

The Urban Wage Premium: Sorting, agglomeration economies or statistical artifact? The problem of sampling from lognormals

Andrés Gómez-Liévano^{a,*}, Vladislav Vysotsky^{c,d}, José Lobo^b

^a*Center for International Development, Harvard University, Cambridge, USA*

^b*School of Sustainability, Arizona State University, Tempe, USA*

^c*Department of Mathematics, University of Sussex, Brighton, UK*

^d*St. Petersburg Division of Steklov Mathematical Institute, St. Petersburg, Russia*

Abstract

Economic explanations for the urban wage premium (UWP) fall in two categories: sorting of more productive individuals into larger cities, or agglomeration externalities. We present a third hitherto neglected mechanism: a statistical artifact arising from sampling lognormally distributed wages. We show how this artificial UWP emerges, systematically and predictably, when the variance of log-wages is larger than twice the log-size of workers sampled in the smallest city. We present an analytic derivation of this connection between lognormals and increasing returns to scale using extreme value theory. We validate our results analyzing simulated data and real data on more than six million real Colombian wages across more than five hundred municipalities. We find that when taking random samples of 1%, or less, of all Colombian workers the estimated real and artificial UWP are both 7%, and become statistically indistinguishable, yet both significantly larger than zero. This highlights the importance of working with large samples of workers. We propose a method to tell whether an estimate of UWP is real or an artifact.

Keywords: urban wage premium, increasing returns to scale, lognormal distribution, heavy-tailed distributions, law of large numbers.

JEL: B23, P25, R12, C46.

*Corresponding author

Email addresses: `Andres_Gomez@hks.harvard.edu` (Andrés Gómez-Liévano),
`V.Vysotskiy@sussex.ac.uk` (Vladislav Vysotsky), `Jose.Lobo@asu.edu` (José Lobo)

1. Introduction

There is a substantial body of theoretical and empirical research on the origins of “Increasing Returns to Scale” in cities (Sveikauskas, 1975; Rosenthal and Strange, 2004; Melo et al., 2009; Behrens et al., 2014; Combes and Gobillon, 2015). This urban productivity premium, whereby workers tend to be more productive in larger cities, can be estimated using wages (Glaeser and Mare, 2001; Combes and Gobillon, 2015) and is thus referred to as the *urban wage premium* (from here on abbreviated as UWP). Similarly, there is an extensive literature on the disparities in productivity across individuals showing that earnings follow a lognormal distribution (Roy, 1950; Aitchison and Brown, 1957; Mincer, 1970; Kleiber and Kotz, 2003; Combes et al., 2012; Eeckhout et al., 2014). No work, however, has demonstrated the link between the lognormality of individual productivity and the UWP. This is the purpose of the present work, and our contributions are, first, to derive and demonstrate using extreme value theory that an artificial UWP can emerge from sampling lognormal random variables under certain conditions, and second, to propose a method to tell apart real from artificial UWP in real data.

To be precise, when total output in a city is a function of its size, $Y = F(n)$, then UWP refers to the situation in which

$$F(\lambda n) > \lambda F(n), \tag{1}$$

for any number $\lambda > 1$. The scale n typically represents the size of the labor force contributing to the total production Y . If $X = f(n) = F(n)/n$ is the productivity per individual, then eq. (1) states that $\lambda n f(\lambda n) > \lambda n f(n)$, which implies

$$f(\lambda n) > f(n).$$

In other words, the UWP literally implies that larger population sizes are associated with more productive individuals. In cities, the processes invoked to explain why larger scales are associated with higher productivity are taken to be one, or a combination, of two general mechanisms (Andersson et al., 2014): productive individuals sorting themselves into larger teams, or larger teams generating more productive individuals.¹ We will refer to these two

¹There may also be selection effects that eliminate the least productive firms.

mechanisms simply as sorting and agglomeration effects, respectively. These effects come from specific economic processes which entail either decisions by, or interactions among, economic agents that, if absent, UWP would also be absent.

We argue in this paper that one can estimate an effect in data that can be mistaken for an UWP in the absence of these mechanisms. This artificial UWP emerges in a systematic and predictable way, but is a statistical artifact. We show when and why this happens, and we propose a method to tell apart real versus artificial UWP based on a simple intuition: that randomization of individuals across cities should eliminate the economic effects but not the artificial one. Crucially, however, there are situations in which not even randomization will remove the artificial effect.

Randomizing individuals across groups (making sure of maintaining constant the groups' sizes) will destroy the information of the way individuals have sorted themselves across groups, and of whom the workers have interacted, or are interacting, with. Hence, if the UWP is real (i.e., has an underlying economic logic), the act of randomizing individuals should eliminate any UWP from our estimations because it will eliminate the built-in dependencies of individuals caused by sorting or agglomeration effects. In contrast, however, an artificial UWP should be statistically invariant to the removal of the causal effects present in the data.

But how can UWP systematically emerge without sorting or agglomeration mechanisms? In other words, how can UWP emerge if wages are independently and identically distributed across cities? We will show this to be the consequence of extreme values of productivity contributing significantly to the total output of the city, which is one of the consequences of productivities being lognormally distributed. Our analytic results are validated on simulated data, as well as on an administrative dataset of all formal workers in Colombia.

To develop the intuition for why individual productivities that are lognormally distributed can give rise to UWP not caused by any sorting or agglomeration effects, it is illustrative to think about a related question. Assume two groups of individuals of different sizes, n_1 and $n_2 = \lambda n_1$, with $\lambda > 1$. In addition, assume we observe a slight UWP, whereby the output generated by the second (larger) group is disproportionately larger than the output of the first (smaller) group. In terms of the average productivity of both groups, we have that $\bar{x}_2 > \bar{x}_1$. Is it more likely that several individuals in the second (larger) group are

slightly more productive than in the first (smaller) group? Or is it more likely that only a few individuals in the larger group are significantly more productive?

To answer such question, we note that the first possibility emphasizes that many individuals contribute to the increase in output, and this has a consequence on its likelihood. The likelihood of this possibility is the probability that individual 1 in the second group is slightly more productive than expected (i.e., slightly larger than \bar{x}_1) *and* individual 2 in the second group is slightly more productive than expected *and* individual 3 in the second group is slightly more productive than expected *and* so on, for several individuals. Hence, the first possibility can be described as a *conjunction of many probable events*. On the other hand, in the second possibility only few individuals are the reason for the increase in output. In this case, the likelihood of this second possibility is the probability that individual 1 in the second group is much more productive than expected *or* individual 2 in the second group is much more productive than expected *or* individual 3 in the second group is much more productive than expected *or* so on, for several individuals. Thus, the second possibility can be described as a *disjunction of improbable events*.

On the one hand, being slightly more productive than expected is easy (i.e., it is a probable event), but in the first possibility we are requiring that all individuals are so. On the other, being an extremely productive individual is a very improbable event, but in the second possibility any individual has a chance. Thus, one can ask when is the disjunction more likely than the conjunction?

These two possibilities (conjunction versus disjunction) are two extreme cases from a spectrum of possibilities. But for the sake of our argument it is useful to realize that sorting and agglomeration effects, the two prevalent explanations in the literature for UWP, are explanations that presume in general a conjunction of events. Sorting and agglomeration, respectively, state that individuals that were *more productive before joining* any group had a preference for larger groups and thus decided to join the second group, or that some positive externalities promoting larger productivities in larger teams *made* individuals in the second group more productive due to the team's comparatively larger size. These two explanations describe a conjunction of probable events because they focus on the contribution of many individuals as opposed to the contribution of some few.

Let us put these two extreme possibilities in mathematical terms. Assuming independent and identically distributed productivities, denoted in general with the letter X , and assuming the difference in average productivities between the groups is $\bar{x}_2 - \bar{x}_1 > \epsilon$, the likelihood of the conjunction is $\Pr(X > \bar{x}_1 + \epsilon)^{n_2}$ while the likelihood of the disjunction is $n_2 \Pr(X > \bar{x}_1 + n_2 \epsilon)$.² If we denote $S(x) = \Pr(X > x)$, and with some minor rearranging, the disjunction is more likely than the conjunction when

$$S(\bar{x}_1 + n_2 \epsilon) > \frac{e^{-n_2 \ln(1/S(\bar{x}_1 + \epsilon))}}{n_2}.$$

From this relation it should be clear that increasing the sample size n_2 will decrease the terms on both sides of the inequality. However, each side of the inequality may fall at different rates. The inequality will hold depending on a balance between the sample size n_2 and how rapidly the tail of $S(x)$ falls.

While these are two extreme situations (e.g., by far not all terms in the conjunction are required to exceed \bar{x}_1 by ϵ), they serve us to build the intuition for why disjunctions can be more probable than conjunctions, or viceversa, depending on (i) the shape of the underlying probability distribution function describing productivities and (ii) the sizes of the samples being analyzed. Even more, the conjunction and the disjunction examples provide a mathematically correct explanation of large deviations probabilities which describe highly unlikely events (e.g., that the average productivity attains an atypically large value above the expected one).³ Of particular relevance for work on cities, we show that if productivities are

²The latter comes from assuming that all individuals in the second group have on average a productivity equal to \bar{x}_1 , except for a single individual that has a large productivity M . The total output in the second group is thus $n_2 \bar{x}_2 = (n_2 - 1)\bar{x}_1 + M$. Or, in other words, the contribution of the highly productive individual is $M = \bar{x}_1 + n_2(\bar{x}_2 - \bar{x}_1)$. If $\bar{x}_2 - \bar{x}_1 > \epsilon$, one requires $M > \bar{x}_1 + n_2 \epsilon$. The probability that the maximum among n random variables is larger than a given x is one minus the probability that the maximum is less or equal than x , which is the probability that all n variables are less or equal than x . Hence, $\Pr(M > \bar{x}_1 + n_2 \epsilon) = 1 - \Pr(M \leq \bar{x}_1 + n_2 \epsilon) = 1 - \Pr(X \leq \bar{x}_1 + n_2 \epsilon)^{n_2}$. For large n_2 , this simplifies into $\Pr(M > \bar{x}_1 + n_2 \epsilon) = 1 - (1 - \Pr(X > \bar{x}_1 + n_2 \epsilon))^{n_2} \approx n_2 \Pr(X > \bar{x}_1 + n_2 \epsilon)$, which is the expression in the main text.

³Conjunctions correspond to light-tailed cases, where an atypically large value of the sum is attained by modifying the vast majority of the terms, while disjunctions correspond to the so-called principle of single big jump valid for heavy-tailed cases.

lognormally distributed with a very large variance, then the disjunction of improbable events is the most likely explanation of the UWP. We show this to be the case when $n_{\min} < e^{\sigma^2/2}$, where n_{\min} is the sample size of workers in the smallest city being considered in the analysis and σ^2 is the variance of log-productivity. The crucial implication is that in this regime of large variances and/or small sample sizes, a large total output is likely to be the result of a large individual contribution rather than the collective result of many small contributions. The statistical origin of the UWP emerges because the maximum among n lognormals grows disproportionately quickly with n , which implies that measuring sample averages across groups of different sizes n can actually track how the maximum grows with n , and thus generate a statistical pattern that can be mistaken for evidence of UWP. We do not claim this is *the* explanation of UWP in the real world. We merely want to raise awareness about its plausibility which, as far as we know, has not been considered in the literature.

The discussion is organized as follows. In the next section we provide a brief overview of the literature on the microfoundations of production functions with an emphasis on the connection with extreme value theory in statistics. Section 3 derives the main analytical results from extreme value theory. Numerical simulations are presented in Section 4 and a real-world application is treated in Section 5 where we analyze wages in Colombian municipalities. Section 6 concludes.

2. Background

Few studies have addressed the relationship between probability distributions and production functions exhibiting increasing returns to scale (IRS). One of the earliest attempts at connecting a probability distribution with a production function was made by Houthakker (1955), who showed that a Cobb-Douglas production function arises when inputs of production are Pareto distributed. Houthakker's production function displays decreasing returns to scale (DRS), but his result is fine-tuned to the Pareto assumption regarding the inputs of production (see also Levhari, 1968). Jones (2005) generalizes Houthakker's results by relaxing the shape of the local production function (e.g., of firms), and derives a global production function using results from extreme value theory. Jones' global production function has constant returns to scale (CRS). However, this property is actually inherited from the

local production functions of firms which may have decreasing, constant, or increasing returns to scale. Similar in spirit to Houthakker’s approach is Dupuy (2012)’s work. Dupuy builds on Rosen (1978), and derives, from a model that matches workers with different skill types and levels to different tasks with varying skill requirements in a market with perfect competition, a production function whose shape depends on the shape of productivity per worker and the density of tasks for different skill requirements. Dupuy’s general production function can accommodate the usual constant elasticity of substitution (CES) production function when tasks follow a Beta distribution. These works of Houthakker, Jones, Rosen and Dupuy are similar to one another in that the decreasing, constants, or increasing returns to scale of their production functions is assumed exogenously, in one way or another.

