UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

THE DISCRETE GENERALIZED BETA DISTRIBUTION: THREE STUDY CASES

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
OSCAR FONTANELLI ESPINOSA

TUTOR PRINCIPAL
PEDRO MIRAMONTES VIDAL (FACULTAD DE CIENCIAS)

MIEMBROS DEL COMITÉ TUTOR
GERMINAL COCHO GIL (INSTITUTO DE FÍSICA)
MANUEL FALCONI MAGAÑA (FACULTAD DE CIENCIAS)

CIUDAD UNIVERSITARIA, CD.MX. ENERO DE 2017

The Discrete Generalized Beta Distribution: Three Study Cases

Oscar Fontanelli

It a pleasurable duty to state my gratitude towards the people and the organizations who facilitated the fulfillment of this work. Prof. Germinal Cocho first came up with the DGBD function and assisted in the conception of the results in chapter 4 and in the difficult tasks of unraveling the mysterious mathematical nature of Nature. Prof. Wentian Li assisted, during very pleasant discussions, the characterization of the Lavalette family of distributions and the project of utilizing the DGBD to describe population distribution. I must acknowledge the absolute and unusual freedom I received from my main advisor, Prof. Pedro Miramontes, in choosing research problems of my own interest and I am thankful for the guidance and disposition to very open discussion during the whole duration of the project. I must cordially thank Prof. Manuel Falconi, who represented an extremely disinterested and careful assistance; I cannot remember one single meeting without a constructive comment from him. For the realization of this work financial support was granted by the CONACyT, Mexico, through grant 239957.

Oscar Fontanelli

Abstract

The Discrete Generalized Beta Distribution is a two-parameter rank-size function which successfully fits data sets that seem to follow a power-law but deviate from it at the tails. In this dissertation we present the results of three research paths that we conducted, all of them concerning this function. First, we utilized it to fit population data of world's administrative units and proposed a model to simulate the creation and evolution of such units; secondly, we calculated the probability density function related to this rank-size function, thus introducing a novel family of probability distributions, which we characterized and whose occurrence we illustrated; thirdly, we formulated a dynamic of random subtractions and proved that this collective phenomenon is a mechanism from which the behavior depicted by our rank-size function emerges.

Contents

Introduction

Cuando una serie de datos aleatorios tiene una cola pesada es habitual representarlos mediante una gráfica de rango-tamaño. A partir de esta representación se han construido algunas funciones de rango-tamaño, de las cuales la ley de potencias es la más popular. Cuando un fenómeno es regido por una ley de potencias se dice que éste es un fenómeno "libre de escala", pues la distribución de la variable de interés se ve igual en todas las escalas. Sin embargo, es muy común observar que las leyes de potencias dejan de cumplirse en el régimen de rangos pequeños. La causa de estas desviaciones es un tema de debate, pueden ser ocasionadas por efectos de tamaño finito o bien porque la ley de potencias no es un modelo adecuado para el fenómeno en cuestión; cada caso debe analizarse detalladamente.

Ha habido varias propuestas con respecto a la tesis de que la ley de potencia no es el modelo más apropiado y que se requiere una corrección en todo el cuerpo de la distribución. La mayoría de las propuestas sugieren funciones de rango-tamaño de dos parámetros, las cuales han tenido diferentes niveles de éxito. De todas ellas la Distribución Beta Discreta Generalizada (DGBD) es, discutiblemente, la más exitosa en términos de su bondad de ajuste y su numerosa cantidad de aplicaciones. Esta función rango-tamaño de dos parámetros fue propuesta originalmente para corregir el ajuste de la distribución del número de citas de diferentes revistas científicas. Sin embargo, su campo de aplicaciones creció rápidamente: se ha reportado en ciencias naturales y sociales, manifestaciones artísticas, fenómenos económicos y un largo etcétera.

Para poder afirmar la validez de un cierto modelo estadístico, en este caso una función de rango-tamaño, hay al menos dos condiciones que deben satisfacerse: lo primero es que debe haber evidencia estadística fuerte y suficiente que fundamente la relación funcional propuesta, lo cual en sí mismo no es algo trivial; en segundo lugar debe haber una teoría sólida que explique por qué debería observarse dicha relación funcional. Estos son dos puntos en los cuales debemos trabajar si queremos sostener que la DGBD o alguna otra ley es un buen modelo para describir una cierta clase de fenómenos. Podemos dividir los trabajos que hasta la fecha se han realizado con respecto a la DGBD en tres tipos: los que la utilizan para describir algún fenómeno particular, los que comparan su bondad de ajuste con otras funciones de rango-tamaño y los que buscan mecanismos de los cuales esta función emerja. Durante el transcurso de este posgrado desarrollamos diferentes líneas de investigación que tocan todos estos puntos.

Después de un breve repaso de lo que se sabe de la función DGBD, presento en esta tesis los resulta-

dos de tres trabajos relativos a esta función. Primero, extendí su campo de aplicaciones al utilizarla para ajustar la distribución poblacional de entidades administrativas en todo el mundo. Dentro de esta investigación propuse utilizar dos técnicas estadísticas que no se habían usado en los trabajos anteriormente mencionados: un enfoque de remuestreo a la prueba de Kolmogorov-Smirnov para medir la bondad de ajuste de la DGBD y el Criterio de Información de Akaike para comparar su ajuste con el de otros modelos. A través de estas técnicas concluimos que la DGBD es un modelo adecuado para este tipo de datos. De manera adicional, propusimos un proceso para simular la formación y el desarrollo de las unidades administrativas. Contamos con evidencia numérica que sugiere que la DGBD surge de este proceso. Presento estos desarrollos y resultados en el capítulo 2.

En el segundo trabajo obtuvimos la función de densidad de probabilidad (pdf) de una variable aleatoria cuya función rango-tamaño es una DGBD. Esto nos llevó a construir una nueva familia de distribuciones de probabilidad, la cual nombramos la *distribución de Lavalette*. Esta distribución es muy parecida a la lognormal en el centro del dominio, es muy difícil distinguirlas, pero la Lavalette tiene una cola mucho más pesada. Caracterizamos esta familia de distribuciones, discutimos la interpretación de sus parámetros, la existencia de sus momentos y damos algunos ejemplos para ilustrar su aplicabilidad en diferentes campos. Todo lo anterior conforma el capítulo 3.

En el capítulo 4 presento una investigación sobre un posible mecanismo detás del comportamiento que representa la DGBD. Un trabajo previo sugiere que la resta de dos variables aleatorias cuya función rango-tamaño es una DGBD es otra variable aleatoria cuya función rango-tamaño es también una DGBD. La prueba formal de esto no es evidente. Motivados por esta cuestión, construimos una amplia familia de funciones, la cual llamamos la familia $\mathcal{J}$, y probamos que es cerrada bajo una cierta operación de resta. Un caso específico de este resultado muestra que la distribución de Lavalette surge después de una resta de dos variables aleatorias de una cierta familia.

When random data presents a heavy tail, it is usual practice to represent them with the rank-size plot. Related to this representation people have constructed some rank-size functions, from which the power-law is by far the most popular. When a phenomenon is ruled by a power-law, it is said to be "scale free" because the distribution of the variable of interest looks the same at all scales. However, it is very common to observe that power-laws halt to hold at the small-size regime. Whether this happens because of finite-size effects or because the power-law is not a suitable model is a matter of debate and each case ought to be analyzed in detail.

Regarding the second possibility, the thesis that power-law is not the most appropriate model and the phenomenon asks for a correction in the whole body of the distribution, there have been some proposals. Most of them suggest two-parameter rank-size functions, which have had different degrees of success. Arguably the most successful one, in terms of its goodness of fit and the width of its application fields, is the Discrete Generalized Beta Distribution (DGBD) [64]. This two-parameter rank-size function was originally introduced to correct the distribution of journal citation rates, but its field of applicability grew very rapidly: it has been reported in natural and social sciences, artistic manifestations, economic phenomena and a long etcetera.

There are at least two necessary conditions before we can claim that a certain statistical model, in this case rank-size function, actually holds: first, there must be strong and enough statistical evidence, which is by itself a nontrivial matter, supporting the hypothetical functional relationship. On the other hand, there must exist a sound theory, a generative mechanism explaining why such a functional relationship should emerge in some particular phenomenon. If we are to claim that the DGBD or any other law holds in some certain kind of phenomena, these are two points on which we must work. We can divide the works done so far about the DGBD in three types: those that apply it for describing some particular phenomenon, those comparing its performance against other models and those looking for mechanisms form which it naturally arises. During the course of our PhD project we developed research paths regarding all these points.

In this thesis I present the results of three investigations concerning the DGBD. First, its field of applications was extended by successfully using it to describe the population distribution of world administrative units. As a part of this work, I proposed two statistical techniques to test the DGBD hypothesis, not previously used in other works regarding this function: a resampling approach to the Kolmogorov-Smirnov test to measure the goodness of the fit and the Akaike Information Criterion to compare its performance with other models. We concluded that DGBD is a suitable function for modeling this kind of data. Additionally, we proposed a computational process to simulate the development of administrative units. We have numerical evidence that suggests that DGBD arises from this process. These results are presented in chapter 2.

For the second inquiry I computed the probability distribution function (pdf) of a random variable whose rank-size function is a DGBD. This yielded a novel family of probability distribution, which we named *the Lavalette distribution function.* This distribution is very close to the lognormal distribution, in the sense that they are very difficult to distinguish from one another, but has a much

heavier tail. We characterize this new family of distributions, discuss the interpretation of its parameters, the existence of its moments, and provide some examples to illustrate its occurrence in real phenomena. This is the matter of chapter 3.

In chapter 4 I present an investigation on a possible mechanism behind the behavior that the DGBD depicts. There is numerical evidence suggesting that the subtraction of two independent random variables whose rank-size function is a DGBD is another random variable whose rank-size function is also a DGBD. The formal proof is not evident. Motivated by this initial question, we constructed a broad family of functions, we called it the $\mathcal{J}$ family, and proved that it is closed under a certain subtraction operation. We give a specific case of this wide result, in which the Lavalette distribution appears after the subtraction of two certain types of random variables.

Chapter 1

What Is Known: Theory and Applications of the DGBD

§1.1 Rank-size functions. The claim of ubiquity of power laws

There are at least two common ways in statistics to graphically present random data. Perhaps the most common one is the histogram, which resembles the probability density function (pdf) from which the data were originated. There is a less common representation, but still popular, which is the rank-size plot. In these kind of plot one takes a sample of N random observations $\{x_i\}_{i=1}^{N}$, arranges them in decreasing order, $x_{[1]} \geq x_{[2]} \geq ... \geq x_{[N]}$ and assigns a *rank* r to each observation: $r = 1$ to the largest one, $r = 2$ to the second largest and so forth. A plot of rank r against the value of the observation $x_{[r]}$ gives a discrete approximation to the reverse-order quantile function. A *rank-size function* gives the size of an observation x in terms of its rank r, $x = x(r)$. By construction, a rank-size function must be non-increasing. In general, the observed or measured quantity of interest is what we refer to as "size"; this measured quantity can be the population of a city, the magnitude of an earthquake, the size of solar flares, the income or wealth of an individual, the relative frequency in which a certain word appears, etc.

A particularly common rank-size function is the power-law, in which the quantity of interest x varies as an inverse power of the rank r, i.e. $x \sim \frac{1}{r^\alpha}$. In probability terms, a power-law states that the probability of making an observation larger than a certain value x is inversely proportional to a certain power of this value. Power-laws characterize phenomena that lack a characteristic size: the number of observations or realizations larger than λx is λ^α times the number of observations larger than x, independently of the value of x [46]. In this sense, a phenomenon governed by a power-law does not possess a characteristic or preferred scale in size or time. For this reason, these kind of phenomena are often described as "scale free". A power law also indicates that the probability distribution of the quantity of interest has a heavy tail, which means that rare events

are more likely to happen than they are in Gaussian, lognormal or exponential distributions [84].

There are indications about a certain ubiquity of power-laws in natural and social sciences. In statistical physics, for instance, these scaling laws arise in some exactly solved models near a critical state. Examples of this are the distribution of clusters in 2d Ising models at the critical state [11], distribution size of percolation clusters at the critical probability [85], avalanche sizes on a sandpile [9] and many others. There are other fields of knowledge where power-laws have been reported with different degrees of statistical evidence, even though they do not appear as an exact solution of certain model: in seismology, the Gutenberg-Richter law establishes that the distribution of earthquake intensities is a power-law [42]; in economics and finance, personal income and wealth [61, 62], size of firms [8] and the magnitude of market crashes [39, 34] are magnitudes power-law distributed; see [71] for an extensive list of examples in many different fields. A very famous example in the literature is the power-law distribution of word frequency in the English language, discovered by the American linguistic George Kingsley Zipf [91]. This behavior has also been observed in many other languages [90]. In fact, a power law with exponent $\alpha = 1$ is referred to as a *Zipf's law*. Of particular interest for our purposes is a purported power-law distribution of city populations [33, 81], first observed by the German physicist Felix Auerbach [6].

However, many reported power-laws do not hold well over the whole regime of observations and it is very common to observe deviations from this law, specially at the high rank regime (about this, see the extensive study made by Clauset et al in [20] and the criticisms to the over-usage of power-laws made in [86]). It is true that perfect power-laws are not expected to hold in real-life systems: in the examples previously mentioned of statistical physics, power-laws occur asymptotically on the limits of infinite energy or infinite size;consequently, when the system is finite, there must be a regime where the power-law does not longer hold and the system is dominated by finite-size effects. On the other hand, the probability density function associated with power-laws, the Pareto distribution [71], diverges at zero, so there must be a point where the power-law is stopped. Nevertheless, most studies in the literature, where data exhibit deviations from power-law at the high rank or small size regime, simply claim the phenomenon to be governed by a power-law, introduce a cut-off value and appeal the deviations to finite size effects, without further investigating the possibility that a correction is needed on the whole body of the distribution [46]. Also, the artificial cut off can have an effect on the estimated parameter of the distribution; we will address this issue on chapter 2.

A thumb rule to investigate the presence of a power-law is to do a log-log plot of rank against size of the observations. If there is a power-law, the data should fall into a straight line, as it happens in the word usage example, which we illustrate in fig. 1.1. Although some authors have used this as the solely indicator of this particular behavior, the truth is that it is by no means a strong statistical evidence and a further analysis must be made before a power-law phenomenon is established. However, this rule should serve to exclude many putative power-laws, but this seldom happens in practice. Deviations from power-laws in real data are ubiquitous and we believe they deserve a much more comprehensive study.

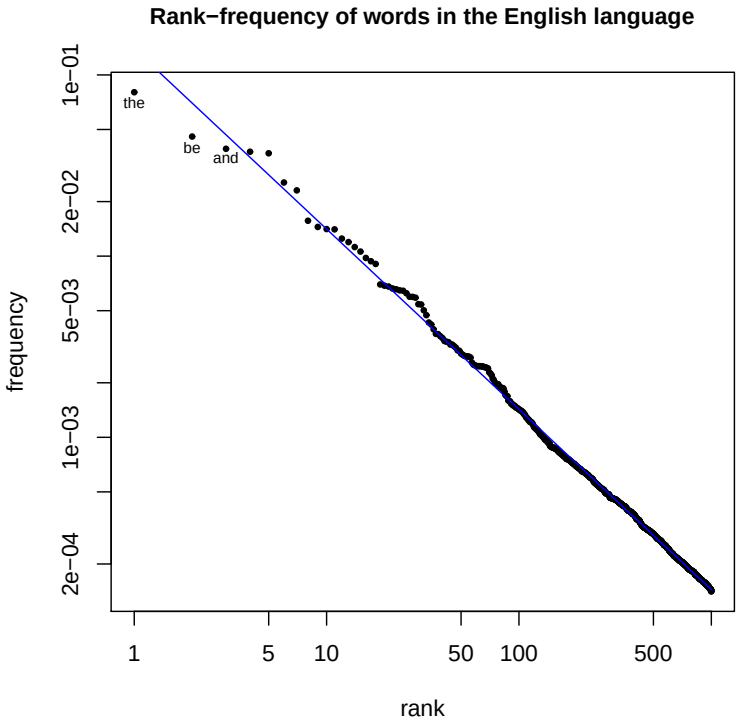


Figure 1.1: **Rank-frequency usage of words in the English language.** The most common word in English language is "the" ($r = 1$), followed by "be" ($r = 2$), "and" ($r = 3$) and so on. If we plot the rank against the relative frequency of the 1 000 most common words in logarithmic representation, the points nearly fall into a straight line. Data were taken from `http://www.wordfrequency.info`.

§1.2 The Discrete Generalized Beta Distribution

Some models have been proposed in order to improve the fitting of data that deviate from power-laws at the high rank regime, for instance the Zipf-Mandelbrot law [68] and the stretched exponential distribution [46]. The Discrete Generalized Beta Distribution (DGBD) was introduced in 2007 to correct deviations from power-laws in the rank-size distribution of journal impact factors [63]. The DGBD is the rank-size function

$$x(r) = A\frac{(N+1-r)^b}{r^a}. \tag{1.1}$$

Here a and b are parameters, N is the maximum rank (it is also the sample size if there are no rank ties) and A has been argued to be a normalization constant (see discussion in [58]). It was originally proposed as a generalization of the Lavalette rank function [47], which will be a matter

of our interest in chapter 3. Notice how the DGBD reduces to a power-law when $b = 0$. The parameter a controls the power-law regime and is related with the scale invariance and long-range correlations in the phenomenon, while b controls the deviation at the high-rank regime and is related with disorder and the absence of long-range correlations [5]. Because this must be by construction a non-increasing function, both parameters a and b must be non-negative. Fig. 1.2 displays various DGBD functions for different sets of parameters.

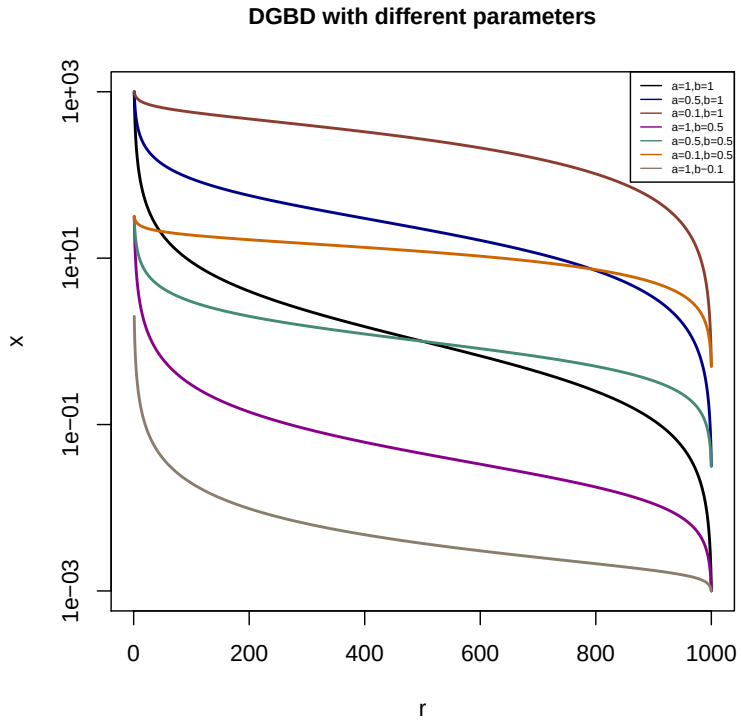


Figure 1.2: **The DGBD rank-size function.** Plots of the DGBD function for different values of its parameters. In all cases $A = 0.1$.

§1.3 Occurrence and applications of the DGBD

Originally, the DGBD was proposed to fit the rank-size order function of journal impact factor [63]. In that work, the authors studied impact factors of journals from a wide variety of scientific disciplines and concluded that the DGBD was a better model than the power-law, the Mandelbrot law and the Lavalette function (even though the latter one is a particular case of DGBD, they stand as different statistical models. This will be discussed in the next chapters). Regarding this

topic, Petersen et al. used the DGBD to describe rank-citation profile of a sample of renowned and assistant professors, concluding there are statistical regularities that could be used to evaluate career progress in academia [76, 77].

The DGBD function has found applications in linguistics. There is a study showing that this functions outperforms power-laws and several two-parameter functions in fitting the rank-frequency distribution of words in the novels *Moby Dick* and *El ingenioso hidalgo Don Quijote de la Mancha* [55], as well as in US and Mexican presidential speeches of the last few centuries [54]. Furthermore, DGBD is a good statistical model for the normalized number of character per syllable in the Chinese language, ranked from high to low [51, 52].

Concerning social phenomena, the number of adherents to the main religions worldwide during the 20th century seems to follow a DGBD [7]. This rank-size function has also been reported in artistic phenomena: Beltrán del Río et.al. successfully utilized the DGBD to fit the rank-frequency order of musical notes over an extensive sample of more than 1,800 musical compositions from different styles and epochs [26]. Finally, there is an extensive study made by Martínez-Mekler et.al. where the authors use the DGBD to fit the frequency of notes in classical pieces of music, sizes of geometric figures in abstract paintings, academic ranking of world universities, crashes of the US stock exchange, municipality and state population in Mexico, Spain and Chile, highway lengths and some other examples [64]. In this particular work, the authors fit the data to the DGBD via a linear regression of the logarithmic transformation of 1.1 and use the coefficient of determination to test the goodness of the fit. In all cases they claim a very high coefficient of determination.

All these examples show the usefulness of the DGBD to fit data. Some authors have seen in this a sign of a certain ubiquity of the DGBD (see for example [27]). However, we have not talked about possible mechanisms that could explain why this particular rank-size distribution should occur.

§1.4 Possible mechanisms behind the tail-decay behavior

What is the reason behind the apparent ubiquity of the DGBD function? Up to this date there are, to our knowledge, four research paths looking for possible answers:

Martínez-Mekler et.al. were the first ones to come up with a possible explanation [64]. They proposed an expansion-modification model which incorporates the basic mechanisms of neutral evolution in a genome sequence: duplications and point mutations. In this process, originally proposed by W. Li in [48], one starts with a chain of variables that can take two values, 0 and 1. Each variable gets transformed in the following way: with probability p it duplicates, $0 \longrightarrow 00$ and $1 \longrightarrow 11$, and with probability $1-p$ it changes its value, $0 \longrightarrow 1$ and $1 \longrightarrow 0$. This generates a sequence of zeros and ones where the length of non-overlapping groups of consecutive elements

has a rank-frequency distribution numerically well fitted by DGBD.

Until today, the only analytic derivation of the DGBD was provided by Naumis et. al. in [70]. The authors start with s random numbers $p_1 \geq ... \geq p_s$, chose an integer N and form the quantities $p_1^{n_1} p_2^{n_2} ... p_s^{n_s}$ with $n_1 + n_2 + ... + n_s = N$. The observed value numbers are arranged in decreasing order. They proved that on the limit of $s \longrightarrow \infty$ (infinitely many random variables) the rank-size function of the resulting numbers is a DGBD. If there is a variable X depending on the quantities $p_1^{n_1} p_2^{n_2} ... p_s^{n_s}$, the first-order approximation of its rank-size function is the DGBD.

A third attempt to explain the purported universality of the DGBD was made by a group under G. Cocho in [25]. Here, the authors propose the following random process: start with N *iid* (independent and identically distributed) random variables X_i and construct the variables $\eta_1 = X_1, \eta_{i+1} = X_{i+1} - \eta_i$. They give numerical evidence that on the limit of $N \longrightarrow \infty$ the conditional distribution $f(\eta_n | \eta_n > 0)$ is such that its related rank-size function is a DGBD.

