}^{\infty}(-1)^m\binom{a}{m}x^m$$

converges. Thus, the exp-Chen cdf can be expanded as

$$H(y;a,\lambda,\beta) = \left[1 - \mathrm{e}^{\lambda(1-\mathrm{e}^{y^\beta})}\right]^a = 1 + \sum_{m=1}^{\infty}(-1)^m\binom{a}{m}[1-G(y;m\lambda,\beta)].$$

By differentiating the last equation,

$$h(y;a,\lambda,\beta) = \sum_{m=1}^{\infty}(-1)^{m+1}\binom{a}{m}g(y;m\lambda,\beta),$$

and then the exp-Chen density is a linear combination of Chen densities. □

Theorem 2. *The pdf of X in Equation (3) can be expressed as*

$$f_{\mathrm{GC}}(x;a,\lambda,\beta) = \sum_{m=1}^{\infty} w_m\, g(x;m\lambda,\beta),$$

where $g(x;m\lambda,\beta)$ is the Chen density with scale $m\lambda$ and shape β and the weights are

$$w_m = w_m(a) = (-1)^{m+1}\sum_{k=0}^{\infty}\frac{\varphi_k(a)}{(a+k)}\binom{a+k}{m}.$$

Proof. The proof comes directly from Equation (4) and Theorem 1. □

By using Theorem 2, the rth moment of X has the form

$$\mathbb{E}[X^r] = \sum_{m=1}^{\infty} w_m \, \mathbb{E}[Y_m^r],$$

where $Y_m \sim \text{Chen}(m\lambda, \beta)$.

If $Y \sim \text{Chen}(\lambda, \beta)$ has pdf (2), we can write (Pogány et al., 2017)

$$\mathbb{E}[Y^r] = \lambda \, \mathrm{e}^{\lambda} \, \mathbb{D}_t^{r\beta^{-1}} \left[\frac{\Gamma(t+1, \lambda)}{\lambda^{t+1}} \right]_{t=0}. \tag{5}$$

Here,

$$\mathbb{D}_t^p \left[\frac{\Gamma(t+1, \lambda)}{\lambda^{t+1}} \right]_{t=0} = \Gamma(p+1) \sum_{k \geq 0} \frac{(2)_k}{k!} \, \Phi_{\mu,1}^{(0,1)}(-k, p+1, 1) \; {}_1F_1(k+2; 2; -\lambda),$$

where $\Phi_{\mu,1}^{(0,1)}(-a, p+1, 1) = \sum_{n \geq 0} \frac{(-a)^n}{n!(n+1)^{p+1}}$ for $\mu \in \mathbb{C}$, ${}_1F_1(a; b; x) = \sum_{n \geq 0} \frac{(a)_n}{(b)_n} \frac{x^n}{n!}$, for $x, a \in \mathbb{C}$ and $b \in \mathbb{C} \setminus Z_0^-$, is the confluent hypergeometric function (Kilbas et al., 2006, page 29, Eq. 1.6.14) and $(\lambda)_\eta = \frac{\Gamma(\lambda+\eta)}{\Gamma(\lambda)}$, for $\lambda \in \mathbb{C} \setminus \{0\}$, is the generalized Pochhammer symbol with $(0)_0 = 1$.

Thus, using (5), the rth moment of X can be reduced to

$$\mathbb{E}[X^r] = \lambda \sum_{m=1}^{\infty} m \, w_m \, \mathrm{e}^{m\lambda} \, \mathbb{D}_t^{r\beta^{-1}} \left[\frac{\Gamma(t+1, m\lambda)}{(m\lambda)^{t+1}} \right]_{t=0}.$$

Figures 3, 4 and 5 provide the plots of the mean and variance of X as functions of a, λ and β, respectively, the other parameters being fixed. The mean and variance of X increase when a increases. In turn, these measures decrease when λ increases. Further, the mean of X increases when β increases and the variance of X increases to a maximum point and starts to decrease.

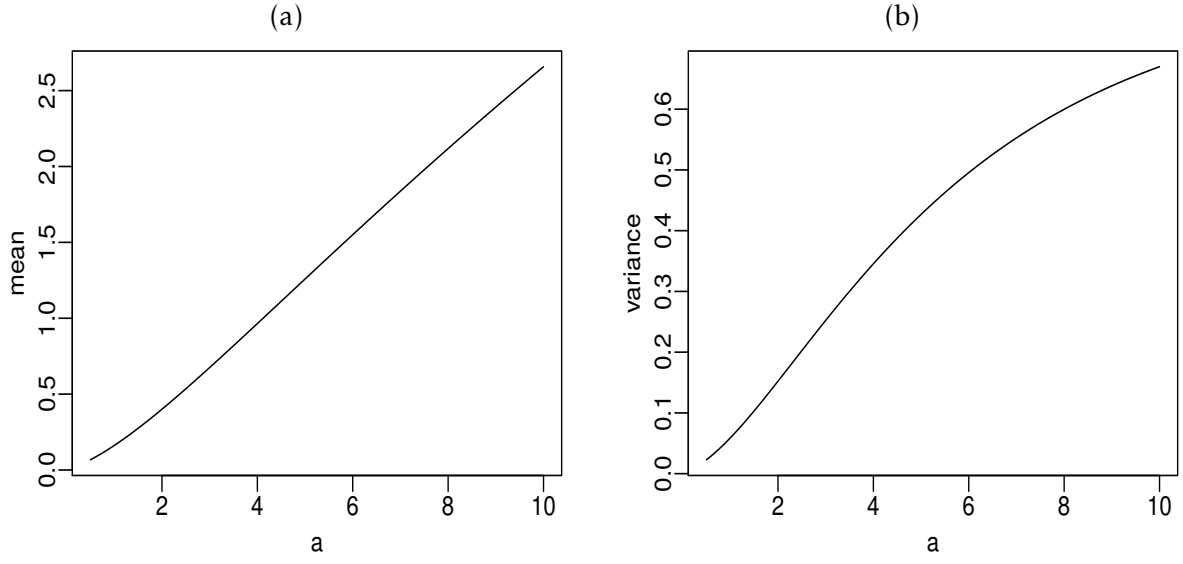

Figure 3: Mean (a) and variance (b) plots of X as functions of a ($\lambda = 2.4, \beta = 0.5$).

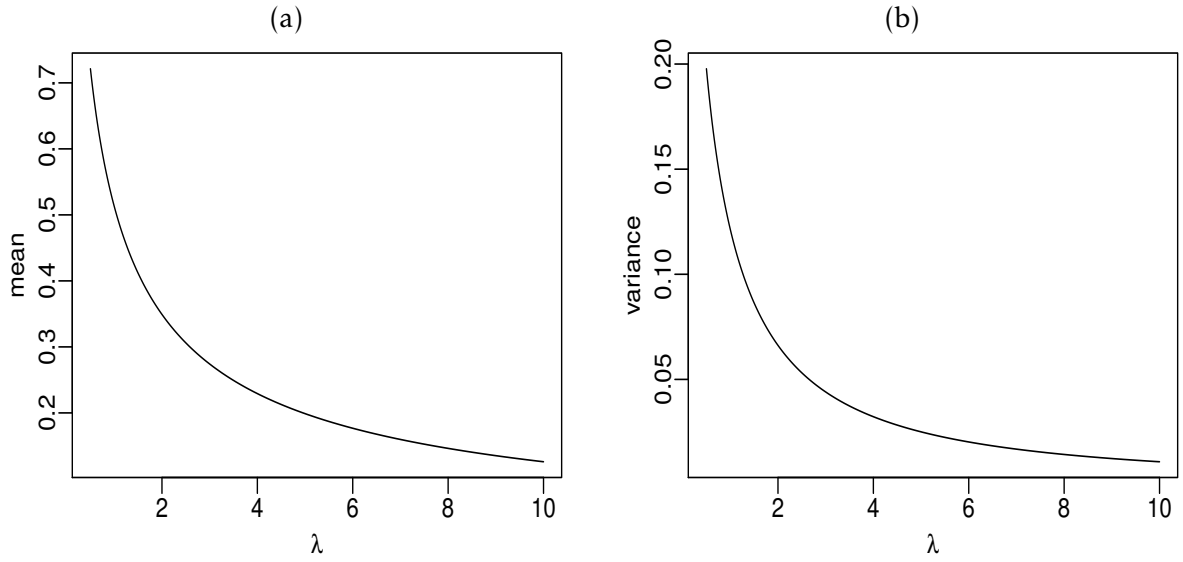

Figure 4: Mean (a) and variance (b) plots of X as functions of λ ($a = 0.7, \beta = 1.4$).

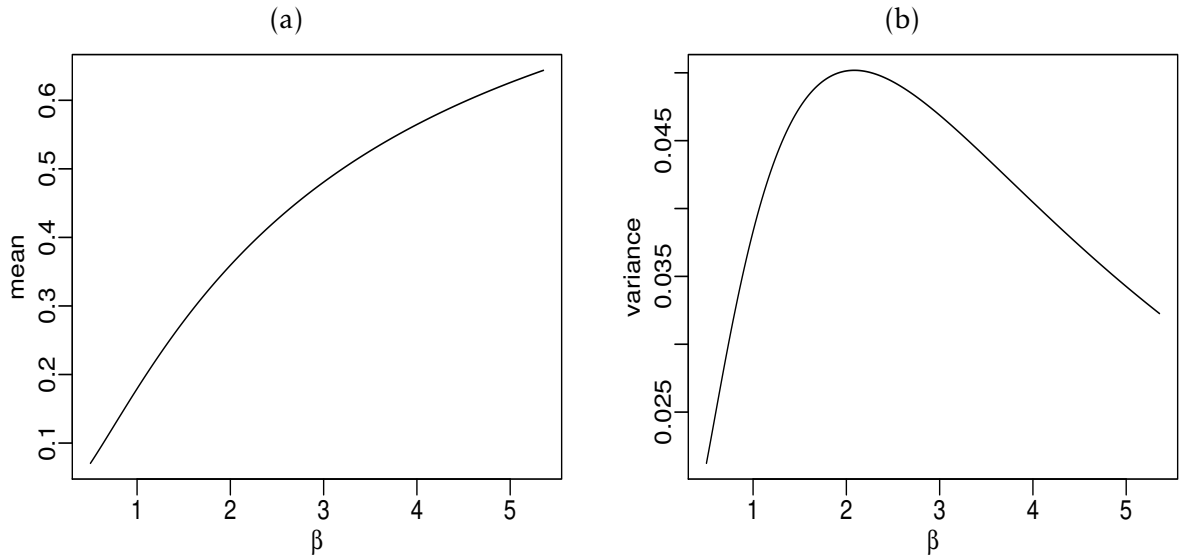

Figure 5: Mean (a) and variance (b) plots of X as functions of β ($a = 0.6, \lambda = 2.7$).

Another type of measure that has great applicability is the incomplete moment. For $z > 0$, the rth incomplete moment of the random variable Y with Chen distribution, say $q_r(z;\lambda,\beta) = \int_0^z y^r g(y;\lambda,\beta)dy$, follows from Pogány et al. (2017) as

$$q_r(z;\lambda,\beta) = \lambda e^{\lambda} \sum_{n,k\geq 0} \sum_{j=1}^{k} \frac{(2)_{n+k}}{(2)_n} \frac{(-1)^{n+j}\,\lambda^n}{n!k!(j+1)^{r\beta^{-1}+1}} \binom{k}{j} \gamma\left(r\beta^{-1}, (j+1)(1-z^{-1})\right). \tag{6}$$

Thus, using Theorem 2 and Equation (6), the rth incomplete moment of X is

$$m_r(z) = \lambda \sum_{m=1}^{\infty} m\, \mathrm{e}^{m\lambda}\, w_m \sum_{n,k\geq 0} \sum_{j=1}^{k} \frac{(2)_{n+k}}{(2)_n} \frac{(-1)^{n+j}\,(m\lambda)^n}{n!k!(j+1)^{r\beta^{-1}+1}} \binom{k}{j} \gamma\left(r\beta^{-1}, (1-z^{-1})(j+1)\right).$$

The first incomplete moment is used to obtain Lorenz and Bonferroni curves and mean deviations.

The generating function (gf) of $Y \sim \mathrm{Chen}(\lambda, \beta)$, $M_Y(-t) = \mathbb{E}[\mathrm{e}^{-tY}]$, $t > 0$, can be written, according to Pogány et al. (2017), by

$$M_Y(-t) = \lambda\beta\, \mathrm{e}^{\lambda}\, t^{-\beta} \sum_{n\geq 0} \frac{(-\lambda)^n}{n!}\, {}_1\Psi_0\left[(\beta,\beta); -; \frac{n+1}{t^{\beta}}\right], \tag{7}$$

where

$${}_1\Psi_0\left[(a,b); -; z\right] = \sum_{n\geq 0} \frac{\Gamma(a+bn)\, z^n}{n!}, \quad z, a \in \mathbb{C}, b > 0,$$

is the generalized Fox-Wright function.

Thus, from Theorem 2 and Equation (7), the gf of X follows as (for $t > 0$)

$$M_X(-t) = \lambda\beta\, t^{-\beta} \sum_{m=1}^{\infty} \sum_{n\geq 0} \frac{m\, \mathrm{e}^{m\lambda}\, (-m\lambda)^n\, w_m}{n!}\, {}_1\Psi_0\left[(\beta,\beta); -; \frac{n+1}{t^{\beta}}\right].$$

The quantile function (qf) of X, say $Q_{\mathrm{GG}}(u; a, \boldsymbol{\eta}) = F^{-1}_{\mathrm{GG}}(u; a, \boldsymbol{\eta})$, can be expressed as (Nadarajah et al., 2015)

$$Q_{\mathrm{GG}}(u; a, \boldsymbol{\eta}) = Q_{\mathrm{G}}\left(1 - \mathrm{e}^{-Q_1(a,u)}; \boldsymbol{\eta}\right), \quad 0 < u < 1,$$

where Q_{G} is the qf of the baseline $G(x; \boldsymbol{\eta})$ and $Q_1(a, u)$ is the inverse function of $\gamma_1(a, w) = \gamma(a, w)/\Gamma(a)$. Further, we can write

$$Q_{\mathrm{GC}}(u; a, \lambda, \beta) = \left\{\log\left[1 + \lambda^{-1} Q_1(a, u)\right]\right\}^{1/\beta}. \tag{8}$$

We can obtain skewness and kurtosis measures of X from Equation (8). The Bowley skewness and Moors kurtosis are based on quartiles and octiles, respectively. Letting $Q_{\mathrm{GC}}(u) = Q_{\mathrm{GC}}(u; a, \lambda, \beta)$, the skewness and kurtosis of X are

$$\mathcal{B}(a, \lambda, \beta) = \frac{Q_{\mathrm{GC}}(3/4) + Q_{\mathrm{GC}}(1/4) - 2Q_{\mathrm{GC}}(2/4)}{Q_{\mathrm{GC}}(3/4) - Q_{\mathrm{GC}}(1/4)}$$

and

$$\mathcal{M}(a, \lambda, \beta) = \frac{Q_{\mathrm{GC}}(7/8) - Q_{\mathrm{GC}}(5/8) - Q_{\mathrm{GC}}(3/8) + Q_{\mathrm{GC}}(1/8)}{Q_{\mathrm{GC}}(6/8) - Q_{\mathrm{GC}}(2/8)},$$

respectively. Plots of these measures as functions of a are displayed in Figure 6, which show that both of them decrease when a increases. Both measures grow when a decreases from one, and they can take negative values and higher positive values when a increases from one.

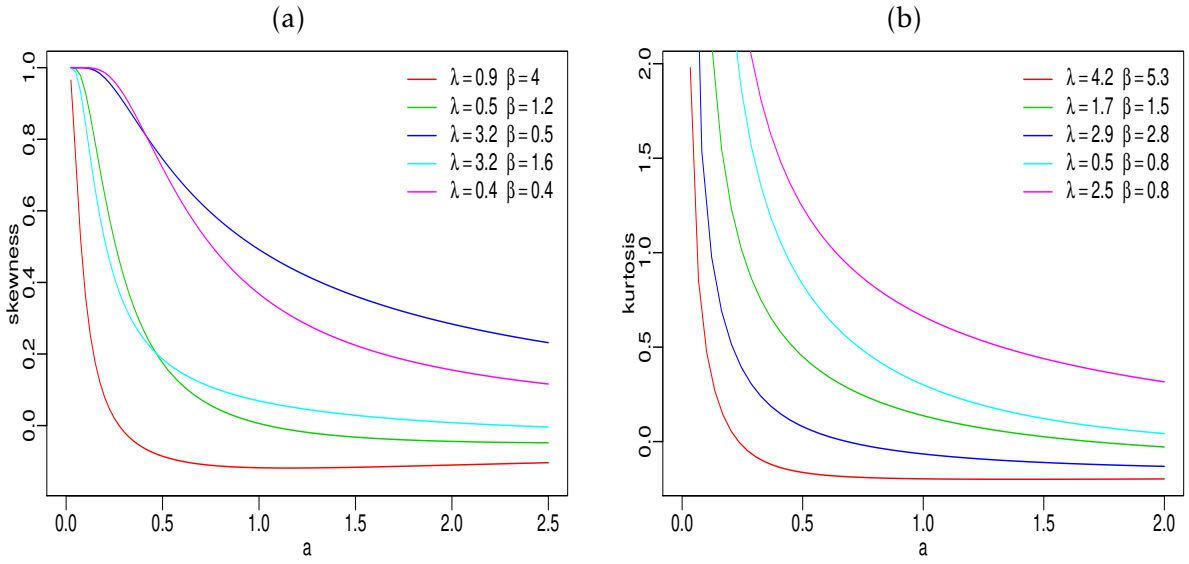

Figure 6: Skewness (a) and kurtosis (b) plots of X as functions of a.

4 Estimation

Let $\boldsymbol{\theta} = (a, \lambda, \beta)^\top$ be the parameter vector of the GC model. Consider the random variables $X_1, \ldots, X_n \sim \mathrm{GC}(a, \lambda, \beta)$ with observed values $x_1, \ldots, x_n$. The log-likelihood function for $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}) = n[a\log\lambda + \log\beta - \log\Gamma(a) + \lambda] + \sum_{i=1}^{n} x_i^{\beta} + (\beta-1)\sum_{i=1}^{n}\log x_i$$
$$+ (a-1)\sum_{i=1}^{n}\log(\mathrm{e}^{x_i^{\beta}} - 1) - \lambda\sum_{i=1}^{n}\mathrm{e}^{x_i^{\beta}}.$$

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, can be found by maximizing $\ell(\boldsymbol{\theta})$ numerically with respect to its components. Some routines such as SAS (`PROC NLMIXED`), R (`optim` function) and Ox (sub-routine `MaxBFGS`) can be used for the maximization.

We now study the behavior of MLEs in the GC model from 1,000 Monte Carlo replications. All simulations are performed using R Project (R Core Team, 2019). The sample sizes chosen are $n = 25$, 50, 100, 200, 300 and 400 and the true parameter vectors are: $(a, \lambda, \beta) = (1.4, 0.7, 1.9)$ for scenario 1, and $(a, \lambda, \beta) = (2.5, 1.5, 0.8)$ for scenario 2. There were no special reasons for choosing these parameters.

Table 2 reports the average estimates (AEs), biases and mean squared errors (MSEs) for both scenarios. The MLEs converge to the true parameters and the biases and MSEs decrease to zero when the sample size n increases, that makes us conclude that the consistency criterion holds.

5 Engineering data

In order to show a superior performance of the new distribution when compared to others already published in the literature, we provide two applications in recent real data sets in the engineering

Table 2: Findings under scenarios 1 to 2.

scenario 1									
Par	$n = 25$			$n = 50$			$n = 100$		
	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
a	1.965	0.565	5.503	1.656	0.256	1.634	1.496	0.096	0.485
λ	1.065	0.365	2.138	0.868	0.168	0.641	0.762	0.062	0.206
β	2.216	0.316	1.074	2.049	0.149	0.415	1.985	0.085	0.167
Par	$n = 200$			$n = 300$			$n = 400$		
	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
a	1.442	0.042	0.155	1.436	0.036	0.108	1.419	0.019	0.081
λ	0.727	0.027	0.068	0.727	0.027	0.045	0.715	0.015	0.029
β	1.942	0.042	0.077	1.917	0.017	0.048	1.916	0.016	0.040
scenario 2									
Par	$n = 25$			$n = 50$			$n = 100$		
	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
a	1.965	0.465	7.624	1.808	0.308	4.813	1.621	0.121	1.326
λ	1.869	0.369	4.307	1.787	0.287	2.159	1.608	0.108	0.806
β	1.393	0.593	2.348	1.043	0.243	0.778	0.891	0.091	0.104
Par	$n = 200$			$n = 300$			$n = 400$		
	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
a	1.539	0.039	0.400	1.547	0.047	0.225	1.477	-0.023	0.345
λ	1.542	0.042	0.246	1.546	0.046	0.155	1.506	0.006	0.134
β	0.851	0.051	0.077	0.819	0.019	0.021	0.839	0.039	0.053

area. In this work, we choose two types of engineering data sets to show the flexibility of the proposed model. One related to ore wagon fleets and the other to natural gas; both sets belong to the production engineering branch. Table 3 presents the descriptive statistics of the two data sets.

The first data set is obtained in a work by (Sivini, 2006), which involves the execution of a pilot project of a reliability data applied in natural gas pressure reducing stations (ERPGN) of a company that operates in Pernambuco (Brazil). The data in question refer to the time until the maintenance time (Tm) in one of the Pressure Reduction and Measurement Stations (ERPM - A) between 10/14/2002 to 05/16/2005.

The second application has a data set taken from the same work (Sivini, 2006). Here, we consider Tm in ERPM B and C, collected between 11/14/2002 and 6/16/2005.

Table 3 gives the descriptive statistics of the two data sets. Note that the two data sets differ widely. The first with a mean of 2.9222 and the second with a mean of 6.2062. Their maximum values and standard deviations (SDs) are also very different.

5.1 Competitive distributions

We compare the GC model with other ten distributions: Chen, exponentiated Weibull (Mudholkar and Hutson, 1993), Kumaraswamy-log-logistic (de Santana et al., 2012), gamma-extended Frèchet (da Silva et al., 2013), beta-log-logistic (Lemonte, 2014), Birnbaum-Saunders (Birnbaum and Saunders, 1969), gamma-Birnbaum-Saunders (Cordeiro et al., 2016), beta Birnbaum-Saunders (Cordeiro and Lemonte, 2011), odd-log-logistic Birnbaum-Saunders (Ortega et al., 2016) and odd-log-logistic Birnbaum-Saunders Poisson (Cordeiro et al., 2018).

Table 3: Descriptive statistics.

Description	data set 1	data set 2
Min.	1.1700	1.0000
1st Qu.	1.1900	3.0625
Median	1.5850	4.8350
Mean	2.9222	6.2062
3rd Qu.	4.9175	6.1250
Max.	8.0000	30.3300
SD	2.2553	6.7021

The choice of the previous distributions is based on suitable ones with good fits to engineering data. We emphasize that other distributions could also be used.

The densities of the exponentiated Weibull (EW), Kumaraswamy-log-logistic (KLL), gamma-extended Frèchet (GEF) and beta-log-logistic (BLL) are (for $x > 0$)

$$f_{\text{EW}}(x;a,\lambda,\beta) = \frac{a\lambda}{\beta}\left(\frac{x}{\beta}\right)^{\lambda-1}\left\{1-\exp\left[-\left(\frac{x}{\beta}\right)^{\lambda}\right]\right\}^{(a-1)}\exp\left[-\left(\frac{x}{\beta}\right)^{\lambda}\right],$$

$$f_{\text{KLL}}(x;a,b,\alpha,\delta) = \frac{ab\delta}{\alpha^{a\delta}}x^{a\delta-1}\left[1-\left(\frac{x}{\alpha}\right)^{\delta}\right]^{-a-1}\left\{1-\left[1-\frac{1}{1+\left(\frac{x}{\alpha}\right)^{\delta}}\right]^{a}\right\}^{b-1},$$

$$f_{\text{GEF}}(x;a,\lambda,\sigma,\alpha) = \frac{\alpha\lambda\sigma^{\lambda}}{\Gamma(a)}x^{-\lambda-1}\exp\left[-\left(\frac{\sigma}{x}\right)^{\lambda}\right]\left\{1-\exp\left[-\left(\frac{\sigma}{x}\right)^{\lambda}\right]\right\}^{\alpha-1} \times\left\{-\log\left\{1-\exp\left[-\left(\frac{\sigma}{x}\right)^{\lambda}\right]\right\}^{\alpha}\right\}^{a-1},$$

and

$$f_{\text{BLL}}(x;a,b,\alpha,\beta) = \frac{\beta\,\Gamma(a)\Gamma(b)}{\alpha\,\Gamma(a+b)}\frac{(x/\alpha)^{a\beta-1}}{[1+(x/\alpha)^{\beta}]^{a+b}},$$

respectively, where all parameters are positive.

The cdf and pdf of the Birnbaum-Saunders (BS) are

$$F_{\text{BS}}(x;\alpha,\beta) = \Phi\left(\frac{1}{\alpha}\left[\left(\frac{x}{\beta}\right)^{\frac{1}{2}}-\left(\frac{\beta}{x}\right)^{\frac{1}{2}}\right]\right), \quad x>0 \tag{9}$$

and

$$f_{\text{BS}}(x;\alpha,\beta) = \frac{\exp(\alpha^{-2})}{2\alpha\sqrt{2\pi\beta}}x^{-\frac{3}{2}}(x+\beta)\exp\left[-\frac{1}{2\alpha^2}\left(\frac{x}{\beta}+\frac{\beta}{x}\right)\right], \tag{10}$$

respectively, where $\alpha, \beta > 0$ and $\Phi(\cdot)$ is the standard normal cdf.

The densities of the gamma-Birnbaum-Saunders (GBS), beta-Birnbaum-Saunders (BBS), odd-log-logistic Birnbaum-Saunders (OLLBS) and odd-log-logistic Birnbaum-Saunders Poisson (OLLBSP) distributions are given by

$$f_{\text{BG}}(x;a,b,\boldsymbol{\eta}) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}g(x;\boldsymbol{\eta})G(x;\boldsymbol{\eta})^{a-1}\bar{G}(x;\boldsymbol{\eta})^{b-1},$$

$$f_{\text{OLLG}}(x;a,\boldsymbol{\eta}) = \frac{a\, g(x;\boldsymbol{\eta})\Big[G(x;\boldsymbol{\eta})\bar{G}(x;\boldsymbol{\eta})\Big]^{a-1}}{\Big[G(x;\boldsymbol{\eta})^a + \bar{G}(x;\boldsymbol{\eta})^a\Big]^2}$$

and

$$f_{\text{OLLG-P}}(x;a,b,\boldsymbol{\eta}) = \frac{ab\, g(x;\boldsymbol{\eta})\Big[G(x;\boldsymbol{\eta})\bar{G}(x;\boldsymbol{\eta})\Big]^{a-1}}{(\mathrm{e}^b-1)\Big[G(x;\boldsymbol{\eta})^a + \bar{G}(x;\boldsymbol{\eta})^a\Big]^2} \exp\left[\frac{b\, G(x;\boldsymbol{\eta})^a}{G(x;\boldsymbol{\eta})^a + \bar{G}(x;\boldsymbol{\eta})^a}\right],$$

respectively, where $a, b > 0$ and $\bar{G}(x;\boldsymbol{\eta}) = 1 - G(x;\boldsymbol{\eta})$.

As for the simulations, we adopt the *open source* computing platform: R Project (R Core Team, 2019). The MLEs of the parameters of the fitted densities are calculated using the `goodness.fit` function of the script `AdequacyModel` (Marinho et al., 2019) available in programming environment R Project (R Core Team, 2019) with the BFGS method. The best models fitted to the data sets are chosen based on the statistics: Cramèr-von Mises (W^*), Anderson-Darling (A^*), Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC) and Kolmogorov-Smirnov (KS) and its p-value.

Initial parameter values are chosen based on a function created using the `GenSA` package from R Project (Xiang et al., 2013). Such a package allows implementing a function that seeks the global minimum of a given function with a large number of optimum points. Therefore, we insert functions in R that take as arguments the data set to be used and the desired density. The functions in question return the initial shots of the parameters in question.

5.2 Findings

Tables 4 and 5 give the MLEs and their standard errors (SEs) in parentheses and the information criteria, respectively. The values for all statistics (except KS) in Table 5 indicate that the GC distribution is the best model to these data. Further, the p-values of the KS statistic also reveal that all distributions (except BS, OLLBS and EW models) can be used to fit the current data. So, the information criteria support that the CG distribution provides the best fit to these data. The plots of the estimated pdfs and cdfs and the Kaplan-Meier (KM) estimate, for the two best models, displayed in Figure 7 reveal that the GC distribution is the most adequate model to these data.

For the data set 2, the MLEs, SEs and information criteria are reported in Tables 6 and 7, respectively. All information criteria also indicate to the CG distribution is the best model when compared to the others. The p-values of the KS statistic show that the BS, OLLBS and OLLBSP models can not be used for the current data. Based on the histogram, the estimated pdfs and cdfs and the KM estimate (Figure 8), we can conclude that the GC distribution provides a better fit to these data.

The likelihood ratio (LR) statistics that compare the GC and Chen models, for the two data sets, are reported in Table 8. For both data sets, the null hypothesis is rejected, and the GC distribution is a more appropriate model for both data sets.

6 Conclusions

We introduce the gamma-Chen (GC) distribution which extends the Chen model. The new distribution adds an extra shape parameter thus giving greater flexibility. We obtain some of its mathematical properties. The hazard rate function of the GC distribution may have increasing, decreasing and bathtub shapes. We show the consistency of the maximum likelihood estimators via Monte Carlo

Table 4: Fitted models to data set 1.

Model	Estimates			
BS(α,β)	29.9850	0.0034		
	(9.6966)	(0.0017)		
Chen(λ,β)	0.1410	0.5849		
	(0.0546)	(0.0685)		
OLLBS(α,β,a)	67.3595	2.1694	95.2768	
	(0.0294)	(0.0294)	(0.1584)	
GBS(α,β,a)	0.6451	3.5546	0.5637	
	(0.1265)	(0.3059)	(0.1775)	
GC(a,λ,β)	159.6704	84.6636	0.0685	
	(0.8199)	(0.1208)	(0.0088)	
EW(a,λ,β)	0.0438	7.9268	40.2123	
	(0.0103)	(0.0890)	(0.0762)	
BLL(a,b,α,β)	7.6014	493.3696	6999.9700	0.5272
	(0.0003)	(<0.0001)	(1.8495)	(0.0109)
GEF$(a,\lambda,\sigma,\alpha)$	1.1000	0.3263	1110.5080	999.9747
	(<0.0001)	(<0.0001)	(<0.0001)	0.0027
KLL(a,b,α,δ)	20.5704	0.1668	0.7889	8.4532
	(<0.0001)	(0.0002)	(0.0004)	(0.0001)
OLLBSP(α,β,a,b)	479.9736	0.0513	182.0844	6.1178
	(0.0004)	(0.0004)	(0.0019)	(0.1940)
BBS(α,β,a,b)	86.2317	0.0551	0.0905	0.0800
	(0.0010)	(0.0120)	(0.0010)	(9.9186)

Table 5: Information criteria for data set 1.

