A Universal model for city-size distributions through rank ordering

Abhik Ghosh\(^1\) and Banasri Basu\(^2\)

\(^1\)Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata 700108, India
\(^2\)Physics and Applied Mathematics Unit, Indian Statistical Institute, Kolkata 700108, India

(Dated: September 25, 2018)

We propose a universal two-parameter Rank-Ordering (RO) distribution for the urban agglomerations of various countries in the world, where the urban agglomerations include the small and mid-sized cities along with the heavily populated cities. Our proposition is validated by an exhaustive study with the data for India and China in 3 decades’ census and USA for a time window of 8 years. Moreover, we have studied the city size distributions for different countries, like Brazil, Italy, Sweden, Australia, Uganda etc., from all the continents around the world according to the availability of the data. The detailed analyses shown here exhibit a unique global pattern for the city size distributions, from low to high size, across the world with various geographic and economic conditions. Furthermore, the analysis based on the entropy of the distribution provides insights on the characteristics of the underlying randomness and spread of the city sizes within a country. Our proposed RO distribution not only fits the data for all the city sizes better than the well-known power law for all the countries considered, through its two parameters, it also helps us to characterize and discriminate their population dynamics and study their evolution over time. It further opens up the possibility of finding the underlying socio-economic drivers, those play vital role in managing the sustainable urban growth – a challenging issue in the present day context.

I. INTRODUCTION

Cities are the examples of complex systems. However, the functioning of this type of complex system depends upon social, economic, and environmental factors. In spite of the diversity of the physical forms of the cities, there is a pattern and order in the expansion of cities. The size and shape of the cities are known to play a fundamental role in social and economic life\(^1\).\(^2\). The study of the size distribution of urban areas in various parts of the world helps us to understand their dynamics and also plays a key role to manage their growth and environmental impacts.

City size distribution for various countries has been extensively studied for several decades. In general it is believed that the city sizes obey a remarkably simple law, known as Zipfs law\(^3\) (alternatively known as Pareto distribution or simply power law) for cities with heavy population. Although a considerable amount of research has been devoted to power law behaviors, in various contexts, however, when real data\(^4\)\(^7\) is analyzed, in most of the cases the power law trend holds only for an intermediate range of values; there is a power law breakdown in the lower distribution tails\(^8\)\(^9\). This has motivated the studies on the size distribution of smaller cities also in different countries\(^10\)\(^19\). Recently, it has been shown that the data for the Indian city size distribution exhibit a strong reverse Pareto in the lower tail, log-normal in the mid-range and a Pareto in the upper tail\(^18\). A relatively simpler mixture of lognormal in the lower tail and Pareto in the upper tail has also been fitted for US city size distributions\(^19\). The existing literature on modeling the sizes of all cities having small to extremely large populations are mostly, up to the knowledge of the authors, if not all, confined with the Pareto law in the upper tail; hence they all needed to use the complex mixture distributional models (with suitable other distribution in lower or middle range). Such mixture distributions are known to be difficult to fit statistically (estimation process is often complex) and often difficult to estimate; also there is no evidence of one such mixture distribution providing significantly good fit to different types of city-size distributions in socio-economically different countries across the world.

As cities are known to be the society’s paramount mechanism of innovation and wealth creation, it is a major challenge to predict a unique statistical distribution that govern the city sizes, with small to large populations, around the world. One can then compare all the cities in a universal framework and better understand the dynamics of urbanization in society, also mitigate their negative impacts on global ecosystem. The global effect of human urbanization requires a global perspective of the problem. It is very important to know how the urban areas are distributed and how the distribution varies around the world. This is typically the goal of the present paper to search for a global pattern of city size distributions along with their discrimination and potential drivers.

In this work we focus on a two-parameter Rank-Ordered (RO) distribution\(^20\) for all cities in different countries across all the continents of the world, including the important economies like India, China, USA, Brazil as well as the recent distribution of the sizes of World Cities with heavy population. The RO distribution is indeed a generalised discrete beta (GDB) distribution, where a given property of a system is ordered according to its importance (rank). Our analysis shows that the
data for all countries under consideration follow this two parameter RO distribution that incorporates the product of two power laws defined over the complete data set, one measured from left to right and another in the opposite direction. For the vanishing value of one particular parameter, the RO distribution simplifies to the usual Pareto law for the cities with heavy population (see Section II.A). To understand the underlying process of the urban morphology, the Shannon entropy of the proposed RO distribution has also been studied. Based on the two parameters of the distribution and the corresponding entropy, one can now describe the discrimination and evolution of different cities in different countries.

The rest of the paper is organised as follows. Sec. II deals with the mathematical formulation of the two-parameter RO distribution. In the respective subsections A, B and C, we define the proposed distribution, corresponding Shannon entropy and provide the methodology for the estimation of the parameters. In Sec. II.D we detail the prediction and the Goodness-of-fit test for examining the performance of our proposal. In Sec. III we provide the description of the data that have been used for our empirical analyses. In Sec. IV we have demonstrated our RO modeling for city sizes with the publicly available data of different countries across the world, including India, China, USA, Brazil, Australia, 3 European countries and 3 African countries at different time point. This section also deals with the analysis with the world wide size distribution of Cities and Countries in the year 2018 (Sec. IV.H). Sec. V describes the discrimination and evolution of the proposed distributions across the countries and also remarks about the potential drivers of those distributions. Finally, in Sec. VI we conclude with a discussion about the future works in this direction.