Finally, there is a model that we will propose in chapter 4 of this thesis. Briefly summarizing it, we will address in an analytic way the stability problem under the subtraction of random variables proposed by G. Cocho. We will construct a very broad family of solutions to this problem and show that DGBD arises as a particular case of it.

§1.5 Perspectives of the studies about the DGBD

The main usage so far of the DGBD is to correct deviations from power-laws at the high rank regime. How do we know that such corrections are needed, when the observed deviations could be the result of finite-size effects? First, these purported finite size effects are seldom studied and justified in actual investigations. Given this situation, we raise the next question: how many of the reported power-laws with deviations at the tail are true power-laws with true finite-size effects? According to the critical essay of Stumpf [86] and the extensive analysis by Clauset [20], many reported power-laws, and some times well accepted, fail the most simple statistical tests and lack a theoretical mechanism explaining why those phenomena should exhibit a scale free behavior; according to these authors, power laws are generally non-justified.

Of course there must be deviations from perfect power-laws because every transformation leading a system to an auto-similar behavior can only happen a finite number of times and because exact power-laws in physical systems only occur asymptotically. As a consequence, a power-law must cease to hold at a certain regime in every finite system. However, our work is *not* about these inevitable errors but explores the thesis that, in many cases, deviations happen at the whole body of the distribution and require a deeper understanding and a whole new characterization of the rank-size function. In this work we provide mathematical and statistical analysis of this

idea and follow the works of those who proposed the DGBD function to correct these distributions.

Why do we propose the DGBD instead of a different one, two or more parameter function? Because there are works showing it outperforms other possibilities. It is true that DGBD is, by its own definition, a very flexible function, capable of fitting many data sets solely by its plasticity. If we are only looking for adequate statistical models this is not an issue, DGBD serves to model the body and the tail of the distribution, common and extreme events alike. If we were asking what is the reason or reasons behind its ubiquity, maybe we should claim that it is not ubiquitous *per se* but it is useful to model a very extended phenomenon: a power-law like behavior over a regime followed by a decay at the tail. What we have intended to do in this work is, briefly summarized, an extension of the applications of the DGBD as a statistical model and a search for mechanisms behind the ubiquitous behavior it reproduces.

Finally, we believe that the goodness of fit methods used so far to test the fittings of the DGBD to real data could be improved. In the current work we will use more sophisticated methods to test the DGBD hypothesis and to compare its performance against other plausible models. Good statistical support and a sound theory to understand it are necessary elements to claim a DGBD to hold; this requires still much more research and we hope we have achieved a few steps of progress towards this direction.

Chapter 2

A New Application: Population Distribution in World Administrative Units

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?"

George Box

In this chapter we use the DGBD function to fit population data of world administrative units. It is a commonly accepted fact that city population follows a power-law, at least in the large population regime [33, 83, 43] . However, very little has been said about population distribution across administrative units (for instance, states and counties in the US, provinces and prefectures in China or regions, departments and arrondissements in France), which are different objects and may follow different models than those used for cities [56, 59, 41, 35, 19]. In order to investigate this seldom studied topic, we fitted DGBD and power-laws to administrative unit population data for 150 countries. We will see that DGBD is a good model for this kind of data in a high number of our samples and a better model than power-law in almost every case. We also present a computational process, which we call the split-merge process, to simulate the formation and evolution of administrative divisions; we will also give numerical evidence that the DGBD arises from this process. The split-merge process together with the DGBD function prove adequate to describe and reproduce administrative unit formation and quantify its effects on population distribution. The structure of this chapter is the following: we begin by defining the rank-size representation, which we use to derive some relationship the DGBD function has with some common probability laws. Then we will discuss our procedures to test the DGBD model and compare its performance with that of the frequently used power-law. After that, we arrive to the core of this chapter: testing the DGBD function for fitting administrative population data and compare it with power-law; we use the fitted parameters of the DGBD to construct a phenomenological characterization of our

sample of countries according to their internal population distribution. Finally, we present the split-merge process for simulating the creation and development of this kind of units.

All these results are presented in a manuscript whose preprint version is available in [31]. The following exposition borrows significantly from this manuscript.

§2.1 The rank-size representation and the DGBD

It is said that a random phenomenon presents a heavy tail when there is a relatively high probability for large, rare events to happen. When this happens, it is customary to describe it via the rank-size or rank-frequency distribution, instead of the probability density function (pdf) [84]. This is often the case for population distributions where a small amount of high populated regions encompass the majority of the population. Thus, we will adopt the rank-size representation to describe the population distribution within a country. This representation does not require binning, avoiding the issue that bins are often under-sampled at the tail. Intuitively, the rank of an observation is high if the probability of making a larger observation is also high and viceversa; following this idea, Sornette takes N sorted observations $x_1 \geq x_2 \geq ... \geq x_n \geq ... \geq x_N$ of a random variable X and defines the rank n of the observation x_n as $n = NP_{>f}(x_n)$, where $P_>(x_n) = \int_x^\infty f(t)dt$ ($f(x)$ is the pdf of X [84]. Following this reasoning, we define the rank of an observation in the following way:

Definition 1. *Let X be a continuous random variable with density $f(x)$ with support in $[x_m, x_M]$ (possibly $x_m \to -\infty$ or $x_M \to +\infty$) and $\{x_{[i]}\}_{i=1}^N$ a collection of N realizations. The* rank *of the observation $x_{[i]}$, denoted by $r(x_{[i]})$ is given by*

$$r(x_{[i]}) = r_m + (r_M - r_m) \int_{x_{[i]}}^{x_M} f(t)dt, \tag{2.1}$$

where r_m and r_M are the minimum and maximum ranks respectively.

In this way, the integer part of $r(x_{[i]})$ is the expected number of values greater or equal than $x_{[i]}$. Note that for a given value $x_{[i]}$, its rank i may have fluctuations in real data (in some samples it can be the largest one, in other samples the second largest one and so on), which this definition does not take into account. This definition has the advantage of determining the expected value of the i-th observation $x_{[i]}$ if the pdf $f(x)$ is known, or computing $f(x_{[i]})$ after measuring $x_{[i]}$. Notice that $r_m = r(x_M)$ and $r_M = r(x_m)$. The values of r_m and r_M are chosen arbitrarily, usually $r_m = 1$ and $r_M = N$. A *rank-size function* is a function that gives the size of an observation $x_{[r]}$ in terms of its rank r, *i.e.* $x = x(r)$. Sometimes the measured quantity is a relative frequency (for example, in qualitative linguistics, where the relative frequency of word appearance is a matter of interest); for this reason, this function is sometimes called a *rank-frequency function*. Now we define the Discrete Generalized Beta Distribution,

Definition 2. *We say that a random variable X follows a Discrete Generalized Beta Distribution, abbreviated DGBD, if it has the rank-size function*

$$x(r) = \frac{C(r_M + r_m - r)^b}{r^a}, \tag{2.2}$$

where b and a are parameters to be estimated and C is a normalization constant.

The way in which the DGBD is defined makes it a very malleable function, capable of well representing data created by different mechanisms. In fact, it reduces exactly to at least three common probability laws: the power-law, the uniform distribution and the delta distribution.

Proposition 1. *The following properties hold:*

1. *A power-law is represented by a DGBD with $b = 0$. If additionally $a = 1$, it is Zipf's law.*
2. *A uniform distribution is represented by a DGBD with $b = 1$ and $a = 0$ on the limit when the number of observations is very large.*
3. *A point located distribution is represented by a DGBD with $b = a = 0$.*

Proof. Point 1) follows direct after the definition of power-law, which is a rank-size function of the form $x = \frac{A}{r^\alpha}$. Next, suppose X is a random variable uniformly distributed over the interval (α, β). This means that X has the pdf $f_X(x) = \frac{1}{\beta-\alpha} 1_{(\alpha,\beta)}$. Thus, for x in (α, β), eq.(2.1) implies that

$$r = 1 + (r_M - 1)\frac{\beta - x}{\beta - \alpha}.$$

Solving this last equation for x we get after some arrangements

$$x = \frac{\beta - \alpha}{r_M - 1}\left(\frac{(r_M - 1)\beta}{\beta - \alpha} + 1 - r\right),$$

which reduces to DGBD with $b = 1$ and $a = 0$ by taking $\beta = \frac{r_M}{r_M - 1}$ and $\alpha = \frac{1}{r_M - 1}$ on the limit when the maximum rank (consequently the number of observations) r_M tends to infinity. Lastly, we observe that for a zero variance distribution, located at x_0 and with a delta pdf function $f_X(x) = \delta(x - x_0)$, eq.(2.1) reduces to $r = r_M$, which is a DGBD with $b = a = 0$ and constant of normalization $C = r_M$. □

Thus, DGBD can be used to represent the population distribution in a completely "flat" country, where all cities or administrative units have the same population ($b = a = 0$) or, at the other extreme, in a completely disordered country where population is distributed fully at random (uniform distribution, $b = 1$ and $a = 0$).

A consequence of our definition of rank 2.1 is that there is a one-to-one relationship between the rank-size and the pdf representation. This means that for every rank-size function there is one

and only one equivalent pdf. However, we cannot in general give a closed form of the pdf related to the DGBD. We will deepen on this subject on chapter 3. Anyway, there are a few cases where a pdf can be derived analytically from the DGBD. One such case is when $a = b$, in which case DGBD is called the Lavalette rank function. When this happens, there is a closed form of the pdf, yielding a probability distribution which closely resembles the lognormal at the center of the distribution, but has fatter tails. This will be the main topic of chapter 3.

§2.2 Testing the DGBD model

In order to test the goodness of fit of the DGBD as a statistical model for a sample of observations, we proceed in the following way: first we discard every sample with less than 10 observations; we do this because we will perform a regression analysis with two independent variables and we need a minimum number of observations to perform this analysis. Next we proceed as in the literature about DGBD: we take the logarithmic transformation of 2.2, $\log x = \log C + b \log(r_M + r_m - r) - a \log(r)$ and perform a linear regression in order to estimate the parameters b and a. A high coefficient of determination R^2 is a first indication that DGBD may be a good model. However, the logarithmic transformation produces an underestimation of the error at the high-rank regime. As a consequence, having a coefficient of determination close to one is by no means enough statistical evidence to claim that DGBD is a good model. Nevertheless, it serves to rule out purported DGBD's.

We propose a resampling approach to the Kolmogorov-Smirnov test, originally proposed in [20]. First we compute the Kolmogorov-Smirnov or K-S statistic, which measures the maximum vertical distance between the empirical and the theoretical cumulative distribution functions, supposing DGBD holds. This gives us a measure of the distance between the actual data and the reference distribution. Next we simulate a large number of DGBD data sets with the fitted parameters and compute the K-S statistic of each simulated sample. We compute the fraction of simulated samples that are farther from DGBD than our population data; this fraction gives an estimate of the p-value of the DGBD hypothesis. A large p-value means random choices mostly lead to a worse fitting, so our fitted function is not bad enough to be rejected. On the other hand, a small p-value implies that our fitting performance is on the worse end among random chances, thus not good enough. With this method, the specific value of the estimation depends on some specific choices, for example how to manage samples with the same K-S distance, the number of replicates, etc. Nevertheless, a large p-value still indicates the fitted model can not be rejected, after the removal of samples with a very low number of observations.

We take the countries for which the p-value is larger than 0.05 (not enough evidence to reject DGBD). We would like to compare the performance of the DGBD in relation with other possible models. Since we will be testing population data, we want to compare our model with the power-law, which is the traditional model for city population. We recall that the DGBD reduces to

power-law when $b = 0$, but they are different models. For instance, DGBD has two adjustable parameters, while power-law only has one. If we had a sample which is a true power-law, both models would be correct, but the power-law ought to be preferred because the lower number of parameters. To compare these two models we use the Akaike information criterion, which measures the relative quality of a statistical model [17, 3, 49]. The model that exhibits a lower $AIC = 2k + N \log\left(\frac{RSS}{N}\right)$ is the better model (k is the number of estimated parameters, N the sample size and RSS the residual sum of squares). We take only those samples where AIC_{DGBD} is less than $AIC_{\text{power law}}$.

In addition to all this, we re-performed all our analysis but estimating the parameters of the DGBD via a nonlinear regression. The reason why we don't use the traditional Maximum Likelihood Estimator (MLE) methodology is that we don't know the pdf of the DGBD except for a few cases, thus we don't have an analytic expression of the likelihood function. However, for the cases where the MLE can be computed they coincide with the results from the regression analysis. The main results and the core of our conclusion hold; we notice that with the nonlinear regression methodology the DGBD model is rejected for a fewer amount of countries. All these results and details of the method are shown in the second appendix to this chapter.

§2.3 DGBD and administrative unit population

While city population in different countries often exhibit power law or lognormal behaviors, when artificial divisions come into consideration it is unclear what the distribution should be. To introduce the subject, we present on Fig.(2.1) several examples of ranked population of administrative units. First there is population of all countries and territories of the world, which we can think as zero-level administrative units (data from the UN, `https://esa.un.org/unpd/wpp/` and from the World Bank, `http://data.worldbank.org/indicator/SP.POP.TOTL`) ; as a matter of example of primary or first level administrative units (PAU from now on) we present ranked population of these units for five different countries; because number of PAU's is low for many countries, we will focus most of our analysis on secondary administrative units. The third panel shows ranked population for provinces, prefectures (primary and secondary administrative units respectively) and cities with more than 750,000 inhabitants in China. Even though cities and administrative units are different objects, in China there is a certain correspondence between them: many administrative units are conformed by a central city plus its neighboring countryside (ref?) and while some cities form PAUs (the four largest ones), some others constitute SAUs. These data are from the 2010 Population Census and were downloaded from `https://www.citypopulation.de/China.html`. Finally, we show ranked population of districts (SAUs) and agglomerations and cities in India; data are from the 2011 Census of India and were taken from `https://www.citypopulation.de/India.html`. Normally, power laws are expected when the rank-size or rank-frequency plot is approximately a straight line in the log-log representation. Although this is by no means a statistical evidence to claim that a data set or phenomenon is a power law, it is a necessary condition that potential power law candidates must fulfill. Deviations from power law behavior can

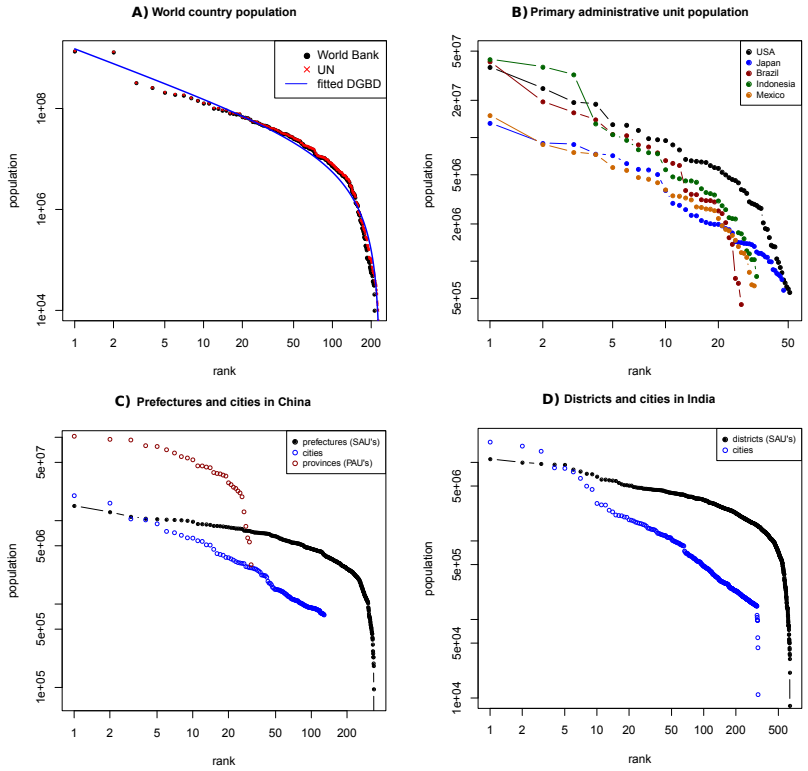


Figure 2.1: **Examples of deviations from power laws**. A) Ranked population of countries and territories of the world according to data of the World Bank and the UN, red line is the fitted DGBD to the UN data; B) ranked population of primary administrative units in five different countries (states in the US, Brazil and Mexico, prefectures in Japan and provinces in Indonesia); C) ranked population of prefectures, cities and provinces in China; D) ranked population of districts and cities in India. All plots are in log-log scale.

be appreciated in Fig.(2.1). For instance, there is a breakdown in the distribution tail of country population, yet DGBD provides a good fit at both tails, as the blue line shows; these deviations are more difficult to appreciate in PAU population, but are more evident for SAU's, as panels C) and D) show. Regarding city population, we see in our examples that deviations from power law appear in the low-population regime, but it is a good model for the largest cities.

We obtained SAU population data from the database Statoids (`http://www.statoids.com`, last consulted on January 2017, detailed information on sources and dates of each data set can be found in this site), which gave us the SAU population for 150 countries. We chose this as our global source because all data sets in this site come from official sources, which are not always accessible in a more direct way. While there are a few cases in which data is more than 10 years old, for most of the countries the information is from 2010 or later. Despite the fact that data within each country or territory was collected by its corresponding census or statistics office at different years, we still get a very general picture of world population and its distribution at

the current time. To check the quality and reliability of data, we compared our global source with three official data sources that were available to us: the 2011 census of Spain (`http://www.ine.es/inebaseDYN/cp30321/cp_inicio.htm`), the 2010 census of the US (`http://www.census.gov/topics/population.html`) and the 2010 census of Mexico (`http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2010/Default.aspx`). For the case of Spain and the US, data sets from Statoids and the official sources are equal; in the case of Mexico, data from Statoids are from the 2005 National Population Count (see `http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2005/`) and it is not updated. We performed our analyses on both data sets, the Statoids and the 2010 Census official source, and noticed a variation of 0.8% and 2.0% on the results (estimated b and a parameters of the fitted DGBD distribution), so the main results and conclusion of our research should remain the same. These comparisons between our global source and the available official ones indicate that our main conclusions would hold if we used only official data. For each country we ordered its SAU population by rank and fitted the data to a DGBD via a linear regression of the logarithmic transform of eq.(2.2). As we show in the first appendix of this chapter, the coefficients of determination R^2 are invariably high, which is a first sign that DGBD may be indeed a good fit. Then we performed the goodness of fit analysis depicted in the previous section: we discarded all countries where the number of SAU's is less than ten; then, we implemented the bootstrap version of the K-S test and picked those countries where the p-value is higher than 0.05, indicating that the DGBD hypothesis is not rejected. Lastly, we calculated the Akaike information criterion (AIC) for the DGBD and the power-law models, picking only those countries where $AIC_{\text{DGBD}} < AIC_{\text{power law}}$. Table (2.1) in the Appendix displays all these results, together with the fitted a and b parameters of the DGBD, which we estimated via the linear regression analysis. There are countries where DGBD performs better than power-law according to AIC, but it is rejected as a consequence of a small p-value. For our study, we looked for countries that satisfied both criteria. Ranked population and fitted DGBD for each country that we studied can be seen in fig.(2.7) in the first appendix.

After these three selection criteria a set of 108 countries remains. Mean and standard deviation of the sample size (number of SAU's) are 176.9 and 251.2 respectively. The mean and standard deviation for the fitted b parameter are 0.58 and 0.35, while the mean and standard deviation for a parameter are 0.53 and 0.25. We show in fig.(2.2) histograms of b and a, which indicate that they are not randomly distributed, but rather clustered around central values. There are two cases, Slovenia and Virgin Island US, for which $a < 0$. When this happens, eq.(2.2) fails to represent a rank-size distribution because it is not monotonous; however, the fittings are still good. Usually, this means that the maximum of the fitted curved is reached below $rank = 1$ [26].

The fitted DGBD b and a parameters for these countries are shown in the scatter plot of fig.(2.3). Countries are indicated by their three letter ISO3 code. The gray, green and red dots represent idealized regions following perfect Zipf's law, delta and uniform distributions respectively. We indicate with purple the vertical line $b = 0$, representing perfect power laws and with blue the line $b = a$, representing the Lavalette distribution which, as we mentioned, has a close resemblance with the lognormal distribution. There has been debate about whether lognormal distribution is a

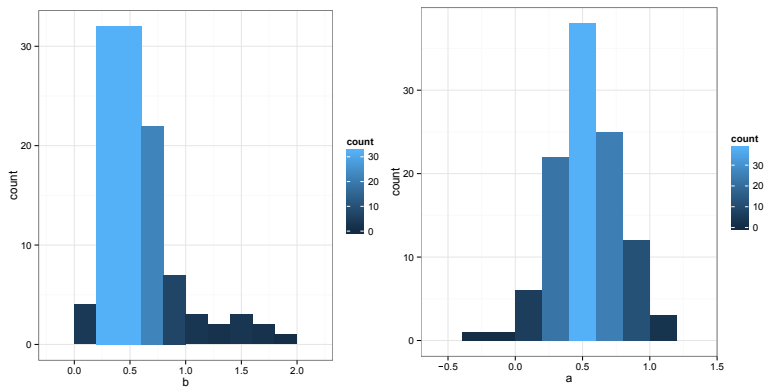


Figure 2.2: **Histograms of fitted b and a for the 108 selected countries**. The parameters b and a were estimated by a linear regression of the logarithm of the DGBD.

better representation for city population than power laws. It might be said that for SAU population, countries with $b < a$ (between the blue and the purple lines on the diagram) are somewhere between these two distributions, $b << a$ indicates that power law fits the data well while countries below the blue line ask for a different model.

The scatter diagram allows a quantitative measure of the distance between the SAU population distribution within a country and an idealized Zipf's law, a completely ordered or disordered population distribution, etc. We propose the euclidean distance on the $< b, a >$ plane to measure this. For example, Myanmar and Mauritania are close to being disordered (uniform distribution) while Burundi and Nigeria are close to being delta distributed (and also close to Lavalette, implying that they have very narrow lognormal-like distributions). These observations are confirmed by observing the histogram densities of Burundi and Mauritania, shown in fig.(2.3). Indeed, we see that Mauritania has a wide pdf, somewhat close to the constant pdf of a uniform random variable, while Burundi exhibits a taller and narrower pdf.

Even if the internal partition of a country is determined by a central administration, the SAU system might not be completely arbitrary, since there may be climatic and geographical factors constricting internal divisions and subdivisions. To investigate this issue we present in fig.(2.4) world maps with the euclidean distance of each country to a Zipf's law and to a country with a uniform probability distribution. Countries and regions shown in white are those for which DGBD is not a good statistical model according to our tests or for which no data was available. In these maps we can see that there is indeed a certain correlation between geographical position and location on the $< b, a >$ plane. For example, countries in East Asia are in general far from Zipf and close to disorder, while countries in South America tend to be closer to ideal Zipf's law.

