# Urban scaling with censored data

Inês Figueira[1,2◑], Rayan Succar[1,2◑], Roni Barak Ventura[1,2], Maurizio Porfiri[1,2,3]

**1** Center for Urban Science and Progress, New York University Tandon School of Engineering, Brooklyn, New York, United States of America
**2** Department of Mechanical and Aerospace Engineering, New York University Tandon School of Engineering, Brooklyn, New York, United States of America
**3** Department of Biomedical Engineering, New York University Tandon School of Engineering, Brooklyn, New York, United States of America

◑These authors contributed equally to this work.
*mporfiri@nyu.edu

## Abstract

In the realm of urban science, scaling laws are essential for understanding the relationship between city population and urban features, such as socioeconomic outputs. Ideally, these analyses would utilize complete datasets; however, researchers often face challenges related to data availability and reporting practices, resulting in datasets that include only the highest observations of urban features (top-$k$). A key question that emerges is: Under what conditions can an analysis based solely on top-$k$ observations accurately determine whether a scaling relationship is truly superlinear or sublinear? To address this question, we conduct a numerical study to explore how relying exclusively on reported values can lead to erroneous conclusions, revealing a selection bias that favors sublinear over superlinear scaling. In response, we develop a method that provides robust estimates of the minimum and maximum potential scaling exponents when only top-$k$ observations are available. We apply this method to two case studies involving firearm violence, a domain notorious for its suppressed datasets, and demonstrate how this approach offers a reliable framework for analyzing scaling relationships with censored data.

## Author summary

Over the past two decades, urban scaling has become essential for understanding the rural-urban continuum by quantifying how urban characteristics evolve with a city's population size. For example, more populous cities are expected to have more patents and wages per capita, but fewer gas stations and road surfaces. Nonetheless, access to incomplete datasets about urban features systematically skews the conclusions derived from this theory. This issue is particularly relevant for features related to health outcomes, which are regularly obtained from partially censored datasets. For instance, data on firearms in the United States remain inaccessible to the public. To address this limitation, we developed a framework that enables urban researchers to draw reliable conclusions about urban scaling, even when dealing with censored datasets. We demonstrate this framework with data on firearm homicide and the number of firearms recovered by authorities in American cities.

# 1 Introduction

Scaling laws are ubiquitous in nature, describing many of the phenomena and processes that surround us. A scaling law describes the behavior of a system through a power-law, connecting certain properties of the system with its size [1]. Scaling laws have been instrumental in characterizing relationships across a wide range of domains, including biological and physical systems. For example, Kleiber's law illustrates how metabolic rates of organisms scale with their body mass [2]. Likewise, scaling laws in the field of ecology indicate that the number of species supported by an ecosystem relates to its area [3]. In the ideal gas law, scaling describes the relationships between pressure, volume, temperature, and the number of molecules [4].

As urbanization rates are ever-increasing [5], understanding scaling of urban features with city population is critical to urban science, management, and planning. Many scaling relationships between the population of a city $X$ and urban feature $Y$ have been documented, which have led to the development of urban scaling theory. Given $N$ cities, an urban scaling law takes the form of $Y_i = CX_i^{\beta} e^{\varepsilon_i}$, with $i = 1, ..., N$, where $C$ is a common baseline, $\beta$ is the scaling exponent that illustrates how an urban feature varies with city size, $e$ is the Napier's constant, and $\varepsilon_i$ represents the deviation of city $i$ from its nominal behavior [6]. The scaling parameters $C$ and $\beta$ are typically computed by logarithmically transforming the scaling law to $\ln Y_i = \ln C + \beta \ln X_i + \varepsilon_i$ and fitting a linear model [6].

Researchers have shown that urban features can scale differently with population size, reflecting systematic relationships across urban and societal metrics. Empirical studies demonstrate that socioeconomic features such as GDP, property values, patents, homicides, and violent crimes exhibit a superlinear dependence on city population ($\beta > 1$) [5–12], meaning that larger (smaller) cities exhibit higher (lower) rates of these features per capita. In contrast, the space occupied by urban infrastructure such as roads, cables and built area scales sublinearly with city population ($0 < \beta < 1$) [13, 14]. Household and individual needs like total employment, housing, and water consumption, instead, typically show a linear dependency on city population ($\beta = 1$) [5, 15].

Over the years, several studies have refined urban scaling and expanded its framework to address methodological limitations. For example, Bettencourt *et al.* distinguished cross-sectional from temporal scaling to capture temporal dynamics beyond pure scale effects [16]. Cross-sectional scaling compares cities at a fixed point in time, whereas temporal scaling tracks changes within cities but can be unstable in cities with slow or negative growth. Finance and Cottineau addressed the issue of null observations in cities during scaling analysis [17]. Although these values may be valid (for example, a city where no patents were filed), the standard practice was to remove them, as the logarithm of zero is undefined [18]. The authors explored alternative methods to ordinary least squares (OLS) for fitting urban models to avoid the exclusion of zero counts. Xiao and Gong argued that spatial dependencies exist between cities that are geographically proximate [19]. They designed a spatial filtering method to account for such dependencies in urban scaling and found that models that do not account for spatial interactions may overestimate GDP in developed regions and underestimate it in underdeveloped ones. In spite of the great strides made in the growing field of urban scaling, the vast majority of existing analyses assume access to a complete data set when fitting the model.

When working with city-level data, access to complete datasets becomes a common challenge. One cause of incomplete data is the obligation of government agencies to prevent the identifiability of sensitive information. For example, the Centers for Disease Control and Prevention Wide-Ranging Online Data for Epidemiological Research (CDC WONDER) publishes data on the underlying causes of death among United States (U.S.) citizens. They provide the yearly counts of each cause of death at the resolution of the entire country, states, and counties. However, to protect individuals' privacy, the agency suppresses counts of nine and lower. Hence, urban scaling research on causes of death in the U.S. are difficult to perform. Similarly, the Tiahrt Amendments [20] impose restrictions on the reporting of data by the U.S. Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF), limiting the disclosure of trace data related to firearms used in crimes to the public. Instead of sharing complete data, the ATF is only allowed to report limited information, such as the top ten cities in each state with the highest number of gun recoveries and the total number of firearms recovered in that state. For both the CDC WONDER and ATF cases, data is censored because they fall below a certain threshold, a situation known as "left-censoring". Such data censoring poses a serious challenge to urban scaling studies on firearm recoveries in the U.S.

Data on cities may also be incomplete due to "missingness", where data points are not available because they are not recorded. The reasons underlying missing data are commonly known as "missing data mechanisms". These mechanisms, as described in [21], fall into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data MCAR occurs when there is no relationship between whether a data point is missing and any values in the dataset, either missing or observed. When the probability of a missing value is dependent on other observed variables but not the value itself, it is considered MAR. In the case of MNAR, the missingness is systematically related to unobserved data or factors not measured by the researcher. For instance, for the CDC WONDER or ATF datasets, data are missing not at random as they are not available when falling below a certain threshold.

Various methods have been devised to address the issue of incomplete data. Recent methodological research [22, 23] has focused on maximum likelihood estimation (MLE) [24, 25], Bayesian estimation [26, 27], and multiple imputation [28, 29]. However, most advanced statistical imputation methods mainly aim at imputing MCAR and MAR and are not suitable for MNAR [30]. Some statistical methods have also been developed for regression analyses when data are MNAR, such as the Tobit model and its variations [31], Powell quantile estimators [32], or othe nonparametric estimators [33]. While effective, these methods are quite general and fail to utilize key information provided by the reporting entity that may be accessible to researchers (for example, the sum of the censored data). Moreover, in the context of urban scaling, the primary focus of a model is whether scaling is superlinear or sublinear, making the precise value of a scaling exponent less critical than its bounds.

In this paper, we aim to address censored data in the context of urban scaling. We focus on data related to firearms and mortality, only available for the highest ("top-$k$") observations due to privacy reasons. We propose a rigorous, yet simple, method tailored for urban scaling analysis that estimates scaling behavior. Along with the top-$k$ observations, the method incorporates the total counts of the feature across the dataset in the form of a constraint, taking advantage of the aggregated observations reported in existing datasets. By solving an optimization problem, we bound the regression slope by providing its minimum and the maximum possible values. This approach not only simplifies the estimation process compared to existing methods, but also provides robust bounds necessary for determining whether an urban feature scales superlinearly or sublinearly. Our method offers a powerful tool for urban researchers, ensuring reliable assessment of scaling behaviors even when working with incomplete data.

In the following, we first conduct numerical simulations using both complete and incomplete synthetic datasets to explore how the use of incomplete data could bias the estimation of scaling laws. We then present an algorithm that iteratively distributes missing values to unknown cities. We apply the developed framework to two case studies. In the first, we inspect suppressed data on firearm homicides from CDC

WONDER and complete data from National Center for Health Statistics' (NCHS) Restricted-Use Vital Statistics Data. We compare the estimates of the scaling exponent when using the incomplete and complete data and validate our $\beta$-bounding method. In the second case study, we apply the bounding method on the partially reported data to conclude whether firearms recovered by the ATF follow a superlinear or sublinear scaling. Our results demonstrate the value of this bounding process in the study of urban scaling laws when datasets suffer from censored observations.

# 2 Results

## 2.1 Assessing bias in urban scaling due to censored data

As a first step to understand how incomplete data can bias the estimation of scaling laws and the inference of superlinearity and sublinearity, we conduct a numerical study using both complete and incomplete synthetic datasets. We simulate the typical case of health-related outcomes where data are only available for a subset of $k$ cities with the highest value of the urban feature reported (top-$k$), and no other information is given regarding other cities except for the total value of the outcome variable in larger spatial units (as reported by CDC WONDER and ATF).

We aim to quantify the deviation of the estimated regression slope $\hat{\beta}^{\mathrm{k}}$ (where a hat refers to an estimated value and superscript $k$ denotes the known partial data) from the true value $\beta$ due to censored data. To this end, we compute the error of the estimation $(\hat{\beta}^{\mathrm{k}} - \beta)$ over a range of changes to key factors that could impact the estimation of $\beta$, including the true scaling law exponent ($\beta$), proportion of known data (top-$k$%), standard deviation of the error ($\sigma$), and complete dataset size ($N$). In addition, we consider two distributions for the population data: normal and log-normal. We generate random synthetic observations while systematically varying these parameters in a factorial design (see Methods for details).

First, by using censored data, we find that the error of the estimation of $\beta$ can be relatively high, and similar for different values of $\beta$ (Fig 1A). Interestingly, we find that the error of the estimation is asymmetric and biased toward sublinear scaling, such that one is more likely to infer a sublinear scaling relationship although a truly superlinear one exists. This asymmetry is engendered by the selection of the top-$k$ cities based on their urban feature (Fig 1B). Specifically, the top-$k$ cities are more likely to have a positive residual with respect to the linear fit on the complete dataset, so that considering only them leads to underestimation of the scaling exponent. In agreement with our expectations, we find that regardless of the population distribution (normal or log-normal) or the value of $\beta$, the magnitude of the error tends to increase as the percent of known data becomes smaller (Fig 2A-B), and as the standard deviation of the noise increases (Fig 2C-D). The error does not change with the size of the complete dataset (Fig 2E-F), although we notice that for larger datasets, the variance of the estimator decreases. Such a decrease does not guarantee the consistency of $\hat{\beta}^{\mathrm{k}}$ (see Section A of S1 Appendix).

