SCIENTIFIC REPORTS

OPEN

The Italian primary school-size distribution and the city-size: a complex nexus

Alessandro Belmonte[1], Riccardo Di Clemente[1,2] & Sergey V. Buldyrev[3]

[1]IMT Institute for Advanced Studies Lucca, Piazza S. Ponziano 6, 55100, Lucca, Italy, [2]Istituto dei Sistemi Complessi - CNR, Via dei Taurini 19, 00185 Rome, Italy, [3]Department of Physics, Yeshiva University, New York, NY 10033, USA.

SUBJECT AREAS:
STATISTICAL PHYSICS
APPLIED PHYSICS
STATISTICS
APPLIED MATHEMATICS

Received
30 January 2014

Accepted
21 May 2014

Published
23 June 2014

Correspondence and requests for materials should be addressed to S.V.B. (buldyrev@yu.edu)

We characterize the statistical law according to which Italian primary school-size distributes. We find that the school-size can be approximated by a log-normal distribution, with a fat lower tail that collects a large number of very small schools. The upper tail of the school-size distribution decreases exponentially and the growth rates are distributed with a Laplace PDF. These distributions are similar to those observed for firms and are consistent with a Bose-Einstein preferential attachment process. The body of the distribution features a bimodal shape suggesting some source of heterogeneity in the school organization that we uncover by an in-depth analysis of the relation between schools-size and city-size. We propose a novel cluster methodology and a new spatial interaction approach among schools which outline the variety of policies implemented in Italy. Different regional policies are also discussed shedding lights on the relation between policy and geographical features.

There is a growing literature that nowadays sheds light on complexity features of social systems. Notable examples are firms and cities[1–4], but many others have been proposed[5,6]. These systems are perpetually out of balance, where anything can happen within well-defined statistical laws[7,8]. Italian schools system seems to not escape from the same characterization and destiny. Despite several attempts of the Italian Ministry of education to reduce the class-size to comply with requirements stated by law[9–11], no improvements have been made and still heterogeneity naturally keeps featuring the size distribution of the Italian primary schools.

In this paper we characterize the statistical law according to which the size of the Italian primary schools distributes. Using a database provided by the Italian Ministry of education in 2010 we show that the Italian primary school-size approximately distributes (in terms of students) as a log-normal distribution, with a fat lower tail that collects a large number of very small schools. Similarly to the firm-size[12,13], we also find the upper tail to decrease exponentially. Moreover, the distribution of the school growth rates are distributed with a Laplassian probability density function (PDF). These distributions are consistent with the Bose-Einstein preferential attachment process. These results are found both at a provincial level and aggregate up to a national level, i.e. they are universal and do not depend on the geographic area.

The body of the distribution features a bimodal shape suggesting some source of heterogeneity in the school organization. The evidence of the bimodality underlies the interplay between different processes that define thresholds and boundaries that are very peculiar for the Italian primary school-size distribution. The question that we attempt to address in this paper is whether such regularity might depend on the complex geographic features of the country that in turn determines the way population (and in particular young people) distributes. We address these questions by analyzing in depth the spatial distribution of the schools, with particular regard to the areas where commuting is more effortful. We then proceed by investigating the complex link between schools and comuni, the smallest administrative centers in Italy, addressed by the introduction of a new binning methodology and a new spatial interaction analysis. Our conclusions indicate that the bimodality of the Italian primary school-size distribution is very likely to be due to a mixture of two laws governing small schools in the countryside and bigger ones in the cities, respectively.

Several examples of different regional schooling organizations are analyzed and discussed. We use GPS code positions for schools in two very different Italian Regions: Abruzzo and Tuscany. We introduce a measure of the average spatial interaction intensity between a school and the surrounding ones. We show that in regions like Abruzzo, that are mainly countryside, a policy favoring small schools uniformly distributed across small comuni has been implemented. Abruzzo small schools are generally located in low density populated zones, belonging to small comuni. They are also very likely to have another small school as closest and the median distance between them is 8 *km* that is also the distance between small comuni. In Tuscany, a flatter region with a very densely populated zone along the metropolitan area composed by Florence, Pisa and Livorno, we conversely find 1) a higher school density; 2) a stronger interaction between small and big schools; 3) a greater average proximity among schools. We address these stylized facts by arguing that the Italian primary school organization is basically the result of a random process in the school choice made by the parents. Primary education is not felt so much determinant to drive housing choice, like in US, because of the absence of any territorial constraint in school choice. Even if there is a certain mobility *within* a comune toward the most appealing schools, primary students generally do not move *across* comuni to attend a school. As a result, school density and school-size are prevalently driven by the population density and then by the geographical features of the territory. This generates a mixture in the schooling organization that turns into a bimodal shape distribution.

Results

Empirical evidence. We analyze a database on the primary school-size distribution in Italy that provides information on public and private schools, locations, and the number of classes and students enrolled. Data are collected, at the beginning of every academic year, by the Italian Ministry of education to be used for official notices. Our dataset covers $N = 17187$ primary schools in 2010 of which 91.31% were public. Almost seven thousands are located in mountain territories, (which represent the 40%) and 4101 are spread among administrative centers (provincial head-towns).

In Italy primary education is compulsory for children aged from six to ten. However, the parents are allowed to choose any school which they prefer, not necessarily the school closest to their home[14]. We define x_i the size of the school $i \in [1, \ldots, N]$ as the number of students enrolled in each school. Fig. 1(a) shows the histogram of the logarithm of the size of all primary schools in Italy. The red solid curve is the log-normal fit to the data

$$P(\ln x) = \exp\left(-\frac{(\ln x - \hat{\mu})^2}{2\hat{\sigma}^2}\right)\frac{1}{\sqrt{2\pi}\hat{\sigma}} \qquad (1)$$

using the estimated parameters $\hat{\mu} = 4.77(\hat{\mu}/\ln(10) = 2.07)$, the mean of the ln x of the number of students per school, and its standard deviation, $\hat{\sigma} = 0.85(\hat{\sigma}/\ln(10) = 0.37)$. On a non-logarithmic scale, $\exp(\hat{\mu}) = 118$ and $\exp(\hat{\sigma}) = 2.34$ are called the location parameter and the scale parameter, respectively[15]. The histogram in Fig. 1(a) suggests that log-normal fits data quite well. However, even a quick glance reveals that there are too many schools with a small dimension and much less mass in the upper tail with respect to the fit, suggesting that the number of students of the largest schools is smaller than would be the case for a true log-normal. In other words, similarly with firms-size distribution[16], tails seem to distribute differently from the log-normal distribution. Also Fig. 1(a) reveals a bimodal shape of the school-size distribution that we will extensively investigate below.

These findings can be detected in a more powerful way by plotting the histogram in a double logarithmic scale, comparing the tails of the log-normal distribution with those of the empirical one. We do this in Fig. 1(b) where y-axes represents the logarithm of the number of schools in the bins whereas in the x-axes the logarithm of the number of students stands. The empirical distribution differs significantly from the theoretical distribution which is a perfect parabola (the red curve), both in the tails and in the central bimodal part. A functional form of the right tail of the empirical distribution is revealed in the inset of Fig. 1(b) where we plot the cumulative distribution $P(X > x)$ of school sizes in semi-logarithmic scale. The straight line fit suggests that the right tail decreases exponentially $P(X > x) = \exp(-x\alpha)$ with a characteristics size $\alpha = \frac{1}{120}$. This in turn means that there are approximately 120 students per school and also that the distribution of large schools declines exponentially. The exponential decay of the right tail of size distribution is consistent with Bose-Einstein preferential attachment process and is observed in the distribution of sizes of universities and firms.

Next we investigate the growth rates of elementary schools. Since temporal data are not currently available, we look at the single academic year, the 2010, and define the growth rate g_i as follows:

$$g_i \equiv \frac{x_i^1 - x_i^5}{\sum_{j=1}^{5} x_i^j} = \lambda_i - \mu_i, \qquad (2)$$

where x_i^j stands for the number of students attending the *j-th* grade in school i, with $j \in [1, 5]$; $\lambda_i \equiv x_i^1 \Big/ \sum_{j=1}^{5} x_i^j$ is the fraction of students that have been enrolled in the first grade at six years old in school i, whereas $\mu_i \equiv x_i^5 \Big/ \sum_{j=1}^{5} x_i^j$ is the fraction of students that exit the school after the 5-*th* grade. Fig. 2(a) shows the relation between growth rate g_i and school-size x_i. The numbers of grades j provided by each school i, named J_i, is defined by the color gradient bar on the right side of the Fig. 2(a). Blue circles identify schools with $J_i = 1$. Such a group collects schools just established only providing the 1-*st* grade, i.e. with $\lambda_i = 1$ *and* $\mu_i = 0$, or that are going to close providing only the 5-*th* grade, i.e. with $\mu_i = 1$ *and* $\lambda_i = 0$. As soon as more grades are provided (colors switching to the warm side of the bar) schools tend to cluster around a null growth rate.

In Fig. 2(b) we investigate the growth/size relationship in depth. We demonstrate the applicability of the Gibrat law that states that the average growth rate is independent on the size[17,18]. We define the average of the school size in each bin c as $\langle x_i \rangle_c$. The number of schools in each bin n_c is represented by the size of the circle and the average number of grades $\langle J_i \rangle_c$ is depicted according to the color gradient on the right side (the same of Fig. 2(a)). Independently from the size and the number of grades provided, schools do not grow on average. Nevertheless, we find more variability in smaller schools, apart from schools with $x_i < 10$, namely hospital-based schools mostly similar to one another, and the standard deviation of the growth rate $\sigma_{g(\langle x_i \rangle_c)}$ is found to be decreasing as $\langle x_i \rangle_c^{-\beta}$ with school-size by a rate of $\beta \approx .60$ (sub Fig. 2(b) inset). This is consistent with what has been found for other complex systems like firms or cities[13,19–23].