The Euclidean distance on the $< b, a >$ plane also provides us a tool for quantifying the proximity of the internal population distributions within the SAU's of any given countries or regions. We

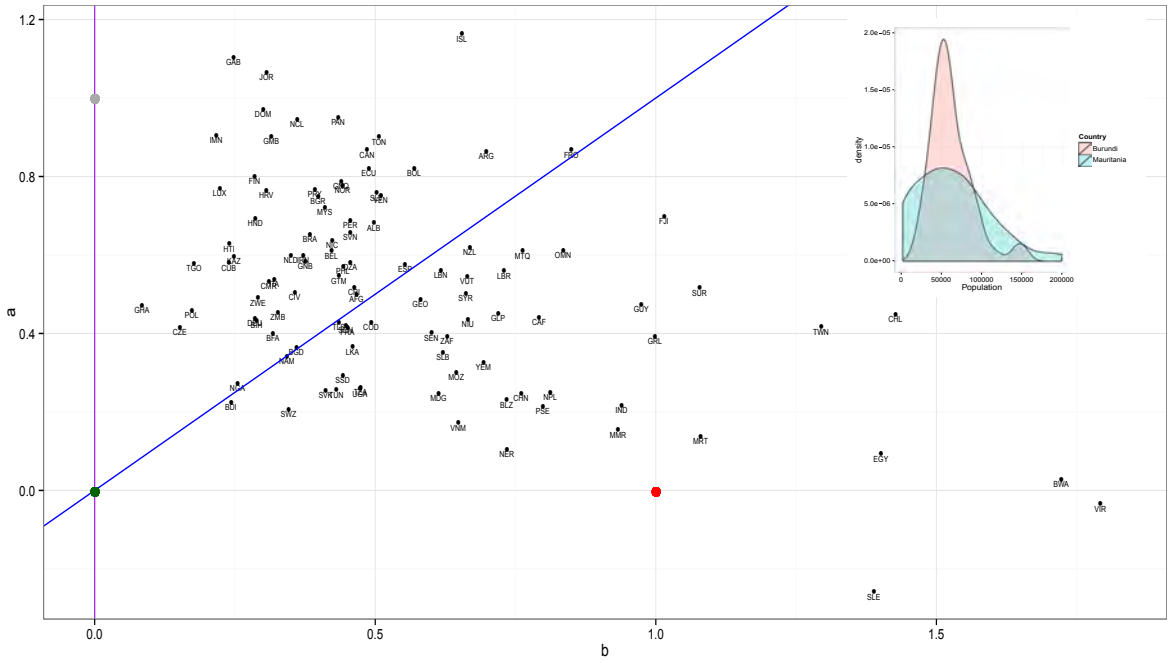


Figure 2.3: **Scatter plot of fitted b and a for the 108 selected countries.** Each country is indicated by its three letter ISO3 code. In purple the line $b = 0$, which represents a perfect power law; in blue the line $b = a$, which is the Lavalette rank-size function; gray point is $(0, 1)$ a pure Zipf's law; green is $(0, 0)$, representing a point located probability distribution; red point is $(1, 0)$, probability uniform distribution. We also show the density histogram for two countries: Mauritania, whose parameters indicate a closeness to a uniform distribution, and Burundi, whose parameters suggest a distribution similar to a very narrow lognormal.

use this distance to measure the similarity of countries according to the internal arrangement of its population. By using the average linkage clustering method (the distance between two clusters is the average of all distances between pairs of objects inside them), we get a measure of the cross-correlation between population distributions. Countries exhibiting high cross-correlations are countries with very similar patterns of where people live. Countries in which people spread similarly may have common challenges and perspectives concerning urban planning, population control, sustainability, etc [14]. See fig.(2.4) for a taxonomic tree of the 108 countries exhibiting good DGBD by this measure. This tree can be divided for any level of correlations into separated families of population-pattern-like countries. By taking the 0.5 decorrelation point as a cut off, we see that five families of kindred countries arise. Furthermore, closeness of countries in this parameter space shows that they have similar administrative division systems, pointing to akin ways in which these countries divide their land.

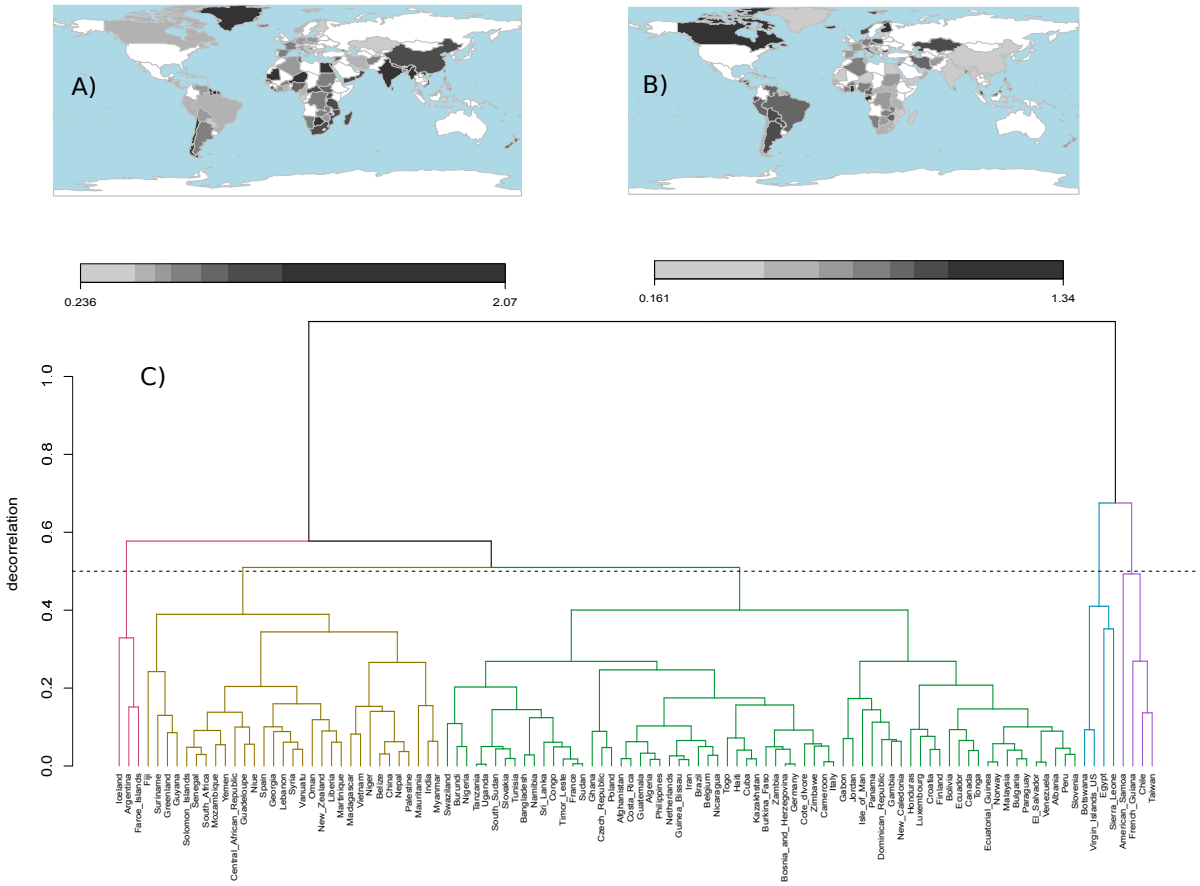


Figure 2.4: **Countries worldwide by their internal population distribution.** A) World map indicating distance in the parameter space to a country with a uniform distribution, B) distance to a perfect Zipf's law, C) dendrogram showing five families of kindred countries according to its internal population patterns.

§2.4 The split-merge process

Now we propose a computational process which is useful to simulate the formation of administrative units. Originally, we introduced this process as a mechanism to break power-law distributions in genetic data [53]. Suppose we look at a delimited territory, a country for example, where population is spread in communities, cities, towns, villages, etc. The population distribution of these aggregations may be well described by a Pareto, a lognormal distribution or some other function, depending on the specific processes that drove the population growth and dispersion in the region. What happens when a politician decides to create artificial boundaries, dividing the territory in well separated administrative units? We call this units "municipalities" during the following discussion. Certainly we are now dealing with a different kind of object and we do not know a priory if the distributions of population in towns and municipalities will be the same. It could happen that a large metropolitan area is disaggregated into several municipalities, or that two different towns are grouped together in a common municipality. We simulated computationally one particular option for this mechanism by means of what we call the *split-merge process*:

1. Start with a sample $\bar{X}_0$ of N_0 observations following some initial probability distribution f_0. These observations represent populations of N_0 human agglomerates; with these observations we create a one-dimensional array, such that every agglomerate has two neighbors, except for the first and the last one that are at the border and have only one neighbor.

2. Pick the two largest values $X^{(1)}$ and $X^{(2)}$ of the sample and split each of them into two new values, $X^{(1)} \longrightarrow p_1 X^{(1)}$, $(1-p_1)X^{(1)}$ and $X^{(2)} \longrightarrow p_2 X^{(2)}$, $(1-p_2)X^{(2)}$ where p_1 and p_2 are random numbers on the interval $(0,1)$.

3. Randomly choose 3% of the remaining observations and pair them with their neighbors, for example, pick X_i and replace them with the single value $X_{i-1} + X_i + X_{i-1}$.

4. With the merged and the split values and the remaining observations construct the new sample $\bar{X}_1$ of size N_1 following distribution f_1.

5. Repeat steps 2,3 and 4 for n iterations.

Step 2 simulates the process in which two large cities are split into two municipalities each; depending on the probability distribution from which p_1 and p_2 are sampled we can simulate different ways to split a city: for instance, if p_1 and p_2 follow a random uniform distribution, cities are divided in an entirely arbitrary manner, yielding to the creation of many low-populated municipalities; on the other hand, if p_1 and p_2 follow a distribution with a peak at the center (for example, a symmetric beta distribution), there is a higher probability for the cities or municipalities to be split in more or less equal sized units, thus simulating the division of a high populated area into different units with substantially less population each. This is typically what happens when authorities

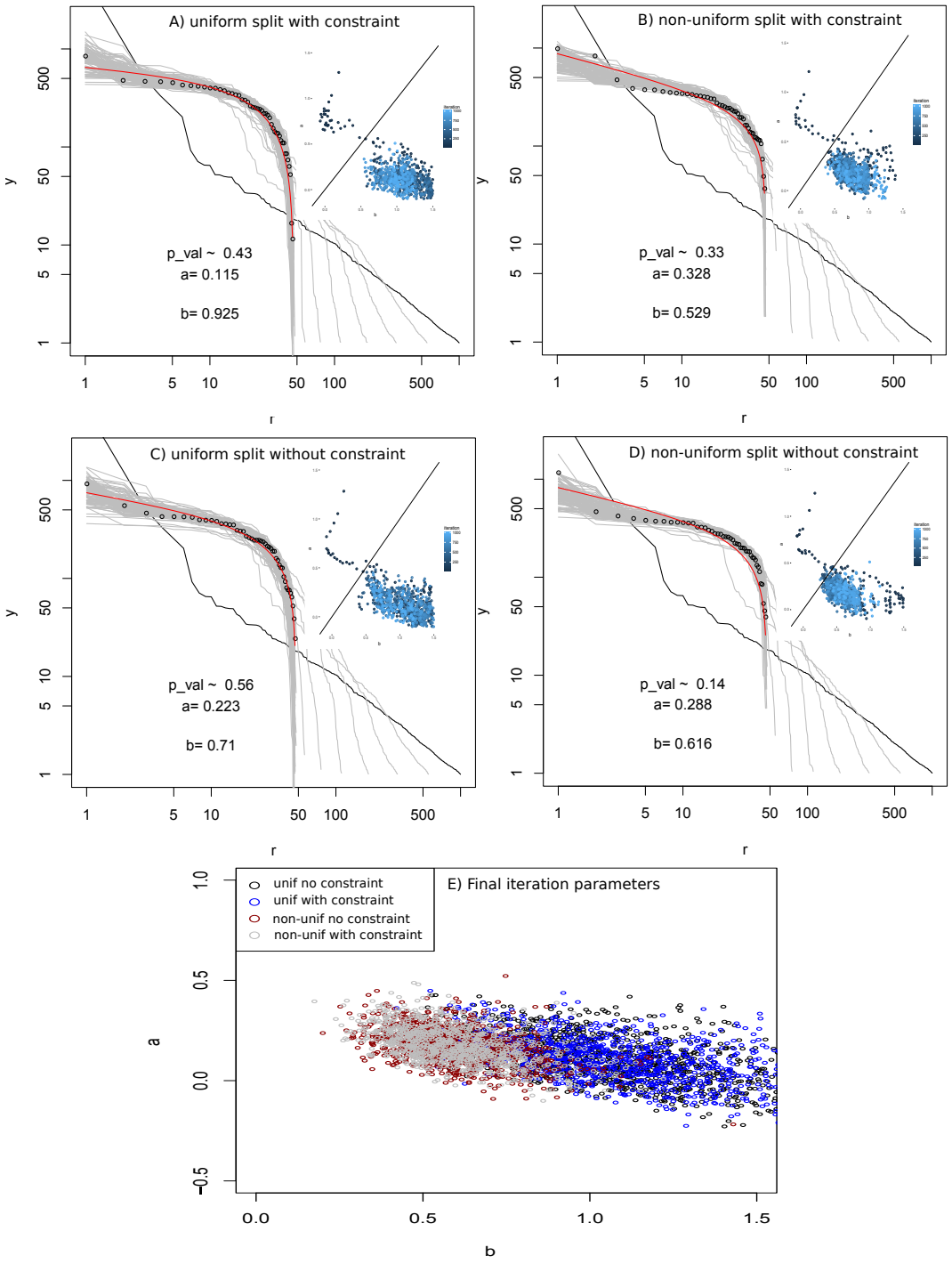


Figure 2.5: **Realizations of the split-merge process.** We see in black the initial rank-size function, in gray a sample of intermediate distributions, in circles the final distribution after 1000 iterations and in red the fitted DGBD to the final sample for: A) , B) uniform and non uniform (beta) unit split with spatial constraint; C) and D) uniform and non-uniform split without the spatial constraint On each plot we display p-value of the DGBD hypothesis for the final sample and fitted parameters. Inserted on each panel we show the fitted b and a parameters for each iteration. Panel E) shows the a and b parameters after 1000 iterations for 1000 different realization of these four processes.

want to improve the administration efficiency in areas with rapid population growth. Step 3 simulates the action in which some neighbor agglomerations are grouped into a single administrative unit or in which neighboring municipalities are merged into new, larger ones. By arranging the observations into a one-dimensional array we impose an spatial constraint and ensure that only adjacent municipalities are merged together or that new municipalities split from a larger one end up being geographically close. This model is a very generic one and does not consider historical and political factors, but its purpose is to give a framework for the simulation and quantification of the general processes behind administrative unit formation and evolution. Every country and region in the world has very specific circumstances that can not be encompassed by any general simulation, yet this model has flexible elements, such as the initial distribution, the number of merged and split units in each step and the way to split the units, that can be changed in order to more realistically reproduce some features and behaviors of this very complex social and political phenomenon. A two dimensional model including elements of cluster analysis or a more realistic GIS based approach are matter of future work.

We simulated computationally one particular option for this mechanism by means of what we call the *split-merge process*: we start with a sample $\underline{X}_0$ of N_0 observations following some initial probability distribution f_0. Then we picked the two largest values $x_{[1]}$ and $x_{[2]}$ of the sample and split each of them into two new values, $x_{[1]} \longrightarrow \big(p_1 x_{[1]},\, (1-p_1)x_{[1]}\big)$ and $x_{[2]} \longrightarrow \big(p_2 x_{[2]},\, (1-p_2)x_{[2]}\big)$ where p_1 and p_2 are uniformly distributed random numbers on the interval $[0, 1]$. With this, we simulated a process in which two large cities are split into two municipalities each. We randomly chose 3% of the remaining observations and paired them with an equal proportion of randomly selected values. With this we simulated the action in which 6% of the villages are merged together into new, larger municipalities. With the merged and the split values and the remaining observations we constructed the new sample $\underline{X}_1$ of size N_1 following distribution f_1. We repeated this process for n iterations; each step represents a system of municipalities in the process of being merged and split in a somewhat arbitrary manner.

We wonder what is the distribution of the municipalities populations after several iterations of this process. There are many options about how to proceed with this mechanism; for the present work we chose a particular version of the split-merge process to illustrate the idea.

We present results for four split-merge process realizations, all of them with the Pareto as the initial distribution. We chose this initial distribution because it is a commonly accepted model of city population distribution. In each process we started with $N_0 = 1000$ initial agglomerations and iterated 1000 times. First we chose uniform random numbers to split the large units, which has the effect of creating many small sized municipalities; then we chose these numbers from a beta distribution $f(x|\alpha, \beta) \sim x^{\alpha-1}(1-x)^{\beta-1}$ with $\alpha = \beta = 2$. This is a symmetric distribution with a maximum at $x = 0.5$, so large units are split into equal or similar sized units much more often. Fig.(2.5) A and B display these results: in each frame we see the rank-size plot for the initial set, for a selection of 100 intermediate set (one every 10 iterations) and for the final set after 1000 iterations. For the final set we also show the fitted DGBD, as well as the fitted b and a

parameters and the estimated p-value of the DGBD hypothesis. To compute this p-value we used the test described in the previous section. Within each frame we show the temporal evolution of the parameters in each realization from early iterations (dark blue) to latter ones (light blue). We also wondered how significant is the spatial constraint for our results. In order to study this, we repeated the simulations with fixed random seeds but this time without the spatial constraint. The results are displayed in Fig.(2.5) C and D. Finally we performed each of these four simulations 1000 times with different random seeds and registered the estimated a and b parameters for the final iteration; a scatter plot of them is provided in Fig.(2.5) E.

The first thing to observe is that in neither case can we reject the DGBD, according to the estimated p-values. Apparently, DGBD is indeed an adequate model for describing the intermediate distributions in this kind of process. Notice how the parameters follow a well defined trajectory on the $< b, a >$ plane at first (dark blue), but begin to move somewhat erratically as times passes (light blue). We also notice how almost invariably $b > a$ for large times. As we mentioned in the previous section, the region $a < b$ represents countries whose internal population distribution is somewhere between power-law and lognormal. We speculate that these distributions break as the territory is artificially divided into municipalities, leading to new distributions where $b > a$. The fact that sub-samples or aggregations of zipfian sets significantly deviate from Zipf's law has already been observed [22]. Here we have a new result in this direction: a perfect zipfian or lognormal set, describing population in natural agglomerations, transforms into a different kind of set of different kinds of objects, artificial municipalities with arbitrary borders, for which the initial distribution no longer holds. Rather than fixed values, the a and b parameters seem to have cluster-like attractors, this means, they end up moving in a bounded area when the number of iterations is very large. As the last panel of Fig.(2.5) shows, these clusters are more extended when units are split with the uniform distribution and more localized when they are split with the beta, non uniform distribution. Thus, the way in which units are divided does have an effect on the final distribution, even though the corresponding clusters overlap. We notice that the clusters for the processes with and without the spatial constraint clearly overlap, so the spatial feature of our model seems to have no effect on the final distribution.

In this study we simulated a very specific split-merge process, whose specific details may show relevant and need a more comprehensive study; nevertheless we see that this process and the DGBD function provide adequate tools for understanding formation and evolution of administrative divisions.

§2.5 Administrative divisions against natural cities

In order to further comprehend the differences between cities and administrative divisions we compared population distribution for cities, primary and secondary administrative divisions in a

country from our sample. We chose China for this analysis for being the most populated country in the world. Population data correspond to the 2010 Population Census by the National Bureau of Statistics of China. As we already mentioned, it has been observed that power law fits this kind of data well for the upper tail (big cities), but fails when smaller towns come into consideration, so a cut off is usually to be introduced [74, 36]. We considered cities with more than 750,000 inhabitants (128 cities) and took sub-samples of the 10, 11,..., 128 largest ones, fitted each of them to a power-law $x(r) = \frac{C}{x^a}$ and estimated the exponent a. Fig.(2.6A) shows the number of cities in the sample (the cut off) against the estimated exponent. We see that the parameter is very sensitive to the threshold value, so the decision of where to make it should be taken with extreme care. This truncation may also introduce deviations from power-law; for each sub-sample, we tested the goodness of fit of the power law model against the DGBD by the Akaike information criterion. We show in red those sub-samples for whom DGBD is better than power-law and in blue those for which power-law is better, according to this criterion. We also show the rank-size plot for the whole sample of cities and its respective DGBD and power-law fits. With these results, we speculate that truncation produces deviations from power-law which are well modeled by DGBD. Aside from truncation, what happens with territorial division comes into play?

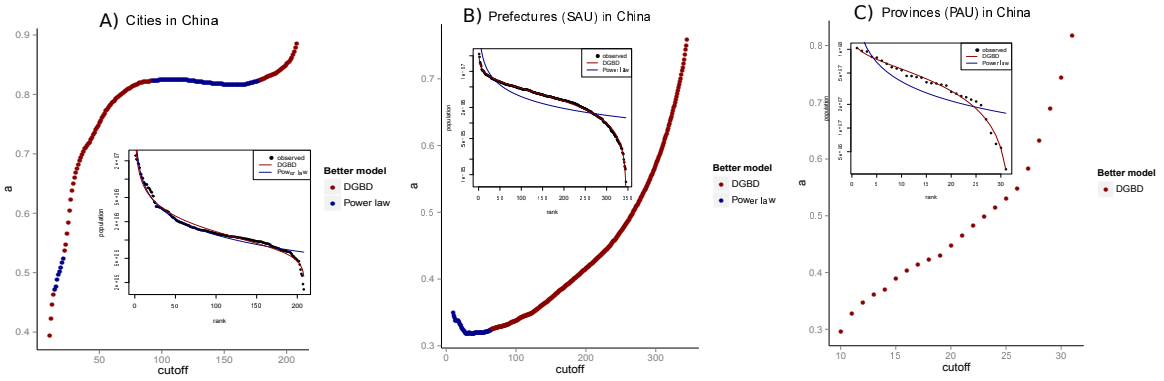


Figure 2.6: **Effects of truncation in China population.** Estimated exponent of power law vs. sample size (taking the "cutoff" largest units in the sample). According to Akaike information criterion, points marked in red and blue are better described by DGBD and power law respectively. We also see rank-size plot for the whole samples for the distribution population in China of its A) cities, B) prefectures and C) provinces.

The same analysis was performed for the prefectures (SAU) population in China. In Fig.(2.6B) it is possible to see again a high sensibility of the exponent in power law to the cut off value. Except for a few number of small sub-samples, DGBD is a better model than power law in almost every occasion. Now we are not only dealing with truncation, but with a process of city splitting as well, as there are natural cities divided into several administration units (for instance, the city of Shanghai is subdivided in 16 SAU's). The splitting mechanism also introduces deviations from power law. Finally we considered provinces (PAU) populations, see Fig.(2.6C). Notice that the four largest cities in China (Beijing, Changqing, Shanghai and Tianjin) are officially PAU; for

this analysis we considered Chinese provinces without these four, so we can clearly distinguish them from natural cities. This time DGBD performs better for almost all sub-samples. There is a certain correspondence between cities and administrative units in China: some administrative units are formed by a large central city plus its neighboring countryside, as is the case of the four largest cities; however, this correspondence ultimately breaks. For instance, the cities of Beijing, Chongqing, Shanghai and Tianjin, which have the status of direct-controlled municipalities, equivalent to that of provinces, are divided in 16, 38, 16 and 16 SAUs (districts and counties) respectively. This accounts for the splitting phenomenon; the effect of merging also becomes visible because there are some sets of cities grouped into the same unit (for example, the province of Anhui has 12 large cities). This merging operation also causes discrepancies between the observed distribution and power law. As we see in the rank-size plot, DGBD is a better model for PAU population.

