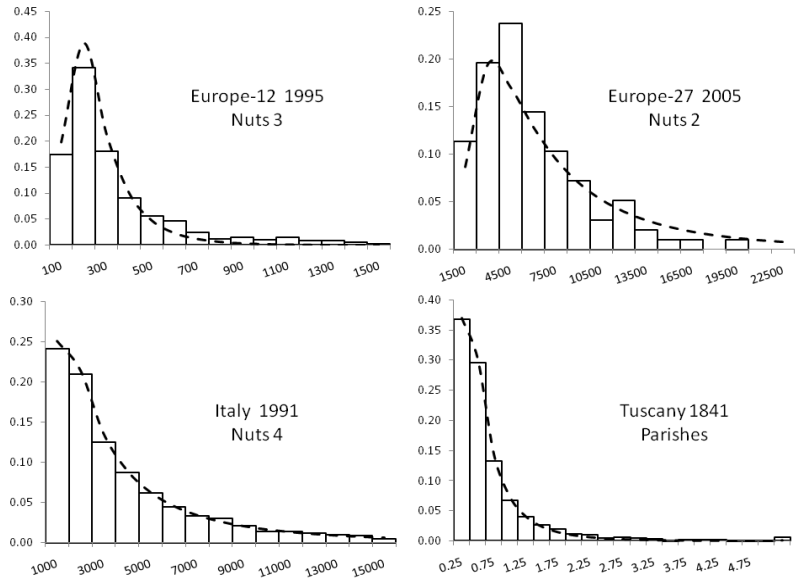
The Dynamic of the Geographical Distribution of a Population

Gustavo De Santis, Giambattista Salinari

(Dept. of Statistics, University of Firenze, Italy)

Let us consider a geographical area divided into an arbitrary number of territorial units. Of these we will consider only one characteristic: population size. As known, the frequency distribution of these units with respect to the size of their population approaches a log-normal distribution. Surprisingly enough, this conclusion is not affected by the type of unit that one uses (large or small), the level of analysis (supranational, national or regional), or the historical epoch. Figure 1, for instance, shows the frequency distribution in four different cases, taken from different epochs and using different geographical grids: a) Europe-12 in 1995, nuts 3[1]; b) Europe-27 in 2005, nuts 2; c) Italy in 1991 communes; d) Tuscany in 1841, by parish.

Figure 1 Geographical distribution of the population with different units of analysis.



Note The dashed line is the best fit of the log-normal model to the empirical distribution. On the x-axis there are population classes (in thousands); on the y-axis, probabilities and relative frequencies.
Sources: Own elaborations on data taken from Eurostat, Istat, and the 1841 census of the Grand Duchy of Tuscany.

[1] "The Nomenclature of Territorial Units for Statistics (NUTS) was established by Eurostat more than 30 years ago in order to provide a single uniform breakdown of territorial units for the production of regional statistics for the European Union" (http://ec.europa.eu/eurostat/ramon/nuts/introduction_regions_en.html)

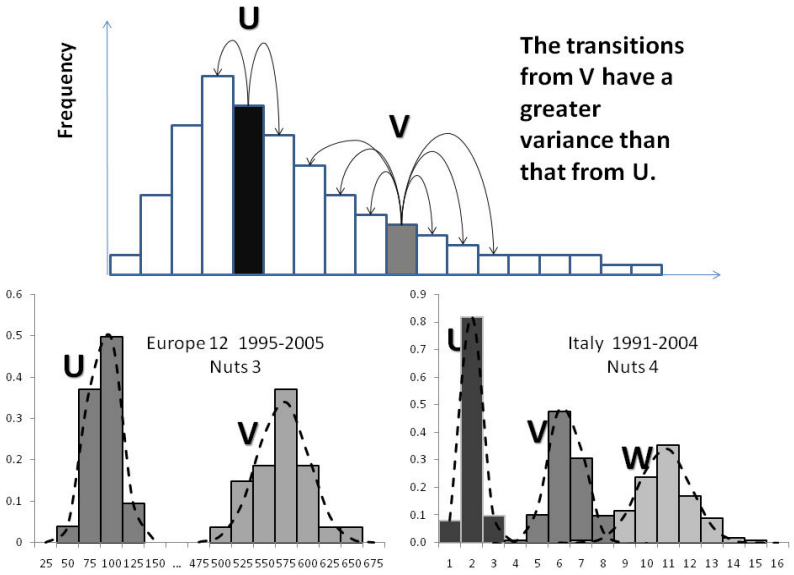
In each case, the general shape of the distribution closely fits a log-normal curve (which, typically, "explains" more than 99 per cent of the total variance), although, of course, with different parameters. The fit not perfect, however, because, as is known, the log-normal distribution tends to slightly underestimate the right tail of the empirical distribution.

What we contend - and this is our original contribution - is that these regularities are not a mere coincidence, and that in each of the four different cases it is possible to recognize the same type of general population dynamic, whose main characteristics are:

1) at a given time *t*, consider a set *U* of units with similar population (for example, the units with population ranging from 5,000 to 10,000 inhabitants). At time *t+1,* the probability distribution of this set of populations approximates a normal curve.

2) at a given time *t*, consider two distinct sets *U* and *V* of units with different population, and let the average population of *V* exceed that of *U* (for example let U represent the units ranging from 5,000 to 10,000 inhabitants, and V represent those ranging from 50,000 to 55,000). At time *t*+1, both the mean and variance of the population distribution of the *V* units are greater than those of the *U* units.

