

JOSEF NOVOTNÝ, VOJTĚCH NOSEK

NOMOTHETIC GEOGRAPHY REVISITED: STATISTICAL DISTRIBUTIONS, THEIR UNDERLYING PRINCIPLES, AND INEQUALITY MEASURES

J. Novotný, V. Nosek: *Nomothetic geography revisited: statistical distributions, their underlying principles, and inequality measures*. – Geografie–Sborník ČGS, 114, 4, pp. 282–297 (2009). – The paper focuses on some issues related to regularities in the statistical distributions of various social and environmental phenomena. Firstly, an older concern with statistical distributions of complex systems is revisited in order to exemplify surprisingly similar findings obtained across different disciplines. This interest has also been reflected in geography with a lot of activity given to the documentation and classification of the regularities but less to their explanations. As such, in the second part, some basic examples of general (statistical rather than context-specific) underlying principles for considered types of distributions are mentioned. The third part addresses related question of the measurement of inequality, which is the most commonly studied quantitative aspect of a statistical distribution. The performance of selected parametric measures of inequality is tested with respect to data coming from differently skewed distributions.

KEY WORDS: complex phenomena – inequality measures – regularity – statistical distribution.

The authors acknowledge support from the Research Grant MSM 0021620831 sponsored by the Czech Ministry of Education, Youth and Sport and from the project GA UK 8388/2008.

1. Introduction

The most common definition of geography refers to the study of spatial differentiation. From a geography-as-spatial-science point of view, a one-dimensional expression of two-dimensional spatial differentiation is statistical distribution showing the inequality in terms of the dispersion of observations around a central value. In the most general sense, the statistical distribution “demonstrates a kind of general regularity in the structure of external world. It contributes to the understanding of how the world is ordered and it thus helps to clarify one of the oldest philosophical tasks” (Korčák 1941, p. 172).

This paper concerns regularities in the statistical distributions of various social and environmental phenomena. It is a topic that has long been studied in a variety of scientific fields. The first step of such endeavour is usually observing and classifying empirical data and searching for general patterns. Typically, this involves the comparison of the data with theoretical models. If there is a pattern (e.g. when a particular functional form fits the data well), there is usually an underlying reason for it. As such, the next step of this research explores possible underlying mechanisms for the distribution in question. In this respect, it makes conceptual sense to distinguish between “general statistical” principles and “context-specific” underlying mechanisms

and factors, when the former may be considered as the “law-like” principles that do not directly depend on the specific context of a particular example. In addition to their academic value, the knowledge of basic regularities in the statistical distributions has an obvious practical appeal. For example, the majority of commonly used quantitative methods are parametric so that they make assumptions about (and depend upon) the distribution from which the analyzed data are drawn.

Given these introductory notes, the three interrelated objectives according to which the paper is organized are as follows. Firstly, in an overview of a literature, a long standing concern with statistical distributions of complex systems is revisited in order to exemplify surprisingly similar findings obtained across different fields of science dealing with complex phenomena. Inevitably, this interest has also been reflected in geography, though the discipline is far from being exceptional in this respect. In fact, while a lot of activity has been given to the empirical documentation of the empirical regularities (such as of the famous Zipf’s law), considerably less effort has been devoted to their explanations. Therefore, the second goal of this paper is to discuss some examples of general statistical (rather than context-specific) underlying principles that can provide some basic explanations for the emergence of considered statistical distributions (hence the term “nomothetic” in the title). The third objective of the article is more practical and concerns the measurement of inequality, which is the most commonly studied quantitative aspect of a statistical distribution. The performance of selected widely used parametric measures of inequality is tested with respect to data coming from distributions with different skewness.

2. Basic notations

Let us now begin with some simple notations. First, let us define a vector of non-negative measurements of some phenomena: $y_1, y_2, y_3 \dots y_n$, where n denotes the number of observations. In addition, let $f(y)$ be the probability density function describing the frequency distribution and $\hat{f}(y)$ be an estimate of this density function (i.e. a smoothed histogram). In the examples of statistical distributions provided in this paper we apply the Gaussian kernel probability density estimates corresponding to:

$$\hat{f}(y) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{\bar{y} - y_j}{h}\right)$$

where \bar{y} is the average value of the measured variable, y_j denotes the same variable for the unit j , h stands for the bandwidth (a parameter that determines smoothness of the density curve) and K is a Gaussian function that integrates to one. In this paper the method of automatic selection of h is applied as described in Silverman (1986). Although there are several other ways of describing statistical distribution, for the sake of simplicity, below we will use the term statistical distribution (or simply distribution) when referring to the estimates of univariate probability density function. Analogously, a bivariate and more generally a multivariate probability density distribution could be considered in order to analyze whether and how the variable in question interact with other measurable variables. The statistical investigation of probabilistic dependencies between different variables is however not ad-

dressed in this paper, as the focus is solely on the regularities in the univariate probability distributions.

3. Statistical distributions of complex systems analyzed in geography and elsewhere

In their introductory textbook on quantitative geography, Cole and King (1968, p. 2) indicate that the limit to the scope of the entities of geographical inquiry is usually reached when they become increasingly divided into smaller elements. In this simple way, the authors suggest that geographers are typically concerned with relatively complex systems and, consequently, with spatial variations in variables pertaining to these systems. Although there is a large variety of such systems, for the present purposes a distinction can be made between “individual objects” (such as cities, firms, lakes and rivers, mountain ranges, etc.) and regional systems (conceptualized, ideally, as spatially contiguous functional regions). While concern with regional systems is somewhat special to geography, the interest in the former group of entities is often common to different fields of science such as biology and ecology, earth sciences, physics, economics, and other disciplines dealing with complex phenomena. Despite the examples given below refer to the distributions of regions, a multidisciplinary engagement with the differentiation of the sets of various other complex systems provides important departures.