In summary, DGBD is a better model than power law for describing administrative unit population. Although there is an initial correspondence between cities and administrative divisions, this correspondence ceases to hold when large cities are divided into several units (split) and when separate units are grouped into larger ones (merge), yielding divisions with more than one large city or town. Cities and administrative units are clearly different objects, whose population are described by different kinds of data sets and in most cases they follow different distributions.

§2.6 Appendix 1 to chapter 2

Fig.(2.7) shows ranked SAU population in logarithmic scale for our whole data base, composed of 150 countries and territories. In the table we show the results of our analysis for the set of 150 countries. For each country, the table displays the fitted b and a parameters, the coefficient of determination of linear regression, number of SAU's N, estimated p-value with the bootstrap-K-S approach and logarithmic difference of Akaike's Information Criterion, AIC= $log(\mathrm{AIC}_{\mathrm{DGBD}}) - log(\mathrm{AIC}_{\text{power law}})$.

Table 2.1: Fitted b and a parameters, R^2 of lin. regression, sample size (number of SAU's) N, p-val estimation for K-S test and AIC criterion for our samples of countries. AIC indicates $log(\mathrm{AIC}_{\mathrm{DGBD}}) - log(\mathrm{AIC}_{\text{power law}})$.

country	b	a	R^2	N	p-val	AIC	country	b	a	R^2	N	p-val	AIC
Afghanistan	0.47	0.5	0.99	382	0.18	-455.37	Macau	1.67	0.16	0.91	7	0.01	0.84
Albania	0.5	0.68	0.98	61	0.19	-55.1	Madagascar	0.61	0.25	0.98	111	0.24	-155.42
Algeria	0.46	0.58	1	1541	0.88	-3068.41	Malaysia	0.41	0.72	0.99	144	0.41	-183.16
American Samoa	1.82	0.72	0.95	15	0.22	-7.69	Mali	1.27	-0.02	0.91	50	0.03	-40.61
Argentina	0.7	0.87	0.99	511	0.2	-823.66	Malta	0.72	0.35	0.98	68	0.01	-91.97
Australia	0.96	0.89	0.97	654	0	-641.16	Marshall Islands	0.89	1.02	0.91	25	0	-4.23
Austria	0.49	0.35	0.9	95	0.05	-43.46	Martinique	0.76	0.61	0.98	34	0.4	-36.41
Bangladesh	0.36	0.37	0.98	64	0.21	-62.84	Mauritania	1.08	0.14	0.97	55	0.51	-67.41
Belgium	0.42	0.61	0.97	43	0.2	-24.13	Mayotte	-0.03	0.8	0.94	16	0.48	4.51
Belize	0.73	0.23	0.97	12	0.12	-6.97	Mexico	0.79	0.93	0.99	2456	0	-3762.12
Benin	0.18	0.38	0.94	77	0.01	-19.45	Morocco	0.8	0.26	0.98	75	0.04	-95.22
Bhutan	0.72	0.3	0.97	262	0.04	-302.71	Mozambique	0.64	0.3	0.99	148	0.45	-258.41
Bolivia	0.57	0.82	0.99	112	0.92	-140.6	Myanmar	0.93	0.16	0.98	63	0.31	-94.98
Bosnia and Herzegovina	0.29	0.44	0.99	105	0.64	-150.06	Namibia	0.34	0.34	0.99	121	0.12	-166.95
Botswana	1.72	0.03	0.97	28	0.38	-33.61	Nepal	0.81	0.25	0.97	75	0.08	-83.06
Brazil	0.38	0.66	0.99	556	0.49	-629.25	Netherlands	0.35	0.6	0.99	504	0.78	-705.13
Bulgaria	0.4	0.75	1	262	0.87	-424.2	New Caledonia	0.36	0.95	0.97	33	0.1	-8.11
Burkina Faso	0.32	0.4	0.97	45	0.24	-29.62	New Zealand	0.67	0.62	0.98	74	0.68	-74.5
Burundi	0.24	0.23	0.99	129	0.54	-217.52	Nicaragua	0.42	0.64	0.98	153	0.14	-144.48
Cameroon	0.31	0.54	0.98	58	0.46	-44.25	Niger	0.73	0.11	0.96	37	0.21	-36.48
Canada	0.49	0.87	0.99	293	0.32	-329.85	Nigeria	0.25	0.28	1	775	0.75	-1694.98
Central A. Republic	0.79	0.44	0.97	72	0.32	-67.95	Niue	0.67	0.44	0.95	14	0.15	-3.99
Chad	0.7	0.17	0.95	62	0.02	-56.57	Norway	0.44	0.78	1	431	0.38	-713.23
Chile	1.43	0.45	0.97	54	0.06	-63.42	Oman	0.83	0.61	0.99	61	0.2	-83.2
China	0.76	0.25	1	345	0.51	-810.67	Pakistan	0.66	0.51	0.95	30	0.05	-17.27
Colombia	0.51	0.7	0.99	1057	0.01	-1183.06	Palestine	0.8	0.22	0.96	16	0.27	-11.47
Congo	0.49	0.43	0.99	100	0.82	-135.53	Panama	0.43	0.95	0.99	76	0.38	-77.53
Costa Rica	0.46	0.52	0.99	81	0.75	-119.18	Papua New Guinea	0.12	0.32	0.98	87	0.05	-52.71
Cote d'Ivore	0.36	0.51	0.95	33	0.37	-12.09	Paraguay	0.39	0.77	1	224	0.59	-372.36
Croatia	0.31	0.77	1	556	0.17	-705.73	Peru	0.46	0.69	0.99	194	0.67	-201.04
Cuba	0.24	0.58	0.99	168	0.38	-134.79	Philippines	0.44	0.57	0.99	1634	0.19	-2559.73
Czech Republic	0.15	0.42	0.94	77	0.07	-14.45	Poland	0.17	0.46	0.99	379	0.7	-350.97
Denmark	0.83	0.18	0.84	99	0	-49.92	Portugal	0.47	0.8	0.98	308	0	-285.39
Dominican Republic	0.3	0.97	1	155	0.73	-157.19	Reunion	0.5	0.79	0.95	24	0.02	-5.78
Ecuador	0.49	0.82	0.99	216	0.39	-251.93	Romania	0.22	0.62	0.97	2951	0	-1273.87
Equatorial Guinea	0.44	0.79	0.97	30	0.28	-11.06	Russia	0.32	0.65	0.98	2581	0	-2172.22
Egypt	1.4	0.1	0.99	367	0.61	-687.37	Rwanda	0.02	0.15	0.98	30	0.04	1.96
El Salvador	0.5	0.76	1	262	0.87	-432.23	Sao Tome and Principe	0.34	0.94	0.96	7	0.25	3.42
Estonia	0.3	0.78	0.96	241	0.04	-95.11	Saudi Arabia	0.39	0.98	0.99	118	0.01	-85.43
Ethiopia	0.93	0.38	0.98	66	0.03	-78.56	Senegal	0.6	0.4	0.98	45	0.06	-51
Faroe Islands	0.85	0.87	0.96	34	0.37	-19.39	Sierra Leone	1.39	-0.26	0.85	15	0.13	-5.16
Fiji	1.01	0.7	0.98	15	0.18	-11.32	Slovakia	0.41	0.26	0.99	79	0.17	-117.57
Finland	0.28	0.8	0.99	69	0.91	-43.7	Slovenia	0.46	0.66	0.99	210	0.53	-263.93
France	0.45	0.42	0.98	96	0.12	-107.51	Solomon Islands	0.62	0.35	0.99	183	0.08	-283.63

Continued on next page

Table 2.1 – continued from previous page

country	b	a	R^2	N	p-val	AIC	country	b	a	R^2	N	p-val	AIC
French Guiana	1.42	0.69	0.99	22	0.54	-27.63	South Africa	0.63	0.39	0.99	52	0.75	-72.7
French Polynesia	0.82	0.88	0.97	49	0.01	-31.78	South Sudan	0.44	0.3	0.98	79	0.17	-93.72
Gabon	0.25	1.11	0.99	48	0.98	-20.67	Spain	0.55	0.58	0.98	52	0.74	-54.25
Gambia	0.31	0.9	0.99	37	0.5	-24.52	Sri Lanka	0.46	0.37	0.99	331	0.12	-581.19
Georgia	0.58	0.49	0.95	66	0.07	-44.94	Sudan	0.45	0.42	0.98	131	0.56	-150.25
Germany	0.29	0.44	0.99	402	0.43	-523.46	Suriname	1.08	0.52	0.96	62	0.08	-58.23
Ghana	0.08	0.47	0.95	110	0.11	-7.05	Swaziland	0.35	0.21	0.98	55	0.4	-69.74
Greece	1.11	0.38	0.99	326	0.01	-572.3	Sweden	0.31	0.69	0.99	289	0.03	-333.33
Greenland	1	0.39	0.96	19	0.06	-13.52	Switzerland	0.49	0.56	0.99	181	0.03	-291.06
Guadeloupe	0.72	0.45	0.98	32	0.32	-35.21	Syria	0.66	0.5	0.91	61	0.06	-25.96
Guatemala	0.44	0.55	0.99	331	0.28	-484.16	Taiwan	1.29	0.42	0.96	22	0.13	-17.8
Guinea Bissau	0.38	0.59	0.95	39	0.5	-14.28	Tajikistan	0.5	0.54	0.97	75	0.01	-58.83
Guyana	0.97	0.48	0.98	117	0.74	-163.06	Tanzania	0.47	0.27	0.99	129	0.33	-246.05
Haiti	0.24	0.63	0.95	42	0.23	-6.34	Thailand	0.41	0.39	0.99	926	0	-1362.43
Honduras	0.29	0.69	0.99	282	0.77	-326.14	Timor Leste	0.44	0.43	0.99	65	0.54	-85.28
Iceland	0.65	1.17	0.99	79	0.96	-97.43	Togo	0.18	0.58	0.92	35	0.45	-0.24
India	0.94	0.22	0.99	638	0.11	-1270.03	Tonga	0.51	0.9	0.97	23	0.56	-7.35
Indonesia	0.58	0.6	0.99	497	0	-623.93	Tunisia	0.43	0.26	0.99	263	0.39	-482.35
Iran	0.37	0.6	0.98	252	0.07	-218.17	Turkey	0.5	0.86	0.99	923	0.01	-902.32
Isle of Man	0.22	0.91	0.98	24	0.65	-1.7	Uganda	0.47	0.26	0.98	160	0.16	-200.82
Israel	1.04	0.12	0.96	15	0.04	-10.35	Ukraine	0.21	0.59	0.99	678	0.01	-501.9
Italy	0.32	0.54	0.99	110	0.66	-121.37	United Kingdom	0.39	0.32	0.96	406	0	-379.09
Japan	0.44	0.6	0.99	1180	0	-1443.85	United States	0.7	0.92	1	3143	0.05	-5487.94
Jordan	0.31	1.07	0.98	89	0.54	-39.15	Vanuatu	0.66	0.55	0.99	62	0.5	-76.95
Kazakhstan	0.25	0.6	0.98	200	0.51	-135.78	Venezuela	0.51	0.75	1	336	0.12	-601.21
Kenya	0.51	0.31	0.98	70	0.04	-89.2	Vietnam	0.65	0.17	0.97	661	0.48	-828.92
Lebanon	0.62	0.56	0.97	26	0.16	-18.63	Virgin Islands US	1.79	-0.03	0.97	20	0.2	-23.04
Lesotho	0.38	0.38	0.95	129	0	-86.26	Wallis and Futuna	0.32	0.22	0.97	5	0.03	1.92
Liberia	0.73	0.56	0.98	136	0.08	-143.22	Yemen	0.69	0.33	1	333	0.46	-709.93
Lithuania	0.42	0.55	0.94	60	0.01	-23.37	Zambia	0.33	0.46	0.97	74	0.34	-47.68
Luxembourg	0.22	0.77	0.99	105	0.43	-77.5	Zimbabwe	0.29	0.49	0.97	63	0.52	-33.52

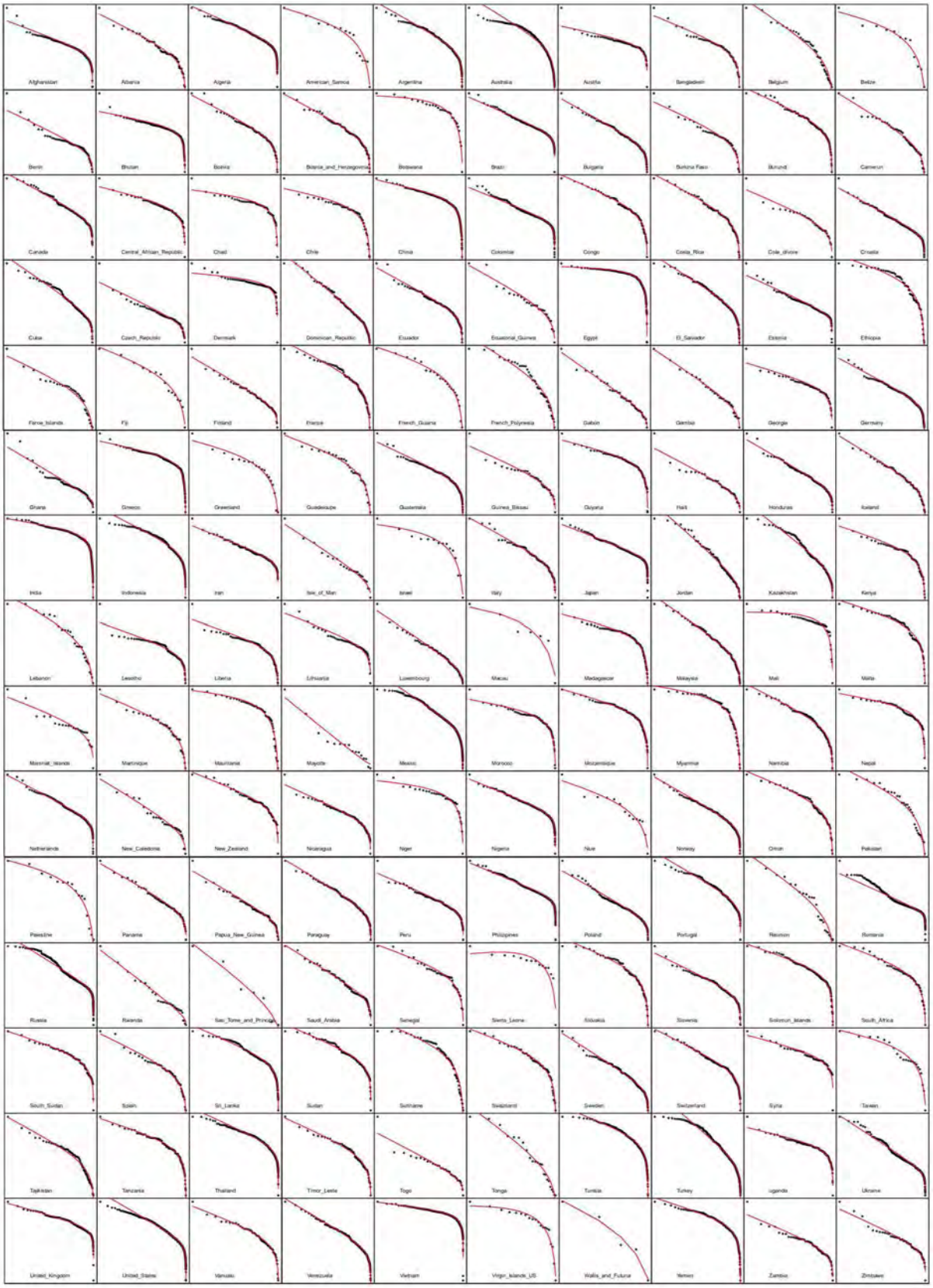


Figure 2.7: **Ranked SAU population in log-log scale for the 150 country data base**. Dots are actual population, red lines are DGBD fits for each data set.

§2.7 Appendix 2 to chapter 2

We performed all analysis of section 2.3 but this time estimating the parameters of the DGBD with a nonlinear regression. The procedure to estimate the parameters a and b is to perform a nonlinear regression of the population f as a function of the rank r with the model given by Eq.(2.2) [10]. We obtained the initial parameters for this iterative method by performing a linear regression of the logarithmic transformation of Eq.(2.2); to minimize the sum square of the residuals we utilized the Levenberg-Marquardt algorithm, which is more robust than Gauss-Newton [37]. We also computed the estimated parameters by performing a linear regression of the logarithmic transform of (2.2). In general, the computed sets of parameters show small variations; our criterion for choosing the nonlinear over the linear regression is that for each of the 150 data sets, the nonlinear gives a smaller sum square of residuals than the linear regression. Only for a few cases, for instance $a = b$, it is possible to analytically compute the probability density function of the DGBD and consequently perform a maximum likelihood estimation; for such cases, MLE and our nonlinear estimation coincide.

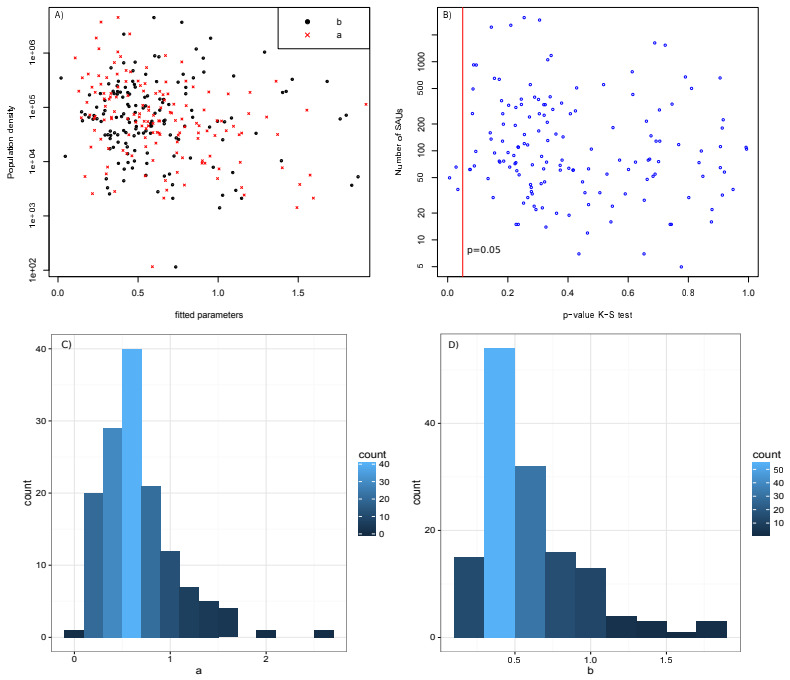


Figure 2.8: **A) shows in semi-logarithmic scale the fitted parameters against population density for the 150 countries and territories we analyzed. B) is a semi-logarithmic plot of the p-value for the K-S test against the number of SAUs for the 150 countries; left to the red line is the rejection zone (p-val<0.05). C) and D) are histograms of fitted a and b for the 141 selected countries.**

Again we discard all countries with less than 10 SAU's and we test the goodness of fit of the DGBD with the bootstrap approach to the Kolmogorov-Smirnov test, rejecting the DGBD whenever the

corresponding p-value is less than 0.05. In the countries where DGBD is not rejected we compare its performance with that of power law in two ways: first we perform a Likelihood Ratio Test (because power-law is a special case of DGBD). Here, a large p-value leads to reject the DGBD in favor of power-law. Following the suggestion in [21] we reject DGBD whe the p-value is higher than 0.001. A different approach is the Akaike information criterion (AIC), already described. There is only one case in which these two criteria do not coincide (Togo, where AIC favors DGBD and LRT favors power law, see table 2.2). After discarding all countries that have not enough SAU's to do a regression analysis the sample reduces to 147 countries; 144 of them show estimated p-values of the K-S test higher than 0.05 (Ethiopia, Mali and Niger are discarded); out of the remaining 144 countries, DGBD is preferred over power law by both criteria in 141 cases (in Mayotte and Rwanda both AIC and LRT favor power law, in Togo there is no coincidence, see table 2.2). Ranked population and fitted DGBD for each country that we studied can be seen in Fig.(2.11).

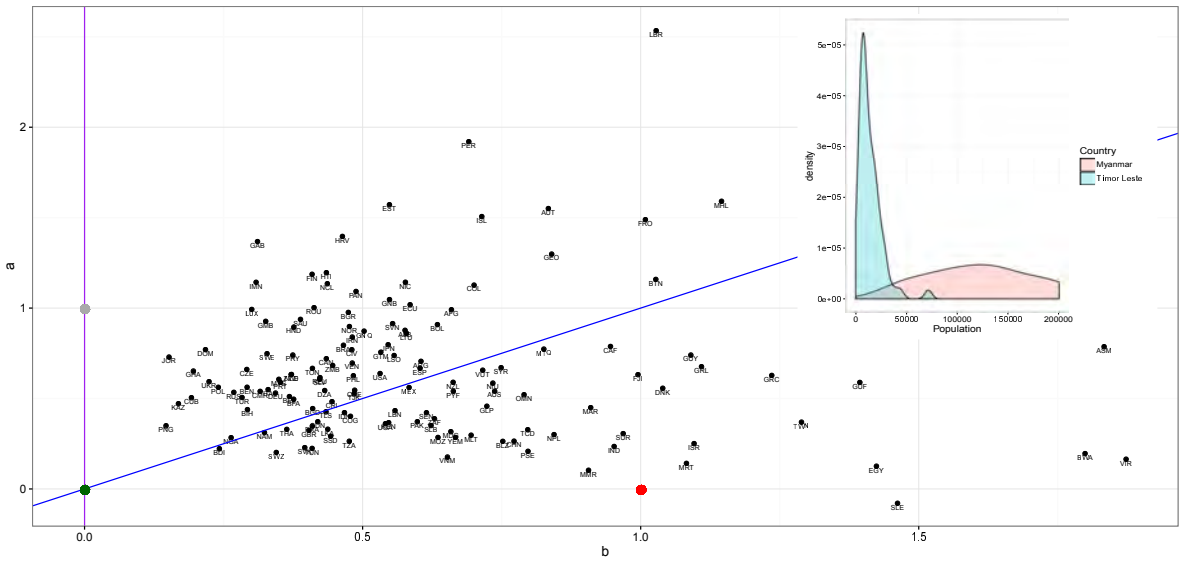
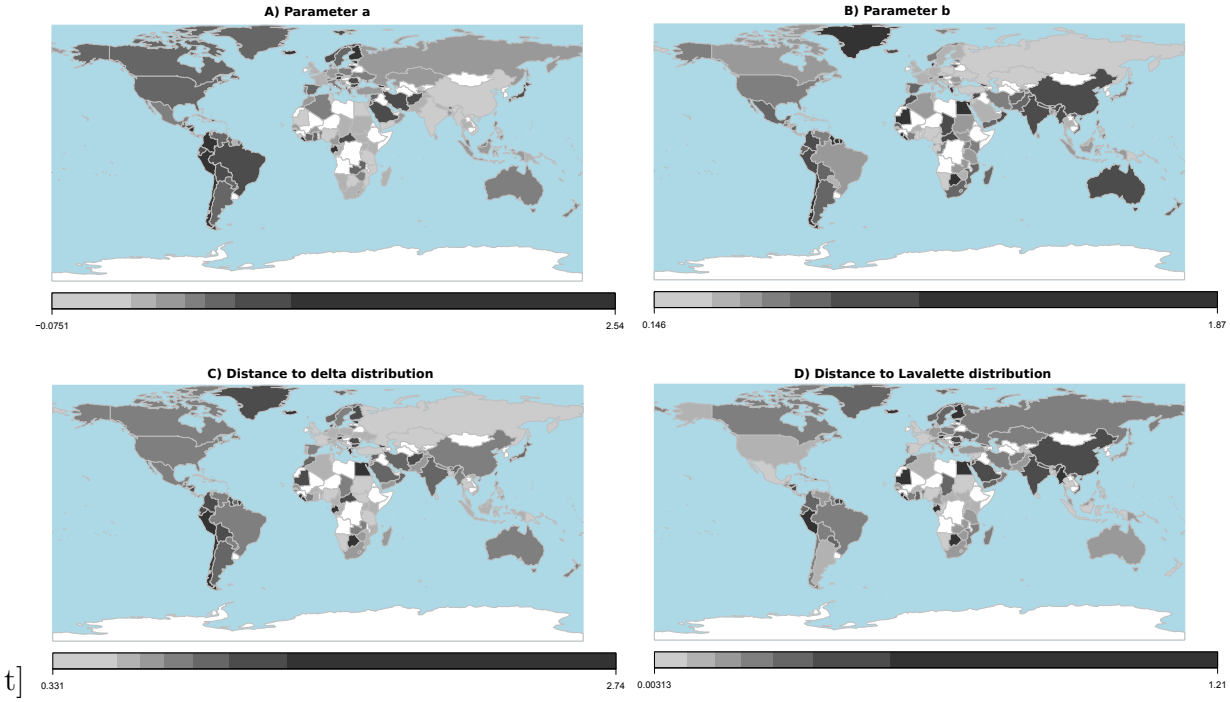


Figure 2.9: **Scatter plot of fitted b and a for the 141 selected countries.** Each country is indicated by its three letter ISO3 code. In purple the line $b = 0$, which represents a perfect power law; in blue the line $b = a$, which is the Lavalette rank-size function; gray point is $(0, 1)$ a pure Zipf's law; green is $(0, 0)$, representing a point located probability distribution; red point is $(1, 0)$, random uniform distribution. We also show the density histogram for two countries: Myanmar, whose parameters indicate a closeness to a random uniform distribution, and Timor Leste, whose parameters suggest a distribution similar to a very narrow lognormal.