Model	W^*	A^*	AIC	CAIC	BIC	HQIC	KS	p-value (KS)
BS	0.3054	1.7138	113.8926	114.6926	115.6733	114.1381	0.7303	<0.0001
Chen	0.3229	1.7662	78.7075	79.5075	80.4882	78.9530	0.2450	0.2300
OLLBS	0.2973	1.6868	117.663	119.3773	120.3341	118.0313	0.7325	<0.0001
GBS	0.2999	1.6924	77.6396	79.3538	80.3107	78.0079	0.2339	0.2782
GC	**0.2613**	**1.5455**	**73.9505**	**75.6648**	**76.6216**	**74.3188**	0.2254	0.3198
EW	0.2775	1.5989	109.3646	111.0789	112.0357	109.7329	0.4288	0.0027
BLL	0.2716	1.5835	76.7815	79.8584	80.3429	77.2725	0.2368	0.2652
GEF	0.2900	1.6348	77.9778	81.0547	81.5392	78.4689	0.2553	0.1911
KLL	0.2673	1.5726	76.6320	79.7089	80.1934	77.1231	0.2255	0.3194
OLLBSP	0.2731	1.6114	78.3789	81.4559	81.9404	78.8700	0.2189	0.3544
BBS	0.2947	1.6678	78.8362	81.9131	82.3976	79.3273	0.2535	0.1976

simulations. We prove empirically that the new distribution is better than ten known distributions by means of two real engineering data sets.

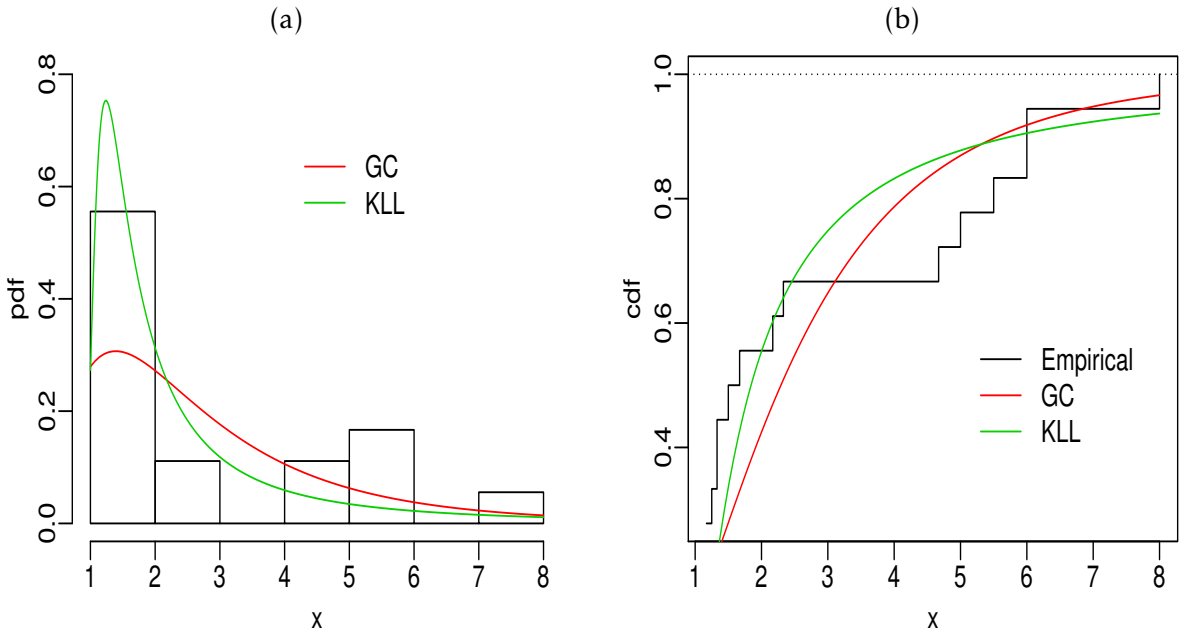

Figure 7: Estimated (a) pdfs and (b) cdfs and empirical cdf for data set 1.

Table 6: Fitted models to data set 2.

Model	Estimates			
BS(α,β)	0.7743	4.7946		
	(0.1678)	(0.0029)		
Chen(λ,β)	0.1247	0.3990		
	(0.0473)	(0.0385)		
OLLBS(α,β,a)	0.9426	1252.8320	0.0075	
	(0.0004)	(0.0319)	(0.0014)	
GBS(α,β,a)	0.6457	9.0578	0.4192	
	(0.0104)	(0.0001)	(0.1064)	
GC(a,λ,β)	18.1587	6.5901	0.1705	
	(<0.0001)	(0.2862)	(0.0178)	
EW(a,λ,β)	0.0685	5.5721	64.0574	
	(0.0171)	(0.0018)	(0.0018)	
BLL(a,b,α,β)	5.4816	465.9178	7001.0060	0.6199
	(0.0001)	(0.0142)	(0.0471)	(0.0153)
GEF$(a,\lambda,\sigma,\alpha)$	0.2547	0.6594	102.7024	78.3443
	(0.0634)	(0.0013)	(0.0051)	(0.0015)
KLL(a,b,α,δ)	20.3658	37.5799	0.1393	0.4325
	(0.0025)	(<0.0001)	(0.0763)	(0.0633)
OLLBSP(α,β,a,b)	96.3598	9.0409	150.5198	0.1610
	(0.0005)	(0.0457)	(1.7901)	(0.0760)
BBS(α,β,a,b)	27.7816	12.3196	827.4882	879.9439
	(7.9993)	(4.5306)	(0.0011)	(0.0016)

Table 7: Information criteria for data set 2.

Model	W^*	A^*	AIC	CAIC	BIC	HQIC	KS	p-value (KS)
BS	0.2187	1.3995	125.2946	126.2177	126.8398	125.3737	0.6726	<0.0001
Chen	0.2981	1.8112	100.0162	100.9393	101.5614	100.0953	0.2800	0.1626
OLLBS	0.1231	0.7648	127.8833	129.8833	130.2011	128.0020	0.6861	<0.0001
GBS	0.1217	0.7848	90.1436	92.1436	92.4614	90.2623	0.2116	0.4709
GC	**0.1109**	**0.7403**	**89.9507**	**91.9507**	**92.2685**	**90.0694**	0.2030	0.5245
EW	0.1154	0.7746	90.3318	92.3318	92.6496	90.4505	0.2156	0.4468
BLL	0.1295	0.8653	93.2454	96.8818	96.3358	93.4037	0.2162	0.4433
GEF	0.2212	1.3587	98.9108	102.5472	102.0012	99.0691	0.2226	0.4059
KLL	0.1179	0.7944	92.7261	96.3625	95.8165	92.8844	0.1992	0.5494
OLLBSP	0.1779	1.0750	193.8526	197.4889	196.9429	194.0108	0.5065	0.0005
BBS	0.1515	0.9833	94.2090	97.8454	97.2994	94.3673	0.2599	0.2300

Table 8: LR test (GC vs Chen).

Description	hypothesis	LR	p-value
data set 1	$\mathcal{H}_0 : a = 1$ vs $\mathcal{H}_1 : a \neq 1$	6.7570	0.0093
data set 2	$\mathcal{H}_0 : a = 1$ vs $\mathcal{H}_1 : a \neq 1$	12.0655	0.0005

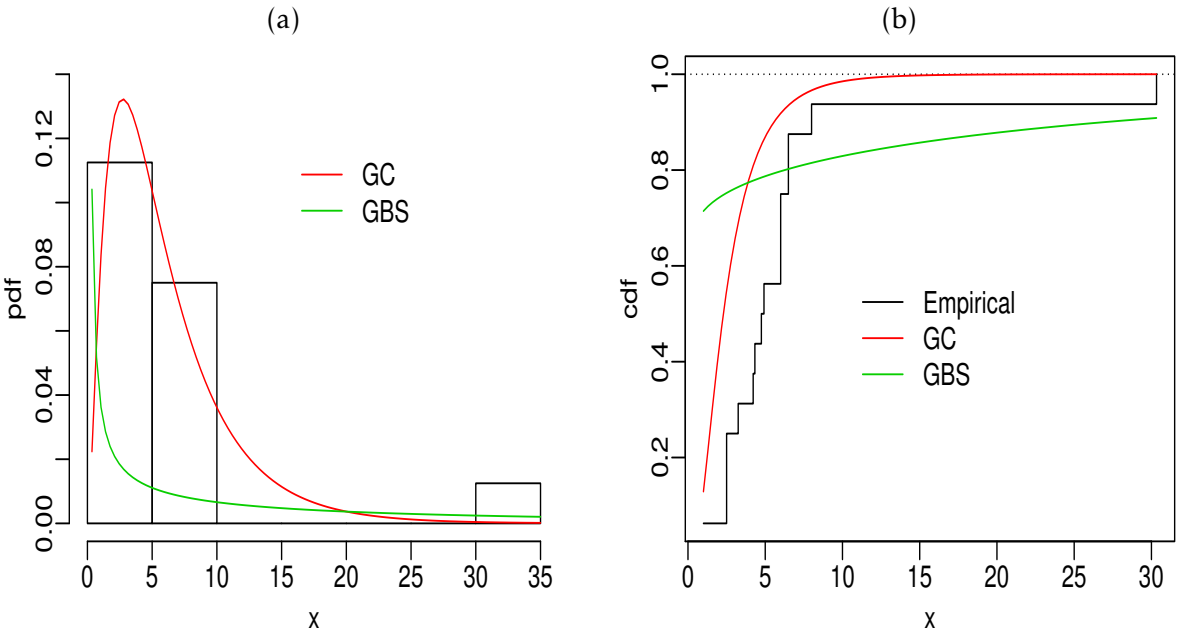

Figure 8: Estimated (a) pdfs and (b) cdfs and empirical cdf for data set 2.

Acknowledgments

The authors would like to thank the Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) for funding the doctoral scholarship. Thanks to Associate Editor and reviewers.

References

Alzaatreh, A., F. Famoye, and C. Lee (2012). Gamma-Pareto Distribution and its Applications. *Journal of Modern Applied Statistical Methods 11*(1), 78–94.

Anzagra, L., S. Sarpong, and S. Nasiru (2020). Odd Chen-G Family of Distributions. *Annals of Data Science 16*(3), 1–23.

Birnbaum, Z.W. and S.C. Saunders (1969). A new family of life distributions. *Journal of Applied Probability 6*(1), 319–327.

Castellares, F. and A.J. Lemonte (2015). A new generalized Weibull distribution generated by gamma random variables. *Journal of the Egyptian Mathematical Society 23*(2), 382–390.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters 49*(2), 155–161.

Cordeiro, G.M. and A.J. Lemonte (2011). The β-Birnbaum-Saunders distribution: an improved distribution for fatigue life modeling. *Computational Statistics & Data Analysis 55*(3), 1445–1461.

Cordeiro, G.M., M.C.S. Lima, A.H.M.A. Cysneiros, M.A.R. Pascoa, R.R. Pescim, and E.M.M. Ortega (2016). An extended Birnbaum-Saunders distribution: Theory, estimation, and applications. *Communications in Statistics-Theory and Methods 45*(8), 2268–2297.

Cordeiro, G.M., M.C.S. Lima, E.M.M. Ortega, and A.K. Suzuki (2018). A new extended Birnbaum-Saunders model: properties, regression and applications. *Stats 1*(1), 32–47.

da Silva, R.V., T.A.N. de Andrade, D.B.M. Maciel, R.P.S. Campos, and G.M. Cordeiro (2013). A new lifetime model: The gamma extended Fréchet distribution. *Journal of Statistical Theory and Applications 12*(1), 39–54.

de Santana, T.V.F., E.M.M. Ortega, G.M. Cordeiro, and G.O. Silva (2012). The Kumaraswamy-log-logistic distribution. *Journal of Statistical Theory and Applications 11*(3), 265–291.

Dey, S., D. Kumar, P.L. Ramos, and F. Louzada (2017). Exponentiated Chen distribution: properties and estimation. *Communications in Statistics-Simulation and Computation 46*(10), 8118–8139.

Khan, M.S., R. King, and I.L. Hudson (2018). Kumaraswamy exponentiated Chen distribution for modelling lifetime data. *Applied Mathematics and Information Sciences 12*(3), 617–623.

Kilbas, A.A., H.M. Srivastava, and J.J. Trujillo (2006). *Theory and applications of fractional differential equations*, Volume 204. Amsterdam: Elsevier.

Lemonte, A.J. (2014). The beta log-logistic distribution. *Brazilian Journal of Probability and Statistics 28*(3), 313–332.

Lima, M.C.S. (2015). *Mathematical properties of some generalized gamma models*. Ph. D. thesis, Universidade Federal de Pernambuco.

Marinho, P.R.D., R.B. Silva, M. Bourguignon, G.M. Cordeiro, and S. Nadarajah (2019). AdequacyModel: An R package for probability distributions and general purpose optimization. *PLOS ONE 14*(8), 1–30.

Mudholkar, G.S. and A.D. Hutson (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability 42*(2), 299–302.

Nadarajah, S., G.M. Cordeiro, and E.M.M. Ortega (2015). The ZografosâĂŞBalakrishnan-G Family of Distributions: Mathematical Properties and Applications. *Communications in Statistics - Theory and Methods 44*(1), 186–215.

Ortega, E.M.M., A.J. Lemonte, G.M. Cordeiro, and J.N. da Cruz (2016). The odd Birnbaum-Saunders regression model with applications to lifetime data. *Journal of Statistical Theory and Practice 10*(4), 780–804.

Pinho, L.G.B., G.M. Cordeiro, and J.S. Nobre (2012). The gamma-exponentiated Weibull distribution. *Journal of Statistical Theory and Applications 11*(4), 379–395.

Pogány, T.K., G.M. Cordeiro, M.H. Tahir, and H.M. Srivastava (2017). Extension of generalized integro-exponential function and its application in study of Chen distribution. *Applicable Analysis and Discrete Mathematics 11*(2), 434–450.

Pogány, T.K. and A. Saboor (2016). The gamma exponentiated exponential-Weibull distribution. *Filomat 30*(12), 3159–3170.

Qian, L (2012). The Fisher information matrix for a three-parameter exponentiated Weibull distribution under type II censoring. *Statistical Methodology 9*(3), 320–329.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sivini, P.G.L. (2006). *Desenvolvimento de banco de dados de confiabilidade: uma aplicação em estações redutoras de pressão de gás natural*. Master thesis, Universidade Federal de Pernambuco.

Tarvirdizade, B. and M. Ahmadpour (2019). A new extension of Chen distribution with applications to lifetime data. *Communications in Mathematics and Statistics 14*(5), 1–16.

Torabi, H. and N.M. Hedesh (2012). The gamma-uniform distribution and its applications. *Kybernetika 48*(1), 16–30.

Xiang, Y., S. Gubian, B. Suomela, and J. Hoeng (2013). Generalized Simulated Annealing for Global Optimization: The GenSA Package. *R Journal 5*(1), 13–28.

Xie, M, Y Tang, and T N Goh (2002). A modified Weibull extension with bathtub-shaped failure rate function. *Reliability Engineering & System Safety 76*(3), 279–285.

Zografos, K. and N. Balakrishnan (2009). On families of beta- and generalized gamma-generated distributions and associated inference. *Statistical Methodology 6*(4), 344–362.

Spanish Journal of Statistics
Vol. 2 No. 1 2020, Pages 41–77
doi:https://doi.org/10.37830/SJS.2020.1.04

Regular Article

Towards a modular end-to-end statistical production process with mobile network data

David Salgado[1, 2], Luis Sanguiao[1], Bogdan Oancea[3, 4], Sandra Barragán[1], Marian Necula[3]
[1]Department of Methodology and Development of Statistical Production, Statistics Spain (INE), Spain
[2]Department of Statistics and Operations Research, Complutense University of Madrid, Spain
[3]Department of Innovative Tools in Statistics, Statistics Romania (INS), Romania
[4]Department of Business Administration, University of Bucharest, Romania

Received: November 3, 2020. Accepted: March 1, 2021.

Abstract: Mobile network data has proved to be an outstanding data source for the production of statistics in general, and for Official Statistics, in particular. Similarly to another new digital data sources, this poses the remarkable challenge of refurbishing a new statistical production process. In the context of the European Statistical System (ESS), we substantiate the so-called ESS Reference Methodological Framework for Mobile Network Data with a first modular and evolvable proposed statistical process comprising (i) the geolocation of mobile devices, (ii) the deduplication of mobile devices, (iii) the statistical filtering to identify the target population, (iv) the aggregation into territorial units, and (v) the inference to the target population. The proposal is illustrated with synthetic data generated from a network event data simulator developed for these purposes.

Keywords: Statistical production, mobile network data, end-to-end process, geolocation, Deduplication, aggregation, inference

MSC: 62-07, 62P25, 62M05, 62F15

1 Introduction

Mobile network data, i.e. digital data generated in a mobile telecommunication network by the interaction between a mobile station (mobile device such as a smartphone or a tablet) and a base transceiver station (commonly known as antenna in an imprecise way) (Miao et al., 2016), constitutes a remarkable source of information for the production of statistics in Social Science, in general, and for Official Statistics, in particular. Many one-off studies can already be found in the literature with applications in different statistical domains (González et al., 2008; Ahas et al., 2010; Phithakkitnukoon et al., 2012; Calabrese et al., 2013; Deville et al., 2014; Louail et al., 2014; Iqbal et al., 2014; Blondel et al., 2015; Douglass et al., 2015; Pappalardo et al., 2016; Raun et al., 2016; Ricciato et al., 2017; Graells-Garrido et al., 2018; Wang et al., 2018) (see Salgado et al. (2020) for a more comprehensive list).

However, the production of official statistics in National and International Statistical Systems requests a standardized and industrialised statistical production process so that this new data source is fully integrated in the daily production framework of statistical offices. This raises remarkable challenges such as the data access conditions, new methodological and quality frameworks, a larger IT infrastructure (both in hardware and in software), a deep revision of the statistical disclosure control, and the identification of relevant aggregates (mostly included as part of legal regulations) for a diversity of stakeholders and users. Although a number of illustrative case studies dealing with official statistics can already be found in the literature (Debusschere et al., 2016; Williams, 2016; Nurmi, 2016; Izquierdo-Valverde et al., 2016; Dattilo et al., 2016; Senaeve and Demunter, 2016; Meersman et al., 2016; Reis et al., 2017; Sakarovitch et al., 2019; Galiana et al., 2018; Lestari et al., 2018), we still lack a production framework with a new statistical process.

In this line of thought, efforts in the international community (UN, 2017) and in the Europen Statistical System (ESS) (Ricciato, 2018) are under way to construct a production framework and some recent examples of an end-to-end statistical production process have been tested in a statistical office (Tennekes et al., 2020). The need for a detailed standardised and harmonised statistical process goes beyond the rise of new digital data sources, since a process-oriented production system instead of a product-oriented or even domain-oriented system is nowadays considered essential to achieve high-quality standards (UNECE, 2011). In this sense, the proliferation of one-off studies with new digital data in different statistical domains may be stressing the risk over statistical offices of reinforcing production silos, thus becoming clearly inefficient and making Official Statistics socially irrelevant (DGINS, 2018).

This article presents the fundamentals of a modular and evolvable statistical process with mobile network data to produce estimates for present population counts and origin-destination matrices as a concrete business case. This proposal constitutes the first step towards the construction of the so-called ESS Reference Methodological Framework for Mobile Network Data (see e.g. Ricciato, 2018), an initiative of the ESS embracing a set of principles to ensure consistency, reproducibilty, portability, and evolvability of data processing methods for this data source, to facilitate interworking between statistical offices and mobile network operators (MNOs) both at technical and organisational levels, and to adapt to the fast-changing technological environment of telecommunications by clearly detaching technology and statistical analysis.

We shall focus on the integral view of the process underlying its functional modularity and evolvability and on the methodological core bringing novel methods in Official Statistics with a clear goal of producing both estimates and their quality indicators (accuracy). We shall illustrate the whole process using synthetic network event data generated by a data simulator developed for these purposes. In section 2 we provide a general description of our approach setting up the general context under which this proposal is thought to be implemented. In section 3 we shall shortly describe the main functionalities of the network event data simulator as of this writing. In section 4 we provide the main contents of each of the modules comprising the statistical process, namely a generic description of the data in subsection 4.1, geolocation of mobile devices in subsection 4.2, deduplication of mobile devices carried by the same individual in subsection 4.3, statistical filtering of individuals in the target population in subsection 4.4, aggregation of device-level data into territorial units in subsection 4.5, and inference with respect to the target population in subsection 4.6. In section 5 we close with some conclusions and future prospects.

2 General description

The development, implementation, and monitoring of a statistical production process with mobile network data entail several complex and highly entangled issues. We need to solve questions regarding the access to data (including the integration with other data sources and even data from several MNOs), the development of statistical methods not traditional used in the production of official statistics, the according update and modernisation of the quality assurance framework, the deployment of the corresponding IT infrastructure, the professional and technical skills of staff necessary to execute this process, and the identification of the key target aggregates to be produced for the public good.

Official statistics play a key role in democratic socities for decision-taking and policy-making. For example, public fund allocation is usually conducted taking into account official population figures published by statistical offices. Thus, high-quality standards must be ensured and verified usually following international frameworks. In this context, in agreement with the ESS Reference Methodological Framework, an official statistical process with mobile network data must comprise the process design from the raw telecommunication data generated in the networks to the final statistical outputs. Acquiring aggregates or data at device-level from unknown preprocessing steps is not considered an option here. This is the first assumption motivating our proposal of an end-to-end process.

Mobile network data are extremely sensitive data and rightful concerns immediately arise to use them for statistical purposes. Data access is indeed an intricately complex set of legal, administrative, technical, and business issues, which we shall not dealt with here. Nonetheless, we assume three principles around which a final solution must be built:

- Privacy and confidentiality: as with any other official statistics produced from any data source by any statistical office, privacy and confidentiality of data holders and respondents must be assured. Indeed, stringent legal conditions have recently arisen to prevent privacy and confidentiality in the European context (European Parliament, 2016).
- Public good: there is an evident socioeconomic interest in extracting different insights from mobile network data valuable for the public good. This is as legitimate as the production of official statistics from traditional data sources.
- Private business interest: the production of statistical outputs and insights from mobile network data stands also as an increasing economic activity providing value and progress to the economy. Indeed, the digital data economy is targeted as a pillar in the European context.

All in all, an aligment of these three principles must be reached in practice. The proposed statistical process herein assumes that a collaborating scenario between statistical offices and MNOs through public-private partnerships, joint ventures, etc. is possible and leaves room for the design, execution, and monitoring of the different modules explained below.

In the context of the ESS, as of this writing no definitive agreement for a fully-fledged sustainable production of official statistics based on mobile network data has been reached between a national statistical office (NSO) and an MNO. Only specific short-term limited agreements for research have been reached[1]. This entails a shortage of data in NSOs to develop the statistical methodology,

[1] A remarkable exception is the compilation of international travel statistics for the balance of payments produced by the National Bank of Estonia (National Estonian Bank, 2020), not a statistical office, though.

the quality frameworks, and the software tools. Furthermore, given the extraordinarily rich and complex data ecosystem associated to a mobile telecommunication network, the identification of concrete data for statistical purposes must be undertaken (Radio network data? Core network data? Network management data? Call Detail Records?). In this sense, our strategy is to produce synthetic network event data together with a ground truth scenario so that all these aspects can be developed and investigated. In this way, more specific data requests can be formulated in agreement with the quality indicators and the ground truth computed in the simulated scenarios. Thus, our starting point will be the generation of these simulated scenarios.

A key feature of the ESS Reference Methodological Framework is the evolvability of the statistical process so that improvements and adaptations of the statistical methods to the underlying technological conditions is always possible and seamless. This justifies the approach of functional modularity (already present in modern proposals of traditional statistical processes (see e.g. Salgado et al., 2018)). By breaking the end-to-end process into modules according to the data abstraction principle we design transparent and independent production steps so that a change in one module will not affect the next module beyond the quality of the input/output interconnecting them through a standardised interface. In this proposal we do not include all necessary modules (e.g. data acquisition, substituted by the simulator) but only those core methodological stages (see Radini et al., 2020, for an architectural point of view):

- Geolocation.- This module focuses on the computation of location probabilities for each device across a reference grid used for the statistical analysis.
- Deduplication.- This module focuses on the computation of multiplicity probabilities for each device, i.e. probabilities of a given device to be carried by an individual jointly with one or several other devices. This is motivated by our interest on individuals of the target population, not on mobile devices.
- Statistical filtering.- This module focuses on the algorithmic identification of mobile devices of individuals of the target population such as domestic tourists, commuters, inbound tourists, etc.
- Aggregation.- This module focuses on the computation of probability distributions for the number of individuals detected by the network (i.e. with mobile devices) across different territorial units.
- Inference.- This module focuses on the computation of probability distributions for the number of individuals of the target population (even with no device) across different territorial units.

A cautious reader will immediately notice how the computation of probabilities is essential across the whole process. The use of probabilities, in our view, is jointly motivated by several relevant reasons. Firstly, probability distributions allow us to account for the uncertainty along the whole process, thus paving the way for the computation of quality indicators, especially those related to accuracy. Secondly, probability models provide a natural way to integrate data through priors and posteriors in a hierarchy of models. This is important because the combination of diverse data sources will not only produce statistical outputs with higher quality but it is also necessary in many cases, in particular, with mobile network data to avoid identifiability problems (see below). Thirdly, probability distributions stand as a flexible module interface between the successive production steps. In this line, we can use the total probability theorem to connect the original input data (raw telco data) with the final output data (population estimates):

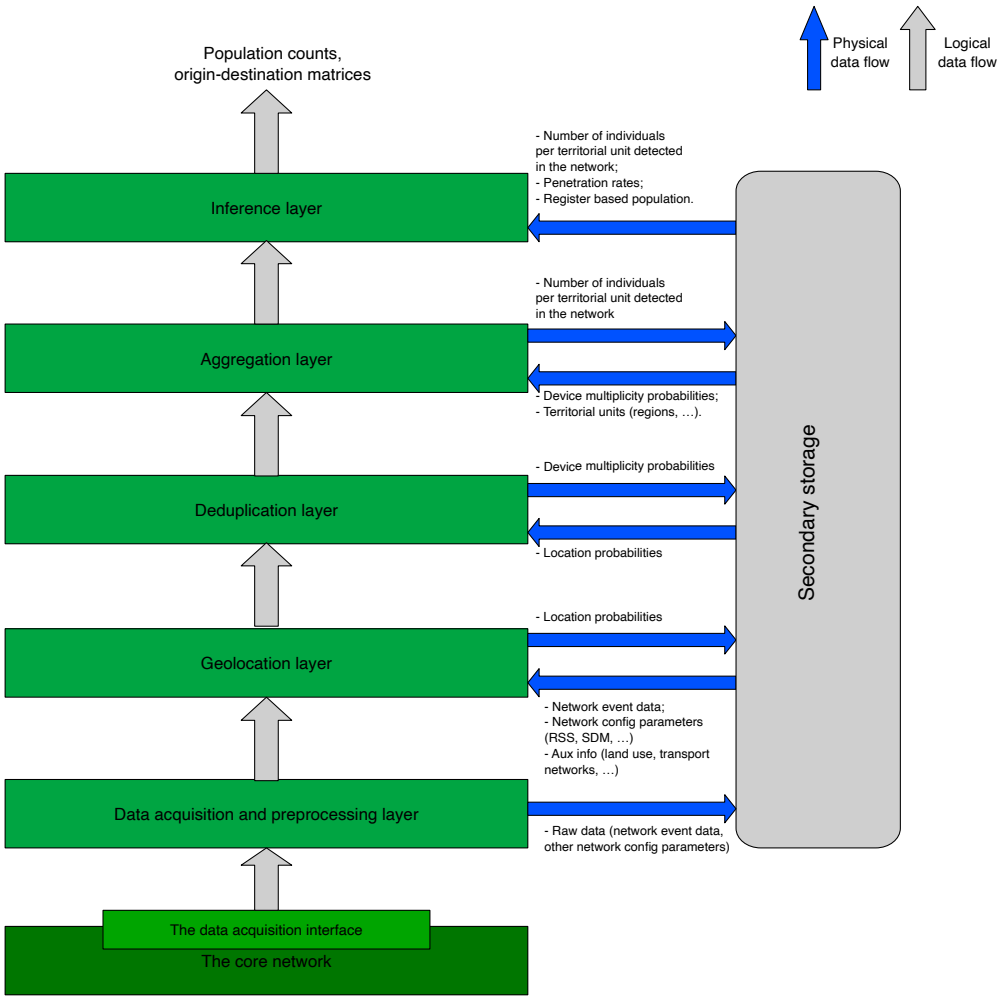

Figure 1: Modular structure of the statistical process and its software tools.

$$\mathbb{P}(z_{out}|z_{in}) = \int \mathrm{d}z_1 \int \mathrm{d}z_2 \cdots \int \mathrm{d}z_N \; \mathbb{P}(z_{out}|z_N) \cdots \mathbb{P}(z_2|z_1)\mathbb{P}(z_1|z_{in}). \tag{1}$$

The modular structure of the methodology is translated into a modular structure for the software tools. The choice of programming languages to develop these tools is motivated by multiple reasons. Firstly, software developed with the intention to be used in the future should be portable at the level of source code. Thus, portability is our first consideration. Secondly, our goal is to produce a software for statisticians, not for computer scientists. Thus, the language(s) of the implementation should be familiar for statisticians and easy to use by them. Thirdly, in the line of software development in the ESS, we planned to use only open source tools like libraries, IDEs, debuggers, profilers, etc. to maintain the software development process under a strict control regarding the associated costs. Moreover, the programming language(s) together with these tools should have a large community of programmers and users which can be seen as a free technical support. Fourthly, the programming language(s) should have support for parallel and distributed computing. Since all the algorithms involved by the our methodological approach are computational intensive, and the size of mobile network data could be very large, this is a mandatory requirement. Last but not least

important, the criteria of programming efficiency and resources needed to run the software even on normal desktops/laptops are also considered.