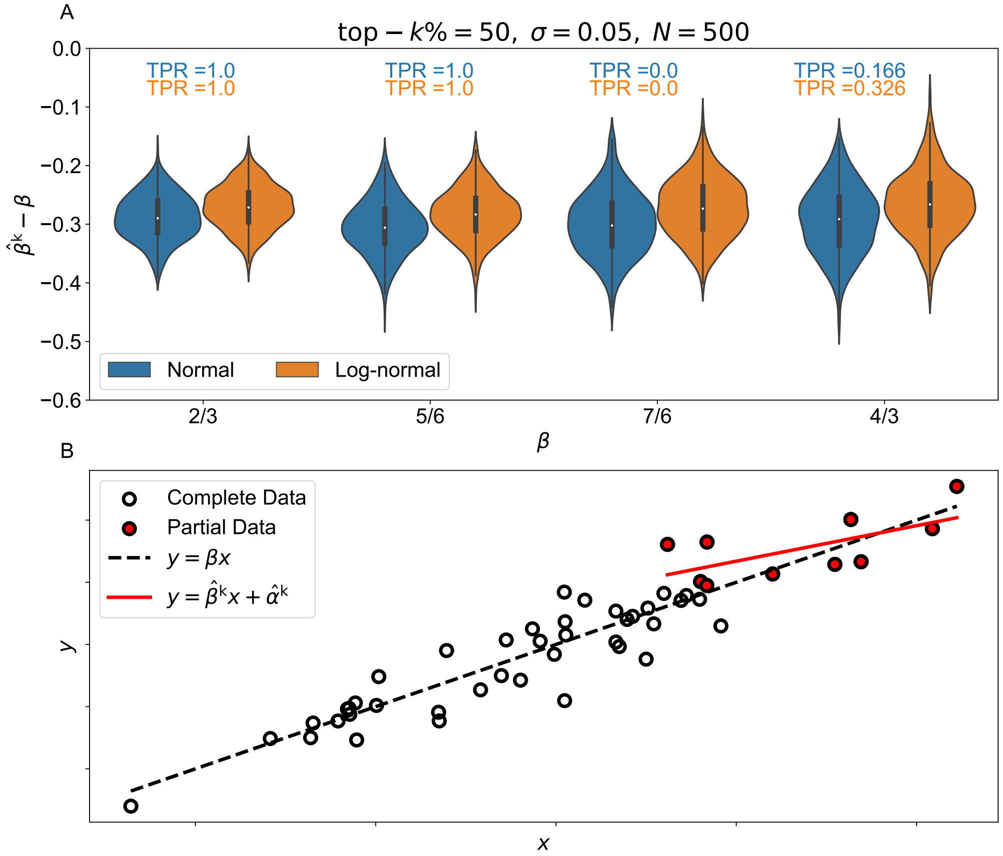

**Fig 1. Bias in estimating the urban scaling exponent with censored data.** (A) Assessment of the estimate of the scaling exponent ($\hat{\beta}^{\mathrm{k}}$) from data generated using a true scaling law ($Y_i = X_i^{\beta} e^{\varepsilon_i}$ for $i = 1, \cdots, N$) with $X$ following either a normal distribution (blue) or log-normal (orange), as a function of the true scaling exponent. The proportion of known data points is selected based on the $k$-highest percent value of the response variables $Y$. The violin plots represent the distribution of the error, while the boxes inside represent the first (Q1) and third (Q3) quartiles, and their whiskers extend to 1.5 times the interquartile range from Q1 and Q3. Each violin plot contains 500 data points. For each violin plot, we also report the true positive rate (TPR) for the inference of sublinear ($\beta < 1$) and superlinear ($\beta > 1$) scaling. (B) Illustration of the reason for bias towards sublinear scaling discovered in (A). Using a censored dataset that only uses the top values of a selected urban feature (red filled circles) incorrectly discounts observations in the complete dataset (open circles) that have negative residual with respect to the true fit (black dashed line), thereby leading to biased model estimation (red solid line).

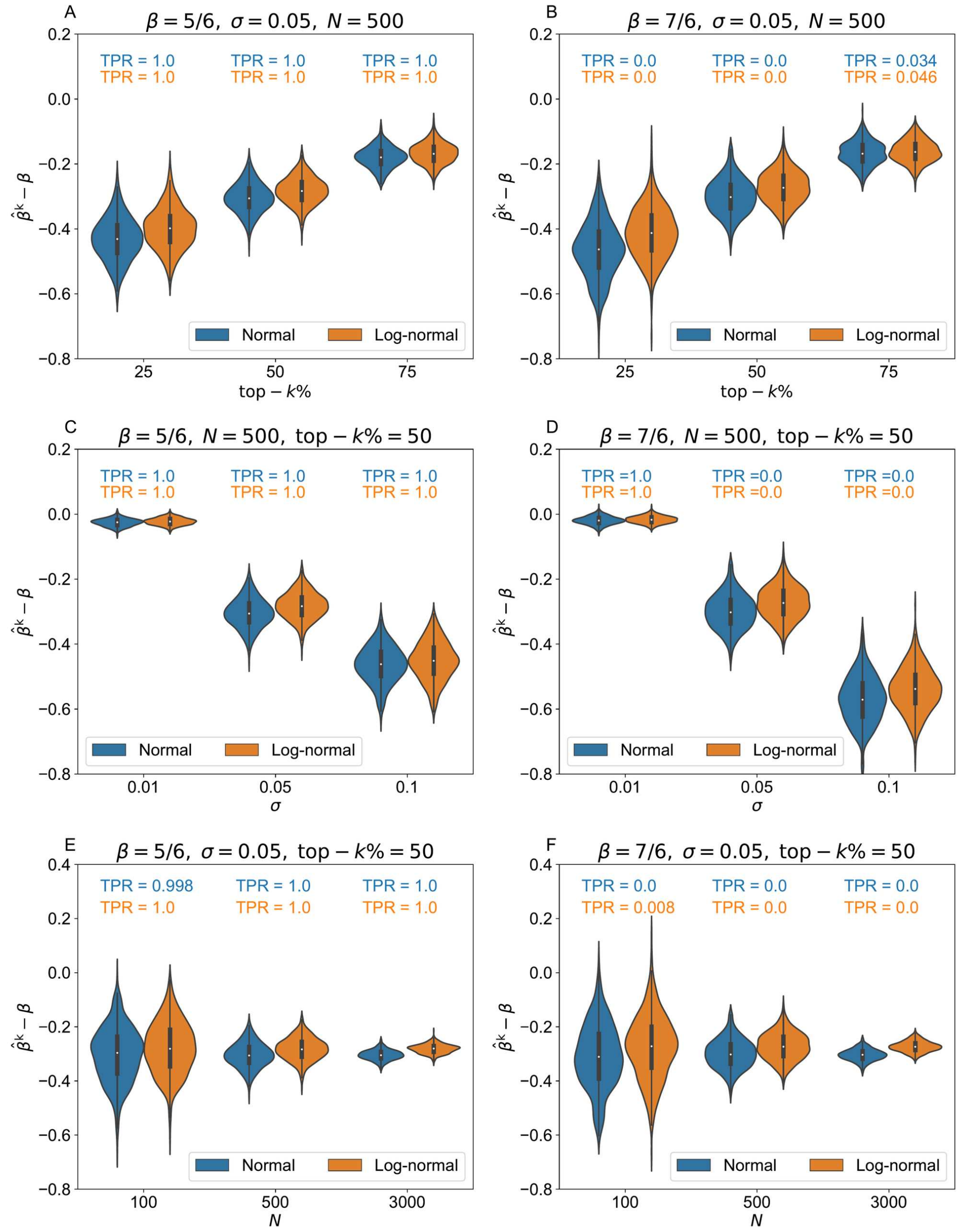

**Fig 2. Factors influencing bias in estimating the urban scaling exponent with censored data.** Assessment of the estimate of the scaling exponent ($\hat{\beta}^{\mathrm{k}}$) from data generated using a true scaling law ($Y_i = X_i^{\beta} e^{\varepsilon_i}$ for $i = 1, \cdots, N$, and $\beta = 5/6$ or $\beta = 7/6$) with $X$ following either a normal distribution (blue) or log-normal (orange), as a function of (A-B) proportion of known data, (C-D) standard deviation of the true error, and (E-F) complete dataset size. The proportion of known data points is selected based on the $k$-highest percent value of the response variables $Y$. The violin plots represent the distribution of the error, while the boxes inside represent the first (Q1) and third (Q3) quartiles, and their whiskers extend to 1.5 times the interquartile range from Q1 and Q3. Each violin plot contains 500 data points. For each violin plot, we also report the true positive rate (TPR) for the inference of sublinear ($\beta = 5/6$) and superlinear ($\beta = 7/6$) scaling.

In all of the simulations, we consider whether regressing with incomplete data causes urban scaling classification errors by looking at the true positive rate (TPR) for true superlinear and sublinear scaling relationships (Fig 1 and Fig 2). The TPR measures the proportion of sublinear (superlinear) cases correctly identified by a model as such, allowing us to evaluate the performance of hypothesis testing regarding sublinear or superlinear dependence on population. For instance, in the case of true superlinear scaling relations, the TPR represents the proportion of correctly identified superlinear relations when only a certain proportion of the data is known ($\beta > 1$ and $\hat{\beta}^{\mathrm{k}} > 1$; see Methods). Due to the asymmetry in the errors ($\hat{\beta}^{\mathrm{k}}$ is underestimated), we find that the TPR for superlinear scaling is less than that for sublinear scaling, potentially being as low as zero.

In Section B of S1 Appendix, we present results in Fig 2 for $\beta = 2/3$ and $4/3$ where similar trends are observed. We also show the relationship between the error in the estimation of the scaling exponent when using censored data and the coefficient of determination of the censored data estimation $(R^{\mathrm{k}})^2$, where we see that the higher $(R^{\mathrm{k}})^2$, the lower the bias.