In Fig. 2(c) we study the growth rate distribution, where the probability density function $P(g = g_i)$ of growth rate has been plotted. The blue line represents the full sample (all the schools) distribution. Black and red colors identify the full capacity schools ($J_i = 5$) and the schools with $J_i < 5$, respectively. Regardless of the number of grades provided, the growth distribution underlines a Laplace PDF in the central part of the sample[24]. The not-fully covered schools show a three peak behavior, where the left peak represents schools which are going to close, the central peak gathers schools that provide several grades but still in equilibrium phase, and the right peak is made up by the growing schools. Fig. 2(d) reports empirical tests for the tails of the PDF of the growth rate of the full sample (the upper one in blue, and the lower one in black). The asymptotic behavior of g can be well approximated by power laws with exponents $\zeta \approx 4$ (the magenta dashed line), bringing support to the hypothesis of a stable dynamics of the process[20]. All these findings are consistent with the Bose-Einstein process according to which the size distribution has an exponential right tail, a tent-shaped distributed growth rate g_i, with a Laplace cap and power law tails, the average growth rate is independent of the size, and the size-variance relationship is governed by the power law behavior with exponent $\beta \approx 0.5$.

City size and school size. Fig. 1(a) features the coexistence of two peaks, the first peak corresponding to $\log_{10} x_i \equiv m_1 = 1.7$ and the second one to $\log_{10} x_i \equiv m_2 = 2.3$, divided by a splitting point in

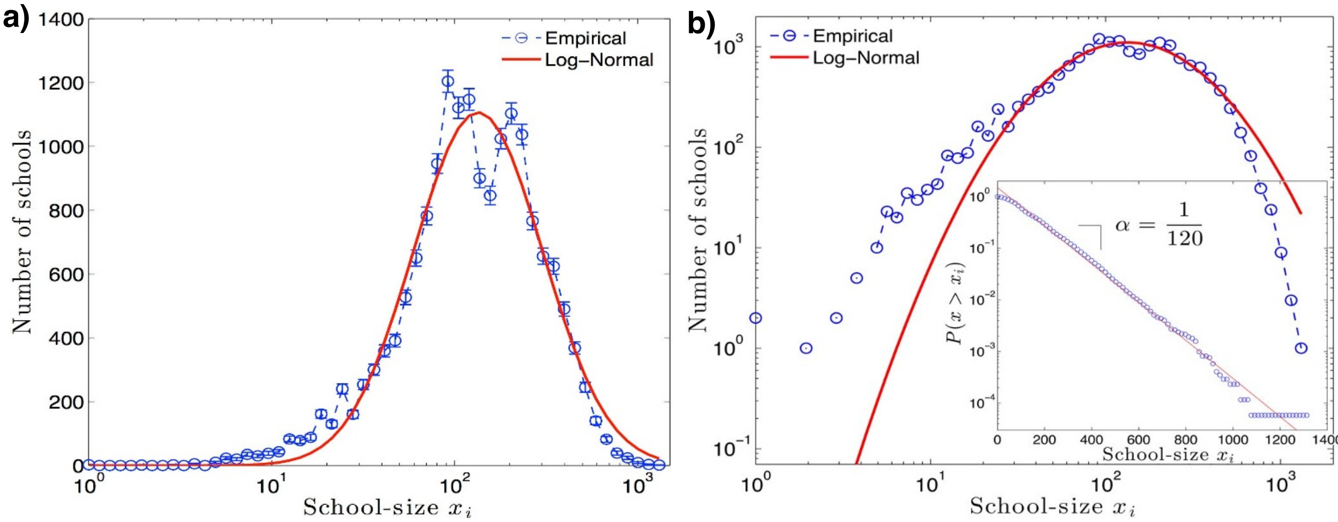

Figure 1 | School-size distribution. (a). Italian primary school-size distribution according to the number x_i of students per school $i \in [1, \ldots, N]$ for the year 2010. The empirical distribution is drawn in blue (each circle is a bin); the red line stands for the Lognormal fit with mean $\hat{\mu} = 4.77$ ($\hat{\mu}/\ln(10) = 2.07$) and standard deviation $\hat{\sigma} = 0.85$ ($\hat{\sigma}/\ln(10) = 0.37$). On a non-logarithmic scale, $\exp(\hat{\mu}) = 118$ and $\exp(\hat{\sigma}) = 2.34$. $N = 17187$. Statistical errors (SE) are drawn in correspondence to each bin, according to $\sqrt{N_{bin}}$. SE are bigger in the body of the distribution and tinier in the tails. Nevertheless, central bins space from the two peaks, $m_1 = 1.7$ and $m_2 = 2.3$, at least 6 times the SE, equals on average to $\sqrt{10^3} = 32$. In this case the probability to have a non bimodal shape under our distribution is 4×10^{-15}. (b). Italian primary school-size distribution in log-log scale. As expected, the theoretical distribution has drawn as a perfect parabola (the red curve), $y = ax^2 + bx + c$, such that $\hat{\mu} = -b/2a$ and $\hat{\sigma} = -1/2a$. Conversely, the empirical distribution does not plot as a parabola, at least for what regards to the tails which deviate from the log-normal. The inset figure shows a functional form of the right tail of the empirical distribution. We plot the cumulative distribution, $P(X > x_i) = \exp(-\alpha x_i)$, of school sizes in semi-logarithmic scale with characteristics size $\alpha = 0.0084$. This in turn means that there are approximately 120 students per school.

correspondence of $\log_{10} x_i \equiv \bar{m} \approx 2.1$. The school sizes corresponding to these features are $\mu_1 = 10^{m_1} = 50$, $\mu_2 = 10^{m_2} = 200$, and $\bar{\mu} = 10^{\bar{m}} = 128$, with $\bar{\mu}$ approximately equal to the average school size. 39% of the Italian primary schools distribute on the right of $\bar{\mu}$, and more than 60% distribute on the left side. We test the alternative hypothesis of unimodality by looking at the probability that the numbers of schools in the two central bins n_1, n_2 are not smaller and the numbers of schools in the next three bins n_3, n_4, n_5 are not larger than a certain number n^* provided that the standard deviation of the number of schools in these bins due to small statistics is $\sqrt{n^*}$. This probability is equal to $p(n^*) = \Pi_i \operatorname{erfc}\left(|n_i - n^*| \big/ \sqrt{2n^*}\right) \big/ 2$ and it reaches maximum $p_{\max} \approx 4 \times 10^{-15}$ at $n^* = 980$. Accordingly, we establish the bimodality with a very high confidence. This is also consistent with the bimodality index that we find to be equal to $\delta = (\mu_1 - \mu_2)/\sigma = .45$ [25].

In this section we investigate the source of this heterogeneity that we find to be related to geographical and political features of the country and remarkably to the size of the comuni, the smallest administrative centers in Italy (information on comuni are provided by the Italian statistical institute, ISTAT), also here referred interchangeably as cities regardless of the size, p_k. A particular treatment is devoted to the nexus between the school-type (private versus public) and the geographical features of the comuni in the supplementary information, where we show that private schools are much less variable in size than public schools and have a narrow unimodal distribution peaked at approximately 100 students which contributes to the left peak of the entire school size distribution (Figure S1).

We denote a comune with letter $k = [1, \ldots, K]$. In 2010, $K = 8,092$ comuni have been counted in Italy, the 40% of which located in the mountains. We define $\mathcal{M}$ the set of mountain comuni and, accordingly, we call school i a mountain school iff it resides in a comune $k \in \mathcal{M}$ (in the Supplementary we explain the laws according to which Italian comuni are classified as mountains). Each city k has $n_k \geq 0$ schools (more than 15% of the cities have no schools) and population p_k, which distributes approximately as a log-normal PDF (see Fig. 3(a)), except for the right tail that is distributed according to a Zipf law, i.e. $p_k \sim r(p_k)^{-\xi}$ with slope $\xi \approx 1$[2,3,26–28]. In Fig. 3(b) we find $\xi \approx .80$, in Italy, that is exactly the slope of the power law $p_k \sim r(n_k)^{-\zeta}$ which links the population p_k with the rank of this city in terms of number of schools n_k (blue circles in Fig. 3b), i.e. $\zeta = \xi \approx .80$. This means that the first city, Rome, has almost the double number of schools than Milan, and triple of Naples, while Rome has almost the double of inhabitants of Milan, and the triple of Naples. This amounts to say that n_k is a good proxy for the city-size.

We use the number of schools to assign comuni to different clusters $h \in [1, \ldots, H]$, according to

$$h = \left\{ \forall\ k \in [1, \ldots, K] : 2^{h-1} \leq n_k < 2^h \right\}. \quad (3)$$

Accordingly, the first bin $h = 1$ gathers all the comuni with only one school; the second one collects all the comuni with $n_k = [2, 3]$, and so on. Though we find the average population $\langle p \rangle_h$ to increase across different city-clusters h, less comuni K_h lie in more populated clusters (the magenta and black lines in Fig. 3(c)). Interestingly, we find the interaction term $K_h \langle p \rangle_h$, the green line in Fig. 3(c), to distribute uniformly across different comuni-clusters, meaning that in small comuni with $n_k = 1$ live the same population than in bigger ones with much more schools.