After analysing different choices, eventually we came to the following two software ecosystems: R (R Core Team, 2020) or Python (Van Rossum and Drake, 2009). Both systems meet our criteria and have a large community of users but while Python is considered to be more computationally efficient, R is better suited for statistical purposes and it seems to gain ground among the official statistics community (Templ and Todorov, 2016; Kowarik and van der Loo, 2018). Since our target audience is the official statistics community, we decided to develop our software modules using R since it has a huge number of available packages, it has support for parallel and distributed processing, it can be easily interfaced and work together with high performance languages like C++ when the performance of plain R is not enough, it can be easily interfaced with computing ecosystems widely used in the Big Data area such as Hadoop (White, 2009) or Spark (Zaharia et al., 2016) and there are several packages allowing a neat interface between R and these systems (Oancea and Dragoescu, 2014; Venkataraman et al., 2016) which means that, if needed, all modules in our software stack can be easily integrated with such systems for a production pipeline.

Thus, to execute the process with simulated data, we have developed an R package for each module implementing the corresponding statistical methods. With a view on scalability through distributed computing and parallelization, we use secondary memory instead of main memory to pass input and output data between modules as well as execution parameters (see figure 1). In the next section we provide details about the contents of each module.

3 Network event data simulator

The simulator is a highly modular software (Oancea et al., 2019) implementing agent-based simulating scenarios with different elements configured by the user. The basic elements are:

- a geographical territory represented by a map;
- a telecommunication network configuration in terms of a radiowave propagation model;
- a population of individuals carrying 0, 1, or 2 mobile devices during their displacement;
- a displacement pattern for individuals;
- a reference grid for analysis.

The simulator works essentially by using a radiowave propagation model (Shabbir et al., 2011) to simulate the connection between the base transceiver stations (loosely, antennas) and each mobile station (device) during the displacement of each carrying individual. The connection mechanism is an extreme simplification of the real world extracting the essential features for statistical analysis. The core output data consists of a time sequence of cell IDs (loosely, antenna IDs) and network event codes (connection, disconnection, etc.) for each device along the duration of the simulation. We simulate signalling data (i.e. passive data not depending on subscribers' behaviour) instead of Call Detail Records or any other active data generated by individuals (call, SMS, Internet connections, ...).

For the time being, since our priority is the simulator as a whole, the different elements implemented so far are kept as simple as possible. Firstly, displacement patterns of individuals are basically a sequence of stays (no movement) and random walks with/without a drift with two

possible speeds (namely, walk and car speeds). The drift, the speeds, and the shares of individuals with 0, 1, and 2 devices are easily configured by the user. Only closed populations can be simulated so far, i.e. individuals cannot abandon or enter into the territory under analysis.

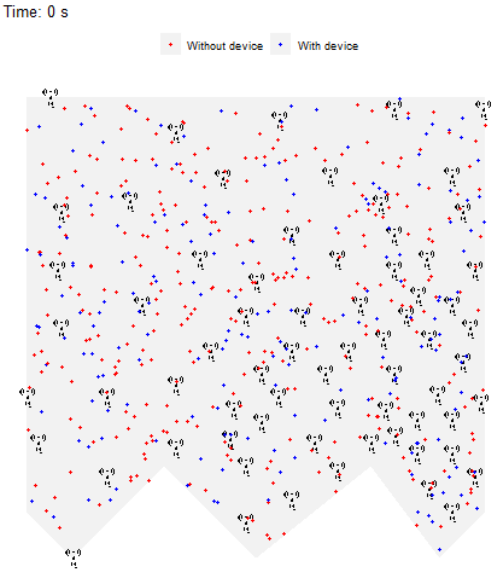

Figure 2: **Animation.** Positions of 70 antennas and drifted displacement pattern of individuals.

Secondly, an extremely simplified radiowave propagation model and a variant thereof is used in terms of the Received Signal Strength (RSS – expressed in dBm), the distance r between the BTS and the device, the emission power P, the so-called path exponent γ (quantifying the loss of signal strength) and some geometrical parameters regarding the BTS orientation (only for directional antennas (see e.g. Tennekes et al., 2020)). For omnidirectional antennas, the model is simply expressed by

$$\mathrm{RSS}(r) = 30 + 10 \cdot \log_{10}(P) - 10 \cdot \gamma \cdot \log_{10}(r). \tag{2}$$

Each device connects to the antenna producing the highest signal strength in each tile until the antenna reaches its maximum capacity. Both the emission power and the path loss are selected as input parameters by the user. A convenient variant introduced by Tennekes et al. (2020) performs a parameterised logistic transformation upon RSS producing the so-called Signal Dominance Measure:

$$\mathrm{SDM}(r) = \frac{1}{1 + \exp\left(-S_{\mathrm{steep}} \cdot (\mathrm{RSS}(r) - S_{\mathrm{mid}})\right)}, \tag{3}$$

where S_{steep} and S_{mid} are chosen according to characteristics of each radio cell. Each device connects to the antenna providing the highest signal dominance measure in each tile until the antenna reaches its maximum capacity. Both S_{steep} and S_{mid} are selected as input parameters by the user, too.

Figure 3 represents the RSS and the SDM for a given antenna in an arbitrary territory depicted as an irregular polygon with a 10 km $\times$ 10 km bounding box.

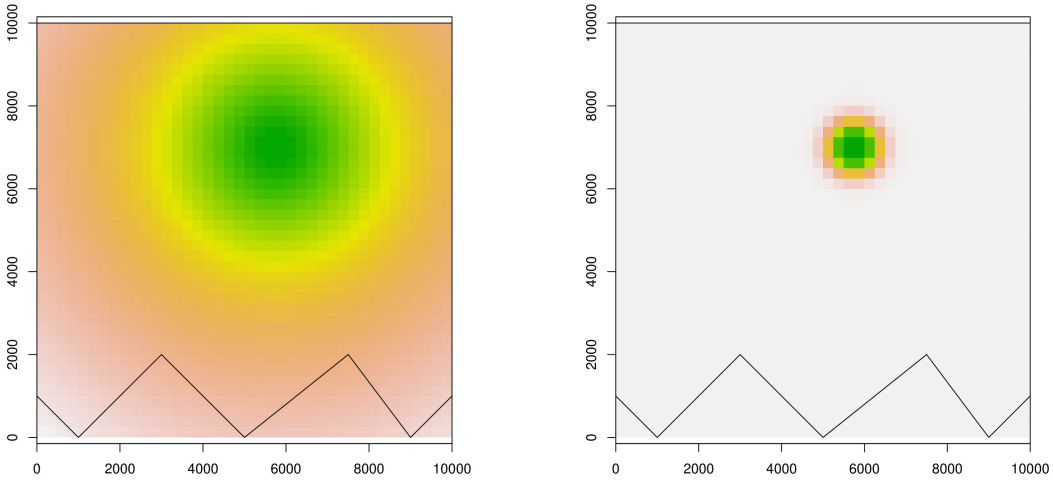

Figure 3: Received signal strength and signal dominance measure for an omnidirectional antenna according to models (2) and (3).

4 Production modules

4.1 Data

When contacting MNOs to access data, the first reaction from telecommunication engineers and data engineers in these companies is to ask "what data?" The data ecosystem of a mobile telecommunication network is extremely complex, derived from its nested cellular structure (see figure 4). Thus, a first step to use mobile network data for statistical purposes is to substantiate the meaning of these data. In this line, the use of a synthetic simulator allows us to devise an end-to-end process and to set up an empirical criterion about specific data to compile statistics accurate enough for official purposes.

Our proposed process helps us to provide a first typology of data required to reach our goal. We identify three types of data (according to organisation which generates them).

4.1.1 Mobile network data

Under this category we embrace two sorts of data related to mobile telecommunication networks. On the one hand, we need data about the configuration of the network. Basically, these are parameters entering the radiowave propagation models used in subsequent stages (see below) such as emission powers, path loss exponents, frequencies and frequency correction factors, base station heights and azimuths,... Notice that these variables do not contain information about the subscribers but they are extremely sensitive for MNOs due to the highly competitive degree of the telecommunication market. Ultimately, the variables to access will depend on the chosen model, which should be in principle chosen according to the accuracy of the final estimates and the associated acquisition costs under the public-private agreement. Access to these data does not mean whatsoever that these data should be made public or even that they have to leave MNOs'

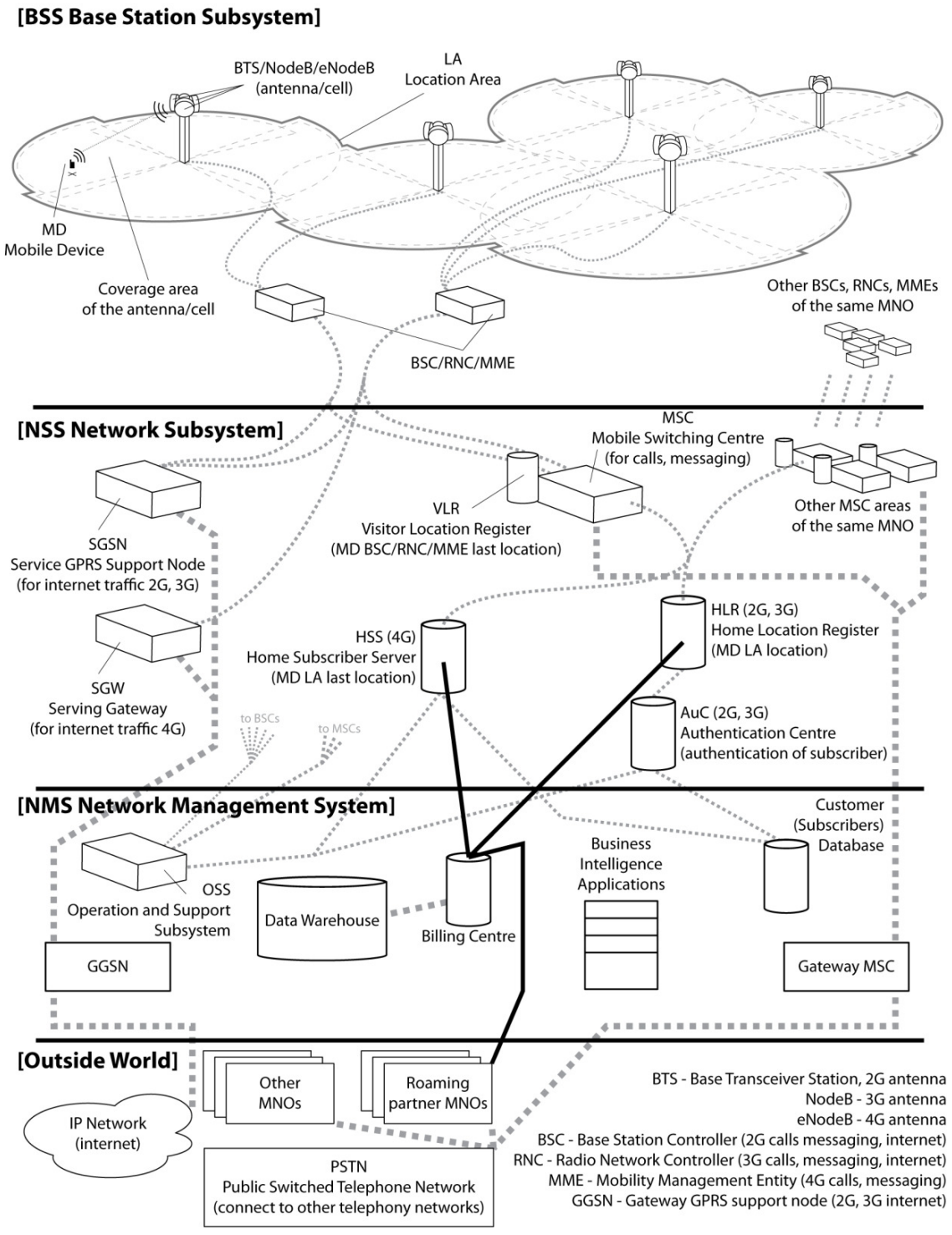

Figure 4: Nested cellular structure of a GSM-like network (taken from Positium (2016)).

information systems. This sensitive information must be kept protected also by NSOs and they are just required to be accessed to produce specific outputs in later stages. Agreements on computing these outputs and their sharing into the statistical process should be enough for our goals (see below).

On the other hand, so-called network event data generated by each mobile station (device) in the network must be accessed. These can be variables such as the cell ID (identifying the cell or sector whether the interaction between a device and a connecting antenna is established), the Time of Arrival (basically collecting the time for a signal to reach a mobile device from the connecting antenna), the Angle of Arrival (measuring the angle of the line-of-sight of a device from the connecting antenna),... These data do contain sensitive information about the subscribers. Again, not only must they be kept private but also they must be preprocessed in the MNOs' original information systems (i.e. no transmission whatsoever to NSOs). Identifying precisely what variables to use will ultimately depend on the accuracy of final estimates and associated accessing and preprocessing costs. Once more, details must be part of agreements between MNOs and NSOs.

In the illustrative example with the data simulator below we will use the emission power and path loss exponent of each base station (network configuration) together with the cell ID of each connection/signal transmission/disconnection and orientation parameters between devices and base stations every 10 seconds.

4.1.2 Auxiliary NSO information on target aggregates

This is information produced by NSOs themselves, thus providing profuse access to microdata for alternative (possibly undisclosed) aggregations in finer territorial units. They may be survey microdata, administrative data, or aggregates from any combination of sources with a relevant relationship with the target outputs of our analysis.

In the illustrative example with the data simulator below to produce present population counts and general-mobility origin-destination matrices we shall use data from the current population register or some other similar demographic operation. It is important to state that the treatment of both data sources makes a difference on their role. Whereas mobile network data will be used as the central source to produce outputs (thus gaining in both spatial and time breakdowns), the population register will enter as an auxiliary prior data source. An equal-footing integration of all data sources to produce, modify, and correct the population register is not pursued here.

4.1.3 Auxiliary (public) information on the geographic territory

As with the production of any other official statistics, the more available information to integrate, the higher expected quality for the output. In this sense, auxiliary information from (usually public) organizations such as land use or transport network configurations and schedules may be profitably integrated in the modelling exercise. For example, for the geolocation of mobile devices, prior location probabilities upon grid tiles can be fixed according to the land use features of each tile. In the wilderness this probabilities will differ a great deal from those in the city centers.

In the illustrative example below, since the geographical territory is just an arbitrary irregular polygon, we shall not use any prior information about land use or transport network. Every tile will be similar to each other.

4.1.4 Privacy-preserving data technologies

As an immediate side-effect of this complex and sensitive data ecosystem, the integration of information in stringent privacy-preserving conditions is a must. A research avenue clearly seems to be arising extending the traditional statistical disclosure control from output aggregates to also input and intermediate data.

This brings the privacy-preserving technologies (Zhao et al., 2019) into scene. However, we would like to pose the following reflection. When considering mobile network data (and probably similarly sensitive new digital data), we detect a change in society about the role of statistical officers in producing official statistics. With more traditional data sources such as survey data and administrative data, statistical officers are undisputedly endowed with the legitimacy of accessing, processing, and integrating ***personal data*** from these diverse sources. Take e.g. the construction of a business register where sensitive information from all business units in a country are compiled for further use in the statistical production process. No privacy-preserving technique is demanded in this case, in spite of which privacy and confidentiality is completely guaranteed and statistical disclosure control is fully effective. In our view, statistical offices must reclaim their traditional role as secure recepcionist of information for the public good.

However, having said this, the challenge of integrating MNOs into the statistical production process includes the management of trust and privacy-preserving techniques stand as an excellent tool in this sense.

4.2 Geolocation

The utility of mobile network data to produce statistics for the public good arises at least from three aspects. Firstly, the geospatial nature of this information makes it ideal to provide population counts and mobility-related statistics at an unprecedented spatial and time breakdown. Different social groups can be targeted (tourists, commuters, present population, etc.) provided algorithms are put into place to identify them within the datasets. Secondly, Internet traffic and the nature of donwloaded mobile apps can provide relevant insights for social analysis (see e.g. Ucar et al., 2019). Finally, and more interestingly in our view, mobile network data can provide an excellent source of network data, i.e. interactions between population units, thus paving the way for the use of network science in the production of novel statistical outputs.

Currently, the main focus of research is centered on the geolocation of mobile devices. Originally, Voronoi tessellations of the geographical territory under analysis were used to partition this territory into disjoint tiles assigning each one to a BTS. In our view, this is an oversimplification of the network, since coverage areas and sector cells of each BTS can often be intersecting (even nested) and directional. To overcome this complexity, we divide the territory into a grid of tiles and using radiowave propagation models compute the so-called event location probabilities $\mathbb{P}(\mathbf{E}_{dt} = \mathbf{e}_j \big| T_{dt} = i)$, i.e. the probability that a device d produces network event data $\mathbf{e}_j$ (e.g. the cell ID of a given BTS to which the device is connected) conditioned on being located at tile i. This conditional probability is used to compute the reverse so-called posterior location probability $\gamma_{dti} = \mathbb{P}(T_{dt} = i \big| \mathbf{E}_d)$ at each time t and each device d. The posterior joint location probabilities $\gamma_{dtij} = \mathbb{P}(T_{dt} = i, T_{dt-1} = j \big| \mathbf{E}_d)$ are also of interest for later modules. Notice that we condition upon all available network information

$\mathbf{E}_d = \{\mathbf{E}_{dt}\}_{t=0,1,\ldots}$

A first direct approach is to make use of Bayes' theorem together with the prior location probabilities $\mathbb{P}(T_{dt} = i)$ (computed according to the prior auxiliary information such as land use or transport network information): $\mathbb{P}(T_{dt} = i \big| \mathbf{E}_{dt} = \mathbf{e}_j) \propto \mathbb{P}(\mathbf{E}_{dt} = \mathbf{e}_j \big| T_{dt} = i) \cdot \mathbb{P}(T_{dt} = i)$. This is the static approach followed by Tennekes et al. (2020).

A superseding alternative is to consider the dynamical behaviour of individuals in the population and to postulate a generic transition model across the reference grid, which together with the event location probabilities computed above, enter into a hidden Markov model (HMM) as transition and emission models, respectively. Upon estimation of these model parameters, we can compute the posterior location probabilities for each device d (see figure 5). Mathematical details are provided in the appendix.

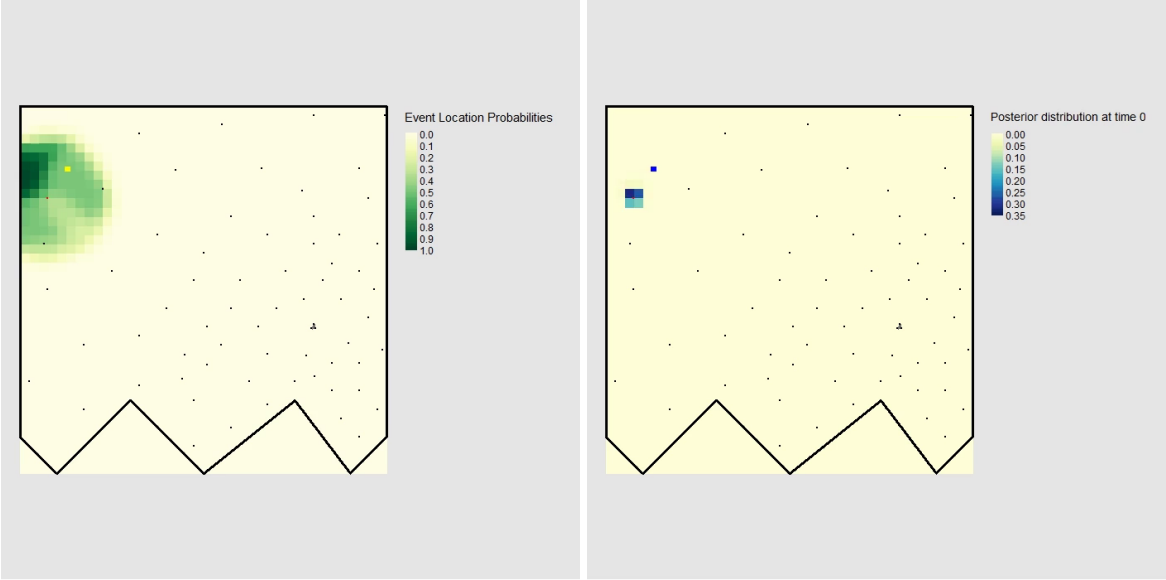

Figure 5: **Animation.** Event location probabilities [left] and posterior location probabilities [right] for a given device. True position also included.

The use of HMMs in the context of this reference grid for analysis is notably versatile and provides a generic framework to deal with multiple aspects. Firstly, at this initial stage of the project, we have defined the HMM state just as the location in the grid, but more complex states can be possibly defined taking into account the velocity, the transport mode, or a classification of anchor points (home, work, second residence, etc.). Secondly, the emission model (i.e. the event location probabilities) is built independently of the transition model, which allows MNOs to concentrate the processing of sensitive network information (antenna localizations, network parameters, etc.) on a this concrete production step. For the HMM, only the output of this step is needed, thus making it possible to undisclose and protect this sensitive information. Finally, the use of probabilities allows us to take into account the uncertainty in the estimation process from the onset. Indeed, we can define familiar accuracy indicators for the geolocation such as bias, standard deviation, and mean squared error as with traditional survey data (see appendix).

4.3 Deduplication

Since we focus on estimating population counts of individuals, not of mobile devices, we need to detect which terminals are carried over by the same individuals. We call it device multiplicity. The goal is to compute a device-multiplicity probability $p_d^{(n)}$ for each device d to be carried over by the same individual together with a total of n devices. In our simulated scenario, for computational ease, with limit the number of devices per individual to 2. Thus, we aim at computing $(p_d^{(1)}, p_d^{(2)} = 1 - p_d^{(1)})$, i.e. the probability that a device d belongs to an individual with 1 or 2 devices, respectively.

The problem of device duplicity has been often recognised as an overcoverage problem. It is usually considered ***after*** the aggregation step producing **number of devices** per territorial area and time interval. Once this aggregation step has been conducted, the challenge is really serious and may easily drive us into an identifiability problem (Lehmann and Casella, 2003) in any model estimating the number of individuals from the number of devices. The reader may easily be convinced with a simple example. Consider a population of $N^{(D)} = 10$ devices, all corresponding to a different individual, i.e. $N = 10$. Consider another population of $N^{(D)} = 10$ devices, where each individual has two devices, i.e. $N = 5$. There is no possible statistical model using only the variable $N^{(D)}$ possibly distinguishing between these two situations. In other words, we run into an identifiability problem unless more parameters are introduced, which will require the use of auxiliary information. In this simple case, we may think of a statistical model based on $(N^{(D)}, R_{dup})$ where we have introduced another parameter R_{dup} standing for the duplicity rate in the population. With these variables, the identifiability problem ameliorates, but the model complexity increases, apart from the issue about data availability (is R_{dup} really available?).

This is why we recommend to address this problem **before the aggregation step**. This has direct implications for the access agreements. According to this recommendation, the number of devices is not a target dataset in the statistical process and the device multiplicity issue must be addressed upon individual information at the device level, thus ideally in MNOs' premises (together with the geolocation step).

Another important consideration arises when considering uncertainty. It is important to remind that we target at the probability $p_d^{(n)}$ of each device d. This probability distribution will indeed be another intermediate distribution in the chain (1). We need to assess the **uncertainty** (i.e. probabilities) and not just to conduct a classification. The relevance of this will be evident in the aggregation step later on.

We have proposed two alternative approaches. On the one hand, we resort to Bayesian reasoning to test the hypothesis that two given devices d_1 and d_2 belong to the same individual. Let us denote by H_{dd} the hypothesis that device d uniquely corresponds to an individual, whereas $H_{d_1 d_2}$ stands for devices d_1 and $d_2 \neq d_1$ belonging to the same individual. Thus, we need to compute $p_d^{(1)} = \mathbb{P}\left(H_{dd} \middle| \mathbf{E}, \mathbf{I}^{\text{aux}}\right)$, where $\mathbf{E} = \{\mathbf{E}_{dt}\}_{t=0,\dots,T}^{d=1,\dots,D}$ is all network event information. We propose two procedures:

- Pair computation.- We compute $p_d^{(1)} = 1 - \max_{d' \neq d} \mathbb{P}\left(H_{dd'} \middle| \mathbf{E}_d, \mathbf{E}_{d'}, \mathbf{I}^{\text{aux}}\right)$, where

$$\mathbb{P}(H_{dd'}|\mathbf{E}_d,\mathbf{E}_{d'},\mathbf{I}^{\text{aux}}) = \frac{\mathbb{P}(\mathbf{E}_d,\mathbf{E}_{d'}|H_{dd',\mathbf{I}^{\text{aux}}})\,\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})}{\mathbb{P}(\mathbf{E}_d|H_{dd},\mathbf{I}^{\text{aux}})\,\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}}) + \mathbb{P}(\mathbf{E}_d,\mathbf{E}_{d'}|H_{dd'},\mathbf{I}^{\text{aux}})\,\mathbb{P}(H_{dd'}|\mathbf{I}^{\text{aux}})}, \tag{4}$$

with $\mathbb{P}(H_{dd'}|\mathbf{I})$, $\mathbb{P}(H_{dd}|\mathbf{I}^{\text{aux}})$ being prior probabilities and $\mathbb{P}(\mathbf{E}_d,\mathbf{E}_{d'}|H_{dd'},\mathbf{I}^{\text{aux}})$, $\mathbb{P}(\mathbf{E}_d|H_{dd},\mathbf{I}^{\text{aux}})$ standing for the likelihoods under each hypothesis, respectively.

- One-to-one computation.- Alternatively, posing $\Omega_d = \bigcup_{d'=1}^{D} H_{dd}$, we compute

$$p_d^{(1)} = \frac{\mathbb{P}\left(\mathbf{E}_d\middle|H_{dd},\mathbf{I}^{\text{aux}}\right)\cdot\mathbb{P}\left(H_{dd}\middle|\mathbf{I}^{\text{aux}}\right)}{\mathbb{P}\left(\mathbf{E}_d\middle|H_{dd},\mathbf{I}^{\text{aux}}\right)\cdot\mathbb{P}\left(H_{dd}\middle|\mathbf{I}^{\text{aux}}\right) + \sum_{d'\neq d}\mathbb{P}\left(\mathbf{E}_d,\mathbf{E}_{d'}\middle|H_{dd'},\mathbf{I}^{\text{aux}}\right)\cdot\mathbb{P}\left(H_{dd'}\middle|\mathbf{I}^{\text{aux}}\right)}. \tag{5}$$

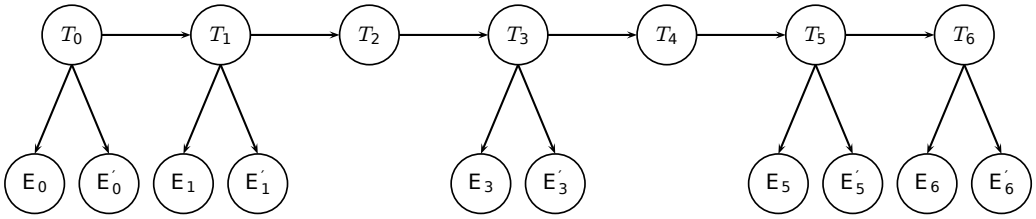

Figure 6: Extended HMM to compute $\mathbb{P}(\mathbf{E}_d,\mathbf{E}_{d'}|H_{dd'},\mathbf{I}^{\text{aux}})$ for a given device d (subscript not included in the graphical model).

In both procedures the probabilities $\mathbb{P}(\mathbf{E}_d|H_{dd},\mathbf{I}^{\text{aux}})$, $\mathbb{P}(\mathbf{E}_d,\mathbf{E}_{d'}|H_{dd'},\mathbf{I}^{\text{aux}})$ are computed with the original HMM and the extended HMM represented in figure 6, respectively. Priors are computed incorporating prior information e.g. from the Customer Relationship Management Database or any other complementary information (see Salgado et al., 2020, for some details).

On the other hand, instead of focusing on the network event variables $\mathbf{E}_{dt}$, we can make use of the random location $\mathbf{R}_{dt} \in \{\mathbf{r}_i^{(c)}\}_{i=1,\ldots,N_T}$ estimated according to the posterior location probabilities γ_{dti}. Then, we can follow the same approach as the Bayesian pair computation case (4) substituting $\mathbf{E}_{dt}$ by $\mathbf{R}_{dt}$ (see Salgado et al. (2020) for details).

In figure 7 we show the results for the Bayesian one-to-one case for our illustrative example. The ROC curves show an excellent performance for the classification of devices according to their duplicity with values of the area under the curve (AUC) above 0.95. Using the simulated ground truth and a threshold of 0.50 we can also notice that very few false positive cases result (and they are due to the short period of time under analysis: basically two individuals following nearly the same sequence of coverage areas), whereas the number of false negative cases are a bit notable. This is due to devices of different individuals staying under the same coverage area during the time period: they are wrongly classified as duplicity cases of analysis. Realistic time periods of analysis will hopefully avoid these problems.

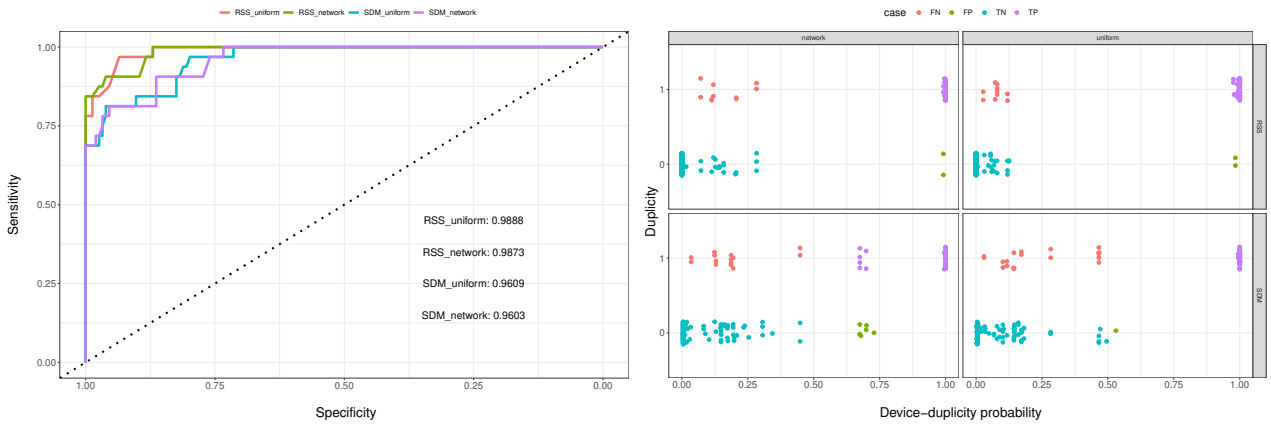

Figure 7: [Left] ROC curve for two emission models (RSS and SDM) and two HMM priors (uniform and network). [Right] Cases for two emission models (RSS and SDM) and two HMM priors (uniform and network).