In a study similar to ours, Gabaix and Landier (2008) used results from extreme value theory to link the distribution of talent in CEOs with their wage. Gabaix and Landier find that, in equilibrium, top talents in firms get paid proportionally to firm size to a certain power (which can be larger than 1). Their result is in fact analogous to the one we present here, although they use Pareto distributions which are known for violating the Central Limit Theorem and the Law of Large Numbers. In our case, we assume a lognormal distribution whose moments are all finite, and therefore the increasing returns to scale that we derive are less trivial.

After Gabaix and Landier (2008)’s study, Gabaix (2011) focused on the fluctuations (as opposed to the levels) of output across firms and showed that if shocks to firm productivity have a heavy-tailed distribution, then shocks do not compensate each other in the aggregate. Crucially, this suggests that aggregate measures are composed of “granular” components (e.g., firms), in such a way that the largest components dominate the aggregate. Gabaix and Landier (2008) and Gabaix (2011) are both studies about the characterization of superstar-like phenomena, which is also our own focus of analysis. Our contribution is in part the further dissemination of these insights in the field of urban economics, by showing a very simple application in which we analyze superstar-like effects on the estimation of the urban productivity premium.

While our modeling framework relates to all the studies previously mentioned in both the general approach and motivation, we differ from them in two main ways. First, we abstract

away any market, equilibrium condition, or coordination mechanism among individuals, since our main claim is that aggregate increasing returns to scale are not necessarily a consequence of any sorting, coordination, interactions or positive externalities. Second, we move away from the frequent use of Pareto (or “power law”) distributions because, as we mentioned above, these can display anomalous behavior, such as undetermined mean and/or variance, that arise from the fact that high-order moments diverge. In contrast, we will derive our results from a model in which all individuals have productivities that are independently and identically sampled from the same distribution whose moments are all fixed and finite. Under this model, we show that total output shows increasing returns for a wide range of scales, a result that emerges purely from sampling effects.

The significance of our results lies not in challenging current models that generate the UWP (e.g., Duranton and Puga, 2004; Combes et al., 2008, 2012; Bettencourt, 2013; De la Roca and Puga, 2017; Gomez-Lievano et al., 2016), but in showing that a UWP can emerge from purely statistical reasons that differ from explanations based on economic mechanisms. Typically, when UWP are empirically inferred, the prevailing assumption is that an “egalitarian” situation applies in which most people living in larger cities are more productive. The prevailing explanation is thus implicitly assuming that what ought to be explained is the change in *average* productivity across different cities. Instead, we argue that what ought to be explained are the changes of the *statistical distributions of productivity* across groups, from which one can identify more clearly the origins of the UWP. As a corollary, per capita metrics of productivity may not be adequate estimates of the productivity of individuals because they can hide the distributional origins of UWP. Their use should thus depend on the underlying distribution of individual productivity as much as on the sample size of the system.

3. Analytic Results

Our analytic results are based on a very simple model whereby individuals, regardless of the city they live in, have productivities *independently* and *identically* distributed (i.i.d.),

sampled from a lognormal distribution $\mathcal{LN}(x_0, \sigma^2)$, whose probability density function is

$$p_X(x; x_0, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \ln x_0)^2}{2\sigma^2}}, \quad (2)$$

where x_0 and σ are positive parameters such that $\ln(x_0) = \mathbb{E}[\ln(X)]$ and $\sigma^2 = \text{Var}[\ln(X)]$. The expected productivity is thus $\mu \equiv \mathbb{E}[X] = x_0 e^{\sigma^2/2}$. We will use upper case letters to denote random variables, and lower case to denote realized values. Total output of a city with population size n will be the sum of the productivities of its inhabitants, $Y(n) = \sum_{i=1}^n X_i$. The intent is to understand the consequences of the fact that output in cities is the sum of heterogeneous contributions. Henceforth, we will assume that there are m cities, indexed as $k = 1, \dots, m$, each with total populations n_1, \dots, n_m .

The i.i.d. assumption about productivities is not adopted for mathematical convenience, but rather by explanatory intent as we want to demonstrate UWP in the absence of interactions between individuals, and in the absence of any structural, compositional, or natural advantages that cities may have. The choice of a lognormal distribution has two purposes. First, there is evidence that the empirical distributions of productivity (Combes et al., 2012) and wages (Eeckhout et al., 2014) for workers are well-fitted by lognormal distributions (see Appendix B for a simple justification for why lognormal productivities would emerge). Second, the lognormal has the properties that enable the emergence of UWP as an artifact: namely, the lognormal belongs to a class of heavy-tailed distributions called “subexponential distributions”. The role of subexponentiality will become evident below.

3.1. Elasticity for a single city

Let us proceed by calculating first the change in the expected value of urban output if population size is increased by $\lambda > 1$ according to this simple model:

$$\begin{aligned} \mathbb{E}[Y(\lambda n)] &= \mathbb{E}\left[\sum_{i=1}^{\lambda n} X_i\right], \\ &= \sum_{i=1}^{\lambda n} \mathbb{E}[X_i], \\ &= \lambda n \mathbb{E}[X_1], \\ &= \lambda \mathbb{E}[Y(n)]. \end{aligned} \quad (3)$$

In per capita terms,

$$\mathbb{E} \left[\frac{Y(\lambda n)}{\lambda n} \right] = \mathbb{E} \left[\frac{Y(n)}{n} \right].$$

From the point of view of expectation values, our model does not display UWP, and the expected per capita output is constant across cities. Specifically, the total expected production in our model is $\mathbb{E}[Y(n)] = Y_0 n^\beta$, with $\beta = 1$ and $Y_0 = \mu$. While the derivation of eq. (3) might seem trivial, what is not so obvious is the realization that $\mathbb{E}[Y(n)]$ is never observable, a fact with practical and measurable consequences when the distribution of X_i has certain properties. In what follows we go beyond relying on expectation values and study how the distribution of X_i determines whether $Y(n)$ may, or may not, display UWP.

Our approach, which draws on the probabilistic notion of “stable laws”, consists of finding sequences c_n and d_n such that the variable $c_n^{-1}(Y(n) - d_n)$ converges to a random variable that has a stable distribution. When we find such sequences, we can state that $Y(n)$ scales with n as d_n does. Hence, we will posit that elasticities can be computed as

$$\frac{d Y(n)/Y(n)}{d n/n} = \frac{d \ln(Y(n))}{d \ln(n)} \approx \frac{d \ln(d_n)}{d \ln(n)}. \quad (4)$$

When X_i are i.i.d. with finite mean and variance, and n is very large, the Central Limit Theorem states that if $d_n = \mathbb{E}[X_1] n$ and $c_n = (\text{Var}[X_1] n)^{1/2}$, the stable law to which $c_n^{-1}(Y(n) - d_n)$ converges to is the Standard Gaussian distribution. Thus, for sizes $n \rightarrow \infty$, the elasticity of total output with respect to size is

$$\begin{aligned} \beta &= \frac{d \ln(Y(n))}{d \ln(n)} \\ &\approx \frac{d \ln(\mu n)}{d \ln(n)} \\ &= 1. \end{aligned} \quad (5)$$

In words, a value $\beta = 1$ means that total output increases proportionally with sample size n , implying the absence of an UWP. This is a result that holds for very large n . The range of values for which n is “large” enough for eq. (5) to hold, however, is determined by the “evenness” or “unevenness” of the distribution of X_1 . By evenness or unevenness of a distribution we mean the extent to which the random variables tend to be highly dispersed, exhibiting extreme values. The formal distinction we will use is between “light-tailed” and “heavy-tailed” distributions, where the difference depends on whether the tails

of the distribution fall faster (light-tail), or slower (heavy-tail), than an exponential tail. When X_i are heavy-tail distributed, $Y(n)$ may not scale as $d_n = \mu n$ except in the limit of extremely large sizes.

In our model, X_i are “unevenly” distributed, and the distribution $p_X(x)$ is heavy-tailed. Specifically, X_i follow a lognormal distribution. Thus, we cannot use eq. (5) naively. Can we find a sequence d_n for heavy-tailed distributions in the regime of “small sizes” (as opposed to “in the limit of large sizes”)?

Lognormals belong to a family of distributions that satisfy the following property:

$$\lim_{t \rightarrow \infty} \frac{\Pr(\max\{X_1, \dots, X_n\} > t)}{\Pr(X_1 + \dots + X_n > t)} = 1, \quad \text{for all } n \geq 2. \quad (6)$$

Distributions that satisfy Equation (6) are called subexponential (Embrechts et al., 2013).

In words, the property of subexponential distributions in eq. (6), widely known as “the single big jump principle”, states that atypically large values of the sum of a fixed number of i.i.d. random terms are achieved by one maximal term that dominates the sum of the other terms. Although, on the contrary, here we are interested in *typical* values of the sum $Y(n)$ of large (increasing to infinity) number of terms n , we use the above property of subexponential distributions as an insight to approximate $Y(n)$ by $\max\{X_1, \dots, X_n\}$. Hence, eq. (6) will motivate our heuristic approach to tract analytically Equation (4), and derive a sequence d_n that we can use to characterize the sum $Y(n)$ of lognormal random variables. We develop the argument in the next subsection.

3.2. The maximum of lognormal random variables

The random variables representing the productivity of individuals in our model can be conveniently represented as $X_i = e^{\sigma Z_i + \ln x_0}$, where Z_1, Z_2, \dots are i.i.d. random variables sampled from the standard normal distribution $\mathcal{N}(0, 1)$. We will consider lognormal distributions with variable parameters $\ln x_0$ and σ related such that $E[X_1] = 1$, where the constant is chosen to be 1 merely for the purpose of convenience. Consequently, $\ln x_0 = -\sigma^2/2$.

The classical Chebyshev inequality states that

$$\Pr(|Y(n) - nE[X_1]| > a) \leq \frac{\text{Var}[X_1]n}{a^2}, \quad a > 0, n \geq 1.$$

This convinces, by taking a to be of order n , that $Y(n)$ grows approximately linearly in n with slope $E[X_1] = 1$ once n becomes large enough so that $n \gg \text{Var}[X_1] = e^{\sigma^2} - 1$. Hence the linear growth assumption (the LLN) can only be violated for cities of population n of order at most e^{σ^2} (or, equivalently, σ must be at least of order $\sqrt{\ln(n)}$). Therefore, from this point on, we assume that σ is large but fixed, and consider the total productivity $Y(n)$ of cities of “small” to “moderate” size n satisfying a slightly stronger constraint $\sqrt{2\ln(n)} \ll \sigma$.

Our idea is to approximate the total productivity $Y(n)$ by the maximal productivity $M(n) := \max\{X_1, \dots, X_n\}$ of individuals in the city. This quantity can be written as $M(n) = e^{\sigma L(n) - \sigma^2/2}$, where $L(n) := \max\{Z_1, \dots, Z_n\}$ denotes the maximum of i.i.d. standard normal random variables. Then

$$\begin{aligned} Y(n) &= \sum_{i=1}^n X_i = \sum_{i=1}^n e^{\sigma Z_i - \sigma^2/2} = e^{\sigma L(n) - \sigma^2/2} \sum_{i=1}^n e^{\sigma(Z_i - L(n))}, \\ &= M(n) \sum_{i=1}^n e^{\sigma(Z_i - L(n))}. \end{aligned} \tag{7}$$

The behavior of $L(n)$ (and hence of $M(n)$) for large n is well-known (Leadbetter et al., 1983; Embrechts et al., 2013): this quantity grows as $\sqrt{2\ln(n)}$ (to be more precise, as $\sqrt{2\ln(n)} - \frac{\ln(\ln(n))}{\sqrt{8\ln(n)}}$) with random fluctuations of order $(\ln(n))^{-1/2}$. Namely,

$$\lim_{n \rightarrow \infty} \Pr\left(L(n) \leq \sqrt{2\ln(n)} + \frac{2x - \ln(\ln(n)) - \ln(4\pi)}{\sqrt{8\ln(n)}}\right) = e^{-e^{-x}}, \quad x \in \mathbb{R},$$

where the right-hand side is the standard Gumbel distribution function. Concretely, in our model, $c_n^{-1}(M(n) - d_n)$ tends to a standard Gumbel, and the sequence that tells us how the maximum scales with size is approximately $d_n \approx \exp\{-\sigma^2/2 + \sigma\sqrt{2\ln(n)}\}$.