## 2.2 Greedy algorithm to bound the scaling exponent

We devise a general bounding framework that uses a greedy optimization to estimate the minimum and maximum possible scaling exponents, $\hat{\beta}_{\mathrm{min}} \leq \hat{\beta} \leq \hat{\beta}_{\mathrm{max}}$. By computing these bounds, we aim to reach a more reliable conclusion about a scaling behavior, while effectively addressing the biases encountered when using OLS on the censored data. Within a system of $N$ cities, we address the case in which the researcher has only access to urban measurements in a subsystem of $k < N$ cities, and the total count of the urban feature $\mathcal{S}$ across all $N$ cities. In order to find the upper bound of the scaling exponent ($\hat{\beta}_{\mathrm{max}}$), we solve the constrained optimization problem

$$\hat{\beta}_{\mathrm{max}} = \max_{\mathbf{Y}^{\mathrm{uk}}} \{ f_{\beta}\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{uk}}\right) \mid \mathcal{S} = \sum_{i=1}^{k} Y_i^{\mathrm{k}} + \sum_{i=k+1}^{N} Y_i^{\mathrm{uk}}, Y_{\mathrm{min},i} \leq Y_i^{\mathrm{uk}} \leq Y_{\mathrm{max},i} \}, \quad (1)$$

where the column vector $\mathbf{X} = [X_1, \cdots, X_N]^T$ contains the population sizes of all $N$ cities, $\mathbf{Y}^{\mathrm{k}} = [Y_1^{\mathrm{k}}, \cdots, Y_k^{\mathrm{k}}]^T$ comprises the $k$ known values of the urban feature, $\mathbf{Y}^{\mathrm{uk}} = [Y_{k+1}^{\mathrm{uk}}, \cdots, Y_N^{\mathrm{uk}}]^T$ consists of the $N - k$ unknown values for which we are optimizing. Similar to city population data, we also consider the urban features to be positive integer numbers. We denote vectors and matrices in bold and use $T$ for matrix transpose. The function $f_{\beta}\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{uk}}\right)$ represents the OLS estimator of the scaling exponent (for further details, see Methods).

In this greedy approach, we pose that the sum of $Y_i^{\mathrm{k}}$ and $Y_i^{\mathrm{uk}}$ over $i$ is equal to the total of the urban feature $\mathcal{S}$. In addition, we constrain $Y_i^{\mathrm{uk}}$ between $Y_{\mathrm{min,i}}$ and $Y_{\mathrm{max,i}}$, the values of which will depend on the reporting and censoring process. The lower bound of the scaling exponent ($\hat{\beta}_{\mathrm{min}}$) can be written equivalently to Eq (7) (see Methods), with "min" instead. Once obtained, the upper and lower bounds can be used to verify the validity of inferences based on partial datasets. In fact, $\hat{\beta}_{\mathrm{max}} < 1$ will offer backing to the inference of sublinear scaling and $\hat{\beta}_{\mathrm{min}} > 1$ to the inference of superlinear scaling. Some insight into the optimal $\mathbf{Y}^{\mathrm{uk}}$ can be garnered by linearizing the objective function and solving the optimization problem analytically (see Section C of S1 Appendix). Such an analysis suggests that bigger cities should be assigned values close to $Y_{\mathrm{max,i}}$ and smaller cities values close to $Y_{\mathrm{min,i}}$, thereby maximizing the contrast between them.

## 2.3 Case studies of urban scaling with censored data

To demonstrate the value of the our bounding scheme in urban research, we apply it to two real datasets with partial observations: firearm homicides from the CDC and firearms recovered by the ATF. In the CDC case study, we obtained access to the uncensored dataset from the National Center for Health Statistics (NCHS) [34] allowing us to validate the scaling conclusions. Such privilege is not granted with the ATF study case. Applying our framework to these datasets, we not only gain a deeper understanding of firearm-related violence and crimes in the U.S., but also demonstrate how this optimization process can be generalized to other censored datasets for estimating scaling laws.

### 2.3.1 Firearm homicides

Similar to Bettencourt *et al.* [16], we perform cross-sectional scaling of firearm homicides with population for U.S. cities, over the five-year period between 2016 and 2020 (Fig 3). The results are presented for cities, encompassing both Metropolitan Statistical Areas (MSAs) and Micropolitan Statistical Areas (MicroSAs). While urban scaling relations are highly sensitive to the spatial boundaries defining a city [35], there is no standardized definition for a city in the U.S. Consequently, both MSAs and MicroSAs are commonly used as functional cities in analyses [36].

Urban scaling for firearm homicides in the U.S. exhibits a power-law relation with city population using both the censored and complete datasets. Using a censored dataset leads to the inference of a sublinear relationship across all years, with the true exponent being consistently underestimated $\hat{\beta}^{\mathrm{k}} < \hat{\beta}$ (Fig 3). With the complete dataset, $\hat{\beta}$ reflects a strictly sublinear relationship for all years, except in the year 2020. In this year, when the reported MSAs and MicroSAs account for about three quarters of the total firearm homicides, $\hat{\beta} = 0.967$, with a 95% confidence interval of $[0.921; 1.013]$ (Table 1). Given the confidence interval, we cannot reject the hypothesis that $\beta = 1$. We also note that the coefficient of determination of the complete model ($R^2$) is larger than that of the partial data ($(R^{\mathrm{k}})^2$), indicating that using OLS regression on the complete dataset could yield better-fitted results (Table 1).

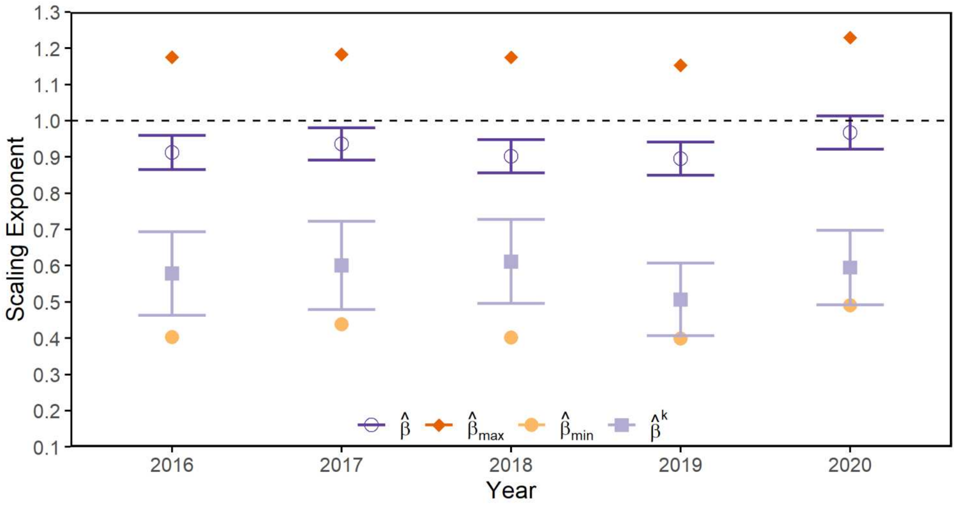

**Fig 3. Urban scaling exponent of firearm homicides in the U.S. MSAs and MicroSAs (2016–2020).** Yellow dots and orange diamonds represent the minimum ($\hat{\beta}_{\text{min}}$) and maximum ($\hat{\beta}_{\text{max}}$) scaling exponent, respectively, obtained by implementing the optimization strategy on the reported data (CDC suppresses firearm homicides in cities where there are fewer than ten incidents). These serve as bounds for the actual $\hat{\beta}$ (dark purple open circles) and $\hat{\beta}^{\text{k}}$ (light purple squares) obtained using only the reported data; horizontal lines (whiskers) denote the limit of the 95% confidence interval. The horizontal dashed line represents the limit above which the scaling relation is superlinear.

**Table 1. Results on urban scaling exponent for firearm homicides in the U.S. MSAs and MicroSAs from 2016 to 2020, using suppressed and complete data.**

| Year | Firearm Homicides - MSA and MicroSA | | | | | | |
|---|---|---|---|---|---|---|---|
| | top-$k$% | $\hat{\beta}^{\text{k}}$ | $(R^{\text{k}})^2$ | $\hat{\beta}$ | $R^2$ | $\hat{\beta}_{\text{min}}$ | $\hat{\beta}_{\text{max}}$ |
| 2016 | 79.5 | 0.578 [0.463; 0.694] | 0.615 | 0.912 [0.865; 0.959] | 0.6869 | 0.402 | 1.175 |
| 2017 | 79.8 | 0.601 [0.479; 0.723] | 0.610 | 0.936 [0.891; 0.981] | 0.7087 | 0.439 | 1.183 |
| 2018 | 79.3 | 0.611 [0.496; 0.727] | 0.635 | 0.902 [0.856; 0.948] | 0.6863 | 0.402 | 1.175 |
| 2019 | 80.4 | 0.506 [0.406; 0.606] | 0.600 | 0.896 [0.850; 0.941] | 0.6863 | 0.399 | 1.153 |
| 2020 | 76.7 | 0.594 [0.492; 0.697] | 0.597 | 0.967 [0.921; 1.013] | 0.7026 | 0.490 | 1.229 |

The second column shows the ratio of reported firearm homicides (top-$k$%). The third and fourth columns provide the $\hat{\beta}^{\text{k}}$ and its adjusted $(R^{\text{k}})^2$. The fifth column presents $\hat{\beta}$ estimates for the complete data along with the adjusted $R^2$. The last two columns refer to the minimum and maximum bounds for $\hat{\beta}$, $\hat{\beta}_{\text{min}}$ and $\hat{\beta}_{\text{max}}$, computed using only the censored data reported by the CDC.

We apply our bounding scheme assuming each suppressed county had between one and nine counts of homicide. Our results indicate that $\hat{\beta}_{\text{max}} > 1$ and $\hat{\beta}_{\text{min}} < 1$ across all years so that when working with partial data, one should be prudent in interpreting their results (Fig 3 and Table 1). In particular, the fact that the upper bound is always greater than 1 indicates that one should not exclude the possibility that their inference based on partial data is incorrect. This is the case for the year 2020, when partial data would yield $\hat{\beta}^{\text{k}} = 0.594$ with confidence $[0.492; 0.697]$ and real data are instead supportive of a linear scaling $\hat{\beta} = 0.967$, with a 95% confidence interval of $[0.921; 1.013]$.

### 2.3.2 Recovered firearms

In the second case study, we investigate the scaling of firearms recovered across the U.S. in 2022 with city population. These yearly data are made available by the ATF, where the top-$k$ cities per state with the most firearms recovered are reported, along with the total number of firearms recovered in the entire state. Using only the reported values, it is difficult to conclude whether firearms recoveries scale sublinearly or superlinearly with population across the U.S states. The small sample size (10 cities for each state except Vermont and Washington) does not allow for precise estimation, resulting in wide confidence intervals (Table 2).