4.4 Statistical filtering

As of this writing, this module is the less developed since more complex and realistic displacement patterns are needed in the simulator to study and analyse different proposals. We limit ourselves to provide a generic view. Again, we shall be focusing on analyses upon the geolocation data, i.e. upon the network event data and location probabilities derived thereof.

First of all, the target mobile network data is assumed to be basically some form of signalling data so that time frequency and spatial resolution are high enough as to allow us to analyse movement data in a meaningful way. In this sense, for example, CDR data only provides information up to a few records per user in an arbitrary day which makes virtually impossible any rigorous data-based reasoning in this line.

The use of HMMs implicitly incorporates a time interpolation which will be very valuable for this statistical filtering exercise. In this way we avoid the issues arising from noncontinuous traces approaches (see e.g. Vanhoof et al., 2018, for home location algorithms). However, a wider analysis is needed to find the optimal time scope. In turn, the spatial resolution issue is dealt with by using the reference grid. This releases the analyst from spatial techniques such as Voronoi tessellation, which introduces too much noise for our purposes. Nonetheless, the uncertainty measures computed from the underlying probabilistic approach for geolocation must be taken into account to deal with precision issues in different regions (e.g. high-density populated vs. low-density populated).

In our view, the algorithms for statistical filtering should be mainly based on quantitative measures of movement data. In particular, from the HMMs fitted to the data (especially the location probabilities) we propose to derive a probability-based coarse-grained trajectory per device which will be the basis for these algorithms. Once a trajectory is assigned to each device, different indicators and measures of movement shall be computed upon which we shall apply algorithms to determine important concepts such as usual environment, home/work location, second home

location, leisure activity times and locations, etc.

A critical issue in the development of this kind of algorithms is the validation procedure. On the one hand, the use of the simulator, once more complex and realistic displacement patterns have been introduced, will offer us in the future a validation against the simulated ground truth. On the other hand, with real data two main problems need to be tackled, namely (i) the use of pseudoanonymised real data will prevent us to link mobile device records with official registers, so only indirect aggregated validation procedures can be envisaged, and (ii) the representativity of the tested sample of devices (e.g. using GPS signals) to validate the algorithm for the whole population needs to be rigorously assessed.

Thus, the starting point is the construction of a probability-based coarse-grained trajectory for each device. In our geolocation model, the state of the HMM was defined in terms of the tile where the device is positioned. Thus, the concept of trajectory follows immediately as the time sequence of states, in which we shall use the coordinates of each tile to build the so-called *path* $\{(x_{dt_0}, y_{dt_0}), (x_{dt_1}, y_{dt_1}), \ldots, (x_{dt_N}, y_{dt_N})\}$, where at each time instant t_i the spatial coordinates x_{dt_i} and y_{dt_i} for device d are specified. In more complex definitions of states, another procedure should lead us to deduce the path from the adopted concept of HMM state.

Given an HMM, it is well-known that at least two different methods can be approached to build a sequence of states, i.e. a trajectory in our case. We can compute either the most probable sequence of states or the sequence of most probable states. In mathematical terms, the former is the sequence

$$T^*_{dt_0:t_N} = \text{argmax}_{T_{dt_0:t_N}} \mathbb{P}\left(T_{dt_0:t_N} \middle| \mathbf{E}_{dt_0:t_N}\right), \tag{6}$$

which can be computed by means of the Viterbi algorithm (see e.g. Murphy, 2012). The second method is indeed given by

$$T^*_{dt_0:t_N} = \left(\text{argmax}_{T_{dt_0}} \gamma_{dt_0}, \text{argmax}_{T_{dt_1}} \gamma_{dt_1}, \ldots, \text{argmax}_{T_{dt_N}} \gamma_{dt_N}\right), \tag{7}$$

where $\gamma_{dt_j} = \mathbb{P}\left(T_{dt_j} \middle| \mathbf{E}_{dt_0:t_N}\right)$ are the posterior location (state) probabilities.

We choose the maximal posterior marginal (MPM) trajectory because it is more robust and because unimodal probabilities are expected so that differences will not be large (Murphy, 2012). Furthermore, coherence with other process modules (e.g. duplicity) using the posterior location probabilities is favoured in this way.

Once a path is assigned to each device we can compute different indicators as well as joint measures. Following Long and Nelson (2013) (see also multiple references therein) we distinguish the following groups of measures:

- Time geography.- This represents a framework for investigating constraints such as maximum travel speed on movement in both the spatial and temporal dimensions. These constraints can be capability constraints (limiting movement possibilities because of biological/physical abilities), coupling constraints (specific locations a device must visit thus limiting movement possibilities), and authority constraints (specific locations a device cannot visit thus also limiting movement possibilities).

- Path descriptors.- These represent measurements of path characteristics such as velocity, acceleration, turning angles. By and large, they can be characterised based on space, time, and space-time aspects.
- Path similarity indices.- These are routinely used to quantify the level of similarity between two paths. Diverse options exist in the literature, some already taking into account that paths are sequences of stays and displacements (see e.g. Long and Nelson, 2013).
- Pattern and cluster methods.- These seek to identify spatialâĂŞtemporal patterns from the whole set of paths. These are mainly used to focus on the territory rather than on individual patterns. They also consider diverse aspects on space, time, and space-time features.
- IndividualâĂŞgroup dynamics.- This set of measures compile methods focusing on individual device displacement within the context of a larger group of devices (e.g. a tourist within a larger group of tourists in the same trip).
- Spatial field methods.- These are based on the representation of paths as space or space-time fields. Different advanced statistical methods can be applied such as kernel density estimation or spatial statistics.
- Spatial range methods.- These are focused on measuring the area containing the device displacement, such as net displacement and other distance metrics.

Diverse indicators can be defined and used within each group (see Salgado et al. (2020) for preliminary examples on our simulated scenario). Further analysis is needed with realistic displacement patterns. With a selected set of movement indicators, we shall be able to provide a computational algorithm to substantiate the concepts of usual environment, home/work location, second home location, etc. Notice that the definition of state for the HMM could be enhanced using these concepts, thus incorporating more information into the geolocation estimation.

4.5 Aggregation

The next step is to aggregate the preceding information at the level of territorial units of analysis. These territorial units usually come from an administrative division of the geographical territory, but in general terms they will be undestood as aggregation of tiles of the reference grid. In this sense, when deciding about the choice of grid, it is highly recommended that the territorial units of analysis are taken into account from the onset. Obviously, the smaller the tiles, the higher the flexibility to define different granular levels of the territorial units.

The bottom line the aggregation step is to avoid making further modelling hypothesis as much as possible. In this line, we use probability theory to define and compute the probability distribution for the number of individuals (not devices) detected by the network using both the posterior location and device-duplicity probabilities.

It is important to make the following general remarks about our approach. Firstly, the aggregation is on the number of ***detected individuals***, not on the number of devices. This is a very important difference with virtually any other approach found in the literature (see e.g. Deville et al., 2014; Douglass et al., 2015). We take advantage of the preceding modules working at the device level to study in particular the duplicity in the number of some devices per individual. This has strong implications regarding agreements with MNOs to access and use their mobile network data for statistical purposes. The methodology devised in the preceding section to study this duplicity (or variants thereof) needs to be applied **before** any aggregation. As we can easily see, working with the number of devices instead of the number of individuals poses severe identifiability problems

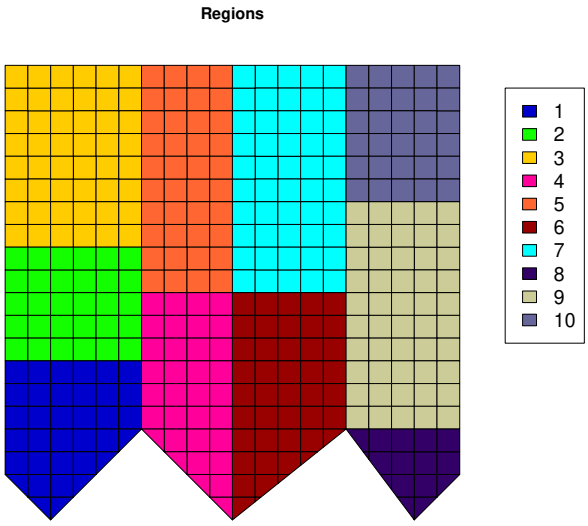

Figure 8: Territorial regions.

requiring more auxiliary information. Let us consider an extremely simplified illustrative example. Let us consider population U_1 of 5 individuals with 2 devices each one and population U_2 of 10 individuals with 1 device each one. Suppose that in order to we make our inference statement about the number N of individuals in the population we build a statistical model relating N and the number of devices $N^{(\mathrm{d})}$, that is, basically we have a probability distribution $\mathbb{P}_N(N^{(\mathrm{d})})$ for the number $N^{(\mathrm{d})}$ of devices dependent on the number of individuals, from which we shall infer N. In this situation we have $\mathbb{P}_{N^{(1)}} = \mathbb{P}_{N^{(2)}}$ even when $N^{(1)} \neq N^{(2)}$. There is no statistical model whatsoever capable of distinguishing between U_1 and U_2 (see Definition 5.2 by Lehmann and Casella, 2003, for unidentifiable parameters in a probability distribution). To cope with the duplicity of devices using an aggregated number of devices we would need further auxiliary information, which furthermore must be provided at the right territorial and time scales.

Secondly, we shall use the language of probability in order to carry forward the uncertainty already present in the preceding stages all along the end-to-end process. In another words, if the geolocation of network events is conducted with certain degree of uncertainty (due to the nature itself of the process) and if the duplicity of a given device (carried by an individual with another device) is also probabilistic in nature, then a priori it is impossible to provide a certain number of individuals[2] in a given territorial unit. For this reason, we shall focus on the probability distribution of the number of individuals detected by the network and shall avoid producing a point estimation. Notice that having a probability distribution amounts to having all statistical information about a random phenomenon and you can choose a point estimation (e.g. with the mean, the mode or the median of the distribution) together with an uncertainty measure (coefficient of variation, credible intervals, etc.).

Thirdly, the problem is essentially multivariate and we must provide information for a set of territorial units. Thus, the probability distribution which we shall provide with our proposed aggregation step must be a multivariate distribution. Notice that this is not equivalent to providing a collection of marginal distributions over each territorial unit. Obviously, there will be a correlation

[2]Notice that this same argument is valid for the number of devices.

structure, the most elementary expression of which is that individuals detected in a given territorial unit cannot be detected in another region, so that the final distribution needs to incorporate this restriction in its construction.

Finally, the process of construction of the final multivariate distribution for the number of detected individuals must make as few modelling assumptions as possible, if any. In case an assumption is made (and this should be accomplished in any use of statistical models for the production of official statistics, in our view), it should be made as explicit as possible and openly communicated and justified. In this line of thought, we shall strongly based the aggregation procedure on the results of preceding modules avoiding any extra hypothesis. Basically, our starting assumptions for the geolocation and the duplicity detection will be carried forward as far as possible without introducing new modelling assumptions of any kind.

To implement the principles outlined above, we shall slightly change the notation used in preceding chapters. Firstly we define the vectors $\mathbf{e}_i^{(1)} = \mathbf{e}_i$ and $\mathbf{e}_i^{(2)} = \frac{1}{2} \cdot \mathbf{e}_i$, where $\mathbf{e}_i$ is the canonical unit vector in $\mathbb{R}^{N_T}$ (with N_T the number of tiles in the reference grid). These definitions are set up under the working assumption of individuals carrying at most 2 devices in agreement with the deduplication step. Should we consider a more general situation, the generalization is obvious, although more computationally demanding.

Next, we define the random variable $\mathbf{T}_{dt} \in \{\mathbf{e}_i^{(1)}, \mathbf{e}_i^{(2)}\}_{i=1,\ldots,N_T}$ with probability mass function $\mathbb{P}(\mathbf{T}_{dt}|\mathbf{E})$ given by

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(1)} | \mathbf{E}_{1:D}\right) = \gamma_{dti} \times p_d^{(1)} \tag{8a}$$

$$\mathbb{P}\left(\mathbf{T}_{dt} = \mathbf{e}_i^{(2)} | \mathbf{E}_{1:D}\right) = \gamma_{dti} \times p_d^{(2)} \tag{8b}$$

where $p_d^{(i)}$ are the device-duplicity probabilities introduced in section 4.3. Notice that this is a categorical or multinoulli random variable. Finally, we define the multivariate random variable $\mathbf{N}_t^{\text{net}} = \left(N_{t1}^{\text{net}}, \ldots, N_{tN_T}^{\text{net}}\right)$ providing the number of individuals N_{ti}^{net} detected by the network at each tile $i = 1, \ldots, N_T$ at time instant t:

$$\mathbf{N}_t^{\text{net}} = \sum_{d=1}^{D} \mathbf{T}_{dt}. \tag{9}$$

The sum spans over the number of devices filtered as members of the target population according to section 4.4. If we are analysing, say, domestic tourism, D will amount to the number of devices in the network classified with a domestic tourism pattern according to the algorithms designed and applied in the preceding module. For illustrative examples, since we have not developed the statistical filtering module yet, we shall concentrate on present population.

The random variable $\mathbf{N}_t^{\text{net}}$ is, by construction, a Poisson multinomial random variable. The properties and software implementation of this distribution are not trivial (see e.g. Daskalakis et al., 2015) and we shall use Monte Carlo simulation methods by convolution to generate random variates according to this distribution.

The reasoning behind this proposal can be easily explained with a simplified illustrative example. Let us consider an extremely simple scenario with 5 devices and 5 individuals (thus, none of them carry two devices), and 9 tiles (a 3×3 reference grid). Let us consider that the location probabilities $\gamma_{dti} = \gamma_{ti}$ are the same for all devices d at each time instant and each tile. In these conditions $p_d^{(2)} = 0$ for all d. Let us focus on the univariate (marginal) problem of finding the distribution of the number of devices/individuals in a given tile i. If each device d has probability γ_{ti} of detection at tile i, then the number of devices/individuals at tile i will be given by a binomial variable Binomial$(5, \gamma_{ti})$. If the probabilities were not equal, then the number of devices/individuals would be given by a Poisson binomial random variable Poisson-Binomial$(5; \gamma_{1ti}, \gamma_{2ti}, \gamma_{3ti}, \gamma_{4ti}, \gamma_{5ti})$, which naturally generalizes the binomial distribution. If we focus on the whole multidimensional problem, then instead of having binomial and Poisson-binomial distributions, we must deal with multinomial and Poisson-multinomial variables. Finally, if $p_d^{(2)} \neq 0$ for all d, we must avoid double-counting, hence the factor $\frac{1}{2}$ in the definition of $\mathbf{e}_i^{(2)}$.

Notice that the only assumption made so far (apart from the trivial question of the maximum number of 2 devices carried by an individual) is the independence for two devices to be detected at any pair of tiles i and j. This independence assumption allows to claim that the number of detected individuals distributes as a Poisson-multinomial variable, understood as a sum of independent multinoulli variables with different parameters. There is no extra assumption in this derivation. The validation of this assumption is subtle, since ultimately it will depend on the correlation between the displacement patterns of individuals in the population. If the tile size is chosen small enough, we claim that the assumption holds fairly well and it is not a strong condition imposed on our derivations. On the other hand, if the tiles are too large (think of an extreme case about a reference grid being composed of whole provinces as tiles), we should expect correlations in the detection of individuals: those living in the same province will have full correlation and those living in different provinces will show near null correlation. Thus, the size of the tiles imposes some limitation to the validity of the independence assumption. Even the transport network in a territory will certainly influence these correlations. Currently, we cannot analyse quantitatively the relationship between the size of the tiles and the independence assumption with the network data simulator because we need both realistic simulated individual displacement patterns and simulated correlated trajectories (probably connected to the sharing of usual environments, home/work locations, etc.).

In figure 9 we show the high-density credible intervals with $\alpha = 0.95$ for the number of individuals N_{rt}^{net} detected by the network in each region r and each time instant t. We can compare with true values from the simulator. Deeper and wider analyses are ongoing to assess this procedure and its relationship with the geolocation and deduplication modules.

4.6 Inference

The last step comprises the estimation of the number of individuals N_{rt} in the target population in each region r at each time instant t. Notice that we aim at estimating also those individuals not detected by the network, namely, those subscribers of other MNOs and those not having a mobile device. To avoid identifiability problems, we need to make use of auxiliary information. This will basically be the register-based population figures N_r^{reg} and the penetration rates P_r^{net}. Notice, however, the different time scales of the register-based population estimates and of the network-

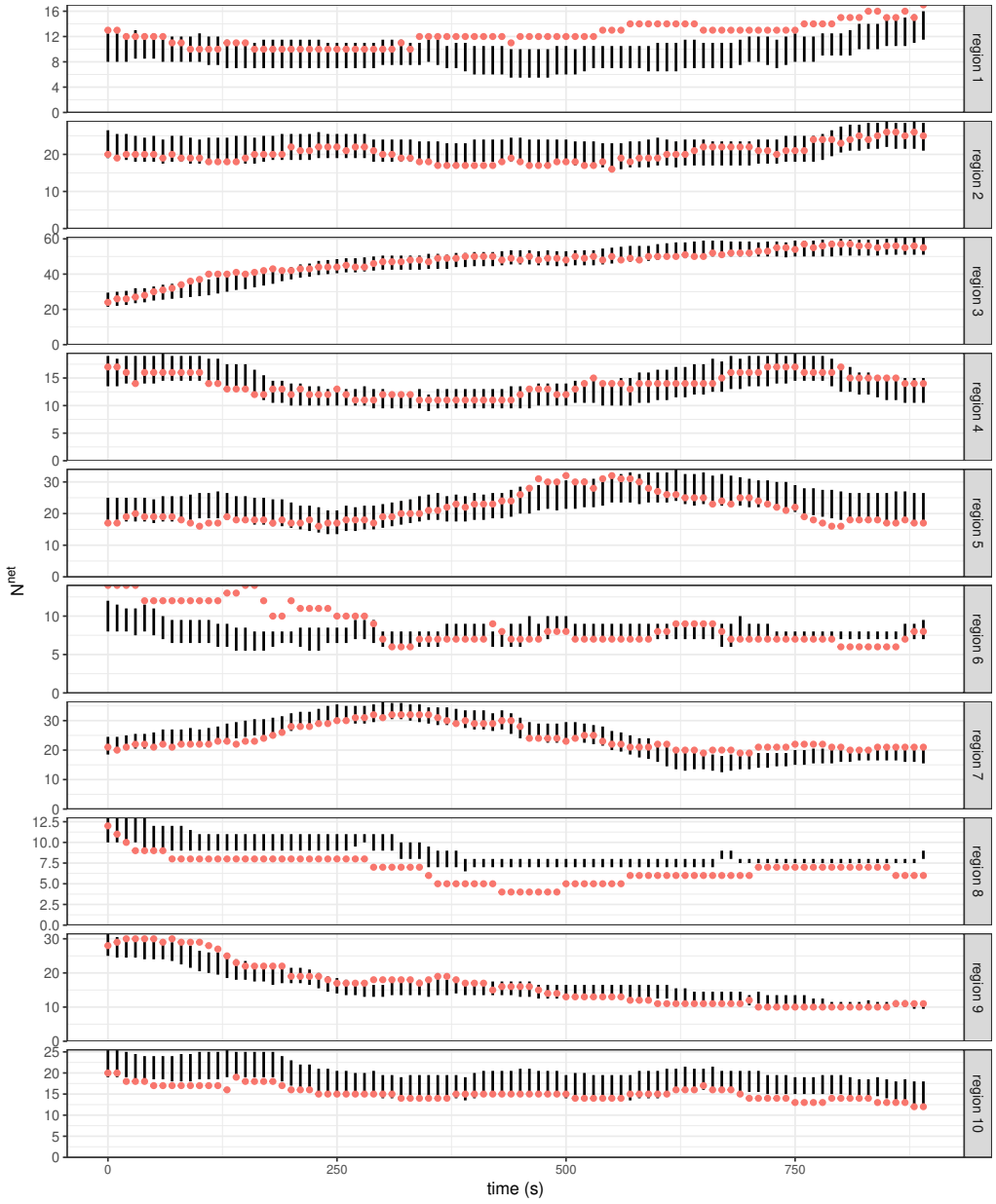

Figure 9: Credible intervals ($\alpha = 0.95$, HDI) for the number of individuals detected by the network. True values in red.

based population estimates. To avoid losing the higher time breakdown from telecommunication networks integrating at the same time the register-based population figures we shall consider a two-stage approach. Firstly, at the initial time t_0 we assume that both the register-based target population and the network-based population can be assimilated in terms of their physical location. Secondly, at later times $t > t_0$ we assume that individuals move over the geographical territory independently of the MNO, i.e. subscribers of MNO 1 will show a displacement pattern similar to those of MNO 2.

Under these general assumptions, following the approach from preceding steps, we propose to use hierarchical models (i) to produce probability distributions, (ii) to integrate all data sources, and (iii) to account for the uncertainty and the differences of concepts (present vs. residential population) and scales (time).

For the first stage the bottom line of our approach is inspired by the approach to estimate the species abundance in Ecology (Royle and Dorazio, 2009). If N_r^{net} and N_r denote the number of individuals detected by the network and in the target population, respectively, in a region r and if p_r denotes the probability of detection of an individual by the network in that region r, then we can model

$$N_r^{\text{net}} \simeq \text{Bin}\,(N_r, p_r), \tag{10}$$

where we have dropped out the time subscript for ease of notation. This model makes the only assumption that the probability of detection p_r for all individuals in region r is the same. This probability of detection amounts basically to the probability for an individual of being a subscriber of the given mobile telecommunication network. This assumption will be further modelled. As a first approximation, we may think of p_r as a probability related to the penetration rate P_r^{net} of the MNO in region r. We shall compute the posterior distribution $\mathbb{P}\left(N_r | N_r^{\text{net}}, \mathbf{I}^{\text{aux}}\right)$, where $\mathbf{I}^{\text{aux}}$ stands for any auxiliary information which we shall integrate into the estimation process.

As a first illustrative example of this reasoning, let us consider p_r as a fixed external parameter and try to compute the posterior probability distribution for N_r in terms of N_r^{net}, i.e.

$$\mathbb{P}\left(N_r | N_r^{\text{net}}, p_r\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}\left(N_r - N_r^{\text{net}}; 1 - p_r, N_r^{\text{net}} + 1\right) & \text{if } N_r \geq N_r^{\text{net}}, \end{cases}$$

where $\text{negbin}\,(k; p, r) \equiv \binom{k+r-1}{k} p^k (1-p)^r$ denotes the probability mass function of a negative binomial random variable k with parameters p and r. Once we have a distribution, we can provide point estimations, posterior variance, posterior coefficient of variation, credible intervals, and as many indicators as possible computed from the distribution. For example, if we use the MAP criterion we can provide as point estimator

$$\widehat{N}_r^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{N_r^{\text{net}}}{p_r} - N_r^{\text{net}} \right\rfloor, \tag{11}$$

With the distribution we can also compute accuracy indicators such as the posterior variance, the posterior coefficient of variation, or credible intervals (see e.g. Gelman et al., 2013).

Moreover, as model assessment we can compute the posterior predictive distribution $\mathbb{P}\left(N_r^{\text{net, rep}} | N_r^{\text{net}}\right)$ and produce some indicators such as[3]

$$ppRB = \frac{\mathbb{E}\left[N_r^{\text{net, rep}} - \widehat{N}_r^{\text{net}} \middle| \widehat{N}_r^{\text{net}}\right]}{\widehat{N}_r^{\text{net}}} \qquad ppRV = \frac{\mathbb{V}\left[N_r^{\text{net, rep}} - \widehat{N}_r^{\text{net}} \middle| \widehat{N}_r^{\text{net}}\right]}{(\widehat{N}_r^{\text{net}})^2} \tag{12}$$

If we take into account the uncertainty in N_r^{net} coming from preceding modules, we can promote these indicators to random variables using the probability distribution $\mathbb{P}\left(N_r^{\text{net}} \middle| \mathbf{E}, \mathbf{I}^{\text{aux}}\right)$ and study

[3] Let us call them posterior predictive relative bias and posterior predictive relative variance.

their mean values and dispersion.

The approach described above took the detection probability p_r as an external fixed parameter built from auxiliary data sources. Furthermore, we assumed that in region r all individuals show the same probability of being a subscriber of the mobile telecommunication network. Also, the number of detected individuals N_r^{net} accumulates the uncertainty from the preceding modules, since the geolocation of mobile devices and the determination of duplicities are probabilistic. To account for this, we propose to further model these quantities, hence the hierarchical approach.

Let us firstly consider how to introduce the uncertainty in N_r^{net}. From the preceding modules we have obtained the posterior probability $\mathbb{P}\left(N_r^{\text{net}}|\mathbf{E},\mathbf{I}^{\text{aux}}\right)$. We still consider the detection probability p_r as an external fixed parameter. Also, we still restrict ourselves to the univariate case. Under these assumptions, the unnormalized posterior probability distribution for the number of individuals in the target population and detected by the network will be

$$\mathbb{P}\left(N_r, N_r^{\text{net}}|\mathbf{E},\mathbf{I}^{\text{aux}}\right) \propto \text{negbin}\left(N_r - N_r^{\text{net}}; 1-p_r, N_r^{\text{net}}+1\right) \times \mathbb{P}\left(N_r^{\text{net}}\middle|\mathbf{E},\mathbf{I}^{\text{aux}}\right). \tag{13}$$

The normalization needs to be carried out numerically. Again, once we have the probability distribution for the random variable of interest, we can provide point estimations (MAP or posterior mean or posterior median) and accuracy indicators (posterior variance, posterior coefficient of variation, posterior IQR, credible intervals). These must be computed numerically.

The uncertainty in the detection probability p_r can be justified straightforwardly. A priori, we can think of a detection probability p_{kr} per individual k in the target population and try to device some model to estimate p_{kr} in terms of auxiliary information (e.g. sociodemographic variables, income, etc.). We would need subscription information related to these variables for the whole target population, which is unattainable. Instead, we may consider that the detection probability p_{kr} shows a common part for all individuals in region r plus some additional unknown terms, i.e. something like $p_{kr} = p_r + \text{noise}$. At a first stage, we propose to implement this idea by modeling $p_r \simeq \text{Beta}(\alpha_r, \beta_r)$ and choosing the hyperparameters α_r and β_r according to the penetration rates P_r^{net} and the official population data N_r^{reg}.

Let us denote by P_r^{net} the penetration rate at region r of the network, i.e. $P_r^{\text{net}} = \frac{N_r^{(\text{devices})}}{N_r}$. Notice that this penetration rate is also subjected to the problem of duplicities (individuals having two devices). To deduplicate, we make use of the duplicity probabilities $p_d^{(n)}$ computed in section 4.3 and of the posterior location probabilities γ_{d0r} in region r for each device d. Notice that $t = 0$ according to our first generic modelling assumption. We define

$$\Omega_r^{(1)} = \frac{\sum_{d=1}^{D} \gamma_{d0r} \cdot p_d^{(1)}}{\sum_{d=1}^{D} \gamma_{d0r}}, \tag{14a}$$

$$\Omega_r^{(2)} = \frac{\sum_{d=1}^{D} \gamma_{d0r} \cdot p_d^{(2)}}{\sum_{d=1}^{D} \gamma_{d0r}}. \tag{14b}$$

The deduplicated penetration rate is defined as

$$\tilde{P}_r^{\text{net}} = \left(\Omega_r^{(1)} + \frac{\Omega_r^{(2)}}{2} \right) \cdot P_r^{\text{net}}. \tag{14c}$$

To get a feeling on this definition, let us consider a very simple situation. Let us consider $N_r^{(1)} = 10$ individuals in region r with 1 device each one, $N_r^{(2)} = 3$ individuals in region r with 2 devices each one, and $N_r^{(0)} = 2$ individuals in region r with no device. Let us assume that we can measure the penetration rate with certainty, so that $P_r^{\text{net}} = \frac{16}{15}$. The devices are assumed to be neatly detected by the HMM (i.e. $\gamma_{d0r} = 1 - O(\epsilon)$) and duplicities are also inferred correctly ($p_d^{(2)} = O(\epsilon)$ for $d^{(1)}$ and $p_d^{(2)} = 1 - O(\epsilon)$ for $d^{(2)}$). Then $\Omega_r^{(1)} = \frac{10}{16} + O(\epsilon)$ and $\Omega_r^{(2)} = \frac{6}{16} + O(\epsilon)$. The deduplicated penetration rate will then be $\tilde{P}_r^{\text{net}} = \frac{13}{15} + O(\epsilon)$, which can be straightforwardly understood as a detection probability for an individual in this network in region r. In more realistic situations, the deduplication factors $\Omega_r^{(i)}$ incorporate the uncertainty in the duplicity determination.

Now, we fix

$$\left.\begin{array}{l} \alpha_r + \beta_r = N_r^{\text{reg}} \\ \frac{\alpha_r}{\alpha_r + \beta_r} = \tilde{P}_r^{\text{net}} \end{array}\right\} \quad \Rightarrow \quad \left\{\begin{array}{l} \alpha_r = \tilde{P}_r^{\text{net}} \cdot N_r^{\text{reg}} \\ \beta_r = \left(1 - \tilde{P}_r^{\text{net}}\right) \cdot N_r^{\text{reg}} \end{array}\right. \tag{15a}$$

There are several assumptions in this choice:

- On average, we assume that detection takes place with probability $\tilde{P}_r^{\text{net}}$. We find this assumption reasonable. Another alternative choice would be to use the mode of the beta distribution instead of the mean.
- Detection is undertaken over the register-based population. We assume some coherence between the official population count and the network population count. A cautious reader may object that we do not need a network-based estimate if we already have official data at the same time instant. We can make several comments in this regard:
  - A degree of coherence between official estimates by combining data sources to conduct more accurate estimates is desirable. By using register-based population counts in the hierarchy of models, we are indeed combining both data sources. In this combination notice, however, that the register-based population is taken as an external input in our model. There exist alternative procedures in which all data sources are combined at an equal footing (Bryant and Graham, 2013). We deliberately use the register-based population as an external source and do not intend to re-estimate by combination with mobile network data.
  - Register-based populations and network-based populations show clearly different time scales. The coherence we demand will be forced only at a given initial time t_0 after which the dynamical of the network will provide the time scale of the network-based population counts without further reference to the register-based population.
  - For the same model identifiability issues mentioned in the aggregation module, to estimate population counts N_r using network-based population counts N_r^{net} we need some external parameter(s). Otherwise, it is impossible. Detection probabilities are indeed these external parameters. We are modelling detection probabilities using penetration rates, which somehow already need register-based population figures. Our pragmatic approach is to identify external data sources already existing to be used in our model. These are penetration rates and register-based population counts easily collected by NSOs.

