Cross Ranking of Cities and Regions: Population vs. Income

Roy Cerqueti[1,*] and Marcel Ausloos[2,3]

[1] University of Macerata, Department of Economics and Law,
via Crescimbeni 20, I-62100, Macerata, Italy
e-mail address: roy.cerqueti@unimc.it

[2]eHumanities group[†],
Royal Netherlands Academy of Arts and Sciences,
Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands

[3]GRAPES[‡],
rue de la Belle Jardiniere 483, B-4031, Angleur Liege, Belgium
e-mail address: marcel.ausloos@ulg.ac.be

Abstract

This paper explores the relationship between the inner economical structure of communities and their population distribution through a rank-rank analysis of official data, along statistical physics ideas within two techniques. The data is taken on Italian cities. The analysis is performed both at a global (national) and at a more local (regional) level in order to distinguish "macro" and "micro" aspects. First, the rank-size rule is found not to be a standard power law, as in many other studies, but a doubly decreasing power law. Next, the Kendall τ and the Spearman ρ rank correlation coefficients which measure pair concordance and the correlation between fluctuations in two rankings, respectively, - as a correlation function does in thermodynamics, are calculated for finding rank correlation (if any) between demography and wealth. Results show non only global disparities for the whole (country) set, but also (regional) disparities, when comparing the number of cities in regions, the number of inhabitants in cities and that in regions, as well as when comparing the aggregated tax income of the cities and that of regions. Different outliers are pointed out and justified. Interestingly, two classes of cities in the country and two classes of regions in the country are found. "Common sense" social, political, and economic considerations sustain the findings. More importantly, the methods show that they allow to distinguish communities, very clearly, when specific criteria are numerically sound. A

---

*Corresponding address: University of Macerata, Department of Economics and Law, via Crescimbeni 20, I-62100, Macerata, Italy. Tel.: +39 0733 258 3246; Fax: +39 0733 258 3205. Email: roy.cerqueti@unimc.it

[†]Associate Researcher

[‡]Group of Researchers for Applications of Physics in Economy and Sociology

specific modeling for the findings is presented, i.e. for the doubly decreasing power law and the two phase system, based on statistics theory, e.g., urn filling. The model ideas can be expected to hold when similar rank relationship features are observed in fields. It is emphasized that the analysis makes more sense than one through a Pearson II value-value correlation analysis.

1 Introduction

Research on rank-size relationships has a long history and has been applied in a wide range of contexts. In this respect, at the inception Zipf's law [1] was illustrated through linguistics considerations, while Pareto's law – a similar hyperbolic power law – finds its origin in finance [2].
However, among several applications, rank-size theory has a prominence in the field of urban studies [1, 3, 4, 5, 6, 7]. In particular, a relevant role is played by the analysis of the geographical-economical variables in the conceptualization of the New Economic Geography, introduced by Krugman [8]. In this respect, see Berry [9], Pianegonda and Iglesias [10] and the extensive surveys of Ottaviano and Puga [11], Fujita et al. [12], Neary [13], Baldwin et al. [14], and Fujita and Mori [15]. Such an analysis can be satisfactorily developed through the study of the rank-size rule for regional and urban areas, on the basis of economical variables and the population size distribution, as shown here below. The interested reader is referred to the monograph of Chakrabarti et al. [16] for outlining interdisciplinary socio-econo-physics points of view.

Several studies proved empirically the validity of Zipf's law [1] (or type-I Pareto distribution [17]): Rosen and Resnick [18], in 1980, analyzed data from 44 Countries, and found a clear predominance of statistical significance of Zipf's law, with R^2 greater than 0.9 (except in one case, Thailand); in Mills and Hamilton [19], data from US city sizes, in 1990, has been taken to show the evidence of Zipf's law ($R^2 \sim 0.99$); see Guerin-Pace [20] also. Other papers, after 2000, which substantially support this type of rank-size rule are Dobkins and Ioannides [21], Song and Zhang [22], Ioannides and Overman [23], Gabaix and Ioannides [24] Reed [25], and Dimitrova and Ausloos [26] more recently, but with warnings [1], just to cite a few. Nitsch [29] provides an exhaustive literature review up to 2005.

As other counter-examples, beside the above-mentioned case of Thailand in [18] and Bulgaria in [26], weak agreement between data and Pareto fit is sometimes pointed out. Peng [30] found a Pareto coefficient of 0.84, not quite close to 1, - when implementing a best fit of data on Chinese city sizes in 1999-2004 with the Pareto distribution. Ioannides and Skouras [31], among others, have argued that Pareto-Zipf's law seems to stand in force only in the tail of

---

[1]In a practically modeling approach, Dimitrova and Ausloos [26] indicated through the notion of the global primacy index of Sheppard [27] that Gibrat (growth) law [28], supposedly at the origin of Zipf's law, in fact, does not hold in the case of Bulgaria cities, but can be valid when selecting various city classes (large or small sizes).

the data distribution. Matlaba et al. [32] also provided such an "evidence that, at least for the analyzed case of Brazilian urban areas over a spectacularly wide period (1907-2008), Zipf's law is clearly rejected. Soo [33] has also empirically shown that the size of Malaysian cities cannot be plotted according to such a rank-size rule, but a suitable collection of them can do it.

A list of other contributions on the inconsistency of Zipf's law in several countries, different periods and under specific economic conditions should include [34], [35] and [36]. Of particular interest, in the present case, is also Garmestani et al. [37], who conducted an analysis for the USA at a regional level. Thus, it seems that the failure of Zipf's law may often depend on the way data are grouped [26, 40]

From the present state of the art point of view, regional agglomerations, commonly ranked in terms of population, may be also sorted out in an order dealing with the economic variables. In fact, Zipf's law is sometimes identified in some "economic" way to rank. As an example, Skipper [38] used such a rank-size relationship to detect well developed countries hierarchy through their national GDP. This result has been also achieved by Cristelli et al. [39], who exhibited evidence of the Zipf's law for the top fifty richest countries in the 1900-2008 time interval.

In fine, the investigation, in the main text, aims to provide some newness, through some recent data; even leading to a better description of a rank-size rule than the Pareto-Zipf's law. Beside, to our knowledge, no statistical evidence of Zipf's law studies has been reported for the economic variables characterizing Italian cities and regions, in the period 2007-2011.

Along such lines of thought and within our statistical physics framework, the paper deals with the rank-size rule for the entire set of municipalities in Italy (IT, hereafter). This country, Italy, is expected, according to its fame, to provide what a physicist looks for, i.e. some universality features but also some non universal ones. Therefore, in an aim toward understanding nature and progressing toward reconciling so called hard and soft science, reliable data is investigated looking for "universality" and "deviations". The IT data are both official, and are given by aggregated income tax (ATI) and number of inhabitants : the former has been provided directly from the Research Center of the Italian Ministry of Economics and Finance (MEF), and cover the quinquennium 2007-2011; the latter comes from the Census 2011, which has been performed by the Italian Institute of Statistics (ISTAT).

Therefore, the *size* to be examined is here defined through two criteria: (i) by the ATI *contribution* that each city has given to the Italian GDP and (ii) by the *population* of each city. First, the 8092 Italian cities are yearly ranked according to such variables. Their related classifications are then compared: (i) at the national level, but also (ii) at the regional level. There are 20 regions in IT with a varied number of cities.

Each specific year of the quinquennium has been examined. However, special attention has been paid to 2011, in order to be somewhat consistent with the year concerned by the Italian census report on population. The census took several years in fact. Therefore, only the ATI averaged over the 5 years interval for each municipality is reported in the main text and discussed through the mean value over 5 years of the yearly ATI data. The conclusions are unaffected, as discussed in an Appendix, except for some mild change in error bars on the numerical parameters, when specific years are selected for examination.

3

Within the statistical physics approach interested in correlation functions, the paper also aims at observing whether there is or not some correlation between ATI and population rankings. For this aim, the Kendall τ and the Spearman ρ rank correlation measures have been computed [41, 42, 43, 44]. The Kendall τ measure compares the number of concordant and non-concordant pairs, to detect the presence of singularities in a possible relationship (here between economy and demography). The Spearman ρ rank correlation is also computed, to provide a more satisfactory interpretation of the relationship between economy and demography in terms of rank [45].

A rank scatter plot of the number of inhabitants in each 8092 city versus the corresponding ATI reveals two regimes. Moreover, two regimes are also observed for the value themselves, distinguishing different "phase states", pointing to regional specificities. It can be stressed [42] that such a rank-rank analysis, implying pair correlations, is the analog of a correlation function, sometimes called susceptibility, in the linear response theory of statistical mechanics. It will be indicated that a "rank" is like a "temperature". A useful methodological paper to read on the rank-rank correlation is by Melucci [46]. It contains also a bibliographic review on previous researches on this theme, up to that time.

In order to obtain analytical expressions simulating the data, whence suggesting a model susceptible of general applications, various simple fits have been attempted. The most classical one, for universality purpose, is the straight line fit on a log-log plot. However, both for the population size and the ATI data, it turns out that, in each region, the main city is an outlier[2]: more precisely 25 cities for the whole country, i.e. about 1 per region. These lowest rank cities are markedly found to occur much above the usual expecting straight line data fit on a log-log plot (the 20 "regional" plots are not shown for saving space). Moreover, the fit visual appearances are not exciting, because our eyes receive the same effects from the low and high rank ranges. Practically, it has been found that the regression coefficient R^2 improves if one removes these outliers. Moreover, the fits visually improve in the asymptotic regimes, - which are very narrow regions, in particular in the high rank range. Therefore, for shortening the paper, the parameters of the fits reported here below only concern fits with a 3-parameter function, discussed below.

To get some perspective, notice that several contributions in the literature propose rank-rank analysis types within different contexts, - all comparing two different rank rules. In a series of papers [47, 48, 49], country ranking due to soccer team ranking (and performance) in UEFA competitions is presented and contrasted with FIFA ranking. In particular, in dealing with the rank-rank correlation [48], the Kendall τ is employed. Interesting is also the application in the context of archeology, with a specific focus on the Aztec settlement distributions, presented in Hare [50]. In [51], Zhou et al. focus on the rank-rank correlation for scientists and scientific journal, in line with the scientometrics literature. In this respect, see also [52], - on the relationship between authors and coauthors, and Stallings et al. [53], - comparing researchers and universities according to different criteria, whence providing also an axiomatization of the rank and rank correlation problems. For what concerns the Spearman ρ coefficient, it measures correlations in the rank deviation from their mean of the measurements. As already said above, for completeness, ρ will be calculated,

_____
[2]This was observed already by Jefferson [3].

even if Kendall τ is usually acknowledged to have better statistical properties than Spearman's ρ [45, 54, 55, 57, 56]. The interpretation of Kendall τ also seems to be easier and more intuitive than that of the Spearman ρ. In this respect, it is important to point out that the average of the ranks is equal to $N/2$, thus representing a measure of the sample size N.

It is of common knowledge that the Pearson Π correlation coefficient is the most frequently used measure when data is supposed to be (or is) normally distributed. On the other hand, nonparametric methods such as Spearman's rank-order and Kendall's correlation coefficients are usually suggested to be better for non-normal data analysis. In fact, all three correlation coefficients are proportional to differently weighted averages of the concordance indicators. The normality constraint is briefly examined in an Appendix, together with a discussion of the Pearson coefficient interest in the present context. It is briefly argued that the Kendall τ brings some information of interest. We consider that this is particularly so, when values are of different natures and are measured with a quite often unknown error bar.