II. MODELING CITY-SIZES USING THE RANK-ORDER DISTRIBUTION

A. The Rank-Ordering (RO) distribution

Let us first recapitulate the mathematical formulation of rank-ordered distribution. Suppose we have data on the sizes of $N$ items (e.g., cities) arranged in decreasing order of “importance” (e.g., size) and let the $i$-th item (or city) has size $n_i$ with rank $r_i$. Note that $(r_1, \ldots, r_N)$ is a permutation of $(1, \ldots, N)$. The most widely used hyperbolic Pareto (Zipf’s) law models the data by the functional form of the probability of rank $r$ as

$$f_P(r) = A \cdot \frac{1}{r^\nu}, \quad r = 1, \ldots, N, \quad (1)$$

where $\nu > 0$ is a model parameter, known as exponent, and $A$ is the normalizing constant. As already noted in Section I this Pareto or power law provides good fit to the data only at low-ranks (large sizes). This is mostly because the empirical data often have a inflection point which is not possible to be captured by the single power law model [1].

In order to accommodate the inflection and the two ends of the ranks with equal importance for a given empirical data containing a wide range of sizes, one can think of a simple two-parameter rank ordering (RO) distribution given by [20]

$$f_{(a,b)}(r) = A \frac{(N + 1 - r)^b}{r^a}, \quad r = 1, \ldots, N, \quad (2)$$

where $a, b$ are two real model parameters and $A$ is again the normalizing constant. The two model parameters $a$ and $b$ indeed characterize the shape of the rank-size distribution at two-ends of high and low ranks, respectively, in respect to the inflection point. Figure 1 representing the shape of this RO distribution for different values of $a, b > 0$, clearly indicates its potentiality in modeling a wide variety of rank-size data structures [21]. In particular, it is interesting to note that the case $b = 0$ yields back the Zipf’s law in (1) and indicates that there is no inflection point in the empirical data so that only one parameter ($a > 0$) is sufficient to model the slope of the data (in log-log scale); see Figure 2 for the form of their distributions. The form of the RO distributions with $a, b < 0$ are symmetrically opposite (in ranks) in the sense that they put more probability to higher ranks (lower sizes) compared to the lower ranks (higher sizes).

\[FIG. 1: \text{The form of Rank-ordering distribution } f \equiv f_{(a,b)} \text{ for different parameter values}\]
It is fascinating to note that the shape of the RO distribution becomes flatter as \(a\) and \(b\) decreases close to zero, just like the effect of \(a\) on power law. However, for any fixed \(a\), the parameter \(b\) in RO distribution allows to fit wide range of different distributional curvatures of the data containing an inflection point. In terms of physical interpretation, the distribution of rank-size becomes more spreader with increasing values of \(a\), \(b\) leading to easier discrimination between individual ranks. On the other hand, as the values of \(a\) and \(b\) decrease, the sizes of the individual ranks become closer (lesser spread) and hence more difficult is their discrimination leading to larger entropy, as detailed in the following subsection.

### B. Shannon Entropy of the RO distribution

A city is not and cannot in any sense be treated as an isolated system since it always exchanges materials, energy, information and people with its surroundings. When related to a probability, the concept of the Shannon-Gibb’s entropy can be used in analysing the frequencies. When related to a probability, the concept of the entropy is directly linked with the Shannon entropy \(S\) in (3) and the corresponding cross-entropy, the MLE is also the same as the corresponding minimum cross-entropy estimator, popular in information sciences and statistical physics. The MLE also has minimum standard error in estimation, asymptotically for larger sample sizes, among a wide class of useful estimators.

Equation (3) lies at the basis of information theory [22].

The contour plots of the exact values of the entropy of the RO distribution with \(N = 200\) and 2000 are plotted over the parameters \((a, b)\) in Figure 3. Clearly, the maximum entropy corresponding to the case \(a = b = 0\) leads to the uniform distribution. The case \(b = 0\), with any \(a > 0\), corresponds to the entropy for the power law distribution. In finite sample data analysis (with finite \(N\)), it turns out to be beneficial to allow for the more general RO model in (2) with possibly different slopes \(a \neq b\) (in log-log scale) for the high and low ended ranks. Further, the pattern of the entropy is almost indifferent for different \(N\), although the value of \(S\) changes with the change of \(N\). Hence, the city-size distributions can be compared in terms of their entropy and distributional structure only through the values of the two parameters \(a\), and \(b\); we will illustrate it further in Section V.

### C. Parameter Estimation: The MLE

We now illustrate how the RO distribution can be used for universal modeling of the city-size distributions and explain the methodology for the estimation of the parameters corresponding to this distribution from the empirical data. Suppose that we have the data for a population of \(N\) cities in order of decreasing sizes and \(x_i, r_i\) denote the actual size and rank of the \(i\)-th city for \(i = 1, \ldots, N\). We can fit the RO distribution to these data by means of estimating the parameters \((a, b)\) which leads the model probabilities to be closest to the empirical (normalized) sizes in an appropriate sense. The best (asymptotically) statistical estimation technique is the maximum likelihood estimator (MLE) which measures the closeness in terms of Kullback-Leibler divergence measure and maximizes the probability of the observed sample-data under the assumed model. Since the Kullback-Leibler divergence is directly linked with the Shannon entropy \(S\) in (3) and the corresponding cross-entropy, the MLE is also the same as the corresponding minimum cross-entropy estimator, popular in information sciences and statistical physics. The MLE also has minimum standard error in estimation, asymptotically for larger sample sizes, among a wide class of useful estimators.