Nevertheless, population is differently composed across city-clusters and a smaller fraction of young people is found in smaller comuni. To see that we also introduce a clusterization of comuni according to population. Each comune is assigned to a cluster $c \in [1, \ldots, C]$ composed by all the comuni k with population p_k ranging from ψ^{c-1} to ψ^c, i.e.

$$c = \left\{ \forall\ k \in [1, \ldots, K] : \psi^{c-1} < p_k \leq \psi^c \right\}. \quad (4)$$

Setting the parameter $\psi = 2$ yields $C = 23$ clusters. Although the first seven sets are empty because no comuni in Italy has less than 128 inhabitants, the first (non-empty) cluster, $c = 8$, collects very small comuni with $p_k \in (128, 256]$. The last one, $c = 23$, conversely, is composed by the biggest cities with $p_k \in (2^{22}, 2^{23}]$. In Fig. 3(d) we plot

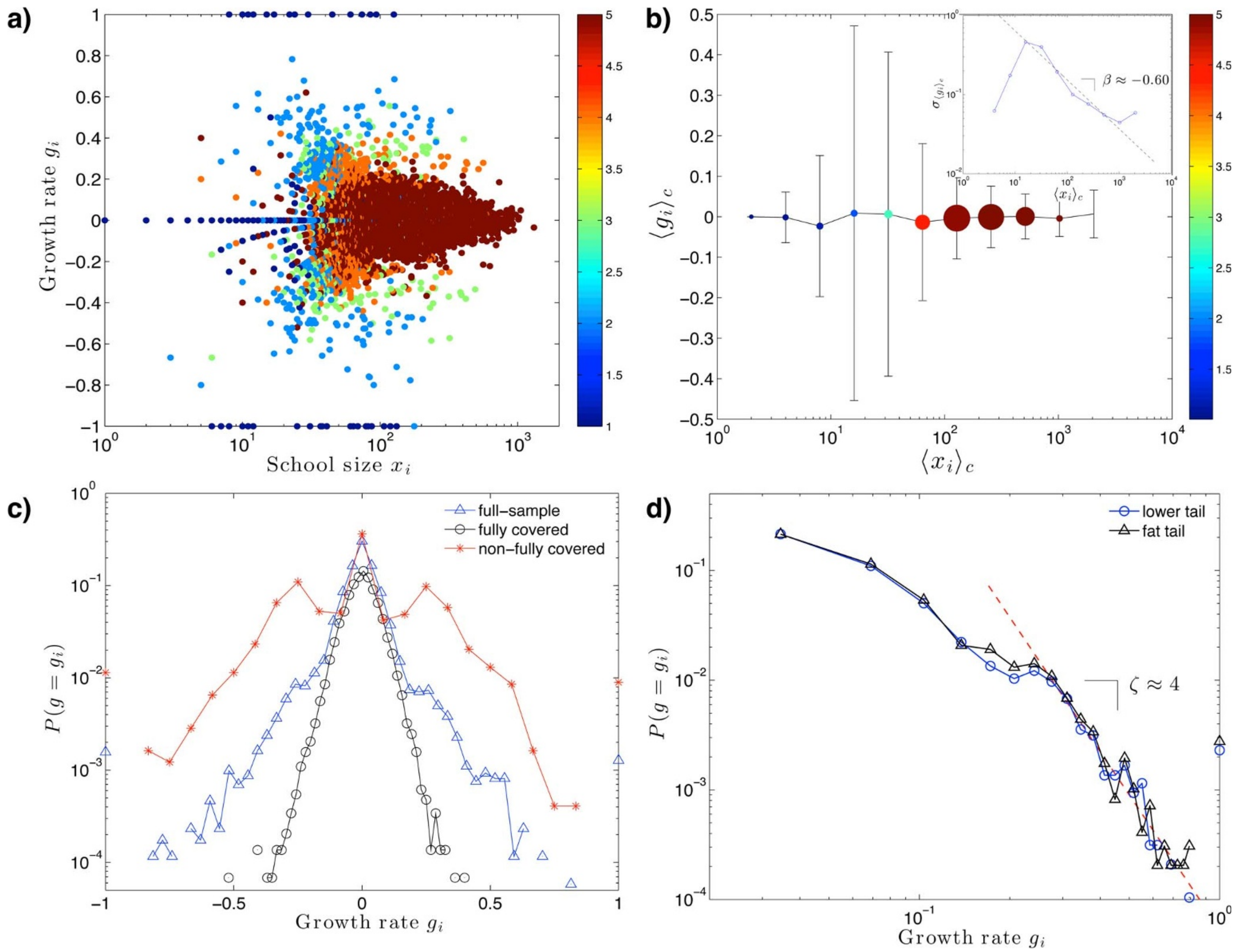

Figure 2 | The growth rate distribution of the Italian primary schools in 2010. The growth rate g_i is defined according to Eq. (2). (a). The growth rate and school-size relationship. Colors, according to the vertical bar on the right-hand side of the graph, are the number of grades J_i provided by the school i. Smaller schools (in blue) with $J_i = 1$ are both the newest one (just created, with $\lambda = 1$) and schools that are going to close (with $\mu = 1$). They can also be schools that do not grow yet providing just one grade (i.e. $j = 3$). (b). The mean growth rate clusters around zero across different subsets c that are differently populated by n_c schools according to the size of the circles. The color of the circles stand for the average number of grades J_i (the same gradient color bar of Fig. 2(a) is used here). The variability within each cluster c is shown in the inset figure. Apart from schools with $x_i < 10$, namely hospital-based schools mostly similar to one another, the standard deviation is found to be decreasing with school-size by a rate of $\beta \approx .60$. (c). The probability density function $P(g = g_i)$ of growth rate has been plotted underlying a Laplace PDF in the body around $P(g) = 1$ and $P(g) \approx 10^{-1.5}$. Blue triangles (△) stand for the full sample distribution, black circles (○) indicate mature schools with $J_i = 5$, and red stars (*) schools with $J_i = 1$. (d). The plot reports empirical tests for the tails parts of the PDF of growth rate, the upper one in blue (○), and the lower one in black (□). The asymptotic behavior of g can be well approximated by power laws with exponents $\zeta \approx 4$ (the magenta dashed line).

the average number of schools $\langle n \rangle_c$ (magenta line) and the average school-size $\langle x \rangle_c$ (the blue line) against the comuni size p_c for each non-empty cluster c. We find that the average number of schools increases as a power law with exponent $\beta = 0.88$. This is consistent with the literature[2,3,26–28] that has stressed the emergence of scale-invariant laws that characterize the city-size distribution. The average school-size increases with the population of the city reaching an asymptotic value at $\langle x \rangle_c \simeq 230$ students per school in the large cities. As expected, the interaction term, representing the average number of school-aged population in comuni belonging to cluster c, $\tilde{s}_c = \langle x \rangle_c * \langle n \rangle_c$, behaves linearly with the comuni size except for small comuni with $p_c < 10^3$, for which the school-aged population constitutes a smaller fraction of the total population than in large cities.

In Fig. 4 we investigate the school-size distribution according to the comuni features. To this end, Fig. 4(a) draws the distributions of $\log_{10} x_i$ conditionally on the number of schools, n_k, in the comune k. It yields 8 curves, one for each cluster h defined in Eq. 3. The first cluster is drawn in red (+) distributing all the schools located in comuni where only one school is provided. The orange line (○) distributes all the schools provided in comuni with two or three schools (i.e. $h = 2$); and so on. The interesting point of Fig. 4(a) is that only the school-size distribution of the smallest comuni (with $n_k = 1$) features a unimodal shape. The reason for that relies on the fact that comuni with only one school are geographically similar: they are the 57% of the total, with little more than 2000 inhabitants, the 81% of which are located in mountain territories.

The relationship between school-size and altitude is investigated in Fig. 4(b). Instead of conditioning on $\mathcal{M}$, here we propose a more consistent procedure according to which comuni are assigned to different bins on the basis of the altitude. In such a way, we can analyze comuni with 1,000 meters above the sea differently to those