In the field of Economic Geography, the rank-rank analysis is quite neglected. A noteworthy exception is Rappoport [58], where population densities and consumption amenities are compared and discussed for U.S. economical-demographical data. The paper of Mori and Smith [59] is also of interest, in that it focuses on the link between economics and demography at a city level by investigating the number of cities inhabitants in presence of established industries. However, to the best of our knowledge, the main text below is the first paper dealing with the application of this theory for discussing the relationship between (Italian) demographical and economical reality.

It is also important to note that the employment of microdata allows to emphasize Italian regional disparities. In this respect, we address the interested reader to De Groot et al. [60] and Melo et al. [61].

In short, the paper is organized as follows: Section 2 contains the description of the data. Section 3 is devoted to the analysis of the whole IT. Two measures, the Kendall's τ and the Spearman's ρ rank correlation coefficients, are proposed and their respective interest discussed. The findings are collected and reworded in Section 4. Such section includes also Subsection 4.1, which serves to emphasize the regularities and disparities between the Italian regions. Thereafter, a specific modeling based on urn filling statistics, is presented in Section 5. It can be expected to hold when rank relationship features similar to those of our findings are observed. The last section (Section 6) serves for conclusions and for offering suggestions for further research lines.

There are three Appendices: App. A contains a technical detail note on data aggregation, as already mentioned, arising from the change in the number of cities in Italy during the ATI measurements. App. B is a short investigation on the (as also pointed here above) negligible, in fact, time dependence. App. C contains a note on the Pearson coefficient and some argument in favor of considering rank-rank correlations instead of value-value correlations.

2 Data

The population data comes from the 2011 Census performed by the Italian Institute of Statistics (ISTAT)[3]. The economic data has been obtained by and from the Research Center of the Italian Ministry of Economics and Finance. Population data consist of number of inhabitants, while economic data are given by IT GDP for recent five years (2007-2011) (the aggregated tax income - ATI). In both cases, we have disaggregated contributions at a municipal level (in IT a *municipality* or *city* is denoted as *comune*, - plural *comuni*).

To provide some better understanding of the paper aims and results, the IT administrative structure is here first described.

IT is composed of 20 regions, more than 100 provinces and more than 8000 municipalities. Each municipality belongs to one and only one province; each province is contained in one and only one region. Administrative modifications due to the IT political system has led to a varying number of provinces and municipalities during the examined quinquennium. The number of cities in each administrative entity has also changed, but the number of regions has been constantly equal to 20.

Therefore, the available yearly ATI data corresponds to a different number of cities. In particular, the number of cities has been yearly evolving respectively as follows : 8101, 8094, 8094, 8092, 8092, - from 2007 till 2011. Details are given in Appendix A.

The number of cities in each region is given as a function of time in Table 1. For making sense, it is necessary to compare identical lists. We have considered this latest 2011 "count" as the basic one. Therefore, we have taken into account a virtual merging of cities, in the appropriate (previous to 2011) years, according to IT administrative law statements (see also $http://www.comuni-italiani.it/regioni.html$).

In the same spirit, the ATI of the resulting cities (and regions) have been linearly adapted, as if these ATI were existing before the merging or city phagocytosis.

A summary of the statistical characteristics for the ATI of all ($r_M \equiv N = 8092$) IT cities in 2007-2011 is reported in Table 2. For a statistical overview of the Italian structure, at the regional level, see Table 3.

It is also worth noting that there is some change in the ATI rank of a city as times goes by. Care was taken that the arithmetics pertain to the same city, when a sum or average was made. For example, there are twice 3 cities with the same name in IT; we carefully distinguished them.

---

[3]Census is an official statistical exploration of the Italian population. It is based on the responses provided by all the Italians, and it is performed every 10 years. However, there were Irregularities: in 1891 and 1941 Census has been not performed (for financial distress in the former case, and due to the Second World War in the latter one), but an adjunctive Census has been provided in 1936. The next Census will be in 2021.

region	$N_{c,r}$		
year:	2007	2009	2011
Lombardia	1546	1546	↓ 1544
Piemonte	1206		
Veneto	581		
Campania	551		
Calabria	409		
Sicilia	390		
Lazio	378		
Sardegna	377		
Emilia-Romagna	341	341	↑ 348
Trentino-Alto Adige	339	↓ 333	333
Abruzzo	305		
Toscana	287		
Puglia	258		
Marche	246	246	↓ 239
Liguria	235		
Friuli-Venezia Giulia	219	↓ 218	218
Molise	136		
Basilicata	131		
Umbria	92		
Valle d'Aosta	74		
Total	8101	↓ 8094	↓ 8092

Table 1: Number N of ($8092 \equiv r_M$) cities in 2011, and in previous years, in the (20) IT regions; such a region ranking by city number corresponds to that illustrated in Fig. 1.

	2007	2008	2009	2010	2011	5yr av.
min. (x10^{-5})	3.0455	2.9914	3.0909	3.6083	3.3479	3.3219
Max. (x10^{-10})	4.3590	4.4360	4.4777	4.5413	4.5490	4.4726
Sum (x10^{-11})	6.8947	7.0427	7.0600	7.1426	7.2184	7.0738
Max. range (r_M)	8092	8092	8092	8092	8092	8092
mean (μ) (x10^{-7})	8.5204	8.7033	8.7248	8.8267	8.9204	8.7417
median (m) (x10^{-7})	2.2875	2.3553	2.3777	2.4055	2.4601	2.3828
RMS (x10^{-8})	6.5629	6.6598	6.6640	6.7531	6.7701	6.682
Std. Dev. (σ) (x10^{-8})	6.5078	6.6031	6.6070	6.6956	6.7115	6.6256
Var. (x10^{-17})	4.2351	4.3601	4.3653	4.4831	4.5044	4.3899
Std. Err. (x10^{-6})	7.2344	7.3404	7.3448	7.4432	7.4609	7.3654
Skewness	48.685	48.855	49.266	49.414	49.490	49.126
Kurtosis	2898.7	2920.42	2978.1	2991.0	2994.7	2955.2
μ/σ	0.1309	0.1318	0.1321	0.1319	0.1329	0.1319
$3(\mu - m)/\sigma$	0.2873	0.2884	0.2883	0.2878	0.2889	0.2879

Table 2: Summary of (rounded) statistical characteristics for ATI (in EUR) of IT cities ($N = 8092$) in 2007-2011.

	$N_{c,r}$
Minimum	74
Maximum	1544
Mean (μ)	404.6
Median (m)	319
RMS	536.998
Std Dev (σ)	362.253
Variance	131 227.52
Std Error	81.0023
Skewness	2.1284
Kurtosis	3.8693
μ/σ	1.117
$3(\mu - m)/\sigma$	0.7089

Table 3: Summary of (rounded) statistical characteristics for the number (N_c = 8092) distribution of IT cities in the various regions ($N_r = 20$) in 2011. The maximum $N_{c,r}$ is 1544 (Lombardia) and the minimum is 74 (Valle d'Aosta), - see Table 1.
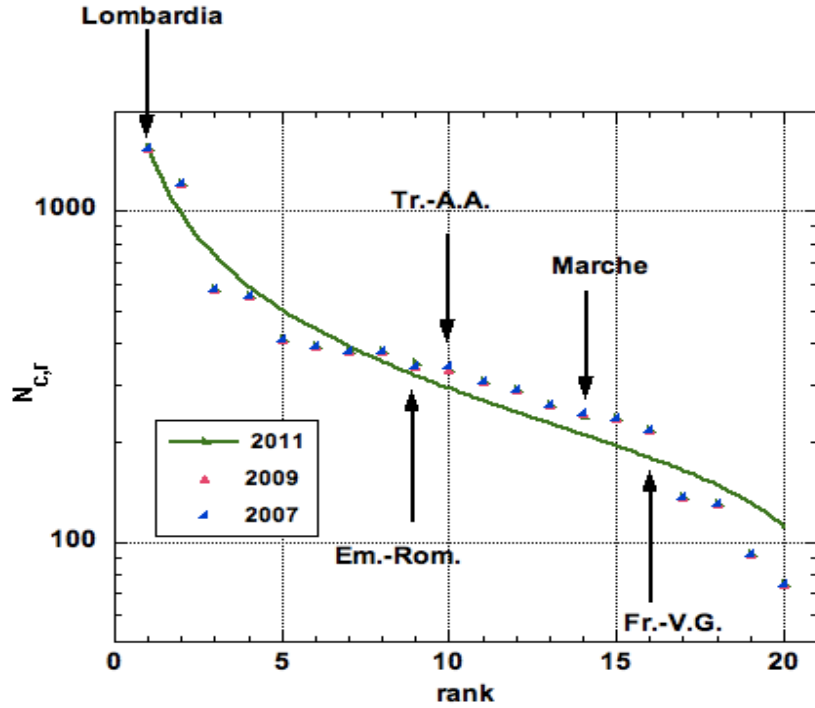
Figure 1: $N_{c,r}$ $vs.$ rank of the region for the years of the quinquennium; the regions having a change in the number of cities are indicated by an arrow ↑ or ↓; the arrow direction is according to the change in $N_{c,r}$ in some year as mentioned in Table 1. The fit corresponds to the function Eq. (3.1); the fit parameters are given in the text.
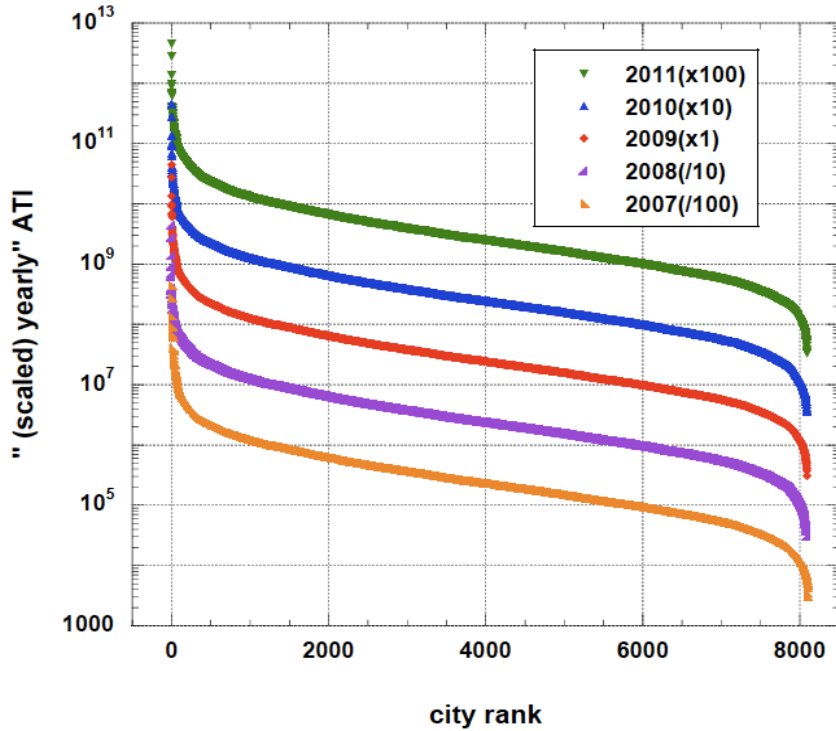
Figure 2: Semi-log plot of the 2007-2011 yearly ATI of the 8902 IT cities ranked according to their "income tax" importance every year; the data is rescaled by a factor 10 or 100, as indicated in the insert, for better visibility. The inflection point is well seen near $r_M/2 \sim 4000$.