- The penetration rates P_r^{net} and the official population counts N_r^{reg} come without error. Should this not be attainable or realistic, we would need to introduce a new hierarchy level to account for this uncertainty.
- The deduplicated penetration rates are computed as a deterministic procedure (using a mean point estimation), i.e. the deduplicated penetration rates are also subjected to uncertainty, thus we should also introduce another hierarchy level to account for this uncertainty.

Then, we can readily compute the posterior distribution for N_r:

$$\mathbb{P}\left(N_r|N_r^{\text{net}},\mathbf{I}^{\text{aux}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negBetaBin}\left(N_r - N_r^{\text{net}}; N_r^{\text{net}}+1, \alpha_r - 1, \beta_r\right) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases}$$

It is a displaced negative beta binomial distribution ($\text{negBetaBin}(k;s,\alpha,\beta) \equiv \frac{\Gamma(k+s)}{k!\Gamma(s)}\frac{\text{B}(\alpha+s,\beta+k)}{\text{B}(\alpha,\beta)}$) with support in $N_r \geq N_r^{\text{net}}$ and parameters $s = N_r^{\text{net}}+1$, $\alpha = \alpha_r - 1$ and $\beta = \beta_r$. The mode is at

$$N^* = N_r^{\text{net}} + \left\lceil \left(\frac{\beta_r - 1}{\alpha_r}\right) \cdot N_r^{\text{net}} \right\rceil.$$

Using (15) we get as a MAP estimate:

$$\widehat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lceil \frac{N_r^{\text{reg}}}{\tilde{P}_r} - N_r^{\text{net}} - \frac{N_r^{\text{net}}}{N_r^{\text{reg}}}\frac{1}{\tilde{P}_r^{\text{net}}} \right\rceil, \tag{16}$$

which is very similar to (11) with the deduplicated penetration rate playing the role of a detection probability and a correction factor coming from the register-based population. The uncertainty is accounted for by computing the posterior variance, the posterior coefficient of variation, or credible intervals.

Notice that when $\alpha_r, \beta_r \gg 1$ (i.e., when $\min(\tilde{P}_r^{\text{net}}, 1-\tilde{P}_r^{\text{net}}) \cdot N_r^{\text{reg}} \gg 1$) the negative beta binomial distribution (16) reduces to the negative binomial distribution

$$\mathbb{P}\left(N_r|N_r^{\text{net}}\right) = \begin{cases} 0 & \text{if } N_r < N_r^{\text{net}}, \\ \text{negbin}\left(N_r - N_r^{\text{net}}; \frac{\beta_r}{\alpha_r+\beta_r-1}, N_r^{\text{net}}+1\right) & \text{if } N_r \geq N_r^{\text{net}}. \end{cases}$$

Notice that $\frac{\beta_r}{\alpha_r+\beta_r-1} \approx 1 - \tilde{P}_r^{\text{net}}$ so that we do not need the register-based population (this is similar to dropping out the finite population correction factor in sampling theory for large populations). The mode is at

$$\widehat{N}^{\text{MAP}} = N_r^{\text{net}} + \left\lfloor \frac{N_r^{\text{net}}}{\tilde{P}_r^{\text{net}}} - N_r^{\text{net}} \right\rfloor,$$

which is similar to (11).

We can make the model more complex by defining a new level in the hierarchy for the hyperparameters α and β (see e.g. Gelman et al., 2013) so that the relationship between these parameters and the external data sources (penetration rates and register-based population counts) is also uncertain.

For example, we can go all the way down the hierarchy, assume a cross-cutting relationship between parameters and some hyperparameters and postulate

$$N_r^{\text{net}} \simeq \text{Bin}(N_r, p_r), \quad \text{for all } r = 1, \ldots, R, \tag{17a}$$

$$p_r \simeq \text{Beta}(\alpha_r, \beta_r), \quad \text{for all } r = 1, \ldots, R, \tag{17b}$$

$$\left(\text{logit}\left(\frac{\alpha_r}{\alpha_r + \beta_r}\right), \alpha_r + \beta_r\right) \simeq \text{N}\left(\mu_{\beta r}(\beta_0, \beta_1; \tilde{P}_r^{\text{net}}), \tau_\beta^2\right) \times \text{Gamma}\left(1 + \xi, \frac{N_r^{\text{reg}}}{\xi}\right), \quad \text{for all } r = 1, \ldots, R, \tag{17c}$$

$$\left(\log\beta_0, \beta_1, \tau_\beta^2, \xi\right) \simeq \text{f}_\beta\left(\log\beta_0, \beta_1, \tau_\beta^2\right) \times \text{f}_\xi(\xi), \tag{17d}$$

where $\mu_{\beta r}(\beta_0, \beta_1; \tilde{P}_r^{\text{net}}) \equiv \log\left(\beta_0 \left[\frac{\tilde{P}_r^{\text{net}}}{1-\tilde{P}_r^{\text{net}}}\right]^{\beta_1}\right)$.

The interpretation of this hierarchy is simple. It is just a beta-binomial model in which the beta parameters α_r, β_r are correlated with the deduplicated penetration rates. This correlation is expressed through a linear regression model with common regression parameters across the regions, both the coefficients and the uncertainty degree. On average, the detection probabilities p_r will be the deduplicated penetration rates with uncertainty accounted for by hyperparameters $\beta_0, \beta_1, \tau_\beta^2$. For large population cells, the hyperparameter ξ drops out so that finally the register-based population counts N_r^{reg} play no role in the model, as above. This further hierarchy is under exploration (see Salgado et al. (2020) for some computational details). Indeed, the hierarchy can be extended also to model N_r (the so-called state process), e.g. by a Poisson distribution with parameter λ_r and keep on modelling λ_r according to some underlying process integrating more auxiliary information (see Salgado et al. (2020) for details).

For the second stage we shall focus only on closed populations, i.e. populations with individuals not allowed to enter or leave the geographical territory during the time period of estimation. This is a first step in agreement with the current status of the network event data simulator.

The basic assumption is that displacement patterns are not dependent on the subscribing MNO providing the data, i.e. individuals in a given MNO network show similar displacement patterns to those in any other network or in the target population in general. We begin by considering a balance equation. Let us denote by $N_{t,rs}$ the number of individuals moving from region s to region r in the time interval $(t-1, t)$. Then, we can write

$$\begin{aligned} N_{tr} &= N_{t-1r} + \sum_{\substack{r_t=1 \\ r_t \neq r}}^{N_T} N_{t,rr_t} - \sum_{\substack{r_t=1 \\ r_t \neq r}}^{N_r} N_{t,r_t r} \\ &= \sum_{r_t=1}^{N_T} \tau_{t,rr_t} \cdot N_{t-1r_t}, \end{aligned} \tag{18}$$

where we have defined $\tau_{t,rs} = \frac{N_{t,rs}}{N_{t-1s}}$ (0 if $N_{t-1s} = 0$). Notice that $\tau_{t,rs}$ can be understood as an aggregate transition probability from region s to region r at time interval $(t-1, t)$ in the target population. According to our general assumption we can use $\tau_{t,rs}^{\text{net}} \equiv \frac{N_{t,rs}^{\text{net}}}{N_{t-1s}^{\text{net}}}$ to model $\tau_{t,rs}$. In particular, we shall

postulate $\tau_{t,rs} = \tau^{\text{net}}_{t,rs}$. The probability distributions of N^{net}_{st-1} and $[\mathbf{N}^{\text{net}}_t]_{sr} = N^{\text{net}}_{t,rs}$ can indeed be already computed in the aggregation module.

Finally, we mention two points. On the one hand, random variables N_{rt} are defined recursively in the time index t, so that once we have computed the probability distribution at time t_0, then we can use equation (18) to compute the probability distribution at later times $t > t_0$. On the other hand, Monte Carlo techniques should be used to build these probability distributions. Once we have probability distributions, we can make point estimations and compute accuracy indicators as above (posterior variance, posterior coefficient of variation, credible intervals).

This same argument can be extended to produce origin-destination matrices. If N_{tr} and $\tau_{t,rs}$ denote, respectively, the number of individuals of the target population at time t in region r and the aggregate transition probability from region s to region r at the time interval $(t-1, t)$, then we can simply define $N_{t,rs} = N_{t-1s} \times \tau_{t,rs}$ and trivially build the origin-destination matrix for each time interval $(t-1, t)$. Under the same general assumption as before, if individuals are to move across the geographical territory independently of their mobile network operator (or even not being a subscriber or carrying two devices), we can postulate $\tau_{t,rs} = \tau^{\text{net}}_{t,rs}$, as before.

One of the advantages of the simulator is that we can analyse the sensitivity of the final outputs with respect to the accuracy of the auxiliary information. In particular, we can provide perturbed values for the register-based population figures in terms of their relative bias $\mathbb{E}\left(\frac{N^{\text{reg}}_r - N^{\text{reg},0}_r}{N^{\text{reg},\,0}_r}\right)$ and their coefficient of variation $\frac{\sqrt{\mathbb{V}\left(N^{\text{reg}}_r\right)}}{N^{\text{reg},0}_r}$.

In figure 10 we represent the high-density credible intervals ($\alpha = 0.95$) for the target population counts in each region r at the initial time instant t_0 using the negative beta binomial model (16) and the integration formula (1). Different values of the relative bias and the coefficient of variation for the register-based population are investigated.

In figure 11 we represent the high-density credible intervals ($\alpha = 0.95$) for the origin-destination matrices of the present population using the negative beta binomial model (16) at the initial time estimation, the integration formula (1) for all time instants, and the assumption $\tau_{t,rs} = \tau^{\text{net}}_{t,rs}$.

5 Conclusions and future prospects

Mobile network data is a complex data source with multiple potential uses in the production of official statistics. To achieve the daily incorporation of this data source into statistical offices many challenges must be overcome. By and large, they are both strategic and technical. To reach a successful solution, strategy and technique must have a two-way interrelation.

From the technical point of view the design and implementation of a modular end-to-end process according to the principles of the ESS Reference Methodological Framework stand as a key element. We have provided a first proposal of such a process where functional modularity and seamless evolvability arise as the main characteristics. Each module is designed to deal with different aspects of the whole complex estimation process from raw telecommunication data to final

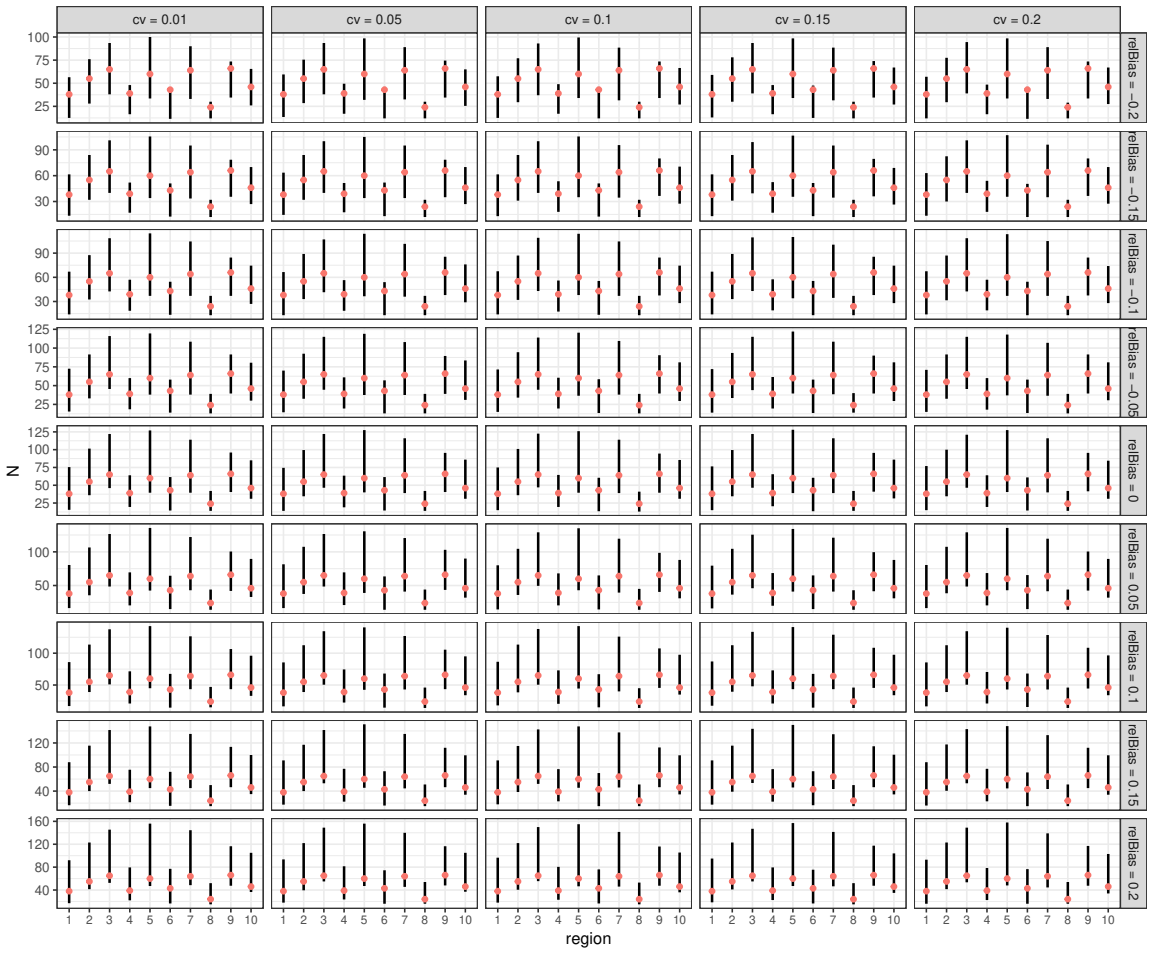

Figure 10: Credible intervals ($\alpha = 0.95$, HDI) for the number of individuals in the target population at the initial time t_0 using the negative beta binomial model and different values for the relative bias and coefficient of variation of the register-based population. True values of target population counts in red.

target population estimates together with accuracy indicators.

Since the access to mobile network data is notably limited for different reasons, we make use of a network event data simulator to provide a proof of concept. Each module needs further development. The use of HMM for geolocation of mobile devices should be further extended with more complex definitions of state and more sophisticated proposals such as continuous-time hidden Markov chains and particle-filter approaches are to be explored. Deduplication procedures and aggregation methods should be accordingly adapted. Statistical filtering algorithms for target inidividuals and anchor point identification in terms of movement analysis indicators need to be proposed and tested using both simulated and real data. Finally, inference models should be extended for open populations and be essentially multivariate including spatial autocorrelations.

Advances in the design of such a process for multiple statistical domains must be taken into account in the management of access and use of this data and agreements with MNOs.

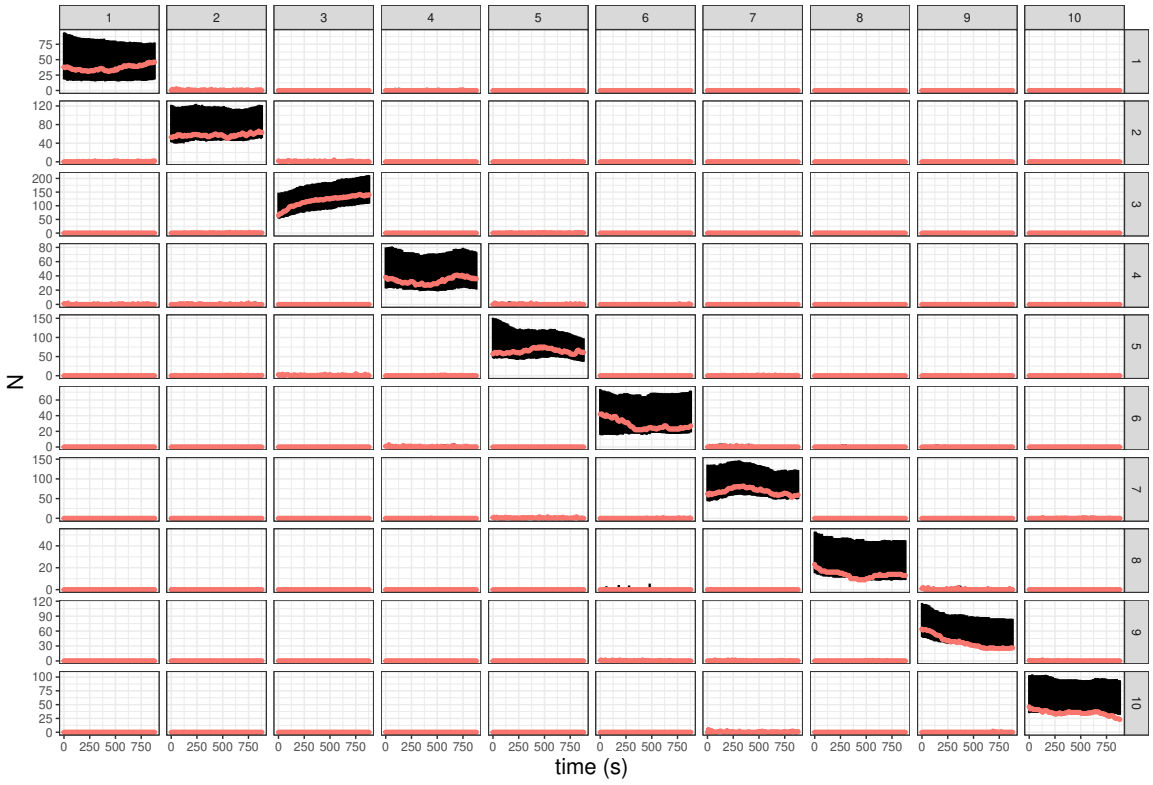

Figure 11: Credible intervals ($\alpha = 0.95$, HDI) for the O-D matrices in the target population using the negative beta binomial model and values for the relative bias 20% and coefficient of variation 20% of the register-based population. True values of target population counts in red.

A Mathematical details

A.1 Geolocation

We include some generic mathematical details to compute the posterior location probabilities from the input data. This is conducted in steps.

A.1.1 Time discretization and padding

We shall work in discrete times. To do this we need to relate three parameters, namely (i) the tile dimension l (we assume a square grid for simplicity), (ii) the time increment Δt between two consecutive time instants, and (iii) an upper bound $v_{\max}$ for the velocity of the individuals in the population. In our transition model we impose that in the time interval Δt, the device d at most can displace from one tile to an adjacent tile. Under this condition, we can trivially set $\Delta t \lesssim \frac{l}{v_{\max}}$. If in the dataset the device d is detected at longer time periods, then we artificially introduce missing values at intervals Δt between every two observed values. This artificial non-response allows us to work with parsimonious models easier to estimate instead of using more complex transition matrices.

Additionally, each observed time instance t is approximated to its closest multiple integer of Δt so that we will have as input data a sequence of time instants at multiples $t_n = \Delta t \cdot n, (n \geq 0)$ and a randomly alternate sequence of missing values and of observed event variables $\mathbf{E}_{dt_n}$.

A.1.2 Construction of the emission model

The emission model is directly built by computing the so-called emission probabilities, i.e. the event location probabilities $\mathbb{P}\left(\mathbf{E}_{t_n} = \mathbf{e}_j \middle| T_{dt_n} = i\right)$, where $\mathbf{e}_j$ is a possible value for the observed event variables $\mathbf{E}_{dt_n}$ and i denotes the tile index. We assume time homogeneity. This conditional probability is computed using the radio wave propagation model of our choice (see e.g. Salgado et al., 2020, for details).

A.1.3 Construction of the transition model

Now we specify a model for the transition between tiles (states) $\{T = i\}_{i=1,\ldots,N_T}$. For ease of explanation and notation, let us change the notation of each tile T_i to a two-dimensional index $T_{(i,j)}$. Accordingly, each tile will be specified in this section by a pair of integer coordinates. The correspondence between both enumerations is arbitrary, but fixed once it has been chosen. We again assume time homogeneity for simplicity. Thus, $\mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right)$ will denote $\mathbb{P}\left(T_{(r,s)}(t_n + \Delta t) \middle| T_{(i,j)}(t_n)\right)$ for any $t_n = 0, 1, \ldots$ We assume a square regular grid for simplicity.

The essential assumption of the model is that an individual can at most reach an adjacent tile in time Δt. Thus,

$$\mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right) = 0 \qquad \max\{|r-i|, |s-j|\} \geq 2, \qquad r, s, i, j = 1, \ldots, \sqrt{N_T}. \tag{19a}$$

Now, we assume that we have no further auxiliary information to model these transitions and impose rectangular isotropic conditions:

$$\mathbb{P}\left(T_{(i\pm1,j)} \middle| T_{(i,j)}\right) = \mathbb{P}\left(T_{(i,j\pm1)} \middle| T_{(i,j)}\right) = \theta_1 \qquad i, j = 1, \ldots, \sqrt{N_T}, \tag{19b}$$

$$\mathbb{P}\left(T_{(i\pm1,j\pm1)} \middle| T_{(i,j)}\right) = \theta_2 \qquad i, j = 1, \ldots, \sqrt{N_T}. \tag{19c}$$

The last set of conditions is row-stochasticity:

$$\begin{aligned} \sum_{r,s=1}^{N_T} \mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right) &= 1, \qquad i, j = 1, \ldots, \sqrt{N_T}, \\ \mathbb{P}\left(T_{(r,s)} \middle| T_{(i,j)}\right) &\geq 0, \qquad i, j, r, s = 1, \ldots, \sqrt{N_T}. \end{aligned} \tag{19d}$$

Now back to the original notation for tiles and using the usual notation for the transition matrix $A = [a_{ij}]$, with $a_{ij} = \mathbb{P}\left(T_{jt} \middle| T_{it}\right)$ (Rabiner, 1989), conditions (19) amounts to having a highly sparse transition matrix A with up to 4 terms equal to θ_1 and θ_2 (each) per row and diagonal entries guaranteeing row-stochasticity.

In our proposed implementation, in order to seek future generalization, we will work with a generic block-tridiagonal matrix where the restrictions (19a) leading to 0 have been included, and

complemented with the rest of restrictions (19b)-(19d) in matrix form. Thus, we write $C \cdot \text{vec}(\tilde{A}) = \mathbf{b}$, where $\text{vec}(\tilde{A})$ stands for the non-null elements of A in vector form. The rows of $[C\ \mathbf{b}]$ encode each of the restrictions (19b), (19c), and (19d). For example, $a_{12} = \theta_1$ and $a_{21} = \theta_1$ produce a row like this

$$C_i \cdot \text{vec}(\tilde{A}) = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & 0 & 1 & 0 & \cdots & 0 & -1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \cdot (\cdots \cdot a_{12} \cdot \cdots \cdot a_{21} \cdot \cdots)^T = b_i = 0.$$

A.1.4 Construction of the initial state (prior) distribution

The HMM prior can be constructed according to any available information. For illustrative purposes, we consider two choices: (i) uniform prior, i.e. $\pi_i^{\text{uniform}} = \frac{1}{N_T}$ and (ii) $\pi_i^{\text{network}} \propto \sum_k (\text{RSS}(d(\mathbf{E}_k, T_i)))$ (where RSS is expressed in watts) or $\pi_i^{\text{network}} \propto \sum_k (\text{SDM}(d(\mathbf{E}_k, T_i)))$, depending on the emission model. Any other choice or combination thereof is also possible (see e.g. Tennekes et al., 2020).

A.1.5 Computation of the likelihood

The likelihood is trivially computed using the numerical proviso of setting emission probabilities equal to 1 when there is a missing value in the observed variables (e.g. due to time padding). The general expression for the likelihood is

$$\begin{aligned} L(\mathbf{E}_d) &= \sum_{i_0=1}^{N_T} \cdots \sum_{i_T=1}^{N_T} \mathbb{P}\left(T_{dt_0} = i_0\right) \prod_{n=1}^{N} \mathbb{P}\left(T_{dt_n} = i_n | T_{dt_{n-1}} = i_{n-1}\right) \mathbb{P}\left(E_{dt_n} \middle| T_{dt_n} = i_n\right) \\ &= \sum_{i_0=1}^{N_T} \cdots \sum_{i_T=1}^{N_T} \mathbb{P}\left(T_{dt_0} = i_0\right) \prod_{n=1}^{N} a_{di_{n-1}i_n}(\boldsymbol{\theta}) \cdot \mathbb{P}\left(E_{dt_n} \middle| T_{dt_n} = i_n\right) \end{aligned} \tag{20}$$

Notice that the emission probabilities only contribute numerically providing no parameter whatsoever to be estimated.

A.1.6 Parameter estimation

The estimation of the unknown parameters $\boldsymbol{\theta}$ is conducted maximizing the likelihood. The restrictions coming from the transition model (19) makes the optimization problem not trivial. Notice that the EM algorithm is not useful. Instead, we provide a taylor-made solution seeking for future generalizations with more realistic choices of transition probabilities incorporating land use information. Formally, the optimization problem is given by:

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \\ & a_k \in [0, 1], \end{aligned} \tag{21}$$

where $\mathbf{a}$ stands for the nonnull entries of the transition probability matrix A, the objective function $h(\mathbf{a})$ is derived from the likelihood L expressed in terms of the nonnull entries of the transition matrix A, and the system $C \cdot \mathbf{a} = \mathbf{b}$ expresses the sets of restrictions from the transition model (19) not involving null rhs terms (restrictions (19b), (19c), and (19d)).

The total number of zeroes in the transition matrix A can be proven to be given by $4 \times (N_T - 4) + 4 \times (\sqrt{N_T} - 2) \times (N_T - 6) + (\sqrt{N_T} - 2)^2 \times (N_T - 9) = N_T^2 - 9 \cdot N_T + 12\sqrt{N_T} - 4$ (Salgado et al., 2020). The number of non-null components of $\mathbf{a}$ in problem (21) is $d = 9 \cdot N_T - 12\sqrt{N_T} + 4$.

The number of restrictions n_r not involving zeroes depends very sensitively on the particular transition model chosen for the displacements. In the rectangular isotropic model considered above, it can also be proven to be $n_r = 4 \cdot N_T - 4\sqrt{N_T} - 1 + 4 \times (\sqrt{N_T} - 1)^2 - 1 + N_T = 9 \cdot N_T - 12\sqrt{N_T} + 2$ (Salgado et al., 2020). Thus, the matrix C will have dimensions $n_r \times d$. Notice that $d - n_r = 2$, as expected, since we have two free parameters θ_1 and θ_2.

The abstract optimization problem is thus

$$\begin{aligned} \max \quad & h(\mathbf{a}) \\ \text{s.t.} \quad & C \cdot \mathbf{a} = \mathbf{b} \quad , \\ & \mathbf{a} \in [0,1]^d, \end{aligned} \tag{22}$$

where $C \in \mathbb{R}^{n_r \times d}$ and $\mathbf{b} \in \mathbb{R}^d$. The objective function $h(\mathbf{a})$ is indeed a polynomial in the non-null entries $\mathbf{a}$. This problem can be further simplified using the matrix QR decomposition. Write $C = Q \cdot R$, where Q is an orthogonal matrix of dimensions $n_r \times n_r$ and R is an upper triangular matrix of dimensions $n_r \times d$. Then we can rewrite the linear system as $R \cdot \mathbf{a} = Q^T \cdot \mathbf{b}$ and we can linearly solve variables $a_1, \ldots, a_{n_r}$ in terms of variables $a_{n_r+1}, \ldots, a_d$:

$$\begin{pmatrix} a_1 & \cdots & a_{n_r} \end{pmatrix}^T = \tilde{C}_{n_r \times (d-n_r)} \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T.$$

The system (22) then reduces to

$$\begin{aligned} \max \quad & \tilde{h}(a_{n_r+1}, \ldots, a_d) \\ \text{s.t.} \quad & 0 \leq \tilde{C} \cdot \begin{pmatrix} a_{n_r+1} & \cdots & a_d \end{pmatrix}^T \leq 1. \end{aligned} \tag{23}$$

In our current software implementation we resort to general-purpose optimizers. It remains for future work to find an optimised algorithm to solve (23). The solution $\mathbf{a}^*$ to problem (22) will be introduced in the transition probability matrix, which will thus be denoted by $\widehat{A}$.

A.1.7 Application of the forward-backward algorithm

Once the HMM has been fitted, we can readily apply the well-known forward-backward algorithm (see e.g. Bishop, 2006) to compute the target location probabilities γ_{dti} and γ_{tij}. No novel methodological content is introduced at this point. For our implementation, we have used the scaled version of the algorithm (see (Bishop, 2006)).

A.1.8 Model evaluation

We propose a bias-variance decomposition of the mean squared error of the estimated location as main figure of merit. We define the center of location probabilities and the root mean squared dispersion. Let us denote by $\mathbf{R}_{dt} \in \{\mathbf{r}_i^{(c)}\}_{i=1,\ldots,N_T}$ the random vector for the position of a device according to the distribution of posterior location probabilities γ_{dti}, where $\mathbf{r}_i^{(c)}$ stands for the coordinates of the center of tile c. Let us shortly denote $\bar{\mathbf{R}}_{dt} \equiv \mathbb{E}\mathbf{R}_{dt} = \sum_{i=1}^{N_T} \gamma_{dti} \mathbf{r}_i^{(c)}$. Let us also denote the true position of device d at time t by $\mathbf{r}_{dt}^*$. Then, we can decompose

$$\begin{aligned} \text{msd}_{dt} \equiv \mathbb{E}\|\mathbf{R}_{dt} - \mathbf{r}^*_{dt}\|^2 &= \mathbb{E}\|(\mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}) + (\bar{\mathbf{R}}_{dt} - \mathbf{r}^*_{dt})\|^2 \\ &= \mathbb{E}\left[\langle \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}, \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}\rangle\right] + \\ &\quad 2 \cdot \mathbb{E}\left[\langle \mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}, \bar{\mathbf{R}}_{dt} - \mathbf{r}^*_{dt}\rangle\right] + \\ &\quad \mathbb{E}\left[\langle \bar{\mathbf{R}}_{dt} - \mathbf{r}^*_{dt}, \bar{\mathbf{R}}_{dt} - \mathbf{r}^*_{dt}\rangle\right] \\ &= \text{rmsd}^2_{dt} + \text{b}^2_{dt}. \end{aligned} \tag{24}$$

This decomposition motivates the definition of bias $\text{b}_{dt} = \|\bar{\mathbf{R}}_{dt} - \mathbf{r}^*_{dt}\|$ and root mean squared deviation

$$\text{rmsd}_{dt} = \sqrt{\mathbb{E}\left[\|\mathbf{R}_{dt} - \bar{\mathbf{R}}_{dt}\|^2\right]}$$

.