The main difficulty for validating the assumption that $Y(n)$ can be approximated by $M(n)$ is in analyzing the last sum in Equation (7). Since it is doubtful that this quantity can be tackled analytically, we suggest the following argument. First write

$$\Delta_n := \sum_{i=1}^n e^{\sigma(Z_i - L(n))} = \sum_{i=1}^n e^{\sigma(L_i(n) - L(n))},$$

where we have re-ordered the terms in the summation such that $L_i(n)$ denotes the i th largest value among Z_1, \dots, Z_n . For the first term, we have $L_1(n) = L(n)$, so $e^{\sigma(L_1(n) - L(n))} = 1$. For the second term, we can use that $L(n) - L_2(n)$ is of order $(\ln(n))^{-1/2}$ (see Leadbetter et al.,

1983, Section 2.3). By our assumption that $\sigma \gg \sqrt{2 \ln(n)}$, the quantity $\sigma(L_2(n) - L(n))$ is negatively large and so $e^{\sigma(L_2(n) - L(n))}$ is close to 0. The remaining terms $e^{\sigma(L_i(n) - L(n))}$ for $i \geq 3$ decay to 0 much faster since so do exponents with larger negative powers.

Thus, we have $\Delta_n \approx 1$ when $\sigma \gg \sqrt{\ln(n)}$. Moreover, our simulations (not shown) reveal that a similar conclusion applies even when σ is larger than, but comparable to, $\sqrt{2 \ln(n)}$, in which case Δ_n is rather close to 1, being of constant order.

Putting everything together and using that fluctuations of the quantity $L(n) - \sqrt{2 \ln(n)}$ vanish for $n \gg 1$, we arrive at the following conclusion. For any fixed σ large enough and any n sufficiently large (so that $L(n) \approx \sqrt{2 \ln(n)}$, but still of order at most e^{σ^2}), we have

$$\begin{aligned} \beta &= \frac{d \ln(Y(n))}{d \ln(n)}, \\ &\approx \frac{d \ln(M(n))}{d \ln(n)}, \\ &\approx \frac{d \ln(d_n)}{d \ln(n)}, \\ &\approx \frac{d \left(-\sigma^2/2 + \sigma \sqrt{2 \ln(n)} \right)}{d \ln(n)}, \end{aligned} \tag{8}$$

which yields

$$\beta(n, \sigma) \approx \frac{\sigma}{\sqrt{2 \ln(n)}}. \tag{9}$$

As we will show, this result derived from the heuristic that $Y(n) \approx M(n)$ is well-supported by simulations.

Figure 1 illustrates the fact that the maximum can indeed become comparable to the sum. We use a proxy of the share $M(n)/Y(n)$ as the ratio of the quantile $Q(1 - 1/n)$ over $\sum_i Q(i/n - 1/n)$, where $Q(\Pr(X \leq x_p)) = x_p$. The figure shows the curves for this proxy of the share $M(n)/Y(n)$ as a function of σ , for four distinct values of n . According to our derivations, for $\sigma = 4$, the maximum is comparable to the sum when $n < e^{\sigma^2/2} \approx 3,000$. Indeed, the figure shows that for $\sigma = 4$ and for $n = 10^3$ (darkest blue line), the maximum can account for 50% of the sum. For $\sigma = 4$ one needs to increase size to $n = 10^7$ (a ten-thousand-fold increase) in order to decrease the dominance of the maximum to about 10% (see lightest blue line). The fact that as population size increases (i.e., as the color of the

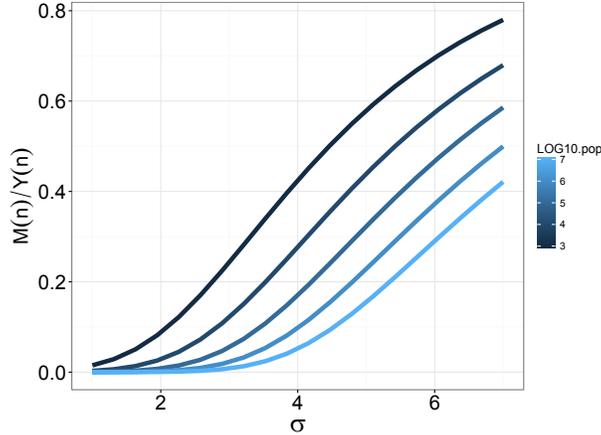


Figure 1: Proxy for the share of the maximum over the total sum, $M(n)/Y(n)$, constructed as the ratio of the quantile of the lognormal associated with the percentile $(n-1)/n$ over the sum of all the n quantiles, as a function of parameter σ . The color of the line represents a fixed population size. We show the curves for $n = 10^3, 10^4, 10^5, 10^6$, and 10^7 . The lighter the blue is, the larger the population is.

lines become lighter) the dominance of the maximum value over the sum decreases is an effect of the law of large numbers. Thus, although eq. (5) holds for very large sizes, fig. 1 provides support to replacing the sum $Y(n)$ with the maximum $M(n)$, and using this to find an approximate result for how the sum scales with size.

Equation (9) provides us with the null expectation of the local elasticity of total output with population size from a purely statistical effect when the distribution of productivities is lognormal, in the neighborhood of a specific sample of size n . If we set $\beta = 1$ (i.e., constant returns to scale), this equation expresses a sort of boundary between the heavy-tailness of the distribution as parametrized by σ and the size of the sample n . When n is above this boundary, the LLN applies, and no increasing returns to scale should be observed. If the distribution of productivities becomes more heavy-tailed (i.e., more unequal), but one wants the LLN to hold, one must increase sample size. Importantly, notice that small increments in σ must be counteracted by very large increases in n . Notwithstanding the fact that eq. (9) is just an approximation, we plot in Figure 2 the boundary that would separate the combination of σ and n values for which one would expect elasticities above one from the combination of values for which we would expect elasticities of one. In the next section we derive the average elasticity across many cities with sizes Pareto-distributed.

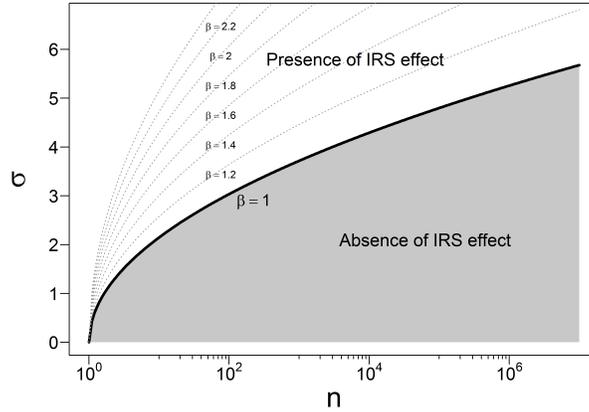


Figure 2: Balance between parameter σ and sample size n . The black solid line is the curve given by eq. (8) when $\beta = 1$. Above the black solid line is the region where sizes n are not large enough to tame the large fluctuations originated by such large values of σ . We assume this behavior gets reflected in the behavior of the sum of n lognormal random variables. Thus, it is the regions where we expect the appearance of UWP purely from a sampling effect. The dotted lines highlight regions where we anticipate increasingly higher elasticities.

3.3. Elasticity from a cross-sectional regression of many cities

The elasticity of the urban productivity premium effect is a relative rate of change of total output with size. This rate may be different for different sizes. In regression analysis, however, we often estimate empirically an average elasticity across many sizes. Assuming the artificial UWP described in the previous section is present in data, what would be the elasticity coefficient we would estimate from a regression line across many cities?

In order to account for all possible elasticities generated by both eqs. (5) and (9) (i.e., the constant returns to scale guaranteed by the law of large numbers for large sizes and small variances *and* the increasing returns to scale generated by the maximum for small sizes and large variances, respectively), we define the piecewise function of total output

$$E[\ln(Y(N)) | \ln(N)] = \begin{cases} \ln(N) & , \quad \text{if } \ln(N) \geq \frac{\sigma^2}{2} \\ -\frac{\sigma^2}{2} + \sigma\sqrt{2\ln(N)} & , \quad \text{if } \ln(N) < \frac{\sigma^2}{2}, \end{cases} \quad (10)$$

where population sizes are now represented by a random variable N . In a simple linear model $f(X) = a + bX$ to fit $E[Y|X]$, the coefficient b of the relation can be re-expressed as the ratio $\text{Cov}[X, Y] / \text{Var}[X]$. In our case, a simple linear regression to estimate the elasticity of

total output with respect population size would require estimating the ratio

$$\text{Cov} [\ln(N), \ln(Y(N))] / \text{Var} [\ln(N)].$$

Let us assume sizes are Pareto distributed, such that the probability density function of sizes is

$$p_N(n; n_{\min}, \alpha) = \frac{\alpha}{n_{\min}} \left(\frac{n}{n_{\min}} \right)^{-\alpha-1}. \quad (11)$$

Small values of α imply city sizes that are very unequal. For example, for $\alpha \leq 2$ variance is infinite, and for $\alpha \leq 1$, both the variance and the expected mean are infinite. When $\alpha = 1$, the distribution is often referred to as ‘‘Zipf’s Law’’. The parameter n_{\min} determines the minimum value above which sizes follow a Pareto distribution.

The regression coefficient can be re-written as

$$\beta_{ave}(n_{\min}, \sigma, \alpha) = \frac{\text{E} [\ln(N) \ln(Y(N))] - \text{E} [\ln(N)] \text{E} [\ln(Y(N))]}{\text{Var} [\ln(N)]}.$$

Using eq. (10) and computing expectation values with eq. (11), we get

$$\beta_{ave}(n_{\min}, \sigma, \alpha) = \begin{cases} 1 & , \text{ for } n_{\min} \geq e^{\sigma^2/2}, \\ \frac{\sigma n_{\min}^{\alpha} \sqrt{2\pi\alpha}(1 - 2\alpha \ln(n_{\min}))}{4} \left[\text{erf} \left(\sqrt{\frac{\alpha\sigma^2}{2}} \right) - \text{erf} \left(\sqrt{\alpha \ln(n_{\min})} \right) \right] + \\ \frac{\alpha\sigma\sqrt{2\ln(n_{\min})}}{2} + n_{\min}^{\alpha} e^{-\alpha\sigma^2/2} (1 - \alpha \ln(n_{\min})) & , \text{ otherwise.} \end{cases} \quad (12)$$

Equation (12) represents our main analytic contribution⁴. Given the values of the three distributional parameters, it provides the null expectation for the urban productivity premium from a cross-section of cities, under the assumption of i.i.d. productivities across individuals and cities. Notice that $\beta_{ave}(n_{\min}, \sigma, \alpha)$ is a function of the parameters of the distributions *only*. The fact that the function $\beta_{ave}(n_{\min}, \sigma, \alpha)$ is piecewise arises from the fact

⁴Equation (12) can be derived manually, but it is a relatively long computation. For the actual computation, and to guarantee a simple representation, we used the Mathematica software.

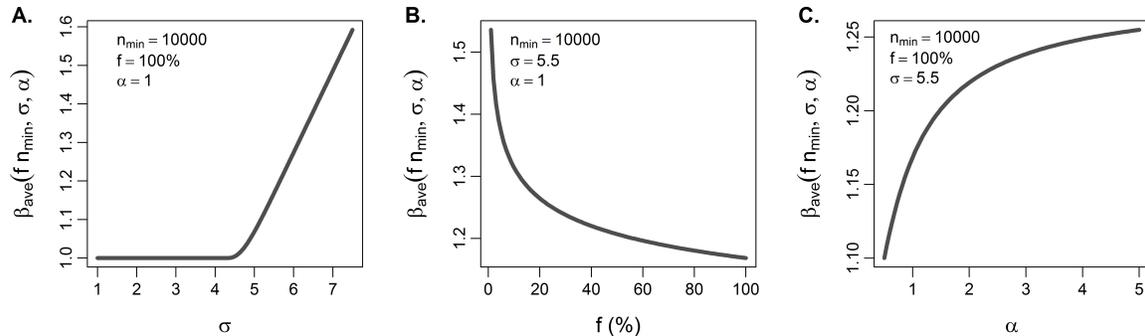


Figure 3: Predicted elasticity of total urban output with respect to city size, generated from a statistical artifact, as a function of three distributional parameters of individual productivities and city sizes. Panel **A**: σ , the standard deviation when the distribution of log-productivities is normal. Panel **B**: f , a fraction to scale down all city population sizes, effectively changing the parameter n_{min} representing the size above which city sizes follow a Pareto distribution. Panel **C**: α , the Pareto coefficient for the distribution of city sizes.

that constant returns to scale (elasticity of 1) will only appear if *all* cities have sizes larger than $e^{\sigma^2/2}$. This piecewise separation given by the specific condition $\sigma > \sqrt{2 \ln(n_{min})}$, it is important to recall, comes from a heuristic argument and is not a sharp boundary between the regimes of increasing returns to scale and constant returns to scale.