	(i) (8092)	(ii) (20)
$p + q$	32 736 186	190
$p - q$	27 778 116	148
p	30 256 042	169
q	2 480 144	21
Kendall τ	0.849	0.779
Spearman ρ	0.9637	0.9098
Pearson Π	0.9849	0.9787

Table 4: Kendall τ, Eq. (3.2) and Spearman ρ, Eq. (3.5) correlation statistics of ranking order between the Number of inhabitants, (i) in (8092) cities or (ii) in (20) regions, according to the 2011 Census, and the corresponding averaged ATI, over the period $2007 - 2011$; the Pearson Π value-value correlation coefficient is given for completeness.
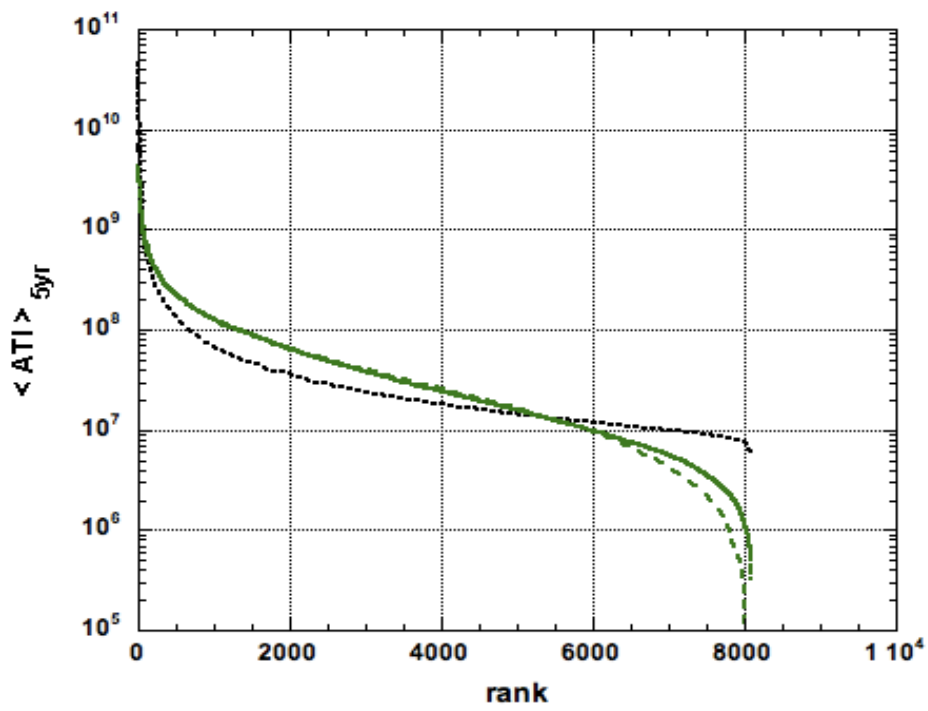
Figure 3: Semi-log plot of the rank-size relationship between each Italian city $< ATI >$ (averaged for the examined quinquenium) and its rank; the black dot line corresponds to the whole (8092) data; the green dash line corresponds to the whole data minus the top 8 city outliers.
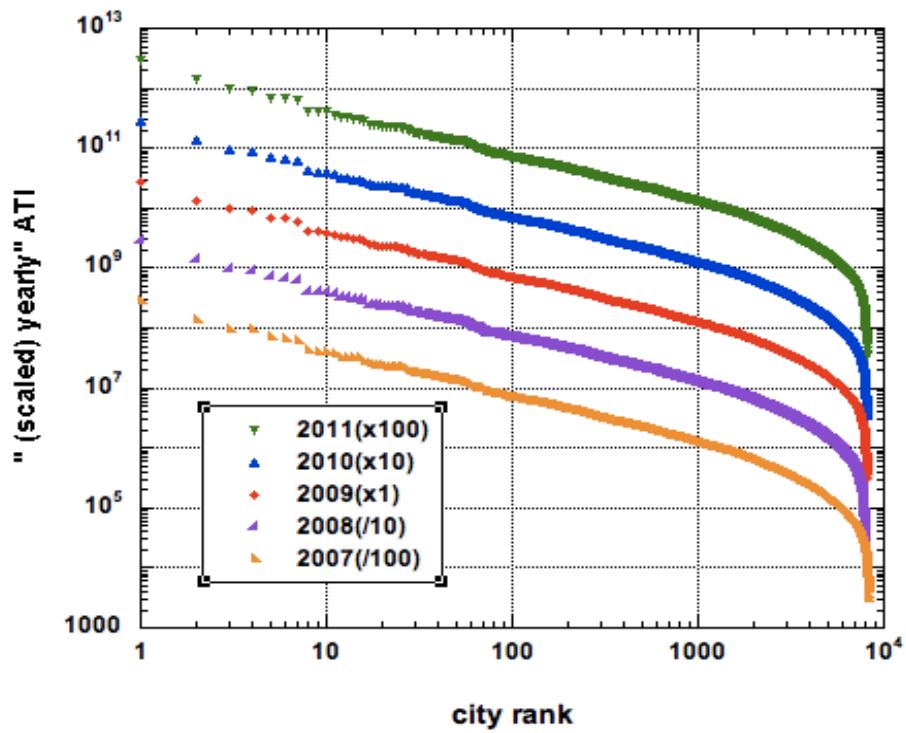
Figure 4: Log-log plot of the 2007-2011 yearly ATI of the 8902 IT cities ranked according to their "income tax" importance every year; the data is rescaled by a factor 10 or 100, as indicated in the insert, for better visibility. The outliers are here better emphasized than on Fig. 2 but the inflection point near $r_M/2 \sim 4000$ not so obvious.

3 City population size and ATI rank order distributions in IT

In this section, *size* is defined under two criteria: either (i) the economic one (averaged ATI over the period 2007-2011)[4] or (ii) the demographic one (population in 2011). The empirical rank-size relationships are first looked for, see Sect. 3.1. The Kendall τ coefficient is next used to compare rank pairing under both criteria, in Sect. 3.2. The Spearman ρ coefficient is next calculated and compared under both criteria, in Sect. 3.2. The Pearson Π coefficient is discussed in App. C.

3.1 Rank-size relationships

We have ranked the regions in decreasing order, according to their respective number of cities. In general, the central part of the data looks like being well fitted by a power law with an exponent a little bit below (-1). However, at low rank, there is usually a jump, while at high rank, there is a sharply marked downward curvature.

Therefore, a rank-size rule fit was attempted with a doubly decreasing power law in order to obtain an inflection point near the center of the data range, i.e. with the analytical form [62]

$$y(r) \;=\; A\, m_1\, r^{-m_2}\, (N - r + 1)^{m_3}, \qquad (3.1)$$

where r is the rank, A is an order of magnitude amplitude, *a-priori* imposed and adapted to the data, without loss of generality, for smoother convergence of the non-linear fit process, and N is the number of regions, of course. The best 3-fit parameters, for $A = 10^3$ and $N = 8092$, have been so obtained: $m_1 = 0.847$; $m_2 = 0.68$; $m_3 = 0.209$: for a regression coefficient $R^2 = 0.957$, and a $\chi^2 \geq$ 106 013, indicating a quite good agreement with the above equation (Fig. 1). Some further discussion on some meaning of the parameters m_1, m_2, and m_3 is postponed to Sect. 5.

A similar fit study, made for the "regional ATI", is given in Fig. 2, on a semi-log plot for each year. The behavior being visually similar to that of Fig. 1 suggests to use Eq. (3.1) as well for further study on economic data.

Next, we made an unweighted average, over the quinquenium, of each city ATI. The rank-size relationship was looked for in Fig. 3. The best fit parameters, for the function in Eq. (3.1), for $A = 10^6$ and $N = 8092$, are $m_1 \sim 27332$; $m_2 \sim 0.938$; $m_3 \sim 1.05$: for a regression coefficient $R^2 = 0.985$, and a $\chi^2 \geq 10^{19}$, indicating a quite good agreement with the above equation (Fig. 3). However, the fit is not very visually appealing. It can be observed on a log-log plot (Fig. 4) that a few big cities (Roma, Milano, Torino, Genova, Napoli, Bologna, Palermo and Firenze) appear as outliers. We have removed these outliers from the overall fits. When the top 8 cities are removed, the best fit leads to $m_1 \sim 1.725$; $m_2 \sim 0.725$; $m_3 \sim 0.055$: for a regression coefficient $R^2 = 0.998$, and a $\chi^2 \geq 10^{15}$ (Fig. 3). The fit is better and more visually appealing.

---

[4]In the Appendix B it is verified that taking the average of the ATI over the considered quinquennium does not bias the analysis.

3.2 Kendall τ coefficient

The Kendall τ measure is hereby discussed. Such a statistical indicator compares the number of concordant pairs p and non-concordant pairs q, i.e. how many times a city occurs, or not, at the same ranks in both (necessarily equal size) lists. This measure is a usual correlation coefficient which allows to find whether the ranking of different measurements possesses some regularity. In other words, the Kendall τ coefficient, measuring the cross correlation between two equal size data, is like the cross-correlation function of two equal size time series [41, 42]. The Kendall τ, thus like the Pearson correlation coefficient, allows a connection with statistical physics theory: in particular, it is an apparatus similar to the linear response theory correlation coefficients. For being more precise, notice that τ is like the off-diagonal generalized susceptibilities, in linear response theory [63, 64], in condensed matter [42], since the variables are two different "fluctuations", an economic and a demographic one here.

By definition,

$$\tau = \frac{p - q}{p + q}, \tag{3.2}$$

thus suggesting how stable the ranking is. Of course, $p+q = N(N-1)/2$, where N is the number of cities (8092 here), or the number of regions (20), in the two (necessarily equal size) sets; thus, $p + q = 32\ 736\ 186$ (cities) or $p + q = 190$ (regions).

For the computation of the Kendall τ, i.e. to find p, q, and $p - q$, e.g., see [65], the procedure in a stepwise form (the case of cities is only outlined) is the following:

- make a 2 column Table: the municipality name in column (1) and the average ATI in column (2);

- do the same for the population data in column (4) and the municipality name in column (3);

- rank the cities in column (1) according to their average ATI, $r_{<ATI>}$, in column (2), for example in decreasing value order;

- rank the cities in column (3) according to their population size, r_{Ninhab} in column (4), also in decreasing order;

- compare the position of cities ("ranked" columns (1) and (3)), i.e. find out how many times cities occur at the same ranks in both ordering (one obtains p) or at different ranks (for q).

Values of the Kendall τ, Eq. (3.2), Z, Eq. (3.3), and other pertinent data for the correlations between the number of inhabitants, according to the 2011 census, and the average ATI over the quinquenium ($< 2007 - 2011 >$) are given in Table 4, as obtained from Wessa algorithm [65]. Observe from Table 4 that $\tau \sim 0.85$.