Given the ranked city-size data \((x_i, r_i)\), the MLE of \((a, b)\) in the best fitted RO model is to be obtained by maximizing the likelihood function given by

\[
L(a, b) = \prod_{i=1}^{N} f_{a,b}(r_i)^{x_i} = \prod_{i=1}^{N} \frac{(N + 1 - r_i)^b}{r_i^a} \frac{r_i^a}{A^a}. \tag{4}
\]

Equivalently, one can also maximize the log-likelihood function having the form

\[
l(a, b) = \log L(a, b) = b \sum_{i=1}^{N} x_i \log(N + 1 - r_i) - a \sum_{i=1}^{N} x_i \log(r_i) + \log(A) \sum_{i=1}^{N} x_i,
\]
or solve the estimating equations
\[
\frac{\partial}{\partial a} l(a, b) = 0 = \frac{\partial}{\partial b} l(a, b).
\]
Noting that, \( A \) is a function of \((a, b)\), we can numerically maximize \( l(a, b) \) or solve the above estimating equations, with respect to \( a, b \in \mathbb{R} \), to obtain their MLE, which we will denote as \((\hat{a}, \hat{b})\).

## D. Prediction and Goodness-of-fit

Once we have obtained the MLE \((\hat{a}, \hat{b})\) of the parameters \((a, b)\) of the RO distribution fitted to a given set of empirical data, one can predict the rank-sizes from this fitted model. The predicted normalized size of rank \(r_i\) is given by
\[
f_{\hat{a}, \hat{b}}(r_i)
\]
and hence, multiplying it by the total size \(\sum_{i=1}^{N} x_i\), one gets the predicted size of \(r_i\) which we denote as \(p_i = \left(\sum_{j=1}^{N} x_j\right) f_{\hat{a}, \hat{b}}(r_i)\).

If the estimated RO model is a good fit to the given data, these predicted values \(p_i\) should be close to the corresponding observed size \(x_i\); hence their plot gives us a visual indication of the goodness-of-fit of our RO model to any data. A quantitative measure of fit can also be obtained by summarizing the error in prediction; we will here consider the measure defined in terms of the Kolmogorov-Smirnov distance between the predicted and observed cumulative frequencies of ranks as given by
\[
KS = \max_{1 \leq i \leq N} \left| \left(\sum_{j:r_j \leq r_i} p_j\right) - \left(\sum_{j:r_j \leq r_i} x_j\right) \right|.
\]
The use of cumulative sizes instead of actual sizes in the definition of KS provides better stability of this goodness-of-fit measure. Note that, this KS measure becomes zero if and only if \(p_i = x_i\) for all \(i = 1, \ldots, N\), i.e., all the predicted sizes coincide exactly with the observed sizes. For a given dataset, the KS measure closer to zero indicates better goodness-of-fit and the model corresponding to minimum KS will be the “best” as a predictive model of the underlying rank-size distribution.

## III. DATA DESCRIPTION

In the last few decades the human population is increasingly clustered in urban areas. The urban population of the world has grown rapidly from 746 million in 1950 to 3.9 billion in 2014. Asia, despite its lower level of urbanization, is home to 53% of the world’s urban population, followed by Europe with 14% and Latin America and the Caribbean with 13% [24]. Today, 55% of the world’s population lives in urban areas, a proportion that is expected to increase to 68% by 2050. Projections show that urbanization, the gradual shift in residence of the human population from rural to urban areas, combined with the overall growth of the world’s population could add another 2.5 billion people to urban areas by 2050, with close to 90% of this increase taking place in Asia and Africa, according to a new United Nations data set launched in May, 2018 [25]. Together, India, China and Nigeria will account for 35% of the projected growth of the world’s urban population between 2018 and 2050. By 2050, it is projected that India will have added 416 million urban dwellers, China 255 million and Nigeria 189 million. In the present day context these data motivate us to study the size-distribution of all cities, or urban settlements in several countries, in particular India and China, once again.

As an effect of urbanization, it is now difficult to distinguish city, suburb or town. All definitions of city has shortcomings, either it is defined as incorporate area, an urban agglomeration, or a human settlement with population density larger than some threshold value [25]. For our analysis, we use the term city to mean any human settlement that is denser than its surroundings; we consider towns, cities, counties, urban agglomerations according to the availability of the data for various countries. Our datasets are public and freely available. Below we describe the data, which we have considered for our analysis in the present paper.

- Indian city size data from the censuses at the years 1991, 2001 and 2011 [29], with minimum inhabitant of 5,000;
- Chinese city and county sizes for three years 1990, 2000 and 2010, as defined in [30];
- USA data for each years between 2010 to 2017, with minimum size of 5,000, where the city populations are estimated based on 2010 census [31];
- Brazil city populations for the years 1991, 2000, 2010 and 2017 (estimated), with minimum of 20,000 inhabitants [32];
- Three African Countries, namely Uganda, Sudan and Algeria for the years 2014, 2008 and 2008, respectively, with minimum sizes of 15,000, 20,000 and 13,000 [33];
- Italian cities for the years 1981, 991, 2001, 2011 and 2017 (estimated) [34], with minimum population of 50,000;
- Sweden city size estimates for each fifth year from 1990 to 2015 and 2017 [35], having population greater than 10,000;
- Switzerland city-sizes for the years 1980, 1990, 2000 and the corresponding estimates for 2010 and 2017 [36], with minimum inhabitants of 10,000;
- Australian urban agglomerations (UA) (all) for the years 2011 and 2016 [37].
• To validate our proposition we have considered the population data for major cities around the world in the year 2018 [38].

• Finally, we have tested our model with the data on the population of 226 countries across the world [38].