**Table 2. Estimates of the scaling exponent ($\hat{\beta}^{\mathrm{k}}$) for recovered firearms in each state of the U.S. (except of Hawaii) and the District of Columbia (D.C.), for the year 2022 based on ATF reported data, along with the corresponding bounds from the optimization.**

| State | top-$k$% | $\hat{\beta}^{\mathrm{k}}$ | $\hat{\beta}_{\min}$ | $\hat{\beta}_{\max}$ | $\hat{\beta}_{\min}$ (95% CI) | | $\hat{\beta}_{\max}$ (95% CI) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Method 1 | Method 2 | Method 1 | Method 2 |
| *Alabama | 61.9 | 0.431 [0.184; 0.679] | -0.472 | 0.625 | [-0.473; -0.472] | [-0.468; -0.464] | [0.625; 0.626] | [0.617; 0.622] |
| Alaska | 89.5 | 0.767 [0.038; 1.495] | - | - | - | - | - | - |
| Arizona | 78.2 | 0.225 [-0.546; 0.996] | -0.437 | 1.071 | [-0.426; -0.420] | [-0.405; -0.393] | [1.050; 1.060] | [1.024; 1.037] |
| *Arkansas | 79.2 | 0.893 [0.054; 1.732] | 0.013 | 0.146 | [0.013; 0.013] | [-0.042; -0.032] | [0.147; 0.147] | [0.194; 0.204] |
| California | 35.1 | 0.352 [0.005; 0.699] | -1.098 | 1.324 | [-1.098; -1.096] | [-1.083; -1.077] | [1.324; 1.326] | [1.306; 1.312] |
| *Colorado | 70.5 | 0.933 [0.506; 1.360] | -0.388 | 0.587 | [-0.389; -0.388] | [-0.391; -0.385] | [0.587; 0.587] | [0.583; 0.589] |
| *Connecticut | 73.6 | 1.167 [0.475; 1.859] | -0.275 | 0.591 | [-0.275; -0.274] | [-0.279; -0.270] | [0.589; 0.590] | [0.585; 0.594] |
| *Delaware | 77.9 | 0.338 [-0.022; 0.698] | -0.127 | 0.697 | [-0.121; -0.117] | [-0.108; -0.100] | [0.650; 0.659] | [0.632; 0.643] |
| D.C | 1.0 | - | - | - | - | - | - | - |
| *Florida | 47.1 | 0.677 [0.224; 1.129] | -0.758 | 0.868 | [-0.759; -0.758] | [-0.765; -0.761] | [0.868; 0.868] | [0.870; 0.874] |
| *Georgia | 48.7 | 0.243 [-0.070; 0.555] | -0.675 | 0.747 | [-0.676; -0.675] | [-0.666; -0.660] | [0.747; 0.748] | [0.732; 0.738] |
| Idaho | 68.0 | -0.028 [-0.417; 0.361] | -0.114 | 0.313 | [-0.114; -0.113] | [-0.105; -0.097] | [0.314; 0.314] | [0.297; 0.305] |
| *Illinois | 58.7 | 0.645 [0.288; 1.001] | -0.290 | 0.401 | [-0.290; -0.289] | [-0.187; -0.173] | [0.401; 0.402] | [0.269; 0.284] |
| *Indiana | 78.0 | 0.237 [-0.129; 0.603] | -0.208 | 0.327 | [-0.208; -0.208] | [-0.171; -0.158] | [0.327; 0.328] | [0.272; 0.287] |
| Iowa | 69.1 | 1.682 [0.376; 2.988] | - | - | - | - | - | - |
| Kansas | 80.0 | 0.852 [0.110; 1.594] | - | - | - | - | - | - |
| *Kentucky | 69.8 | 0.550 [0.295; 0.805] | -0.393 | 0.499 | [-0.394; -0.393] | [-0.336; -0.327] | [0.501; 0.502] | [0.437; 0.446] |
| *Louisiana | 58.3 | 0.706 [0.420; 0.992] | -0.637 | 0.825 | [-0.637; -0.637] | [-0.634; -0.63] | [0.825; 0.826] | [0.819; 0.822] |
| Maine | 79.3 | -0.335 [-1.094; 0.423] | - | - | - | - | - | - |
| *Maryland | 40.6 | 0.440 [-0.12; 0.999] | -0.668 | 0.846 | [-0.647; -0.637] | [-0.608; -0.595] | [0.792; 0.805] | [0.754; 0.769] |
| *Massachusetts | 39.8 | 0.170 [-0.232; 0.571] | -0.550 | 0.701 | [-0.551; -0.550] | [-0.560; -0.553] | [0.700; 0.701] | [0.703; 0.711] |
| *Michigan | 79.8 | 0.266 [-0.324; 0.857] | -0.037 | 0.158 | [-0.037; -0.037] | [-0.118; -0.105] | [0.160; 0.161] | [0.237; 0.253] |
| Minnesota | 56.8 | 0.612 [-0.073; 1.296] | - | - | - | - | - | - |
| *Mississippi | 47.6 | 0.168 [-0.306; 0.641] | -0.534 | 0.648 | [-0.535; -0.534] | [-0.537; -0.531] | [0.647; 0.648] | [0.645; 0.650] |
| *Missouri | 78.2 | 0.371 [-0.150; 0.893] | -0.127 | 0.250 | [-0.127; -0.126] | [-0.132; -0.120] | [0.251; 0.251] | [0.241; 0.255] |
| *Montana | 76.6 | 0.387 [-0.004; 0.778] | -0.003 | 0.314 | [-0.003; -0.003] | [-0.001; 0.009] | [0.316; 0.316] | [0.304; 0.313] |
| Nebraska | 89.5 | 0.997 [0.657; 1.338] | - | - | - | - | - | - |
| *Nevada | 97.3 | 0.617 [0.141; 1.093] | 0.102 | 0.833 | [0.119; 0.125] | [0.217; 0.231] | [0.772; 0.784] | [0.698; 0.712] |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| New Hampshire | 74.5 | 0.371 | [-0.19; 0.931] | - | - | - | - | - | - |
| *New Jersey | 44.6 | 0.529 | [0.094; 0.965] | -0.515 | 0.694 | [-0.514; -0.514] | [-0.515; -0.509] | [0.693; 0.694] | [0.688; 0.695] |
| *New Mexico | 86.8 | 0.777 | [0.216; 1.338] | -0.193 | 0.495 | [-0.192; -0.191] | [-0.149; -0.138] | [0.498; 0.500] | [0.449; 0.459] |
| *New York | 68.3 | 0.674 | [0.310; 1.038] | -0.264 | 0.445 | [-0.264; -0.263] | [-0.239; -0.221] | [0.445; 0.446] | [0.391; 0.411] |
| *North Carolina | 51.0 | 0.730 | [0.398; 1.061] | -0.648 | 0.781 | [-0.649; -0.649] | [-0.642; -0.639] | [0.782; 0.782] | [0.772; 0.775] |
| North Dakota | 86.8 | 0.475 | [0.065; 0.886] | - | - | - | - | - | - |
| *Ohio | 65.4 | 0.646 | [0.307; 0.985] | -0.419 | 0.540 | [-0.419; -0.419] | [-0.429; -0.422] | [0.539; 0.540] | [0.542; 0.550] |
| *Oklahoma | 83.7 | 0.908 | [0.369; 1.446] | -0.031 | 0.173 | [-0.031; -0.030] | [-0.075; -0.063] | [0.175; 0.175] | [0.209; 0.221] |
| *Oregon | 60.4 | 1.037 | [0.589; 1.485] | -0.417 | 0.645 | [-0.417; -0.417] | [-0.413; -0.407] | [0.645; 0.646] | [0.635; 0.641] |
| *Pennsylvania | 55.7 | 0.290 | [-0.144; 0.724] | -0.461 | 0.564 | [-0.461; -0.459] | [-0.385; -0.369] | [0.563; 0.565] | [0.471; 0.488] |
| Rhode Island | 86.2 | 0.420 | [-0.312; 1.151] | -0.078 | 1.063 | [-0.044; -0.035] | [-0.003; 0.015] | [0.968; 0.985] | [0.928; 0.949] |
| *South Carolina | 58.1 | 0.159 | [-0.414; 0.731] | -0.527 | 0.730 | [-0.527; -0.526] | [-0.526; -0.520] | [0.729; 0.730] | [0.724; 0.729] |
| South Dakota | 73.4 | 0.390 | [-0.049; 0.828] | - | - | - | - | - | - |
| *Tennessee | 78.1 | 1.171 | [0.602; 1.739] | -0.512 | 0.687 | [-0.512; -0.511] | [-0.502; -0.492] | [0.688; 0.688] | [0.668; 0.679] |
| *Texas | 62.2 | 0.874 | [0.561; 1.188] | -0.621 | 0.751 | [-0.622; -0.621] | [-0.611; -0.606] | [0.751; 0.752] | [0.736; 0.741] |
| *Utah | 54.2 | 1.053 | [0.264; 1.843] | -0.365 | 0.631 | [-0.366; -0.365] | [-0.364; -0.360] | [0.631; 0.632] | [0.626; 0.630] |
| Vermont | 63.9 | 0.286 | [-0.023; 0.595] | - | - | - | - | - | - |
| *Virginia | 57.3 | 0.698 | [0.143; 1.253] | -0.634 | 0.809 | [-0.635; -0.634] | [-0.637; -0.632] | [0.809; 0.809] | [0.806; 0.811] |
| *Washington | 57.7 | 0.580 | [0.252; 0.907] | -0.446 | 0.654 | [-0.446; -0.446] | [-0.442; -0.438] | [0.654; 0.654] | [0.646; 0.650] |
| *West Virginia | 54.2 | 0.141 | [-0.208; 0.490] | -0.323 | 0.448 | [-0.323; -0.322] | [-0.324; -0.315] | [0.448; 0.449] | [0.440; 0.449] |
| *Wisconsin | 71.2 | 0.708 | [0.149; 1.266] | -0.147 | 0.366 | [-0.147; -0.147] | [-0.099; -0.086] | [0.368; 0.369] | [0.289; 0.304] |
| Wyoming | 72.4 | 0.026 | [-0.491; 0.544] | - | - | - | - | - | - |

ATF annually reports the top-$k$ cities per state where the most firearms are recovered, along with the total number of firearms recovered in the state. Here, top-$k$% refers to the ratio of firearms recovered in those $k$ cities relative to the total number of firearms recovered in the entire state. Of the 49 states, it is not possible to compute $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ for 11. This issue arises because, in these 11 states, the number of cities not in top-$k$ exceeds the number of recovered firearms there, indicating some of them have zero firearms recovered. The optimization procedure used to compute $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ operates under the assumption that every city within a state has at least one firearm recovered. When this assumption is violated - that is, when there are cities with zero recovered firearms — these bounds cannot be computed. The estimated 95% confidence interval of each bound, are computed using two different methods (see Methods for details).

To address this issue and bound the exponent $\hat{\beta}$, we apply the developed optimization algorithm with the assumption that each city has at least one firearm recovered. For 11 of the 49 states (all states except of Hawaii, see Methods), it is not possible to apply the optimization scheme since the number of cities other than the reported top-$k$ exceeds the number of recovered firearms outside of the top-$k$ cities, violating the underlying assumption. Out of the remaining 38 states, only three (Arizona, California, and Rhode Island) have $\hat{\beta}_{\max} > 1$. Therefore, we cannot reject the hypothesis of superlinearity or linearity for these states. For the remaining states, $\hat{\beta}_{\max} < 1$, indicating a sublinear behavior of firearm recoveries with respect to city population.

Figure 4 shows the bounds for the scaling relation when considering the combined 38 states and the District of Columbia (D.C.), where $\hat{\beta}_{\min} = -0.284$ and $\hat{\beta}_{\max} = 0.556$, reflecting the trend of sublinearity in the country. For this case study, we numerically explore the global optimality of the solution through exhaustive perturbations (see Section D of S1 Appendix).