For the purposes of this paper a complex system is conceived in a usual way as any system composed from a large number of interacting components (particles grouping into the active agents) which form an integrated whole. Some other constitutive properties of complex systems are the non-linearity of interactions, the adaptive behavior (self-organization under selective pressures and a capacity to learn from history), the emergence (new “qualities” or properties arising out of a multiplicity of relatively simple interactions), the flexibility of boundaries, or the tendency to organize in hierarchies (see e.g. Halloy 1998; Amaral, Ottino 2004; O’Sullivan 2004). Please note that this meaning is not identical with the definition of the term complex system as follows

from “the primary classification of real systems” (Hampl, 1998, p. 196), in which the author stressed a high structural complexity of geographical systems in terms of their higher qualitative completeness in comparison to other real systems.

Being a subject of numerous context-specific influences, the actual distribution of any set of complex systems is always to a large extent a matter of empirical observation. However, this is not to deny the existence of some general regularities such as that of an elementary difference between the distributions based on the measures of the “inner structure” or “inner quality” of some complex systems and those based on variables which

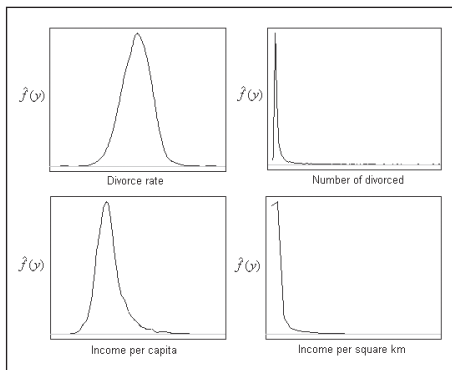


Fig. 1 – Distributions of 3,141 US counties according to the selected structure/quality characteristics and respective size characteristics. Source of data: US Census Bureau.

measure their size or magnitude (Thomas, Huggett 1980; Hampl 1971, 1998, 2000; Dostál, Hampl 1995). The distributions based on the former type of characteristics reveal relatively more symmetric shapes signifying a “typological” similarity of these entities compared to the quite asymmetric – considerably right-skewed – distributions according to their size measures. Some illustrative examples are shown in Figure 1 that depicts distributions of US counties based on the selected structure/quality characteristics (divorce rate, income per capita) and their respective size characteristics (number of divorced, income per square km).

Although a normal distribution is often considered as an appropriate functional form for the distributions based on the former type of variables, few empirical observations fit it perfectly in reality. In addition to discrepancies such as the floor (or ceiling) effect that may easily obscure the applicability of a normal distribution, often a more dynamic view is needed. Taking the time dimension and the character of the respective diffusion processes into account, we may often observe that the distribution is situated in some point of transition from one “state of symmetry” to another with various asymmetries in between. Usually, such distributions tend to proceed through the right-skewed to the left-skewed shape according to the diffusion of some “innovation”. Figure 2 provides an illustrative example that shows the historical evolution of the distribution of world countries according to their estimated life expectancy in the period 1800–2007. In this case, the epidemiological, nutrition, and demographic transitions (often conditioned by other factors such as economic and socio-cultural) may be considered as the abovementioned “innovations” whose diffusion determines the actual shape of the curve.

In contrast to the relative typological “homogeneity” of complex systems with respect to their structure/quality characteristics signified by their more or less symmetric distributions, they tend to be considerably differentiated with regard to their size measures. The statistical expression of this behavior is a class of highly right-skewed statistical distributions with a large number of low values and few high values. The size distributions of numerous relevant examples include cities and towns (Auerbach 1913, Zipf 1949, Simon 1955), firms (Gibrat 1931; Simon, Bonini 1958), land according to its value (Kaizoji 2003, Andersson et al. 2006), wars and terrorist events (Richardson 1948; Roberts, Turcotte 1998; Clauset et al. 2007), tourism arrivals (Ulubaşoğlu, Hazari 2004), islands, lakes, catchment areas, or lengths of rivers in river networks among other landforms (Korčák 1938, 1941; Turcotte 1995; Downing et al. 2006), disasters such as earthquakes (Gutenberg, Richter 1944), floods or wildfires (Malamud et al. 1998, Newman 2005), abundance of biological species (Willis, Yule 1922; Simon 1955; Preston 1960), worldwide and national transportation networks (Guimerà et al. 2005; Li, Cai 2004), or traffic jams (Nagel, Paczuski 1995).

Interestingly, this non-exhaustive list of literature indicates a surprisingly similar interest in certain types

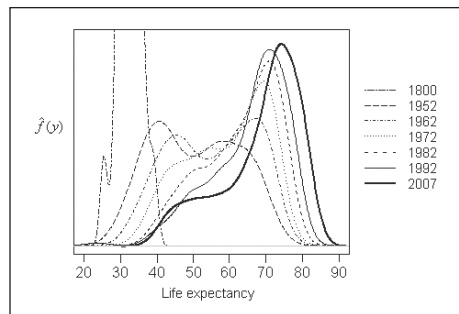


Fig. 2 – Evolution of the distribution of the world countries according to their estimated life expectancy. Source of data: dataset compiled by Gapminder.org basing on different sources and numerous estimates (see Johansson 2008).

of considerably right-skewed (positively asymmetric) distributions across different fields of scientific study, which dates back well into the nineteenth century. To go beyond the examples quoted above, Galton (1879) had previously emphasized the fact that the “law of arithmetic mean” (introduced into social science by Quetelet 1835) is inappropriate for distributions of many “social and vital” statistics. He thus expressed a concern similar to numerous subsequent authors who dealt with many other social, as well as natural phenomena. Not incidentally, this concern has also been reflected in geography especially with regard to a substantive body of research devoted to the empirical regularity known as Zipf’s law for cities. However, it appears that the first who described the general importance of asymmetric forms of variability in terms of highly skewed frequency distributions for geographical phenomena was Korčák (1938, 1941) who was inspired by Láska (1928) who proposed a method for a map-scale determination based on the examination of the frequency distributions.