After these selection criteria a set of 141 countries remains. Mean and standard deviation of the sample size (number of SAUs in these 141 countries) are 278.4 and 510.9 respectively. The mean and standard deviation for the fitted b parameter are 0.61 and 0.34, while the mean and standard deviation for a parameter are 0.66 and 0.39. It can be seen in Eq.(2.2) that along the small rank

regime the parameter a has a more marked effect on the shape of the distribution than b, which has a greater repercussion when the rank is high. This means that a describes the high populated units regime (the tail of the distribution) while b characterizes the behavior of the low populated units. Together they give a sense of the internal arrangement of population across municipalities within a given country or territory. Another crucial variable for describing this is population density, which involves the number of inhabitants and the available space, so a natural question to ask is whether there is a correlation between the DGBD parameters and population density. We show in Fig.(2.8) A) a plot of the fitted parameters against population density in semi-logarithmic scale. A visual inspection of this plot together with linear correlation coefficients (0.05 for b and population density, -0.22 for a and density) suggest there is no relationship between the fitted parameters and density. We wonder if the goodness of fit of the DGBD is sensible to sample size, i.e. number of SAUs in each country or territory. We chose the p-value of the K-S test as a measure of the goodness of fit (there is no clear equivalent to the coefficient of determination for nonlinear regression) and show in Fig.(2.8) B) a plot in semi-logarithmic scale of this p-value vs number of SAUs. The linear correlation coefficient between these two variables is -0.11, which jointly with a visual inspection of the plot suggest there is no correlation. Panels C) and D) of this figure display histograms of a and b, indicating that they are not randomly distributed but rather clustered around central values. There is one case, Sierra Leone, for which $a < 0$. When this happens, Eq.(2.2) fails to represent a rank-size distribution because it is not monotonous; however, the fittings are still good; usually this means that the maximum of the fitted curved is reached below $rank = 1$ [26]. The corresponding scatter plot of the fitted parameters is shown in Fig. (2.9). Maps shown the fitted parameters and distances to a random uniform and to a Lavalette distribution are given in Fig. (2.10).



t]

Figure 2.10: A) and B) show the fitted a and b parameters respectively of the DGBD for the internal SAU population of each selected country and territory, C) displays the euclidean distance in the parameter plane $< b, a >$ to a delta distribution in which all SAUs have the same population and D) shows the distance to a Lavalette distribution.

Table 2.2: Fitted b and a parameters, sample size (number of SAU's) N, p-value estimation for K-S test, AIC criterion between DGBD and power law and p-value for the Likelihood Ratio Test between DGBD and power law (small p-value indicates DGBD is preferred).

country	b	a	N	K-S	AIC	LRT	country	b	a	N	K-S	AIC	LRT
Afghanistan	0.66	0.99	382	0.24	-455.37	0	Macau	1.76	0.44	7	0.44	0.84	0.005
Albania	0.58	0.88	61	0.42	-55.10	0	Madagascar	0.66	0.32	111	0.23	-155.42	0
Algeria	0.43	0.55	1541	0.72	-3068.41	0	Malaysia	0.35	0.61	144	0.38	-183.16	0
American Samoa	1.83	0.79	15	0.74	-7.69	0	Mali	1.40	0.23	50	0.01	-40.61	0
Argentina	0.60	0.71	511	0.43	-823.66	0	Malta	0.70	0.30	68	0.09	-91.97	0
Australia	0.74	0.54	654	0.15	-641.16	0	Marshall Islands	1.15	1.59	25	0.47	-4.23	3E-05
Austria	0.83	1.55	95	0.16	-43.46	0	Martinique	0.83	0.78	34	0.51	-36.41	0
Bangladesh	0.41	0.45	64	0.40	-62.84	0	Mauritania	1.08	0.15	55	0.53	-67.41	0
Belgium	0.37	0.51	43	0.31	-24.13	0	Mayotte	0.04	0.95	16	0.54	4.51	0.677
Belize	0.75	0.27	12	0.46	-6.97	1E-06	Mexico	0.58	0.56	2456	0.14	-3762.12	0
Benin	0.29	0.57	77	0.17	-19.45	0	Morocco	0.91	0.45	75	0.17	-95.22	0
Bhutan	1.03	1.16	262	0.18	-302.71	0	Mozambique	0.64	0.29	148	0.68	-258.41	0
Bolivia	0.63	0.91	112	0.91	-140.60	0	Myanmar	0.91	0.11	63	0.47	-94.98	0
Bosnia and Herzegovina	0.29	0.44	105	0.99	-150.06	0	Namibia	0.32	0.32	121	0.25	-166.95	0
Botswana	1.80	0.20	28	0.65	-33.61	0	Nepal	0.84	0.30	75	0.33	-83.06	0
Brazil	0.47	0.80	556	0.52	-629.25	0	Netherlands	0.37	0.64	504	0.81	-705.13	0
Bulgaria	0.47	0.98	262	0.41	-424.20	0	New Caledonia	0.44	1.14	33	0.59	-8.11	0
Burkina Faso	0.38	0.50	45	0.32	-29.62	0	New Zealand	0.66	0.59	74	0.83	-74.50	0
Burundi	0.24	0.23	129	0.69	-217.52	0	Nicaragua	0.58	1.15	153	0.25	-144.48	0

Continued on next page

Table 2.2 – continued from previous page

country	b	a	N	K-S	AIC	LRT	country	b	a	N	K-S	AIC	LRT
Cameroon	0.32	0.54	58	0.92	-44.25	0	Niger	0.78	0.19	37	0.03	-36.48	0
Canada	0.43	0.72	293	0.35	-329.85	0	Nigeria	0.26	0.29	775	0.61	-1694.98	0
Central African Republic	0.95	0.79	72	0.22	-67.95	0	Niue	0.73	0.59	14	0.33	-3.99	3E-05
Chad	0.80	0.33	62	0.07	-56.57	0	Norway	0.48	0.90	431	0.61	-713.23	0
Chile	1.68	1.36	54	0.24	-63.42	0	Oman	0.79	0.52	61	0.37	-83.20	0
China	0.77	0.27	345	0.38	-810.67	0	Pakistan	0.60	0.38	30	0.15	-17.27	0
Colombia	0.70	1.13	1057	0.33	-1183.06	0	Palestine	0.80	0.21	16	0.88	-11.47	0
Congo	0.48	0.41	100	0.84	-135.53	0	Panama	0.49	1.10	76	0.70	-77.53	0
Costa Rica	0.44	0.49	81	0.67	-119.18	0	Papua New Guinea	0.15	0.35	87	0.31	-52.71	0
Côte d'Ivore	0.48	0.77	33	0.28	-12.09	0	Paraguay	0.37	0.74	224	0.92	-372.36	0
Croatia	0.46	1.40	556	0.27	-705.73	0	Peru	0.69	1.92	194	0.22	-201.04	0
Cuba	0.19	0.51	168	0.30	-134.79	0	Philippines	0.48	0.63	1634	0.69	-2559.73	0
Czech Republic	0.29	0.66	77	0.19	-14.45	0	Poland	0.24	0.57	379	0.30	-350.97	0
Denmark	1.04	0.56	99	0.09	-49.92	0	Portugal	0.35	0.59	308	0.23	-285.39	0
Dominican Republic	0.22	0.77	155	0.36	-157.19	0	Reunion	0.42	0.62	24	0.29	-5.78	5E-06
Ecuador	0.59	1.02	216	0.66	-251.93	0	Romania	0.41	1.01	2951	0.30	-1273.87	0
Equatorial Guinea	0.50	0.88	30	0.80	-11.06	0	Russia	0.27	0.54	2581	0.21	-2172.22	0
Egypt	1.42	0.13	367	0.18	-687.37	0	Rwanda	0.02	0.15	30	0.27	1.96	0.023
El Salvador	0.42	0.61	262	0.30	-432.23	0	Sao Tome and Principe	0.32	0.87	7	0.65	3.42	0.098
Estonia	0.55	1.57	241	0.27	-95.11	0	Saudi Arabia	0.39	0.94	118	0.77	-85.43	0
Ethiopia	0.86	0.26	66	0.03	-78.56	0	Senegal	0.61	0.42	45	0.45	-51.00	0
Faroe Islands	1.01	1.49	34	0.46	-19.39	0	Sierra Leone	1.46	-0.08	15	0.23	-5.16	1E-07
Fiji	1.00	0.64	15	0.74	-11.32	0	Slovakia	0.40	0.23	79	0.38	-117.57	0
Finland	0.41	1.19	69	0.37	-43.70	0	Slovenia	0.55	0.92	210	0.33	-263.93	0
France	0.41	0.35	96	0.20	-107.51	0	Solomon Islands	0.62	0.35	183	0.55	-283.63	0
French Guiana	1.39	0.59	22	0.88	-27.63	0	South Africa	0.63	0.39	52	0.85	-72.70	0
French Polynesia	0.66	0.54	49	0.13	-31.78	0	South Sudan	0.44	0.30	79	0.57	-93.72	0
Gabon	0.31	1.37	48	0.66	-20.67	0	Spain	0.60	0.67	52	0.68	-54.25	0
Gambia	0.33	0.93	37	0.95	-24.52	0	Sri Lanka	0.44	0.33	331	0.32	-581.19	0
Georgia	0.84	1.30	66	0.21	-44.94	0	Sudan	0.42	0.37	131	0.33	-150.25	0
Germany	0.34	0.53	402	0.29	-523.46	0	Suriname	0.97	0.31	62	0.07	-58.23	0
Ghana	0.20	0.66	110	0.23	-7.05	2E-06	Swaziland	0.34	0.21	55	0.69	-69.74	0
Greece	1.24	0.63	326	0.20	-572.30	0	Sweden	0.33	0.75	289	0.70	-333.33	0
Greenland	1.11	0.68	19	0.40	-13.52	0	Switzerland	0.49	0.55	181	0.91	-291.06	0
Guadeloupe	0.72	0.46	32	0.91	-35.21	0	Syria	0.75	0.67	61	0.42	-25.96	0
Guatemala	0.53	0.76	331	0.32	-484.16	0	Taiwan	1.29	0.37	22	0.29	-17.80	0
Guinea Bissau	0.55	1.05	39	0.28	-14.28	0	Tajikistan	0.48	0.53	75	0.62	-58.83	0
Guyana	1.09	0.74	117	0.27	-163.06	0	Tanzania	0.48	0.27	129	0.70	-246.05	0
Haiti	0.43	1.20	42	0.27	-6.34	3E-06	Thailand	0.36	0.33	926	0.09	-1362.43	0
Honduras	0.38	0.90	282	0.42	-326.14	0	Timor Leste	0.43	0.43	65	0.91	-85.28	0
Iceland	0.71	1.51	79	0.67	-97.43	0	Togo	0.37	1.06	35	0.28	-0.24	0.002
India	0.95	0.24	638	0.17	-1270.03	0	Tonga	0.41	0.67	23	0.31	-7.35	1E-06
Indonesia	0.47	0.43	497	0.08	-623.93	0	Tunisia	0.41	0.23	263	0.08	-482.35	0
Iran	0.48	0.84	252	0.32	-218.17	0	Turkey	0.28	0.51	923	0.09	-902.32	0
Isle of Man	0.31	1.15	24	0.55	-1.70	4E-04	Uganda	0.54	0.36	160	0.14	-200.82	0
Israel	1.10	0.25	15	0.23	-10.35	0	Ukraine	0.22	0.60	678	0.79	-501.90	0
Italy	0.33	0.55	110	0.99	-121.37	0	United Kingdom	0.40	0.33	406	0.34	-379.09	0
Japan	0.55	0.80	1180	0.34	-1443.85	0	United States	0.53	0.64	3143	0.25	-5487.94	0
Jordan	0.15	0.73	89	0.22	-39.15	0	Vanuatu	0.72	0.66	62	0.60	-76.95	0
Kazakhstan	0.17	0.47	200	0.18	-135.78	0	Venezuela	0.48	0.70	336	0.75	-601.21	0
Kenya	0.55	0.37	70	0.28	-89.20	0	Vietnam	0.65	0.18	661	0.90	-828.92	0
Lebanon	0.56	0.44	26	0.25	-18.63	0	Virgin Islands US	1.87	0.17	20	0.36	-23.04	0
Lesotho	0.56	0.74	129	0.18	-86.26	0	Wallis and Futuna	0.32	0.22	5	0.78	1.92	0.013
Liberia	1.03	2.54	136	0.14	-143.22	0	Yemen	0.67	0.29	333	0.24	-709.93	0
Lithuania	0.58	0.87	60	0.23	-23.37	0	Zambia	0.45	0.69	74	0.23	-47.68	0
Luxembourg	0.30	1.00	105	0.48	-77.50	0	Zimbawe	0.37	0.64	63	0.32	-33.52	0

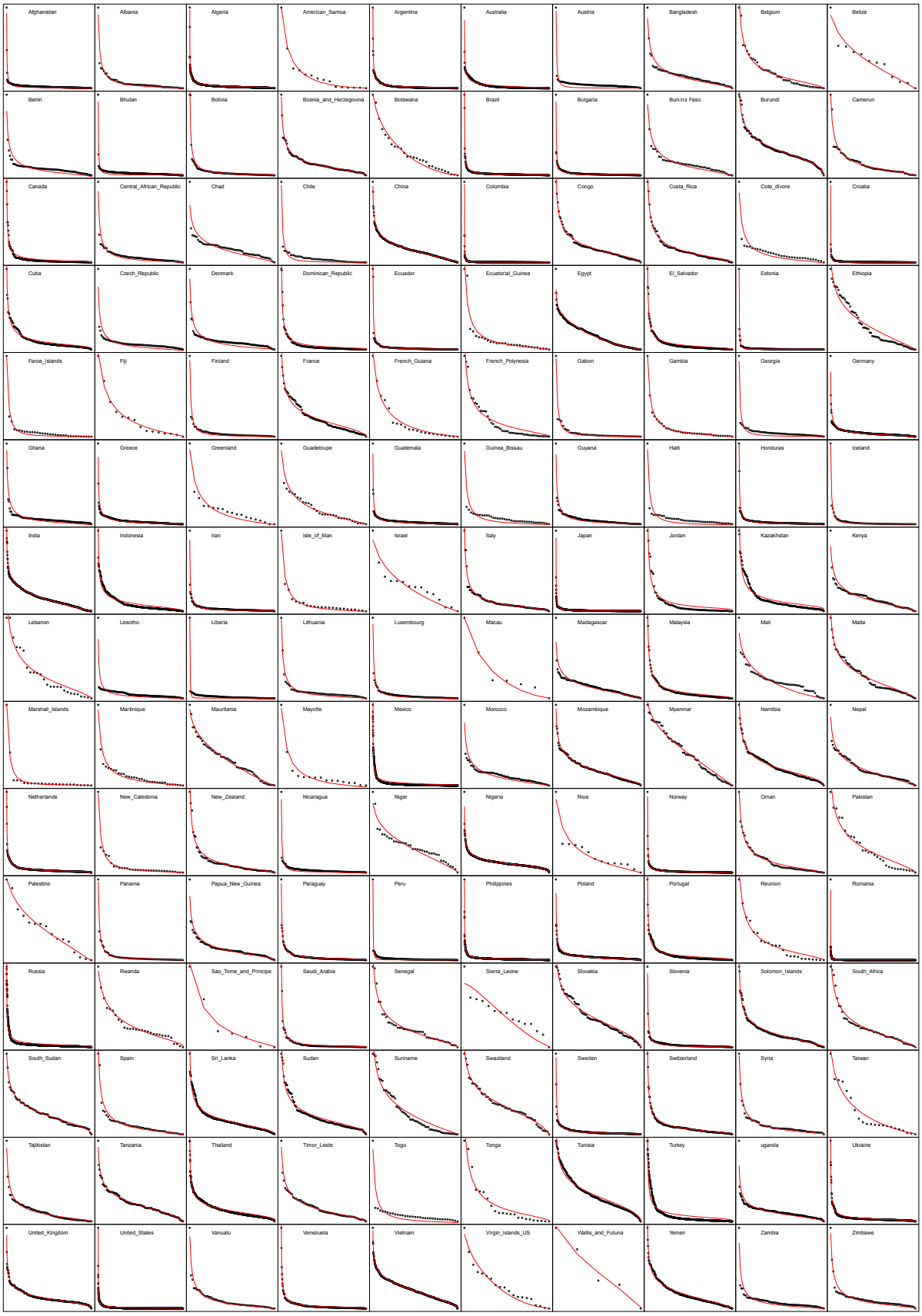


Figure 2.11: **Ranked SAU population in log-log scale for the 150 country data base**. Dots are actual population, red lines are DGBD fits for each data set.

Chapter 3

A Different Perspective: the Lavalette Distribution

In fact, all epistemological value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, non random regularity.

—B.V. Gnedenko and A.N. Kolmogorov, *Limit Distribution for Sum of Independent Random Variables*

In this chapter we present the Lavalette family of continuous distributions. The question that motivated the construction of this distribution is the following: what is the pdf of a random variable whose rank-size function is a DGBD? We will start, in section 3.1, by establishing a one to one relationship between the pdf and the rank-size representations; even though the pdf related to the DGBD exists, we will see that it is not possible, in general, to give its closed form except for a few cases. One such case is when the parameters of the DGBD are equal, $a = b$, in which case this is called the Lavalette rank function. For this particular case we compute the pdf, yielding the novel Lavalette distribution function. Section 3.2 exhibits this family of distributions and discusses the interpretation of its parameters and the existence of its moments. Then we will study, in section 3.3, a resemblance between the Lavalette and the lognormal distributions, which prove to be very difficult to distinguish from one another, specially at the center of the distribution. This fact is of particular importance for the possible applications of the Lavalette distribution, making it a suitable alternative for data that does not clearly follow a lognormal or a power -law. After this we present three examples of occurrence of the Lavalette distribution in real data coming from different phenomena; finally, we discuss the goodness of fit tests used to evaluate the performance of the Lavalette distribution as a statistical model. All the results presented in this chapter are included in our original research article [32]. The present exposition closely follows this last reference.

§3.1 Connection between the rank-size and the pdf representations

In chapter 2 we defined the rank of an observation as being proportional to the number of observations greater than it or, equivalently, to the probability of making a larger observation. This gave us the rank as function of size, $r = r(x)$; then we defined the rank-size function as the inverse $x = r^{-1}(r(x)) = x(r)$. From the definition of rank, we compute the associated pdf by differentiating eq.2.1 with respect to x, yielding $f(x) = \frac{1}{r_m - r_M}\frac{dr}{dx}$. By construction, $r = r(x)$ is a strictly decreasing function, so it is invertible on the interval where it is defined, $[r_m, r_M]$. We will now show that this relationship between the pdf and the rank-size representation is one-to-one.

Proposition 2. *The relationship between the pdf $f(x)$ and the rank-size function $x = x(r)$ established by*

$$\begin{aligned} r(x) &= r_m + (r_M - r_m)\int_x^{x_M} f(t)dt, \\ f(x) &= \frac{1}{r_m - r_M}\frac{dr}{dx} \end{aligned} \tag{3.1}$$

is one-to-one.

Proof. Lets start with two continuous random variables X_1 and X_2 with the same pdf $f(x)$ over $[x_m, x_M]$, where possible $x_M \longrightarrow \infty$. Let $x_{1,1} \leq x_{1,2} ... \leq x_{1,N}$ and $x_{2,1} \leq x_{2,2} ... \leq x_{2,N}$ be two ordered samples of these random variables, each with N observations. We can normalize the ranks of this observations to make $r(x_{1,1}) = r(x_{2,1}) = r_M$ and $r(x_{1,N}) = r(x_{2,N}) = r_m$. Since X_1 and X_2 have the same pdf, $P[X_1 \geq x] = P[X_2 \geq x]$ for all $x \in [x_m, x_M]$ and therefore $r_m + (r_M - r_m)P[X_1 \geq x] = r_m + (r_M - r_m)P[X_2 \geq x]$. By definition of rank, we see that $r(x_1) = r(x_2)$, and because $r(x)$ is by construction an invertible function on the interval $[x_m, x_M]$, we conclude that $x_1(r) = x_2(r)$ for all $r \in [r_m, r_M]$.

Now, let X_1 and X_2 be continuous random variables and $x_{1,1} \leq x_{1,2} ... \leq x_{1,N}$ and $x_{2,1} \leq x_{2,2} ... \leq x_{2,N}$ sets observations with the same rank-size function, $x_1(r) = x_2(r)$ for all $r \in [r_m, r_M]$. Then $x_1'(r) = x_2'(r)$ and, by means of the inverse function theorem, $\left(x_1^{-1}\right)'(x_1(r)) = \left(x_2^{-1}\right)'(x_2(r))$ for all $r \in [r_m, r_M]$. Therefore $\frac{1}{r_m - r_M}\frac{1}{\frac{dx_1}{dr}} = \frac{1}{r_m - r_M}\frac{1}{\frac{dx_2}{dr}}$, thus $f_1(x) = f_2(x)$. □

From now on we make the convention $r_m = 1$ and $r_M = N$. Suppose we have the DGBD rank-size function $x(r) = C\frac{(N+1-r)^b}{r^a}$ and we want to compute its corresponding pdf. According to our definitions, we would have to compute the inverse function $r(x)$, differentiate and normalize. We must start by solving

$$r = \left(\frac{C}{x}\right)^{\frac{1}{a}}(N + 1 - r)^{\frac{b}{a}},$$

and we come to a dead end because, by the Abbel-Ruffini theorem, we cannot solve this equation for r by radicals except for a few cases ($\frac{b}{a} = 1, 2, 3, 4, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$). Therefore it is impossible, in the general case, to give a closed-form of the pdf associated to the DGBD. This does not mean that such pdf does not exist, only that there is no way to express it in terms of simple functions. However, there are some cases where we are able to do it.