Acknowledgments

The authors acknowledge M.Á. Martínez-Vidal, S. Lorenzo, M. Suárez-Castillo, R. Radini, T. Tuoto, M. Offermans, M. Tennekes, S. Hadam, and F. Ricciato for invaluable insights and debates. This work was partially supported by the European Commission through the European Statistical System [Grant Agreement Number 847375-2018-NL-BIGDATA (ESSnet on Big Data II)].

References

Ahas, Rein, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru (2010, April). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology 17*(1), 3–27.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Cambridge: Springer-Verlag New York Inc.

Blondel, Vincent D., Adeline Decuyper, and Gautier Krings (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science 4*, 10.

Bryant, John R. and Patrick J. Graham (2013, September). Bayesian demographic accounts: Subnational population estimation using multiple data sources. *Bayesian Analysis 8*(3), 591–622.

Calabrese, Francesco, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti (2013, January). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies 26*, 301–313.

Daskalakis, C., G. Kamath, and C. Tzamos (2015). On the structure, covering, and learning of poisson multinomial distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 1203–1217.

Dattilo, B., R. Radini, and M. Sabato (2016, November). How many SIM in your luggage? A strategy to make mobile phone data usable in tourism statistics. In *14th Global Forum on Tourism Statistics*.

Debusschere, Marc, Jan Sonck, and Michail Skaliotis (2016, November). Official Statistics and mobile network operator partner up in Belgium. In *OECD Statistics Newsletter*, Number 65, pp. 11–14.

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem (2014, October). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*(45), 15888–15893.

DGINS (2018). Bucharest memorandum.

Douglass, Rex W, David A Meyer, Megha Ram, David Rideout, and Dongjin Song (2015). High resolution population estimates from telecommunications data. *EPJ Data Science 4*, 4.

European Parliament (2016). *EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1*.

Galiana, Lino, Benjamin Sakarovitch, and Zbigniew Smoreda (2018, October). Understanding sociospatial segregation in french cities with mobile phone data. DGINS18.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, and Aki Vehtari (2013). *Bayesian Data Analysis*. Taylor & Francis Ltd.

González, Marta C., César A. Hidalgo, and Albert-László Barabási (2008, June). Understanding individual human mobility patterns. *Nature 453*(7196), 779–782.

Graells-Garrido, Eduardo, Diego Caro, and Denis Parra (2018, December). Inferring modes of transportation using mobile phone data. *EPJ Data Science 7*, 49.

Iqbal, Md. Shahadat, Charisma F. Choudhury, Pu Wang, and Marta C. González (2014, March). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies 40*, 63–74.

Izquierdo-Valverde, M., J. Prado Mascuñano, and M. Velasco-Gimeno (2016, November). Same-day visitors crossing borders a big and data approach using traffic control. In *14th Global Forum on Tourism Statistics*, Venice, Italy.

Kowarik, A. and M. van der Loo (2018). Using R in the Statistical Office: the experiences of Statistics Netherlands and Statistics Austria. *Romanian Statistical Review 2018*(1), 15–29.

Lehmann, Erich L. and George Casella (2003). *Theory of Point Estimation*. New York: Springer New York.

Lestari, Titi Kanti, Siim Esko, Sarpono, Erki Saluveer, and Rifa Rufiadi (2018, November). Indonesia's experience of using signaling mobile positioning data for official tourism statistics. In *15th World Forum on Tourism Statistics*, Cusco, Peru.

Long, Jed A. and Trisalyn A. Nelson (2013, February). A review of quantitative methods for movement data. *International Journal of Geographical Information Science 27*(2), 292–318.

Louail, Thomas, Maxime Lenormand, Oliva G. Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy (2014). From mobile phone data to the spatial structure of cities. *Scientific Reports 4*, 5276.

Meersman, Freddy De, Gerdy Seynaeve, Marc Debusschere, Patrick Lusyne, Pieter Dewitte, Youri Baeyens, Albrecht Wirthmann, Christophe Demunter, Fernando Reis, and Hannes I. Reuter (2016, June). Assessing the quality and of mobile and phone data as a source of statistics. In *European Conference on Quality in Official Statistics (Q2016)*, Madrid.

Miao, G., J. Zander, W. Sung, and S.B. Slimane (2016). *Fundamentals of Mobile Data Networks*. Cambridge: Cambridge University Press.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. Boston, MA: MIT Press.

National Estonian Bank (2020). Methodology for the compilation of international travel statistics.

Nurmi, Ossi (2016). Improving the accuracy of outbound tourism statistics with mobile positioning data. In *15th Global Forum on Tourism Statistics*, Number from, Cusco, Peru.

Oancea, B. and R. Dragoescu (2014). Integrating R and Hadoop for Big Data analysis. *Romanian Statistical Review 2014*(2), 83–94.

Oancea, Bogdan, Marian Necula, Luis Sanguiao, David Salgado, and Sandra Barragán (2019, December). A simulator for network event data. Technical report, Statistics Romania (INS) and Statistics Spain (INE). Deliverable I.2 of Work Package I of ESSnet on Big Data II.

Pappalardo, Luca, Maarten Vanhoof, Lorenzo Gabrielli, Zbigniew Smoreda, Dino Pedreschi, and Fosca Giannotti (2016, June). An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics 2*(1-2), 75–92.

Phithakkitnukoon, Santi, Zbigniew Smoreda, and Patrick Olivier (2012, June). Socio-geography of human mobility: A study using longitudinal mobile phone data. *PLoS ONE 7*(6), e39253.

Positium (2016). Technical documentation for required raw data from mobile network operator for official statistics. ESSnet WP5 internal technical report.

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

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Radini, Roberta, Tiziana Tuoto, Raffaella M. Acrari, and D. Salgado (2020). WPI DeliverableI.6. Quality - A proposal for a statistical production process with mobile network data.

Raun, Janika, Rein Ahas, and Margus Tiru (2016, December). Measuring tourism destinations using mobile tracking data. *Tourism Management 57*, 202–212.

Reis, Fernando, Gerdy Seynaeve, Albrecht Wirthmann, Freddy de Meersman, and Marc Debusschere (2017, March). Land use classification based on present population daily profiles from a big data source.

Ricciato, Fabio (2018). Towards a Reference Methodological Framework for processing MNO data for Official Statistics. In *15th World Forum on Tourism Statistics*, Number opera-.

Ricciato, Fabio, Peter Widhalm, Francesco Pantisano, and Massimo Craglia (2017, February). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing 35*, 65–82.

Royle, A.J. and R.M. Dorazio (2009). *Hierarchical modelling and inference in Ecology*. New York: Elsevier.

Sakarovitch, Benjamin, Marie-Pierre de Bellefon, Pauline Givord, and Maarten Vanhoof (2019). Estimating the residential population from mobile phone data, an initial exploration. *Economie et Statistique / Economics and Statistics 505-506*, 109–132.

Salgado, David, M. Elisa Esteban, María Novás, Soledad Saldaña, and Luis Sanguiao (2018, December). Data organisation and process design based on functional modularity for a standard production process. *Journal of Official Statistics 34*(4), 811–833.

Salgado, David, Luis Sanguiao, Sandra Barragán, Bogdan Oancea, and Milena Suarez-Castillo (2020). WPI DeliverableI.3. Methodology - A proposed production framework with mobile network data.

Salgado, D., L. Sanguiao, B. Oancea, S. Barragán, and M. Necula (2020). An end-to-end statistical process with mobile network data for official statistics. Submitted to EPJ Data Science.

Senaeve, Gerdy and Christophe Demunter (2016, November). When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics. In *14th Global Forum on Tourism Statistics*, Venice, Italy.

Shabbir, Noman, Muhammad T Sadiq, Hasnain Kashif, and Rizwan Ullah (2011, September). Comparison of radio propagation models for long term evolution (LTE) network. *International Journal of Next-Generation Networks 3*(3), 27–41.

Templ, M. and V. Todorov (2016, Feb). The Software Environment R for Official Statistics and Survey Methodology. *Austrian Journal of Official Statistics 45*(1), 97–124.

Tennekes, Martijn, Yvonne A.P.M. Gootzen, and Shan H. Shah (2020, May). A Bayesian approach to location estimation of mobile devices from mobile network operator data. resreport, Statistics Netherlands (CBS).

Ucar, I., M. Gramaglia, M. Fiore, Z. Smoreda, and E. Moro (2019). Netflix or youtube? regional income patterns of mobile service consumption. In *NetMob 2019*, Oxford, UK.

UN (2017). Handbook on the use of mobile phone data for official and statistics.

UNECE (2011, June). Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics. UNECE (Ed.), 59th Plennay Session of Conference of European Statisticians, Item 4. High-Level Group for the Modernisation of Official Statistics.

Van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vanhoof, Maarten, Fernando Reis, Thomas Ploetz, and Zbigniew Smoreda (2018, December). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics 34*(4), 935–960.

Venkataraman, Shivaram, Zongheng Yang, Davies Liu, Eric Liang, Hossein Falaki, Xiangrui Meng, Reynold Xin, Ali Ghodsi, Michael Franklin, Ion Stoica, and Matei Zaharia (2016). SparkR: Scaling R Programs with Spark. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD âĂŹ16, New York, NY, USA, pp. 1099âĂŞ1104. Association for Computing Machinery.

Wang, Zhenzhen, Sylvia Y. He, and Yee Leung (2018, April). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society 11*, 141–155.

White, Tom (2009). *Hadoop: The Definitive Guide* (1st ed.). OâĂŹReilly Media, Inc.

Williams, Susan (2016). Statistical uses for mobile phone data: literature review. Technical report, Office for National Statistics.

Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica (2016, oct). Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM 59*(11), 56âĂŞ65.

Zhao, C., S. Zhao, M. Zhao, Z. Chen, C.-Z. Gao, H. Li, and Y. Tang (2019). Secure multi-party computation: Theory, practice and applications. *Information Sciences 476*, 357–372.

Spanish Journal of Statistics
Vol. 2 No. 1 2020, Pages 79–107
doi:https://doi.org/10.37830/SJS.2020.1.05

Regular Article

Commonly used methods for measuring output quality of multisource statistics

Ton de Waal[1], Arnout van Delden[2], Sander Scholtus[3]
[1]Statistics Netherlands and Tilburg University, t.dewaal@cbs.nl
[2]Statistics Netherlands, a.vandelden@cbs.nl
[3]Statistics Netherlands, s.scholtus@cbs.nl

Received: November 4, 2020. Accepted: March 3, 2021.

Abstract: Estimation of output quality based on sample surveys is well established. It accounts for the effects of sampling and non-response errors on the accuracy of an estimator. When administrative data are used or combinations of administrative data with survey data, more error types need to be taken into account. Moreover, estimators in multisource statistics can be based on different ways of combining data sources. That partly affects the methodology that is needed to estimate output quality. This paper presents results of the ESSnet project Quality of Multisource Statistics that studied methods to estimate output quality. We distinguish three main groups of methods: scoring methods, (re)sampling methods and methods based on parametric modeling. Each of those is split into methods that can be used for both single and multisource statistics and methods that can be applied to multisource statistics only. We end the paper by discussing some of the main challenges for the near future. We argue that estimating output quality for multisource statistics is still more an art than a technique.

Keywords: bias, bootstrap, coherence, data integration, parametric modeling, quality framework, sampling theory, variance

MSC: 62D05, 62D10, 62F40, 62H12, 62P20, 62P25

1 Introduction

The fundamental reason for existence of National Statistical Institutes (NSIs) is that they are responsible for the official figures for policy making. It is therefore crucial that NSIs produce reliable estimates of societal phenomena. That implies that NSIs should be able to monitor the quality of the output that they produce. Estimation of output quality for single source statistics as a function of sampling error is well established. When administrative data sources are used, or a combination of administrative and survey data, more error types, such as measurement and linkage errors, need to be estimated and taken into account. How the effects of those error types can be estimated, the

methodology, will partly depend on how the data are combined. One example is that a target variable with measurement error is available at micro-level in multiple sources. Another example is that estimates of primary statistics, again with measurement error, are reconciled into an integrated set of values that fulfill balancing equations. One needs different methods to measure the output quality in these two situations.

Estimation of output quality of multisource statistics has been studied in the ESSnet project Quality of Multisource Statistics (also referred to as Komuso). The Komuso project lasted from January 2016 until October 2019. It was part of the ESS.VIP Admin Project. The main objectives of that latter project were: (i) to improve the use of administrative data sources, and (ii) to support the quality assurance of output produced using administrative sources. Partners in Komuso were Statistics Denmark (overall project leader of the ESSnet), Statistics Norway, ISTAT (the Italian national statistical institute), Statistics Lithuania, Statistics Austria, the Hungarian Central Statistical Office, the Central Statistical Office of Ireland, and Statistics Netherlands.

The main aim of Komuso was to produce quality guidelines for NSIs that are specific enough to be applied in statistical production by those NSIs. These guidelines take the entire production chain into account (input, process, and output) and cover a variety of situations in which NSIs work: various error types and different basic data configurations (BDCs, see Subsection 2.2). The guidelines list a variety of potential indicators/measures, indicate for each of them their applicability and in what situation they are preferred or not, and provide an ample set of examples of specific cases and decision-making processes.

Work Package 3 (WP 3) of Komuso focused on measuring the quality of statistical output based on multiple data sources. Measuring the quality of statistical output differs fundamentally from measuring the quality of input data since one ideally wants to take into account all processing and estimation steps that were taken to achieve the output. The problem encountered in WP 3 was not so much how to define the quality measures, but rather how these quality measures should be computed for a given set of input datasets and a certain procedure for combining these input datasets. At the moment, there is no all-encompassing theory or framework that can be used as a basis for quality measures for multisource statistics and for methods to calculate such measures. Constructing quality measures for multisource statistics and calculating them is still more of an art than a technical recipe that one can simply follow.

The present paper concentrates on methods to compute output quality measures. Those quality measures and their computational methods were examined and described in WP 3 of Komuso. They form an appendix to the above-mentioned quality guidelines which were developed in WP 1 of Komuso. Section 2 describes the approach taken in WP 3 of Komuso. Section 3 focuses on scoring methods for measuring output quality, Section 4 on methods based on (re)sampling, and Section 5 on methods based on (parametric) modelling. Each of these sections is split into two parts: methods that can be used to measure output quality for single and multisource statistics, and methods that have been developed for multisource statistics only. Section 6 concludes this paper with a brief discussion.

Due to the large variety in situations and methods we consider in this paper, the notation varies slightly over the various (sub)sections. We hope that this will not confuse the reader.

2 Approach taken in Komuso

In WP 3 of Komuso the work was subdivided into three consecutive steps:

1. In the first step a literature review or suitability test was carried out. In a literature review existing quality measures and recipes to compute them were studied and described. In a suitability

test also data were used to test quality measures and the recipes to compute them. Suitability tests were mainly used for newly proposed quality measures, but also for some already known quality measures to learn more about their properties, or for already known quality measures that were applied to a new field. In such a suitability test, practical and theoretical aspects of a quality measure and the accompanying calculation recipe were examined.
2. In order to make the results of Step 1 easily accessible, in Step 2 so-called quality measures and computation methods (QMCMs) were produced. Such a QMCM is a standardized, short description of a quality measure and the accompanying calculation recipe as well as a description of the situation(s) in which the quality measure and accompanying recipe can be applied. In total, 32 QMCMs were developed in the Komuso project.
3. In Step 3 hands-on examples were developed in the Komuso project for 31 of the 32 QMCMs. The one exception for which no example was provided concerned a general description of error types.

In order to cover different situations of different NSIs, and for ease of finding the results, the quality measures were structured along five classifications:

- quality dimensions;
- BDCs;
- error types;
- general approaches;
- computational methods.

The first four classifications are discussed in Subsections 2.1 to 2.4. The fifth classification is discussed extensively in Sections 3 to 5.

2.1 Quality dimensions

WP 3 of Komuso focused on four quality dimensions: accuracy, timeliness, coherence and relevance. The selected quality dimensions can more or less be quantified. *Accuracy* is "the degree of closeness of computations or estimates to the exact or true values that the statistics were intended to measure" (Eurostat, 2014). *Timeliness* was operationalized as "the time lag between the date of the publication of the results and the last day of the reference period of the estimate of the event or phenomenon they describe" [Komuso (ESSnet Quality of Multisource Statistics) (2019), in line with Eurostat (2014)]. *Coherence* "measures the adequacy of statistics to be combined in different ways and for various uses" (Eurostat, 2014). *Relevance* is defined as "the degree to which statistical outputs meet current and potential user needs" (Eurostat, 2014). "It refers to whether all the statistics that are needed are produced and the extent to which concepts used (definitions, classifications etc.) reflect user needs" (Komuso (ESSnet Quality of Multisource Statistics), 2019).

2.2 Basic data configurations

As already mentioned above, WP 3 of Komuso used a breakdown into a number of BDCs that are most commonly encountered in practice. In Komuso, we identified six BDCs [for more information on BDCs and methods to produce multisource statistics we refer to De Waal et al. (2020)]:

- BDC 1: multiple non-overlapping cross-sectional microdata sources that together provide a complete dataset without any under-coverage problems;
- BDC 2: same as BDC 1, but with overlap between different data sources;

- BDC 3: same as BDC 2, but now with under-coverage of the target population;
- BDC 4: microdata and aggregated data that need to be reconciled with each other;
- BDC 5: only aggregated data that need to be reconciled;
- BDC 6: longitudinal data sources that need to be reconciled over time (benchmarking).

2.3 Error types

There exist many different schemes of error categories in survey and administrative sources; see for instance Zhang (2012). The error categories that are distinguished also depend on the level of detail that is used. Table 1 provides an overview of the error categories that were distinguished in Komuso.

Error category	Type of error included	Survey	Admin
Validity error	Specification error	X	
	Relevance error		X
Frame and source error	Under-coverage	X	X
	Over-coverage	X	X
	Duplications	X	X
	Misclassification in the contact variables	X	
	Misclassification in the auxiliary variables	X	X
Selection error	Error in terms of the selected sampling units	X	
	Unit non-response	X	
	Missing units in the accessed dataset		X
Measurement error and Item missingness	Arising from: respondent, questionnaire, interviewer, data collection	X	
	Fallacious or missing information in admin source		X
Processing error (*)	Data entry error	X	
	Coding or mapping error or misclassification	X	X
	Editing and imputation error	X	X
	Identification error		X
	Unit error		X
	Linkage errors	X	X
Model error (examples, non-exhaustive)	Editing and imputation error, record linkage error, ...	X	X
	Model based estimation error (Small Area Estimation, Seasonal Adjustment, Structural Equation Modeling, Bayesian approaches, Capture-Recapture or Dual System Estimation, Statistical Matching, ...)	X	X

Table 1: Main sources of errors in multisource statistics [from Komuso (ESSnet Quality of Multisource Statistics) (2019)]. (*) Processing errors are errors occurring with manual activities. These include also trivial errors, e.g., typographical errors in writing a procedure or errors in specifying a variable in the program (also in a model). When the processing steps mentioned are done via a model they may result in model errors.

2.4 General strategies to measuring output quality

In Komuso, four different general strategies were identified that one can take with respect to measuring quality: (1) using a quality framework without trying to quantify quality measures, (2) us-

ing generic quality measures that do not rely on any underlying model or design, (3) using non-parametric models to quantify quality measures (including sampling theory), and (4) using parametric models to quantify quality measures.

1. A quality framework is often used for measuring the quality of an entire statistical production chain, including data collection and data processing steps, since constructing statistical models for all steps in the statistical production chain is generally not feasible. A quality framework does not rely on statistical distributions nor on distribution-free quality measures. Instead, a quality framework often tries to combine various pieces of information on the quality of the produced output, such as expert opinions, into a set of quality measures.
2. Generic quality measures that do not rely on any underlying model or design are often used when one wants to measure quality for a single step in the statistical production chain, rather than measure the quality of the entire statistical production chain as a quality framework aims to do. Such quality measures are often used for steps in the statistical production that are hard or impossible to capture in a statistical model. An example is the difference between an earlier estimate and the latest revisions, which quantifies the effect of revisions (see Subsection 3.1.2). The QMCMs using generic quality measures that do not rely on any underlying model or design that have been developed in Komuso all concern the dimension "coherence".
3. Sampling theory, including resampling techniques, is the most common form of non-parametric models for measuring output quality, in any case within the Komuso project. Generally, (re)sampling theory is used to estimate bias and variance for estimators based on random samples.
4. Parametric models are often used when one wants to measure quality for a single step in the statistical production chain, and such a step can be captured in a statistical model. Examples of such methods are constrained optimization, latent class modeling, and structural equation modeling.

Within each general strategy one can use different methods to calculate quality measures. An overview of some of the methods encountered in Komuso is given in Sections 3 to 5.

3 Scoring methods: "Art is born of the observation and investigation of nature"

Scoring methods basically just base any quality measure directly on the observations, without relying, for instance, on statistical models. Scoring methods remind us of a quote by Marcus Tullius Cicero (Roman statesman, lawyer and scholar): *"Art is born of the observation and investigation of nature."*

3.1 Scoring methods for single and multisource statistics

3.1.1 Qualitative methods

The usual objective of qualitative methods is to collect non-numerical data such as reasons, (expert) opinions, and motivations. Examples of qualitative methods are individual interviews and group discussions.

Example: Two-phase and three-phase error framework

A multisource production process may consist of several administrative registers that are linked and harmonized by means of micro-integration [see, e.g., Bakker (2011)] for constructing a statistical register and deriving the variable of interest and related variables. The steps in this production process are usually complicated. Zhang (2012) developed a two-phase error framework for the situation where data from multiple sources are integrated to create a statistical micro dataset (see Table 2).

	Measurement dimension	**Representation dimension**
Phase 1	Validity error	Frame error
	Measurement error	Selection error
	Processing error	Missing/redundancy
Phase 2	Relevance error	Coverage error
	Mapping error	Identification error
	Comparability error	Unit error

Table 2: Error sources in the two-phase error framework.

The first phase shows the stages and error sources during the construction of each input source (e.g., an administrative register). For the process along the measurement perspective, the errors are validity error, measurement error, and processing error; along the representation perspective, the errors are frame error, selection error, and missing/redundancy. The model of this phase is adapted from the total survey error model of Groves et al. (2004).

The second phase starts with the target concept and target population, defined according to the purpose of the integrated statistical data. These targets are typically different from the first phase where the targets were defined and generated according to the purpose of the data owner. Because of this, it is sometimes even necessary to "swap" the two measurement and representation dimensions when moving from the first stage to the second stage; e.g., employment can be considered either as representation (of the population of employed people) or as measurement (of employment in the labor force population). During the second phase, the errors related to the measurement dimension are relevance error, mapping error, and comparability error; for the representation dimension the errors are coverage error, identification error, and unit error.

Reid et al. (2017) proposed an extension of this two-phase error framework with a third phase: the estimation phase. According to Reid et al. (2017), phase 2 ends with a unit-level records file containing units and values of the variables. The third phase then describes inaccuracies that can be made during the actual estimation process, in which one may try to correct for errors made during the first two phases. Furthermore, phase three includes estimation of the quality of the output estimates.

To what extent errors are treated can be measured as a simple proportion in this framework, where 1 stands for complete treatment of all the error sources and 0 if none are treated. These values may be based on expert knowledge. Alternatively, any Likert scale measure can be defined subjectively by the expert. Reid et al. (2017) give three examples of how their three-phase error framework was used to qualitatively compare different possible designs for statistical output, treating both single-source and multisource statistics. An example of giving scale measures to errors sources can be found in Biemer et al. (2014). Rocci et al. (2018) also applied an error framework to a multisource statistic, and for different error types they estimated the fraction of units in which that error type occurred.

3.1.2 Descriptive summary statistics

A descriptive summary statistic quantitatively describes features of collected data. A descriptive statistic aims to summarize the observed data and is generally quite simple. Commonly used de-

scriptive summary statistics are the minimum and maximum values of the variables, the means, medians and modes of the variables, the standard deviation and variances of the variables and the correlation between two variables.

In the Komuso project several descriptive summary statistics were examined. Below we give examples of three such statistics. These examples have been developed by ISTAT in the Komuso project.

Example: Cross-domain and sub-annual versus annual statistics coherence

A descriptive summary statistic to measure the coherence of estimates for the same parameter/variable of interest based on cross-domain or sub-annual statistics versus annual statistics is the relative difference between the "main" estimate and the "comparison" estimate. It is computed as

$$I = \frac{y_A - y_B}{y_B} \times 100,$$

where y_A is the main estimate and y_B the comparison estimate. For instance, y_B may be the estimate based on annual statistics and y_A the estimate for the same parameter/variable of interest based on cross-domain or sub-annual statistics. Generally, y_B is based on the most accurate/trustable source, unless there is no reason to consider any of the sources as the most accurate. In the latter case one could consider setting y_B equal to the average of the two estimates.

The above indicator I can be used after the final point estimates have been computed and one or more estimates for the same parameter/variable of interest are available from different sources or from processes with a different frequency.

Example: (Change of) sign, size, bias and variability of revisions and discrepancies

In order to quantify the effect of revisions and discrepancies on statistical estimates one can simply calculate the difference between the latest estimate and earlier estimates in the case of a revision, or the difference between estimates for similar domains in the case of discrepancies. For convenience we will only discuss revisions, but the same holds for discrepancies.

The difference is simply computed as $R^t = L^t - P^t$, where R^t denotes the revision for moment t, L^t the latest estimate for a variable/parameter of interest, and P^t a preliminary estimate for the same variable/parameter of interest. The later calculated estimate L^t is generally considered more reliable than the preliminary estimate P^t. Here, P^t may, for instance, denote the estimated period-on-period growth rate in a certain period, and L^t a later calculated, more accurate, estimated period-on-period growth rate.

Given the calculated revisions R_i^t for several statistics i, the following descriptive summary statistics can, for instance, be estimated: the change of sign due to a revision, the size of the revisions (mean of absolute revisions, median of absolute revisions, mean of relative absolute revisions), bias of the revisions (revision mean and its statistical significance, revision median) and the variability of the revisions (root mean square error, range, min, max, ...).

Seasonally adjusted estimates may be taken into account to calculate descriptive summary statistics. Such seasonally adjusted estimates can, for instance, be obtained by applying available seasonal adjustment software on the unadjusted data.

The descriptive summary statistics are, for instance, applied at ISTAT when monthly seasonally adjusted data of industrial production indices are estimated by means of a direct approach at domain level, and quarterly seasonally adjusted output of the industrial sector is based on disaggregation techniques on annual data with seasonally adjusted industrial production indices. At least two approaches can be applied in such a situation: one can use the quarterly averages of the disseminated seasonally adjusted indices or one can use seasonally adjusted quarterly averages of the unadjusted

monthly indices. The descriptive summary statistics offer some help in choosing between these two (and possibly other) approaches.

Example: Scalar measure of coherence in a reconciled demographic balancing equation
For the situation where estimates related by linear constraints need to be reconciled, descriptive summary statistics are also available. An example of such a situation is when stocks and flows of a population have to be balanced. Another example is when macro-economic figures, for instance figures for the National Accounts, that are connected by accounting equations need to be reconciled.

We will illustrate the descriptive summary statistics that have been examined in the Komuso project by the demographic balancing equation. More in detail, the population sizes in a domain i at times t (P_i^t) and $t+1$ (P_i^{t+1}) (stocks), and flows (migrations, birth and deaths) within the period $[t,\ t+1]$ need to satisfy the demographic balancing equation $P_i^{t+1} = P_i^t + N_i + M_i$, where $N_i = B_i - D_i$ is the natural increase, with B_i and D_i the number of births, respectively the number of deaths in period $[t,\ t+1]$, and $M_i = \sum_j M_{ij}$, where M_{ij} is the number of people who immigrated to domain i from domain j minus the number of people who emigrated from domain i to domain j and the sum is taken over all domains j. Here a domain may be any disjoint partitioning of the population, for instance region by sex by age class.

Let us assume that the estimates for domains i and j are given by $\hat{P}_i^t$, $\hat{P}_i^{t+1}$, $\hat{N}_i$ and $\hat{M}_{ij}$. A simple descriptive summary statistic for the degree of incoherence is then the average over the domains of the differences between the direct estimate for the population $\hat{P}_i^{t+1}$ and the corresponding estimate obtained by the estimates of stock of the population at time t and the flows in period $[t,\ t+1]$, that is by $\hat{P}_i^t + \hat{N}_i + \hat{M}_i$.

An indicator for the degree of incoherence for domain i is

$$d_i = \left|\hat{P}_i^{t+1} - (\hat{P}_i^t + \hat{N}_i + \hat{M}_i)\right|.$$

A descriptive summary statistic for the global measure of coherence is then given by the sum of the differences standardized with respect to the average of the two estimates of the population P_i^{t+1}, i.e., by

$$C = \frac{2}{D} \sum_i \frac{d_i}{\hat{P}_i^{t+1} + \hat{P}_i^t + \hat{N}_i + \hat{M}_i},$$

where D denotes the number of domains and the summation is over all domains $i = 1, \ldots, D$.

The above indicator and descriptive summary statistic for the global measure of coherence do not examine the impact of reconciliation. We will now consider an indicator and descriptive summary statistic for the impact of reconciliation. Let $(\tilde{P}_i^t,\ \tilde{P}_i^{t+1},\ \tilde{N}_i,\ \tilde{M}_{ij})$ be reconciled values that satisfy the demographic balancing equation. Indicators for the impact of reconciliation for the separate variables for each domain i are then given by $\left(\tilde{P}_i^t - \hat{P}_i^t\right)/\hat{P}_i^t$, $\left(\tilde{P}_i^{t+1} - \hat{P}_i^{t+1}\right)/\hat{P}_i^{t+1}$, $\left(\tilde{N}_i^t - \hat{N}_i^t\right)/\hat{N}_i^t$, and $\left(\tilde{M}_i^t - \hat{M}_i^t\right)/\hat{M}_i^t$. The indicators obviously depend on the reconciled values, and hence on the reconciliation method used.

A descriptive summary statistic for the impact of reconciliation based on these indicators is the average of the above four indicators over all domains, i.e.,

$$CR = \frac{1}{4D} \sum_i \left(\left| \frac{\tilde{P}_i^t - \hat{P}_i^t}{\hat{P}_i^t} \right| + \left| \frac{\tilde{P}_i^{t+1} - \hat{P}_i^{t+1}}{\hat{P}_i^{t+1}} \right| + \left| \frac{\tilde{N}_i - \hat{N}_i}{\hat{N}_i} \right| + \left| \frac{\tilde{M}_i - \hat{M}_i}{\hat{M}_i} \right| \right).$$

CR can also be seen as a measure of incoherence, since it quantifies the overall change in values required to obtain the reconciled values. Like the four underlying separate indicators, CR depends on the reconciliation method.