The empirical findings of Korčák have been further elaborated by Hampl (1971, 1998, or 2000 among other works) into a consistent theory. In addition to the difference between the distributions of structural and size variables that is discussed in this paper, the focus of Hampl was also on the difference between generally more symmetric size distributions of “elements” (relative similarity of internal/personal capacities of individuals) in comparison to a transient distribution of “semi-complexes” (partial heterogeneity of narrowly defined social systems) and strongly right-skewed distributions of “complex geographical systems” (i.e. externally determined hierarchies typical for societal organization in environment). Also notably, some of the Korčák’s (1938) empirical findings have had an interesting international impact that has so far been ignored within the Czech geographical community. As indicated by a well known mathematician B. Mandelbrot in personal communication with the authors, it was a French mathematician Fréchet who noticed the Korčák’s empirical findings on the size distributions of lakes and islands (Fréchet 1941) and who made them available to Mandelbrot. Later on, Mandelbrot acknowledged Korčák’s findings in some of his works on fractals (e.g. Mandelbrot 1975a, 1975b). This gave rise to a dozen of the subsequent references to original Korčák’s paper (including the very recent ones) that have appeared in the literature from various fields (mostly environmental sciences).

Despite the existence of several other functional forms for strongly right-skewed distributions, the above mentioned empirical distributions are most commonly approximated either by a power law function/distribution¹ (including the special cases of Pareto and Zipf’s distributions) or by a lognormal distribution. Sometimes, a combination of these two functional forms provides a realistic approximation when the upper tail is estimated by a power law and the body by a lognormal distribution (Levy, Solomon 1996; Gabaix 1999; Mitzenmacher 2003). As such, these two families, which are the most familiar from the class of highly skewed distribution functions, may be considered as the approximate (if not natural²) “attractors” for the size distributions of many complex systems including those studied in geography.

¹ Power law is a scale invariant polynomial relationship between two variables (x and y) that can be described generally as $y = ax^{-k}$, where a is a constant and k is a scaling exponent. Note that by transforming both sides of the equation to logarithms we get: $\log(y) = \log(a) - k \times \log(x)$, which describe a line in a log-log plot with slope $-k$.

Two illustrative examples of three different displays of empirical distributions that may be approximated by a power law or lognormal functional forms are shown in Figure 3. It depicts the distributions of Czech municipalities and US counties according to their population size. The upper plots capture conventional kernel density estimates based on the untransformed values, the middle plots show the same observations after logarithmic transformations, and the lower plots display log-log rank-size relationships. A simple way of inspecting whether the empirical data exhibit the properties of a power law is their comparison to Zipf's models that correspond to the straight lines in the lower plots in Figure 3. As the visual inspection suggests, the distribution of Czech municipalities seems to obey the Zipf's functional form very well. This is additionally confirmed numerically by almost perfect fit of a regression line corresponding to: $\log pop_{cz} = 6.65 - 1.19 \log rank_{cz}$; $R^2 = 0.94$ (note the almost ideal slope of the regression line). In addition, the Vuong's test suggests the suitability of power law explanation in comparison with the application of other highly skewed distributions. By contrast, an explanation allowing for non-linearity in the log-log rank-size relationship seems to be more plausible for the distribution of US counties, which also documents the fit of polynomial regression: $\log pop_{us} = 5.59 + 1.00 \log rank_{us} - 0.44 (\log rank_{us})^2$; ($R^2 = 0.94$). At the same time, the concavity of the rank-size plot indicates the applicability of a lognormal functional form (Ulubaşođlu and Hazari 2004, p. 465). Although conventional statistical tests do not allow us to conclude (at the usual significance level) that the empirical data strictly follow some particular functional form, a lognormal function seems to be the closest of the other commonly used distributions that were tested. These findings may suggest a hybrid explanation for the population size distributions in terms of simultaneous operation of generative principles which are usually proposed for power laws and lognormal distributions. Some of the basic generative mechanisms will be briefly described in the following section.

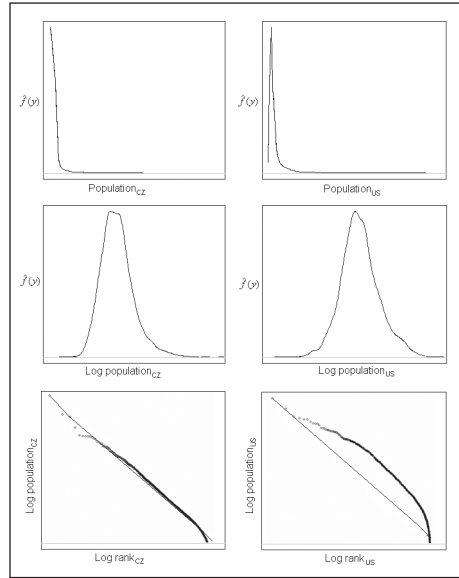


Fig. 3 – Different displays of the distributions of 6,258 Czech municipalities (left side plots) and 3,141 US counties (right side plots) according to their population size. The straight lines on the log-log rank-size plots (lower plots) represent theoretical Zipf's distributions. Sources: Czech Statistical Office (Census 2001), US Census 2000.

² Halloy (1998) proposed theoretical framework explaining the general convergence of the size distribution of (certain sort of) complex systems towards a “polo” distribution – that means simultaneously to a power law rank-size distribution and a lognormal frequency distribution.

4. Basic underlying principles

Although the observation and categorization of regularities according to which the external world is organized is undoubtedly an important task, it is only the first step which has to be followed by an effort directed towards providing explanation for these regularities. In the case of statistical distributions discussed in the previous section such explanation resides in examining processes from which these distributions may arise. Although the actual shape of each empirical distribution is usually subject to a number of specific influences (consider impacts of the epidemiological, nutrition, and demographic transitions in the example of world life expectancy distribution in Figure 2), it is often thought that there are some more general generative principles that work beyond the specific contexts of particular empirical examples. These principles may be regarded as determining the stochastic attraction of the empirical distributions to the properties of functional forms mentioned above.

In this regard, the very label “skewed distribution” personifies the fact that a given distribution deviates from the normal distribution symmetry. It thus seems reasonable to begin with the classical explanation of normal distribution on the basis of the central limit theorem (CLT) that may serve a simple null proposition against which different mechanisms that skew so many empirical distributions to the right can be contrasted. In this way, we will briefly outline two groups of such general mechanisms including the random multiplicative process and the spatial analogies of the CLT.

4.1. Random multiplicative process

The classical CLT states that the normal distribution arises from the summation of many independent random variables (supposing some other conditions are in place). It is a result of the cumulative addition of many small independent effects distributed relatively symmetrically around both sides of the average observation and, as such, it has also become known as “the law of error”. A number of real processes are additive in character such as most of those underlying the endogenous variability of the traits and abilities of individuals within a genera (not incidentally, examples from genetics have been frequently mentioned – see Galton 1869).