§3.2 The Lavalette distribution and its properties

There are seven cases where the pdf associated to the DGBD can be analytically derived. It is not our current purpose to study each of them in detail; notwithstanding, there is one where the resulting pdf yields a useful distribution function with interesting properties. This is the case $a = b$ where, as we already commented, the DGBD reduces to the Lavalette rank function $x(r) = C\left(\frac{N+1-r}{r}\right)^a$.

Proposition 3. *A random variable X whose rank-size function is the Lavalette rank function has the pdf*

$$f(x|a, C) = \frac{1}{aC} \frac{\left(\frac{x}{C}\right)^{\frac{1}{a}-1}}{\left(1 + \left(\frac{x}{C}\right)^{\frac{1}{a}}\right)^2}, \tag{3.2}$$

for $x \geq 0$.

Proof. Solving the Lavalette rank function for r we get $r = \frac{N+1}{1+\left(\frac{x}{C}\right)^{\frac{1}{a}}}$. Differentiating and multiplying by $\frac{1}{r_m - r_M} = \frac{1}{N-1}$ we get

$$\frac{1}{r_m - r_M} \frac{dr}{dx} = \frac{N+1}{N-1} \frac{1}{aC} \frac{\left(\frac{x}{C}\right)^{\frac{1}{a}-1}}{\left(1 + \left(\frac{x}{C}\right)^{\frac{1}{a}}\right)^2},$$

which converges to the right side of 3.2 when $N \longrightarrow \infty$. □

We say that a continuous and positive random variable X has a *Lavalette distribution* if its pdf is 3.2. Here $a, C \geq 0$ are parameters. The cumulative distribution function or cdf is $F(x|a, C) = 1 - \frac{1}{1+\left(\frac{x}{C}\right)^{\frac{1}{a}}}$; notice how $F(x|a, C) = F(x/C|a, 1)$, which means that C is a scale parameter. On the other hand, a is a shape parameter, which controls and affects the whole body of the distribution.

In fig.3.1 we show plots of the Lavalette density function for different values of its parameters: they all have identical $C = 1$ but $a = \frac{1}{3}, \frac{1}{5}$ (unimodal cases) and $a = 1, 2, 3, 4$ (in such cases, the pdf is monotonically decaying). We can analytically obtain the $i-$th moment of the distribution,

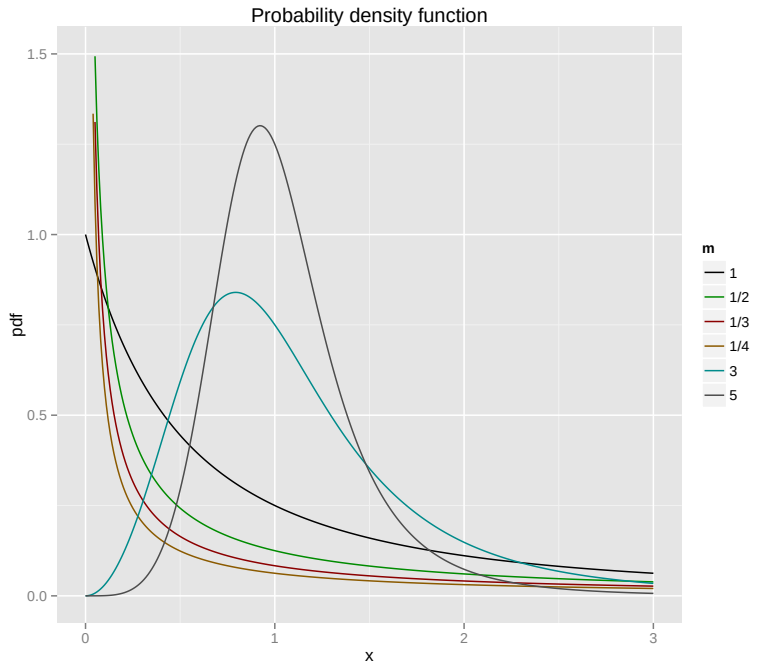


Figure 3.1: **Pdf of the Lavalette distribution.** Some Lavalette probability density functions with identical parameter $C = 1$ but with a=1/5, 1/3, 1, 2, 3, and 4 ($m = 1/a = 1/b$).

$$\begin{aligned} E[x^i] &= \int_0^\infty x^i p(x)dx = \frac{C^i}{a}\int_0^\infty \frac{(x_1)^{1/a-1+i}}{(1+(x_1)^{1/a})^2}dx_1 \\ &= iC^i\int_0^\infty \frac{(x_1)^{i-1}}{1+(x_1)^{1/a}}dx_1 \\ &= aiC^i\int_0^\infty \frac{(x_2)^{ai-1}}{1+x_2}dx_2 = \frac{aiC^i\pi}{\sin(ai\pi)}. \end{aligned}$$

Here, we made the change of variables $x_1 = x/C$ and $x_2 = x_1^{1/a}$. The integral converges provided $i < \frac{1}{a}$ (for the last step of this computation see for example [40]). In particular, the mean of a Lavalette random variable is

$$E[x] = \frac{\pi aC}{\sin(\pi a)},$$

which is finite if $a < 1$, while its variance is

$$Var[x] = \pi aC^2\left(\frac{2}{\sin(2a\pi)} - \frac{\pi a}{\sin^2(a\pi)}\right),$$

which exists and is finite if $a < \frac{1}{2}$. Notice how the spread of the distribution depends quadratically on the parameter C. It also comes to our attention that the $i-$th moment is finite only if $i < \frac{1}{a}$, indicating that the parameter a controls the weight of the tail: the larger value a has, the heavier is the tail of the Lavalette distribution.

§3.3 Resemblance between the Lavalette and the lognormal distributions

One distinctive feature of the DGBD function is its flexibility: the way in which it is defined and the two adjustable parameters make it capable of "resembling" other distributions. What happens when $a = b$, which is our current case of study? Is it behaving like some other distribution that we know? In order to investigate this, we generated data from 14 distributions (beta, binomial, χ^2, exponential, gamma, geometric, hypergeometric, lognormal, Mandelbrot, negative binomial, Pareto, Poisson, uniform and Weibull), and fitted the ranked data by the DGBD function via linear regression of the logarithmic transformation of eq.(2.2). The estimated parameter values for a and b are shown in fig.3.2. Remarkably, whenever we simulate lognormally distributed observations, choosing its parameters at random, we get a good fit to DGBD with $a \approx b$, that is, we get a good fit to the Lavalette rank function. This suggests that the lognormal and the Lavalette distribution may have similar appearances, akin behaviors and might be difficult to distinguish from one another, as is in fact the case. We will show this in a moment.

We use a novel argument from statistics to explain why the Lavalette and lognormal distributions may be difficult to distinguish within a certain interval of their domain. There are two models for the probability of a binary variable $y \in (0, 1)$: on the one hand there is the probit model [15]: $P = P(y = 1) = \Phi(z)$ where Φ is the cdf of standard normal distribution; on the other hand there is the logit model or logistic regression [66]: $P = 1/(1 + e^{-z})$. These two regression models for binary variables (regressed over an independent variable z) usually lead to similar results [4, 2], which can be written as (after the logistic variable being re-scaled by a factor α):

$$\Phi(z) \approx \frac{1}{1 + e^{-\alpha z}}, \quad or, \quad \frac{\Phi(z)}{1 - \Phi(z)} \approx e^{\alpha z}. \tag{3.3}$$

Here, the $\approx$ symbol stands for "approximately equal". The α can be $\sqrt{8/\pi} \approx 1.596$ to achieve the best fit near the midpoint [73], or ≈ 1.7 to best fit the whole range, or $\pi/\sqrt{3} \approx 1.81$ which is the standard deviation of the variable from the logistic distribution [2]. The standard normal variable can be converted to a lognormal distribution variable x: $z = (\log(x) - \mu)/\sigma$, and re-expressing eq.(3.3) in x becomes:

$$e^{\mu} \left(\frac{\Phi((\log(x) - \mu)/\sigma}{1 - \Phi((\log(x) - \mu)/\sigma)} \right)^{\sigma/\alpha} \approx x, \tag{3.4}$$

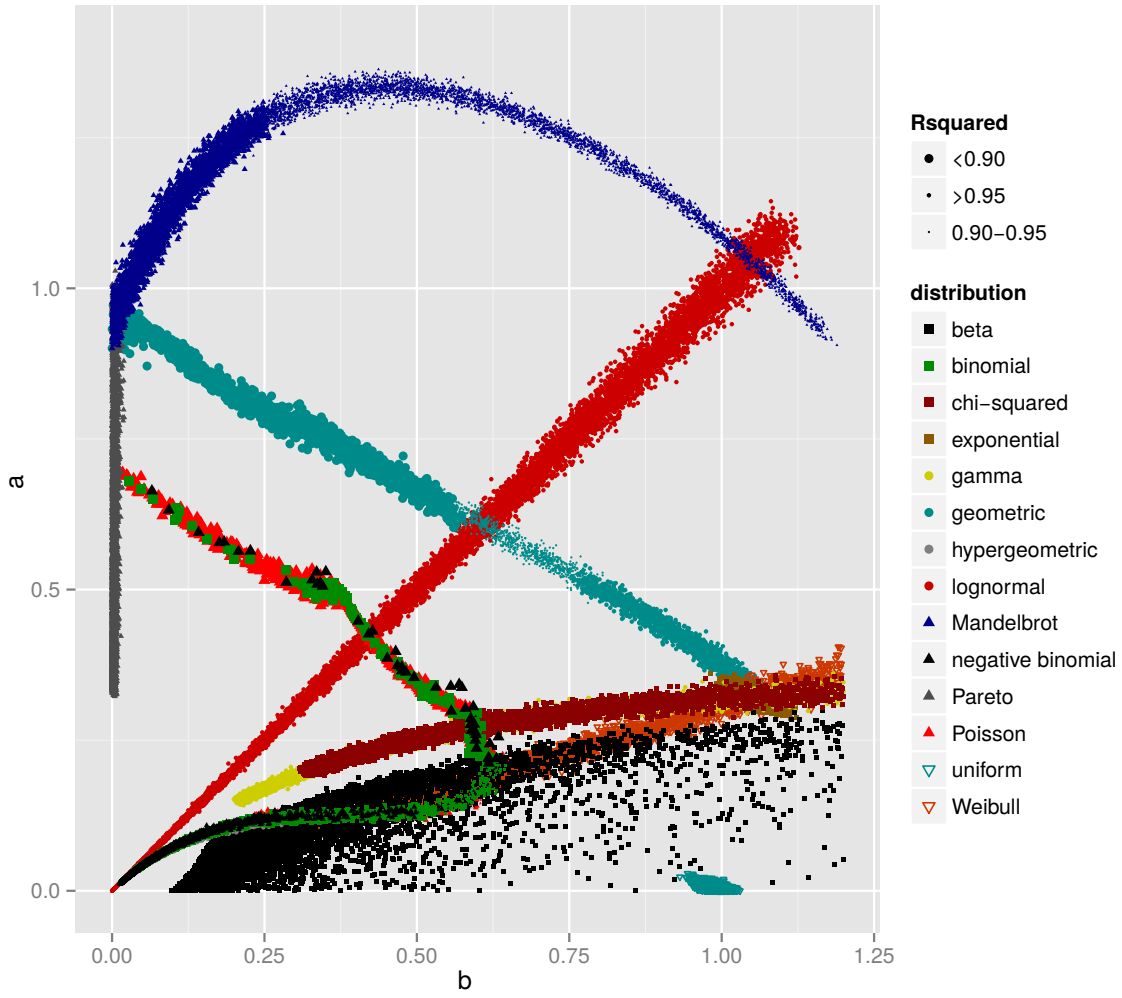


Figure 3.2: **DGBD fits for various distributions.** The estimated a and b parameter values in DGBD for data generated by well known distributions. Size of the dots indicate the coefficient of determination R squared. The dots around the $a = b$ diagonal line are for data generated by the lognormal distribution.

which we recognize as the Lavalette rank function over variable x (because $1-\Phi$ is proportional to the rank). This derivation also points out that the parameter $a = b$ is the standard deviation of the lognormal distribution divided by $\alpha(= 1.6 \sim 1.8)$, whereas the log-mean of the lognormal distribution is related to the scaling parameter by $C = e^{\mu}$. Since probit and logistic regression are not the same, we conclude that the Lavalette and the lognormal distributions cannot be identical. Indeed, the Lavalette and lognormal distributions have qualitatively different behaviors at the tails. All moments of the lognormal distribution exist, while the Lavalette has only finite moments of order $i < 1/a$, as we previously showed. If there is enough data to sample the tail, they cannot be mistaken into one another.

Nonetheless, the Lavalette and the lognormal distribution have very similar behaviors, specially at the center of the distribution, as we illustrate in in fig.3.3. The cdf's of lognormal distribution and the corresponding Lavalette distribution are plotted at three different parameter values ($\mu = 0$ with $\sigma = 0.1$, 0.5 and 1 for the lognormal, corresponding $C = e^{\mu}$, $a = \sigma\sqrt{3}/\pi$ for the Lavalette). Besides the difference at the tails (which is not visible from the cdf plot because the difference along the y-axis is very small for extreme values), the two functions also deviate slightly from each other in the middle range. This deviation is equivalent, after a transformation, to that between the cdf of standard normal distribution and logistic function. It has been proposed that a modification of the logistic function, $1/(1+e^{-1.5876x-0.070566x^3})$, is a very good approximation of the cdf of standard normal distribution [44, 73]. The small coefficient of the high-order term is another indication that the cdf of normal and logistic function, or equivalently, that the lognormal and the Lavalette distributions are close.

§3.4 Occurrence of the Lavalette distribution

The next step is to investigate whether the Lavalette distributions occurs in real phenomena or of it is a good statistical model for actual data. We examined several datasets, besides the impact factor and citation data used in [47, 80], and found some examples where this novel distribution can be applied. These examples comprehend three different phenomena: population data, amino acid mutation rates and codon usage data. The parameters were estimated through linear regression of the logarithmic transformation of eq.(1.1), which in our case gives very similar results to maximum likelihood estimators. In the following section we will discuss our goodness of fit tests, which closely resemble those delineated in the previous chapter.

The first set of examples is about administrative units of population, which were widely discussed in the previous chapter. From the extensive sample that we studied in chapter 2, we found three examples where the Lavalette is a good model: second level administrative units (SAU) in Nigeria and third level administrative units (TAU) in the Spanish provinces of Madrid and Cádiz (notice that TAU's were not part of our former studies). Fig.3.4(A,B) shows the rank-size plot of the

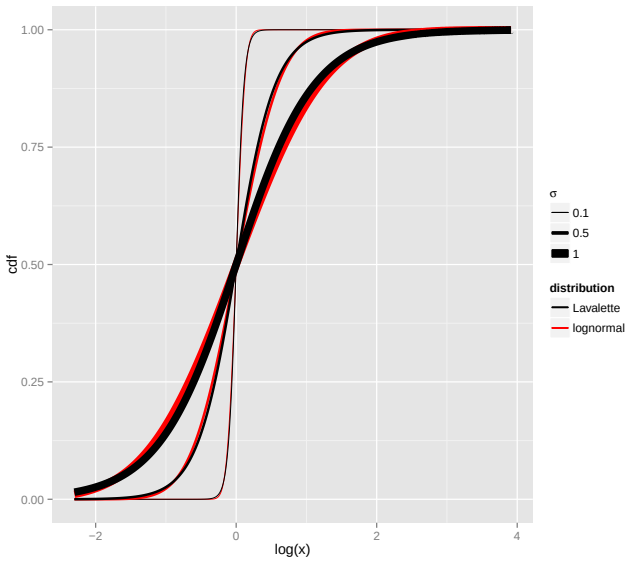


Figure 3.3: **Lognormal vs Lavalette cdf.** Cumulative distribution function for lognormal and Lavalette distributions, being $\mu = 0$ and $\sigma = 0.1, 0.5$ and 1 the parameters of the lognormal. The x axis is in logarithmic scale. We see that over an important interval of the domain, it may be difficult to distinguish a lognormal from a Lavalette distribution.

NRG/SAU, Madrid/TAU and Cádiz/TAU population in logarithmic scale. The fitted parameter values (a, b) by eq.(2.2) are (0.275, 0.255) for NRG/SAU, (1.249, 1.049) for Madrid/TAU, (0.901, 0.906) for Cádiz/TAU, all with $a \approx b$.

The second example are the amino acid mutation rates [24] based on the amino acid changing (missense) variants in the 1000 Genomes Project [1]. A missense mutation is a point mutation which results in the codification of a different amino acid. Because the variants are observed in normal human population with a short evolutionary history, it can be considered as an instantaneous mutation rate. The substitution rate between different species, such as the point accepted mutation (PAM) [23], cover a much longer evolutionary history with stronger selection constraints. Out of 380 ($= 20 \times 19$) possible mutations between 20 amino acids, only $N = 150$ are allowed from the single base mutation in the DNA sequence, due to the nature of the genetic code. Fig.3.4(C) shows the ranked amino acid to amino acid frequencies derived from the missense variants in DNA sequence of the 1000 Genomes Project. Fig.3.4(C) shows fittings by the DGBD with $a \approx 0.650$ and $b \approx 0.615$, which suggests again a good Lavalette function.

The third example is the codon usage of $N = 61$ non-stop codons, with data from the Codon Usage Database [69]. Codon usage refers to the frequency of occurrence each type of codon within a DNA sequence. We picked the two examples best demonstrating Lavalette function: genes in plant organelles (9221 species) and in non-primate, non-rodent mammalian nucleus (433 species). The codon frequencies are averaged over all species in plant organelle and mammalian separately. The

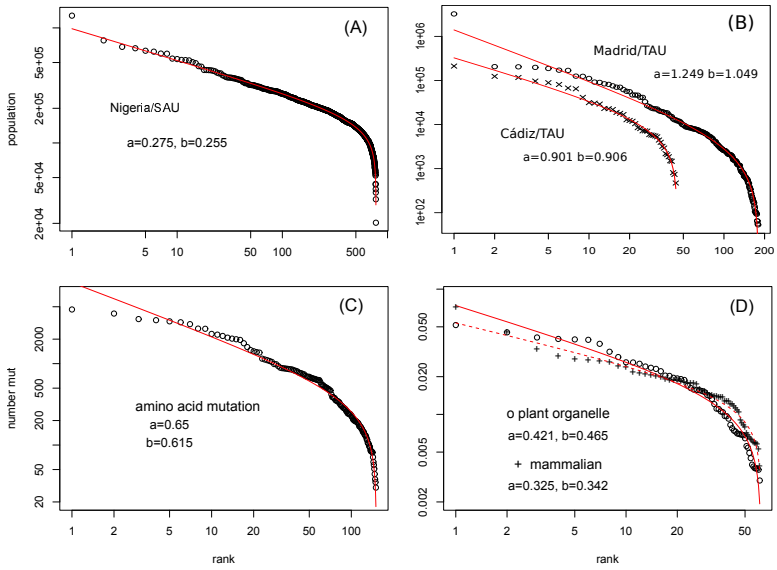


Figure 3.4: **Ranked datasets fitted by Lavalette rank function.** (A) Nigeria (NRG) local government area (the secondary administrative unit (SAU)) population; (B) Madrid and Cádiz municipality (the tertiary administrative unit (TAU)) population; (C) Amino acid to amino acid mutation counts in the 1000 Genomes Project; (D) Averaged codon usage (excluding the three stop codons) of plant organelles and mammals.

three stop codons are discarded. The (a, b) are (0.422, 0.465) for plant organelles, and (0.325,0.342) for mammalian (fig.3.4(D)).

With the previous examples we have illustrated the occurrence of the Lavalette distribution. Finally, we will expose our statistical criterion to discern if this distribution is consistent with the data.

§3.5 Goodness of fit tests

The first clue that a certain dataset may be well described by the Lavalette distribution is to fit the data to the DGBD, estimate the parameters and check if $a \approx b$. If this is the case, the data set is a candidate for the Lavalette distribution. This is a first criterion and it serves to rule out many datasets; however, it is by no means strong statistical evidence to claim the the Lavalette is a good model for the data.

To test more rigorously whether a Lavalette function fits the observed data well, we use the resampling approach to the K-S test explained in the previous chapter, which can also be called a *bootstrap* [28]. We first fit the data by the Lavalette function. The difference between the

observed and fitted value is measured by the Kolmogorov-Smirnov distance. Using the fitted Lavalette distribution, artificial data (replicates) are generated multiple times: each time a new Lavalette distribution is fitted and K-S distance calculated. The proportion of replicates with larger K-S distances than the observed one is the empirical p-value. A large empirical p-value indicates that there is not enough evidence to reject the Lavalette function. Empirical p-values from 1 000 replicates are 0.49 for NRG/SAU, 0.91 for Madrid/TAU, 0.88 for Cádiz/TAU, 0.06 for mutation rate, and 0.4 for codon usage in both plant organelle and mammals. The empirical p-value we have indicate that the Lavalette distribution is a good fitting model for these data.

There have been debates in the literature whether power-laws results from the central limit theorem [75, 87, 67]. Given a data set, a possible answer to that debate would be to pick the better fitting model between power-law and lognormal distribution [60]. The approximate equivalence between Lavalette distribution and lognormal distribution provides us with a simple method in deciding if a set of data follows a power-law or a lognormal distribution. For the fitting of ranked data by the DGBD, if $b \approx 0$, the power-law is better; if $a \approx b > 0$, lognormal is better; and if $a \neq b \gg 0$, neither are good fitting functions.

For our examples to illustrate the Lavalette distribution in real data, it is obvious that lognormal distribution is a better fitting function than the power-law. We can further quantify the fitting performance by model selection techniques such as Akaike information criterion, which we also utilized in the previous chapter [18, 3, 50]. We will prefer the model that exhibits the lower AIC value. The $\mathrm{AIC}_{\mathrm{lav}}$-$\mathrm{AIC}_{\text{power-law}} = N \log(SSE_{\mathrm{lav}}/SSE_{\text{power-law}})$ [54], where SSE is the sum squared error, is -3284.6,-410.3, -108.1 for the NRG/SAU, Madrid/TAU, Cádiz/TAU data, -353.7 for the amino acid mutate data, and -101.7, -114.7 for plant organelle, mammalian codon frequencies, all representing an overwhelming support to the Lavalette function over the power-law.

Chapter 4

An Inquiry on its Origins: the $\mathcal{J}$ Family of Functions

If only I had the Theorems! Then I should find the proofs easily enough.

Bernhard Riemann

§4.1 Motivation of the problem

The DGBD is a good function to model a behavior that is observed under many different circumstances: a power-law at the high-rank regime followed by a decay at the low-rank region. Which kind of mechanisms could produce such a widely extended phenomenon? Look for instance at the universality of the Gaussian distribution; this can be explained by means of the Central Limit Theorem, a remarkable result in Probability Theory which establishes that the normalized sum of finite-variance and independent random variables converges in probability to the normal distribution [30]. Thus, this distribution emerges for very general and unspecific reasons, when many random effects sum up.