From a purely statistical perspective, under the null hypothesis of independence of the rank sets, the sampling would have an expected value $\tau = 0$. For large samples, it is common to use an approximation to the normal distribution,

with mean zero and variance, in order to emphasize the coefficient τ significance, through calculating:

$$Z = \frac{\tau}{\sigma_\tau} \equiv \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}}. \tag{3.3}$$

Here, in the case of cities, $N = 8092$ and $\sigma_\tau = 0.00741$. Note that $\tau \simeq 0.85$ ($\simeq 1$) and $Z \simeq 115$. Thus, it can be concluded that there is a strong regularity in the pair ranking from a mere statistical point of view, – even though there are different regimes. In a thermodynamic sense, the system presents different phases.

3.3 Spearman's rank correlation coefficients

This section contains the computation of the Spearman ρ and the related discussion.

It is firstly needed to recall the definition of the Pearson coefficient, as the ratio of the covariance of two variables x and y to the product of their respective standard deviations, i.e.

$$\Pi = \frac{\Sigma xy - N\ (\Sigma x)(\Sigma y)}{\sqrt{[(\Sigma x)^2 - N\ \Sigma x^2].[(\Sigma y)^2 - N\ \Sigma y^2]}} \tag{3.4}$$

in usual notations.

It is easy to show that the Pearson coefficient measures the correlations between deviations form the mean, i.e. correlations between fluctuations, like the transport coefficients in linear response theory. In the present case, Π, like τ, corresponds to the off-diagonal terms. Thus, it has also some direct statistical physics appeal.

The Spearman's rank-order correlation coefficient ρ is the rank-based version of the Pearson correlation coefficient, i.e. the values x and y of the measured quantities are replaced by their corresponding rank in Eq. (3.4) (for computing the ranks, see the first four bullets of the algorithm listed in the previous Section):

$$\rho = \frac{\Sigma r_x r_y - N\ (\Sigma r_x)(\Sigma r_y)}{\sqrt{[(\Sigma r_x)^2 - N\ \Sigma r_x^2].[(\Sigma r_y)^2 - N\ \Sigma r_y^2]}} \equiv \frac{\Sigma(r_x - <r_x>)(r_y - <r_y>)}{\sqrt{\Sigma(r_x - <r_x>)^2 \Sigma(r_y - <r_y>)^2}} \tag{3.5}$$

It is worth noting that, except for the product of the rank fluctuations appearing in the Spearman's form of Eq. (3.4), the other terms are simply related to the number N of measurements; e.g. $\Sigma r_y = N(N+1)/2$. In contrast, Kendall τ reflects the number of concordances and discordances regardless of the cardinality of the dataset, hence being a sort of probability measure. Necessarily, Kendall's τ seems to contain more information on the distribution, and seems more reliable in view of a statistical conclusion: indeed, a few incorrect value data have less influence on the number of wrongly discordant pairs than the wrong absolute values would have on a Pearson, whence Spearman also, coefficient, – especially for finite size samples [66].

The Spearman coefficient has been calculated both at a municipal level ($\rho \sim 0.9637$) and at a regional level ($\rho \sim 0.9098$), see Table 4.
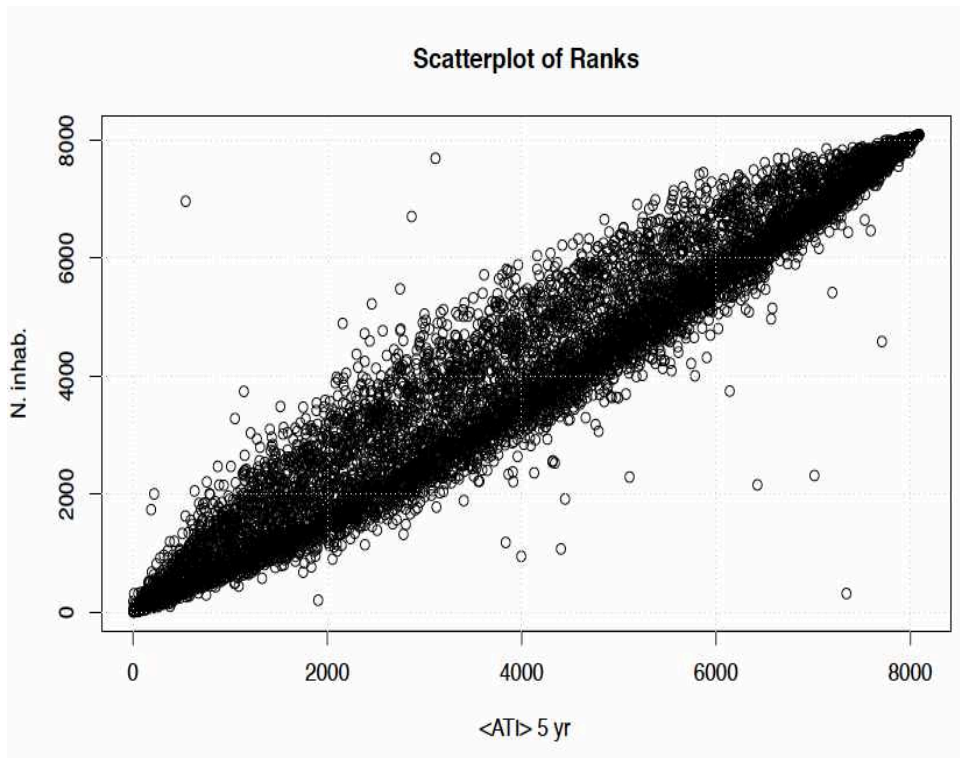
Figure 5: Scatter plot of the city ranks for $<ATI>$ (averaged for the examined quinquennium) and the number of inhabitants in each Italian city.

High values have been found, as expected. In fact, usually Spearman ρ has a larger value than that of Kendall τ, and in our computations Kendall τ is rather large (see Table 4). Therefore, Spearman ρ confirms Kendall τ outcomes on the regularity between economic and demographic data both at a municipal as well as at a regional level.

4   Results and discussion

In this Section, the empirical findings are commented upon taking into account numerical, economic, historical, demographic and political considerations.

The scatter plot of cities rank-rank correlation (average ATI *vs.* population) is shown in Fig. 5, as obtained using [65]. A large number of cities are found to have approximately the same rank, within the elongated cloud of points, but there are marked deviations. The main "inertia axis" can be obtained: it reads: $r_{Ninhab} = 178.35(\pm 15) + 0.956(\pm 0.003) \, r_{<ATI>}$. Some deviation from symmetry along the inertia axis is observed. A fine statistical analysis has shown us that the difference distribution $r_{Ninhab} - r_{<ATI>}$ is slightly negative skewed; skewness ~ -0.57; the median $= 92$. Neglecting the outlier tails, the distribution presents a smooth variation on the negative $r_{Ninhab} - r_{<ATI>}$ side, followed by sharp peak in the near 0 regime, itself followed by a sharp decay on the $r_{Ninhab} - r_{<ATI>}$ range. This implies that the probability to find a