However, the classical CLT can hardly explain situation when the observations of the same phenomena differ by orders of magnitude. It is thus suggested that “the law of error has two forms, and resulting normality may be arithmetic where equivalent positive and negative deviations from expectation differ by equal amounts, or normality may be geometric where equivalent deviations differ by equal proportions” (Gingerich 2000, p. 201). The term geometric normality refers to the fact that underlying processes are multiplicative rather than additive or, put differently, they are additive acting on logarithms because: $\log(x \times y) = \log(x) + \log(y)$. It then follows from the basis of the additive CLT that $\log(x \times y)$ approaches a normal distribution implying that the product itself ($x \times y$) approaches a log-normal distribution (according to the multiplicative version of CLT). As such, lognormal distribution is sometimes emphasized to have virtually as fundamental a position in the real world as does the normal distribution (Aitchinson and Brown 1957, Limpert et al. 2001).

A simple example of such a multiplicative process that is often used as an explanation of the genesis of the size distributions of many complex systems

is the mechanism of random multiplicative growth. The essence of this concept has also become known as the “law of proportionate effect” (Gibrat 1931, Simon 1955) as it assumes that absolute increments are proportional to the size of a system in question (i.e. multiplicative effect). In addition, it proves that even purely random fluctuations in growth rates within a set of complex systems (e.g. due to specific environmental or social factors) will on the basis of the multiplicative CLT lead to a highly right-skewed size distribution. Moreover, this distribution is thought to have lognormal properties according to the Gibrat’s original formulation or power law properties if some further specifications are added.³ The mechanism of a stochastic multiplicative growth and its various modifications⁴ has gained a prominent position among several other explanations including the concepts of preferential attachment or optimizing behavior (e.g. Mitzenmacher 2003). It has also been frequently mentioned with regard to phenomena studied in geography such as those concerning the formation and modeling of urban systems and distribution of population or economic activity (Gabaix 1999; Davis, Weinstein 2002; Gabaix, Ioannides 2003) or differentiation of tourism arrivals (Ulubaşoğlu, Hazari 2004) among other applications.

4.2. Spatial analogies to the central limit theorem

Although the concept of random multiplicative growth provides an intuitively appealing explanation for the size differentiation of many complex systems, it seems to be less comprehensible when the spatial context is explicitly taken into account. In particular, it doesn’t consider spatial interactions and different geographical extent (spatial levels) at which the distributions can be observed (Andersson et al. 2006).

Both of these issues have been recently addressed in an interesting explanation of a strongly right-skewed species abundance distributions proposed by Sizing et al. (2009). The authors show that the distribution arises “from below” by a repetitive additive splicing of the abundances in many non-overlapping neighboring subplots (i.e. small regions) into the abundances pertaining to regions of progressively larger sizes (regions at higher spatial scales). Up to this point the process is analogous to the summation of random variables indicating the applicability of the additive CLT. However, it is shaped by specific spatial arrangements described by two parameters: the structure of spatial autocorrelation among the abundances, and spatial turnover of species, which determine a convergence towards strongly right-skewed distribution (Sizing et al. 2009, p. 6691). Despite this bottom-up model having been proposed and validated in the context of community ecology, it may suggest a very general

³ In fact, only a small change in the specification of the underlying generative model can change the result from a power law to a lognormal distribution and conversely (Levy, Solomon 1996; Gabaix 1999; Mitzenmacher 2003).

⁴ For example Davis and Weinstein (2002) propose an interesting hybrid explanation for the historical development of the distribution of regional population in Japan combining a random growth process with the concepts of locational fundamentals and increasing returns. They showed that rather than city growth itself being purely random, it is the locational fundamentals (geographical attractiveness in terms of close distance to rivers, ports etc.) that are random. The locational fundamentals established the spatial pattern of relative regional densities and increasing returns help to determine the degree of spatial differentiation.

(statistical rather than biological) principle of the summation (splicing) of spatially unevenly distributed variables which may be applicable in other fields dealing with spatial phenomena.

Nevertheless, even this model provides an explanation for the shape of the “parent distribution” of some “individual objects” (whether they are species, cities, or firms etc.) and doesn’t tell us much about the distribution of regions (regional characteristics). We can thus propose another simple spatial analogy to the CLT based on its slightly different interpretation that justifies convergence of the sampling distribution of the mean towards a normal distribution. More concretely, the CLT suggests that, given a parent population with a mean μ and variance σ^2 , the distribution of the means pertaining to random independent samples drawn from this parent population (that can be of arbitrary distribution if some basic conditions hold) converges with increasing sample size (N) to a normal distribution with the same mean and variance σ^2/N . Now imagine that regions (subpopulations delimited by regional boundaries) could be considered as these independent samples drawn from some parent population (e.g. country). Then the convergence of the distribution of regional means to a normal distribution would be expected with increasing the levels of aggregation (i.e. increasing N) or geographical scale of considered regions irrespective of the parent distribution shape. Although this may be considered a simple null proposition, the process is again mostly shaped by various spatial interactions which lead to spatially autocorrelated data.⁵ Because of the prevailing positive spatial autocorrelation the sampling variance predicted on the basis of CLT (i.e. σ^2/N) tends to be underestimated. At the same time, the stronger the spatial autocorrelation the larger the underestimation and, consequently, a larger deviation of the actual distribution from the normal distribution symmetry can be expected⁶. This simple argument lies beyond the empirically documented regularity that relative spatial unevenness (relative spatial concentration) of natural and social phenomena generally increases with increasing geographical scale of observation (see figure 12 in Hampl 1998, p. 86). While the aforementioned proposition seeks to indicate links between the spatial autocorrelation and the shape of respective distribution (and thus also between the measures of spatial autocorrelation and inequality – see Netrudová, Nosek 2009), the exact nature of this relationship still requires further investigation.