IV. EMPIRICAL ILLUSTRATIONS: UNIVERSALITY IN SPACE AND TIME

In this section, we will present the results of the applications of the proposed RO modeling to different countries’ city-size data described in the previous section. For comparison, the popular Pareto model, corresponding to the Zipf’s law, has also been fitted to the same sets of data through a linear regression between logarithms of size and rank, and the superiority of our proposed RO model has been illustrated in all the cases. The highly significant fits of the RO distribution for all such cases indicate its universality over time as well as its global nature to model different kinds of city size distributions from different economies and geographies across the world.

A. City size distribution for Indian cities

Let us start our demonstration of the proposed RO distribution with Indian cities, using the census data for the years 1991, 2001 and 2011, and also remark on the viability of our prediction. According to the most recent census in 2011 [26], India is the second largest country in the world with more than 1.2 billion people, and it is predominantly rural country with about 67% population living in rural areas [38]. The country has a total of 39 cities that each has population exceeding one million residents. Of these cities, Mumbai and Delhi have populations that exceed 10 million. However, the country also has smaller but still very populated cities, including 388 that have populations exceeding 100,000, and a whopping 2,483 cities with populations of over 10,000.

As noted earlier in Section III, we consider Indian cities to be defined as human settlements with population more than 5,000. The whole data is divided into different classes (I-V) according to its population. Let us fit the proposed RO model for all these city-sizes (Class I-V) and for some subset of cities starting with a larger minimum size. In particular, we consider the subset of cities with population at least 100,000 (i.e., Class I cities), or at least 50,000 (i.e., Class I and II cities) or at least 10,000 (i.e., Class I-IV cities). The estimated parameter values and the goodness-of-fit measures (KS) for all four cases and three years are reported in Table I. The corresponding fitted (and actual) sizes are plotted in Figure 4 for the most recent year 2011; the plots for other two years are very similar and hence not presented to save space. The fitted values and goodness-of-fit measures corresponding to the Pareto modeling (along with the estimated exponent $\nu$), for the same sets of data, are also shown in Figure 4 and Table I respectively.

![Figure 4: Plots of the actual and the predicted sizes over ranks for different sets of Indian cities in 2011 (Black square: observed sizes, blue circle: RO fits, red star: Pareto fits)](image-url)

| City Classes | N   | $\hat{a}$ | $\hat{b}$ | KS | $\nu$ | KS |
|--------------|-----|-----------|-----------|----|-------|----|
| 2011         |     |           |           |    |       |    |
| I            | 468 | 0.8552    | 0.1613    | 0.0308 | 0.9684 | 0.1252 |
| I-II         | 942 | 0.8791    | 0.1641    | 0.0313 | 0.9890 | 0.1470 |
| I-IV         | 4001 | 0.9122    | 0.1899    | 0.0310 | 1.0190 | 0.1923 |
| I-V          | 5749 | 0.9144    | 0.3051    | 0.0304 | 1.0856 | 0.4418 |
| 2001         |     |           |           |    |       |    |
| I            | 394 | 0.8842    | 0.0671    | 0.0275 | 0.9328 | 0.0506 |
| I-II         | 798 | 0.8940    | 0.0610    | 0.0256 | 0.9333 | 0.0462 |
| I-IV         | 3307 | 0.9079   | 0.1349    | 0.0229 | 0.9923 | 0.1560 |
| I-V          | 4186 | 0.9056    | 0.2411    | 0.0220 | 1.0535 | 0.3605 |
| 1991         |     |           |           |    |       |    |
| I            | 289 | 0.8072    | 0.1374    | 0.0204 | 0.9129 | 0.0912 |
| I-II         | 614 | 0.8347    | 0.1166    | 0.0233 | 0.9182 | 0.0838 |
| I-IV         | 2533 | 0.8638   | 0.1823    | 0.0226 | 0.9760 | 0.1731 |
| I-V          | 3138 | 0.8614    | 0.2964    | 0.0216 | 1.0380 | 0.3642 |

The superiority of the goodness-of-fit by the proposed RO model over the usual Pareto model is clearly visible from Figure 4 and Table I in all cases. In particular, when we consider a wider range of cities having sizes significantly small as well as significantly large (i.e., enough observations and variety in both the high and low end of the ranks leading to larger $N$), the Pareto model fails...
significantly as also observed in several past literatures [10–19]. But, even in such cases, the proposed RO model approach can provide remarkably better fit with very little prediction error (KS) that is almost the same as the case with only large sizes (low-ends of the rank). This observation, along with establishing the significance of the proposed approach over the existing Zipf’s law in modeling city size distributions, also indicate the universality of the RO model in predicting the city distributions with much wider ranges (in both extremes) equally well.

B. Chinese City-size Distribution

It has been already noted the importance of China in urban planning and city-size analysis as it is the most populous country in the world with 1,415,045,928 people according to 2018 estimate [38]. With 44.2 million people in its metro area, Guangzhou is the biggest city in China followed by Shanghai, with a total population of 35.9 million. We consider the sizes of all Chinese cities and counties as recorded in the years 1990, 2000 and 2010, and model them through both the proposed RO and the usual Pareto modeling. We apply the proposed model once to those marked as City and County, in [30], separately and once to their combined pool. The fitted sizes obtained in both the approaches are plotted in Figure 5, whereas the estimated parameters and KS measures are reported in Table II.