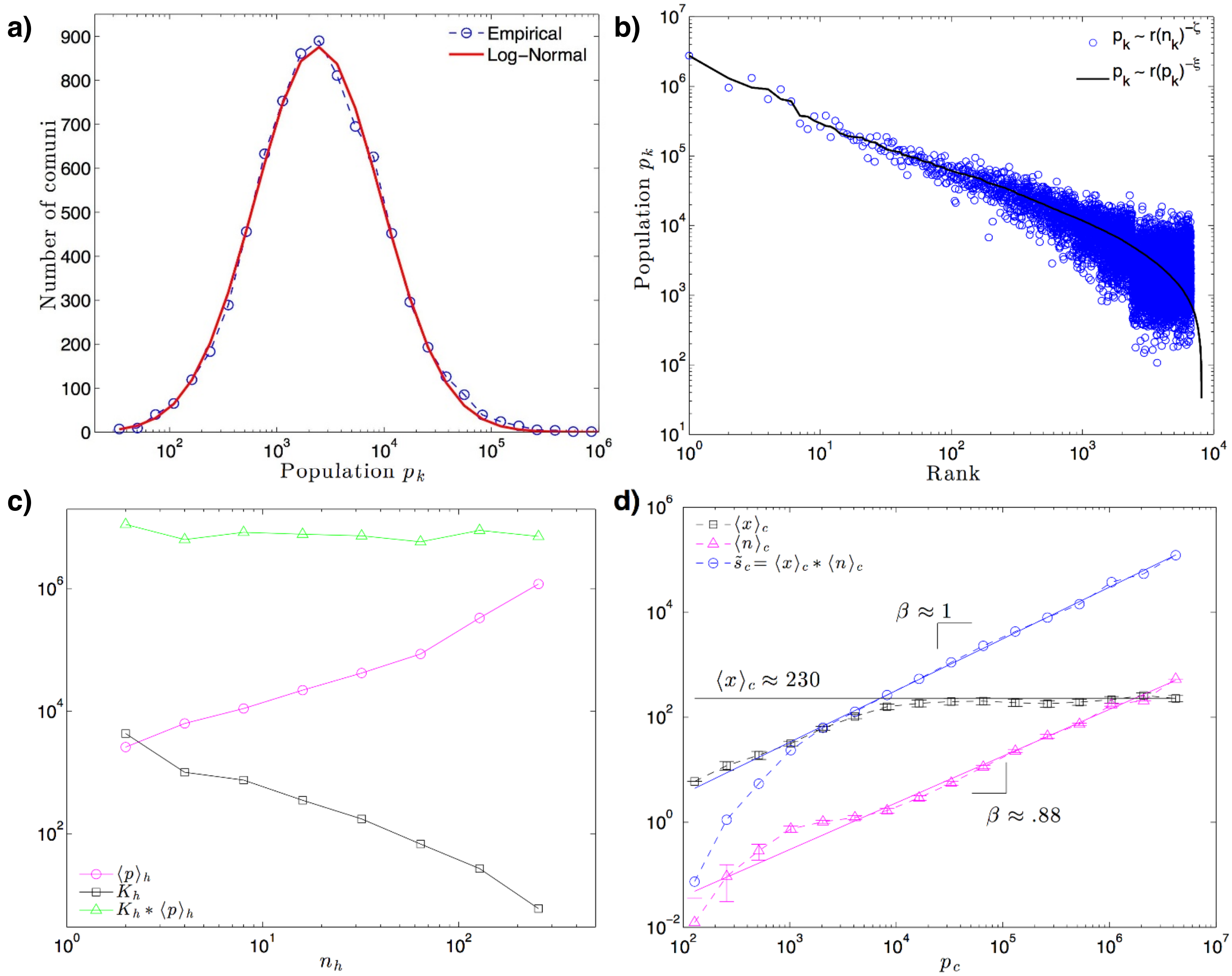

Figure 3 | **Population and cities features.** (a). The Italian city-size distribution for $K = 8092$ observations. Blue circles stand for each city-bin whereas the red solid line draws the log-normal fit of the data. Conversely to the school-size distribution depicted in Fig. 1(a), the city-size PDF features single-peakedness, but similarly it has a power-law decay in the upper tail. (b). Zipf plot for Italian cities according to the size p_k and the number of schools n_k. The black line draws the classical Zipf plot $p_k \sim r(p_k)^{-\xi}$, with cities ranked according to population p_k. Blue circles instead depict the Zipf plot $p_k \sim r(n_k)^{-\zeta}$, with cities ranked according to the number of schools n_k. Consequently, the sample reduces to $M = 6726$ over $N = 8092$ since more of the 15% of the cities have no schools. (c). Each comune is assigned to 8 clusters, according to Eq. 3, and scattered against population, the magenta line (○) and the number of cities K_h, the black line (□). The interaction term, $K_h * \langle p \rangle_h$, the green line (△), represents the total population living in each city-cluster h. (d). According to Eq. 4 K cities are assigned to $C = 16$ clusters. In the x-axis the number of inhabitants in cluster $c = \{7, 22\}$ is scattered against the average number of schools (magenta line (△)) and the average school-size $\langle x \rangle_c$ (the black line (□)). The interaction term (○), representing the typical number of schooling-aged population in cluster c, $\tilde{s}_c = \langle x \rangle_c * \langle n \rangle_c$ distributes as a power law with coefficient $\beta \approx 1$ for cities bigger than 10^3 inhabitants, and it is drawn in green. For smaller comuni, instead, the line drops meaning that a smaller fraction of young people features them.

with 600 meters of altitude that would be gathered in the same cluster $\mathcal{M}$ throwing away informative heterogeneity. It yields 5 bins: the first bin (drawn as a blue + line) gathers all the comuni whose altitude is lower than 125 meters above the see level (labeled 125 in Fig. 4(b)). Comuni with an altitude between 125 and 250 meters above the sea level composed the second bin (the green ○ line). These two distributions cluster around the second mode m_2 and in the Supplementary Information we additionally demonstrate that the hypothesis of bimodality can be rejected for the latter distribution. However, the greater the altitude of the comuni the greater is the shift of the corresponding school-size distribution toward the small schools and the greater is the contribution of these comuni to the first mode m_1. Such a shift becomes evident for comuni with an altitude between 250 m and 500 m (red Δ line). Comuni located between 500 m and 1000 m (cyan × line) and above 1000 m (purple □) line) clusterize around m_1.

Even the largest cities are very different from each other in terms of their school size distribution. This heterogeneity is very likely to be driven by geographical features. We argue this point in Fig. 4(c), where we restrict our interest on the largest Italian cities belonging to cluster $h = 8$ (and to the first two bins in terms of altitude in Fig. 4(b)). These cities provide a number of schools n_k between 127 and 255, whose overall size distribution shows a three-peak shape with a third peak around 300 students absent in smaller cities (the bottom violet Δ line in Fig. 4(a)). The presence of the three peaks around 100, 200 and 300 students might suggest the presence of architectural standards of school buildings supporting these particular sizes. However, by plotting the distribution by city we show that

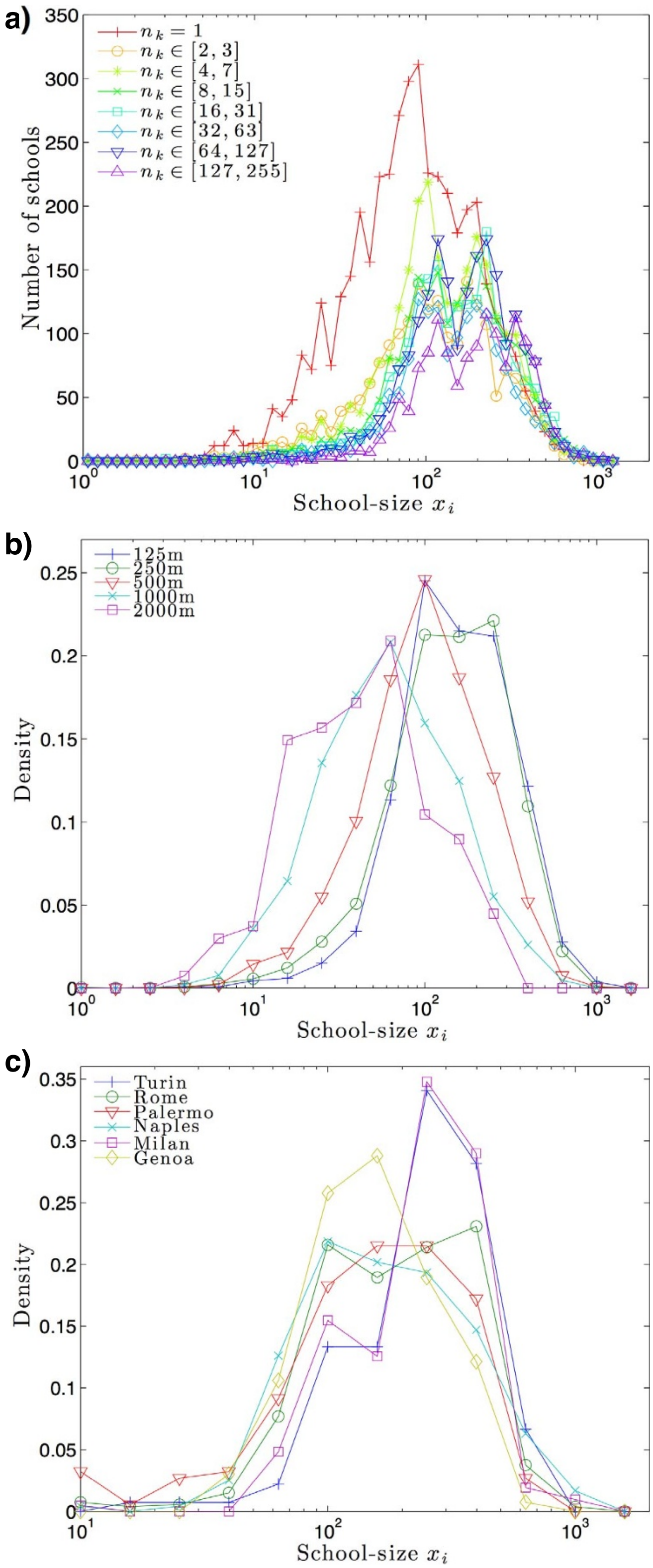

Figure 4 | School-size distribution conditional on comuni features. (a). School-size distribution for different city-samples clustered according to the number of schools, i.e. to Eq. 3. Only comuni with $n_k = 1$ show a single peak school-size distribution, clustered around m_1 (the +-red line on the top). They have an average population of 2000 inhabitants and the 81% are located in mountain territories. (b). School-size distribution for different city-samples clustered according to the altitude. The altitude of the comune shift the school-size distribution (shift location effect) as higher comuni are generally smaller schools. (c). School-size distribution in the six biggest Italian cities. With the exception of Rome, the hypothesis of unimodality may not be rejected in none of the biggest cities. In particular, flatter cities, such as Milano and Torino, mostly contribute to second mode m_2, whereas in Genova, Italian city built upon mountains that steeply ended on the sea, all the school-size distribution stands on the left side.

all the traces of trimodality disappear. In particular flatter cities, such as Milano and Torino, mostly contribute to second mode m_2, whereas in Genova, an Italian city built upon mountains that steeply slope towards the sea, the school-size distribution is unimodal contributing mostly to the first mode m_1.