5. Sensitivity of inequality measures to data from skewed distributions

While the graphical displays used above provide complete information about the entire course of a statistical distribution, often these tools are impractical because they do not allow for more exact quantitative comparisons. Therefore, measures of inequality which quantify the character of dispersion of observations over the whole distribution are employed. While a number of such measures have been developed for various applications, among other prop-

⁵ The pervasiveness of spatial autocorrelation determined by the spatial dependency of a majority of real-world processes and events led Tobler (1970, p. 236) to invoke “the first law of geography” arguing that “everything is related to everything else, but near things are more related than distant things.”

⁶ In fact, the gap between the actually observed between-region variance (σ^2/N) refers to the relative significance of the spatial dimension of inequality (see Novotný 2007).

erties, they may perform differently with regard to the presence of extreme values (Cowell, Flachaire 2007). Given the considerations of different forms of distributions that can be expected for geographical phenomena, here we will examine the behavior of selected widely used inequality indices in the relation to the distribution skewness. For these purposes, the following six measures were selected, of which each has certain appeal either because of its mathematical properties or intuitive transparency⁷:

1. Coefficient of variation (*CV*):
$$CV = \frac{1}{\bar{y}} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$
2. Theil coefficient (*T*):
$$T = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\bar{y}} \ln \frac{y_i}{\bar{y}}$$
3. Mean logarithmic deviation (*MLD*):
$$MLD = \frac{1}{n} \sum_{i=1}^n \ln \frac{\bar{y}}{y_i}$$
4. Gini coefficient (*G*):
$$G = \frac{1}{2\bar{y}n^2} \sum_{i=1}^n \sum_{j=1}^n |(y_i - y_j)|$$
5. Rate of heterogeneity (*H*) which quantifies value at the 50th percentile of the Lorenz curve
6. Robin Hood (or Hoover) index (*RHI*)⁸:
$$RHI = \frac{1}{2} \sum_{i=1}^n \left| \frac{\bar{y} - y_i}{y_i} \right|$$

The design of the testing procedure is as follows: Firstly, a large number of random numbers drawn from lognormal distributions with the same (geometric) mean (corresponding to one) but different standard deviations (σ) is generated. The selection of the lognormal family is important not only because of its real-world significance (e.g. Limpert et. al 2001). The parameter σ determines the extent of the right-tail skewness of particular lognormal distributions which thus represent a continuum of distributions going from those with almost symmetric shapes very close to the normal one (for small σ) to considerably right-skewed shapes similar to some power law functions for large σ (the results were calculated for 13 values of σ ranging from 0.001 to 10). Moreover, it is suggested that, dissimilarly to some other functional distributions, the Lorenz curves for lognormal distributions specified by different σ are nonintersecting and the parameter σ therefore allows for unambiguous ordering of these distributions according to their inequality (Aitchinson, Brown 1957).

Secondly, the six measures of inequality listed above are calculated for each lognormally distributed dataset defined by different σ . As the results for the considered inequality measures mostly fall into different ranges, their standardized values $I(y)$ are calculated by dividing each of the results by the respective value of the same measure calculated for the normal distribution with

⁷ Note that we consider only the relative measures of inequality that focus on differences in relative proportions instead of absolute differentials and, at the same time, do not depend on the units of measurement.

⁸ Note that *RHI* corresponds to the amount that would have to be redistributed (from the upper half of a distribution to the lower half) in order to get uniform (equal) distribution.

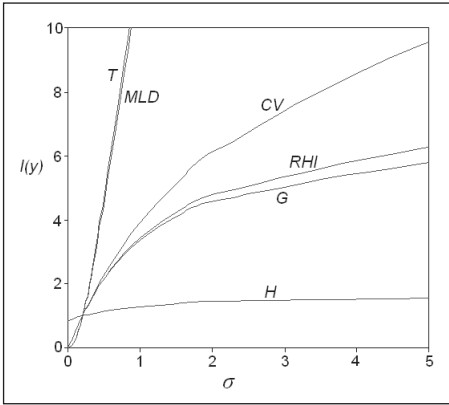


Fig. 4 – Sensitivity of inequality measures to data from differently skewed distributions; σ stands for the standard deviation of considered lognormal distribution and $I(y)$ denotes standardized values of particular measures of inequality for this distribution.

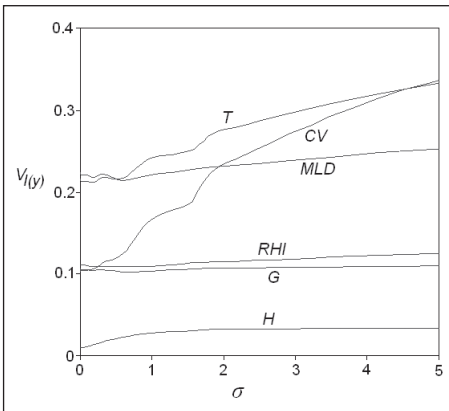


Fig. 5 – Stability of inequality measures with respect to random fluctuation in data from differently skewed distributions; σ stands for the standard deviation of considered lognormal distribution and $V_{I(y)}$ denotes relative variability of particular indicators of inequality (measured by the coefficient of variation) calculated 1,000 times for 13 different levels of σ .

In addition, Figure 5 depicts another interesting property of the considered inequality measures in terms of their stability with regard to random fluctua-

a mean of 10 and standard deviation of 2. This should make the behavior of particular indicators more easily mutually comparable.

Thirdly, using the Monte Carlo simulation technique the same procedure as described above is repeated 1,000 times and each result recorded. In this way, the $1,000 \times 13 \times 6$ results (the set of 1,000 results for each of 13 values of σ and each of the six inequality measures) are gathered. The average values of the inequality measures are then computed from these 13×6 sets of results (the repetitions should obviously restrict impacts of random fluctuations)⁹. These average values are shown in Figure 4 and indicate the behavior of the considered inequality measures in relation to σ .

The figure suggests that the Theil coefficient, *MLD* and, to a certain extent, also the coefficient of variation are considerably sensitive to data from skewed distributions. It may limit the applications of these indices when there are large outlying observations. On the other hand, the rate of heterogeneity *H* is quite insensitive to changes in the distribution shape, at least as far as the skewness is considered. While in some cases this may be regarded as an advantage (e.g. when we can expect some measurement errors or existence of some influential large outliers that are not the object of interest), in many other cases, the lack of sensitivity to different levels of σ disqualifies *H* as a transparent measure of inequality. The Gini coefficient and *RHI* are then in an intermediate position as they are reasonably sensitive to changes in σ until it reaches some “critical” level (approximately $\sigma = 2$) from which their marginal increases are minimal.