Often, research is carried using a percentage sample of the total populations in a country, typically because full census microdata is not available (a case will be discussed in more detail in Section 5.2). Hence, researchers often estimate average productivities across cities using samples of the city populations. To model this situation we only need to introduce a new parameter f , a number between 0 and 1, premultiplying n_{min} , using the convenient property that taking a fixed percentage f of all city sizes does not change probabilities of events $p_N(n)dn$ based on the Pareto density in eq. (11). For example, if one is working with a 1% census sample, it suffices to multiply the parameter n_{min} in eq. (12) by $f = 0.01$.

Figure 3 shows three graphs, plotting $\beta_{ave}(f, n_{min}, \sigma, \alpha)$ as a function of one of the parameters, keeping constant the other parameters. The plots show the variation due to σ , f , and α , respectively in panel A, in panel B, and in panel C.

Panel A in fig. 3 is as we would expect, showing that for distributions of productivity with thin tails (small σ) the returns from scale should be constant (i.e., $\beta = 1$), but for heavy-tailed distributions (large σ), they should increase (i.e., $\beta > 1$). Panel B confirms the

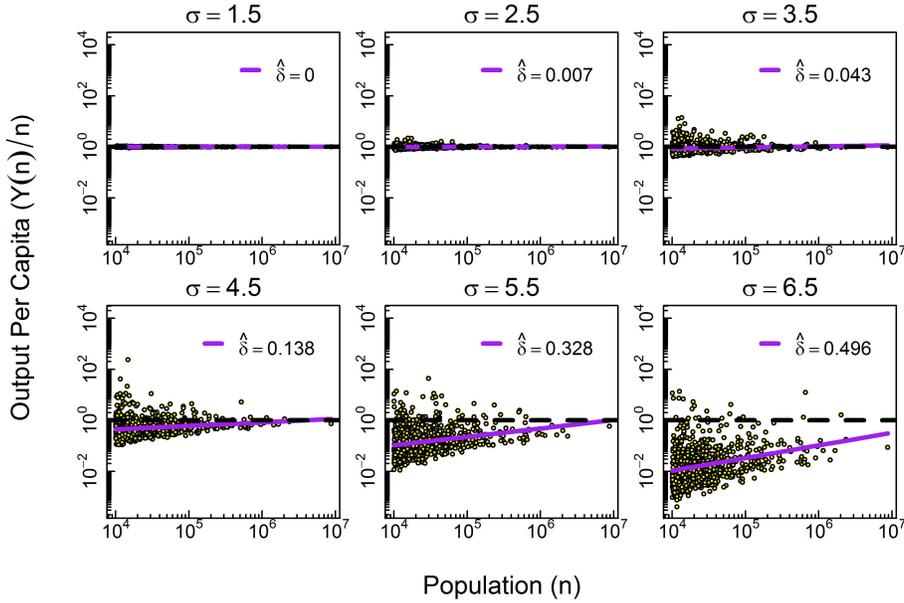


Figure 4: Effect of σ on urban productivity per capita in null model. In these simulations $m = 900$ cities were used with populations generated from a Pareto distribution of parameters $n_{\min} = 10,000$ and $\alpha = 1$. The parameter x_0 was adjusted for each value of σ so that $E[X] = 1$ is fixed (black dashed line). The average increase in productivity per capita with increasing population size is shown using an Ordinary Least Square (OLS) regression line of $\ln(Y(n)/n)$ against $\ln(n)$ (purple solid line).

effect of the law of large numbers, whereby taking larger percentages of the city populations reduces the artificial UWP. Finally, Panel C shows that β increases with α , which suggests that the average elasticity will be larger the less unequal are city sizes, and the less likely extremely large sizes are generated.

Of all three parameters, α has the weakest effect on β . Its effect in practice is probably negligible given the fact the estimated values from data barely deviate from $\hat{\alpha} \approx 1$. In contrast, parameters σ and f strongly affect the values of β . In the following sections, we will analyze the effect of σ through simulations, and the effect of f with real world data.

4. Simulations

We simulate $m = 900$ synthetic cities using our model. Each city has a population n_k taken from a Pareto distribution with parameters $n_{\min} = 10,000$ and $\alpha = 1$, and for each of them, we generate n_k productivities sampled i.i.d. from eq. (2). We will generate simulations

for different values of the model parameter σ , and we will compare $y^{(s)}(n_k)/n_k$ against n_k , for all cities $k = 1, \dots, 900$, where we will use the superscript (s) to make explicit the fact that the output of cities is simulated.

These populations we use for cities range from ten thousand to close to twenty million inhabitants. Based on fig. 1 we expect the maximum productivity to be dominant when $\sigma > 4$ for these population sizes. And based on figs. 2 and 3, we thus anticipate that an artificial UWP will emerge when $\sigma > 4$.

Figure 4 shows the results of such simulations, plotting per capita productivity with respect to population size, using logarithmic axes. The black dashed line is the theoretical expected value of average productivity, which we set to $\mu = 1$.⁵ In spite of eq. (3), we observe that the simulated data in fig. 4 are well described by the relation $y(n)/n = Y_0 n^\delta$. The purple solid line is the OLS fit of

$$\ln(y^{(s)}(n_k)/n_k) = \ln(Y_0) + \delta \ln(n_k) + \varepsilon_k,$$

where $\widehat{\delta}$ ranges between 0 and 0.5 depending on the value of σ . Therefore, the average total output of the city under our model is well approximated by the function

$$F_{model}(n) = Y_0 n^\beta, \tag{13}$$

where $\beta = 1 + \delta$. As can be observed, the parameter σ controls the estimated elasticity β of output with population size, exactly as we anticipated from Section 3. We also note that the estimated Y_0 decreases as σ increases. That all these effects emerge when the variance of the log-productivity is very large is also reflected on the fact that the overall dispersion around the average trend increases, meaning that the goodness-of-fit of eq. (13) decreases, as evidenced by the low R^2 values in Figure 5C below.

Figure 5 plots more systematically the departure of $\widehat{\beta}$ and $\widehat{\ln(Y_0)}$ from their theoretical values, $\beta = 1$ and $\ln(Y_0) = \ln(\mu) = 0$, and their dependence on σ . That is, these graphs show the concomitant emergence of UWP with the departure from the LLN. It was constructed by

⁵As σ changes one has to adjust the value of x_0 so that $E[X] = \mu$ is kept constant. Changing the value of x_0 in order to keep $E[X]$ fixed across individuals and across allows us to isolate the effect of changes in the variance of log-productivities.

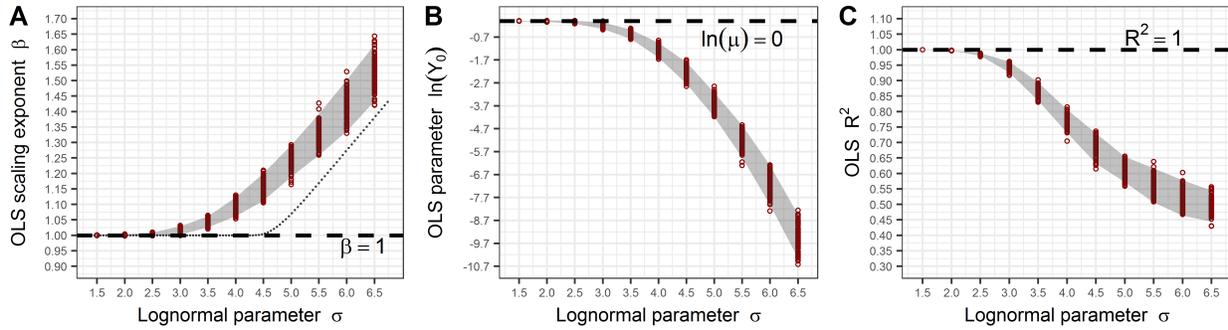


Figure 5: Increasing returns to scale driven by a departure from the Law of Large Numbers (LLN) for a lognormal distribution of productivities as they become more heavy-tailed. Each point represents the results of a linear regression for a cross-section of cities, like one of the panels in fig. 4, for which we show the OLS estimations of: (panel **A**) the scaling exponent β , (panel **B**) the intercept $\ln(Y_0)$, and (panel **C**) the R^2 of the regression. For each value of σ we generated 100 simulations. The values of $\hat{\beta}$, $\widehat{\ln(Y_0)}$ and R^2 change systematically with the value of the σ -parameters of the individuals' lognormal. As σ increases, the scaling exponent also increases, the intercept decreases, and the goodness of the linear fit decreases, evidencing increasing violations of the LLN. The gray areas show the regions where 95% of the point estimates fell. The dotted line in panel **A** is the elasticity predicted by Equation (12).

simulating 100 different runs of the model (i.e., 100 different cross-sections of m cities defined by the ordered pair $(n_k, y^{(s)}(n_k))$) per each value of σ between 1.5 and 6.5. We observe that the value of $\hat{\beta}$ starts to depart from 1.0 when $\sigma \approx 3.0$, qualitatively following the predictions from eq. (12) and fig. 3. The point $\sigma \approx 3.0$ is also where $\widehat{\ln(Y_0)}$ deviates from 0. In panel A of fig. 5, the dotted line represents the analytical curve predicted by $\beta_{ave}(n_{\min}, \sigma, \alpha)$ for $n_{\min} = 10,000$ and $\alpha = 1.0$. It is important to note that, for each value σ , the gray area representing the region where 95% of estimated values of β fell is relatively narrow, which means that the average elasticity can indeed, *significantly and systematically*, depart from 1. These departures from the theoretical values are associated with a larger unexplained variance of the OLS regression, which we observe as a monotonically decreasing R^2 .

The discussion in this section provided support to the analytic results presented in Section 3. In particular, the simulations verified the effects of increasing the variance of log-productivity on the elasticity of total output with respect to size, for a fixed number of cities with fixed populations. Doing these allowed us to study eq. (12) when we change σ but we left the other distributional parameters fixed. As was demonstrated, the null expectation

when assessing the presence of the city size productivity premium was not always the absence of UWP. Instead, UWP are observed under certain conditions even though the data generating process does not have the putative underlying mechanisms, and we explored the particular situation in which individual log-productivity had very large variances.

In the next section we will analyze the other important part of our results: the effects of changing f , while keeping σ fixed. The prediction is that the city size premium will artificially become larger with increasingly smaller sample sizes (see panel B of fig. 3). For this, we will analyze real data on Colombian wages.

5. An Application

In Section 3 we derived eq. (9), which highlights two important effects. On the one hand, that the elasticity β of total output in a city with respect to its population size will increase if the standard deviation of log-productivity, σ , increases. On the other, it tells us that β will also increase when population sizes n are small. The former prediction was analyzed in the last section through simulations. The latter prediction is studied in this section. We will show it has consequences on real world data analyses, especially if they rely on surveys or subsamples based on census data.

In this section we have two specific goals. The first is to confirm the prediction that UWP increase when n decreases. The second, to check whether this artificial UWP arises in typical real world data. Using administrative data on nominal wages of individual formal sector workers in Colombia, and information of the municipality they live in, we will thus assess whether this sampling effect is present in real data, and to what degree.

The basic methodology utilized to assess whether the phenomenon of increasing returns to scale (IRS), as a statistical sampling effect, is present in real world data is *geographical randomization of individuals*. The methodological choice is motivated by the consideration that, as we argue below in more detail, the statistical effect should be robust to locating individuals randomly in space.

Whether wages are an appropriate measure of productivity depends on the question one is asking (Combes and Gobillon, 2015). For the purpose of estimating the city size productivity premium, nominal wages tend to capture almost all of the effects from either sorting or

agglomeration. In the particular context of the statistical effect we are studying here, the distinction between different measures of productivity (such as Total Factor Productivity, real or nominal wages, or GDP per worker) is only relevant when, and if, they differ in their σ parameter. While we use nominal wages as the urban productivity measure there is no a priori reason to think that the findings are specific to this measure.

5.1. Data, Descriptives and Distributions

The data used here is the 2014 administrative records of the social security system in Colombia (the Spanish acronym is PILA, for *Integrated Report of Social Security Contributions*). We refer the reader to the Appendix A for the source, details and preparation of the data, as the dataset has been cleaned and prepared for the analysis of average monthly wages across formal workers in all Colombian municipalities. After the preprocessing of the data, the study population of analysis consists of 6,713,975 workers employed in the formal sector, geographically distributed in 1,117 municipalities that cover almost the entire Colombian territory.⁶ We quantify the size of a municipality using the count of formal employment, defined as the number of workers in our data that reported the municipality as the last place of work in 2014.