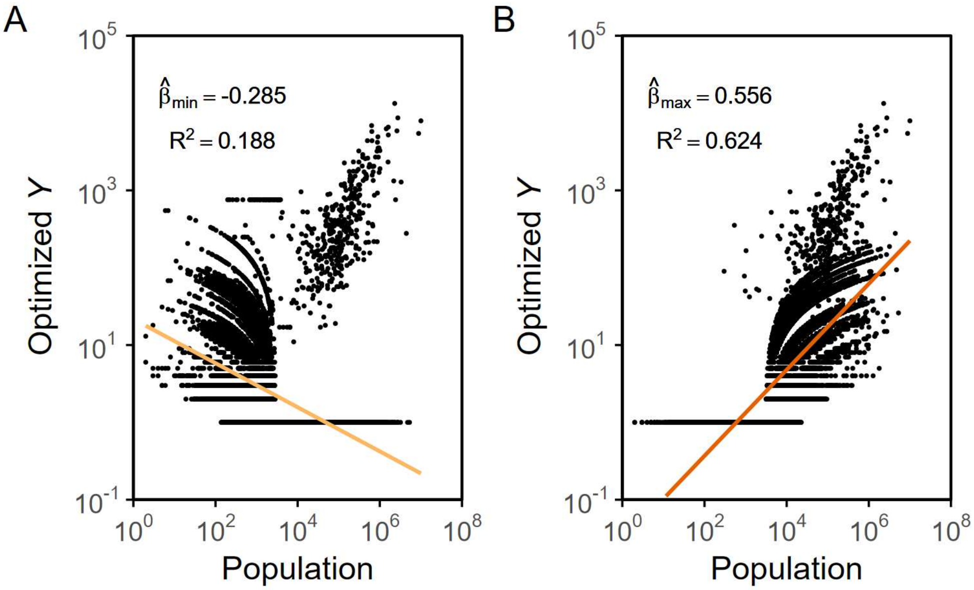

**Fig 4. Urban scaling results for recovered firearms in the U.S. in 2022 after optimization.** The dots identify the optimal number of recovered firearms as a function of the population in 28,970 Census Incorporated Places and Minor Civil Divisions. The number of unknown recovered firearms in each of the 38 states and D.C. was optimally distributed among the different states to compute the minimum (A) and the maximum (B) scaling exponent $\beta$. All places were assumed to have at least one recovered firearm. Of the 49 states, it was not possible to compute $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ for 11. This issue arises because, in these 11 states, the number of cities not in top-$k$ exceeds the number of recovered firearms there, indicating some of them had zero firearms recovered. The bounding procedure used to compute $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ operates under the assumption that every city within a state has at least one firearm recovered. When this assumption is violated, these bounds cannot be computed.

## 2.4 Sensitivity analysis

The proposed bounds may be prone to error due to noise in the data. In order to assess the robustness of these bounds, we conduct Monte Carlo simulations to estimate 95% confidence intervals in the ATF dataset. We perform two variations of the simulations. The first assumes the sum of simulated values are within 5% difference with respect to the real data, and the second that the top-$k$% of the synthetic data matches the real. Each of these methods preserves different characteristics of the data (see Methods for details) and allows us to estimate the 95% confidence intervals of $\hat{\beta}_{\text{min}}$ and $\hat{\beta}_{\text{max}}$. We observe narrow confidence intervals for both simulations, indicating robustness of the bounding scheme (Table 2). For all states where $\hat{\beta}_{\text{max}} < 1$, the confidence intervals are below 1, reinforcing our claim of sublinear scaling (Table 2). For Rhode Island, despite $\hat{\beta}_{\text{max}} > 1$, the confidence interval using different methods are below 1, indicating potential ambiguity in the scaling interpretation for this state.

# 3 Discussion

Urban scaling is a fundamental tool used in urban science, yielding interesting power laws that capture the relationship between urban features and city population. Ideally, urban scaling needs a complete dataset to derive accurate scaling exponents; however, legal and ethical considerations often lead to censoring of data, thereby presenting significant challenges to the estimation of urban scaling relationships. Censored data differently affect cities as a function of their count of an urban feature, whereby small cities are more prone to be characterized by smaller value of some urban features, potentially below the minimum that agencies can share with the public.

In numerical simulations, we explore five factors that could impact the estimation of the scaling exponent. Our results indicate that two factors critically affect the estimation of scaling exponents: the proportion of known data and the variance of the noise. While the role of these factors in the estimation of scaling is intuitive as both determine the quality of a dataset, we also find that scaling exponents are consistently underestimated. Therefore, one is more likely to correctly infer a sublinear relationship and fail to infer a superlinear one. Arguably, performing OLS fitting using a top-$k$ dataset leads to systematic underestimation of the scaling exponent. For sufficiently dispersed datasets (ones with high noise), the cities experiencing the largest values of the urban feature under investigation may not be the most populous ones. Thus, a linear model with only the top-$k$ cities could omit cities with large populations but values of the urban feature lower than the top-$k$. These cities have a negative residual with respect to the fit on the complete dataset; discarding them will lead to underestimating the scaling exponent. In real datasets, such a discrepancy may also result from data segmentation, where different population segments have been found to exhibit different scaling behaviors [37–39]. To address the biases that result from censored data, we devise a bounding method that determines the minimum and maximum possible scaling exponents, and apply it to two case studies.

The first case study focuses on the scaling of firearm homicides over the five-year period between 2016 and 2020, where we compare the performance of a left-censored dataset against that of a complete one. For the complete dataset, we find a sublinear relationship for all years except for 2020, when the COVID-19 pandemic started and an increase in firearm purchases and violence has been documented [40, 41]. In 2020, the data are, in fact, indicative of a linear scaling. Using the left-censored dataset, we are not able to recover such a change in time, whereby we consistently register sublinear scaling of firearm homicides for all the years. Our bounding scheme successfully casts a doubt on the validity of the sublinear trend. In fact, our lower bound is below one and our upper bound is above one, so that prudence is needed when drawing conclusion on the scaling exponent with partial data. Interestingly, aggregating the data over multiple years to mitigate zero counts may not resolve the issue of data missingness in the scaling. First, aggregation could skew the inference towards superlinear scaling, by systematically under counting firearm homicides in small cities without affecting the counting in large cities. Second, the aggregation does not capture time trends in the scaling, such as the one observed herein due to the COVID-19 pandemic (see Section E of S1 Appendix). Both these factors are likely the reasons for which several studies support firearm homicides to be more frequent in urban rather than rural settings [12, 42–45].

In the second case study, we investigate the scaling of recovered firearms in the year 2022. In the absence of a complete dataset, the fit of an OLS model produces extremely wide confidence intervals, ranging from negative to values greater than one in some cases. Thus, conclusive interpretations of scaling behavior become virtually impossible. However, implementing our proposed bounding scheme allows to shed light on scaling of firearm recoveries. Our results support the sublinear scaling of firearm recoveries in the U.S., hinting that firearms might be more prevalent in rural areas. This notion aligns with the sublinear behavior of firearm ownership and federal firearm-selling licenses reported by Succar and Porfiri [12]. Similarly, a recent Pew Research Center survey has shown that 46% of people who reside in rural areas reported themselves as firearm owners, compared to 19% of people who live in urban areas [46]. The observed sublinear scaling in firearm recoveries could also be attributed to varying strategies for tracking and recovering firearms across different jurisdictions. The Tiahrt Amendment prohibits federal agencies from creating searchable firearm databases, making the ATF's firearm recovery efforts extremely inefficient [47]. Under these circumstances, records of completed firearm sales have become invaluable for regional law enforcement, especially when maintained and retained permanently in a central database. For instance, handgun sales records in California are stored in a state Department of Justice database, enabling law enforcement agencies to swiftly trace the ownership of handguns recovered in crimes [48]. California is also one of the three states where we observe a potential superlinear relationship between city population and the number of firearms recovered by the ATF ($\hat{\beta}_{\max} = 1.324$). Additionally, it is tenable that recovering firearms in smaller cities is easier than in larger ones due to familiarity among locals [49, 50], their investment in creating a safe environment through community policing [51, 52], and higher trust and cooperation between citizens and authorities [51].

While both study cases demonstrate its value in firearm research, our bounding method could also be implemented in domains other than urban science. For example, recent work suggests that metabolic rates of eusocial systems scale sublinearly with the mass of a colony [53, 54]. Yet, practical limitations have hindered validation of this proposition across a wide range of colony sizes, as measurements of metabolic rates for small colonies are difficult to capture by typical respirometry apparatuses. Similarly, performing experiments on large colonies is challenged by housing requirements in the laboratory. As such, data in these metabolic studies are left- and right-censored. Our approach could help overcome those data limitations by bounding the scaling exponents of partial datasets and inferring the metabolism laws of colonies. Another possible application is in the field of environmental studies, where concentration of pollutants or chemicals is often left-censored because analytical instruments have detection limits below which pollutants cannot be accurately measured [55]. Instead of reporting an exact concentration, values below the detection limit are commonly recorded as "less than" the limit or the percent detected [56, 57]. By applying our approach to these left-censored datasets, environmental scientists could bound the scaling exponents that describe the relationship between chemical concentrations and various environmental factors.

Our study has five significant limitations. First, in the numerical simulations we consider the residuals to be normally distributed. This assumption may not always be appropriate [58], hindering the generalization of our conclusions to scenarios where the errors do not follow a normal distribution. The second limitation concerns the acquisition of data on city populations in the firearm recoveries case study. The ATF does not have a consistent definition of a city. While most of the cities included in the top-$k$ list correspond to census incorporated places and minor civil divisions, some do not. This is the case for 22 areas, such as Eagle River in Alaska (a community within the Municipality of Anchorage). In our analysis we consider only census incorporated places and minor civil divisions [59], thereby excluding these 22 other areas. This inconsistency in defining cities complicates the analysis, making it challenging to accurately define all possible cities not mentioned in the top-$k$ list. The third limitation relates to our bounding method's assumption that there is at least one observation in each city. For the ATF case study, we were unable to bound the scaling exponent for 11 states because the number of cities not included in the top-$k$ list exceeds the number of firearms recovered in those areas. This limitation could be addressed in a future study by combining the proposed problem with the work of Finance and Cottineau [17] that employ estimation techniques to handle datasets with zero counts so that our bounding method accounts for the possibility of zero observations. Fourth, our approach assumes that cities are independent of each other in line with classical urban scaling theory. As a result, we apply standard OLS for the estimation. We envision integrating our approach with the one proposed by Xiao and Gong [19] to account for spatial interactions between cities, by generalizing the objective function of our optimization. Finally, the proposed optimization framework based on a greedy algorithm was developed for scaling with specific cases, which may limit the generalization of the algorithm to other problems. These problems may include datasets with large variances or a high number of outliers, different types of constraints, or scaling that requires estimators other than OLS.

In conclusion, our work identifies a potential flaw in the current use of partial data to draw conclusions about scaling relationships in urban data. We offer compelling evidence that censored data may lead to inaccurate predictions of scaling exponents, where sublinear relationships could be erroneously identified as superlinear ones. We put forward a simple methodology to bound the scaling exponent from censored observations, based on the solution of a constrained optimization problem that assumes absence of zeros in the dataset and leverages information on the sum of all counts. We propose that future reporting of urban scaling relationships in technical papers (especially sublinear ones) include explicit information about the number of inaccessible data points along with an estimation of the expected effect of such a data missingness. The latter can be pursued through the implementation of a bounding scheme like the one proposed in this work (when possible) or stress tests on the scaling exponent through Monte Carlo simulations.