15

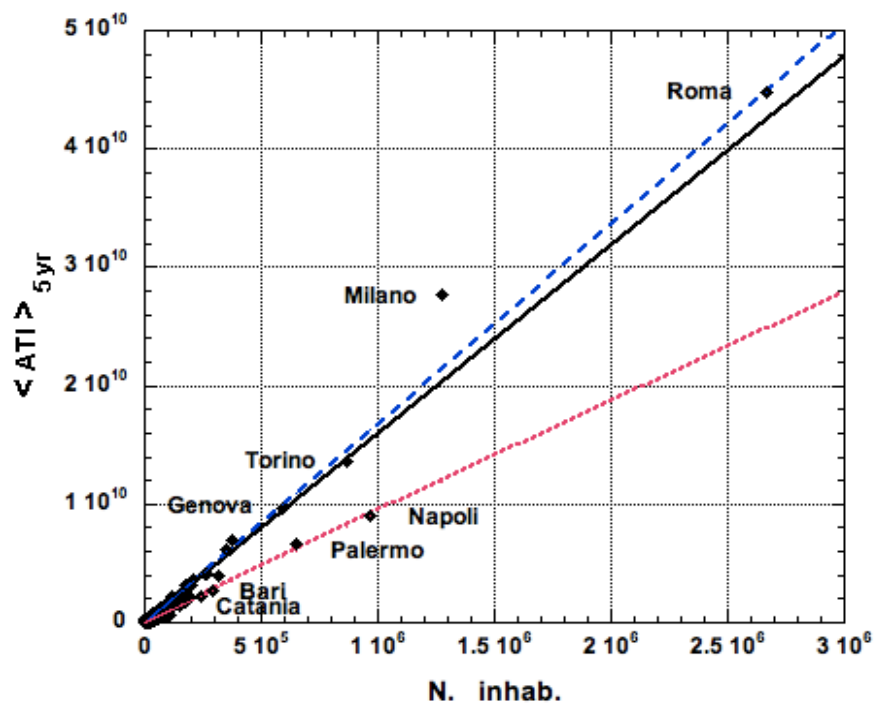Figure 6: Scatter plot of the $< ATI >$ (averaged for the examined quinquennium) and the number of inhabitants in all Italian cities; two sets of cities are emphasized from linear fits.
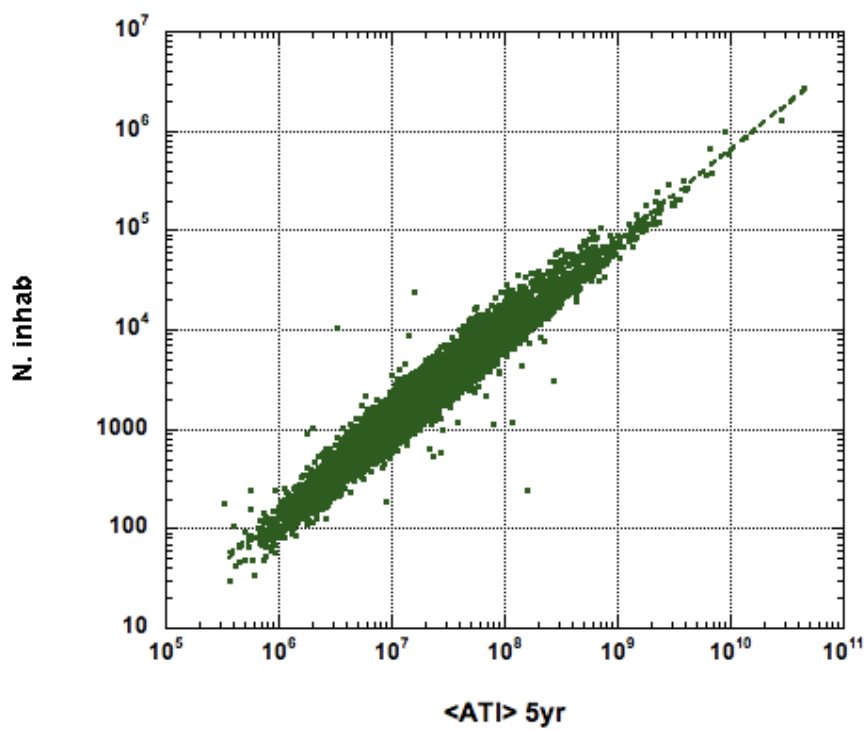
Figure 7: Log-log scatter plot of the number of inhabitants and the $< ATI >$ (averaged for the examined quinquennium) in Italian cities; the main inertia axis is indicated.
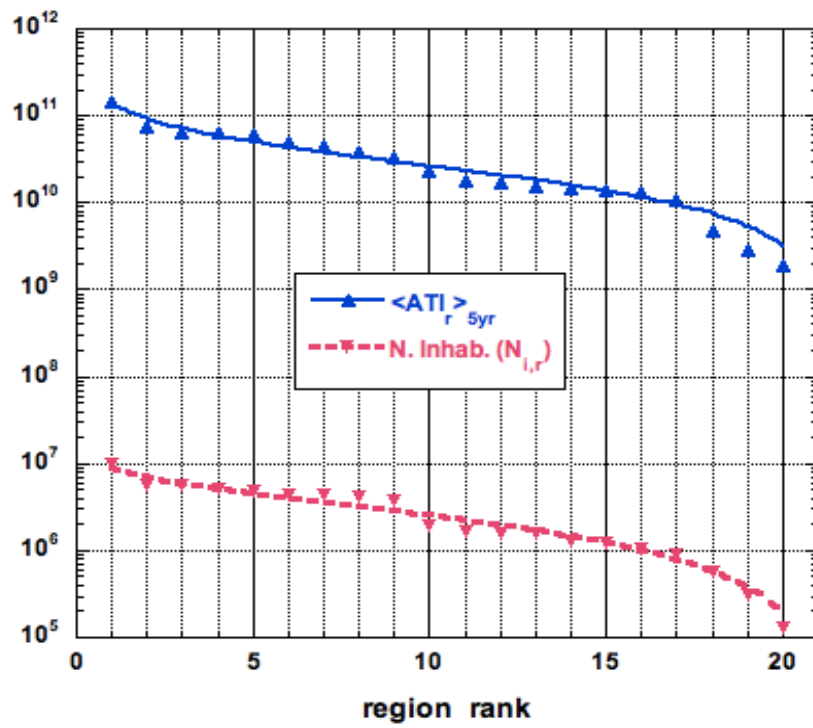
Figure 8: $N_{i,r}$ in IT regions and averaged regional ATI *vs.* the rank of the region for the years of the quinquennium. The fits correspond to the function Eq. (3.1); the fit parameters are given in the text.
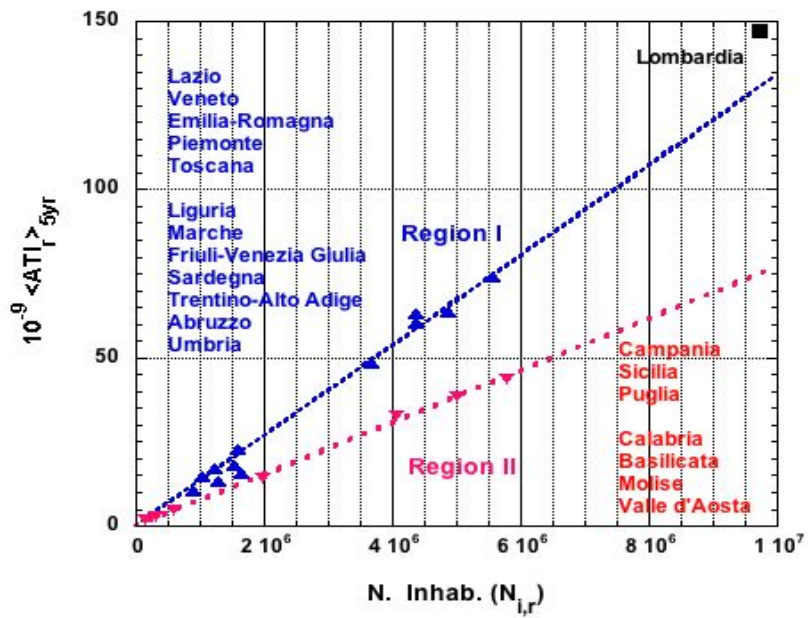
Figure 9: Scatter plot of the region $< ATI_r >$ and the number of inhabitants $(N_{i,r})$ in Italian cities of the different regions; two sets of regions are emphasized from linear fits, - plus the outlier Lombardia.
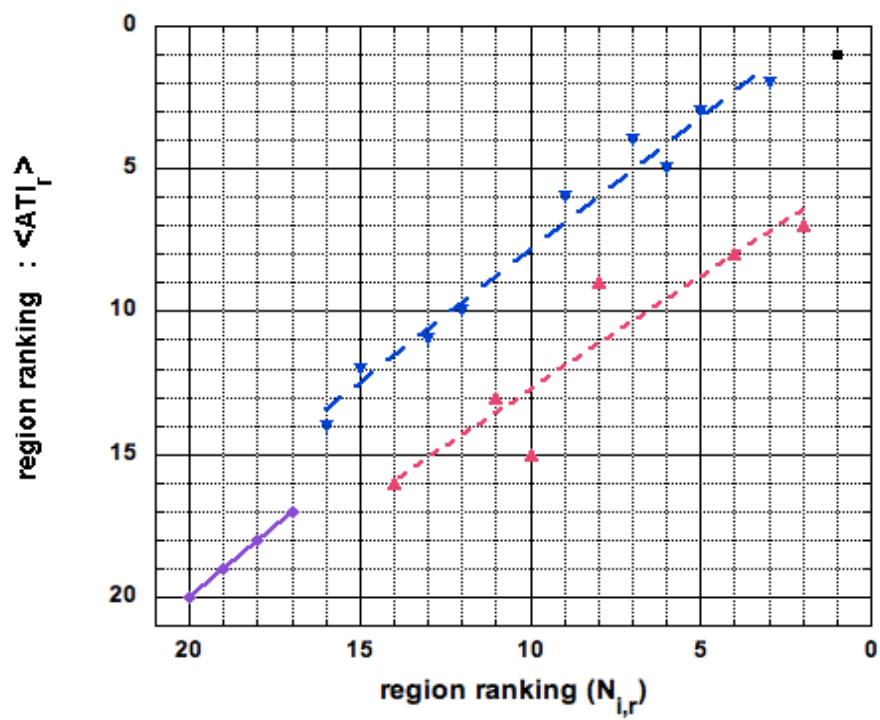
Figure 10: Scatter plot of the region rank for $< ATI >$ with respect to the number of inhabitants in a region rank; sets of regions are emphasized by linear best fits.

$r_{Ninhab} \leq r_{<ATI>}$ is about 40%. This suggests a superposition of two homogeneous/similar distributions.

It is also of interest to observe the scatter plot of the $< ATI >$ (averaged for the examined quinquennium) and the number of inhabitants in all Italian cities; this is shown in Fig. 6. Some structure inside the cloud of data points can be emphasized: two sets of cities seem existing. This is finely shown by visually distinguishing two sets of data points, and subsequent linear fits: one has (i) (blue dash line) $y = 16\ 791.15\ x$, and (ii) (red dot line) $y = 9\ 311.28\ x$. An overall fit gives the proportionality (iii) (black continuous line) $y = 15\ 942.30\ x$. The fits cannot be compared through their regression coefficient; they are all close to 0.96, but we emphasize that the visual inspection leads to some evidence. Notice that (i) Milano appears to be an outlier; (ii) the red dot line seems to point to a set of cities "from the South"; (iii) in contrast to the blue dash line, pointing cities "from the North".

The scatter plot of the number of inhabitants and the $< ATI >$ is also presented in Fig. 7, but on a log-log scale. This allows to emphasize the low values. The two different regimes are not well seen. It is like Fig. 6 with a change in x and y axes. A power law fit through the cloud leads to the main inertia axis equation given by $y \simeq 0.456\ 10^{-3}\ x^{0.915}$, with a regression coefficient $R^2 = 0.963$.

Let us now again take two focussing points: (i) the cities in the whole country and (ii) the regions.

Recall that the regions having a change over the quinquennium in the number of cities are indicated by an arrow $\uparrow$ or $\downarrow$; the arrow direction is according to the change in $N_{c,r}$ in some year as mentioned in Table 1. The fit in Fig. 1 captures the administrative changes. It is based on the function in Eq. (3.1); the fit parameters are given in the text. Administrative changes are usually due to local tensions grounded on historical motivations. Discussing such aspects is far beyond the scopes of this paper. However, it is important to note that the definition of the bounds of the IT regions derive often from the administrative structure of Kingdoms and States in the Italian territory after the Holy Roman Empire. In this respect, the influence of the historical facts occurred in Italy (Napoleon, the evolution of the Papacy, etc.) played also a relevant role.

Rank plots can be produced on classical, semi-log or log-log axes. In the first case, the data looks like a mere decaying convex function. However on semi-log (Fig. 2) and on log-log (Fig. 4), the ATI (and usually other data) shows some structure. An inflection point is well seen near $r_M/2 \sim 4000$, on the 2007-2011 yearly ATI of the 8902 IT cities ranked according to their "income tax" importance (Fig. 2). Some jumps between $r = 7$ and $r = 8$ are well marked on the log-log plot (Fig. 4). The rank plots are particularly meaningful in describing the economical structure of Italy under the point of view of the municipalities. The widest part of Italian cities has comparable small amount of ATI; this explains the inflection point at $r_M \sim 4000$ and why the yearly ATI decreases vertically with the rank for rank high enough. The jumps in the highest rank cities identifies the great differences among the cities with highest values of $< ATI >$. Such a difference is reduced for low ranked cities; this leads to some understanding of the polarization of the aggregated (citizen) income values in the main urban areas.

Such results are confirmed also by visually inspecting the semi-log plot of the rank-size relationship between each Italian city $< ATI >$ (averaged for

the examined quinquennium) and its rank (Fig. 3). Some departure of the data from the empirical fit can be noticed, mainly after the inflection point. Specifically, this happens for the cities ranked above $r \simeq 6000$, corresponding to an (averaged) ATI $\sim 10^7$. These 2000 or so cities contribute to $\sim 1.2 \, 10^{10}$; mean $\mu \sim 5.5 \, 10^6$; standard deviation $\sigma \sim 2.6 \, 10^6$. Thus $\mu/\sigma \sim 2.1$ for these cities. These cities (roughly) correspond to those having less than 1000 inhabitants (the border rank is at 6154.5), and $\mu \sim 543$, $\sigma \sim 256$; $\mu/\sigma \sim 2.1$ also. Such numbers are quite interesting, mainly if one notes that the total IT population is about $5.957 \, 10^7$; $\mu \sim 7361$, $\sigma \sim 40262$; $\mu/\sigma \sim 0.183$. Substantially, small cities have a number of inhabitants which is, relatively to the mean, less volatile when compared to that of IT. This further confirms the polarization of Italian inhabitants in a small number of highly populated cities.

The demographic, ATI relationship displayed through the scatter plot of the city ranks for $< ATI >$, in Fig. 5, indicates a rather huge variation. However the pair concordance is very high $\tau \simeq 0.849$ and $Z \simeq 114.6$.

Interesting findings are seen in Fig. 6, for the scatter plot of the $< ATI >$ (averaged for the examined quinquennium) and the number of inhabitants in all Italian cities. Two sets of cities are emphasized from linear fits. Such straight lines capture a relevant aspect of Italian reality, which is divided into different income distribution areas, the South being much poorer than the North. The red dot line includes cities showing a low proportion between rank for $< ATI >$ and rank for population. In the cities belonging to the blue dash line, such a proportion is high. Cities in the former case (Bari, Catania, Palermo, Napoli) are poorer than those of the latter one (Torino, Genova). Specifically, Torino is less populated and richer than Napoli (and, similarly, Genova is less populated and with a higher ATI than Palermo). The reasons for this can be found in the well-documented distortion of GDP due to illegal activities and organized crime, which is more pervasive in the South than in the North [67, 68, 69, 70].