⁹ The authors are aware that similar results can probably also be obtained numerically. However, mainly because of a lack of appropriate mathematical skills, the simple “experimental” way was employed here.

tions in repetitively generated data. While the x -axis has remained the same as in the previous figure, the value of $V_{I(\sigma)}$ in the y -axis shows variability of results obtained by particular inequality indicators within the sets of 1,000 repetitions for each of 13 levels of σ . Although the stability of particular inequality measures is evidently related to their sensitivity to different σ examined previously in Figure 4, here additionally the poor performance of the coefficient of variation has been uncovered.

6. Conclusion

The reviewed literature from different disciplines (of which only a fraction is referred to in this paper) has indicated that existing research on the regularities in statistical distributions of various social and environmental phenomena involves two interrelated steps. The first step usually comprises the identification and empirical documentation of the regularities as well as their classification (e.g. the approximation of observed data by functional forms). In the second step, possible explanations in terms of underlying models are searched, proposed, and possibly validated. Obviously, the identification and description of a distribution is as consequential as the knowledge of processes and circumstances from which such a distribution can emerge. Now we argue that most of this research in quantitative geography (including the Czech case) has focused on identifying, documenting, and categorizing but less on the latter step. In this regard, it lags behind some other disciplines dealing with complex systems where the current emphasis is much more on the propositions of underlying models (both statistical and context-specific) and on their validation on real or simulated data. To a certain extent this gap may be justified by the specific character of geographical inquiry. However, a general remark which goes beyond the topics discussed here is that findings obtained elsewhere still have to be acknowledged and critically examined with regard to their applicability in the specific geographical context.

In this paper we have sought to provide some basic illustrations pertaining to both of the aforementioned steps in the research on regularities in statistical distributions. Departing from the elemental difference between the distributions according to structural and size variables, the regularity of considerably right-skewed size distributions was discussed as pertaining to almost any set of (certainly defined) complex systems including those investigated in geography. Subsequently, some examples of basic underlying principles that may provide very general (statistical rather than context-specific) explanations for the considered distributions have briefly been mentioned. More concretely, we have touched on the well-known random multiplicative process and more innovative spatial analogies of the central limit theorem. With respect to the latter, a challenge for future research has been indicated in terms of the investigation of the relationship between the structure and extent of spatial autocorrelation and the convergence of the distribution of regions (spatially contiguous groups of individual observations) to a skewed (more unequal) figures with increasing the levels of aggregation (spatial scale).

A more practically-minded person might regard the above discussion as worthless. However, it is suggested that the regularities such as those in statistical distributions are important because they provide some general global constraints for locally specific processes and patterns. At the same time, they cannot be understood as some absolute laws but rather in a softer sense as a

stochastic framework that may often be anticipated beyond empirical facts and that is open to incorporation of context-specific details and factors (Andersson et al. 2006).

Bearing these general remarks in mind, in this paper we nevertheless intended to appease more practically-oriented readers too. Therefore, in section 4 we tested the performance of six common measures of inequality with regard to data coming from distributions of different skewness¹⁰. Using a simple Monte Carlo simulation method a considerably different sensitivity and stability of the analyzed inequality measures was detected. The Theil coefficient, the mean logarithmic deviation, and to certain extent also the coefficient of variation are quite sensitive to data from skewed distributions. By contrast, the rate of heterogeneity H has been found to be impractically insensitive. In addition, the Theil index and especially the coefficient of variation were uncovered to be prone to instabilities in data. For researchers or policy-makers who may wish to select an appropriate measure of inequality we thus advise application of the Gini coefficient or the Robin Hood index if skewed data are expected. A simple comparison of the distance between the mean and median may reveal a lot in this respect. Obviously, although the focus here was on the measures of inequality, similar arguments may apply regarding the usage of many other statistical techniques that require normally distributed data.

References:

- AMARAL, L. A. N., OTTINO, J. M. (2004): Complex networks: Augmenting the framework for the study of complex systems. *The European Physical Journal B*, 38, No. 2, pp. 174–162.
- ANDERSSON, C., FRENKEN, K., HELLERVIK, A. (2006): A complex network approach to urban growth. *Environment and Planning A*, 38, No. 10, pp. 1941–1964.
- AITCHINSON, J., BROWN, J. A. C. (1957): *The Lognormal Distribution*. Cambridge University Press, Cambridge, 176 p.
- AUERBACH, F. (1913): Das gesetz der bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59, No. 1, pp. 74–76.
- CLAUSET, A., YOUNG, M., GLEDITSCH, K. S. (2007): On the frequency of severe terrorist events. *Journal of Conflict Resolution*, 51, No. 1, pp. 58–87.
- COLE, J. P., KING, C. A. M. (1968): *Quantitative geography: techniques and theories*. Wiley, London, 692 p.
- COWELL, F. A., FLACHAIRE, E. (2007): Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, 141, No. 2, pp. 1044–1072.
- DAVIS, D. R., WEINSTEIN, D. W. (2002): Bones, bombs, and break points: the geography of economic activity. *American Economic Review*, 92, No. 5, pp. 1269–1289.
- DOSTÁL, P., HAMPL, M. (1995): Geographical organization and societal development: searching for an integral approach. *Acta Universitatis Carolinae Geographica*, XXX, No. 1–2, pp. 21–42.
- DOWNING, J. A., PRAIRIE, Y. T., COLE, J. J., DUARTE, C. M., TRANVIK, L. J., STRIEGL, R. G., MCDOWELL, W. H., KORTELAINEN, P., CARACO, N. F., MELACK, J. M., MIDDELBURG, J. (2006): The global abundance and size distribution of lakes, ponds, and impoundments. *Limnology and Oceanography*, 51, No. 5., pp. 2388–2397.
- FRÉCHET, M. (1941): Sur la loi de répartition de certaines grandeurs géographiques. *Journal de la Société de Statistique de Paris*, 82, 114–122.
- GABAIX, X. (1999): Zipf's law for cities: an explanation. *Quarterly Journal of Economics*, 114, No. 3, pp. 739–767.