As we showed in Section 3.3, if UWP come from eq. (9), the average elasticity that will be estimated from a regression with many cities (see eq. (12)) is determined by the properties of two distributions: that of productivities and that of population sizes. In our derivation, we assumed that the former were lognormally distributed and the latter were Pareto distributed. When working with real datasets, we thus want to characterize the empirical distributions of wages and sizes in our data, and assess whether a lognormal and a Pareto stand as reasonable approximations.

Figure 6 plots the full empirical complementary cumulative distribution function of municipality sizes. We have followed the methodology proposed by Clauset et al. (2009) to visualize and fit Pareto distributions. Plotting both axes on logarithmic scales, the straight line on the right tail reveals a Pareto distribution. We observe in this empirical distribution a natural small-size scale, determined by the estimated minimum size, n_{\min} , above which the

⁶There were 1,122 municipalities in the country officially as of 2014.

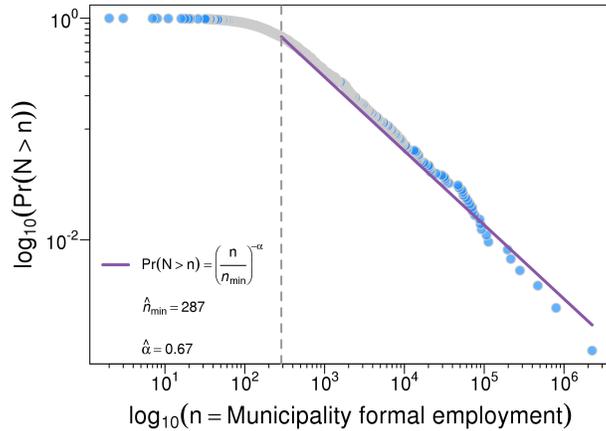


Figure 6: The complementary cumulative empirical distribution of number of workers across municipalities (blue circles) is well-fit by a Pareto distribution (solid purple line).

Pareto distribution is well fit (see Clauset et al., 2009 for how to estimate this parameter). We show this value as the vertical dashed line. We will carry out all our subsequent analyses on the municipalities above that small-scale size, and on the individuals that live in those municipalities. Dropping the small-sized municipalities allows us to satisfy the assumption we used for Equation (12), that city sizes are Pareto distributed. Truncating the data in this way does not change our results qualitatively.⁷ Dropping municipalities that have less than 287 formal workers, means dropping from our analysis 80,526 workers (only 1.2 percent of total workers in our sample) and 564 municipalities (approximately half of all municipalities).

In Appendix C we present a comparison of the goodness-of-fit statistics for several probability distribution functions to model municipality sizes (table 2) and wages (table 3). Both quantities are truncated from below, so we fitted accordingly some truncated probability distribution functions through Maximum Likelihood Methods. For both sizes and wages, the two best fits were obtained by a truncated-lognormal and a Pareto distribution. Specifically, wages were best fit by the former while sizes were best fit by the latter, as quantified by three criteria: the largest likelihood, the minimum Akaike Information Criterion (AIC), and the

⁷Dropping the municipalities with the smallest sizes is typically done as this reduces the potential bias introduced by the fact that their formal employment is overrepresented by public servants whose wages are not determined by economic forces.

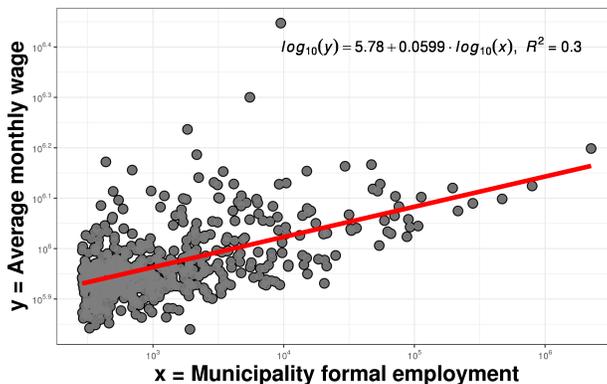


Figure 7: Larger municipalities have, on average, workers with higher monthly wages.

minimum Bayesian Information Criterion (BIC). Figures 9 and 10 in Appendix C show the diagnostic graphical comparison for the distributions of sizes and wages, respectively, fitted by a truncated-lognormal, a Pareto, and a normal distributions, along with some descriptive statistics.

These fitted distributions yielded the estimated values $\hat{\sigma} \approx 2.00$, $\hat{\alpha} \approx 0.67$ and $\widehat{n_{\min}} \approx 287$, which we can use to get a sense for whether to expect UWP from sampling in this data (see Appendix C for estimated confidence intervals of these parameters). Using eq. (12), we get $\beta_{ave}(287, 2, 0.67) = 1$. However, replicating the exercise exemplified in panel B of fig. 3 but with these fitted parameters, we can obtain a value above one if we reduce sizes using fractions of the data for f less than 0.01, approximately. For example, for $f = 0.005$, we get $f n_{\min} = 1.435$, which yields $\beta_{ave}(1.435, 2, 0.67) = 1.082$. In other words, our results predict that we will observe a realistic urban wage premium of about 8.2% if we consider a 0.5% sample of the Colombian population of formal workers. We will study the effect of taking smaller samples of workers in more detail below.

5.2. Telling Apart Real Versus Artificial UWP

Figure 7 plots the cross-section of the average monthly wage per municipality with respect to municipality size. There is clearly a positive and significant elasticity $\hat{\delta} \approx 0.06$.

The strategy to estimate whether an elasticity of wages with respect to size, such as the one observed in fig. 7 in Colombian municipalities, is due to an artificial sampling effect is to randomize workers geographically. The reasoning behind this is fairly clear: while random-

izing individuals should eliminate the empirical evidence for urban productivity premiums given by the built-in dependencies of individuals caused by sorting or agglomeration effects, the artificial UWP effect should be statistically invariant to the removal of the causal effects present in the data. Randomizing will destroy the information of the way workers have sorted themselves across cities, and of who the workers have interacted, or are interacting, with. In other words, the causal effects are removed by randomizing the spatial location of workers, but the distributional effects are not. After we randomize the municipalities where workers work, any UWP arising from a regression must come from the statistical sampling effect of the distribution.

Notice that randomization does not change workers' wages. In this sense, we have not destroyed *all* of the information, since the distribution of wages is itself a consequence of the socioeconomic causes related to people moving, agglomerating, and learning from each other. Hence, we are not claiming that the geographical randomization of people assumes workers would have earned that same wage had they worked in that new location. We are also not claiming that the distribution of wages should be invariant to the presence or absence of sorting or agglomeration effects. We are just saying that there is a distribution of wages, which we acknowledge may arise from local processes, but that *given that distribution* UWP could arise naturally in a regression, *even* after destroying the local information attached to where people are located.

For the real and the randomized versions of the data, we will estimate the following basic regression:

$$\ln \left(\bar{w}_k^{(f,j)} \right) = \alpha + \delta \ln (n_k) + \varepsilon_k, \quad (14)$$

where the dependent variable is the natural logarithm of the average wage in municipality k , $\bar{w}_k^{(f,j)}$, with $j \in \{real, randomized\}$, where “*real*” indicates that we compute the average wage from the actual individuals that work in municipality k , whereas “*randomized*” indicates that we are taking the average *after* randomly permuting the location of individuals across municipalities. The superscript f is to indicate that the average wage (real or randomized) was taken over a subsample of all workers. We will take $f = 0.1\%$, 0.5% , 1% , 5% , 10% and 100% samples. In the regression given by Equation (14), however, the size of

formal employment n_k for each municipality is kept constant across sample percentages. To summarize the procedure, first, we will sample *without* replacement a fraction f of all workers, second, we will estimate the real unconditional elasticity, and third, we will randomize several times the locations of individuals by applying random permutations of the location of individuals in the sample, estimating the elasticity for each randomization.

Table 1: Results of four OLS regressions comparing real versus randomized location, and carrying out the analysis with all workers versus a small sample of them.

<i>Dependent variable: log(Average monthly wage)</i>				
	All workers		0.1% of all workers	
	Real locations	Random locations	Real locations	Random locations
	(1)	(2)	(3)	(4)
log(Employment size)	.060*** (.004)	.001 (.001)	.072*** (.013)	.070*** (.018)
Constant	13.317*** (.029)	14.084*** (.011)	13.151*** (.099)	13.342*** (.140)
Municipalities	553	553	342	342
Num. workers	6,633,449	6,633,449	6,633	6,633
Adj. R ²	.294	-.001	.086	.041
F Stat.	230.6*** (df = 1; 551)	.3 (df = 1; 551)	33.1*** (df = 1; 340)	15.6*** (df = 1; 340)

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 1 shows the comparison between regressions carried on the real locations and on randomized locations, and for all workers or just a sample of them. The first column is the same result shown in fig. 7, except in the figure the equation shows the regression equation using base 10 logarithms, whereas table 1 shows the results of regressions using natural logarithms (this change in the bases of logarithms does not change the estimated elasticity, $\hat{\delta} \approx 0.06$). The second column in the table is the same regression but with the locations of workers randomized (see eq. (14)). The third and fourth columns are exactly the same

as before, except we have computed the average monthly wage using only a 0.1% sample of workers ($f = 0.001$). Taking a random subsample of workers from the full population keeps the distribution of productivities and the distribution of municipality sizes approximately fixed (i.e., the parameters σ and α of the distributions stay approximately constant), except we are reducing the sample sizes of municipalities, and thus we are scaling down the parameter n_{\min} .

Comparing the coefficient of log Employment size between the first and second columns, we observe that randomizing the spatial location of individuals effectively destroys the urban size effect, as expected. However, when we restrict our analysis to a small random subsample of the workers (both third and fourth columns), randomizing individuals geographically does not destroy the urban size effect. While the fourth column represents a single geographical randomization among many, these results connect with, and confirm, our analytical expectations as revealed by eq. (12), as well as our numerical prediction about what to expect when using $f < 0.01$. Namely, that reducing the sample size increases the estimated elasticity of total output with respect to size.

Crucially, the statistical significance of the elasticity of the log Employment size in column four in table 1 calls into question the estimation reported in the third column. Statistically, the coefficients in columns three and four are not significantly different (more on this below). It implies that if we only had access to a 0.1% of workers in Colombia, and the cross-section of where they work, there is a chance we would not be able to reject the possibility that the city size premium was a statistical artifact.

We systematize the type of analysis in table 1 to understand the robustness of these results. As was explained, the method is to (i) take a subsample f from the full population of workers, with $f = 0.1\%$, 0.5% , 1% , 5% , 10% and 100% , (ii) compute the elasticity of average monthly wage with respect to employment size and refer to this as the “real” elasticity, (iii) randomize the location of individuals, (iv) estimate the elasticity of average monthly wage with respect to employment size and refer to this as a “randomized” elasticity. For a given subsample, we do (iii) and (iv) 1,000 times. Furthermore, we repeat this whole process, (i)-(iv), 10 times so that we can compare different subsamples determined by the same f . After estimating the elasticity for each randomization, we obtain a distribution of possible

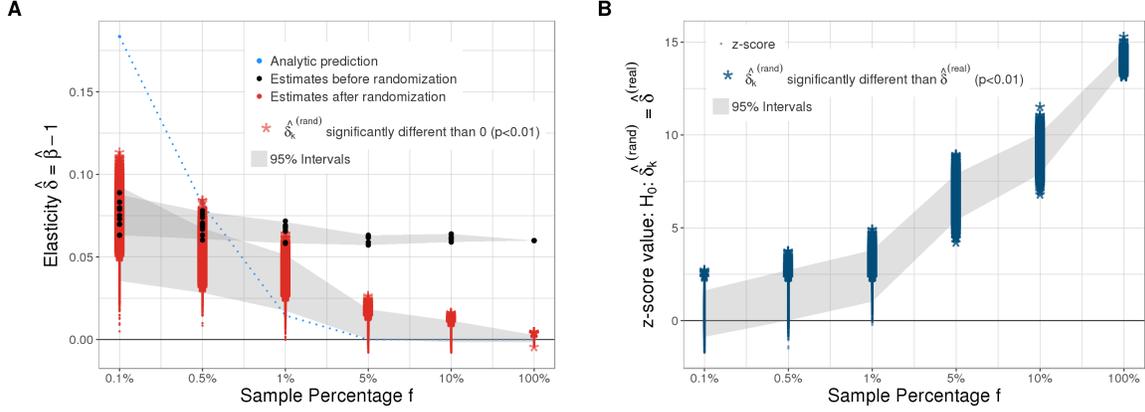


Figure 8: Effects on elasticity from decreasing sample sizes by reducing the number of individuals per city. Panel **A** plots the elasticities (y-axis) calculated before randomizations (black dots) and after randomizations (red dots), for a given subsample of workers determined by each of the values of parameter f (x-axis). It is observed that as the percentage sample f decreases, the distribution of elasticities increase (see main text for details about the procedure). Those red dots that are statistically different from zero have been highlighted by a red star. The dotted blue line is the elasticity $\delta = \beta_{ave}(f \times n_{\min}, \sigma, \alpha) - 1$ predicted by Equation (12), with values $n_{\min} = 287$, $\sigma = 2.0$, $\alpha = 0.67$, and $f = 0.001, 0.005, 0.01, 0.05, 0.10$ and 1.00 . Panel **B** plots the values of the z-score statistic for each elasticity from the OLS regression after individuals have been randomized, constructed in order to test the null hypothesis that, given a subsample of the workers, the elasticity after individuals have been randomized is equal to elasticity before randomization. Elasticities of the randomized samples that are statistically different from the corresponding elasticity without randomization have been highlighted by a blue star.

elasticities, all due to sampling effects.