The difference between the slopes of the red and blue lines in Fig. 5 may be useful in providing a measure of the entity of shadow economy. Milano represents an outlier for a simple reason: even if it is not the political capital of Italy (it is Rome), it is the financial one (the Italian Stock Market is in Milano). This explains the high value of ATI. Moreover, Milano has a highly populated hinterland, with many big cities (like Sesto San Giovanni or Rho). Hence, it is the center of a highly populated area, even though the municipality of Milano itself is not excessively populated *per se*, i.e. with respect to its ATI value.

4.1 Regional disparities

Note for completeness, that the number of provinces in 2007, i.e. 103, has increased by 7 units (BT, CI, FM, MB, OG, OT, VS)[5] to 110 provinces in 2011. In this time window, it is worth to point out that 228 municipalities have changed from a province to another one. Nevertheless, they remained in the same region, except for 7 cities from PU (the province of Pesaro and Urbino) in the Marche region, to RN (province of Rimini) in the Emilia Romagna region (Casteldelci, Maiolo, Novafeltria, Pennabilli, San Leo, Sant' Agata Feltria, Talamello). By looking at the data, after calculating either the number ($N_{i,r}$) of inhabitants in a region or the regional ATI (ATI_r), i.e. the sum for the relevant cities, in each

---

[5]e. g. see ISO code: *http://en.wikipedia.org/wiki/Provinces_of_Italy*.

year and the subsequent average, the change in regional membership appears to be very weakly relevant.

Thus, the regions can be also ranked each year according to their $N_{i,r}$ and displayed on a plot (Fig. 8), corresponding to $N_{c,r}$ in Fig. 1. Similarly, the regions can be ranked according to their yearly ATI_r. For conciseness, this is shown on this same Fig. 8. The fit parameters to Eq. (3.1), with A pre-imposed to be $= 10^9$, are respectively $m_1 = 0.445 \ 10^{-3}$; $m_2 = 0.287$; $m_3 = 1.006$ with $R^2 = 0.954$ for $N_{i,r}$, and $m_1 = 16.20$; $m_2 = 0.54$; $m_3 = 0.719$, with $R^2 = 0.966$ for $< ATI_r >$.

It should not be necessary to repeat that the rank of a region is not the same when ranking the $< ATI_r >$ (averaged for the examined quinquennium) and when considering the number of inhabitants $N_{i,r}$. The comparison of the region respective ranks is however quite illuminating: first, the Kendall τ calculation can be easily performed; results are given in Table 4, column (ii). The τ and ρ values are large ($\tau \sim 0.78$, $\rho \sim 0.9098$), but they are smaller than when not distinguishing regions.

Remarkably, the data and fits (to the function in Eq. (3.1)) for $N_{i,r}$ in IT regions and averaged regional ATI $vs.$ the rank of the region for the years of the quinquennium in Fig. 8 indicate a coherence with respect to Fig. 1, although the data transformation is not that trivial. This shows again that a rank-size rule is of great interest, showing structures not seen when absolute value-size relations are displayed or analyzed.

This is emphasized in the scatter plot of the region $< ATI_r >$ (averaged for the examined quinquennium) and the number of inhabitants ($N_{i,r}$) as well as the classical scatter plots in Italian cities of the different regions; Fig. 9 and Fig. 10. Remarkably, it is visually found that IT regions belong to different types of sets. These sets of regions can be emphasized also through linear fits: (i) a classical scatter plot points to *three* sets of regions, beside the outlier (Lombardia). Furthermore, the scatter plot of ranks indicate the existence of subregions. Those sets are characterized by a ratio between the ATI and the number of inhabitants, either greater or smaller than an "equilibrium point".

Fig. 9 provides a regional confirmation of the analysis carried out at the municipality level. The poor regions are the Southern ones, while the cities in the North are those belonging to the qualified group of high ATI. Valle d'Aosta and Sardegna are peculiar cases of wrong classes (Valle d'Aosta is a rich region belonging to the South group, for Sardegna the converse applies). These findings appear to us not so meaningful, being Valle d'Aosta and Sardegna positioned at the origin of the Cartesian plan in Fig. 9. As for the cities, Milano is an outlier; for regions, Lombardia plays a similar outlier role. These outcomes describe well the situation of highly productive regions in the North of Italy, with a South affected by the organized crime and poor government institutions distorting economical resources.

Also in this case, the gap between the slopes of the blue and the red lines may provide a good idea on how the ratio between population and aggregated tax income should be (blue line, the North), but how it is presently in the South (red line).

The rank-rank scatter plot of the region rank for $< ATI >$ (averaged for the examined quinquennium) with respect to the number of inhabitants in a region rank, Fig. 10, is very interesting, and fits well with results obtained by the reading of Fig. 9. Regions are confirmed to be clustered in two groups (sets

of regions are emphasized by linear best fits). This is in agreement with the brief discussion here above on the Italian socio-economic differences between the Northern regions and the Southern ones.

To conclude, it is worth noting the high variance, positive skewness and kurtosis of the distributions of ATI (see Table 2.). Observe also the μ/σ values and time evolution in this Table 2 : an increase in this variable indicates a sort of tendency to peaking of the sample distribution. Nevertheless, its small value indicates a quite large variety in intrinsic ATI values for the various cities. This effect is much obscured when looking at the regional level, since then $\mu/\sigma \simeq 1.12$.

5   Model

A specific modeling is presented based on arguments derived from statistics.

First of all, it may be a wonder why a form like Eq. (3.1) is used. Observe that it can be related to a power law with exponential cut-off [71], e.g., to the Yule-Simon distribution [72, 17]

$$y(r) = h\, r^{-\alpha}\, e^{-\lambda r}, \tag{5.1}$$

appropriately describing settlement formation (following the classical Yule model [72]) and its subsequent geographical distribution.

However, due to the finite size of the number N of data points, - there cannot reasonably be an infinite amount of cities in a region, the upper r regime should be considered as rather collapsing at the highest rank $r_M \equiv N$. This characterizes a function with an inflection point: for such a case, the Yule-Simon distribution can be adapted, bearing upon the fact that $h\, e^{-\lambda r} \equiv \hat{d}\, e^{\lambda(r_M - r)}$, and

$$e^{\lambda(r_M - r)} \simeq 1 + \lambda(r_M - r) \simeq [1 + (r_M - r)]^\lambda \tag{5.2}$$

for $r \to r_M$, thereby leading Eq. (5.1) to be written in the new form, that of Eq. (3.1),

$$y(r) = \kappa_3\, \frac{(N\, r)^{-\gamma}}{(N - r + 1)^{-\xi}} \tag{5.3}$$

The parameter κ_3 (or m_1 in Eq. (3.1)), is like the average amplitude of the data, see h in Eq. (5.1) also. Some meaning of the exponent γ of the hyperbola, (or m_2 in Eq. (3.1)), can be obtained from the decay exponent α in Eq. (5.1). Similarly, ξ (or m_3 in Eq. (3.1)), has the meaning of the decay exponent of an order parameter at a phase transition [73, 74, 75, 76]

Usually, the parameters (exponents), like $m_2 \leftrightarrow \gamma$ and $m_3 \leftrightarrow \xi$, designate the statistical physics model nowadays used for interpreting properties of a complex system, e.g. through phase transitions studies.

Having such ideas in mind, we suggest how to interpret the ($m_2 \leftrightarrow \gamma$ and $m_3 \leftrightarrow \xi$) parameters through mathematical statistics theories, i.e. the incomplete Beta function, as follows.

Recall that a preferential attachment process is an urn process in which additional balls (e.g, settlement locations) are added continuously to the system and are distributed among the urns (e.g., areas) as an increasing function of the number of balls the urns already have. In the most general form of the process,

balls are added to the system at an overall rate of m new species for each new urn. This leads to the so called Yule-Simon probability distribution

$$f(a;b) = b\,B(a;b+1). \tag{5.4}$$

where $B(x;y)$ is the Euler Beta function

$$B(x;y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{5.5}$$

$\Gamma(x)$ being the standard Gamma function [77, 78].

In practical words, a newly created urn (= region) starts out with k_0 balls (= cities) and further balls are added to urns at a rate proportional to the number k that they already have plus a constant $a \geq -k_0$. With these definitions, the fraction $P(k)$ of urns (areas) having k balls (cities) in the limit of long time is given by

$$P(k) = \frac{B(k+a;b)}{B(k_0+a;b-1)} \tag{5.6}$$

for $k \geq 0$ (and zero otherwise). In such a limit, the preferential attachment process generates a "long-tailed" distribution following a hyperbolic (Pareto) distribution, i.e. power law, ($\sim r^{-\alpha}$ or $\sim r^{-\gamma}$).

Moreover, a two-parameter generalization of the original Yule-Simon distribution replaces the Beta function with the incomplete Beta function:

$$B_\epsilon(a,b) = \int_0^\epsilon x^a \, (1-x)^b \, dx \tag{5.7}$$

In statistics, the expression $x^a(1-x)^b$ describes the probability of randomly selecting $a+b$ real numbers in [0,1] such that the first a are in $[0,x]$ and the last b are in $[x,1]$. The integral $\int_0^1 x^a \, (1-x)^b \, dx$ then describes the probability of randomly selecting $a+b+1$ real numbers such that the first number is x, the next a numbers are in $[0,x]$, and the next b numbers are in $[x,1]$.

It is worth noting that, in most of the studied examples concerning applications of the Beta-function in statistical physics, the number of urns increases continuously, although this is not a necessary condition for a "preferential attachment". In fact, it is unconceivable that an infinite number of urns regions) can be created. Moreover, an increase in the number of settlements (cities) is limited by "available resources", e.g. by the socio-economic need for optimizing the useful distances between settlements.

The product of two terms in Eq. (3.1) and the above reasoning remind of the Verhulst's modification [79] of the Keynes population expansion equation, when introducing a "capacity factor" .

6 Conclusions

This paper applies ideas of statistical mechanics in order to deal with an analysis of cities and regions. Specifically, the demographic (number of inhabitants, from Census 2011 - ISTAT) and the economic (ATI, averaged over the quinquennium 2007-2011, from MEF) ranking are compared and discussed for Italian cities.

Two statistical physics-like instruments have been mainly employed: (i) a (new) rank-size rule found as a doubly decreasing power law type (see Eq.(3.1) and (ii) the computation of the Kendall τ and Spearman ρ coefficients for finding fluctuation correlations, and phase state discrimination.

- It is found that both cities and regions, within the country, can be clustered in two categories, which mirror the Italian reality of a rich North and a poor South. Milano (city) and Lombardia (region) represent outliers, and cannot be indeed properly inserted in the resulting clusters. Some social, economical and political arguments might be carried out to explain these findings. A few sentences have been introduced to suggest reasoning outside the scope of this paper. It has seemed ppropriate to propose a statistical physics-like model, based on a number of evolving urn filling.