What happens with the ubiquity of power-laws? While some authors argue that they appear everywhere because there are many possible mechanisms leading to them [71], it is also true that power-laws correspond to Lévy distributions [84]; according to the general version of the Central Limit Theorem, given originally by Paul Lévy and later by Kolmogorov and Gnedenko, these are limit distributions of the sum of infinitely many random variables with non-finite moments [38]. Thus, one expects power-laws to appear just as naturally as Gaussian distributions. The Gaussian and the Lévy laws together conform the stable family of probability distributions, which is closed under the convolution operator that adds random variables. In fact, this closedness property is a necessary and sufficient condition for a family of distributions to be the limit of sums of random variables [72, 92, 88].

What can we say about the behavior that DGBD depicts? It could be that they are all power-laws with a decay at the high-rank regime explained by finite size effects. It could also be that some phenomenon well described by DGBD are not scale-free, not even in principle, and there are different dynamics conducting the phenomenon at different scales [70]. Look for instance at the phenomenon of turbulence, where Kolmogorov's power-law is observed at the big scale regime, followed by a decay at the small regime [38]. Here, energy injection dominates at the high-energy scale, while energy dissipation dominates at the low-energy region. Similarly, maybe there are many different dynamics leading to the power-law-followed-by-decay behavior, or perhaps there is a very generic collective phenomenon explaining its emergence. Up to our knowledge, these questions have been not properly answered.

In chapter 1 we mentioned the different research paths there are in this direction. In this work we explore the following question, originally proposed by G. Cocho: are there any generic functions or probability distributions resulting from a phenomenon of subtracting random effects? If so, are they in any way related to the power-law-followed-by-decay behavior? We remark that, according to numerical evidence provided in [25], this could be a promising research direction.

Our main result of this chapter is the construction and characterization of a very broad family of complex-valued functions that is closed under an integral operator that, when restricted to real, positive and normalized functions, gives the pdf of the subtraction of two independent random variables. As an application of this very general result, we give a random dynamics that lead to the Lavalette distribution, which is a particular case of the probability distribution associated to the DGBD. We will start by defining and showing some properties of our integral operator, which we call the pseudo cross-correlation. Then we will need to give some theory about Meijer's G function, which is a generalization of the generalized hypergeoemtric function with many applications in mathematical physics; following this, we construct a very wide family of functions, which we call the $\mathcal{J}$ family, and prove its closedness under the pseudo cross-correlation operator. Finally, we give specific examples of our main results from which the Lavalette distribution emerges. With this we contribute to the understanding of the possible mechanisms giving rise to the pseudo cross-correlation phenomenon and the DGBD.

§4.2 The pseudo cross-correlation operator

Lets start by computing the pdf of the subtraction of two independent and continuous random variables.

Proposition 4. *Let X_1 and X_2 be continuous and independent random variables with probability densities f_1 and f_2. Then the random variable $X_2 - X_1$ has the pdf*

$$f_{X_2-X_1}(x) = \int_{-\infty}^{\infty} f_1(y) f_2(x+y) dy. \tag{4.1}$$

Proof. Because of the independence assumption, we can write the joint pdf of X_1 and X_2 as $f_{X_1,X_2}(u,v) = f_1(u)f_2(v)$. We compute the cumulative distribution function of the subtraction $X_2 - X_1$,

$$F_{X_2-X_1}(x) = P[X_2 - X_1 \leq x] = \int\int_D f_1(u)f_2(v)dudv,$$

where $D = \{(u,v) \in \mathbb{R}^2 | v - u \leq x\}$. By making a change of variable and using Tonelli's theorem, we derive

$$\begin{aligned} F_{X_2-X_1}(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{u+x} f_2(v)f_1(u)dvdu \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_2(u+v)f_1(u)dvdu \\ &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_1(u)f_2(u+v)dudv. \end{aligned}$$

Differentiating this last expression we get the desired result. □

The fact that $f_{X_2-X_1}(x)$ integrates one is a direct consequence of Fubinis theorem,

$$\begin{aligned} \int_{-\infty}^{\infty} f_{X_2-X_1}(x)dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(y)f_2(x+y)dydx \\ &= \int_{-\infty}^{\infty} f_1(y)\int_{-\infty}^{\infty} f_2(x+y)dxdy \\ &= 1. \end{aligned}$$

We also notice that $f_{X_2-X_1}(x)$ is everywhere nonnegative, so it is a probability density function. The right side of eq. 4.1 defines an integral operation between two functions f_1 and f_2. This operation does not need its arguments to be pdf's (real valued, nonnegative and normalized) in order to be defined. It will be useful to our purposes to define this operation for complex valued functions,

Definition 3. *We define the pseudo cross correlation between two complex functions f_1 and f_2 as*

$$f_2 \ominus f_1(x) = \int_{-\infty}^{\infty} f_1(y)f_2(x+y)dy. \tag{4.2}$$

This operation resembles the cross-correlation operator, hence its name pseudo cross-correlation, which is widely used in signal analysis [89]. Lets recall some properties of this integral operation. First, the cross-correlation of two complex functions f_2 and f_1, denoted $f_2 \star f_1$, is defined by

$$f_2 \star f_1(x) = \int_{-\infty}^{\infty} f_2^*(y)f_1(x+y)dy. \tag{4.3}$$

It is well known this integral operation satisfies the following properties:

- If both f_1 and f_2 are Hermitian, then $f_1 \star f_2 = f_2 \star f_1$.
- If f_2 is Hermitian, then $f_2 \star f_1 = f_2 * f_1$, where $*$ denotes the convolution operator.
- $(f_2 \star f_1) \star (f_2 \star f_1) = (f_2 \star f_2) \star (f_1 \star f_1)$.

Unlike the cross-correlation, the pseudo cross-correlation does not take complex conjugate inside the integral. We also see that the order in which the functions f_1 and f_2 appear inside the integral is reversed. Thus, if f_1 is a real valued function, then $f_2 \ominus f_1 = f_1 \star f_2$. Like the cross-correlation, the pseudo cross-correlation is not (in general) a commutative operator. We notice that if both f_1 and f_2 are even, or if both are odd functions, then $f_2 \ominus f_1 = f_1 \ominus f_2$, which can be easily verified by changing the sign of the integration variable in the definition of $\ominus$. From the definitions of the pseudo and the cross-correlation operators we also see that

$$f_2 \ominus f_1 = f_1^* \star f_2 \quad \text{and} \quad f_2 \star f_1 = f_1 \ominus f_2^*. \tag{4.4}$$

The associative property of the $\star$ operator, together with the relationships 4.4, imply that if both f_2 and f_1 are real valued functions, then $(f_2 \ominus f_1) \ominus (f_2 \ominus f_1) = (f_2 \ominus f_2) \ominus (f_1 \ominus f_1)$. We summarize these properties in the following

Proposition 5. *The pseudo correlation operator $\ominus$ satisfies the following properties:*

1. *If both f_1 and f_2 are even or both are odd functions, then the operator is commutative $f_2 \ominus f_1 = f_1 \ominus f_2$.*
2. *If both f_1 and f_2 are real valued, then $(f_2 \ominus f_1) \ominus (f_2 \ominus f_1) = (f_2 \ominus f_2) \ominus (f_1 \ominus f_1)$,.*

The empirical importance of the pseudo cross-correlation comes from the fact that if f_1 and f_2 are probability density functions (pdf) of independent random variables X_1 and X_2 respectively, then $f_2 \ominus f_1$ is the pdf of the subtraction $X_2 - X_1$, as we proved at the beginning of this section.

At this point we introduce the two main questions that motivated the present investigation: *1) which is the general class of functions that is closed under the pseudo cross-correlation operator? (by "closed" we mean that if two functions belong to the same family, so does its pseudo cross-correlation.) 2) Does this tell us something about the origins of the DGBD?.*

Although we did not characterize the general class of functions that is closed under this operation, we gave a step toward the solution by giving a very broad family with the desired property. Before we introduce it, we need to recall some theory about another very general class of functions, called Meijer's G-functions.

§4.3 Meijer's G function

The Meijer G function is a very general function of one variable that serves as a generalization of the hypergeometric function [12]. Due to the great freedom in its parameters, many special functions in physics can be represented in terms of this functions and the analysis of many problems is much simplified. For these reasons, this function has found applications in numerous fields such as statistic distributions [82], nuclear physics [78], theoretical and mathematical physics [79], fractional calculus [45], symbolic integration [16], etc.

For the next discussions z will represent a complex-valued variable and x and y real-valued ones. Meijer's G function of order (m, n, p, q), where $m \leq q$, $n \leq p$ and $p + q \leq 2(m + n)$ is a function defined by the Mellin-Barnes integral

$$G_{p,q}^{m,n}\left(z \left| \begin{array}{c} \mathbf{a_p} \\ \mathbf{b_q} \end{array} \right.\right) = \frac{1}{2\pi i}\int_L z^{-s}\frac{\prod_{j=1}^{m}\Gamma(b_j+s)\prod_{j=1}^{n}\Gamma(1-a_j-s)}{\prod_{j=n+1}^{p}\Gamma(a_j+s)\prod_{j=m+1}^{q}\Gamma(1-b_j-s)}ds.$$

The complex parameters $\mathbf{b_q} = (b_1, ..., b_q)$ and $\mathbf{a_p} = (a_1, ..., a_p)$ are such that no pole of $\Gamma(b_j + s)$, $j = 1, ..., m$ coincides with any pole of $\Gamma(1 - a_k - s)$, $k = 1, ..., n$; L is a contour separating the poles of $\Gamma(b_j + s)$, $j = 1, ..., m$ from the poles of $\Gamma(1 - a_k - s)$, $k = 1, ..., n$ [29, 65]; whenever an empty product occurs, we take it equal to 1. Situations for which $G(z)$ exists are the following [29, 65]:

i) $q \geq 1$, $q > p$, for all z, $z \neq 0$.

ii) $q \geq 1$, $q = p$, for $|z| < 1$.

iii) $p \geq 1$, $p > q$, for all z, $z \neq 0$.

iv) $p \geq 1$, $p = q$, for $|z| > 1$.

If no confusion can occur, we denote $G_{p,q}^{m,n}\left(z \left| \begin{array}{c} \mathbf{a_p} \\ \mathbf{b_q} \end{array} \right.\right)$ simply by $G(z)$. $G(z)$ is an analytic function of the complex variable z with a branch at the origin [57]. For the results in the next section we will need the following properties of the G function. Proofs of these results are found in the references [57, 65].

Lemma 1. *The G function satisfies the property*

$$z^{\rho}G_{p,q}^{m,n}\left(z \left| \begin{array}{c} \mathbf{a_p} \\ \mathbf{b_q} \end{array} \right.\right) = G_{p,q}^{m,n}\left(z \left| \begin{array}{c} \mathbf{a_p}+\rho \\ \mathbf{b_q}+\rho \end{array} \right.\right),$$

where ρ is any complex constant.

A property of the G function of fundamental importance in its applications is the next result, known as the multiplication convolution property of the G function,

Lemma 2. *The product of two arbitrary real-valued G functions integrated over the positive domain can be represented as another G function,*

$$\int_0^\infty G_{p,q}^{m,n}\left(\eta x \,\middle|\, \begin{matrix}\mathbf{a_p}\\ \mathbf{b_q}\end{matrix}\right) G_{\sigma,\tau}^{\mu,\nu}\left(\omega x \,\middle|\, \begin{matrix}\mathbf{c}_\sigma\\ \mathbf{d}_\tau\end{matrix}\right) dx =$$

$$\frac{1}{\eta} G_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(\frac{\omega}{\eta} \,\middle|\, \begin{matrix}-b_1,...,-b_m,\mathbf{c}_\sigma,-b_{m+1},...,-b_q\\ -a_1,...,-a_n,\mathbf{d}_\tau,-a_{n+1},...,-a_p\end{matrix}\right),$$

under very general conditions (explicitly enumerated in [57]).

Finally, we will use the following special cases of the G function. They are deduced immediately after the definition and can be found in [65],

Lemma 3. *The following holds,*

i)

$$G_{0,1}^{1,0}\left(x \,\middle|\, \begin{matrix}-\\ a\end{matrix}\right) = x^a e^x.$$

ii)

$$\Gamma(a) G_{1,1}^{1,1}\left(x \,\middle|\, \begin{matrix}1-a\\ b\end{matrix}\right) = \frac{x^b}{(1+x)^{(a+b)}}.$$

§4.4 The $\mathcal{J}$ family of functions

We define the J function of order (m,n,p,q) and complex parameters $\mathbf{b_q} = (b_1,...,b_q)$, $\mathbf{a_p} = (a_1,...,a_p)$, where $m \leq q$, $n \leq p$ and $p+q \leq 2(m+n)$ as the composition of the G function with the exponential,

$$J_{p,q}^{m,n}\left(z \,\middle|\, \begin{matrix}\mathbf{a_p}\\ \mathbf{b_q}\end{matrix}\right) = G_{p,q}^{m,n}\left(e^z \,\middle|\, \begin{matrix}\mathbf{a_p}\\ \mathbf{b_q}\end{matrix}\right) = \frac{1}{2\pi i}\int_L e^{-zs} \frac{\prod_{j=1}^m \Gamma(b_j+s)\prod_{j=1}^n \Gamma(1-a_j-s)}{\prod_{j=n+1}^p \Gamma(a_j+s)\prod_{j=m+1}^q \Gamma(1-b_j-s)} ds. \tag{4.5}$$

Since we have not made changes in the variable of integration from the definition of the G function, the contour of integration L is the same, that is, a contour that separates the poles of $\Gamma(b_j+s)$, $j=1,...,m$ from those of $\Gamma(1-a_k-s)$, $k=1,...,n$. The parameters must be such that no pole of $\Gamma(b_j+s)$, $j=1,...,m$ coincides with any pole of $\Gamma(1-a_k-s)$, $k=1,...,n$. Conditions for which the integral converges and the J function exists are the following:

i) $q \geq 1$, $q > p$, for all x.

ii) $q \geq 1$, $q = p$, for $\Re x < 0$.

iii) $p \geq 1$, $p > q$, for all x.

iv) $p \geq 1$, $p = q$, for $\Re x > 0$.

We define the $\mathcal{J}$ family of $J-$ functions as the set of all complex functions that can be represented by 4.5. Now we will give an example of functions in this family.

Proposition 6. *Let Y be a gamma random variable with pdf $f_Y(y) = \frac{1}{\Gamma(\theta)} y^{\theta-1} e^{-y}$ and $\theta > 0$. Then the random variable $X = \log Y$ has a pdf that belongs to the $\mathcal{J}$ family.*

Proof. By means of the change of variable theorem $f_X(x) = f_Y(y(x))|\frac{dy}{dx}|$, the r.v. X has the pdf

$$f_X(x) = \frac{1}{\Gamma(\theta)} e^{\theta x} e^{-e^x}. \tag{4.6}$$

The first point of lemma implies that $f_X(x) = J_{0,1}^{1,0}\left(x \left| \begin{array}{c} - \\ \theta \end{array} \right. \right)$ and thus belongs to the $\mathcal{J}$ family. □

In analogy to the lognormal distribution, whose logarithm is normally distributed, we will say that the r.v. X has an *exp-gamma* distribution if it has the pdf 1 (its support is the complete real axis).

Now we establish our main result: if J_1 and J_2 belong to the $\mathcal{J}$ family, so does $J_2 \ominus J_1$. After this result, we say that the $\mathcal{J}$ family is closed under the pseudo cross correlation.

Proposition 7. *If J_1 and J_2 belong to the $\mathcal{J}$ family, then the pseudo cross correlation $J_2 \ominus J_1$ belongs to $\mathcal{J}$. Specifically,*

$$J_{\sigma,\tau}^{\mu,\nu}\left(\cdot \left| \begin{array}{c} \mathbf{c}_\sigma \\ \mathbf{d}_\tau \end{array} \right. \right) \ominus J_{p,q}^{m,n}\left(\cdot \left| \begin{array}{c} \mathbf{a_p} \\ \mathbf{b_q} \end{array} \right. \right)(x) =$$

$$J_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(x \left| \begin{array}{c} -b_1+1,...,-b_m+1,\mathbf{c}_\sigma,-b_{m+1}+1,...,-b_q+1 \\ -a_1+1,...,-a_n+1,\mathbf{d}_\tau,-a_{n+1}+1,...,-a_p+1 \end{array} \right. \right),$$

under the same circumstances specified in theorem 2, taken over the variable e^x. Using integral notation,

$$\begin{aligned} &\int_{-\infty}^{\infty} J_{p,q}^{m,n}\left(y \left| \begin{array}{c} \mathbf{a_p} \\ \mathbf{b_q} \end{array} \right. \right) J_{\sigma,\tau}^{\mu,\nu}\left(x+y \left| \begin{array}{c} \mathbf{c}_\sigma \\ \mathbf{d}_\tau \end{array} \right. \right) dy = \\ &J_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(x \left| \begin{array}{c} -b_1+1,...,-b_m+1,\mathbf{c}_\sigma,-b_{m+1}+1,...,-b_q+1 \\ -a_1+1,...,-a_n+1,\mathbf{d}_\tau,-a_{n+1}+1,...,-a_p+1 \end{array} \right. \right). \end{aligned} \tag{4.7}$$

Proof. First we take the left side of (4.7) and take the definition of the $\mathcal{J}$ functions, writing the integrand as a product of Gs.

$$\int_{-\infty}^{\infty} J_{p,q}^{m,n}\left(y \,\middle|\, \begin{matrix}\mathbf{a_p}\\ \mathbf{b_q}\end{matrix}\right) J_{\sigma,\tau}^{\mu,\nu}\left(x+y \,\middle|\, \begin{matrix}\mathbf{c}_\sigma\\ \mathbf{d}_\tau\end{matrix}\right) dy = \int_{-\infty}^{\infty} G_{p,q}^{m,n}\left(e^y \,\middle|\, \begin{matrix}\mathbf{a_p}\\ \mathbf{b_q}\end{matrix}\right) G_{\sigma,\tau}^{\mu,\nu}\left(e^{x+y} \,\middle|\, \begin{matrix}\mathbf{c}_\sigma\\ \mathbf{d}_\tau\end{matrix}\right) dy.$$

The next step is to use lemma 1 to take $\frac{1}{2}$ powers out of each G function and change the variable of integration to $\log y$, yielding

$$\int_{-\infty}^{\infty} e^{\frac{y}{2}} G_{p,q}^{m,n}\left(e^y \,\middle|\, \begin{matrix}\mathbf{a_p}-\frac{1}{2}\\ \mathbf{b_q}-\frac{1}{2}\end{matrix}\right) e^{\frac{x+y}{2}} G_{\sigma,\tau}^{\mu,\nu}\left(e^{x+y} \,\middle|\, \begin{matrix}\mathbf{c}_\sigma-\frac{1}{2}\\ \mathbf{d}_\tau-\frac{1}{2}\end{matrix}\right) dy =$$

$$\int_0^{\infty} y^{\frac{1}{2}} y^{\frac{1}{2}} x'^{\frac{1}{2}} G_{p,q}^{m,n}\left(y \,\middle|\, \begin{matrix}\mathbf{a_p}-\frac{1}{2}\\ \mathbf{b_q}-\frac{1}{2}\end{matrix}\right) G_{\sigma,\tau}^{\mu,\nu}\left(x'y \,\middle|\, \begin{matrix}\mathbf{c}_\sigma-\frac{1}{2}\\ \mathbf{d}_\tau-\frac{1}{2}\end{matrix}\right) \frac{dy}{y},$$

where we have made the change of variable $x' = \log x$. By means of prop. (2), we can express the multiplicative convolution of the two G functions as another G function,

$$x'^{\frac{1}{2}} \int_0^{\infty} G_{p,q}^{m,n}\left(y \,\middle|\, \begin{matrix}\mathbf{a_p}-\frac{1}{2}\\ \mathbf{b_q}-\frac{1}{2}\end{matrix}\right) G_{\sigma,\tau}^{\mu,\nu}\left(x'y \,\middle|\, \begin{matrix}\mathbf{c}_\sigma-\frac{1}{2}\\ \mathbf{d}_\tau-\frac{1}{2}\end{matrix}\right) dy =$$

$$x'^{\frac{1}{2}} G_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(x' \,\middle|\, \begin{matrix}-b_1+\frac{1}{2},\ldots,-b_m+\frac{1}{2},\mathbf{c}_\sigma-\frac{1}{2},-b_{m+1}+\frac{1}{2},\ldots,-b_q+\frac{1}{2}\\ -a_1+\frac{1}{2},\ldots,-a_n+\frac{1}{2},\mathbf{d}_\tau-\frac{1}{2},-a_{n+1}+\frac{1}{2},\ldots,-a_p+\frac{1}{2}\end{matrix}\right).$$

We use again eq. (1) and the definition of the J function to finally arrive to the right side of eq. (4.7).

$$G_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(x' \,\middle|\, \begin{matrix}-b_1+1,\ldots,-b_m+1,\mathbf{c}_\sigma,-b_{m+1}+1,\ldots,-b_q+1\\ -a_1+1,\ldots,-a_n+1,\mathbf{d}_\tau,-a_{n+1}+1,\ldots,-a_p+1\end{matrix}\right) =$$

$$G_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(e^x \,\middle|\, \begin{matrix}-b_1+1,\ldots,-b_m+1,\mathbf{c}_\sigma,-b_{m+1}+1,\ldots,-b_q+1\\ -a_1+1,\ldots,-a_n+1,\mathbf{d}_\tau,-a_{n+1}+1,\ldots,-a_p+1\end{matrix}\right) =$$

$$J_{q+\sigma,p+\tau}^{n+\mu,m+\nu}\left(x \,\middle|\, \begin{matrix}-b_1+1,\ldots,-b_m+1,\mathbf{c}_\sigma,-b_{m+1}+1,\ldots,-b_q+1\\ -a_1+1,\ldots,-a_n+1,\mathbf{d}_\tau,-a_{n+1}+1,\ldots,-a_p+1\end{matrix}\right).$$

□

Thus, we have proved the closedness of the $\mathcal{J}$ family under the $\ominus$ operator. How does this gives us light on the origins of the DGBD, which was our original motivation?

§4.5 Application: the $\mathcal{J}$ family and its relationship with DGBD

The previous is a very general result, since the freedom we have to choose the parameters of the J function is very broad. Also, the Js are in general complex valued functions. Our original motivation was a problem from Probability Theory: which families of random variables are closed under subtraction? The $f_2 \ominus f_1$ operation gives the pdf of the subtraction of two random variables, provided of course f_1 and f_2 are pdfs. Our result is much more extended, since the $\mathcal{J}$ family contains many functions that are not pdfs. However, we found an specific example that helps us understand the possible dynamics that make the Lavalette distribution emerge. As we previously discussed, this distribution is a particular case of the pdf related to DGBD. The following result is a direct consequence of prop. 7.

Corollary 1. *If* $f(x|\theta) = e^{\theta x} Exp[-e^x]$, $x \in (-\infty, +\infty)$, *then the correlation* $f(\cdot|b) \ominus f(\cdot|a)(x)$ *equals* $\Gamma(a+b)\frac{e^{bx}}{(1+e^x)^{a+b}}$, *that is*

$$e^{bx}e^{-e^x} \ominus e^{ax}e^{-e^x} = \Gamma(a+b)\frac{e^{bx}}{(1+e^x)^{a+b}}, \tag{4.8}$$

provided $Re(e^x) > 1$ *and* $Re(a+b) > 0$.