This procedure we propose assumes a data generating process in which individuals locate themselves randomly in different “buckets” which we call municipalities. There are two expected results involved in this procedure. First, we expect the estimated elasticities after having randomized the geographical location of workers will be different than zero (i.e., UWP without sorting and agglomeration economies). Given this expected result, we want to test a null hypothesis that asserts the elasticities are equal to zero. Second, we expect for small values of f that the estimated elasticities we get from this data generating process will not be statistically different from the elasticity we observe in real data (i.e., without randomization). Thus, we want to test a null hypothesis that, given a specific subsample of workers, asserts the elasticities after randomization are equal to the elasticities before randomization.

The first hypothesis is tested automatically when we carry our OLS regressions using a t -statistic. The second hypothesis we test it by constructing a z -score (see Clogg et al., 1995; Paternoster et al., 1998). For the latter, assume a specific subsample of workers, and let $\widehat{\delta}^{(\text{real})}$ be the estimated elasticity without the randomization, and $\widehat{\delta}_k^{(\text{rand})}$ the elasticity after one specific k -th randomization. Since these are OLS estimates they are assumed to be drawn from a normal distribution, and there is a standard error associated with them, $se^{(\text{real})}$ and $se_k^{(\text{rand})}$, respectively. Under the null hypothesis that these two estimated elasticities are equal, and assuming the number of municipalities large, we can construct the following z -score:

$$z_{\text{stat}} = \frac{\widehat{\beta}^{(\text{real})} - \widehat{\beta}_k^{(\text{rand})}}{\sqrt{(se^{(\text{real})})^2 + (se_k^{(\text{rand})})^2}}$$

which will follow approximately a standard normal distribution.

Figure 8, panel A, plots the elasticities before and after randomizations. Those elasticities after randomization that are statistically significant (at a level $p < 0.01$) have been highlighted with a red large star marker. Since for each subsample we generate 1,000 randomizations, we also show the bands between which 95% of the elasticities fall. The blue dots show our analytic prediction. As can be observed, we confirm that the elasticities after randomizing individuals increase steadily as smaller samples are taken, until they are not significantly different anymore from the real ones for fractions below $f < 0.01$. This is shown in panel B of fig. 8, which plots the z -score. Those elasticities after randomization that are statistically different (at a level $p < 0.01$) than the corresponding elasticity before the randomization) have been highlighted with a blue large star marker.

6. Discussion and Conclusions

The argument presented here provides evidence for there being a mechanism, previously unaccounted in the urban economics literature, that can give rise to a statistical effect which can be mistaken for a city-size productivity premium. Our goal is to persuade the reader that increasing returns to scale exhibited by data can be generated under certain conditions, systematically, without recourse to sorting or agglomeration effects. To analytically derive the strength of this effect we exploited (i) that measures of individual productivity are

lognormally distributed, and (ii) that lognormals belong to the class of heavy-tailed distributions (specifically, subexponential distributions). Through the use of both simulations and empirical analysis, we confirmed the predictions that result from our analytical derivations.

Underpinning our argument was the assumption that individual productivities are inherently stochastic and drawn from the same distribution with the same fixed parameters across places. A key assumption for the validity of our results was to assume, in addition, that productivities were mutually independent. This assumption was not made for mathematical convenience, but precisely to facilitate our main claim: that increasing returns to scale (IRS) can emerge in the total absence of self-sorting, externalities, or interactions, which are mechanisms that would induce dependencies and correlations between productivities. This is not to say that productivities in real settings are independent, just that independence itself does not guarantee the absence of UWP. The specific assumption of lognormally distributed output at the level of individuals, on the other hand, is by itself uncontroversial (see Kleiber and Kotz, 2003, pp. 126–130). Lognormal variability is often the result of a variety of independent factors acting multiplicatively (that is to say, strongly interacting) when generating individual-level productivities (Limpert et al., 2001). An early argument about the distribution of productivity being lognormal was given by Roy (1950) albeit not formally. In the context of scientific output, one of the first to recognize a lognormal distribution describing the productivity of individual researchers was Shockley (1957). It must be noted that these assumptions do not necessarily apply only to people, but also to larger organizations like households, firms, or industrial clusters.⁸

We speculate that UWP without sorting or agglomeration would still emerge if productivities were not lognormally distributed, conditioned they still follow a subexponential distribution. The reason subexponential distributions of productivity may naturally induce UWP is because this class of distributions have the special property that sums of random variables can have magnitudes comparable to single extreme values. Because of this, the average of a sample that is larger than expected is more likely due to a single variable in the

⁸Interestingly, lognormal distributions have even been observed at the level of whole cities (Bettencourt et al., 2010; Gomez-Lievano et al., 2012; Alves et al., 2013b,a; Mantovani et al., 2013; Alves et al., 2014).

sample having an extreme value, than to the contribution of many variables being slightly larger than expected. In other words, for subexponential distributions the disjunction of many improbable events (with potentially very large influence) becomes more probable than the conjunction of many probable events (but with small influence each). If we add to this the fact that the maximum among n random variables may grow faster than linearly, we obtain that sample averages may systematically grow as sample sizes grow (this growth will eventually stop after some large size *if* the distribution has finite mean and variance due to the law of large numbers).

In our analytical results, we showed that the elasticity emerging from the effect we presented here depends positively on the standard deviation of log-productivities, and negatively on the sample size of the sample considered. We derived a precise formula to compute the null model elasticity for both a single city and a cross section of cities, the latter being solely a function of the distributional parameters of productivities (σ) and city sizes (n_{\min} and α). Our approach shifts attention away from the study of averages to the analysis of probability distributions. While many aggregate phenomena can be understood well enough by studying averages, a more complete understanding comes from studying how aggregates emerge from the properties of the underlying full probability distribution (Gould, 1996; Gabaix, 2009; Gomez-Lievano et al., 2012; Behrens and Robert-Nicoud, 2015)). As we demonstrated, the phenomenon of UWP is an aggregate phenomenon that can be obtained from a heavy-tailed distribution of individual productivity, even when having expected mean and variance that are fixed and independent of scale. Our investigation therefore contributes to our understanding of the effects of heterogeneity in cities (e.g. Behrens and Robert-Nicoud, 2015).

We studied the practical relevance of our results using administrative data at the worker level in Colombia. Duranton (2016) already presented a rigorous analysis of agglomeration effects in Colombian municipalities (and metropolitan areas). Duranton warns, however, about the use of administrative data at the worker level for developing countries (such as Colombia) since in these countries less than half the working age population is employed formally. This is a valid concern, indeed, and in our case our dataset consists of 6.79 million formal workers, which represents a restricted sample out of the 31.3 million individuals between ages 15 and 64 who represent the total working age population in the country (see

the studies about formality in Colombia by O’Clery and Lora, 2016; O’Clery et al., 2016). To quantify the true association between city size and the earnings of workers one must have, ideally, data on the informal sector as well. To account for this, Duranton (2016) uses a survey of Colombian households that contains individuals that work in both the formal and informal sector.

A comparison with the framework presented in Duranton (2016)’s was beyond the scope of this paper, much less a full replication. Given that we have access to the full population of Colombian formal workers in 2014, we used our dataset instead as a case study to illustrate the statistical emergence of UWP without sorting or agglomeration effects in real data. Our study confirmed the presence of this effect, given the broad distribution of wages, for subsamples smaller than 1% of the total population of workers in our data. This is, indeed, a very small sample of 66,335 workers. Conversely, we found that the artificial UWP disappears for samples larger than that, implying that a causal explanation is required. Since it is likely that the distribution of income in the informal sector is less uneven than the distribution of income in the formal sector, by rejecting the presence of a statistical effect in the elasticity of wages with respect to city size in our data we are likely to give additional support to Duranton’s results. In other words, Duranton (2016)’s samples are sufficiently large, meaning that his estimated elasticity of 11% (from a simple OLS regression of wages with respect to city size not controlling for individual characteristics) is probably free of the artificial UWP our paper is about.⁹

Further work should be devoted to studying the effects of adding control variables. The effective sample size per city can be reduced, for example, if too many controls are included (e.g., in order to do in-group regressions). Including several demographics may effectively partition the population in several subgroups, and this may increase the likelihood of an artificial UWP appearance.

In general, our present study highlights the importance of analyzing with care data from small samples, or surveys. One must understand the distributional properties that describe

⁹This estimate is somewhat higher than what is typically observed elsewhere (Rosenthal and Strange, 2004; Puga, 2010).

individuals, in particular how the variance relates to the possible sample sizes. We regard the effect we have studied here as a bias, but we distinguish it from other types of statistical biases in that the sampling effect is a real tangible property of a statistical distribution which has consequences on measures of central tendency such as the sample mean. As mentioned in the Background section, this has already been studied in other contexts, for example by Gabaix and Landier (2008); Gabaix (2011). This line of research warns about naive interpretations of what increasing returns to scale mean from a statistical point of view when measures of individual output are unevenly distributed. As a corollary, the equivalence between “total output increases more than proportionately with size” and “individual productivity increases with larger sizes” is only applicable when the law of large numbers is valid. This means, moreover, that per capita transformations can give misleading information about the average individual productivity.

In this work, it is important to emphasize, we do not seek to refute the relevance of sorting and agglomeration for explaining the well-established city size productivity premium. The main consequence of our work, rather, is methodological. We argue that theoretical models and statistical analyses involving increasing returns to scale should have as their objects of study the statistical distributions of productivity, especially in cities. This is because cities are highly heterogeneous places, yet since they are not infinitely sized, the assumption that the law of large numbers always holds is not guaranteed. It means that our estimates of the urban productivity premium may carry a bias arising from a statistical artifact. It also means that our null expectation should not be the statistical absence of an urban size effect, but rather the presence of it. We hope that further analysis of the effect of urban size on productivity will account for these distributional effects.

Acknowledgements

We gratefully acknowledge useful comments from Ricardo Hausmann, Luis M.A. Bettencourt, Rachata Muneeppeerakul, Dario Diodato, and Michele Coscia. Special thanks go to Frank Neffke, for many helpful comments and suggestions, which improved the clarity in the presentation of some of the central results in this project. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest: none.

References

Aitchison, J., Brown, J. A. C., 1957. The Lognormal Distribution with Special Reference to Its Uses in Economics. University of Cambridge, Department of Applied Economics, Monograph 5. Cambridge University Press, London.

Alves, L., Ribeiro, H., Lenzi, E., Mendes, R., 2014. Empirical analysis on the connection between power-law distributions and allometries for urban indicators. *Physica A: Statistical Mechanics and its Applications* 409 (0), 175 – 182.

URL <http://www.sciencedirect.com/science/article/pii/S0378437114003641>

Alves, L. G., Ribeiro, H. V., Mendes, R. S., 2013a. Scaling laws in the dynamics of crime growth rate. *Physica A: Statistical Mechanics and its Applications* 392 (11), 2672 – 2679.

URL <http://www.sciencedirect.com/science/article/pii/S0378437113001416>

Alves, L. G. A., Ribeiro, H. V., Lenzi, E. K., Mendes, R. S., 08 2013b. Distance to the Scaling Law: A Useful Approach for Unveiling Relationships between Crime and Urban Metrics. *PLoS ONE* 8 (8), e69580.

Andersson, M., Klaesson, J., Larsson, J. P., 2014. The sources of the urban wage premium by worker skills: Spatial sorting or agglomeration economies? *Papers in Regional Science* 93 (4), 727–747.

Behrens, K., Duranton, G., Robert-Nicoud, F., 2014. Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy* 122 (3), 507–553.