- Moreover, the above considerations and findings also serve as a demonstration of the advantage and interest of the Kendall τ and Spearman ρ coefficients to analyze and understand various (equal size) lists of variables measured according to various criteria. It has been pointed out that such a measure is similar to the fluctuation correlation coefficient in the linear response theory of statistical physics. Interestingly, it provided an indication of phase structures in the "sample" (=country).

- For completeness, the Pearson Π coefficient has been calculated. It has been argued that when the measurements are of so different natures, contain debatable error bars, rank-rank correlations are more meaningful, in contrast to the corresponding linear response theory coefficients in condensed matter physics.

It is worth saying, in concluding, that the analysis at a provincial level might be of interest, but this leads to complications in the data and subsequent analysis. The impact of the creation of new provinces ($103 \rightarrow 110$) in the considered time period might be interesting, - such an administrative act, similar to the application of an external field, providing an extra axis for investigations.

Appendix A. On data reorganization

During the examined time interval, several cities have merged into new ones, other were phagocytized. For completeness and thereby explaining some "finalized data reorganization", we give here below the various cases "of interest". We use official IT acronyms for the regions:

(i) Campolongo al Torre (UD) and Tapogliano (UD) have merged after a public consultation, held on Novembre 27th, 2007, into Campolongo Tapogliano (UD); thus 2 cities $\rightarrow$ 1 city only;

(ii) Ledro (TN) was the result of the merging (after a public consultation, held on Novembre 30th, 2008) of Bezzecca (TN), Concei (TN), Molina di Ledro

(TN), Pieve di Ledro (TN), Tiarno di Sopra (TN) and Tiarno di Sotto (TN) as far as it is explained e.g. in $http://www.tuttitalia.it/trentino-alto-adige/18-concei/$; thus $6 \rightarrow 1$;

(iii) Comano Terme (TN) results from the merging of Bleggio Inferiore (TN) and Lomaso (TN), in force of a regional law of November 13th, 2009; thus $2 \rightarrow 1$;

(iv) Consiglio di Rumo (CO) and Germasino (CO) were annexed by Gravedona (CO) on May 16th, 2011 and February 10th, 2011, to form the new municipality of Gravedona ed Uniti (CO); thus $3 \rightarrow 1$.

To sum up: 13 cities (in 2007) $\rightarrow$ 4 cities (in 2011). Thus, the number 8092 taken as our reference number of municipalities in the main text.

Appendix B. A short note on time dependence

In this Appendix, it is verified that we do not bias the analysis when we take the average of the ATI over the five year interval. In order to verify the point, we have examined the ATI year-year Kendall τ correlations with respect to each other as well as with respect to the average[6]. From [65] the τ, Eq. (3.2), and Z, Eq. (3.3), values are easily obtained; the results are given in Table 5 and Table 6. As mentioned in the main text, variations do exist but are rather mild.

Furthermore, a bonus is obtained in doing this time dependence examination. The relevant quantities given in Tables 5-6 readily indicate a rather stable system of city hierarchies within the examined time interval; e.g. $q/p \simeq 0.01$.

Scatter plots for every pair and for the scatter plots of pair of ranks are available from the authors upon request.

Appendix C. Pearson coefficient

The Pearson Π coefficient, Eq.(3.4), is classically used to estimate a correlation between (supposedly normally distributed) sets of measures [80]. The Π values for the case of the 8092 cities and the 20 regions, i.e. the value correlations between the number of inhabitants (according to the 2011 Census) and the corresponding averaged ATI (over the period 2007-2011) are given in Table 4. Values are in the same range as those of the Spearman ρ, but again much differ from those of the Kendall τ. It can be briefly argued that this arises from the fact that the measurements are of different natures and found in intervals with a quite often unknown error bar, - like many econo-sociological surveys: there are not even similar orders of magnitudes in measurements; there are outliers; the units are wholly different ones. Moreover, for calculating a Pearson Π coefficient, and deducing its meaning, the measurements should conform to some normality criterion; this is not quite the case here. Figures 11-12 and Fig. 13 show that there are outliers and the measurements are not normality distributed.

In fact, rankings are thought to be more illuminating and appealing, the more so when there is a sort of "competition". As a case surely already met

---

[6]A Spearman ρ has not been computed in such cases; it is not expected to provide further insights. Indeed, Spearman ρ is usually larger than Kendall τ; τ is already remarkably large as seen in Table 6.
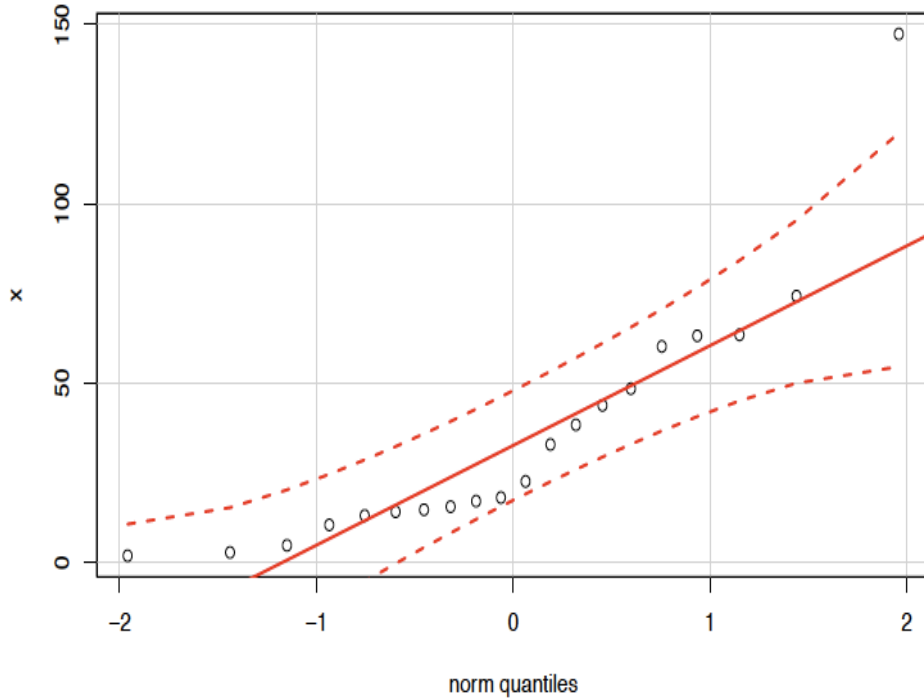
Figure 11: Distribution of normalized quantiles for x, i.e., the average ATI for cities throughout the 20 IT regions.

by all readers, consider a hiring process in academia or a grant funding scheme to research groups: the hiring is strictly based on some ranking (and committee member consensus, of course or through some vote procedure); the funding is usually not based on the quantitative or qualitative values of groups (they are usually measured through different indicators, like for the case in our text), relative to each other. All conclusions mainly depend on some ranking correlations through the chosen indicators. One (*in fine*) does not compare qualities. What is used is the rank. The measured quality has (alas) been by-passed. A Pearson correlation coefficient is rather irrelevant. The same is true in other "competitions", like in sport. The gap in points, the quality or value, at the ned of a "season", is masked by the ranking, when one wants to glorify a team or an agent. This goes also in many other cases. Recently, Raschke et al. [81] have also concluded that the rank correlation is a more robust measure, in the field of complex networks. This is exactly what we claim for the present case, but which is not *per se* a network.

References

[1] G.K. Zipf, *Human Behavior and the Principle of Least Effort : An Introduction to Human Ecology* (Addison Wesley, Cambridge, Mass., 1949).
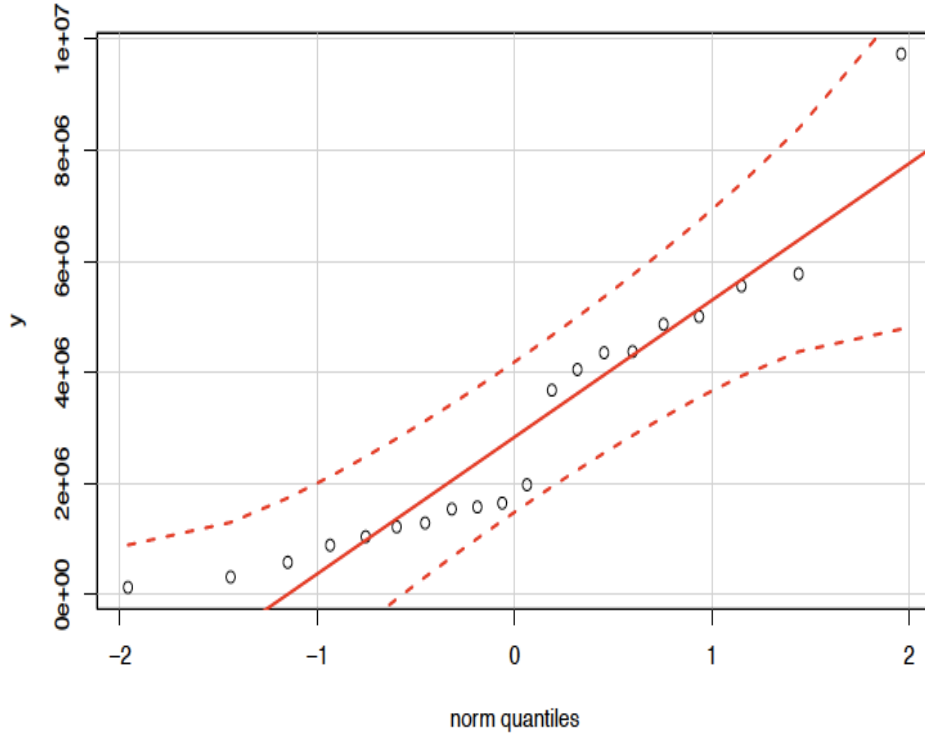
Figure 12: Distribution of normalized quantiles for y, i.e. the number of inhabitants in cities throughout the 20 IT regions.

$q \setminus p$	2007	2008	2009	2010	2011		$< 5yav >$
2007	-	32322584	32191636	32162292	32128014		32294107
2008	413602	-	32476840	32430158	32381578		32561455
2009	544550	259346	-	32544918	32466870		32581689
2010	573894	306028	191268	-	32530208		32572191
2011	608172	354608	269316	205978	-		32521347
$< 5yav >$	442079	174731	154497	163995	214839		-

Table 5: The number of concordant pairs p (above the diagonal) and that of non-concordant pairs q (below the diagonal) of the 8092 cities, according to their ATI value, for every year pairs.
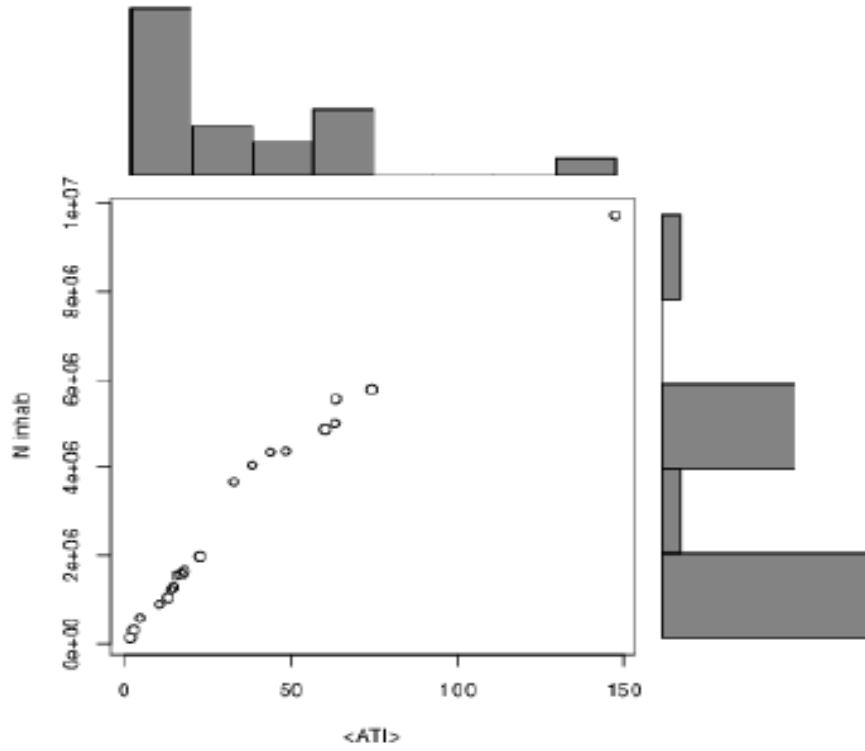
Figure 13: Relationship and distribution of the number of inhabitants and the average ATI in corresponding IT regions.