![Figure 5](image)

**TABLE II: Estimated measures for Chinese Cities**

| Year | N  | \( \hat{a} \) | \( \hat{b} \) | KS   | \( \nu \) | KS   |
|------|----|------------|------------|------|--------|------|
| City Only | 1990 | 731         | 0.6939     | 0.5314 | 0.0081 | 1.037 | 0.4959 |
|        | 2000 | 735         | 0.6786     | 0.3695 | 0.0106 | 0.9212| 0.2665 |
|        | 2010 | 687         | 0.6899     | 0.3786 | 0.0207 | 0.9346| 0.2632 |
| County Only | 1990 | 1057        | 0.3073     | 0.4435 | 0.007  | 0.6005| 0.1303 |
|        | 2000 | 1074        | 0.2932     | 0.4016 | 0.0104 | 0.5603| 0.1033 |
|        | 2010 | 1028        | 0.2685     | 0.2801 | 0.0132 | 0.4627| 0.0622 |
| City + County Combined | 1990 | 1788        | 0.7382     | 0.5271 | 0.0188 | 1.0637| 0.6371 |
|        | 2000 | 1809        | 0.7086     | 0.47   | 0.0106 | 1.0061| 0.5065 |
|        | 2010 | 1715        | 0.7176     | 0.2333 | 0.0159 | 0.8675| 0.1737 |

C. Yearly City-size Distribution of USA

USA is the 3rd populous country in the world according to the 2018 ranking [35]. New York City, is the most populous city in USA (world city ranking 2) with a present population of 8,550,405 and expected to reach 9 million by 2040.

USA has a developed economy in contrast to the developing ones in India and China; hence its city size distribution is expected to be different from those in China or India, if we consider the lower populated cities also. To illustrate the universality and global nature of the proposed RO distribution in modeling city sizes of any kind of economy, let us now apply our proposal for USA city size data (estimated), as already described in Section III. We consider only those entries having size greater than or equal 5,000 as a city. The parameter estimates and goodness-of-fit measures (KS) obtained by both the Pareto and proposed RO modeling are given in Table III for each years from 2010 to 2017; results for 2010 are obtained both from the true census data as well as the estimated data for comparison purpose. However, due to the similarity between all these results, the plots of the corresponding fitted sizes are shown in Figure 6 only for the years 2010 and 2017, both with estimated data.

Once again the proposed RO approach fits the USA
TABLE III: Estimated measures for USA Cities (2010* denotes the census data; others are estimated data)

| Year | N    | \( \hat{a} \) | \( \hat{b} \) | KS  | \( \nu \) | KS  |
|------|------|---------------|---------------|-----|----------|-----|
| 2010*| 16330| 0.8667        | 0.3474        | 0.0276| 1.0741   | 0.7011|
| 2010 | 16397| 0.8663        | 0.3466        | 0.0274| 1.0736   | 0.7020|
| 2011 | 16412| 0.8665        | 0.3504        | 0.0274| 1.0757   | 0.7128|
| 2012 | 16418| 0.8667        | 0.3474        | 0.0277| 1.0778   | 0.7235|
| 2013 | 16443| 0.8669        | 0.3586        | 0.0278| 1.0800   | 0.7359|
| 2014 | 16436| 0.8670        | 0.3614        | 0.0279| 1.0816   | 0.7436|
| 2015 | 16449| 0.8672        | 0.3650        | 0.0280| 1.0836   | 0.7545|
| 2016 | 16456| 0.8673        | 0.3685        | 0.0281| 1.0853   | 0.7640|
| 2017 | 16459| 0.8673        | 0.3709        | 0.0282| 1.0865   | 0.7708|

The accuracy obtained by the RO approach is extremely small and quite stable, but the accuracy of Pareto modeling is about 25-30 times larger and also increases over the year due to the increase in the variation of city sizes.

TABLE IV: Estimated measures for Brazilian Cities with at least 20000 inhabitants (* denotes estimated data)

| Year | N    | \( \hat{a} \) | \( \hat{b} \) | KS  | \( \nu \) | KS  |
|------|------|---------------|---------------|-----|----------|-----|
| 1991 | 709  | 0.886         | 0.079         | 0.021| 0.947    | 0.096|
| 2000 | 922  | 0.845         | 0.157         | 0.012| 0.962    | 0.176|
| 2010 | 1096 | 0.837         | 0.170         | 0.011| 0.964    | 0.198|
| 2017*| 1204 | 0.833         | 0.190         | 0.013| 0.974    | 0.228|

D. City-size Distribution of Brazil

Next we consider another large major developing country, Brazil, from South American continent. In the population wise global ranking of 2018 estimate, Brazil is in 5th position with a population of 211,179,426 [32]. We use the census data of the Brazilian city sizes as described in Section III and apply the proposed RO modeling along with the usual Pareto modeling for the cities having at least 20,000 inhabitants (as per the availability of data). Resulting estimates are shown in Table IV which again show appreciably improved fit by our proposal with RO distribution as compared to the usual Pareto law. The fitted values obtained by the RO modeling is again seen to be much closer to the original city-sizes, as before, compared to the power law fits; hence these plots are not presented here for brevity.

E. Cities within Three African Countries

Next we try to test if the proposed model for the city sizes fits well for the African countries with underdeveloped economy. In March 2013, Africa was identified as the world’s poorest inhabited continent. Although, in 2017 the African Development Bank reported Africa to be the world’s second-fastest growing economy, and estimates that average growth will rebound to 3.4% in 2017, while growth is expected to increase by 4.3% in 2018 [39]. In this respect, it is interesting to study the city size distribution of some countries in Africa. In most such countries, proper census data is rarely available. We here present the analysis for three such countries, namely Uganda, Algeria and Sudan, where the latest census data have been available for the years 2014, 2008 and 2008, respectively; based on the availability of these data, we could only consider the cities having at least 15,000, 13,000 and 20,000 inhabitants, respectively, for the three countries [33]. We would like to make a note of a point here that though it is known that in coming years, Nigeria will become a good choice for the urban dwellers [24], but due to the unavailability of the data to the authors, the illustration for their city-size distributions cannot be given here.