Figure 2 General aspects of the dynamics of geographical distribution.



Note. The histogram on the top represents the theoretical distribution of the territorial units of a region according to their population at time *t*. The arrows represent the transitions that take place in the time interval (*t*, *t*+1) between the different classes. On the bottom we provide two examples of actual dynamics. On the left) Europe-12: we selected the European nuts 3 that in 1995 had a population between 50,000 and 100,000 inhabitants (set U) and those between 500,000 and 550,000 inhabitants (set V). Then we computed, for the two sets of units, the distribution according to their population in 2005. On the right) Italy: we selected all the Italian communes that in 1991 had a population between 1,000 and 2,000 inhabitants (set U), those between 5,000 and 6,000 inhabitants (set V) and those between 9,000 and 10,000 inhabitants (set W). Then we computed for each of the three sets of

units the distribution according to their population in 2005. The dashed line indicates the best fit of a normal model to the empirical data.

Both empirical observations and theoretical arguments (developed in the complete paper) suggest that the relation between the mean population at time *t* (m_t) and the variance of the population at time *t*+1 (s^2_{t+1}), for each set of units *U*, is

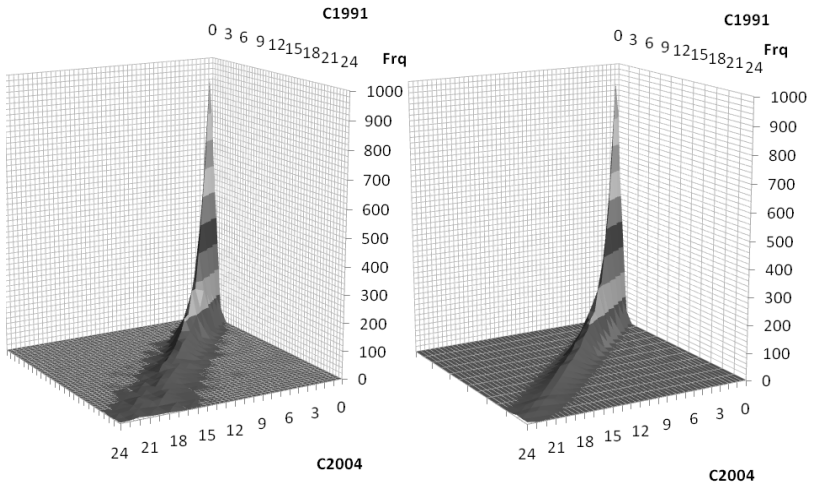
1) $$s^2_{t+1} = \frac{a}{b} m_t - \frac{a}{b^2} m_t^2$$

where *a* e *b* are parameters. With these findings, we can build a markovian model that mimics with precision, and with two parameters only, the evolution of the population of the territorial units of a given geographical area, and that explains why the log-normal model closely fits empirical geographical distributions of virtually every population. This model can be summarized by the following equation:

2) $$P_{t+1} = P_t + \mathcal{R}_{P_t}$$

Where P_t and P_{t+1} represent the population of a unit in time *t* and *t+1* respectively, and where $\mathcal{R}_{P_t}$ is a normal random variable, the parameters of which (mean and variance) both depend on the original population P_t. The variance, in particular, is greater if the average of the starting population is greater too. Incidentally, this characteristic (varying parameters) distinguishes this type of model from a first order autoregressive one, where the random variable has a constant parameter. The results of an application of this model to the evolution of Italian communes from 1991 to 2004 is shown in Figure 3.

Figure 3 Empirical and theoretical dynamic of Italian communes 1991-2004.



Actual evolution Italy 1991-2004 communes

Estimated evolution Italy 1991-2004 communes

Note. This is the graphical representation of the empirical (left) and theoretical (right) transition matrix. C1991 indicates the population class to which a unit belonged in 1991. C2004 indicates the population class to which a unit belonged in 2004. Frq indicates the number of transition from class *i* to class *j* over the 14 years considered.

The R^2 calculated comparing the theoretical matrix with the empirical one is 0.993. *Source*: Own elaboration on ISTAT data.

The goodness of fit between the empirical and the theoretical distribution is not a mere coincidence: the model applies satisfactorily to the geographical distribution of virtually all European populations in several years, and using different units of analysis (see also Figures 1 and 2). More empirical data, together with the theoretical background of this model are presented in the complete paper.

References

CASWELL H. (2001) *Matrix population models*, Sinauer Associetes, Inc. Publishers Sunderland, Massachusetts.

EECKHOUT J. (2004) *Gibrat's Law for (All) Cities*, in «The American Economic Review», vol. 94, n. 5, pp. 1429-1451.

GABAIX X. (1999a) *Zipf's Law for Cities, an Explanation*, in «Quaterly Journal of Economics», 114, pp. 739-767.

GABAIX X. (1999b) *Zipf's Law and the Growth of Cities*, in «The American Economic Review», vol. 89, n.2, pp. 129-132.

LE BRAS H. (2000) *Essai de géométrie sociale*, Editions Odile Jacobe, Paris.

NEWMAN M. E. (2003) *Power laws, Pareto distributions and zipf's law*, in «Contemporary Physics», n. 46 pp. 323-351.

PRIEUR C., SALINARI G. (forthcoming) *Social Distances or What Lies Beneath Preferential Attachment*, in «Mathematics and Social Sciences».

ZIPF G. K. (1949) *Human Behavior and the principle the Principle of Least Effort. An Introduction to Human Ecology,* Addison-Wesley, Cambridge Ma.