# 4 Methods

## 4.1 Urban scaling law

Given $N$ cities, an urban scaling law is a relationship between some urban feature of interest and the city population of the form

$$Y_i = C X_i^{\beta} e^{\varepsilon_i} \tag{2}$$

where $i = 1, \ldots, N$, $Y_i$ and $X_i$ are the urban feature and population size for city $i$, $C$ is a common baseline, and $\varepsilon_i$ is the deviation of city $i$ from its nominal behavior. This scaling law can be written in linear form [15],

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{3}$$

where we introduce log-transformed variables $y_i = \ln Y_i$, $x_i = \ln X_i$, and $\alpha = \ln C$. Since urban scaling relations are linear on the log-log scale, we can estimate the parameters of the scaling relationship by using OLS, which minimizes the sum of squared errors [6]. Such a minimization yields

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{4}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{5}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated values of $\alpha$ and $\beta$, and $\bar{x}$ and $\bar{y}$ are the averages of the log-transformed population size and urban feature, respectively. One of the limitations of OLS regression is that it requires complete data for all variables included in the model to ensure unbiased estimation. If there are missing data points, OLS may result in biased and unreliable regression coefficients [21].

## 4.2 Assessing bias in urban scaling due to censored data

The synthetic data are simulated according to a true scaling law, $Y = X^{\beta}e^{\varepsilon}$, $\varepsilon \sim \mathcal{N}(0, \sigma)$, with zero intercept ($\ln C = 0$). This true scaling law serves as a baseline for comparing estimated scaling laws when using $\hat{\beta}^{\mathrm{k}}$, allowing us to quantify the bias more accurately. The synthetic data are generated to account for different scenarios of scaling that could be produced by a real dataset. Specifically, we identify the population distribution, the true slope ($\beta$), standard deviation of the error ($\sigma$), size of the dataset ($N$), and the proportion of known data points (top-$k$%) out of the dataset as parameters that could meaningfully alter the estimation of the scaling exponent.

To assess whether the marginal distribution of $X$ affects the estimated $\hat{\beta}^{\mathrm{k}}$, two population distributions are considered, normal ($\mathcal{N}(10^5, 10^4)$) and log-normal ($\mathcal{LN}(10, 0.1)$), and sampled using *numpy* (version 1.26.4; [60]). For each population distribution, we employ a factorial design varying the other four parameters: $\beta \in \{2/3, 5/6, 7/6, 4/3\}$, $\sigma \in \{0.01, 0.05, 0.1\}$, $N \in \{100, 500, 3000\}$, and top$-k\% \in \{25\%, 50\%, 75\%\}$. The values of $\beta$ were selected based on the literature on urban scaling laws, which have helped identify typical scaling exponents as a function of the city organization and type of urban feature [13, 61]. In total, the factorial design for each distribution contains 108 combinations (216 in total).

For each possible combination in the factorial design, linear regression is performed on the subset of the known top$-k\%$ data points to obtain a value of $\hat{\beta}^{\mathrm{k}}$. We simulate the experiment on the entire design 500 times, totaling $108,000$ observations. To further assess how the bias resulting from using censored data affects the estimation of scaling relationships, we also look at the TPR of real superlinear and sublinear scaling relations. Specifically, for sublinear scaling cases ($\beta < 1$), we consider estimates as true only when the upper bound of $\hat{\beta}^{\mathrm{k}} < 1$. Similarly, for superlinear cases ($\beta > 1$), we consider estimates as true only when the lower bound of $\hat{\beta}^{\mathrm{k}} > 1$. TPR is then computed as the fraction of correct estimates out of all 500 estimates.

### 4.3 Greedy algorithm to bound the scaling exponent

The bounding method consists of optimizing over $\mathbf{Y}^{\mathrm{uk}}$ to estimate the scaling exponent using the OLS estimator derived from Eq (4), denoted as

$$\begin{aligned} f_\beta\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{uk}}\right) = & \frac{N\left(\sum_{i=1}^{k} \ln X_i \ln Y_i^{\mathrm{k}} + \sum_{i=k+1}^{N} \ln X_i \ln Y_i^{\mathrm{uk}}\right)}{N\sum_{i=1}^{N}\left(\ln X_i\right)^2 - \left(\sum_{i=1}^{N} \ln X_i\right)^2} \\ & - \frac{\sum_{i=1}^{N} \ln X_i \left(\sum_{i=1}^{k} \ln Y_i^{\mathrm{k}} + \sum_{i=k+1}^{N} \ln Y_i^{\mathrm{uk}}\right)}{N\sum_{i=1}^{N}\left(\ln X_i\right)^2 - \left(\sum_{i=1}^{N} \ln X_i\right)^2}. \end{aligned} \tag{6}$$

To find the maximum or minimum regression slopes, we construct the unknown observations $Y_i^{\mathrm{uk}}$. As an initial step, we assign $Y_i^{\mathrm{uk}} = Y_{\min,i}$, where $Y_{\min,i} \geq 1$ in accordance with the assumption that all cities must have non-zero values for their feature, which may be violated in reality. To find the $\mathbf{Y}^{\mathrm{uk}}$ entries that result in $\hat{\beta}_{\max}$ ($\hat{\beta}_{\min}$), we iteratively increase the value of each entry by one, without surpassing $Y_{\max,i}$, and seek the largest increase (decrease) of $f_\beta$. In other words, for each iteration over the entries of $\mathbf{Y}^{\mathrm{uk}}$, we compare the values of $f_\beta$ for all updated entries and identify the entry that results in the largest (or smallest) value of $\hat{\beta}$. If two or more entries produce the same result for $f_\beta$, the algorithm will select the first entry that appears in the order of iteration. We end the process when the sum of known and unknown values matches $\mathcal{S}$. This greedy scheme is detailed in Algorithms 1 and 2. To gain a better intuition about the procedure, we describe it using the following equation:

$$\boldsymbol{\mathcal{Y}}^t = \boldsymbol{\mathcal{Y}}^{t-1} + \underset{\mathbf{e}\in\xi}{\operatorname{argmax}}\{f_\beta\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \boldsymbol{\mathcal{Y}}^{t-1} + \mathbf{e}\right) \mid Y_{\min,i} \leq \mathcal{Y}_i^{t-1} \leq Y_{\max,i}\}, \tag{7}$$

for $t = 1, \cdots, t_f$ , where $t_f = \mathcal{S} - \sum_{i=1}^{k} Y_i^{\mathrm{k}} - \sum_{i=k+1}^{N} \mathcal{Y}_i^0$, $\boldsymbol{\mathcal{Y}}^0 = [Y_{\min,k+1}, \cdots, Y_{\min,N}]^T$, and $\xi$ is the set of all standard basis vectors of length $N-k$, that is, $\{[1,0,...,0]^{\mathrm{T}}, ..., [0,0,...,1]^{\mathrm{T}}\}$. The maximum regression slope is found during the last iteration, $\hat{\beta}_{\max} = f_\beta\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \boldsymbol{\mathcal{Y}}^{t_f}\right)$. The optimization in Eq (7) is executed through exhaustive search, that is, searching over the entire set $\xi$.

**Algorithm 1** Greedy algorithm to find $\hat{\beta}_{\max}$

```
Input: $\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}_{\min}, \mathbf{Y}_{\max}, f_\beta, \mathcal{S}$
Output: $\hat{\beta}_{\max}$
Initialization: $\mathbf{Y}^{\mathrm{uk}} \leftarrow \mathbf{Y}_{\min}, sum\_unknown \leftarrow \mathcal{S} - \sum_{i=1}^{k} Y_i^{\mathrm{k}}$
for $iteration$ from 1 to $sum\_unknown - \sum_{i=k+1}^{N} Y_{\min,i}$ do
  $\hat{\beta}_{\max} \leftarrow -\infty$
  $i_{\max} \leftarrow 0$
  for $i$ from $k+1$ to $N$ do
    $\mathbf{Y}^{\mathrm{aux}} \leftarrow \mathbf{Y}^{\mathrm{uk}}$
    $Y_i^{\mathrm{aux}} \leftarrow Y_i^{\mathrm{aux}} + 1$
    Ensure: $Y_i^{\mathrm{aux}} \leq Y_{\max,i}$
    if $f_\beta\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{aux}}\right) > \hat{\beta}_{\max}$ then
      $\hat{\beta}_{\max} \leftarrow f_\beta\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{aux}}\right), i_{\max} \leftarrow i$
    end if
  end for
  $Y_{i_{\max}}^{\mathrm{uk}} \leftarrow Y_{i_{\max}}^{\mathrm{uk}} + 1$
end for
return $\hat{\beta}_{\max}$
```

**Algorithm 2** Greedy algorithm to find $\hat{\beta}_{\min}$

```
**Input:** $\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}_{\min}, \mathbf{Y}_{\max}, f_{\beta}, \mathcal{S}$
**Output:** $\hat{\beta}_{\min}$
**Initialization:** $\mathbf{Y}^{\mathrm{uk}} \leftarrow \mathbf{Y}_{\min}, sum\_unknown \leftarrow \mathcal{S} - \sum_{i=1}^{k} Y_i^{\mathrm{k}}$
**for** $iteration$ from 1 to $sum\_unknown - \sum_{i=k+1}^{N} Y_{\min,i}$ **do**
  $\hat{\beta}_{\min} \leftarrow \infty$
  $i_{\min} \leftarrow 0$
  **for** $i$ from $k+1$ to $N$ **do**
    $\mathbf{Y}^{\mathrm{aux}} \leftarrow \mathbf{Y}^{\mathrm{uk}}$
    $Y_i^{\mathrm{aux}} \leftarrow Y_i^{\mathrm{aux}} + 1$
    **Ensure:** $Y_i^{\mathrm{aux}} \leq Y_{\max,i}$
    **if** $f_{\beta}\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{aux}}\right) < \hat{\beta}_{\min}$ **then**
      $\hat{\beta}_{\min} \leftarrow f_{\beta}\left(\mathbf{X}, \mathbf{Y}^{\mathrm{k}}, \mathbf{Y}^{\mathrm{aux}}\right), i_{\min} \leftarrow i$
    **end if**
  **end for**
  $Y_{i_{\min}}^{\mathrm{uk}} \leftarrow Y_{i_{\min}}^{\mathrm{uk}} + 1$
**end for**
**return** $\hat{\beta}_{\min}$
```

## 4.4 Case studies of urban scaling with censored data

### 4.4.1 Firearm homicides

Firearm homicide data are obtained from the CDC WONDER database and NCHS's Restricted-Use Vital Statistics Database. For both data sets, we query for incidents of firearm homicides using the following ICD-10 Codes: X93 (Assault by handgun discharge), X94 (Assault by rifle, shotgun, and larger firearm discharge), and X95 (Assault by other and unspecified firearm discharge). We filter the data for years between 2016 and 2020, and group the results by year and county. Population counts in each county are returned with the query.