Algeria is Africa’s largest country and has several cities with population of over a million residents. The country’s capital and largest city is Algiers with a population of around 3.5 million. The city of Oran is the second-largest in Algeria, with a population of around 650,000 residents. Uganda is the 81st largest country by area in the world, but it has the 36th largest population. The largest city in Uganda is Kampala, with a population of 1,353,189 people. Uganda has one city with more than a million people, two cities with people between 100,000 to 1 million, and 62 cities with people between 10,000 to 100,000. Sudan has two cities with more than a million people, 19 cities with people between 100,000 to 1 million, and 42 cities with people between 10,000 to 100,000. The largest city in Sudan is Khartoum, with a population of 1,974,647 people.

We apply both the RO and the Pareto modeling to the city size data of these three countries, and the resulting estimates and fitness measures are reported in Table V. For Uganda and Algeria, the proposed RO modeling performs as a close contender of the Pareto law which also gives reasonably well fit; note that the estimated \( \hat{b} \)
parameter in RO models are close to zero in both the cases indicating the good behavior of the Pareto modeling. But, for Sudan, the estimate of $b$ is slightly away from zero and the corresponding RO modeling provides a better fit with almost half the KS error compared to the Pareto modeling.

| Country Year | RO Model | Pareto |
|--------------|----------|--------|
|             | $N$      | $\hat{a}$ | $\hat{b}$ | $\nu$ | KS |
| Uganda 2014 | 105      | 0.917  | 0.029 | 0.048 | 0.936 | 0.046 |
| Algeria 2008| 180      | 0.800  | 0.024 | 0.044 | 0.798 | 0.046 |
| Sudan 2008  | 63       | 1.033  | 0.187 | 0.043 | 1.157 | 0.094 |

**F. City-size distributions with significantly higher entropy ($b < 0$): Three European Countries**

Next we consider the population of the cities of three smaller European countries, namely Italy, Sweden and Switzerland, all having strong developed economy. The current population of Italy is 59,277,324 based on the latest United Nations estimates \[25\]. Italy population is equivalent to 0.78% of the total world population and Italy ranks number 23 in the list of countries (and dependencies) by population. 71.8% of the population is urban (42,587,390 people in 2018 estimate). The most populated city is Roma (Rome) with population 2,648,843, followed by Milano (Milan) with 1,305,591 \[34\]. On the other hand, almost all of Sweden’s population lives in urban areas. Sweden has one city with more than a million people, 9 cities with between 100,000 and 1 million people, and 141 cities with between 10,000 and 100,000 people. The largest city in Sweden is Stockholm, with a population of 1,515,017 people \[35\]. Finally, Switzerland has no city with more than a million people, 5 cities with between 100,000 and 1 million people, and 145 cities with between 10,000 and 100,000 people. The largest city in Switzerland is Zurich, with a population of 344,730 people \[36\].

We repeat our application of the proposed RO modeling along with the Pareto distribution for the census data of Italy at the years 1981, 1991, 2001 and 2011, and the corresponding estimated city sizes for the year 2017, where we restrict our attention to the human settlements having at least 50,000 inhabitants. We also apply the same for the yearly official estimate of city sizes of Sweden, from the year 1990 to 2015 and 2017, having at least 10,000 inhabitants. For Switzerland we fit the model for the census data of the year 1980, 1990 and 2000, and the official estimates for the year 2010 and 2017, where again we restrict only to the settlements having at least 10,000 humans inhabitants. The resulting estimates and the KS measure of fits are provided in Table VI for all the three countries, which again illustrate the significantly improved goodness of the RO modeling over the usual Pareto law. However, interestingly, we now have the estimated $b$ values in the fitted RO distribution as negative compared to its positive values for India, China, Brazil or USA, which lead to significantly higher entropy and help us to discriminate between these two groups of countries in terms of their city size distributions; this higher entropy is because of the more uniform city sizes in these European countries. We will elaborate further on this discrimination in Section VII.

**G. Distribution of Australian Agglomerations**

We then focus our attention on the 6th largest nation in the world, with very low population density; we test the proposed RO modeling to characterize the distributions of Urban Agglomeration (UA) of Australia. Australia has one of the most urbanised societies in the world. As of 2018, Australia has an estimated population of 24.77 million, up from the official 2011 census results of 21.5 million. Australia is the 54th largest country in the world in terms of population and the most populous country in Oceania, three times more populous than its neighbor Papua New Guinea (8.2 million) and 5 times more populous than New Zealand (4.5 million).