¹⁰ Note that inequality is inherently related to the form of statistical distribution and, at the same time, inequality and its measurement has many practically important consequences (e.g. Novotný 2006).

- GABAIX, X., IOANNIDES, Y.M. (2003): The evolution of city size distributions. In: Henderson, J. V., Thisse, J. F. (ed.): Handbook of urban and regional economics, IV: Cities and geography, North-Holland, Amsterdam, pp. 2341–2378.
- GALTON, F. (1869): Hereditary Genius: an Inquiry into its Laws and Consequences. London, Macmillan, 390 p.
- GALTON, F. (1879): The geometric mean, in vital and social statistics. Proceedings of the Royal Society 29, pp. 365–367.
- GIBRAT, R. (1931): Les inégalités économiques. Librairie du Recueil Sirey, Paris.
- GINGERICH, P. D. (2000): Arithmetic or geometric normality of biological variation: an empirical test of theory. Journal of Theoretical Biology, 204, No. 2, pp. 201–221.
- GUIMERA, R., MOSSA, S., TURTSCHI, A., AMARAL, L. A. N. (2005): The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. PNAS, 102, No. 22, pp. 7794–7799.
- GUTENBARG, B., RICHTER, R. F. (1944): Frequency of earthquakes in California. Bulletin of the Seismological Society of America, 34, No. 4, pp. 185–188.
- HALLOY, S. (1998): A theoretical framework for abundance distributions in complex systems. Complexity International, 6, No. 12, pp. 1–12.
- HAMPL, M. (2000): Reality, society and geographical/environmental organization: searching for an integrated order. Charles University. Prague, 112 p.
- HAMPL, M. (1998): Realita, společnost a geografická organizace: hledání integrálního řádu. Přírodovědecká fakulta UK, Praha, 110 p.
- HAMPL, M. (1971): Teorie komplexity a diferenciace světa. Praha, Univerzita Karlova, 183 p.
- JONASSON, K. (2008): Documentation for life expectancy at birth (years) for countries and territories. Gapminder Documentation 004, Stockholm, www.gapminder.com.
- KAIZOJI, T. (2003): Scaling behavior in land markets. Physica A: Statistical Mechanics and its Applications, 326, No. 1–2, pp. 256–264.
- KORČÁK, J. (1941): Přírodní dualita statistického rozložení. Statistický obzor, 22, pp. 171–222.
- KORČÁK, J. (1938): Deux types fondamentaux de distribution statistique. Prague, Comité d'organisation, Bull. de l'Institute Int'l de Statistique, vol. 3, pp. 295–299.
- LÁSKA, V. (1928): Zpráva o zeměpisně-statistickém atlasu. Věstník Československé akademie věd a umění, pp. 61–67.
- LEVY, M., SOLOMON, S. (1996): Power laws are logarithmic Boltzmann laws. International Journal of Modern Physics C, 7, No. 4, pp. 595–601.
- LI, W., CAI, X. (2004): Statistical analysis of airport network of China. Physical Review E, 69, No. 4, pp. 1–6.
- LIMPERT, E., STAHEL, W. A., ABBT, M. (2001): Log-normal distributions across the sciences: keys and clues. Bioscience, 51, No. 5, pp. 341–352.
- MALAMUD, B., MOREIN, G., TURCOTTE, D. (1998): Forest fires: an example of self-organized critical behavior. Science, 281, No. 5384, pp. 1840–1842.
- MANDELBROT, B. B. (1975a): Earth's relief, shape and fractal dimension of coastlines, and number area for islands. PNAS, 72, No. 10, pp. 3825–3838.
- MANDELBROT, B.B. (1975b): Les Objets Fractals, Forme, Hasard et Dimension. Flammarion, Paris, 190 p.
- MITZENMACHER, M. (2003): A brief history of generative models for power law and log-normal distributions. Internet Mathematics, 1, No. 2, pp. 226–251.
- NAGEL, K., PACZUSKI, M. (1995): Emergent traffic jams. Physical Review E, 51, No. 4, pp. 2909–2918.
- NETRDOVÁ, P., NOSEK, V. (2009): Přístupy k měření významu geografického rozměru společenských nerovnoměrností. Geografie, 114, No. 1, pp. 52–65.
- NEWMAN, M. E. J. (2005): Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46, No. 5, pp. 323–351.
- NOVOTNÝ, J. (2006): Negativní vlivy společensko-ekonomických nerovností a mechanismy jejich regulace. Ekonomický časopis, 54, No. 7, pp. 709–724.
- NOVOTNÝ, J. (2007): On the measurement of regional inequality: does spatial dimension of income inequality matter? Annals of Regional Science, 41, No. 3, pp. 563–580.
- O'SULLIVAN, D. (2004): Complexity science and human geography. Transactions of the Institute of British Geographers, 29, No. 3, pp. 282–295.
- PRESTON, F. W. (1948): The commonness and rarity of species. Ecology, 29, No. 3, pp. 611–627.

- QUETELET, A. (1835): Sur l'homme et le développement de ses facultés, ou essai de physique sociale. Livre second. Bachelier, Paris, 327 p.
- RICHARDSON, L. F. (1948): Variation of the frequency of fatal quarrels with magnitude. *Journal of the American Statistical Association*, 43, No. 244, pp. 523–546.
- ROBERTS, D., TURCOTTE, D. (1998): Fractality and self-organized criticality of wars. *Fractals*, 6, No. 4, pp. 351–357.
- SILVERMAN, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 212 p.
- SIMON, H. A. (1955): On a class of skew distribution functions. *Biometrika*, 42, No. 3/4, pp. 425–440.
- SIMON, H. A., BONINI, CH. (1958): The size distribution of business firms. *The American Economic Review*, 48, No. 4, pp. 607–617.
- ŠIZLING, A. L., ŠTORCH, D., ŠIZLINGOVÁ, E., REIF, J., GASTON, K. J. (2009): Species-abundance distribution results from a spatial analogy of central limit theorem. *PNAS*, 106, No. 16, pp. 6691–6695.
- THOMAS, R. W., HUGGETT, R. J. (1980): *Modelling in geography: a mathematical approach*. Barnes and Noble, New Jersey, 338 p.
- TOBLER, W. (1970): A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, No. 2, 234–240.
- TURCOTTE, D. (1995): Scaling in geology: landforms and earthquakes. *PNAS*, 92, No. 15, pp. 6697–6704.
- ULUBAŞOĞLU, A., HAZARI, B. R. (2004): Zipf's law strikes again: the case of tourism. *Journal of Economic Geography*, 4, No 4, pp. 459–472.
- WILLIS, J. C., YULE, G. U. (1922): Some statistics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, 109, pp. 177–179.
- ZIPF, G. K. (1949): *Human behaviour and the principle of least effort*. Addison-Wesley, Reading MA.