We conduct scaling analyses for each year, at the level of MSA and MicroSA. We begin with an OLS regression on logarithmically transformed variables to compute $\hat{\beta}^{\mathrm{k}}$ from the left-censored dataset and $\hat{\beta}$ from the complete dataset. The bounds $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ for $\hat{\beta}$ are estimated using the greedy optimization algorithm described earlier for the censored data. Cities with null values are removed from the analysis.

For the MSA and MicroSA level analysis, we first convert county level data to MSAs and MicroSAs. We rely on the U.S. Bureau of Labor Statistics' Quarterly Census of Employment and Wages County-MSA-CSA Crosswalk [62] to aggregate counts of firearm homicides in counties to MSAs and MicroSAs, based on county codes. The total number of homicides ($\mathcal{S}$) in all the MSAs and MircroSAs is also reported. After grouping the counties into their respective MSA or MicroSA, we take the ones that do not have any suppressed counties and construct the vector $\mathbf{Y}^{\mathrm{k}}$. Each element of $\mathbf{Y}^{\mathrm{uk}}$ consists of an MSA/MicroSA that has at least one suppressed county. Let $h_i$ represent the total homicides reported in MSA/MicroSA $i$, and $sc_i$ represent the number of suppressed counties. Within each MSA/MicroSA $i$, the entries of $\mathbf{Y}^{\mathrm{uk}}$ are constrained by $Y_{\min,i} = h_i + sc_i$ and $Y_{\max,i} = h_i + 9sc_i$, since the CDC suppresses values between one and nine for each county, while reporting the counties with zero homicides. For example, if MSA/MicroSA $i$ has 3 suppressed counties and 14 homicides reported in total, we constrain the entries of $i$ in the range 17 to 41, corresponding to one or nine homicides in each of the suppressed counties.

### 4.4.2 Recovered firearms

For the analyses of the total number of firearms recovered, we manually collect data from the "U.S. Firearms Trace Data by State" provided by the ATF [63]. The dataset includes the total number of firearms recovered and traced by state in 2022, along with the top-$k$ cities in terms of recoveries within each state ($k$=10 for all states, except for Vermont and Washington where $k$=15 and 11, respectively). Due to limited data on population size in its cities, Hawaii is excluded from this analysis. The population data are collected from the Census "Incorporated Places and Minor Civil Divisions Datasets" [59]. Although there is no standardized definition for a city in the U.S. [36], the cities included in this dataset encompass various administrative divisions such as incorporated places, minor civil divisions, and census-designated places, among others, leading to an inconsistent definition of what constitutes a city.

The scaling estimates $\hat{\beta}^{\mathrm{k}}$ for each state are calculated using OLS regression on logarithmically transformed data. The bounds $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ could not be computed for 11 states because they have more cities than recovered firearms, indicating that in some places no firearm is recovered. This situation is not accounted for in the algorithms because we assume that the number of firearms recovered in each of the unknown cities is between one and the smallest of the top-$k$, that is, $Y_{\min,i} = 1$ and $Y_{\max,i} = \min(\mathbf{Y}^{\mathrm{k}})$.

To extend the scaling analyses to the entire U.S., we must account for the fact that each state has a different total number of recoveries. We re-define $\mathcal{S}$ in Equation (7) as

$$\mathcal{S} = \sum_{j=1}^{G} \left( \sum_{i=1}^{k_j} Y_{j,i}^{\mathrm{k}} + \sum_{i=k_j+1}^{N_j} Y_{j,i}^{\mathrm{uk}} \right). \tag{8}$$

Here, the observations are organized into $G$ states such that there are $k_j$ reported cities out of the total of $N_j$ cities in state $j$. We apply the optimization algorithm with vectors $\mathbf{Y}^{\mathrm{k}}$ and $\mathbf{Y}^{\mathrm{uk}}$ being constructed by stacking each state's $\mathbf{Y}_j^{\mathrm{k}}$ and $\mathbf{Y}_j^{\mathrm{uk}}$, respectively, where $j = 1, ..., 38$ is the index of each state. We constrain the elements of $\mathbf{Y}_j^{\mathrm{uk}}$ so that $1 \leq Y_{j,i}^{\mathrm{uk}} \leq \min(\mathbf{Y}_j^{\mathrm{k}})$, where we account for the different states having different constraints depending on the top cities reported.

## 4.5 Sensitivity analysis

To investigate the effects of small perturbations on the optimal bounds, we compute the 95% confidence intervals for the ATF case study for each state separately. We rely on Monte Carlo simulation to estimate the variance and the confidence intervals. Specifically, for each simulation, we generate a set data points that resembles the known real data reported by the ATF by sampling $k$ synthetic urban features from the power-law distribution $\widetilde{Y}_i^{\mathrm{k}} = \exp(\hat{\alpha}^{\mathrm{k}}) X_i^{\hat{\beta}^{\mathrm{k}}} \exp(\varepsilon_i)$, with $\varepsilon \sim \mathcal{N}(0, \sigma)$. Here, $\hat{\alpha}^{\mathrm{k}}$, $\hat{\beta}^{\mathrm{k}}$, and $\hat{\sigma}^{\mathrm{k}}$ are the parameters estimated from the real known data. Each time we sample the vector $\widetilde{\mathbf{Y}}^{\mathrm{k}}$, we optimize accordingly to obtain a distribution for $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$. We estimate the variance of $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ from $1,000$ realizations of the Monte Carlo simulations using the *var(·)* function in R. Assuming the distributions of $\hat{\beta}_{\min}$ and $\hat{\beta}_{\max}$ to be Gaussian, we compute the confidence intervals using the standard normal approximation, which calls for scaling the standard error by 1.96 [64].

We note that sampling $\widetilde{\mathbf{Y}}^{\mathrm{k}}$ according to the estimated power law may not preserve the sum of the unknown values. We propose two methods to address this issue. In Method One, we disregard samples with more than a 5% difference with respect to the sum of the known data, specifically

$$\frac{\left| \sum_{i=1}^{k} Y_i^{\mathrm{k}} - \sum_{i=1}^{k} \widetilde{Y}_i^{\mathrm{k}} \right|}{\sum_{i=1}^{k} Y_i^{\mathrm{k}}} \leq 0.05. \tag{9}$$

Hence, while optimizing, we assume that the sum of the entries of $\widetilde{\mathbf{Y}}^{\mathrm{uk}}$ equals the difference between the reported total and the sum of our generated indicators,

$$\sum_{i=k+1}^{N} \widetilde{Y}_i^{\mathrm{uk}} = \mathcal{S} - \sum_{i=1}^{k} \widetilde{Y}_i^{\mathrm{k}}. \tag{10}$$

In Method Two, we posit that

$$\sum_{i=k+1}^{N} \widetilde{Y}_i^{\mathrm{uk}} = \frac{\sum_{i=1}^{k} \widetilde{Y}_i^{\mathrm{k}}}{\sum_{i=1}^{k} Y_i^{\mathrm{k}}} - \sum_{i=1}^{k} \widetilde{Y}_i^{\mathrm{k}}, \tag{11}$$

retaining all samples and ensuring that the top-$k$% of the generated known data matches the real one,

$$\frac{\sum_{i=1}^{k} \widetilde{Y}_i^{\mathrm{k}}}{\widetilde{\mathcal{S}}} = \frac{\sum_{i=1}^{k} Y_i^{\mathrm{k}}}{\mathcal{S}}. \tag{12}$$

## Acknowledgments

The authors would like to thank Christopher Buglino for useful discussion on the methodology and for porting the optimization algorithm from R to Rcpp.

## Code and data availability

Codes and datasets are available on the Dynamical Systems Laboratory's Github (Link: https://github.com/dynamicalsystemslaboratory/Urban-scaling-with-missing-data, Accession Number Link: https://github.com/dynamicalsystemslaboratory/Urban-scaling-with-missing-data/releases/tag/V2.0). The only dataset not shared in this repository is the NCHS's Restricted-Use Vital Statistics Database: readers who wish to use this dataset can request access from NCHS at https://www.cdc.gov/nchs/nvss/nvss-restricted-data.htm#anchor_1553801903.

## References

1. Barenblatt GI. Scaling. 1st ed. Cambridge: Cambridge University Press; 2003.

2. West GB, Brown JH, Enquist BJ. A general model for the origin of allometric scaling laws in biology. Science. 1997;276(5309):122–126. doi:10.1126/science.276.5309.122.

3. García Martín H, Goldenfeld N. On the origin and robustness of power-law species–area relationships in ecology. Proc Natl Acad Sci USA. 2006;103(27):10310–10315. doi:10.1073/pnas.0510605103.

4. Sengers JMHL, Greer WL, Sengers JV. Scaled equation of state parameters for gases in the critical region. J Phys Chem. 1976;5(1):1–52. doi:10.1063/1.555529.

5. Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB. Growth, innovation, scaling, and the pace of life in cities. Proc Natl Acad Sci USA. 2007;104(17):7301–7306. doi:10.1073/pnas.0610172104.

6. Bettencourt LM. Introduction to urban science: evidence and theory of cities as complex systems. 1st ed. Cambridge: MIT Press; 2021.

7. Oliveira M. More crime in cities? On the scaling laws of crime and the inadequacy of per capita rankings–a cross-country study. Crime Sci. 2021;10(1):27. doi:10.1186/s40163-021-00155-8.

8. Lobo J, Bettencourt LM, Strumsky D, West GB. Urban scaling and the production function for cities. PLOS One. 2013;8(3):e58407. doi:10.1371/journal.pone.0058407.

9. Alves LG, Ribeiro HV, Lenzi EK, Mendes RS. Distance to the scaling law: a useful approach for unveiling relationships between crime and urban metrics. PLOS One. 2013;8(8):e69580. doi:10.1371/journal.pone.0069580.

10. Meirelles J, Neto CR, Ferreira FF, Ribeiro FL, Binder CR. Evolution of urban scaling: Evidence from Brazil. PLOS One. 2018;13(10):e0204574. doi:10.1371/journal.pone.0204574.

11. Bilal U, de Castro CP, Alfaro T, Barrientos-Gutierrez T, Barreto ML, Leveau CM, et al. Scaling of mortality in 742 metropolitan areas of the Americas. Sci Adv. 2021;7(50):eabl6325. doi:10.1126/sciadv.abl6325.

12. Succar R, Porfiri M. Urban scaling of firearm violence, ownership and accessibility in the United States. Nat Cities. 2024;1(3):216–224. doi:10.1038/s44284-024-00034-8.

13. Bettencourt LM. The origins of scaling in cities. Science. 2013;340(6139):1438–1441. doi:10.1126/science.1235823.

14. Angel S, Parent J, Civco DL, Blei A, Potere D. The dimensions of global urban expansion: Estimates and projections for all countries, 2000–2050. Prog Plann. 2011;75(2):53–107. doi:10.1016/j.progress.2011.04.001.

15. Bettencourt LM, Lobo J. Urban scaling in Europe. J R Soc Interface. 2016;13(116):20160005. doi:10.1098/rsif.2016.0005.

16. Bettencourt LM, Yang VC, Lobo J, Kempes CP, Rybski D, Hamilton MJ. The interpretation of urban scaling analysis in time. J R Soc Interface. 2020;17(163):20190846. doi:10.1098/rsif.2019.0846.