Behrens, K., Robert-Nicoud, F., 2015. Chapter 4 - agglomeration theory with heterogeneous agents. In: Gilles Duranton, J. V. H., Strange, W. C. (Eds.), *Handbook of Regional and Urban Economics*. Vol. 5 of *Handbook of Regional and Urban Economics*. Elsevier, pp. 171 – 245.

URL <https://www.sciencedirect.com/science/article/pii/B9780444595171000040>

- Bettencourt, L. M. A., June 2013. The Origins of Scaling in Cities. *Science* 340, 1438.
- Bettencourt, L. M. A., Lobo, J., Strumsky, D., West, G. B., 11 2010. Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities. *PLoS ONE* 5 (11), e13541.
URL <http://dx.doi.org/10.1371/journal.pone.0013541>
- Clauset, A., Shalizi, C. R., Newman, M., 2009. Power-law distributions in empirical data. *SIAM Review* 51 (4), 661–703.
- Clogg, C. C., Petkova, E., Haritou, A., 1995. Statistical methods for comparing regression coefficients between models. *American Journal of Sociology* 100 (5), 1261–1293.
- Combes, P.-P., Duranton, G., Gobillon, L., 2008. Spatial wage disparities: Sorting matters! *Journal of Urban Economics* 63 (2), 723–742.
- Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., Roux, S., 2012. The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica* 80 (6), 2543–2594.
- Combes, P.-P., Gobillon, L., 2015. Chapter 5 - the empirics of agglomeration economies. In: Gilles Duranton, J. V. H., Strange, W. C. (Eds.), *Handbook of Regional and Urban Economics*. Vol. 5 of *Handbook of Regional and Urban Economics*. Elsevier, pp. 247 – 348.
URL <https://www.sciencedirect.com/science/article/pii/B9780444595171000052>
- De la Roca, J., Puga, D., 2017. Learning by working in big cities. *The Review of Economic Studies* 84 (1), 106–142.
- Dupuy, A., 2012. A microfoundation for production functions: Assignment of heterogeneous workers to heterogeneous jobs. *Economica* 79 (315), 534–556.
- Duranton, G., 2016. Agglomeration effects in Colombia. *Journal of Regional Science* 56 (2), 210–238.

- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. *Handbook of regional and urban economics* 4, 2063–2117.
- Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. *Journal of Political Economy* 122 (3), 554–620.
- Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. *Modelling extremal events: for insurance and finance*. Vol. 33. Springer Science & Business Media.
- Gabaix, X., 2009. Power Laws in Economics and Finance. *Annual Review of Economics* 1, 255–293.
- Gabaix, X., 2011. The granular origins of aggregate fluctuations. *Econometrica* 79 (3), 733–772.
- Gabaix, X., Landier, A., 2008. Why has ceo pay increased so much? *Technology* 49.
- Glaeser, E. L., 2011. *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier*. Penguin Press, New York.
- Glaeser, E. L., Mare, D. C., 2001. Cities and skills. *Journal of labor economics* 19 (2), 316–342.
- Gomez-Lievano, A., Patterson-Lomba, O., Hausmann, R., 2016. Explaining the prevalence, scaling and variance of urban phenomena. *Nature Human Behaviour* 1, 6.
- Gomez-Lievano, A., Youn, H., Bettencourt, L. M. A., 07 2012. The Statistics of Urban Scaling and Their Connection to Zipf’s Law. *PLoS ONE* 7 (7), e40393.
URL <http://dx.plos.org/10.1371/journal.pone.0040393>
- Gould, S. J., 1996. *Full House: The Spread of Excellence from Plato to Darwin*. Harvard University Press, Cambridge, MA.
- Houthakker, H. S., 1955. The pareto distribution and the cobb-douglas production function in activity analysis. *The Review of Economic Studies* 23 (1), 27–31.

- Jones, C. I., 2005. The shape of production functions and the direction of technical change. *The Quarterly Journal of Economics* 120 (2), 517–549.
- Kleiber, C., Kotz, S., 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley Series in Probability and Statistics. John Wiley & Sons.
- Leadbetter, M. R., Lindgren, G., Rootzén, H., 1983. *Extremes and related properties of random sequences and processes*. Springer, Berlin.
- Levhari, D., 1968. A note on houthakker’s aggregate production function in a multifirm industry. *Econometrica: journal of the Econometric Society*, 151–154.
- Limpert, E., Stahel, W. A., Abbt, M., May 2001. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* 51 (5), 341–352.
- Mantovani, M. C., Ribeiro, H. V., Lenzi, E. K., Picoli, S., Mendes, R. S., Aug 2013. Engagement in the electoral processes: Scaling laws and the role of political positions. *Phys. Rev. E* 88, 024802.
URL <http://link.aps.org/doi/10.1103/PhysRevE.88.024802>
- Melo, P. C., Graham, D. J., Noland, R. B., 2009. A meta-analysis of estimates of urban agglomeration economies. *Regional Science and Urban Economics* 39 (3), 332–342.
- Mincer, J., 1970. The distribution of labor incomes: a survey with special reference to the human capital approach. *Journal of economic literature* 8 (1), 1–26.
- O’Clery, N., Gomez, A., Lora, E., 2016. The path to labour formality: Urban agglomeration and the emergence of complex industries. Under Review.
- O’Clery, N., Lora, E., 2016. City size, distance and formal employment creation. CAF (Development Bank of Latin America) Working Paper.
- Paternoster, R., Brame, R., Mazerolle, P., Piquero, A., 1998. Using the correct statistical test for the equality of regression coefficients. *Criminology* 36 (4), 859–866.
- Puga, D., February 2010. The Magnitude and Causes of Agglomeration Economies. *Journal of Regional Science* 50 (1), 203–219.

- Rosen, S., 1978. Substitution and division of labour. *Economica* 45 (179), 235–250.
- Rosenthal, S. S., Strange, W. C., 2004. Evidence on the nature and sources of agglomeration economies. *Handbook of regional and urban economics* 4, 2119–2171.
- Roy, A. D., Sep. 1950. The Distribution of Earnings and of Individual Output. *The Economic Journal* 60 (239), 489–505.
- Shockley, W., March 1957. On the Statistics of Individual Variations of Productivity in Research Laboratories. *Proceedings of the IRE* 45, 279–290.
- Sveikauskas, L., August 1975. The Productivity of Cities. *The Quarterly Journal of Economics* 89 (3), 393–413.
URL <http://www.jstor.org/stable/1885259>
- Wirth, L., 1938. Urbanism as a Way of Life. *American Journal of Sociology* 44 (1), 1–24.

Appendix A. Data

In Section 5 we use data of the formal workforce in Colombia to analyze the unconditional elasticity of nominal wages on municipality population size. These come from the administrative records of the social security system in Colombia (abbreviated as PILA in Spanish, meaning the *Integrated Report of Social Security Contributions*). The PILA is maintained by the Colombia Ministry of Finance and Public Credit (“Ministerio de Hacienda y Crédito Público”). PILA consists of individual contributions to health and pensions reported by workers, firms, public institutions, and other formal entities like associations, universities, cooperatives and multilateral organizations.

The dataset was obtained from the Ministry of Finance and Public Credit, under a data use agreement that is part of the development of www.datlascolombia.com, a joint project between the Center for International Development and the Colombian Foreign Trade Bank (Bancoldex) to map the industrial economic activity in Colombia. The data are stored on secure computers at the Harvard-MIT Data Center. Access is restricted to identified and authorized researchers by means of a confidential account. The use of the PILA for research purposes has been reviewed by the Harvard’s Institutional Review Board (IRB). In the database individuals and firms have been previously anonymized in order to protect their habeas data. Harvard IRB determined that this dataset is not human subjects as defined by the Department of Health and Human Services (DHHS) regulations.

Each row of the dataset consists of a monthly contribution to the social security system, with more than seventy different fields with information about the worker and the firm, and with the values of the contribution to health and pension, according to the days the worker worked at the firm in that month. The raw microdata consists of 122,287,562 rows (i.e., social security contributions), from 10,535,587 unique workers (i.e., each worker had an average of 11.6 contributions per year). As explained below, we aggregate and keep a subset of all these observations, and we only use two fields for this study: the list of nominal wages earned, and the municipalities of work to which the wage values were attached.

As a start, these data must be cleaned, as is often the case with datasets built from observations resulting from administrative transactions. Common problems include mis-reported or missing wages, no municipality of work reported, no age reported, duplicated

observations, or missing contribution to pension or health. In addition to dropping these problematic observations, we keep only those workers that are categorized as “dependent” or “independent”, which means they are either employed in a firm or are self-employed, respectively (by keeping these type of social security contributors we exclude those individuals that contribute to social security through means other than a formal job). Finally, we keep those individuals who worked for at least 30 days during the whole year, and had ages between 15 and 64.

We compute the monthly average wage of workers by first adding their net wage earned during the year, then dividing it by the total number of days worked, and finally multiplying by thirty. By law, firms are required to pay a minimum wage to workers, or more. However, there exist special cases in the dataset in which this does not hold. Hence, we make sure this is the case by dropping observations which report average monthly wages below the minimum wage (\$616,000 Colombian Pesos, or COP, in 2014). At the end, our population of analysis consists of 6,713,975 formal Colombian workers (approximately 64% of the unique individuals that appear originally in the dataset).

Appendix B. The null model

The model of lognormally distributed productivities is built on arguments given by Roy (1950) about the distribution of the earnings of individuals. Roy’s central assumptions are that (i) earnings are proportional to the output of workers, and (ii) that output is the product of many internal characteristics of the individuals, in the same way that weights and volumes are the product of linear dimensions. Our model is also based on the one proposed by Shockley (1957) to explain the distribution of individual scientific output in research laboratories, except Shockley finds that Roy’s first assumption does not hold, i.e., “rewards do not keep pace with increasing production”. We do not make this distinction, and we assume, as Roy did, that earnings are proportional to output.

We embed individuals in cities, where UWP have been extensively studied. Cities have long been regarded as the setting where a large and dense population of heterogeneous individuals dynamically assembles to interact socially, economically, and politically (Wirth, 1938). The three most salient characteristics of cities are thus the *physical proximity* that

enables social interactions, the *openness* which allows the flow of people, money, goods and information, and *heterogeneity* between individuals which highlights the socioeconomic complexity of human agglomerations (Glaeser, 2011). UWP in cities have been typically attributed to the first two. Our model exhibits the emergence of UWP under the *absence* of the first two, i.e., proximity and interaction, and we demonstrate the emergence of UWP under the presence of the third characteristic only, i.e., the heterogeneous distribution of productivities.

Based on these ideas, the assumptions of our model are:

Assumption 1: Let a city be defined as the collection of n individuals, $i = 1, \dots, n$. We ignore physical proximity.

Assumption 2: Let each citizen i in the city be defined by a large set of innate, not directly observable, characteristics, $\xi_1^{(i)}, \dots, \xi_S^{(i)}$, where $S \gg 1$, and $\xi_s^{(i)}$ are independent and identically distributed (i.i.d.) positive random variables with finite mean and variance, for all $i = 1, \dots, n$ and $s = 1, \dots, S$. The i.i.d. assumption here removes the possibility of any interaction or correlation between individuals.

Assumption 3: Let the output of individual i be $X_i = \prod_{s=1}^S \xi_s^{(i)}$. Because of Assumption 2, X_i are also i.i.d. random variables.

Assumption 4: Let the total output of the city be $Y(n) = \sum_{i=1}^n X_i$. Hence, the output of each city is the sum of heterogeneous independent individual contributions.

We assume the Central Limit Theorem (CLT) applies in Assumptions 2 and 3. Thus, we have that $\ln(X_i) = \sum_s \ln(\xi_s^{(i)})$ is the sum of $S \gg 1$ terms, and we assume that the number S of innate characteristics that affect human productivity is large enough, such that $\ln(X_i)$ approximately is a normal random variable. Hence, we can restate Assumptions 2 and 3 into a single statement: individuals have a productivity X_i i.i.d. sampled from a lognormal distribution $\mathcal{LN}(x_0, \sigma^2)$ with probability density function

$$p_X(x; x_0, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \ln x_0)^2}{2\sigma^2}}, \quad (15)$$

where x_0 and σ are some positive parameters. The expected productivity is thus $E[X] = \mu \equiv \int x p_X(x) dx = x_0 e^{\sigma^2/2}$.

We use upper case letters to denote random variables, and lower case to denote realized values, i.e., y will represent particular numerical values of Y . We assume that there are m cities, indexed as $k = 1, \dots, m$, each with total populations n_1, \dots, n_m . To numerically analyze our model, we generate a total of $\sum_{k=1}^m n_k$ i.i.d. random variables according to eq. (15), with fixed parameters x_0 and σ . It is worth repeating that the i.i.d. condition in Assumption 3 is not adopted for mathematical convenience, but rather by explanatory intent as we want to demonstrate UWP in the absence of interactions.