As an illustration, let us consider the data on the sizes of 101 UA at the years 2011 and 2016, obtained from \[37\]. When these Australian UA sizes are modeled by the proposed RO and the usual Pareto distributions, as
before, the RO modeling gives significantly improved fit compared to the Pareto fit; see Table VII. All parameter estimates are surprisingly stable in both the year indicating the stability in the urban dynamics of Australia.

| Year | $N$  | $\hat{a}$ | $\hat{b}$ | $\nu$  | KS  | $\nu$  | KS  |
|------|------|----------|----------|-------|-----|-------|-----|
| 2011 | 101  | 1.259    | 0.418    | 0.104 | 1.420 | 0.158 |
| 2016 | 101  | 1.267    | 0.434    | 0.106 | 1.434 | 0.163 |

H. World-wide Size distributions of Cities and Countries in the year 2018

Currently, Tokyo is the world’s largest city with an agglomeration of 37 million inhabitants, followed by New Delhi with 29 million, Shanghai with 26 million, and Mexico City and So Paulo, each with around 22 million inhabitants. Today, Cairo, Mumbai, Beijing and Dhaka all have close to 20 million inhabitants. By 2020, Tokyo’s population is projected to begin to decline, while Delhi is projected to continue growing and to become the most populous city in the world around 2028. By 2030, the world is projected to have 43 megacities with more than 10 million inhabitants, most of them in developing regions. However, some of the fastest-growing urban agglomerations are cities with fewer than 1 million inhabitants, many of them located in Asia and Africa. While one in eight people live in 33 megacities worldwide, close to half of the world’s urban dwellers reside in much smaller settlements with fewer than 500,000 inhabitants. As the world continues to urbanize, sustainable development depends increasingly on the successful management of urban growth, especially in low-income and lower-middle-income countries where the pace of urbanization is projected to be the fastest. Understanding the key trends in urbanization likely to unfold over the coming years is very crucial and this motivated us to focus our analysis for the size of the top 20 cities of 233 countries around the world for the year 2018 [35]. As before, we have restricted our modeling to the cities with a minimum number of 5000 habitants, but still the filtered data of 3371 cities have a very wider spectrum of population varying from 5017 to 37,468,302. These are the estimated data projected from the actual census data collected from the data of previous years that is different for different countries. The table shows that in the rank wise list of top cities around the world, though Tokyo tops the list, but as expected there are many cities from India, China and USA.

We combine the population of the above mentioned 3371 cities across the world and apply the proposed RO modeling along with the usual Pareto modeling. Since there are wide varieties of city sizes, as expected the Pareto modeling fails miserably (KS measure 7.474; $\nu = 1.504$). But, the universal RO modeling again gives extremely better fit of the distribution of world cities with an KS error measure of only 0.04 (Figure 7a); the parameters of the fitted RO distribution are $a = 0.632$ and $b = 1.963$. Note that, here we have a significantly larger values of $b$ indicating wider spectrum of the city sizes.

![FIG. 7: Plot of the actual and the predicted sizes over ranks for Worldwide city sizes (estimated) in 2018 (Black square: observed sizes, blue circle: RO fits, red star: Pareto fits)](image)

Due to the inability of the existing Pareto distribution to model the low-end cities (in sizes), a commonly used alternative had been to model the cities with sizes larger than a specific limit. For example, next, we consider all cities around the world having populations of at least 3 lakh in the year 2018 and these 171 cities are again modelled by the RO and Pareto distributions. As expected, Pareto modeling gives a reasonable fit to these high-end city sizes (KS=0.703, $\nu = 1.178$) but the proposed RO modeling yields even a better fit than Pareto with a KS value of 0.041 ($\hat{a} = 0.426$, $\hat{b} = 1.302$); see Figure 7b for the corresponding fitted sizes with their original population. This further supports our claim on the universality of the RO distribution in modeling any kind of city size distributions.

![FIG. 7: Plot of the actual and the predicted sizes over ranks for World-wide size distributions of Cities and Countries in the year 2018 (Black square: observed sizes, blue circle: RO fits, red star: Pareto fits)](image)

Finally, we end our illustrations with a remark on the modeling of the country sizes. We note that while analysing the populations of 226 major countries across the world as per the 2018 estimate [35], as expected the Pareto modeling again fails tremendously (KS=14.091, $\nu = 2.257$) due to their wider spectrum. On the other hand, the RO modeling perform much well in this case as well, leading to the KS error of only 0.061 ($\hat{a} = 0.976$, $\hat{b} = 1.499$). This result justifies further the universal nature of the RO distribution. Probably, the countries around the world have a similar dynamics as the city dynamics within a country and we should be able to model them as well with a city size law.
V. ENTROPY OF THE DISTRIBUTION:
DISCRIMINATION, EVOLUTION AND THEIR
FUNDAMENTAL DRIVERS

An important usefulness of the proposed RO modeling is its ability to characterize the city size distribution through the parameters \((a, b)\). It is noted in Sec. II.A that the two parameters \(a\) and \(b\) completely define the RO distribution and hence the characteristics of the city sizes along with the entropy of the underlying process. Therefore, the estimated values of the parameters of the RO distributions fitted to city-sizes of different countries enable us to characterize and discriminate the corresponding city-size distributions and hence the underlying dynamics and its entropy (in consultation with Fig. 3): the changes seen in these estimated parameters at different time points for any fixed country also indicate the evolution of the corresponding process within that country.

For illustration, we have plotted the estimated values of \((a, b)\) for all our illustrative examples with 11 countries across the seven continents at different years in Figure 8, which provides a vividly clear comparison of their city-size distributions and their evolution. The heat-map of the entropy of the RO distribution with the corresponding ranges of the parameters \((a, b)\) are also presented in Figure 8 for the entropic analysis of the countries’ fitted city size-distributions.

One can clearly see from the figures that the large countries like India, China, USA, Australia or Brazil have positive \(b\) and hence are completely different from the smaller but developed countries like Italy, Sweden or Switzerland having negative values of \(b\). Thus, the first group of countries clearly have much lower entropy and hence lower chaos in their city size distribution compared to the second group; the spectrum of city sizes within the first group is much wider than those in the second group.