$Z \setminus \tau$	2007	2008	2009	2010	2011		$< 5yav >$
2007	-	0.9747	0.9667	0.9649	0.9628		0.9730
2008	131.49	-	0.9842	0.9813	0.9783		0.9893
2009	130.42	132.77	-	0.9883	0.9835		0.9906
2010	130.18	132.38	133.32	-	0.9874		0.9900
2011	129.89	131.98	132.68	133.21	-		0.9869
$< 5yav >$	131.26	133.47	133.63	133.55	133.13		-

Table 6: Rank correlation test of the 8092 cities, according to their ATI value: τ (above the diagonal), Eq. (3.2), and Z (below the diagonal), Eq. (3.3), for every year pairs.

[2] Newman M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, **46**, 323 - 351.

[3] M. Jefferson, Geogr. Rev. **29**, 226 (1939).

[4] M. Beckmann, Economic Development and Cultural Change **6**, 243 (1958).

[5] X. Gabaix, Q. Econ. **114**, 739 (1999); Am. Econ. Rev. **89**, 129 (1999).

[6] L.M.A. Bettencourt, J. Lobo, G.B. West, Eur. Phys. J. B **63**, 285 (2008).

[7] F. Semboloni, Eur. Phys. J. B **63**, 295 (2008).

[8] P. Krugman, *Geography and Trade* (The MIT Press, Cambridge MA, 1991).

[9] B.J.L. Berry, Economic Development and Cultural Change **9**, 573 (1961).

[10] S. Pianegonda, J.R. Iglesias, Physica A: Statistical Mechanics and its Applications **342** 193 (2004).

[11] G. Ottaviano, D. Puga, The World Economy **21**, 707 (1998).

[12] M. Fujita, P. Krugman, and A.J. Venables, *The Spatial Economics: Cities, Regions, and International Trade* (The MIT Press, Cambridge, MA, 1999).

[13] J.P. Neary, Journal of Economic Literature **39**, 536 (2001).

[14] R. Baldwin, R. Forslid, P. Martin, G. Ottaviano, F. Nicoud, *Economic geography and public policy.* (Princeton University Press, Princeton, NJ, 2003).

[15] M. Fujita, T. Mori. Papers in Regional Science **84**, 377 (2005).

[16] B.K. Chakrabarti, A. Chakraborti, S.R. Chakravarty, A. Chatterjee, *Econophysics of income and wealth distributions* (Cambridge University Press, 2013).

[17] N.K. Vitanov, Z.I. Dimitrova, *Bulgarian Cities and the New Economic Geography* (Vanio Nedkov, Sofia, 2014).

[18] K.T. Rosen, M. Resnick, Journal of Urban Economics **8**, 165 (1980).

[19] E.S. Mills, B.W. Hamilton, *Urban Economics* (Prentice Hall, 1994).

[20] F. Guérin-Pace, Urban Studies **32**, 551 (1995).

[21] L.H. Dobkins, Y.M. Ioannides, Regional Science and Urban Economics **31**701 (2001).

[22] S. Song, K.H. Zhang, Urban Studies **39**2317 (2002).

[23] Y.M. Ioannides, H.G. Overman, Regional Science and Urban Economics **33**, 127 (2003).

[24] X. Gabaix, Y.M. Ioannides, *The Evolution of City Size Distributions, in: Henderson, J.V., Thisse, J.-F. (Eds.), Handbook of Regional and Urban Economics*, Vol. 4, (Elsevier, Amsterdam, 2004).

[25] W.J. Reed, Journal of Regional Science **42**, 1 (2002),

[26] Z. Dimitrova, M. Ausloos, Primacy analysis of the system of Bulgarian cities, $arXiv : 1309.0079$ (2013).

[27] E. Sheppard, Nonlinearities, Geographical Analysis **17**, 47 (1985).

[28] R. Gibrat, *Les inégalités économiques: Applications aux inégalités des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel.* (Sirey, Paris, 1931).

[29] V. Nitsch, Journal of Urban Economics **57** 86 (2005).

[30] G. Peng, Physica A: Statistical Mechanics and its Applications **389**, 3804 (2010).

[31] Y.M. Ioannides, S.Skouras, Journal of Urban Economics **73**, 18 (2013).

[32] V. J. Matlaba, M. J. Holmes, P.McCann, J. Poot, Review of Urban and Regional Development Studies **25**, 129 (2013).

[33] K.T. Soo, Urban Studies **44**, 1 (2007).

[34] J.-C. Cordoba, Urban Economics **63**, 177 (2008).

[35] A.S. Garmestani, C.R. Allen, C.M. Gallagher, J.D. Mittelstaedt, Urban Studies **44**, 1997 (2007).

[36] M. Bosker, S. Brakman, H. Garretsen, M. Schramm, Regional Science and Urban Economics **38**, 330 (2008).

[37] A.S. Garmestani, C.R. Allen, C.M. Gallagher, Journal of Economic Behavior and Organization **68**, 209 (2008).

[38] R.K. Skipper, McNair Scholars Undergraduate Research Journal **3**, 217 (2011).

[39] M. Cristelli, M. Batty, L. Pietronero, Scientific Reports **2**, 812 (2012).

[40] K. Giesen, J. Südekum, Journal of Economic Geography **11**, 667 (2011).

[41] M. Kendall, Biometrika **30**,81 (1938).

[42] H. Abdi, *The Kendall Rank Correlation Coefficient. In Neil Salkind (Ed.): Encyclopedia of Measurement and Statistics.* (Sage, Thousand Oaks (CA), 2007).

[43] C. Spearman, American Journal of Psychology **15**, 201 (1904).

[44] S.-D. Bolboacaa, L. Jantschi, Leonardo Journal of Sciences **9**, 179 (2006).

[45] M. Bland, *An Introduction to Medical Statistics,* Oxford University Press; Oxford, New York, Tokyo, pp. 205-225, (1995).

[46] M. Melucci, SIGIR Forum **41**, 18 (2007).

[47] M. Ausloos, Central European Journal of Physics **12**, 773 (2014).

[48] M. Ausloos, R. Cloots, A. Gadomski, N.K. Vitanov, International Journal of Modern Physics C **25**, 1450060 (2014).

[49] M. Ausloos, A. Gadomski, N.K. Vitanov, Phys. Scr. **89**, 108002 (2014).

[50] T.S. Hare, The Archaeology of Communities: A New World Perspective , 78 (2000).

[51] Y.-B. Zhou, , L. Lu, M. Li, M., New Journal of Physics **14**, 033033 (2012).

[52] M. Ausloos, Scientometrics **95**, 895 (2013).

[53] J. Stallings, E Vance, J Yang, MW Vannier, J. Liang, L. Pang, L. dai, I. Ye, G. Wang, Proceedings of the National Academy of Science of the United States of America **110**, 9680 (2013).

[54] J.-J. Hsieh Computational Statistics & Data Analysis **54**(6), 1613-1621 (2010).

[55] B. Yang, Y. Peng, H. Chi-Ming Leung, S.-M. Yiu, J.-C. Chen, F. Yuk-Lun Chin BMC Bioinformatics 11 (suppl. 2), S5, (2010).

[56] M. Rezapour, N. Balakrishnan Statistical Methodology 15, 55-72 (2013).

[57] R.J. Schmitz, M.D. Schultz, M.A. Urich, J.R. Nery, M. Pelizzola, O. Libiger, A. Alix, R.B. McCosh, H. Chen, N.J. Schork, J.R. Ecker Nature 495, 193198 (2013).

[58] J. Rappoport, Regional Science and Urban Economics **38**, 533 (2008).

[59] T. Mori, T.E. Smith, Journal of Regional Science **51**, 694 (2011).

[60] H.L.F. De Groot, J. Poot, M.J. Smit, *Agglomeration, innovation and regional development: theoretical perspectives and meta-analysis. In: R. Capello and P. Nijkamp (eds), Handbook of regional growth and development theories* (Edward Elgar, Cheltenham, 2009.) 256-281.

[61] P.C. Melo, D.J. Graham, R.B. Noland, Regional Science and Urban Economics **39**, 332 (2009).

[62] M. Ausloos, J. Appl. Quant. Meth. **9**, 1 (2014).

[63] R Kubo, Journal of the Physical Society of Japan **12**, 570 (1957).

[64] R Kubo, Reports on Progress in Physics **29**, 255 (1966).

[65] P. Wessa, Kendall τ Rank Correlation (v1.0.11) in Free Statistics Software (v1.1.23-r7), Office for Research Development and Education, *http : //www.wessa.net/rwasp_kendall.wasp/*(2012).

[66] D.G. Bonett, T.A. Wright , Psychometrika **65**, 23 (2000).

[67] G Brosio, A Cassone, R Ricciuti, Public Choice **112**, 259 (2002).

[68] C.V. Fiorio, F. D'Amuri, Giornale degli Economisti e Annali di Economia **64**, 247 (2005).

[69] F. Calderoni, Global Crime **12**, 41 (2001).

[70] R. Galbiati, G. Zanella, Journal of Public Economics, **96**, 485 (2012).

[71] C. Rose, D. Murray, D. Smith, *Mathematical Statistics with Mathematica* (Springer, New York, 2002), p. 107.

[72] G.U. Yule, Phil. Trans. Roy Soc. London B **213**, 21 (1922).

[73] C.J. Thompson, *Mathematical statistical mechanics* (Macmillan, London, 1971).

[74] H.E. Stanley, *Phase transitions and critical phenomena* (Clarendon Press, London. 1971).

[75] M.E. Fisher, Rev. Mod. Phys. **70**, 653 (1998).

[76] H.E. Stanley, Rev. Mod. Phys. **71**, 358 (1999).

[77] I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series and Products* (Academic Press, New York, 2000).

[78] M. Abramowitz, I. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970)

[79] P.F. Verhulst, *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles* **18**, 1 (1845).

[80] W. L. Hays, *Statistics*, (CBS College Publishing, New York, 1988).

[81] M. Raschke, M. Schläpfer, R. Nibali Phys. Rev. E **82**, 037102 (2010),