Another way to look at the effect of geography on the comunal school-size is to compute the fraction of large schools within each comune k:

$$P_k(x_i > \bar{\mu} | \forall\ i \in k) \equiv \frac{n_k(x_i > \bar{\mu})}{n_k} \quad \forall\ i \in k, \tag{5}$$

where $n_k(x_i > \bar{\mu})$ stands for the number of schools that, in each comune k, are larger than the minimum $\bar{\mu}$ of the school-size distribution shown in Fig. 1(a). It can also be interpreted as the contribution rate of a comune k to the second mode m_2. The upper panel of Fig. 5(a) diagrammatically explains how $P_k(\cdot)$ is computed.

We firstly study the relationship between $P_k(\cdot)$ and population, then looking at the spatial distribution across the Italy. In Fig. 5(a), we clusterize comuni according to Eq. 3, and for each bin h we compute the average $\langle P_k(x_i > \bar{\mu} | \forall\ i \in k)\rangle_h$ and population $\langle p_k \rangle_h$. Interestingly, the plot shows that $P_k(\cdot)$ does not increase monotonically with population, demonstrating the existence of two city-patterns. More precisely, cities with less than 10^4 inhabitants follow a pattern according to which the fraction of big schools, with $x_i > \bar{\mu}$, increases, on average, with population at a rate of $\beta_1 \approx .22$; in cities with more than 10^5 we find the effect of population to be smaller, corresponding to $\beta_2 \approx .15$. For the cities with population between 10^4 and 10^5, the fraction of large schools does not increase with size suggesting that exogenous shocks such as altitude, rugged terrain and age might shift a city in this transition zone to either mode m_1 or mode m_2.

Overall, the distribution of $P_k(x_i > \bar{\mu} | \forall\ i \in k)$ is strongly correlated with the geographical features of the comuni territory. The map in Fig. 5(b) clarifies this point; all the mountain territories, Apennines that represent the spine of the peninsula and the Alps on the northern side, turns to be comuni with small schools, since the share of small schools in mountain comuni is equal to $P(x_i \leq \bar{\mu} | k \in \mathcal{M}) = 0.72$. As soon as the probability to contribute to m_2 increases the colors get warmer; but this is very unlikely to be in the mountain territories, because less than 30% of mountain comuni contribute to the antimode. Some regional patterns are also shown in the insets. The first upper panel depicts the area around Milan, which is surrounded by warm colors that mostly dye the Pianura Padana around. On the south side, Appennines approach and colors get blue with a lot of comuni with no schools (depicted in white). This pattern is more evident in the lower panel, which maps the region of Apulia, flat and mostly red, and the Basilicata on the left side, mountainous and mostly blue colored.

Countryside versus dense regions. In this last section, we bring more evidence on the effect of geography and comuni organization on the school-size by restricting our attention at two Italian regions: Abruzzo and Tuscany. But same results stand by looking at regions with the same geographical features. The two regions have very peculiar and representative geographical and administrative characteristics. Abruzzo is a mostly mountain region with a little flat seaside; it has four main head towns divided from each other by mountains. Conversely, Tuscany has many flat zones in the center

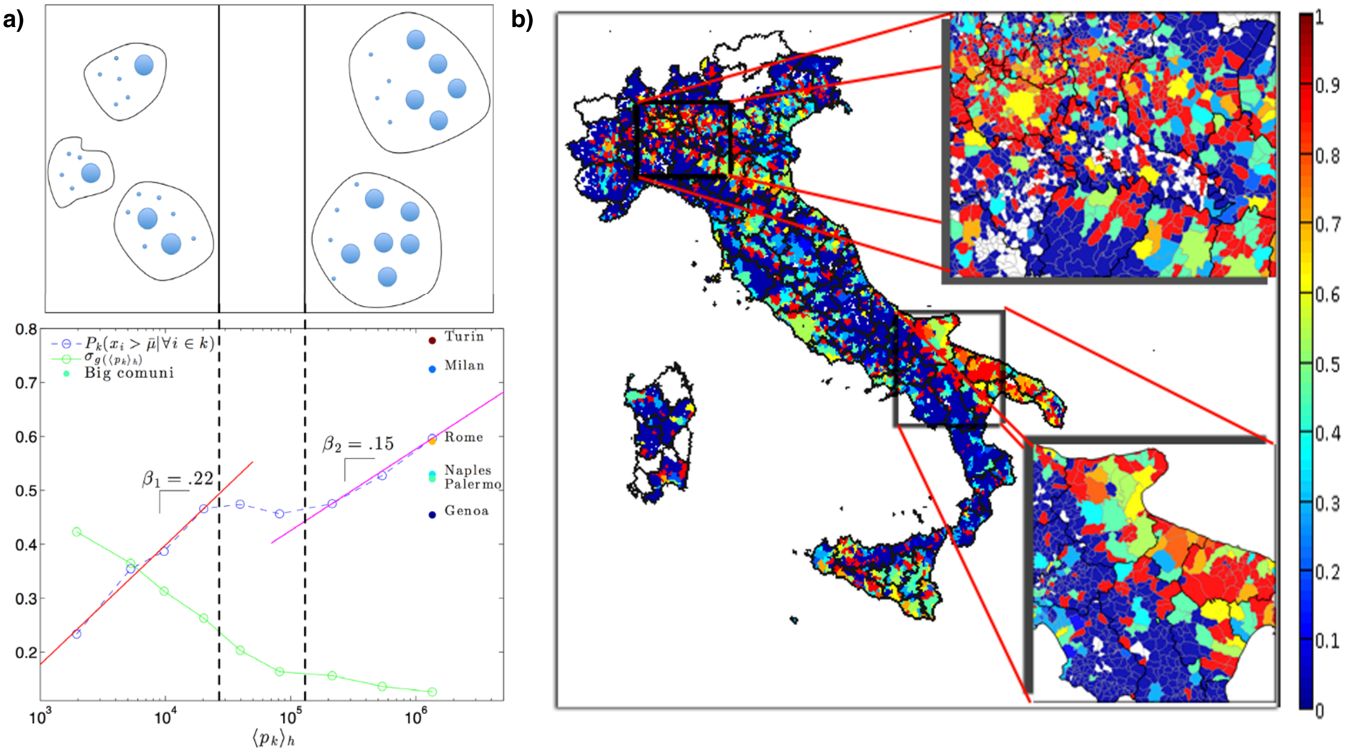

Figure 5 | **Fraction of large schools in comune *k*.** (a). The panel above shows the process according to which each comune, with population p_k defined by the size of the the black circles, is assigned to either patterns on the basis of the size of the schools provided in there (the small blue circles). The panel below shows that more populated clusters of cities are, on average, more likely to have schools sized around m_2. The relationship, depicted in blue, is however non monotonic. In correspondence of each bin h, the standard deviations has been computed, underlining the outstanding variability in very small cities (the green line). (b). Spatial distribution of cities according to $P_k(x_k > \bar{\mu}|\forall i \in k)$. Warmer territories stand for cities more likely of having schools distributed around m_2. The two figure inset underline the region around Milan (in the North), on the top, and the regions of Basilicata (mostly mountain, at the left side) and of Apulia (mostly flat, at the right side), on the bottom. Maps generated with Matlab.

and the mountain areas shape the region boundaries. Remarkably, it has a very high densely populated zone along the metropolitan area composed by Florence, Pisa and Livorno.

These two regions also differ in terms of administrative organizations, Abruzzo favoring the establishment of comuni with a smaller size due to the presence of mountains. Figure 6(a) shows the comuni population distributions in Tuscany and Abruzzo. We clusterize comuni using the algorithm in Eq. 4. As Fig. 6(a) makes clear, comuni distribute approximately as a log-normal PDF in both regions, i.e. as a parabola in a log-log scale (the green-○ line stands for Abruzzo PDF, the magenta-Δ for Tuscany). Nevertheless, Tuscany has bigger cities. Figure 6(a) also shows the average number of schools as function of the population. The fact that $\langle n_K \rangle$ is less than one for small comuni reflects the fact that many of these comuni do not provide schools. Abruzzo has a larger number of small comuni that do not provide schools. The first bin collects comuni with a bit more than 100 inhabitants. They are 7 in Abruzzo (none in Tuscany), none of them providing any school services. The second bin accumulates 10 comuni in Abruzzo with 300 inhabitants (none in Tuscany), of which only one has a school. Comuni with about 600 inhabitants are 40 in Abruzzo and only 7 in Tuscany. Only 30% of them have one school in Abruzzo while 80% of them have at least one school in Tuscany. Overall, there are 53 comuni in Abruzzo without schools; only 3 in Tuscany.

Such differences reflects on the school-size distribution, depicted in Fig. 6(b). Although primary schools distribute in both regions in terms of size with two peaks, both Abruzzo m_1 and m_2 are shifted on the left with respect to the Tuscany ones. The average school-size is smaller in Abruzzo ($\hat{\mu}_{ABR} = 4.56$ ($\hat{\mu}_{ABR}/\ln(10) = 1.98$) versus $\hat{\mu}_{TOS} = 4.91$ ($\hat{\mu}_{TOS}/\ln(10) = 2.13$)), and, remarkably, the lower tail is fatter in the former region. The cutoff for splitting the mixed distributions amounts to 128 in Abruzzo and 151 in Tuscany, and 31% of the schools are clustered in the second peak in the former region; $P(x_i > \bar{\mu}_{TOS}|\forall\ i \in TOS) = 0.38$ in the latter.