Further, within the first group of countries, China has a different city size distribution compared to India and US which are more similar: China has quite high entropy as well. In terms of evolution, both India and China have a drastic change in its urban dynamics between the twentieth and twenty-first centuries, which is clear from the difference in their \((a, b)\) estimates. But, in twenty-first century, both China and USA become quite stable in terms of their city size distributions, but that of India is still changing significantly from the year 2001 to 2011. The city-dynamics of Brazil, on the other hand, was in the same entropy range over the years but structurally becoming more closer to the India-USA cluster after an abrupt change in the last decades of twentieth century.

Next, noting that the case \(b = 0\) coincides with the Power law and the entropy formula of RO distribution, it is evident that the cases with \(b\) negative implies further higher entropy of the underlying distribution. Hence, the city sizes of European countries like Italy, Sweden and Switzerland has much smaller distribution spectrum and lower inequality. Among them, the city dynamics of Italy and Switzerland are pretty similar and have the possible highest entropy indicating their equal distribution of city sizes. Italian cities also have more stable size distributions over time compared to the Switzerland where entropy is further increasing over the year. Sweden, on the other hand, has quite a different city size distribution with comparatively lower entropy; this is because of the few large cities in Sweden. This makes their city size distribution having entropy close to that of USA but structurally very different from them.

Clearly, the African countries are slightly in between these two groups, with Sudan having significantly lower entropy and randomness in its city size distribution compared to Algeria or Uganda. However, the distribution of Australian Agglomeration is completely different from all these countries and have the lowest possible entropy among all; this is because of the wide range of the spectrum of Australian human settlements.

This exhaustive study helps us to conclude that the proposed RO distribution, along with its entropy formulation, can characterize the evolution of the urbanization and discriminate between the city size distributions of different countries. All the above characterization and discrimination between the city size distributions of different countries are made possible with greater clarity through the RO modeling, which was not possible through the Pareto modeling. Further, one can also use these estimated values of \((a, b)\) and link them with the associated factors of the countries to further find out the drivers of the city size growth. Such an analysis will further justify the discrimination between different countries’ underlying city dynamics and help us to control whenever needed. As a follow-up study of this global pattern for the city sizes, including the small and medium sized urban settlements, we intend to do some more in-depth driver analysis in future.

VI. CONCLUSIONS AND DISCUSSIONS

There are many competing criteria in the study of cities. This has made the science of city planning such a challenging one. Despite the enormous complexity and diversity of human behavior and extraordinary geographic variability, we have shown that size distribution of cities around the world follows a universal law. A two-parameter distribution, with the parameter values varying within a very short range, can illustrate the empirical data for the city sizes across the world. Moreover, the entropy analysis presented here is used to quantify the variation in the distribution in city sizes across the world as well as city evolutions. Wider spread of city sizes in a country having few extremely populous cities along with several smaller cities can be characterized with low entropy of the underlying city-size distribution of the country; this is possibly further linked to the lesser uncertainty in such country’s human settlements focusing more on clearly larger cities (e.g., Swedish cities, Australian UA). On the other hand, the uncertainty and in-
ternal movement within cities are expected to be higher for countries with almost equal city sizes or with lower spread of city sizes; they further leads to the comparatively higher entropy of the underlying RO distribution (e.g., Switzerland, Italy or China).

Sustainable urbanization is key to successful development and as such managing urban areas has become one of the most important development challenges of the 21st century. The small and mid-sized cities require the same attention, as the large cities of any country, for successful sustainability planning. We believe that the proposed universal rank-order modeling for the cities sizes (comprising the whole data set, starting from small number of inhabitants to the most populous city in the country) around the world will open up new avenues of research in this area. Information resulting from our analysis will be vital for setting policy priorities to promote inclusive, equitable and sustainable development for urban and rural areas alike. Till today, urban size for small/mid-sized cities has been a neglected dimension to sustainability and livability planning. A focus on smaller cities might offer an especially productive diversification of the sustainability planning. We believe that the proposed RO distribution. These two parameters can be thought of as two fundamental latent drivers of the city dynamics which could be further associated with observable economic, social or environmental factors; these observables can then be controlled for any suitable planning of the city-sizes in a country as per the requirements. Furthermore, these two parameters, the fundamental drivers of city populations, can be linked with the underlying entropy and the corresponding maximum entropy principle, which might shed further lights on their physical significance. Such an analysis may be explored with the inverse problem of characterizing the proposed RO distribution as a maximum entropy distribution under appropriate constraints. We hope to pursue some of these interesting and useful extensions of our work in future.

Acknowledgments

The work of the first author, AG, is supported by the INSPIRE Faculty Research Grant from the Department of Science and Technology, Government of India, India.
FIG. 8: (Top) Estimated parameters \((a, b)\) of the fitted RO distribution for different countries and different years; (Bottom) Entropy of the RO distribution at different values of the parameters \((a, b)\).

[33] http://www.citypopulation.de/Sudan, http://www.citypopulation.de/Uganda, http://www.citypopulation.de/Algeria, http://www.citypopulation.de/Italy, http://www.citypopulation.de/Sweden, http://www.citypopulation.de/Switzerland, http://www.citypopulation.de/Australia.html, https://www.citypopulation.de/Australia-Agglo.html, https://www.citypopulation.de/Australia-AggloEst.html, http://worldpopulationreview.com/countries/australia-population, http://worldpopulationreview.com/countries/italy-population, http://worldpopulationreview.com/countries/sudan-population, http://worldpopulationreview.com/countries/sweden-population, http://worldpopulationreview.com/countries/switzerland-population, http://worldpopulationreview.com/countries/uganda-population, http://worldpopulationreview.com/world-cities, http://worldpopulationreview.com/us-cities, African Economic Outlook 2017, African Development Bank.