S h r n u t í

K VYBRANÝM OTÁZKÁM NOMOTETICKÉ GEOGRAFIE: STATISTICKÁ ROZLOŽENÍ, JEJICH VÝCHOZÍ PRINCIPY A MÍRY NEROVNOUČNOSTÍ

Nejběžnějším jednorozměrným vyjádřením dvourozměrné prostorové diferenciace určitého jevu je jeho statistické rozdělení (frekvenční, resp. pravděpodobnostní rozdělení). Důležitost tohoto vyjádření, které zachycuje některé obecné pravidelnosti v uspořádání vnějšího světa, zdůraznil v kontextu geografického zkoumání již Korčák (1938, 1941), jehož myšlenky pak dále rozvedl Hampl (1971, 1998, 2000 atd.). Vybranými otázkami, které se týkají pravidelností ve statistických rozloženích různých sociálních a environmentálních jevů se zabývá i předkládaný článek, který je strukturovaný do tří souvisejících částí. První část je založena na rešerši literatury a několika empirických příkladech. Východiskem je dříve zdůrazněné rozlišení mezi rozloženími podle strukturálních/„kvalitativních“ a velikostních znaků komplexních systémů. Cílem je zde mimo jiné poukázat na dlouhodobý (a mnohdy paralelní) zájem řady různých vědních disciplín studujících komplexní systémy o obdobné typy jejich statistického rozdělení na základě velikostních znaků s významnou pravostrannou šikmostí (pozitivní asymetrií).

Uvedený zájem byl také reflektován v geografii, byť v tomto ohledu nelze naši disciplínu považovat za výlučnou. Zatímco velký důraz byl položen na empirickou dokumentaci a případně pak i klasifikaci pravidelností v uvedených rozloženích, relativně méně pozornosti bylo (oproti některým jiným disciplínám) věnováno jejich potenciálním vysvětlením. Ve druhé části článku proto diskutujeme dva příklady jednoduchých obecných (statistických spíše než kontextuálních) principů, které mohou zmíněná rozdělení podmiňovat. Jde jednak o známý mechanismus multiplikativního náhodného procesu (a jeho varianty) a dále pak tzv. prostorové analogie k principu centrální limitní věty. V rámci zmíněných prostorových analogií k centrální limitní větě je naznačena obecná souvislost mezi prostorovou autokorelací a formou statistického rozdělení řady komplexních jevů včetně regionálních systémů. Podrobnější prozkoumání charakteru tohoto vztahu představuje potenciální téma dalšího výzkumu.

V návaznosti na diskuse různé zešikmených rozdělení v prvních dvou částech článku je jeho třetí část zaměřena aplikačně. Zabývá se testováním vybraných parametrických měř

nerovnoměrností, jakožto v praxi patrně nejčastěji kvantifikovaného aspektu statistického rozdělení. Sledována je citlivost a stabilita šesti hojně používaných indikátorů vůči datům pocházejícím z různě zešíkmených rozdělení a vůči náhodným fluktuacím v těchto datech. Výsledky v tomto ohledu poukazují na značnou citlivost a nestabilitu Theilova koeficientu, průměrné logaritmované odchylky a variačního koeficientu a na druhé straně neprakticky nízkou citlivost míry heterogenity H . Z hlediska testovaných vlastností lze při předpokladu existence asymetrického rozdělení analyzovaných dat pro sledování nerovnoměrnosti (variability) doporučit použití Giniho koeficientu či Robin Hood (Hooverova) indexu.

- Obr. 1 – Rozdělení 3 141 amerických okresů (counties) podle vybraných charakteristik struktury/kvality a odpovídajících velikostních charakteristik. Vlevo nahoře – míra rozvodovosti, vpravo nahoře – počet rozvodů, vlevo dole – příjem na osobu, vpravo dole – příjem na km².
- Obr. 2 – Vývoj rozdělení zemí světa podle odhadované naděje dožití mezi lety 1800 a 2007.
- Obr. 3 – Různé způsoby znázornění rozdělení 6 258 českých obcí (grafy vlevo) a 3 141 amerických okresů (grafy vpravo) podle jejich populační velikosti. Přímkovy ve spodních grafech („log-log rank-size grafy“) odpovídají teoretickým rozdělením podle Zipfova modelu.
- Obr. 4 – Citlivost měr nerovnoměrností v ohledu k datům z různě zešíkmených rozdělení; σ na ose x označuje standardní odchylku uvažovaného lognormálního rozdělení a $I(y)$ na ose y zachycuje standardizované hodnoty uvažovaných měr nerovnoměrností.
- Obr. 5 – Stabilita měr nerovnoměrností v ohledu k náhodným fluktuacím v datech generovaných z různě zešíkmených distribucí; σ na ose x označuje standardní odchylku uvažovaného lognormálního rozdělení a $V_{I(y)}$ na ose y zachycuje relativní variabilitu jednotlivých měr nerovnoměrností (měřenou variačním koeficientem) opakovaně 1 000krát vypočítaných pro jednotlivé hodnoty σ .

Authors are with Department of Social Geography and Regional development, Faculty of Science, Charles University, Albertov 6, 128 43 Praha 2, Czechia; e-mail: pepino@natur.cuni.cz, vojtaa.nosek@seznam.cz.

Arrived to the editorial board on July 13, 2009