In Fig. 6(c) we show, following the same clustering technique used in Fig. 5(a), that in both regions the fraction of big schools within comune k, $P_k(x_i > \bar{\mu}|\forall\ i \in k)$, increases monotonically with respect to the number of inhabitants for $\langle p_k \rangle_h < 20000$. In this interval, a comparison with figures for entire Italy, plotted in Fig. 5(a), reveals that both regions follow the same national pattern. Yet, mountain regions, such as Abruzzo, have a significantly smaller concentration of big schools. In particular in Abruzzo, only about 1/10 of comuni with just one school, with an average population of roughly 2000, have a school with more than 125 students. In Tuscany, they are the 25%, about the same as national ratio. In larger comuni, with an average population of 5000 and two schools provided (the second bin), the probability of having big schools raises to 0.2 in Abruzzo, still smaller than Tuscany where $\langle P_k(x_i > \bar{\mu}|\forall i \in k) \rangle_{h=2} = 0.3$.

Small schools are mainly located in the countryside, and for that reason they cluster together, i.e. it is more likely to find a small school near a small one. In Abruzzo this clustering effect is stronger than in Tuscany. We investigate this point in Fig. 6(d), where we compute, and plot on the x-axis, the cumulative probability $P(x_i \leq x^*)$, as function of x^*, and the correspondent conditional probability $P(x_{\underline{i}} \leq x^*|x_i \leq x^*)$, on the y-axis, which is the fraction of schools with the size smaller than x^* among the schools closest to a school of size x^*. This quantity is equal to 74% and 65% for $x^* \equiv \bar{\mu}_{reg}$ in Abruzzo and Tuscany respectively, meaning that there is a greater probability that a small school matches with another of the same kind in the former region. If the conditional probability were equal to the cumulative, as indicated by the black line in Fig. 6(d), the sizes of

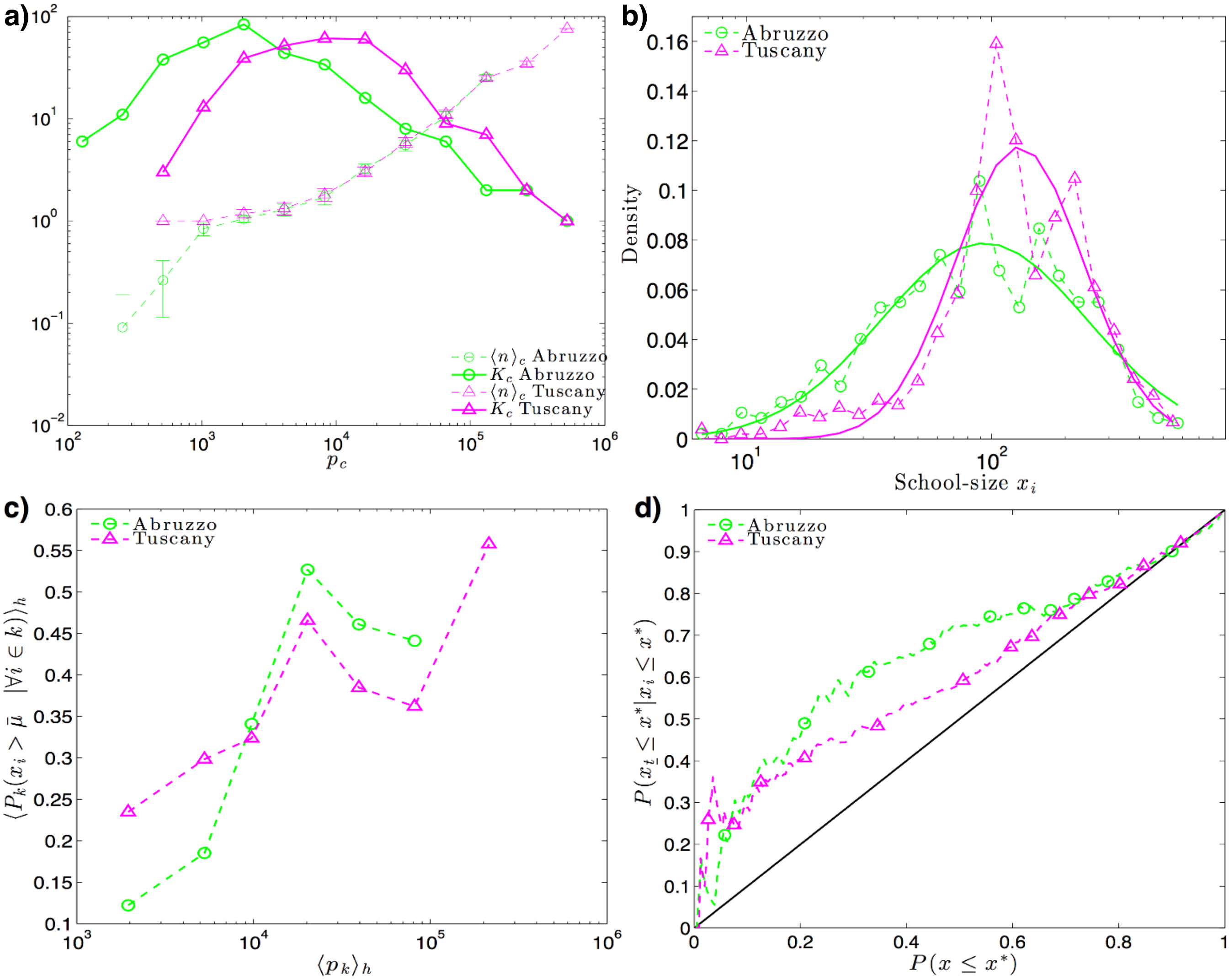

Figure 6 | **Regional analysis.** (a). The figure distributes the city-size in Abruzzo (○-green) and Tuscany (△-magenta) by plotting the number of comuni, K_c, against the number of inhabitants, p_c. Also shown is the average number of schools in a comune in Abruzzo and Tuscany, belonging to a bin c defined by Eq. 4, by the circled- and triangled-connected lines respectively. (b). School-size distribution in Abruzzo (○-green) and Tuscany (△-magenta). Both PDFs are approximately lognormal and bimodal with splitting point equal to 128 and 151 students per school respectively. (c). Average fraction of big schools in each comuni bin, defined by Eq. 3, in Abruzzo (○-green) and Tuscany △-magenta). The plot shows that more populated comuni are, on average, more likely to have schools sized around m_2, in both regions. Yet, in mountain regions, such as Abruzzo, smaller comuni have also smaller schools on average. (d). The conditional probability is plotted in the y-axis, for an arbitrary school size x^*, as function of x^* against the cumulative probability $P(x_i \le x^*)$. The conditional probability is equal to the cumulative in correspondence of the black line. Along these points, there is no attraction between schools of the same size. This is not the case in both the two regions.

neighboring schools would be independent. This is not the case in either of the two regions. The probability that a small school has a smaller nearest neighbor is larger than the probability that any school is smaller than a given one. Indeed, the two curves (green for Abruzzo and magenta for Tuscany) are significantly above the 45 degree line for $P(x_i < x^*) < 0.6$ in Tuscany and for $P(x_i < x^*) < 0.7$ in Abruzzo. These probability values roughly correspond to the probabilities $P(x_i < \bar{\mu})$ in respectively Tuscany and Abruzzo, indicating that in both regions small schools are likely to belong to the small mountainous comuni, whose nearest neighbors are of the same class.

We further study the attraction intensity among small schools by disentangling the effect between the countryside and dense zones. To this end, we analyze the GPS location of the schools in the two regions and, for each school i, we compute the number of schools n^i_m belonging within a circle of radius r_m centered at each school j. We exclude from n^i_m all the schools which do not belong to Tuscany or Abruzzo, respectively. To eliminate the effect of region's boundaries, we also compute areas D^i_m as the areas of the intersections of these circles with a given region (Abruzzo or Tuscany). Thus $D^i_m \le \pi(r^i_m)^2$, because these areas do not include the seaside and administrative territories of other regions. The difference between two subsequent circles yields the area of the annulus $A^i_m = D^i_m - D^i_{m-1}$. The density of schools in the area A^i_m is then defined as:

$$\rho^i_m = \frac{n^i_m - n^i_{m-1}}{A^i_m}, \tag{6}$$

and the average density of schools as function of a distance to a randomly selected school is

$$\langle \rho_m \rangle_i = \frac{\sum_N n^i_i - \sum_N n^i_{m-1}}{\sum_N A^i_m}. \tag{7}$$

In Fig. 7(a) red lines represent the average school-density around *all* the schools in Tuscany and Abruzzo, which are 472 in the former and

1037 in the latter region. Green lines describe the average school density around a small school with $x_i \leq \bar{\mu}$, named S_1, whereas the blue lines describe the density around large schools, S_2 with $x_i > \bar{\mu}$. 64% of the schools in Abruzzo belong to the S_1 group, 53% in Tuscany. Fig. 7(a) collects evidence about the fact that small schools S_1 are located in low school density zones and, accordingly, have a smaller probability to be surrounded by competitor schools than large schools (S_2) located in densely populated areas. In both regions, in fact, the green line goes under the blue one, for at least first 50 *km*. In particular, within this distance, in Abruzzo the density stays almost constant at approximately 0.053 meaning that 1 school is provided every 20 km^2. In Tuscany, this figure goes up to 0.07, because of a generally higher population density, but yet small.