Applying Assumption 4, our model yields the output of all m cities, $Y(n_1), \dots, Y(n_m)$. Per capita productivities for each city can in turn be obtained by simply dividing total output over population sizes.

Appendix C. Tables for goodness-of-fit statistics for monthly wages and municipality sizes in Colombia

Table 2: Distributions fitted to municipality sizes. The list of the distributions are ordered from top to bottom by increasing AIC values.

dist	numobs	loglik	AIC	BIC	Parameter 1	C.I.	Parameter 2	C.I.
powerlaw	553	-4,733.51	9,469.02	9,473.33	$\hat{\alpha} = 0.67$	[0.61, 0.72]		
truncnorm	553	-4,733.02	9,470.05	9,478.68	$\widehat{\ln(x_0)} = -22.96$	[-50, 0.44]	$\hat{\sigma} = 6.87$	[3.46, 9.63]
truncweibull	553	-4,733.08	9,470.16	9,478.79	$\hat{a} = 0.04$	[0.03, 0.11]	$\hat{b} = 0$	[0, 0]
truncauchy	553	-4,755.53	9,515.06	9,523.69	$\hat{l} = 0.001$	[0, 186.88]	$\hat{s} = 428.74$	[354.78, 531.55]
truncgamma	553	-4,931.69	9,867.38	9,876.02	$\hat{a} = 0$	[0, 0]	$\hat{\lambda} = 0.0000$	[0, 0]
lnorm	553	-4,942.60	9,889.21	9,897.84	$\widehat{\ln(x_0)} = 7.16$	[7.04, 7.27]	$\hat{\sigma} = 1.44$	[1.35, 1.52]
weibull	553	-5,115.31	10,234.61	10,243.24	$\hat{a} = 0.50$	[0.47, 0.55]	$\hat{b} = 2,882.57$	[2377.31, 3424.19]
gamma	553	-5,299.62	10,603.24	10,611.87	$\hat{a} = 0.31$	[0.28, 0.34]	$\hat{\lambda} = 0.0000$	[0, 0]
truncgumbel	553	-5,937.73	11,879.45	11,888.09	$\hat{a} = 0.0002$	[0, 3980.75]	$\hat{b} = 10,684.55$	[9403.29, 11307.02]
trunclogis	553	-6,003.80	12,011.61	12,020.24	$\hat{m} = 0.0001$	[0, 5064.15]	$\hat{s} = 10,458.00$	[9169.54, 11040.73]
gumbel	553	-6,188.90	12,381.81	12,390.44	$\hat{a} = 2,211.40$	[1374.19, 3351.65]	$\hat{b} = 10,582.62$	[9954.13, 11198.11]
norm	553	-7,178.59	14,361.17	14,369.80	$\hat{\mu} = 11,995.39$	[4325.07, 22479.46]	$\hat{\sigma} = 105,053.80$	[100241.97, 111087.68]

Table 3: Distributions fitted to individual wages. The number of total workers (1,325,950 observations) analyzed in this table differ from the number mentioned in the main text (6,633,449) because wages are clustered on the minimum wage. The fits of continuous distributions to data with repeated values, such as the minimum value which is repeated several times, was much improved when we removed repeated values. The list of the distributions are ordered from top to bottom by increasing AIC values.

dist	numobs	loglik	AIC	BIC	Parameter 1	C.I.	Parameter 2	C.I.
truncnorm	1,325,950	-19,940,740	39,881,484	39,881,508	$\widehat{\ln(x_0)} = 10.23$	[2, 10.15]	$\hat{\sigma} = 2.00$	[1.99, 2.02]
powerlaw	1,325,950	-19,952,047	39,904,095	39,904,108	$\hat{\alpha} = 1.16$	[1.16, 1.16]		
truncauchy	1,325,950	-19,960,472	39,920,948	39,920,972	$\hat{l} = 120,766.20$	[190129.96, 114817.91]	$\hat{s} = 190,130.00$	[174740.73, 206786.87]
truncgamma	1,325,950	-20,018,427	40,036,859	40,036,883	$\hat{a} = 0.0000$	[0, 0]	$\hat{\lambda} = 0.0000$	[0, 0]
truncweibull	1,325,950	-20,077,580	40,155,164	40,155,188	$\hat{a} = 0.72$	[0.72, 1109695.37]	$\hat{b} = 1.11 \times 10^6$	[1109654.53, 1111878.12]
truncgumbel	1,325,950	-20,330,914	40,661,832	40,661,856	$\hat{a} = 0.13$	[0.13, 1332540.63]	$\hat{b} = 1.33 \times 10^6$	[1330795.1, 1332942.21]
lnorm	1,325,950	-20,344,669	40,689,342	40,689,366	$\widehat{\ln(x_0)} = 14.19$	[0.76, 14.19]	$\hat{\sigma} = 0.76$	[0.76, 0.76]
gamma	1,325,950	-20,626,561	41,253,126	41,253,151	$\hat{a} = 1.41$	[0, 1.41]	$\hat{\lambda} = 0.0000$	[0, 0]
weibull	1,325,950	-20,667,064	41,334,133	41,334,157	$\hat{a} = 1.05$	[1.05, 2219332.54]	$\hat{b} = 2.22 \times 10^6$	[2219270.9, 2222105.32]
gumbel	1,325,950	-20,825,629	41,651,261	41,651,285	$\hat{a} = 1.30 \times 10^6$	[1135954.33, 1297516.1]	$\hat{b} = 1.14 \times 10^6$	[1134178.33, 1137906.37]
logis	1,325,950	-21,163,576	42,327,155	42,327,179	$\hat{m} = 1.61 \times 10^6$	[1016667.06, 1607560.29]	$\hat{s} = 1.02 \times 10^6$	[1015054.41, 1017887.2]
norm	1,325,950	-21,750,706	43,501,417	43,501,441	$\hat{\mu} = 2.17 \times 10^6$	[3220115.06, 2161041.21]	$\hat{\sigma} = 3.22 \times 10^6$	[3216296.98, 3223985.48]

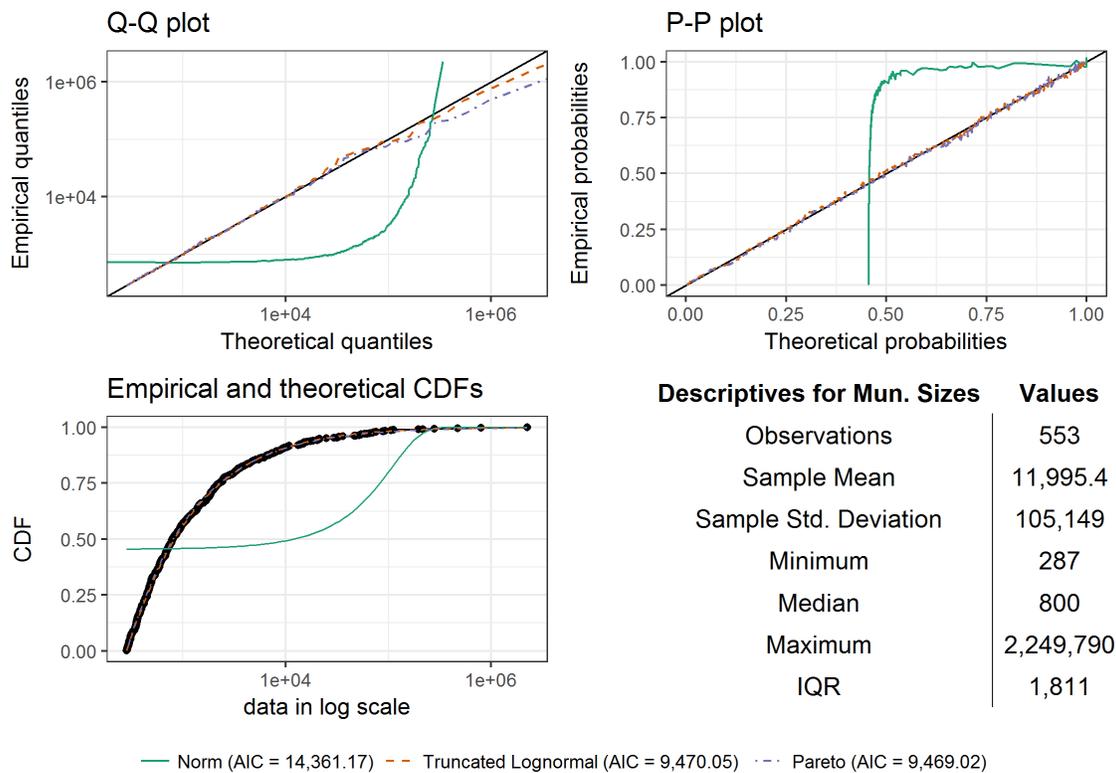


Figure 9: Diagnostic graphical comparison for the distributions of municipality sizes (with sizes above $n_{\min} = 287$), fitted by a truncated-lognormal, a Pareto, and a normal distributions, along with some descriptive statistics. Distributions that fit well the data should line up with the black solid line in the Q-Q and P-P plots. Clearly, the normal distribution (green line) is not a good fit for the distribution of municipality sizes. Ultimately, the relative best fit among many alternative distributions is given by the smallest AIC, according to which the Pareto distribution is the preferred model for Colombian municipality sizes.

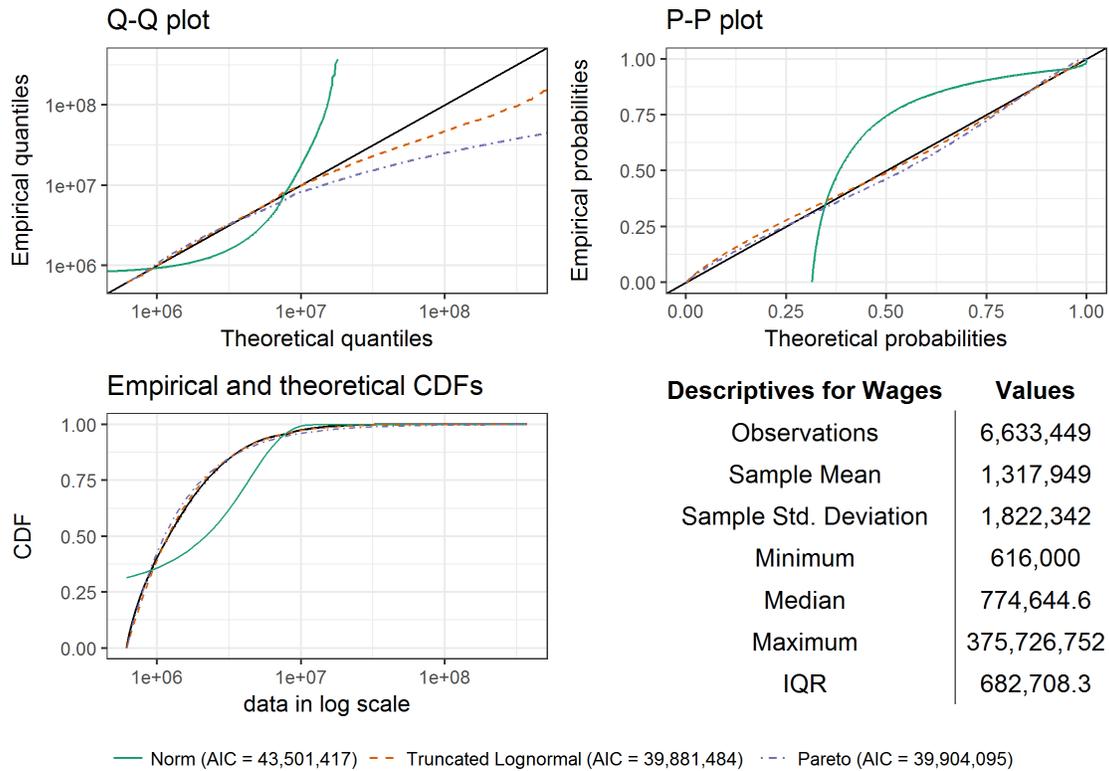


Figure 10: Diagnostic graphical comparison for the distributions of individual monthly wages (for workers living in municipalities with sizes above $n_{\min} = 287$), fitted by a truncated-lognormal, a Pareto, and a normal distributions, along with some descriptive statistics. Distributions that fit well the data should line up with the black solid line in the Q-Q and P-P plots. Clearly, the normal distribution (green line) is not a good fit for the distribution of monthly wages across workers. Ultimately, the relative best fit among many alternative distributions is given by the smallest AIC, according to which the (truncated) log-normal distribution is the preferred model for monthly wages among Colombian formal workers.