Proof. According to the first point of lemma 3 and the definition of J,

$$e^{\theta x}e^{-e^x} = J^{1,0}_{0,1}\left(x \middle| \begin{matrix} - \\ \theta \end{matrix}\right),$$

so we immediately see from proposition 7 that

$$J^{1,0}_{0,1}\left(\cdot \middle| \begin{matrix} - \\ b \end{matrix}\right) \ominus J^{1,0}_{0,1}\left(\cdot \middle| \begin{matrix} - \\ a \end{matrix}\right) = J^{1,1}_{1,1}\left(x \middle| \begin{matrix} 1-a \\ b \end{matrix}\right),$$

which we can express, in the light of the second point of lemma 3, as the right side of eq. (4.8), yielding the result. □

Previously, we studied the properties and applications of a novel continuous random variable whose distribution we called the *Lavalette distribution* and whose pdf we recall,

$$f(x) = \frac{\mu}{C}\frac{\left(\frac{x}{C}\right)^{\mu-1}}{\left(1+\left(\frac{x}{C}\right)^{\mu}\right)^2}, \tag{4.9}$$

defined over the semipositive axis and where $\mu, C > 0$ are parameters. We proved that this distribution fits satisfactorily data coming from countries internal population, codon frequencies, etc.

Proposition 8. *The difference of two independent an identically distributed random variables* X_2 *and* X_1, *following an exp-gamma distribution with parameter* $\theta = 1$ *is a continuous random variable* $X = X_2 - X_1$ *whose exponential transformation* $Y = Ce^{\frac{x}{\mu}}$ *is Lavalette distributed.*

Proof. Let $f(x|\theta)$ be the pdf of an exp-gamma r.v. By using corollary 1 we see that

$$f(\cdot|1) \ominus f(\cdot|1) = \frac{e^x}{(1+e^x)^2}.$$

The right side of this equation is the pdf of $X = X_2 - X_1$. Now we take the variable $Y = Ce^{\frac{x}{\mu}}$. By using change of variables theorem, we get

$$f_Y(y) = \frac{\left(\frac{y}{c}\right)^{\mu-1}}{\left(1+\left(\frac{y}{c}\right)^{\mu}\right)^2}.$$

□

In order to illustrate this result, we generated two sets of $n = 10\,000$ observations X_1 and X_2 of gamma distributed random variables, both with shape parameter $\theta = 1$. Then we generated the set $X_3 = C * e^{\frac{(\log X_2 - \log X_1)}{\mu}}$, taking $C = 1$ and $\mu = 0.30$ for this particular example. According to proposition 8, X_3 ought to be Lavalette-distributed with parameters $C = 1$ and $\mu = 3$. First we fitted X_3 to a DGBD and got $b \approx 0.34$ and $a \approx 0.32$, which is consistent with the Lavalette distribution. Then we fitted X_3 to this distribution and estimated its parameters via MLE, getting $\hat{m} = 0.33$ and $\hat{C} = 0.99$, both in agreement with our result. We tested the goodness of fit of the Lavalette distribution for this set of data by using the resampling approach to the K-S test that we described in chapter 3, getting a p_{val} of ~ 0.61, also in agreement with our result. This same test rejects the lognormal distribution as a model for the data, with a p_{val} of ~ 0.00. Fig. 4.1 shows the density histogram of X_3 and the pdf of the Lavalette distribution with the corresponding parameters.

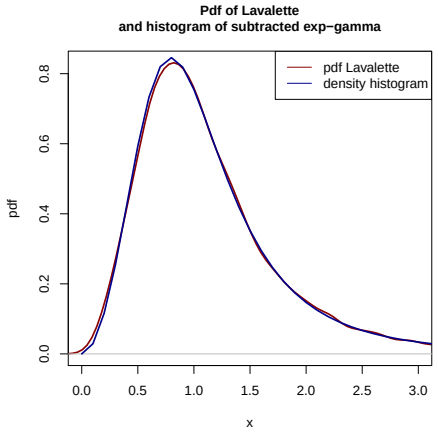


Figure 4.1: **Simulation of the model for the Lavalette distribution** Blue is the density histogram of the simulated data, generated by subtracting and transforming exp-gamma random variables. Red is the Lavalette density with the estimated parameters from the data.

This result has the following interpretation: if we look at the logarithm of gamma-distributed random variables, its subtraction in the logarithmic representation is a random variable whose distribution is a Lavalette after an expansion and a translation. In other words, within the logarithmic representation the subtraction of independent gamma random variables has a distribution of the same kind as the Lavalette distribution.

Conclusions

When shall we three meet again?
In thunder, lightning, or in rain?
When the hurlyburly's done,
When the battle's lost and won.

William Shakespeare, *Macbeth*

In this work we conducted three research projects concerning the DGBD, which is a two parameter rank-size function with a growing number of applications.

In our first inquiry we developed a novel application of the DGBD by utilizing it to describe population data of administrative units around the world. With this we addressed a rarely discussed topic: structure and population distribution of administrative divisions. These kind of units are artificially defined and in constant evolution, so we expected a high degree of heterogeneity. After comparing, for a set of 150 countries, the performance of the DGBD against the power-law, which is the most accepted model for city population, we concluded that DGBD is a good model in 73.5% of the cases, there is only one case which exhibits a a good power-law, representing 0.7% of the total, and in 25.8% of the cases neither is a good model. However, within those countries where both models are rejected, DGBD is better than power-law in 37 out of 38 cases. These results support the idea that DGBD is an appropriate model for fitting administrative unit population data. The DGBD is a very flexible function, capable of giving a good representation of a large number of data sets. In consequence, it is a good candidate for characterizing and comparing population distributions in different parts of the world, since they could be following different dynamics, making it difficult to propose a more universal model. We proposed the split-merge process to simulate the action of bureaucrats and politicians partitioning the territory of a country in administrative units. Because some cities are divided into distinct units and some units have more than one city, the correspondence between cities and administrative divisions breaks, so we do not have the same population distribution for these two types of data. Computational simulations show how initial power-laws and lognormal distributions of cities evolve into other kind of probabilistic laws for municipalities as the territory gets partitioned; distributions in latter stages are well represented by DGBD with $b > a$. With our numerical evidence, we conjecture that DGBD rises form the split-merge process. Further analysis focusing on the details of the process and a possible analytic derivation of DGBD are still needed. To conclude this part, we can say that DGBD together with the split-merge process prove adequate to describe and comprehend the formation, evolution and

population distribution in administrative units.

In our second investigation we constructed a novel family of continuous distribution functions, the Lavalette distribution, and showed it is a suitable alternative for data that does not clearly follow a power-law or a lognormal. This is the distribution of a random variable whose rank-size function is a DGBD with equal parameters. This distribution yields a very good approximation to the lognormal at the center of the distribution, they are difficult to distinguish from each other, but it may decay much slower depending on the value of one of its parameters. Thus, it models a random variable with a lognormal-like behavior at the central regime, followed by a fat tail (a relative high probability for extreme or rare events to happen). There are several phenomena in natural, social and economic sciences where a certain variable of interest seems to be logormally distributed over an extended regime, but displays a power-law behavior at the tail [13, 60]. This is precisely the form of the Lavalette distributions, so it natural to propose it to model such phenomena, where the debate is usually centered on which model is better, the lognormal ot he power-law. We showed the occurrence of the Lavalette distribution in some examples of real data in different phenomena: three from population distribution (municipality population in two Spanish provinces, municipality population in Nigeria) and two from genetic data (non-stop codons usage in non-primate and non-rodent mammalian nucleus, amino acid to amino acid mutation rates in DNA sequences). We utilized a resampling approach to the K-S goodness of fit test and the Akaike information criterion to test its suitability as a statistical model and compare its performance with that of other models. To further document its occurrence and comprehend possible mechanisms from where the Lavalette distribution emerges are topics of future study.

In our third study case we investigated on the possible causes of the extended power-law-followed-by-decay behavior, which the DGBD very well represents. Specifically, we provided partial answers to the following two questions: 1) how should a distribution family be in order to be closed under a subtraction operator? 2) what does this tells us about the behavior depicted by DGBD? These questions are already delineated in [27]. This is not the only open research path in this direction, but previously given numerical evidence suggested it is a promising one [25]. We introduced a very broad family of complex-valued functions, the $\mathcal{J}$ family, and proved its closedness under an integral operation that, when restricted to pdfs, gives the pdf of a subtraction of two independent random variables. This is a very general family of solutions and it encompasses many functions that are not probability functions. However, it provided us an specific example that helps understanding some possible machinery behind the DGBD behavior: if there are two independent random variables whose logarithm is gamma distributed, its subtraction is a third random variable whose logarithm is Lavalette distributed up to an expansion and a translation (they are of the same kind). This may prove to be a small but important step towards a full understanding of the random mechanisms behind the power-law-followed-by-decay phenomenon. There is still much to be done in this direction.

Which other applications may the DGBD function have? What is the reason of the extensiveness of the behavior it depicts? Are there many mechanisms leading to it, or is there a very unspe-

cific collective behavior, such as the one described by the center limit theorem? Does it emerge analytically from some version of the split-merge process? In which other fields can the Lavalette distribution be of utility? What else can it contribute to the "lognormal or power-law" debates? Which other mechanisms are there from where this distribution rises? There are other cases in which the pdf of the DGBD can be expressed in its closed form; are they useful or illuminating? Which properties do they have? There are still many open questions but, as somebody once said, *"it is ambition enough to be employed as an under-labourer in clearing the ground a little, and removing some of the rubbish that lies in the way to knowledge"*.

Bibliography

[1] 1000 Genomes Project Consortium, R. Durbin, G. Abecasis, D. Altschuler, A. Auton, and et al., *A map of human genome variation from population-scale sequencing*, Nature, 467 (2010), pp. 1061–1073.

[2] A. Agresti, *Categorical Data Analysis*, Wiley, Hoboken, NJ, USA, third ed., 2013.

[3] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automatic Control, 19 (1974), pp. 716–723.

[4] J. Aldrich and F. Nelson, *Linear Probability, Logit, and Probit Models*, Sage Pub., Newbury Park, CA, USA, first ed., 1984.

[5] R. Alvarez-Martinez, G. Martinez-Mekler, and G. Cocho, *Order-disorder transition in conflicting dynamics leading to rank-frequency generalized beta distributions*, Physica A, 390 (2011), pp. 120–130.

[6] F. Auerbach, *Das gesetz der bevolkerungskonzentration*, Petermans geograpische Mitteilungen, 49 (1913), pp. 73–76.

[7] M. Ausloos, *Two-exponent lavalette function: A generalization for the case of adherents to a religious movement*, Phys. Rev. E, 89 (2014), p. 062803.

[8] R. Axtell, *Zipf distribution of us firm sizes*, Science, 293 (2001), pp. 1818–20.

[9] P. Bak, *How Nature Works*, Copernicus Press, NY, 1996.

[10] D. M. Bates and D. G. Watts, *Nonlinear Regression Analysis and its Applications*, John Wiley and Sons, US, second ed., 2008.

[11] R. Baxter, *Exactly Solved Models in Statistical Mechanics*, Academic Press, London, 1982.

[12] R. Beals and J. Szmigielski, *Meijer g-functions: a gentle introduction*, Notices of the AMS, 60 (2013), pp. 866–872.

[13] M. Bee, M. Riccaboni, and S. Schiavo, *Pareto versus lognormal: A maximum entropy test*, PRE, 84 (2011), pp. 026104–1–11.

[14] L. Bettencourt, J. Lobo, D. Strumsky, and G. West, *Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities*, PLOS ONE, 5 (2010), p. e13541.

[15] C. Bliss, *The method of probits*, Science, 79 (1934), pp. 38–39.

[16] F. Bornemann, D. Laurie, S. Wagon, and J. Waldvogel, *Vom Lösen numerischer Probleme*, Springer, 2006.

[17] K. Burnham and D. Anderson, *Model Selection and Multi-Model Inference*, Springer-Verlag, New York, USA, second ed., 2003.

[18] ——, *Model Selection and Multi-Model Inference*, Springer-Verlag, New York, USA, second ed., 2003.

[19] J. Christenson and C. Sachs, *The impact of government size and number of administrative units on the quality of public services*, Administrative Science Quarterly, 25 (1980), pp. 89–101.

[20] A. Clauset, C. Shalizi, and M. Newman, *Power-law distributions in empirical data*, SIAM Rev., 51 (2009), pp. 661–703.

[21] D. Colquhoun, *An investigation of the false discovery rate and the misinterpretation of p-values*, Royal Society Open Science, 1 (2014), p. 140216.

[22] M. Cristelli, M. Batty, and L. Pietronero, *There is more than a power law in zipf*, Scientific Reports, 2 (2012), p. srep00812.

[23] M. Dayhoff, R. Schwartz, and B. Orcutt, *A model of evolutionary change in proteins*, in Atlas of Protein Sequence and Structure, M. Dayhoff, ed., Natl. Biomed. Res. Found., Washington, DC, USA, 1978, pp. 345–362.

[24] T. de Beer, R. Laskowski, S. Parks, B. Sipos, N. Goldman, and J. Thornton, *Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset*, PLoS Genet., 9 (2013), p. e1003382.

[25] M. B. del Río, G. Cocho, and R. Mansilla, *General model of subtraction of stochastic variables. attractor and stability analysis*, Physica A, 390 (2011), pp. 154–160.

[26] M. B. del Río, G. Cocho, and G. Naumis, *Universality in the tail of musical note rank distribution*, Physica A, 387 (2008), pp. 5552–5560.

[27] M. B. del Río García, *Ubicuidad de la distribución beta de dos parámetros. Fenomenología y modelos.*, PhD thesis, Universidad Nacional Autónoma de México, 2010.

[28] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London, UK, 1993.

[29] A. Erdély, *Higher trascendental functions Vol.1*, McGraw-Hill, US, 1953.

[30] W. Feller, *An Introduction to Probability Theory and its Applications Vol II*, John Wiley, NY, 1970.

[31] O. Fontanelli, P. Miramomtes, G. Cocho, and W. Li, *Population patterns in world's administrative units*, arXiv (preprint), (2016), p. arXiv:1610.02708.

[32] O. Fontanelli, P. Miramontes, Y. Yang, G. Cocho, and W. Li, *Beyond zipf's law: The lavalette rank function and its properties*, PLOS ONE, 11 (2016), p. e0163241.

[33] X. Gabaix, *Zipf's law and the growth of cities*, Am. Econ. Rev., 89 (1999), pp. 129–132.

[34] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. Stanley, *A theory of power-law distributions in financial market fluctuations*, Nature, 423 (2003), pp. 267–270.

[35] F. Gantner, B. Waldvogel, R. Meile, and P. Laube, *The basic formal ontology as a reference framework for modeling the evolution of administrative units*, Trnasactions in GIS, 17 (2013), pp. 206–226.

[36] K. Giesen, A. Zimmermann, and J. Suedekum, *The size distribution across all cities – double pareto lognormal strikes*, Journal of Urban Economics, 68 (2010), pp. 129–137.

[37] P. E. Gill, W. Murray, and M. Wright, *Practical Optimization*, Academic Press, San Diego CA, 1997.

[38] B. Gnedenko and A. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Cambridge, US, 1954.

[39] P. Gopikrishnan, V. Plerou, L. Amaral, M. Meyer, and E. Stanley, *Scaling of the distributions of fluctuations of finantial market indices*, Phys Rev E, 60 (1999), pp. 5305–5316.

[40] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, Burlington, MA, USA, seventh ed., 2007.

[41] G. Grossman and J. Lewis, *Administrative unit proliferation*, American Political Science Review, 108 (2014), pp. 196–217.

[42] B. Gutenberg and C. Richter, *Seismicity of the Earth and Associated Phenomena*, Princeton University Press, Princeto, NJ, second ed., 1954.

[43] B. Jiang and T. Jia, *Zipf's law for all the natural cities in the united states: a geospatial perspective*, Int. J. Geograph. Info. Sci., 25 (2010), pp. 1269–1281.

[44] N. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, John Wiley & Sons, New York, USA, second ed., 1994.

[45] A. Kilbas, O. Repin, and M. Saigo, *Generalized fractional integral transforms with gauss function kernels as g-transforms*, Integral Transforms and Special Functions, 13:3 (2010), pp. 285–307.

[46] J. Laherre and D. Sornette, *Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales*, Eur. Phys. J. B., 2 (1998), pp. 525–539.

[47] D. Lavalette, *Facteur dimpact: impartialit ou impuissance?*, INSERM preprint, U350 (1996).

[48] W. Li, *Spatial 1/f spectra in open dynamical systems*, Europhys. Lett., 10 (1989), pp. 395–400.

[49] ——, *New stopping criteria for segmenting dna sequences*, Phys. Rev. Lett., 86 (2001), pp. 5815–5818.

[50] ——, *New stopping criteria for segmenting dna sequences*, Phys. Rev. Lett., 86 (2001), pp. 5815–5818.

[51] ——, *Fitting chinese syllable-to-character mapping spectrum by the beta rank function*, Physica A, 391 (2012), pp. 49–53.

[52] ——, *Characterizing ranked chinese syllable-to-character mapping spectrum: a bridge between spoken and written chinese language*, J. Quant. Ling., 20 (2013), pp. 153–167.

[53] W. Li, O. Fontanelli, and P. Miramontes, *Size distribution of function-based human gene sets and the split–merge model*, Royal Society Open Science, 3 (2016), p. 160275.

[54] W. Li and P. Miramontes, *Fitting ranked english and spanish letter frequency distribution in us and mexican presidential speeches*, J. Quant. Ling., 18 (2011), pp. 337–358.

[55] W. Li, P. Miramontes, and G. Cocho, *Fitting ranked linguistic data with two-parameter functions*, Entropy, 12 (2010), pp. 1743–1764.

[56] F. J. López-Pellicer, A. J. Florczyk, J. Lacasta, F. J. Zarazaga-Soria, and P. R. Muro-Medrano, *Administrative units, an ontological perspective*, in Proceedings of the ER 2008 Workshops (CMLSA, ECDM, FP-UML, M2AS, RIGiM, SeCoGIS, WISM) on Advances in Conceptual Modeling: Challenges and Opportunities, ER '08, Berlin, Heidelberg, 2008, Springer-Verlag, pp. 354–363.

[57] Y. Luke, *The special functions and their approximations Vol.1*, Academic Press, US, 1969.

[58] R. Álvarez Martínez, *Transiciones orden-desorden en dinámicas en conflicto vía la función beta generalizada en rango-frecuencia*, PhD thesis, Universidad Nacional Autónoma de México, 2012.

[59] L. J. Ma, *Urban administrative restructuring, changing scale relations and local economic development in china*, Political Geography, 24 (2005), pp. 477–497.

[60] Y. Malevergne, V. Pisarenko, and D. Sornette, *Testing the pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of citie*, Phys. Rev. E, 83 (2011), p. 036111.

[61] B. Mandelbrot, *The pareto-lévy law and the distributionof income*, International Economic Review, 1 (1960), pp. 79–106.

[62] ——, *The stable paretian income distribution when the apparent exponent is near two*, International Economic Review, 4 (1963), pp. 111–115.

[63] R. Mansilla, E. Köppen, G. Cocho, and P. Miramontes, *On the behavior of journal impact factor rank-order distribution*, J. Informetrics, 1 (2007), pp. 155–160.

[64] G. Martínez-Mekler, R. Martínez, M. B. del Río, R. Mansilla, P. Miramontes, and G. Cocho, *Universality of rank-ordering distributions in the arts and sciences*, PLOS ONE, 4 (2009), p. e4791.

[65] A. Mathai and R. Saxena, *Generalized Hypergeometric Functions with Applications in Statistics and Physical Sciences*, Springer, Germany, 1973.

[66] P. McCullagh and J. Nelder, *Generalized Linear Models*, Chapman and Hall/CRC, London, second ed., 1989.

[67] M. Mitzenmacher, *A brief history of generative models for power law and lognormal distribution*, Internet Math., 1 (2004), pp. 226–251.

[68] M. Montemurro, *Beyond the zipf–mandelbrot law in quantitative linguistics*, Physica A, 300 (2001), pp. 567–578.

[69] Y. Nakamura, T. Gojobori, and T. Ikemura, *Codon usage tabulated from international dna sequence databases: status for the year 2000*, Nucl. Acids Res., 28 (2000), p. 292.

[70] G. Naumis and G. Cocho, *The tails of rank-size distributions due to multiplicative processes: from power laws to stretched exponentials and beta-like functions*, New Journal of Physics, 9 (2007).

[71] M. Newman, *Power laws, pareto distributions and zipf's law*, Journal of Contemporary Physics, 46 (2005), pp. 323–351.

[72] J. P. Nolan, *Stable Distributions - Models for Heavy Tailed Data*, Birkhauser, Boston, 2015. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

[73] E. Page, *Approximations to the cumulative normal function and its inverse for use on a pocket calculator*, Appl. Stat., 26 (1977), pp. 75–76.

[74] G. Peng, *Zipf's law for chinese cities: Rolling sample regressions*, Physica A, 389 (2010), pp. 3804–3813.

[75] R. Perline, *Zipf's law, the central limit theorem, and the random division of the unit interval*, Phys. Rev. E, 54 (1996), pp. 220–223.

[76] A. Petersen, H. Eugene, and S. Succi, *Statistical regularities in the rank-citation profile of scientists*, Sci. Rep., 1 (2011), p. 181.

[77] A. Petersen and S. Succi, *The z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile*, Journal of Informetrics, 7 (2013), pp. 823–832.

[78] A. Pishkoo, *The meijer's g function convenient for describing beta and gamma decays*, Journal of Nuclear Physics, 2 (2014), pp. 25–31.

[79] A. Pishkoo and M. Darus, *Fractional differintegral transformations of univalent meijer's g-functions*, Journal of Inequalities and Applications, 2012:36 (2012), pp. 1–10.

[80] I. Popescu, *On the lavalette ranking law*, Romanian Rep. Phys., 49 (1997).

[81] A. Saichev, Y. Malevergne, and D. Sornette, *Theory of Zipf's law and beyond*, Springer Verlag, 2010.

[82] M. Samta and H. Singh, *Statistical distributions involving meijer's g-function of matrix argument in the complex case.*

[83] K. Soo, *Zipf's law for cities: a cross-country investigation*, Reg. Sci. Urb. Econ., 35 (2005), pp. 239–263.

[84] D. Sornette, *Critical Phenomena in Natural Sciences*, Springer-Verlag, Berlin, second ed., 2006.

[85] D. Stauffer and A. Aharony, *Introduction to Percolation Theory*, Taylor and Francis, second ed., 2003.

[86] M. Stumpf and M. Porter, *Critical truths about power laws*, Science, 335 (2012), pp. 665–666.

[87] G. Troll and P Beim Graben, *Zipf's law is not a consequence of the central limit theorem*, Phys. Rev. E, 57 (1998), pp. 1347–1355.

[88] V. Uchaikin and V. Zolotarev, *Chance and Stability*, VSP, The Netherlands, 1999.

[89] R. R. Yarlagadda, *Analog and Digital Signals ans Systems*, Springer, US, 2010.

[90] G. Zipf, *The Psycho-Biology of Languages*, Houghtion-Mifflin, Boston, MA, 1935.

[91] ——, *Human Behaviour and the Principle of Least-Effort*, Addison-Wesley, Cambridge, US, 1949.

[92] V. Zolotarev, *One Dimensional Stable Distributions*, AMS, US, 1986.