The correlation coefficients between the school size and the distance to it nearest neighbor are negative in both regions, but the magnitude is quite different, equal to 0.34 in Abruzzo, that is 1.7 times greater than in Tuscany (0.20). To reduce the noise, we proceed by clusterizing schools according to their size. Fig. 7(b) confirms this pattern by showing that small schools have on average more distant nearest schools. We look at the size of each school in both regions, and we define the geodesic euclidean distance between the school i and its nearest neighboring school (which we denote by subscript $\underline{t}$) as $d(x_i, x_{\underline{t}})$. The binning algorithm used is to base 2:

$$l = \{\forall i \in [1, \ldots, N] : 2^{l-1} \leq x_i < 2^l\}. \quad (8)$$

This clusterization yields 8 bins, with different average sizes plotted on the x-axis of Fig. 7(b). On the y-axis, we plot the average distance between the school i, that belongs to the bin l, and its nearest, i.e. $\langle d(x_i, x_{\underline{t}}) \rangle_l$. Each school-bin l is depicted by green circles for Abruzzo and magenta triangles for Tuscany. The average distance between the closest schools decreases with respect to the average school size $\langle x_i \rangle_l$ for $\langle x_i \rangle_l > 32$ in both regions meaning that, in general, small schools are sparser than large schools that are more likely to be located in very dense zones, like cities. The non-monotonic behavior of this quantity for $\langle x_i \rangle_l < 32$ in Tuscany can be explained by the fact that such small schools in Tuscany are usually hospital schools which are located in densely populated areas. Whereas the schools provided in small islands, at least 20 *km* far from the coast, have been removed in order to eliminate any artificial bias from the spatial analysis. The three first magenta bins are all below the green ones, confirming, in accordance with the geographical features of the two regions, that in Abruzzo small schools are sparser and more likely to be located in the countryside where the school density is low (see Fig. 7(a)). Moreover, small schools on average have a distance to the nearest neighbor of 4–5 *km* which is the average distance between a small comune and a more school-dense one (see the Methods section).

The two regions then outline very different patterns of the school system in the countryside. In Abruzzo small schools are uniformly distributed across small comuni, as a result of a policy favoring the disaggregation of the comuni and school organization, due to a tight geographical constraint. In Tuscany, instead, a different system has been implemented, according to geographic features and a higher population density, where small comuni are larger and do not necessarily have small schools, especially if they stand in very populated zones.

Discussion

We have studied the main features of the size distribution of the Italian primary schools, including the sources of the bimodality, and we have investigated its relation to the characteristics of the Italian cities. The fat left tail of the distribution is the consequences of political decisions to provide small schools in small (mostly countryside) comuni, instead of increasing the efficiency of public transportations. This is most probably caused by the topographical features of the hilly terrain making transportation of students dangerous and costly. The evidence of this conclusions is that hilly cities like Palermo, Napoli, an, above all, Genoa, with steep mountains that end up into the sea, have higher fraction of small schools than mainly flat cities like Torino and Milano.

The analysis of schools growth rates highlights that the schools dynamics follows the Gibrat law, and both the growth rate distribution and the size distribution are consistent with a Bose-Einstein

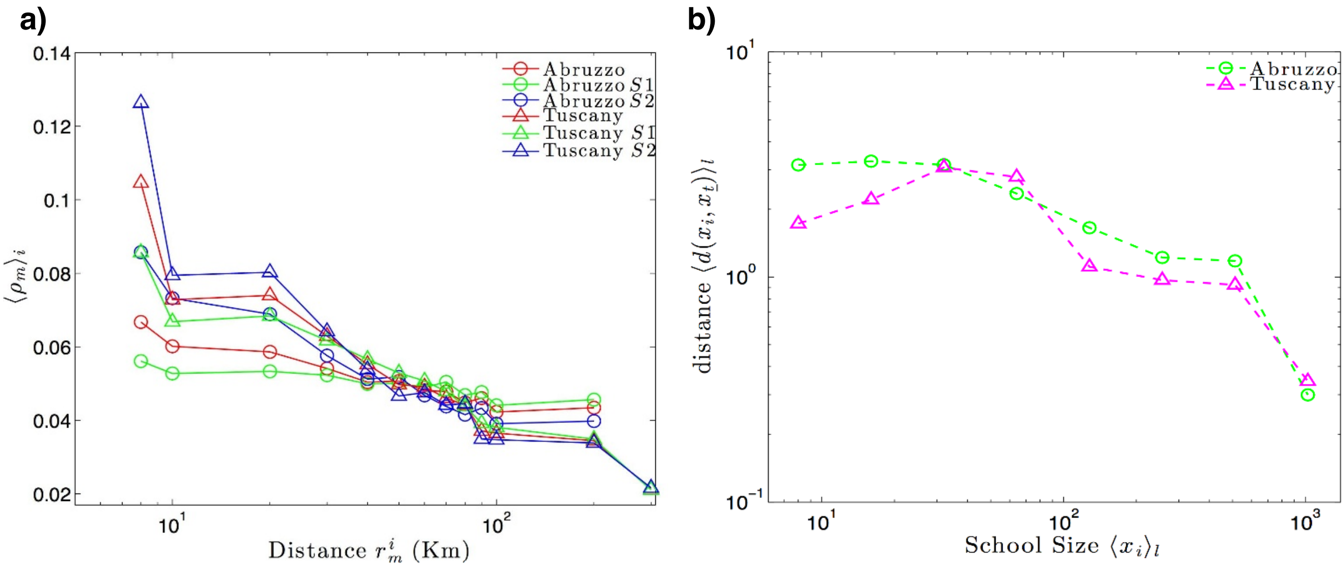

Figure 7 | Regional spatial analysis. (a). $\langle \rho_m \rangle_i$ has been plotted, based on Eq. 6, and 7, for the region of Abruzzo (□) and Tuscany (△). The red line draws the trajectory averaging among *all* the schools in Italy. Green and blue lines stand for small schools, i.e. $x_i \leq \bar{\mu}$, called S_1, and big schools, i.e. $x_i > \bar{\mu}$, called S_2, respectively. (b). The average distance, in km, between the closest schools, $\langle d(x_i, x_{\underline{t}}) \rangle_l$, is plotted in Abruzzo (○-green) and Tuscany (△-magenta) with respect to the average size, $\langle x_i \rangle_l$. Each cluster l has been obtained by aggregating schools with near size according to Eq. 8. In Tuscany, the schools provided in small islands, at least 20 *km* far from the coast, have been removed in order to eliminate any artificial bias from the spatial analysis, whereas the 18% of the schools, with no address provided in the MIUR dataset, have been geocoded in Tuscany according to the GPS localization of the city hall of the comune in which they stand. The average distance between the closest schools decreases in both regions with respect to the average size meaning that, in general, small schools are sparser than large schools that are more likely to be located in very dense zones, like cities.

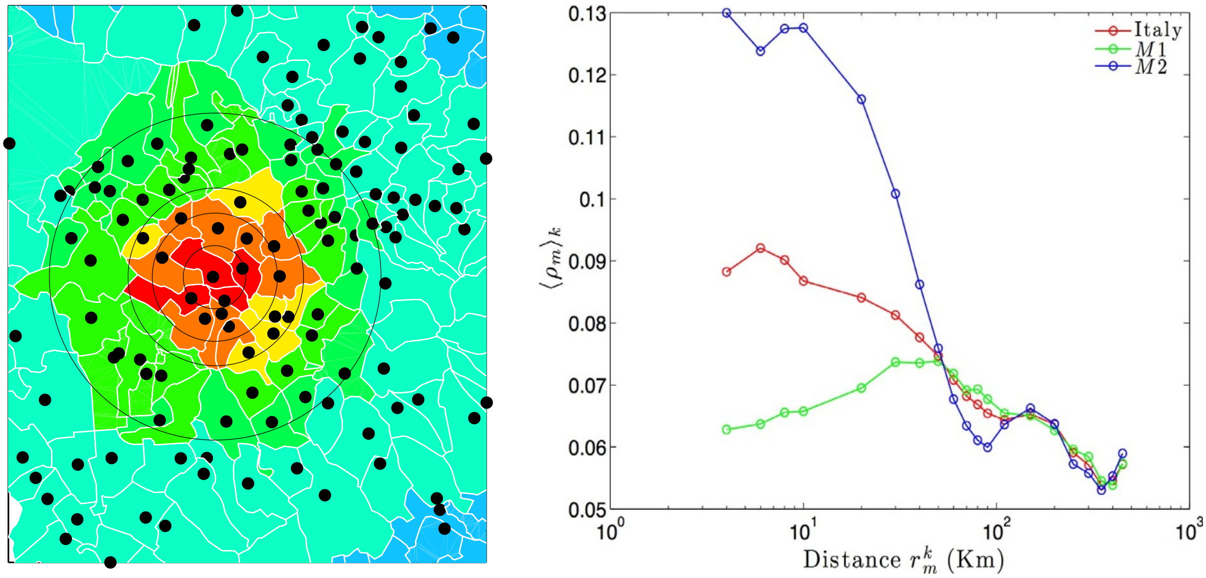

Figure 8 | Spatial analysis. (a). Graphical example for a small comune in Abruzzo of the algorithm used in Fig. 8(b), based on the Eq. 11, 12, and 13. Different comuni are colored according to the annulus in which they belong. (b). $\langle \rho_m \rangle_k$ has been plotted for a radius r_m^k of length 10^3 across Italy. The red line draws the trajectory averaging among *all* the cities in Italy. Green and blue lines stand for cities with probability $P_k(x_i > \bar{\mu}|\forall i \in k) \leq 1/2$, labeled $M1$, and $P_k(x_i > \bar{\mu}|\forall i \in k) > 1/2$, labeled $M2$, respectively. Maps generated with Matlab.

process. Alternatively, the exponential decay of the upper tail can be explained by a constraint by the size of the building or a traveling distance and transportation cost.