When only a subset of the demographic variables $\hat{P}_i^t$, $\hat{P}_i^{t+1}$, $\hat{N}_i$ and $\hat{M}_i$ are reconciled, or when they are reconciled for a subset of domains i only, the descriptive summary statistic should be computed on that subset only.

CR allows one to compare the impact of several reconciliation methods to each other. By zooming in on specific subsets one can study the impact of reconciliation for certain (groups of) domains or on certain variables.

3.2 Scoring methods developed especially for multisource statistics

3.2.1 Dempster-Shafer theory

Dempster-Shafer theory offers a general methodological framework for dealing with uncertainty. It enables one to combine, possibly conflicting, information from different sources. With Dempster-Shafer theory one can take all available information into account and quantify the degree of belief in a certain outcome by means of a belief function. Such a belief function relates the plausibility of a certain answer to a certain question to the plausibilities of answers to a related question. These degrees of belief may be subjective. For instance, they may be based on expert opinions. Dempster-Shafer theory gives rules for combining degrees of belief that are based on independent sources.

Dempster-Shafer theory can be used in many cases. For instance, the theory can be used when one wants to combine information from several experts or when one wants to combine expert opinions with information based on observed data. This makes Dempster-Shafer theory a very broadly applicable methodological tool.

For more on Dempster-Shafer theory, and on an application of Dempster-Shafer theory to the Austrian Population Census, we refer to Berka et al. (2010), Berka et al. (2012), Schnetzer et al. (2015), and Asamer et al. (2016).

4 Methods based on (re)sampling: "Design is the intermediary between information and understanding"

Methods based on (re)sampling generally use the design, for instance the sampling design, by which the data are collected, to base quality measures upon. For such quality measures the "*design is the intermediary between information and understanding*," a quote by the German painter Hans Hoffman.

Sampling theory allows one to compute the sampling variance for a large number of sampling designs (Särndal et al., 1992). Assuming a mechanism for the non-response process, sampling theory may in some cases also be used to estimate non-response variance, besides sampling variance. In sampling theory, one usually derives analytical formulae to compute sampling (and non-response) variance.

For single-source statistics, calculating the sampling (and non-response) variance is often the only realistic way to estimate the quality of the output. Sampling theory is also very useful for multisource statistics. In the case of multisource statistics, different situations may arise for which one may want to estimate sampling (and non-response) variance than for single-source statistics (see, e.g., the first example in Section 4.2).

Resampling can often be used to estimate the variance (and bias) of an estimator. The advantage of resampling methods is that, while analytic variance formulae need to be derived for different kinds of estimators separately and can become quite complex, resampling methods offer a relatively simple computational procedure for obtaining variance estimates that is general enough to be applicable to many estimation problems.

There are several resampling techniques, such as the jack-knife, balanced repeated replication and subsampling (Wolter, 2007). For example, in the jack-knife, one systematically recomputes estimates for the statistic of interest, leaving out one or more observations at a time from the dataset. From the obtained set of replicates of the statistic, estimates for the bias and variance of the statistic can be obtained.

One of the most frequently used resampling methods is the bootstrap (Efron and Tibshirani, 1993). Bootstrapping is a method of repeated sampling from either a sample (non-parametric bootstrapping) or from an estimated parametric distribution (parametric bootstrapping). Under certain conditions, the variance over the set of bootstrap outcomes is an approximately unbiased estimator for the variance of the original estimator. Likewise, the difference between the mean of the bootstrap estimates and the estimate derived from the original sample is often an approximately unbiased estimator of the bias of the original estimator.

For examples of applications of non-parametric bootstrapping in a multisource context, see, e.g., Kuijvenhoven and Scholtus (2011) and Scholtus and Daalmans (2020). In these papers, the bootstrap is used to estimate the variance of an estimated frequency table involving the highest attained level of education based on combined administrative and survey data, where missing values of education in the target population are accounted for either by weighting (the first reference) or mass imputation (the second reference). An example of an application of a parametric bootstrap method will be given in Section 4.2.

4.1 Methods based on (re)sampling for single and multisource statistics

In this section, we give an example of a case where sampling theory can be applied to both single and multisource statistics. This example was developed by Statistics Lithuania in the Komuso project.

Example: Effect of frame under-coverage / over-coverage on the estimator of a total and its accuracy measures

Sampling theory can be applied to a case where a sample is taken from a frame that is kept constant for a longer time. This may occur in business statistics: some NSIs use a business register that is "frozen" for a year, i.e., certain population changes are stored during the year and they are effectuated only once, at the beginning of a year.

In the case of intra-annual estimators (month, quarter), the corresponding population is likely to have changed compared to the sampling frame. With respect to the true population, the sampling frame suffers both from under- and over-coverage. Assume that an up-to-date administrative source is available that does not suffer from coverage errors. Using this administrative source, an adjusted estimator can be calculated. Next, metrics on differences between the original and the adjusted estimator quantify the sensitivity of the original estimator to coverage errors in the frozen sampling frame.

Suppose we have a sampling frame of a population U of size N, divided into non-overlapping strata U_h of size N_h, with $h = 1,\ldots,H$. From each stratum U_h a random sample s_h is taken of size n_h. For the sample units we collect information on a target variable y. For instance, we are interested in quarterly gross earnings of enterprises (y) by economic sector h, estimated by a sample survey drawn from a frozen business register U. Furthermore we have an auxiliary variable x, for instance the number of employees per enterprise. This variable is assumed to be available for all units in the sampling frame U as well as in a population V of a social insurance inspection data base which contains an up-to-date population for the number of employees. Population V consists of non-overlapping strata

V_h of size M_h, with $h = 1,\ldots,H$. As a result of under- and over-coverage, we find that $U \setminus V \neq \emptyset$ and $V \setminus U \neq \emptyset$.

Now assume that we are interested to study the effect of the frozen register on an estimate of the true population total t_y, the quarterly gross earnings. Based on the sampling frame U we obtain the Horvitz-Thompson estimator for the total $\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh}$, with $\hat{t}_{y_h} = \frac{N_h}{n_h'} \sum_{i=1}^{n_h'} y_{hi}$ where n_h' denotes the number of responding (and alive) enterprises ($n_h' \leq n_h$). For the auxiliary variable x, we know the totals for the frozen register: $t_x = \sum_{h=1}^{H} t_{xh}$ and $t_{xh} = \sum_{i=1}^{N_h} x_{hi}$. Similar to variable y, the Horvitz-Thompson estimators of the totals for x are given by $\hat{t}_x = \sum_{h=1}^{H} \hat{t}_{xh}$ with $\hat{t}_{xh} = \frac{N_h}{n_h'} \sum_{i=1}^{n_h'} x_{hi}$. Furthermore, we have more up-to-date totals based on V: $\tilde{t}_x = \sum_{h=1}^{H} \tilde{t}_{xh}$ and $\tilde{t}_{xh} = \sum_{i=1}^{M_h} x_{hi}$.

Economic sector	$\hat{R}_t$		$\hat{R}_{\text{Var}}$	
	Q1	Q4	Q1	Q4
A: Agriculture	−0.64	−1.03	−0.46	−0.18
B: Mining and quarrying	−1.04	−1.21	−2.08	−2.41
C: Manufacturing	−0.52	−0.73	−0.78	−1.26
D: Electricity, gas, steam and air conditioning supply	−0.59	−1.17	+9.66	+16.33
E: Water supply; sewerage; waste management and remediation activities	−0.25	−0.51	+0.70	+1.36
F: Construction	−1.49	−2.42	−2.77	−4.48
G: Wholesale and retail trade; repair of motor vehicles and motorcycles	−1.34	−1.71	−2.68	−3.40
H: Transportation and storage	−0.94	−1.19	−0.58	−0.21
I: Accommodation and food service activities	−1.36	−2.08	−2.67	−4.07

Table 3: Changed totals ($\hat{R}_t$) and changed variances ($\hat{R}_{\text{Var}}$) for separate ratio estimators of quarterly gross earnings for the first and fourth quarter of 2015 in Lithuania for a selection of economic sectors, obtained from Krapavickaitė and Šličkutė-Šeštokienė (2017).

We can now use a separate ratio estimator or a combined ratio estimator for t_y and compare the original estimator (based on U) with an updated version (based on V). The original separate ratio estimator and its updated version are given by:

$$\hat{t}_y^{(sep)} = \sum_{h=1}^{H} t_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}},$$

$$\tilde{t}_y^{(sep)} = \sum_{h=1}^{H} \tilde{t}_{xh} \frac{\hat{t}_{yh}}{\hat{t}_{xh}}.$$

Likewise, the original combined ratio estimator and its updated version are given by

$$\hat{t}_y^{(comb)} = t_x \frac{\hat{t}_y}{\hat{t}_x},$$

$$\tilde{t}_y^{(comb)} = \tilde{t}_x \frac{\hat{t}_y}{\hat{t}_x}.$$

Note that when the population has not changed, i.e., when V is the same as U, then $t_{xh} = \tilde{t}_{xh}$ (for $h = 1,...,H$), and the separate ratio and combined ratio estimators based on V are indeed equal to the corresponding estimators based on U.

For both estimators, i.e., the separate ratio estimator and the combined ratio estimator, we can quantify the effect of the change of the frame by looking into the extent that the totals have changed, and the extent that their variances have changed:

$$\hat{R}_t = \frac{\tilde{t}_y - \hat{t}_y}{\hat{t}_y},$$
$$\hat{R}_{\text{Var}} = \frac{\widehat{\text{Var}}\left(\tilde{t}_y\right) - \widehat{\text{Var}}\left(\hat{t}_y\right)}{\widehat{\text{Var}}\left(\hat{t}_y\right)},$$

where we have omitted the superscripts "(sep)" and "(comb)". The variances can be derived analytically with standard sampling theory; see, e.g., Särndal et al. (1992, p. 253 and pp. 270-271). Note that $\hat{\Delta}_y = \hat{t}_y - \tilde{t}_y$ is an estimate of the bias of the total $\hat{t}_y$ due to the use of a frozen sampling frame. Note further that $\hat{\Delta}_y / \sqrt{\widehat{\text{Var}}(\hat{\Delta}_y)}$ could be used to test whether the difference between the two estimators is significant.

The indicators $\hat{R}_t$ and $\hat{R}_{\text{Var}}$ for the separate ratio estimator have been applied to the Lithuanian survey on earnings in 2015. Outcomes for the first (Q1) and last quarter (Q4) are shown in Table 3 for the first nine economic sectors; more results can be found in Krapavickaitė and Šličkutė-Šeštokienė (2017). The results on $\hat{R}_t$ show that the estimator becomes more sensitive to frame updates from Q1 to Q4 when coverage errors have increased.

4.2 Methods based on (re)sampling developed especially for multisource statistics

In this section, we give two examples of methods based on (re)sampling theory that have been developed especially for multisource statistics.

Example: Variance of cell values in estimated frequency tables
The Repeated Weighting (RW) estimator was developed to ensure numerical consistency among tables estimated from different combinations of administrative data and sample surveys (Houbiers, 2004). The basic idea of RW is to use the regression estimator to calibrate table estimates to any marginal tables that have been estimated previously. Calculation of the variances of the resulting estimates can be rather complicated.

In general, the RW estimation procedure consists of the following three steps:

1. Specify the set of target tables to be estimated, and order them in descending order of available information.
2. Estimate each table separately from an appropriate subset of the available data, called a block. In general, a block will consist of the largest survey or combination of surveys in which all variables of the table are observed. It is assumed that a regression estimator is used in this step, based on auxiliary variables that are observed throughout the population.
3. Perform reweighting: adjust each table consecutively using the regression estimator, in the order specified in step 1, so that numerical consistency is achieved for any part of the table (including its margins) that overlaps with a previously estimated table.

It should be noted that tables that are estimated from the same block using the same regression estimator (as is done in Step 2) are automatically numerically consistent. Reweighting is applied when not all tables can be estimated from the same block.

As an illustration, consider the following example taken from Knottnerus and van Duin (2006). There are three categorical variables, X, Y and Z. The available data consist of one register and two without-replacement samples S_1 and S_2 of sizes n_1 and n_2, respectively. The auxiliary variable X is observed in the register for all N units in the population. The target variable Y is observed only in S_2; the target variable Z is observed in S_1 and S_2. We want to find consistent estimates for two tables: t_Z and $t_{Z\times Y}$. Note that the first table is actually a margin of the second table. Since more information is available for estimating t_Z than for estimating $t_{Z\times Y}$, the two tables will be estimated in this order from different blocks. [Technically, we are using the so-called "splitting up" variant of RW here (Knottnerus and van Duin, 2006).]

Let S_{12} denote the union of S_1 and S_2. We begin by deriving initial estimates for the two tables using the regression estimator with X as auxiliary information. This yields: $\hat{t}_Z^{REG(S_{12})}$, estimated from S_{12}, and $\hat{t}_{Z\times Y}^{REG(S_2)}$, estimated from S_2. In general, the estimated margin for Z from $\hat{t}_{Z\times Y}^{REG(S_2)}$ will be numerically inconsistent with $\hat{t}_Z^{REG(S_{12})}$. In the third step of the RW procedure, $\hat{t}_{Z\times Y}^{REG(S_2)}$ is therefore reweighted with respect to its Z-margin, using the regression estimator. This yields:

$$\hat{t}_{Z\times Y}^{RW} = \hat{t}_{Z\times Y}^{REG(S_2)} + \hat{\boldsymbol{B}}'_{w;Z}\left(\hat{t}_Z^{REG(S_{12})} - \hat{t}_Z^{REG(S_2)}\right).$$

Here, $\hat{\boldsymbol{B}}_{w;Z}$ denotes a matrix of estimated regression coefficients, and $'$ denotes the transpose. More generally, if there was also additional information available outside S_2 about the Y-margin, then the table would simultaneously be reweighted with respect to this margin, leading to a third term in the above expression for $\hat{t}_{Z\times Y}^{RW}$.

To estimate the variance of this RW estimator, Knottnerus and van Duin (2006) noted that, under certain regularity assumptions, $\hat{t}_{Z\times Y}^{RW}$ can be approximated by

$$\hat{t}_{Z\times Y}^{RW} = t_{Z\times Y} + \hat{t}_{e(Z\times Y)}^{HT(S_2)} + \boldsymbol{B}'_Z\left(\hat{t}_{e(Z)}^{HT(S_{12})} - \hat{t}_{e(Z)}^{HT(S_2)}\right) + O_p\left(N/n_2\right),$$

where $e(.)$ denotes a vector of residuals from a regression of $(.)$ on the register variable X, the superscript HT denotes a Horvitz-Thompson estimator, and $\boldsymbol{B}_Z$ is the matrix of population regression coefficients estimated by $\hat{\boldsymbol{B}}_{w;Z}$. Now, assuming for simplicity that $1 \ll n_1, n_2 \ll N$ and that the two samples S_1 and S_2 are independent, the variance-covariance matrix of $\hat{t}_{Z\times Y}^{RW}$ can be estimated by

$$\widehat{\mathrm{cov}}\left(\hat{t}_{Z\times Y}^{RW}\right) = \sum_{i\in S_1}\left(d_i^{(S_1)}\right)^2 \epsilon_{1i}\epsilon'_{1i} + \sum_{i\in S_2}\left(d_i^{(S_2)}\right)^2 \epsilon_{2i}\epsilon'_{2i},$$

where $d_i^{(S_k)}$ is the design weight of unit i in sample S_k $(k = 1, 2)$, $\epsilon_{1i} = \lambda_1 \boldsymbol{B}'_Z e_i(Z)$, $\epsilon_{2i} = e_i(Z\times Y) - \lambda_1 \boldsymbol{B}'_Z e_i(Z)$, and λ_1 is a weighting factor that reflects the relative reliability of S_1 in S_{12}. A simple choice that is often made is to set $\lambda_1 = n_1/(n_1+n_2)$. In particular, the square roots of the diagonal elements of $\widehat{\mathrm{cov}}\left(\hat{t}_{Z\times Y}^{RW}\right)$ provide standard errors for the cells of the estimated table $\hat{t}_{Z\times Y}^{RW}$.

The variables ϵ_{1i} and ϵ_{2i} in the above expression are examples of "super-residuals", which are linear combinations of ordinary regression residuals. More generally, Knottnerus and van Duin (2006) showed that the variance-covariance matrix of an RW estimator for a frequency table can always be approximated by means of super-residuals. If the above assumptions that $n_1, n_2 \ll N$ and/or that the two samples are independent do not hold, other variance estimators from sample survey theory can be used (Knottnerus and van Duin, 2006). Unlike the above variance estimator, these other variance estimators require that all second-order inclusion probabilities are known,

which may be difficult to achieve in practice.

Example: Bias and variance of parameter estimates affected by classification errors
NSIs often publish statistics that are obtained by aggregating numerical variables separately for each domain defined by a classification variable. For instance: the total turnover of businesses by type of economic activity, or the average hourly income of employed persons by highest attained education level. If the numerical variables are observed accurately for all units in the target population (e.g., in an administrative dataset), then the main issue affecting the quality of these statistics may be errors in the assignment of units to the right domain (classification errors). Van Delden et al. (2016) proposed a parametric bootstrap method to evaluate the bias and variance of statistics due to classification errors, under the simplifying assumption that these are the only errors that occur.

Let $i = 1,\ldots,N$ denote the units in the target population. Given the classification of interest, each unit has an unknown true code $s_i \in \{1,\ldots,H\}$, where H is the total number of classes. For each unit in the population, we observe a code $\hat{s}_i \in \{1,\ldots,H\}$ which may or may not be equal to the true code. We suppose that random classification errors occur, independently across units, according to a (possibly unit-specific) transition matrix $\boldsymbol{P}_i = (p_{ghi})$, with $p_{ghi} = \mathrm{P}(\hat{s}_i = h \mid s_i = g)$. The true codes are considered fixed.

In general, we write the target parameter as a function $\theta = f(y_1,\ldots,y_N;s_1,\ldots,s_N)$, where y_i denotes the value of a numerical target variable for unit i (or, more generally, a vector of numerical variables). For instance, a domain total can be written as $\theta_g = \sum_{i=1}^{N} y_i I\{s_i = g\}$, where $I\{.\}$ equals 1 if its argument is true and 0 otherwise. We assume that all $y_1,\ldots,y_N$ are known. An important special case where this is trivially true occurs when the target parameter is the number or proportion of units per domain (i.e., a domain total with $y_i \equiv 1$ and $y_i \equiv 1/N$, respectively). Given the assumption that no errors occur besides classification errors, θ can be estimated from the observed data by $\widehat{\theta} = f(y_1,\ldots,y_N;\hat{s}_1,\ldots,\hat{s}_N)$. We are interested in the bias and variance of $\widehat{\theta}$ as an estimator for θ.

As a preliminary step towards evaluating the bias and variance due to classification errors, we need to estimate the probabilities in the transition matrix $\boldsymbol{P}_i$. Typically, this requires the collection of additional data on the classification variable. Possible approaches include:

- Draw a random audit sample of units for which, in addition to $\hat{s}_i$, the true code s_i is obtained, e.g., by manual verification.
- Use process information from an editing step during regular production, where $\hat{s}_i$ may have been checked and corrected for certain units.
- Use multiple measurements of s_i from different, independent sources (e.g., a population register, an external administrative source and a sample survey). Under certain conditions, the error probabilities can be estimated from these multiple measurements using latent class analysis (see also Section 5.2.2).

In general, some model assumptions have to be introduced to reduce the number of unknown parameters. In this way, the unit-specific transition matrix $\boldsymbol{P}_i$ can be estimated as a function of a limited number of background variables. An example can be found in Van Delden et al. (2016). In applications where the codes $\hat{s}_i$ are predicted from a machine-learning algorithm, an estimate for $\boldsymbol{P}_i$ may be obtained naturally when the quality of the algorithm is evaluated. See, e.g., Meertens et al. (2020) for an example of such an application.

Having obtained an estimated transition matrix $\widehat{\boldsymbol{P}}_i = (\hat{p}_{ghi})$, we can apply the bootstrap method. For each unit, we draw a new code $\hat{s}^*_{ir}$ given the original observed code $\hat{s}_i$, using probabilities that mimic (our best estimate of) the original process by which $\hat{s}_i$ was generated from s_i:

$$\mathrm{P}\left(\hat{s}^*_{ir} = h \mid \hat{s}_i = g\right) \equiv \widehat{\mathrm{P}}(\hat{s}_i = h \mid s_i = g) = \hat{p}_{ghi}.$$

Based on the obtained values $\hat{s}^*_{1r},\ldots,\hat{s}^*_{Nr}$, we compute the bootstrap replicate $\widehat{\theta}^*_r = f(y_1,\ldots,y_N;\hat{s}^*_{1r},\ldots,\hat{s}^*_{Nr})$. This procedure is repeated R times, yielding replicates $\widehat{\theta}^*_1,\ldots,\widehat{\theta}^*_R$. From these replicates, the bias and variance of $\widehat{\theta}$ are estimated as follows [see also Efron and Tibshirani (1993)]:

$$\hat{B}^*_R\left(\widehat{\theta}\right) = m_R\left(\widehat{\theta}^*\right) - \widehat{\theta},$$
$$\hat{V}^*_R\left(\widehat{\theta}\right) = \frac{1}{R-1}\sum_{r=1}^{R}\left\{\widehat{\theta}^*_r - m_R\left(\widehat{\theta}^*\right)\right\}^2,$$

with $m_R\left(\widehat{\theta}^*\right) = R^{-1}\sum_{r=1}^{R}\widehat{\theta}^*_r$. In the bootstrap literature, it is often recommended to take $R \geq 200$ for variance estimates and $R \geq 1000$ for bias estimates.

An advantage of the bootstrap is that the above algorithm can be applied to many different types of estimators in the same way. For instance, θ could be a regression coefficient or a median. For certain simple target parameters, it is possible to derive an explicit formula for the analytical bias and variance estimators to which $\hat{B}^*_R\left(\widehat{\theta}\right)$ and $\hat{V}^*_R\left(\widehat{\theta}\right)$ converge for $R \to \infty$. As an example, Van Delden et al. (2016) obtained the following formulas for the estimated bias and variance of an estimated domain total $\widehat{\theta}_h = \sum_{i=1}^{N} y_i I\{\hat{s}_i = h\}$:

$$\hat{B}^*_\infty\left(\widehat{\theta}_h\right) = \sum_{i=1}^{N} y_i\left[(\hat{p}_{hhi}-1)\,I\{\hat{s}_i = h\} + \sum_{g=1,g\neq h}^{H}\hat{p}_{ghi}I\{\hat{s}_i = g\}\right],$$
$$\hat{V}^*_\infty\left(\widehat{\theta}_h\right) = \sum_{i=1}^{N} y_i^2\sum_{g=1}^{H}\hat{p}_{ghi}(1-\hat{p}_{ghi})I\{\hat{s}_i = g\}.$$

It can be shown that, in general, these bias and variance estimators are biased for the true bias and variance of $\widehat{\theta}_h$; hence, the same holds for the above bootstrap estimators based on a finite number of replicates. For simple target parameters such as a domain total, it is possible to correct for this bias in the estimated bias and variance, although this typically leads to bias and variance estimates that are less stable. See Van Delden et al. (2016) and Kloos et al. (2020) for more details.

5 Methods based on (parametric) modelling: "Artists can spend a lifetime searching for a perfect model"

Finally, we consider methods for calculating quality measures based on (parametric) modelling. Such methods usually construct a model for the target variable(s) to be estimated, and then use properties of the estimated model, such as the bias and variance of an estimator based on the model, as (basis for) quality measures. In some cases, methods based on (parametric) models give excellent results. In other cases, a suitable model may be difficult or even impossible to construct: "*Artists can spend a lifetime searching for a perfect model*," a quote by the American painter Robert Liberace.

5.1 Methods based on (parametric) modelling for single and multisource statistics

Besides methods that are used for measuring output quality directly, there are also some supporting methods, which do not measure output quality directly but are often used in combination with other methods that do measure quality directly. Examples of such supporting methods are estimating equations, log-linear modelling and mixture models.

Estimating equations specify how the parameters of a statistical model should be estimated. Examples of estimating equations are the method of moments, minimum distance methods like least squares estimation, Bayesian methods and (in some cases) maximum likelihood estimation [see, for example, Van der Vaart (1998)]. The idea of the estimation equations method is to find a set of simultaneous equations, involving observed data and model parameters of a statistical model, that need to be solved in order to find estimates of the model parameters.

In a log-linear model [see, for example, Bishop et al. (1975)], a logarithm of a certain variable equals a linear combination of the parameters of a statistical model. The technique is often used to study the relationship between several categorical variables. It can be used to build a statistical model as well as for statistical hypothesis testing.

Mixture models are often used in statistics when there are several subpopulations with different characteristics within the population [see, for example, McLachlan and Peel (2000)]. When using such a mixture model, it is not necessary to identify to which subpopulation each individual observation belongs. Mixture models can, for instance, be used when there are different subpopulations within the population with different kinds or different rates of measurement errors.

Below we discuss constrained optimization, a (parametric) modelling method that can be used for both single and multisource statistics.

5.1.1 Constrained optimization

Constrained optimization aims to optimize an objective function with respect to some variables, given constraints on those variables. These constraints can be either hard constraints, i.e., constraints that need to be satisfied, or soft ones, i.e., constraints that preferably – but not necessarily – should be satisfied. Soft constraints are often taken into account by incorporating them into the objective function, and penalizing the violation of such soft constraints.

Constrained optimization can, for instance, be used to adjust the values of some variables so they satisfy (or nearly satisfy) certain hard or soft constraints. Constrained optimization can also be used to benchmark data over time, i.e., to ensure that high-frequency time series data are reconciled with low-frequency time series data. These kinds of problems are quite common in, for instance, National Accounts.

Constrained optimization can also be used for single-source statistics, for instance when one wants to impute missing data such that constraints within individual records are satisfied (De Waal et al., 2011, Chapter 10).

Example: Macro integration / Data reconciliation

Macro-integration is often used for National Accounts and other kinds of statistical accounts. It can be carried out by means of several methods. We start by describing Stone's method [see, e.g., Stone et al. (1942) and Bikker et al. (2011)]. After describing Stone's method, we will focus on the univariate Denton method and then on the multivariate Denton method.

Suppose that $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ is a vector of high frequency, say quarterly, numerical data. We denote the corresponding estimated covariance matrix by $\boldsymbol{V}$. Let us assume that our aim is to ensure by means of macro-integration that the reconciled components of $\boldsymbol{x}$ sum up to the values of low frequency, say annual, data $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)'$, where $n = 4m$. We then have to fulfill the following constraints

$$\sum_{j=4(k-1)+1}^{4k} x_j = b_k, \qquad \text{for } k = 1, \ldots, m. \tag{1}$$

These constraints are generally violated by the initial high frequency data $\boldsymbol{x}$. The vector $\boldsymbol{x}$ is therefore adjusted so the adjusted vector $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)'$ does satisfy (1).

In matrix notation (1) can be written as

$$\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{A}$ is an $m \times n$ matrix given by

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{j} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{j} & \ldots & \boldsymbol{0} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{j} \end{pmatrix}$$

with $\boldsymbol{j} = (1,1,1,1)$ and $\boldsymbol{0} = (0,0,0,0)$.

In Stone's method the adjustment is done by minimizing the quadratic distance function

$$\min_{\tilde{\boldsymbol{x}}} (\tilde{\boldsymbol{x}} - \boldsymbol{x})' \boldsymbol{V}^{-1} (\tilde{\boldsymbol{x}} - \boldsymbol{x}) \tag{3}$$

subject to the constraints (1). The quadratic optimization problem (3) subject to (1) can be solved by the Lagrange multiplier method. The thus reconciled vector $\tilde{\boldsymbol{x}}$ is given by

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{V}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1} (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \tag{4}$$

and the variance $\tilde{\boldsymbol{V}}$ of $\tilde{\boldsymbol{x}}$ is

$$\tilde{\boldsymbol{V}} = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{V}. \tag{5}$$

(5) is a measure for the quality of the reconciled data.

A drawback of distance function (3) in the case of quarterly time series and annual time series is that discontinuity may arise between the last quarter of one year and the first quarter of the next year. The univariate Denton method (Denton, 1971) aims to avoid this discontinuity by minimizing a quadratic function based on differences between the first order differences, i.e., on $\Delta^{(1)} x_j = \Delta\tilde{x}_j - \Delta x_j$ where $\Delta\tilde{x}_j = \tilde{x}_j - \tilde{x}_{j-1}$ and $\Delta x_j = x_j - x_{j-1}$, rather than on differences between the levels of original and reconciled time series. The underlying idea is to preserve as much as possible the original quarter to quarter changes (the movement preservation principle). Note that Δx_1 is undefined and a value needs to be specified for Δx_1. Denton proposed to use $\Delta x_1 = x_1$. So, the univariate Denton method consists of solving

$$\min_{\tilde{\boldsymbol{x}}} \sum_{j=1}^{n} \left(\Delta\tilde{x}_j - \Delta x_j\right)^2 \tag{6}$$

subject to the boundary condition $\Delta x_1 = x_1$.

That Δx_1 needs to be fixed to a value can be seen as a disadvantage of the univariate Denton method. This disadvantage can be overcome by using the Cholette adaptation. For the Cholette adaptation we first rewrite (6) as

$$\min_{\tilde{\boldsymbol{x}}} (\tilde{\boldsymbol{x}} - \boldsymbol{x})' (\boldsymbol{D}'\boldsymbol{D}) (\tilde{\boldsymbol{x}} - \boldsymbol{x}), \tag{7}$$

where

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

That is, we use $\boldsymbol{V}^{-1} = \boldsymbol{D}'\boldsymbol{D}$ in formula (3). In the Cholette adaptation, the first line from the matrix $\boldsymbol{D}$ is deleted and the generalized inverse of $\boldsymbol{D}'\boldsymbol{D}$ needs to be taken to find $\boldsymbol{V}$ in (4) and (5).

Denton's idea can be taken further: the quadratic distance function can also be based on the second order differences $\Delta^{(2)}x_j = \Delta^{(1)}\tilde{x}_j - \Delta^{(1)}x_j$ (Sax and Steiner, 2013), i.e., instead of minimizing (6) one can minimize

$$\min_{\tilde{\boldsymbol{x}}} \sum_{j=1}^{n} \left(\Delta^{(1)}\tilde{x}_j - \Delta^{(1)}x_j\right)^2, \tag{8}$$

or even on third or higher order differences. Again, the Cholette adaptation can be used for the alternative distance function (8) as well.