17. Finance O, Cottineau C. Are the absent always wrong? Dealing with zero values in urban scaling. Environ Plan B Urban Anal City Sci. 2019;46(9):1663–1677. doi:10.1177/2399808318785634.

18. Leitao JC, Miotto JM, Gerlach M, Altmann EG. Is this scaling nonlinear? R Soc Open Sci. 2016;3(7):150649. doi:10.1098/rsos.150649.

19. Xiao Y, Gong P. Removing spatial autocorrelation in urban scaling analysis. Cities. 2022;124:103600. doi:10.1016/j.cities.2022.103600.

20. 114 Congress. 114 HR 1449 IH: Tiahrt Restrictions Repeal Act; 2015 [cited 2024 Aug 3]. Available from: `https://www.congress.gov/bill/114th-congress/house-bill/1449`.

21. Little RJ, Rubin DB. Statistical analysis with missing data. vol. 793. 3rd ed. Hoboken, New Jersey: John Wiley & Sons; 2019.

22. Enders CK. Missing data: An update on the state of the art. Psychol Methods. 2023;doi:10.1037/met0000563.

23. Enders CK. Applied missing data analysis. 2nd ed. New York: Guilford Publications; 2022.

24. Savalei V, Falk CF. Robust two-stage approach outperforms robust full information maximum likelihood with incomplete nonnormal data. Struct Equ Modeling. 2014;21(2):280–302. doi:10.1080/10705511.2014.882692.

25. Lüdtke O, Robitzsch A, West SG. Analysis of interactions and nonlinear effects with missing data: A factored regression modeling approach using maximum likelihood estimation. Multivar Behav Res. 2020;55(3):361–381. doi:10.1080/00273171.2019.1640104.

26. Du H, Enders C, Keller BT, Bradbury TN, Karney BR. A Bayesian latent variable selection model for nonignorable missingness. Multivar Behav Res. 2022;57(2-3):478–512. doi:10.1080/00273171.2021.1874259.

27. Levy R, Mislevy RJ. Bayesian psychometric modeling. New York: Chapman and Hall/CRC; 2017.

28. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. Behav Res Methods. 2016;48:640–649. doi:10.3758/s13428-015-0590-3.

29. Chan KW, Meng XL. Multiple improvements of multiple imputation likelihood ratio tests. Stat Sin. 2022;32(3):1489–1514. doi:10.48550/arXiv.1711.08822.

30. Baraldi AN, Enders CK. An introduction to modern missing data analyses. J Sch Psychol. 2010;48(1):5–37. doi:10.1016/j.jsp.2009.10.001.

31. Amemiya T. Advanced econometrics. Cambridge: Harvard university press; 1985.

32. Powell JL. Censored regression quantiles. J Econom. 1986;32(1):143–155. doi:10.1016/0304-4076(86)90016-3.

33. Lewbel A, Linton OB. Nonparametric censored regression; 1998. Available from: `https://elischolar.library.yale.edu/cowles-discussion-paper-series/1434/`.

34. Centers for Disease Control and Prevention. 1999-2020: Underlying Cause of Death by Bridged-Race Categories; 2013 [cited 2024 Aug 15]. Available from: `https://wonder.cdc.gov/Deaths-by-Underlying-Cause.html`.

35. Molinero C, Thurner S. How the geometry of cities determines urban scaling laws. J R Soc Interface. 2021;18(176):20200705. doi:10.1098/rsif.2020.0705.

36. Bettencourt L, West G. A unified theory of urban living. Nature. 2010;467(7318):912–913. doi:10.1038/467912a.

37. Hanley QS, Lewis D, Ribeiro HV. Rural to urban population density scaling of crime and property transactions in English and Welsh parliamentary constituencies. PLOS One. 2016;11(2):e0149546. doi:doi.org/10.1371/journal.pone.0149546.

38. Ribeiro HV, Hanley QS, Lewis D. Unveiling relationships between crime and property in England and Wales via density scale-adjusted metrics and network tools. PLOS One. 2018;13(2):e0192931. doi:10.1371/journal.pone.0192931.

39. Sutton J, Shahtahmassebi G, Ribeiro HV, Hanley QS. Rural–urban scaling of age, mortality, crime and property reveals a loss of expected self-similar behaviour. Sci Rep. 2020;10(1):16863. doi:10.1038/s41598-020-74015-x.

40. Schleimer JP, McCort CD, Shev AB, Pear VA, Tomsich E, De Biasi A, et al. Firearm purchasing and firearm violence during the coronavirus pandemic in the United States: a cross-sectional study. Inj Epidemiol. 2021;8:1–10. doi:10.1186/s40621-021-00339-5.

41. Sun S, Cao W, Ge Y, Siegel M, Wellenius GA. Analysis of firearm violence during the COVID-19 pandemic in the US. JAMA Netw Open. 2022;5(4):e229393–e229393. doi:10.1001/jamanetworkopen.2022.9393.

42. Branas CC, Nance ML, Elliott MR, Richmond TS, Schwab CW. Urban–rural shifts in intentional firearm death: different causes, same results. Am J Public Health. 2004;94(10):1750–1755. doi:10.2105/AJPH.94.10.1750o.

43. Crifasi CK, Merrill-Francis M, McCourt A, Vernick JS, Wintemute GJ, Webster DW. Association between firearm laws and homicide in urban counties. J Urban Health. 2018;95:383–390. doi:10.1007/s11524-018-0273-3.

44. Siegel M, Solomon B, Knopov A, Rothman EF, Cronin SW, Xuan Z, et al. The impact of state firearm laws on homicide rates in suburban and rural areas compared to large cities in the United States, 1991-2016. J Rural Health. 2020;36(2):255–265. doi:10.1111/jrh.12387.

45. Reeping PM, Mak A, Branas CC, Gobaud AN, Nance ML. Firearm death rates in rural vs urban US counties. JAMA Surg. 2023;158(7):771–772. doi:10.1001/jamasurg.2023.0265.

46. Parker K, Horowitz JM, Igielnik R, Oliphant JB, Brown A. America's complex relationship with guns. Pew Research Center's Social and Demographic Trends Project. Pew Research Center.; 2017 Jun 22 [cited 2024 Aug 16]. Available from: `https://www.pewresearch.org/social-trends/2017/06/22/the-demographics-of-gun-ownership`.

47. Horwitz S, Grimaldi JV. ATF's oversight limited in face of gun lobby. Washington Post.; 2010 Oct 26 [cited 2024 Aug 2]. Available from: `https://www.washingtonpost.com/wp-dyn/content/article/2010/10/25/AR2010102505823.html?sub=AR`.

48. Giffords Law Center to Prevent Gun Violence. Maintaining Records of Gun Sales in California; 2023 [cited 2024 Aug 7]. Available from: `https://giffords.org/lawcenter/state-laws/maintaining-records-of-gun-sales-in-california/`.

49. Hummon DM. In: Altman I, Low SM, editors. Community attachment. Boston, MA: Springer US; 1992. p. 253–278.

50. Belanche D, Casaló LV, Rubio MA. Local place identity: A comparison between residents of rural and urban communities. J Rural Stud. 2021;82:242–252. doi:10.1016/j.jrurstud.2021.01.003.

51. Sozer MA, Merlo AV. The impact of community policing on crime rates: Does the effect of community policing differ in large and small law enforcement agencies? Police Pract Res. 2013;14(6):506–521. doi:10.1080/15614263.2012.661151.

52. Everytown Research  Policy. Community-Led Public Safety Strategies; 2022 [cited 2024 Aug 14]. Available from: `http://www-cs-faculty.stanford.edu/~uno/abcde.html`.

53. Waters JS, Holbrook CT, Fewell JH, Harrison JF. Allometric scaling of metabolism, growth, and activity in whole colonies of the seed-harvester ant Pogonomyrmex californicus. Am Nat. 2010;176(4):501–510. doi:10.1086/656266.

54. Porfiri M, De Lellis P, Aung E, Meneses S, Abaid N, Waters JS, et al. Reverse social contagion as a mechanism for regulating mass behaviors in highly integrated social systems. PNAS Nexus. 2024;3(7). doi:10.1093/pnasnexus/pgae246.

55. Hernandez-Vargas G, Sosa-Hernández JE, Saldarriaga-Hernandez S, Villalba-Rodríguez AM, Parra-Saldivar R, Iqbal HM. Electrochemical biosensors: a solution to pollution detection with reference to environmental contaminants. Biosensors. 2018;8(2):29. doi:10.3390/bios8020029.

56. Shen G, Preston W, Ebersviller SM, Williams C, Faircloth JW, Jetter JJ, et al. Polycyclic aromatic hydrocarbons in fine particulate matter emitted from burning kerosene, liquid petroleum gas, and wood fuels in household cookstoves. Energy Fuels. 2017;31(3):3081–3090. doi:10.1021/acs.energyfuels.6b02641.

57. Yu Y, Katsoyiannis A, Bohlin-Nizzetto P, Brorstrom-Lunden E, Ma J, Zhao Y, et al. Polycyclic aromatic hydrocarbons not declining in Arctic air despite global emission reduction. Environ Sci Technol. 2019;53(5):2375–2382. doi:10.1021/acs.est.8b05353.

58. Sutton J, Shahtahmassebi G, Hanley QS, Ribeiro HV. A heteroscedastic Bayesian generalized logistic regression model with application to scaling problems. Chaos Solit Fractals. 2024;182:114787. doi:10.1016/j.chaos.2024.114787.

59. United Stated Census Bureau. Incorporated Places and Minor Civil Divisions Datasets: Subcounty Resident Population Estimates: April 1, 2020 to July 1, 2023 (SUB-EST2023); 2024 [cited 2024 May 11]. Database: City and Town Population Totals: 2020-2023 [Internet]. Available from: `https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html`.

60. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–362. doi:10.1038/s41586-020-2649-2.

61. Ortman SG, Cabaniss AH, Sturm JO, Bettencourt LM. The pre-history of urban scaling. PLOS One. 2014;9(2):e87902. doi:10.1371/journal.pone.0087902.

62. United States Bureau of Labor Statiscs. COUNTY-MSA-CSA CROSSWALKS, 1990-2012  2013-2023; 2024 [cited 2024 July 7]. Database: City and Town Population Totals: 2020-2023 [Internet]. Available from: `https://www.bls.gov/cew/classifications/areas/county-msa-csa-crosswalk.htm`.

63. Bureau of Alcohol, Tobacco, Firearms and Explosives. U.S. Firearms Trace Data by State; 2022 [cited 2024 May 11]. Database: Data  Statistics [Internet]. Available from: `https://www.atf.gov/resource-center/data-statistics`.

64. Devore JL, Berk KN, Carlton MA, et al. Modern mathematical statistics with applications. vol. 285. 3rd ed. Cham, Switzerland: Springer; 2012.

## Supporting information

**S1 Appendix.** This appendix consists of five sections that provide additional details supporting the claims made in the main manuscript. Section A: Assessing consistency in urban scaling. Section B: Assessing bias in urban scaling with alternative 12 values of $\beta$ and $N$. Section C: Optimization problem. Section D: Validity of the greedy algorithm solution. Section E: Urban scaling of firearm homicides with complete data.