Despite our results are conducted using data on Italian primary schools, they predict that schooling organization would be different in another country with different geographical features. Flat territory would lead to open schools in the main villages allowing the children residing in the smallest ones to travel daily. This result is additionally supported by the fact that no territorial constraint has been imposed to the schooling choice. Despite parents can enroll children in the most preferred school, primary students generally do not move *across* comuni to attend a school. Accordingly, we find that school density and school-size are prevalently driven by the population density and then by the geographical features of the territory, as a result of a random process in the school choice made by the parents. This goes in the opposite direction with what has been found in other countries such as USA where school choices influence residential preferences of parents and drive the real estate prices in townships depending on the quality of their schools[29].

The availability of new longitudinal school data will be relevant to a more in-depth analysis and further discussions. Moreover, the availability of data for other similar countries would favor comparison and would be useful to assert our theory. We believe that this study, and future research, can lead to a higher level of understanding of these phenomena and can be useful for a more effective policy making.

Methods

In this section we propose a novel algorithm for the analysis of spatial distribution of primary schools in entire Italy. This algorithm is needed if the exact coordinates of individual schools are not available, but instead, the centers and the territories of all the comuni are known. For each comune k, we define a gravity center g_k of its territory corresponding to the GPS location of its city hall, and t_k as the area of the comune administration. In Italy the city hall is located in the center of the densely populated part of the administrative division, in order to be easily reachable by the majority of inhabitants. We develop a novel spatial-geographical approach consisting of a sequence of geographic regions bounded by two concentric circles, that we exemplified in Fig. 8(a) for a comune in Abruzzo. First we define a set Z_m^k of comuni whose city halls are within a circle of radius r_m^k centered at the city hall of comune k. Formally,

$$Z_m^k = \{\forall\, j \in [1, \ldots K] : d(g_k, g_i) \leq r_m^k\}. \quad (9)$$

Next we compute the number of schools provided by the comuni which are members of set Z_m^k that is defined by

$$n_m^k = \sum_{j \in Z_m^k} n_j \quad (10)$$

and their area

$$D_m^k = \sum_{j \in Z_m^k} t_j, \quad (11)$$

where t_j is the area of comuni j. Next we compute the area associated with all the comuni in the m-th concentric annulus surrounding comune k as the difference between the area associated with the larger circle m of radius r_m^k and the area associated with the smaller circle $m-1$ of radius r_{m-1}^k, i.e. $A_m^k = D_m^k - D_{m-1}^k$. In Fig. 8(a), each comune territory is colored with different colors according to the annulus in which they belong.

The density of schools in the area A_m^k is then defined as:

$$\rho_m^k = \frac{n_m^k - n_{m-1}^k}{A_m^k} \quad (12)$$

Then we compute the average density of schools around any school in Italy as:

$$\langle \rho_m \rangle_k = \frac{\sum_K n_m^k - \sum_K n_{m-1}^k}{\sum_K A_m^k} \quad (13)$$

In Fig. 8(b), we plot $\langle \rho_m \rangle_k$ averaged over all the $K = 8092$ Italian comuni as a function of the radius r_m that goes up to 10^3 Km across the entire Italy. The red line represents the average school-density among *all* the cities in Italy. On average, Italian comuni stand within very dense zones providing almost 1 school per 10 km^2. The dense zones generally last for 10 km and, after that, a smoothed depletion zone is experienced. However, the average distance between a comune k and a very large city with many schools is about 100 km, accordingly we see a second peak in the average school density at distance 100 km.

The full sample analysis basically averages heterogeneous characteristics that feature different types of comuni. The interaction among schools can be better understood by splitting the sample according to $P_k(x_i > \bar{\mu}|\forall i \in k)$. In Fig. 8(b), comuni with $P_k(x_i > \bar{\mu}|\forall i \in k) \leq 1/2$, i.e. with predominantly small schools, are named $M1$. The others, with predominantly big schools, are called $M2$.

- $M2$-comuni, the blue line, are (on average) more likely to be surrounded by school-dense comumi. These comuni are located in densely populated areas (depicted in red in Fig. 5(b)) where the school density is large (1.3 schools stand on average within 10 km^2). As far as the distance increases mountainous areas (and hence $M1$-comuni) are encountered and, as a result, the density of schools is found to dramatically decrease.
- The green line describes instead comuni labeled $M1$ where a smaller school density is found. Within 10 km, in fact, almost one school per 20 km^2 is encountered on average, about the half of what we find for the $M2$-comuni. This is

because $M1$-comuni mainly stand along the countryside (those depicted in blue in Fig. 5(b)) where school density slowly increases with distance and reach a maximum at approximately 40 km, which can be interpreted as a typical distance to a densely populated area in a neighboring mountain valley. After this distance the density of schools around $M1$ and $M2$ comuni behave approximately in the same way.

1. Gabaix, X. Power Laws in Economics and Finance. *Annu. Rev. Econ.* **1**, 255–93 (2009).
2. Gabaix, X. Zipf's Law for Cities: An Explanation. *Q J Econ.* **114**, 739–67 (1999).
3. Allen, P. M. *Cities and regions as self-organizing systems: models of complexity.* (Routledge, 1997).
4. Amaral, L. A. N. *et al.* Power Law Scaling for a System of Interacting Units with Complex Internal Structure. *Phys. Rev. Lett.* **80**, 1385–1388 (1998).
5. Byrne, D. *Complexity theory and the social sciences: an introduction,* (Routledge, 2002).
6. Caves, R. E. Industrial Organization and New Findings on the Turnover and Mobility of Firms. *J Econ Lit.* **36**, 1947–82 (1998).
7. Bak, P. *How nature works* (Oxford University Press, 1997).
8. Kauffman, S. *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity: The Search for the Laws of Self-Organization and Complexity.* (Oxford University Press, 1996).
9. Decreto del Presidente della Repubblica del 20 marzo 2009. *Gazzetta Ufficiale* **81**, (2 luglio 2009).
10. Disposizioni concernenti la riorganizzazione della rete scolastica, la formazione delle classi e la determinazione degli organici del personale della scuola. *D. M.* **331**, (24 luglio 1998).
11. Belmonte, A. & Pennisi, A. Education reforms and teachers needs: a longterm territorial analysis. *IJRS.* **12**, 87-114 (2013).
12. De Wit, G. Firm Size Distributions: An Overview of Steady-State Distributions Resulting form Firm Dynamics Models. *Int J Ind Organ.* **23**, 423–50 (2005).
13. Fu, D. *et al.* The growth of business firms: Theoretical framework and empirical evidence. *Proc Natl Acad Sci USA.* **102**, 18801–18806 (2005).
14. Decreto del Presidente della Repubblica dell'8 marzo 1999. *Gazzetta Ufficiale* 275, (10 Agosto 1999).
15. Pitman, E. J. G. The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika* **30**, 391–421 (1939).
16. Stanley, M. H. R. *et al.* Zipf plots and the size distribution of firms. *Econ Lett.* **49**, 453–457 (1995).
17. Gibrat R. *Les Inégalités économiques,* (Paris: Recueil Sirey 1931).
18. Sutton, J. Gibrat's Legacy. *J Econ Lit.* **35**, 40–59 (1997).
19. Fu, D. *et al.* A Generalized Preferential Attachment Model for Business Firms Growth Rates-I. Empirical Evidence. *Eur Phys J B.* **57**, 127–130 (2007).
20. Pammolli, F. *et al.* A generalized preferential attachment model for business firms growth rates: II. Mathematical treatment. *Eur Phys J B.* **57**, 131–138 (2007).
21. Axtell, R. L. Zipf Distribution of U.S. Firm Sizes. *Science* **293**, 1818–20 (2001).
22. Growiec, J., Pammolli, F., Riccaboni, M. & Stanley, H. E. On the Size Distribution of Business Firms. *Econ Lett.* **98**, 207–12 (2007).
23. Stanley, M. H. R. *et al.* Scaling behaviour in the growth of companies. *Nature* **379**, 804–6 (1996).
24. Ayebo, A. & Kozubowski, T. J. An asymmetric generalization of Gaussian and Laplace laws. *J. Probab. Stat. Sci.* **1**, 187–210 (2003).
25. Wang, J., Wen, S., Symmans, W. F., Pusztai, L. & Coombes, K. R. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Inform* **7**, 199–216 (2009).
26. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM review.* **51**, 661–703 (2009).
27. Eeckhout, J. Gibrat's Law for (All) Cities. *Am Econ Rev.* **94**, 1429–51 (2004).
28. Reed, W. J. The Pareto, Zipf and Other Power Laws. *Econ Lett.* **74**, 15–9 (2001).
29. Black, S. E. Do better schools matter? Parental valuation of elementary education. *Q J Econ.* **114**, 577–599 (1999).

Acknowledgments

We acknowledge Stefano Sebastio for many interesting and important discussions. We are also grateful to all the participants at the LIME seminars in IMTLucca. S.V.B. thanks the Dr. Bernard W. Gamson Computational Science Center at Yeshiva College for support.

Author contributions

A.B. & R.D.C. analyzed the data. R.D.C. created the maps. A.B., R.C.D. & S.B. devised the research, wrote and revised the main manuscript text.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Belmonte, A., Di Clemente, R. & Buldyrev, S.V. The Italian primary school-size distribution and the city-size: a complex nexus. *Sci. Rep.* **4**, 5301; DOI:10.1038/srep05301 (2014).