Thus far we discussed *additive* Denton methods. Besides additive Denton methods also proportional Denton methods exist. We first define

$$\Delta_{prop}\left(x_j\right) = \frac{\tilde{x}_j - x_j}{x_j} - \frac{\tilde{x}_{j-1} - x_{j-1}}{x_{j-1}} = \frac{\tilde{x}_j}{x_j} - \frac{\tilde{x}_{j-1}}{x_{j-1}}.$$

Instead of solving (6), we then solve

$$\min_{\tilde{\boldsymbol{x}}} \sum_{j=1}^{n} \left(\Delta_{prop}\left(x_j\right)\right)^2.$$

In a similar way a proportional version of (8) can be formulated.

The univariate Denton method can be extended to multiple variables. The multivariate Denton method allows linear restrictions for separate variables as well as linear restrictions involving several variables. So, besides reconciliation of quarterly data to annual data, relationships between different variables can be taken into account.

By $\boldsymbol{x}$ we now denote a vector consisting of M variables that are observed k times per year for T years, i.e., $\boldsymbol{x}$ is an MkT-dimensional vector. The multivariate Denton method consists of solving a quadratic optimization problem (3) subject to (2). The only difference with Stone's method is that $\boldsymbol{x}$ now consists of several variables, and that $\boldsymbol{A}$ and $\boldsymbol{b}$ are now a matrix, respectively vector, that together describe the reconciliation constraints and constraints between different variables. $\boldsymbol{V}$ is again the covariance matrix of vector $\boldsymbol{x}$. The solution is given by (4) with variance given by (5). For more details on the multivariate Denton method we refer to Di Fonzo and Marini (2003) and Bikker et al. (2011).

5.2 Methods based on (parametric) modelling developed for multisource statistics only

In this section we discuss five commonly used methods for measuring output quality that can be used for multisource statistics only, and were examined in Komuso.

5.2.1 Capture-recapture methodology

Capture-recapture methodology is commonly used to estimate the size of a population. The methodology originated in ecology where it is used to estimate an animal population's size. In that context, a number of animals are captured, marked and then released. Later, again a number of animals are captured. The number of marked animals in the second sample can then be used to obtain an estimate for the total number of animals.

The methodology is also used by NSIs to estimate the size of a population. To estimate the total number of individuals that possess a certain characteristic, one records the individuals with that

characteristic occurring in a certain dataset, for example, a census. Next, one counts how many of the recorded individuals occur in another dataset, for example, a post-enumeration survey. This allows one to obtain an estimate for the total number of individuals with this characteristic in the population. Instead of a census or survey, administrative data may also be used.

Example: Capture-recapture methodology in its basic form
We will describe the basic form of capture-recapture methodology. We assume that two datasets A and B of the same fixed population size N are linked. We also assume that the following five technical assumptions are satisfied:

1. inclusion of an element into dataset A is independent of its inclusion in dataset B;
2. inclusion probabilities of units in at least one of the datasets are homogeneous, i.e., all units have an equal probability to be included in this dataset;
3. the population is closed;
4. it is possible to link the elements of datasets A and B perfectly;
5. The datasets do not contain units that do not belong to the target population ("erroneous captures"), nor do they contain duplicates.

Table 4 below describes how many units in datasets A and B occur in both datasets (n_{AB}), in dataset A only (n_A), in dataset B only (n_B), and how many units in the population occur in neither of the two datasets (n_{00}). The value of n_{00} is unknown and has to be estimated. Once the value of n_{00} is estimated, the population size can easily be estimated.

		Dataset B	
		Yes	No
Dataset A	Yes	n_{AB}	n_A
	No	n_B	n_{00}

Table 4: Numbers of units in datasets A and B.

When all five above-mentioned assumptions are valid, n_{00} can be estimated by means of the Petersen estimator [see, e.g., Sekar and Deming (1949)]. The Petersen estimate for n_{00} is

$$\hat{n}_{00} = \frac{n_A n_B}{n_{AB}}.$$

The Petersen estimate for the population size is then given by

$$\hat{N} = n_A + n_B + n_{AB} + \hat{n}_{00}.$$

An estimator for the variance of $\hat{N}$ is given by [see, e.g., Sekar and Deming (1949) and Bishop et al. (1975)]:

$$\widehat{\mathrm{Var}}\left(\hat{N}\right) = \frac{(n_A + n_{AB})(n_B + n_{AB})\, n_A n_B}{(n_{AB})^3}.$$

Van der Heijden et al. (2012) and Gerritse et al. (2015) consider more complicated approaches involving covariates. Those approaches are based on log-linear modelling instead of the Petersen estimator to estimate the unknown population size.

In general, the saturated log-linear model for a contingency table as in Table 4 would be given by

$$\log(n_{00}) = \mu, \tag{9}$$
$$\log(n_A) = \mu + \mu_A, \tag{10}$$
$$\log(n_B) = \mu + \mu_B, \tag{11}$$
$$\log(n_{AB}) = \mu + \mu_A + \mu_B + \mu_{AB}, \tag{12}$$

where μ_A, μ_B, and μ_{AB} indicate that the number of units in the corresponding cell depends on dataset A, on dataset B, or on both. However, in our case, equation (12) of the saturated log-linear model has to be replaced by

$$\log(n_{AB}) = \mu + \mu_A + \mu_B, \tag{13}$$

since the interaction term μ_{AB} cannot be identified. Assuming that datasets A and B are independent (i.e., the first technical assumption made above) this term can be set to zero.

We can estimate the model parameters μ, μ_A, and μ_B using those relations for which we know the cell totals, i.e., (9), (10) and (11) in our case. In general, we can estimate the model parameters by means of maximum likelihood estimation. In our simple example we can compute μ, μ_A and μ_B directly. From (9) to (11) and (13), we directly obtain $n_{00} = \exp(\mu)$, $n_A = \exp(\mu + \mu_A)$, $n_B = \exp(\mu + \mu_B)$ and $n_{AB} = \exp(\mu + \mu_A + \mu_B)$. This means that

$$\hat{n}_{00} = \exp(\hat{\mu}) = \frac{\exp(\hat{\mu} + \hat{\mu}_A)\exp(\hat{\mu} + \hat{\mu}_B)}{\exp(\hat{\mu} + \hat{\mu}_A + \hat{\mu}_B)} = \frac{n_A n_B}{n_{AB}}.$$

That is, the saturated log-linear model under the assumption that datasets A and B are independent gives the same estimate as the Petersen estimator.

An advantage of using log-linear models instead of the Petersen estimator is that they are easy to extend to more general cases, such as:

- three or more datasets instead of two;
- using auxiliary data besides the cell counts, for instance using an available auxiliary variable "gender" to differentiate between counts for women and counts for men, which may improve the quality of the final estimate for the population size;
- adding interaction terms between counts and available auxiliary variables.

For a given situation one can base one or more log-linear models on substantive knowledge, and then select the "best" model. What is considered "best" in a given situation may depend on the model fit (e.g., one can use a chi-square distribution where observed values are compared to expected values), the number of model parameters, and substantive considerations.

As already mentioned, the parameters of log-linear models can be estimated by means of maximum likelihood estimation. In this estimation procedure it may be necessary to set some model parameters to zero beforehand, since otherwise some parameters cannot be identified.

The variance of population size estimates based on a log-linear model can be estimated by means of a bootstrap procedure. For an example where log-linear models are used to estimate the population size, and bootstrapping is used to estimate the variance of the estimated population size, we refer to Van der Heijden et al. (2012) and Gerritse et al. (2015). For more on log-linear modelling in general, see, e.g., Bishop et al. (1975) and Agresti (2013).

5.2.2 Latent variable/class modeling and structural equation modeling

A latent variable model is a statistical model that relates a set of observable variables to a set of non-observable variables. The non-observable variables are called latent variables; the observable variables manifest variables. In a latent class model (Hagenaars and McCutcheon, 2002; Biemer, 2011), the latent variables are categorical.

Latent variable modeling is strongly related to structural equation modeling. In structural equation modeling, one also relates one or more unobserved latent variables to a set of observed variables. Structural equation modeling can be applied to both categorical and numerical variables. In fact, latent class analysis can be considered as a type of structural equation modeling for categorical data.

In both latent class modeling and structural equation modeling, true values can be fitted as a function of background variables or they can be modeled over time. In latent class modeling one estimates the probability that a certain value is observed given the true value. In structural equation modeling, each observed value is considered to be a function of the latent true value plus an error.

In the context of measuring the quality of multisource statistics, latent class modeling and structural equation modeling can be used in a situation where one has several datasets measuring the same target variable with measurement error. One can then see these error-prone measurements as observed indicators for an unobserved (latent) variable that represents the true values. Quality assessments based on these latent class models or structural equation models can then be used to assess the quality of output based on the observed indicators. If the model is trusted sufficiently, one could also correct output for measurement error using the predicted latent variable.

Example: Variance of estimates based on microdata reconciled by means of latent class analysis
Suppose that we have observed data on $S \geq 3$ categorical variables for the same units, where all variables are intended to measure the same categorical target variable. These multiple measurements could be obtained by linking data from different sources (e.g., administrative datasets, or a combination of administrative datasets and a survey) or by asking multiple questions about the same construct in a single survey. We use $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_S)'$ to denote the observed variables in general and $\boldsymbol{y} = (y_1, y_2, \ldots, y_S)'$ for a particular realization of values. The underlying target variable that these variables attempt to measure is denoted by X (with a particular value x) and is considered to be unobserved (latent) for all units. For simplicity we assume here that the set of categories is the same for all variables Y_j and X, denoted by $\{1, \ldots, L\}$.

Using some standard rules of probability theory, the marginal probability of observing $\boldsymbol{Y} = \boldsymbol{y}$ can be written as

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{L} \mathrm{P}(X = x, \boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{L} \mathrm{P}(X = x)\,\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} \mid X = x),$$

where $\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} \mid X = x)$ denotes the conditional probability of observing $\boldsymbol{Y} = \boldsymbol{y}$ when the true value of the target variable is x.

In latent class analysis, it is often assumed that each Y_j is measured independently of the other observed variables. Thus, the errors in different observed variables for the same unit are assumed to be independent. This assumption is known as "local independence" or "conditional independence". Under this assumption, we can write:

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y} \mid X = x) = \mathrm{P}(Y_1 = y_1 \mid X = x)\ \mathrm{P}(Y_2 = y_2 \mid X = x) \cdots \mathrm{P}(Y_S = y_S \mid X = x).$$

Applying the local independence assumption to the above expression for $\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y})$, we obtain the formula that describes the basic latent class model:

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{x=1}^{L} \mathrm{P}(X = x) \prod_{j=1}^{S} \mathrm{P}\left(Y_j = y_j \mid X = x\right).$$

To estimate the latent class model, we need to estimate the unknown probabilities on the right-hand side of this expression from the observed values for the probabilities on the left-hand side. This can be done, for instance, by maximum likelihood estimation for incomplete data (Hagenaars and McCutcheon, 2002; Biemer, 2011). Without additional assumptions, the latent class model is not identified with $S < 3$ observed variables.

Note that each factor $\mathrm{P}\left(Y_j = y_j \mid X = x\right)$ can be interpreted as a model for classification errors in one of the observed variables. Hence, estimates of these probabilities can provide information about the quality of each observed variable as an indicator for the true target variable X. For instance, the probability that a unit with true value $X = 1$ is misclassified on observed variable Y_j is given by $\mathrm{P}\left(Y_j \neq 1 \mid X = 1\right) = 1 - \mathrm{P}\left(Y_j = 1 \mid X = 1\right)$. In this way, a latent class model could provide input for the parametric bootstrap method discussed in Section 4.2.

The estimated model also provides predictions for the probability that a unit with a certain combination of observed values belongs to a particular category of the true target variable. These predictions can be obtained by the formula

$$\mathrm{P}(X = x \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\mathrm{P}(X = x) \prod_{j=1}^{S} \mathrm{P}\left(Y_j = y_j \mid X = x\right)}{\sum_{x'=1}^{L} \mathrm{P}(X = x') \prod_{j=1}^{S} \mathrm{P}\left(Y_j = y_j \mid X = x'\right)},$$

which follows from the previous expressions by Bayes' rule.

Boeschoten et al. (2017) proposed the MILC method to construct an error-corrected estimator based on the probabilities $\mathrm{P}(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$, along with a variance estimate for this estimator. The acronym MILC stands for the combination of multiple imputation (MI) and latent class (LC) analysis. An application of the MILC method consists of the following steps:

1. From the original dataset containing all observations of $\boldsymbol{Y}$, select M bootstrap samples.
2. For each bootstrap sample, estimate the latent class model. Denote the predicted probabilities $\mathrm{P}(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$ from the parameter estimates for the m^{th} bootstrap sample by $\widehat{\mathrm{P}}_m(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$.
3. In the original dataset, construct M multiply imputed versions of X, based on the predicted probabilities from the bootstrap samples. That is, create M empty variables $(W_1, \ldots, W_M)$ and impute variable W_m by drawing one category from $\{1, \ldots, L\}$ for each unit based on $\widehat{\mathrm{P}}_m(X = x \mid \boldsymbol{Y} = \boldsymbol{y})$.
4. Obtain M estimates for the parameter of interest based on the imputed variables $W_1, \ldots, W_M$.
5. Apply Rubin's rules for multiple imputation to pool the estimates from the previous step [see Rubin (1987)]. These pooling rules yield both a final estimate and an associated variance estimate. This variance estimate reflects the uncertainty about the true value of the parameter of interest due to classification errors in the observed variables and, if relevant, also due to sampling error.

The basic latent class model as described here can be extended in many ways. For instance, auxiliary variables can be added to the model if these are available. Depending on the available data, the local independence assumption can sometimes be relaxed to a certain extent. Boeschoten

et al. (2017) also discuss how to incorporate edit restrictions – e.g., that certain values for X are impossible given a particular value for an auxiliary variable – into the MILC method so that these are automatically satisfied by the imputed values.

So far, we have assumed that all variables are categorical and refer to the same point in time. A particular type of latent class model can be applied when multiple measurements are available over a period of time (e.g., each month or each quarter) and the latent variable can also change over time. This is known as a Hidden Markov Model (HMM). See, e.g., Pavlopoulos and Vermunt (2015) for an application of an HMM to model classification errors in linked data from two sources. Finally, when the observed and latent variables are numerical, structural equation models can provide a similar approach; see, e.g., Scholtus et al. (2015) and Oberski et al. (2017).

5.2.3 Statistical hypothesis testing

In statistical hypothesis testing, one uses observed data to determine the likelihood that a posited hypothesis holds true. In order to do so, one must formulate a null hypothesis and the alternative hypothesis, which says that the null hypothesis is not valid (in a particular way). One then computes how likely the observed data are, assuming the null hypothesis. The likelihood that the observed data were obtained as a realization of the null hypothesis is used as a measure for the validity of the null hypothesis and its alternative.

Hypothesis testing can, for instance, be used to test whether and to which extent the quality of revised estimates improves. In that case, the null hypothesis would be that there is no change in the quality of the estimates. One would then attempt to reject this hypothesis in favour of the alternative hypothesis that there is an improvement in the quality of the estimates. For instance, Fosen (2017) describes a test applied to revised employment statistics that are derived from gradually completing register data. As another example, suppose that a new estimation method has been proposed to replace an existing method, but it is not clear a priori whether the new method is an improvement. Here, the null hypothesis would again be that there is no difference in the quality of the estimates between the two methods, but now the alternative hypothesis may be that the quality with the new method either increases or decreases (i.e., a two-sided alternative).

5.2.4 Using estimated model parameters

Using estimated model parameters is a mix of descriptive summary statistics and statistical hypothesis testing. When using estimated model parameters for assessing quality, one uses a statistical model as in statistical hypothesis testing, but one does not use a posited statistical model to test a hypothesis. Instead, one directly uses estimated model parameters and calculates some descriptive summary statistics for them. For example, one can use a model to estimate the probability that a target variable in a certain unit has an incorrect value. One can then, for instance, take the average of these probabilities over all units in the dataset as an overall quality measure.

5.2.5 Small area estimation

Small area estimation is an umbrella term for several statistical techniques aiming to estimate parameters for small areas [see, e.g., Rao (2003) for an excellent introduction to small area estimation and descriptions of many small area estimators]. The main problem in small area estimation is usually the lack of observations for such small areas, which prevents one from using more standard estimation techniques, such as standard survey weighting [see, e.g., Särndal et al. (1992)].

In general, the term "small area" refers to a small geographical area such as a municipality, but it may also refer to other kinds of "small domains", such as small groups of individuals in the population.

6 Discussion

In this paper we have focused on the work with respect to the measurement of output quality that has been carried out in the Komuso project. In the Komuso project, we examined a large number of different basic data configurations, error types, and methods to assess output quality for some important quality dimensions (accuracy, timeliness, coherence, and relevance). We hope that the QMCMs that were produced in the Komuso project – of which we gave several examples in the current paper – are directly useful for many practical cases the readers of this paper are confronted with, and in other cases may form a source of inspiration to develop similar methods.

In the introduction to this paper, we mentioned that constructing quality measures for multisource statistics and calculating them is still more of an art than of a technical recipe. An illustration of this point is that, even with all the QMCMs and corresponding hands-on examples that have been developed in the Komuso project, one still needs to use one's instinct, or expert knowledge, to decide on what the most important error sources could be in a given situation. For instance, in some cases, measurement error may affect data quality the most, whereas in other cases sampling error or linkage error may affect data quality the most. Besides using one's instinct or expert knowledge, exploratory analyses and the use of scoring methods (see Section 3) may also provide some insight in the most important error sources. Something similar holds for the interaction between different kinds of errors. In many cases it may be reasonable to assume that the different kinds of errors are more or less independent, whereas in other cases this is definitely not the case. Again, using one's instinct or expert knowledge is important in order to distinguish between these situations. Although expert knowledge is not directly quantifiable, it can be valuable because it may capture years of experience. This may be illustrated by an anecdote about Pablo Picasso. When Picasso was asked by an admirer to scribble something on a napkin, Picasso complied and asked a large amount of money. The admirer was astonished by the large sum: "But you did that in thirty seconds." He replied: "No, it has taken me forty years to do that."

The QMCMs developed in the Komuso project provide quality measures and methods to calculate them for separate steps, or building blocks, in the statistical production process. We hope that in the, hopefully near, future, an all-encompassing theory or framework to base quality measures for multisource statistics upon will be developed. Such an all-encompassing theory or framework should be able to handle several different types of error sources at the same time and, preferably, use the same statistical theory to treat these error sources. Possible examples of approaches that may able to deal with several error sources simultaneously are, for instance, based on Bayesian techniques [see, e.g., Bryant and Graham (2015)] and over-imputation (Blackwell et al., 2015a,b).

We see two potential paths towards the development of such an all-encompassing theory or framework. The first potential path is the further development of Total Survey Error frameworks [see, e.g., Amaya et al. (2020), Biemer (2010), Biemer et al. (2014), Reid et al. (2017), Rocci et al. (2018), and Zhang (2012)], and the development of quality measures and methods to calculate them for the separate steps in such a framework.

The second – fundamentally different – potential path we see is not to specify and examine all the separate error sources, but develop a quality measure that covers several error sources at once. A landmark paper for such an approach is Meng (2018) [see also Rao (2020)]. Meng (2018) focuses

on the inclusion of units in the datasets under consideration, and hence on sampling error, inclusion error, non-response error, et cetera; measurement error and related error types are not considered. In the approach by Meng (2018) basically only three factors determine the quality of a certain estimated target parameter. In Meng's terminology these factors are referred to as "data quantity" (i.e., the amount of data), "problem difficulty" (i.e., the variation in the target variable), and "data quality" (i.e., the correlation between the target variable and possible inclusion in the data source).

Biemer and Amaya (2018) have extended the approach by Meng (2018) to include measurement error. For multisource statistics, it may be useful to further extend the approach to include linkage error.

A major challenge appears to be the application of Meng's approach in practical situations. In particular, the "data quality", and to a lesser extent the "problem difficulty", can be (very) hard to estimate in a practical situation.

Personally, we feel that this latter approach proposed by Meng (2018) may be the most promising path of the two paths towards an all-encompassing theory to base quality measures for multisource statistics upon as it avoids having to consider all possible error sources and their interactions. As noted above, despite the promising nature of Meng's approach, quite some research needs to be done before it can be applied in the day-to-day practice at, for instance, an NSI.

In any case for the next few years, we expect that measuring the output quality of multisource statistics will remain a field in motion, and a field that still is more an art than a technique.

Acknowledgments

We sincerely thank our colleagues in the ESSnet on Quality of Multisource Statistics. It has been a pleasure and a privilege for us to work with them in the ESSnet. Without the work done in that project we could not have written this paper.

We also thank a reviewer of our paper for carefully reading it.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Part of the work underlying this paper has been carried out as part of the ESSnet on Quality of Multisource Statistics, funded by the European Commission (FPA 07112.2015.003-2015.226: SGA 07112.2015.015-2015.705, SGA 07112.2016.019-2017.144 and SGA 07112.2018.007-2018.0444).

References

Agresti, A. (2013). *Categorical data analysis* (3 ed.). Hoboken, New Jersey: John Wiley & Sons.

Amaya, A., P. P. Biemer, and D. Kinyon (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology 8*, 89–119.

Asamer, E., F. Astleithner, P. Ćetković, S. Humer, M. Lenk, M. Moser, and H. Rechta (2016). Quality assessment for register-based statistics - results for the Austrian census. *Austrian Journal of Statistics 45*(2), 3–14.

Bakker, B. F. M. (2011). Micro-integration: State of the art. In *ESSnet on Data Integration, Report on WP1*, pp. 77–107. Available at http://ec.europa.eu/eurostat/cros/content/essnet-di-final-report-wp1_en.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer (2010). A quality framework for statistics based on administrative data sources using the example of the Austrian census 2011. *Austrian Journal of Statistics 39*(4), 299–308.

Berka, C., S. Humer, M. Lenk, M. Moser, H. Rechta, and E. Schwerer (2012). Combination of evidence from multiple administrative data sources: Quality assessment of the Austrian register-based census 2011. *Statistica Neerlandica 66*(1), 18–33.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly 74*, 817–848.

Biemer, P. P. (2011). *Latent class analysis of survey error*. Hoboken, New Jersey: John Wiley & Sons.

Biemer, P. P. and A. Amaya (2018). A total error framework for hybrid estimation. Paper presented at the BigSurv18 conference, Barcelona, Spain.

Biemer, P. P., D. Trewin, H. Bergdahl, and L. Japec (2014). A system for managing the quality of official statistics. *Journal of Official Statistics 30*(3), 381–415.

Bikker, R., J. Daalmans, and N. Mushkudiani (2011). Macro integration. Data reconciliation. Technical report, Statistical Methods (201104), Statistics Netherlands, The Hague. Available at https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/output/output/macro-integration-data-reconciliation.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Blackwell, M., J. Honaker, and G. King (2015a). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods & Research 46*(3), 342–369.

Blackwell, M., J. Honaker, and G. King (2015b). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research 46*(3), 303–341.

Boeschoten, L., D. Oberski, and T. de Waal (2017). Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *Journal of Official Statistics 33*, 921–962.

Bryant, J. R. and P. Graham (2015). A Bayesian approach to population estimation with administrative data. *Journal of Official Statistics 31*(3), 475–487.

De Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of statistical data editing and imputation*. Hoboken, New Jersey: John Wiley & Sons.

De Waal, T., A. van Delden, and S. Scholtus (2020). Multisource statistics: Basic situations and methods. *International Statistical Review 88*, 203–228.

Denton, F. T. (1971). Adjustment of monthly to quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association 66*(333), 99–102.

Di Fonzo, T. and M. Marini (2003). Benchmarking systems of seasonally adjusted time series according to Denton's movement preservation principle. Technical report, University of Padova. Available at http://www.oecd.org/dataoecd/59/19/21778574.pdf.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall/CRC.

Eurostat (2014). ESS handbook for quality reports, 2014 edition. Technical report, Eurostat. Available at http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf.

Fosen, J. (2017). Output quality for statistics based on several administrative sources. Technical report. Deliverable of WP 3 of the ESSnet on Quality of Multisource Statistics (SGA 1), available at https://ec.europa.eu/eurostat/cros/system/files/st2_7.pdf.

Gerritse, S., P. G. M. van der Heijden, and B. F. M. Bakker (2015). Sensitivity of population size estimation for violating parameter assumptions in log-linear models. *Journal of Official Statistics 31*(3), 357–379.

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2004). *Survey methodology*. New York: John Wiley & Sons.

Hagenaars, J. A. and A. L. McCutcheon (Eds.) (2002). *Applied latent class analysis*. New York: Cambridge University Press.

Houbiers, M. (2004). Towards a social statistical database and unified estimates at Statistics Netherlands. *Journal of Official Statistics 20*(1), 55–75.

Kloos, K., Q. Meertens, S. Scholtus, and J. Karch (2020). Comparing correction methods to reduce misclassication bias. Paper presented at the BNAIC/Benelearn conference, Leiden, The Netherlands.

Knottnerus, P. and C. van Duin (2006). Variances in repeated weighting with an application to the Dutch labour force survey. *Journal of Official Statistics 22*(3), 565–584.

Komuso (ESSnet Quality of Multisource Statistics) (2019). Quality guidelines for multisource statistics (QGMSS). Technical report. Available at https://ec.europa.eu/eurostat/cros/content/quality-guidelines-multisource-statistics-qgmss_en.

Krapavickaitė, D. and M. Šličkutė-Šeštokienė (2017). Effect of the frame under-coverage / over-coverage on the estimator of total and its accuracy measures in the business statistics. Technical report. Deliverable of WP 3 of the ESSnet on Quality of Multisource Statistics (SGA 1), available at https://ec.europa.eu/eurostat/cros/system/files/st2_5.pdf.

Kuijvenhoven, L. and S. Scholtus (2011). Bootstrapping combined estimators based on register and sample survey data. Technical report, discussion paper, Statistics Netherlands, The Hague. Available at http://www.cbs.nl/nl-nl/achtergrond/2011/39/bootstrapping-combined-estimator-based-on-register-and-sample-survey-data.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Meertens, Q. A., C. G. H. Diks, H. J. van den Herik, and F. W. Takes (2020). A data-driven supply-side approach for estimating cross-border internet purchases within the European Union. *Journal of the Royal Statistical Society Series A 183*(1), 61–90.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics 12*(2), 685–726.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association 112*, 1477–1489.

Pavlopoulos, D. and J. K Vermunt (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology 41*, 197–214.

Rao, J. N. K. (2003). *Small area estimation*. Hoboken, New Jersey: John Wiley & Sons.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B: The Indian Journal of Statistics*, In press.

Reid, G., F. Zabala, and A. Holmberg (2017). Extending TSE to administrative data: A quality framework and case studies from Stats NZ. *Journal of Official Statistics 33*(2), 477–511.

Rocci, F., R. Varriale, and O. Luzi (2018). A proposal of an evaluation framework for processes based on the use of administrative data. Paper presented at the UNECE Workshop on Statistical Data Editing, NeuchÃćtel, Switzerland. Available at https://www.unece.org/index.php?id=47802.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

Sax, C. and P. Steiner (2013). Temporal disaggregation of time series. *The R Journal 5*(2), 80–87.

Schnetzer, M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk, and M. Moser (2015). Quality assessment of imputations in administrative data. *Journal of Official Statistics 31*(2), 231–247.

Scholtus, S., B. F. M. Bakker, and A. van Delden (2015). Modelling measurement error to estimate bias in administrative and survey variables. Technical report, discussion paper, Statistics Netherlands, The Hague. Available at https://www.cbs.nl/nl-nl/achtergrond/2015/46/modelling-measurement-error-to-estimate-bias-in-administrative-and-survey-variables.

Scholtus, S. and J. Daalmans (2020). Variance estimation after mass imputation based on combined administrative and survey data. *Journal of Official Statistics*. Accepted for publication.

Sekar, C. C. and W. E. Deming (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association 44*(245), 101–115.

Stone, J. R. N., D. A. Champernowne, and J. E. Maede (1942). The precision of the national income accounting estimates. *Review of Economic Studies 9*, 111–125.

Van Delden, A., S. Scholtus, and J. Burger (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics 32*(3), 619–642.

Van der Heijden, P. G. M., J. Whittaker, M. J. L. F. Cruyff, B. F. M. Bakker, and H.N. van der Vliet (2012). People born in the middle east but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics 6*, 831–852.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wolter, K. M. (2007). *Introduction to variance estimation* (2 ed.). New York: Springer.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica 66*(1), 41–63.

Acknowledgement to Reviewers

The Editors of Spanish Journal of Statistics gratefully acknowledge the assistance of the following people, who reviewed the manuscripts of this volume.

Manuela Alcañiz, Universitat de Barcelona.
Subrata Chakraborty, Dibrugarh University, India.
Victoriano José García García, Universidad de Cádiz.
Victoria López, CUNEF Universidad.

Spanish Journal of Statistics

General Information

The Spanish Journal of Statistics (SJS) is the official journal of the National Statistics Institute of Spain (INE). The journal replaces Estadística Española, edited and published in Spanish by the INE for more than 60 years, which has long been highly influential in the Spanish scientific community. The journal seeks papers containing original theoretical contributions of direct or potential value in applications, but the practical implications of methodological aspects are also welcome. The levels of innovation and impact are crucial in the papers published in SJS.

SJS aims to publish original sound papers on either well-established or emerging areas in the scope of the journal. The objective of papers should be to contribute to the understanding of official statistics and statistical methodology and/or to develop and improve statistical methods; any mathematical theory should be directed towards these aims. Within these parameters, the kinds of contribution considered include:

- Official Statistics.
- Theory and methods.
- Computation and simulation studies that develop an original methodology.
- Critical evaluations and new applications
- Development, evaluation, review, and validation of statistical software and algorithms.
- Reviews of methodological techniques.
- Letters to the editor.

One volume is published annually in two issues, but special issues covering up-to-date challenging topics may occasionally be published.

Author guidelines

The Spanish Journal of Statistics publishes original papers in the theory and applications of statistics. A PDF electronic version of the manuscript should be submitted to José María Sarabia, Editor in chief of SJS via email to sjs@ine.es. Submissions will only be considered in English.

Manuscripts must be original contributions which are not under consideration for publication anywhere else. Its contents have been approved by all authors and. A single-blind refereeing system is used, so the identity of the referees is not communicated to the authors. Manuscripts that exceed 30 journal pages are unlikely to be considered for publication. More detailed information can be found at https://www.ine.es/sjs.

Published by the Spanish National Statistical Institute