



Urban population distribution in China: Evidence from internet population[☆]

Huixuan Li^a, Jing Chen^{b,*}, Zihao Chen^c, Jianguo Xu^d

^a School of Economics, Beijing Technology and Business University, No.33 Fucheng Road, Haidian District, Beijing 100048, PR China

^b College of Economics and Management, China Agricultural University, No.17 Qinghuadong Road, Haidian District, Beijing 100083, PR China

^c Yinhua Fund Management Co. Ltd, 15/F Tower C2 Oriental Plaza No.1 East Chang An Ave., Dong Cheng District, Beijing 100738, PR China

^d National School of Development, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing 100871, PR China

ARTICLE INFO

JEL classification:

J61
R12
P21

Keywords:

Internet population
Population distribution
Zipf's law
Public resource distortions
Big data

ABSTRACT

Based on mobile internet user data, we construct an "Internet population" measure and reexamine spatial population distribution in China. The location based service (LBS) data of mobile internet users is able to capture the accurate location of users' residence and solve the underestimation problem of missing migrants. We have three main findings. First, contrary to previous studies based on traditional population statistics, city size distribution of Internet population fits well into Zipf's law with a R^2 of 90.7%. Second, the Internet population indicator is superior to traditional population statistics in explaining inelastic household consumption such as water consumption, electricity consumption, and garbage disposal. It suggests that the "Internet population" is a better proxy of actual city population. Third, the traditional population statistics systematically overestimate population in small cities and underestimate population in large cities. It indicates that the public resource distortions will continue to exist or even worsen off in China if the allocation process relies greatly on traditional population statistics. Although no measures are perfect, our new population measure provides important incremental information for future discussion.

1. Introduction

Geographic population agglomeration promotes economic efficiency and facilitates resource allocation (Glaeser & Gottlieb, 2009; Mera, 1973). Many studies have emphasized that urbanization is one of the major driving forces for China's economic growth and production efficiency improvement (Bertinelli & Black, 2004; Brandt & Zhu, 2010; Young, 2003; Zhu, 2012). In urban studies, accurate measurement of city population is of great importance for policy making and allocation of public resources. However, the official census data suffers severe drawbacks, i.e., the omission of migrants, which leads to inefficient resource allocation. For example, the allocation of intergovernmental fiscal transfer payment is largely based on the official population statistics, which might be quite inaccurate (Zhang & Martinez-Vazquez, 2003). In this paper, we obtain the Location Based Service (LBS) data of the largest social

[☆] The authors gratefully acknowledge the financial support from the National Natural Science Foundation of China (#71472006, #71772004) and the National Social Science Foundation of China (#20CJY045). All remaining errors and inaccuracies are our own responsibility.

* Corresponding author.

E-mail addresses: lihuixuan@btbu.edu.cn (H. Li), Jing.chen@cau.edu.cn (J. Chen), chenzihao@pku.edu.cn (Z. Chen), jgxu@nsd.pku.edu.cn (J. Xu).

media company in China to construct a better complementary proxy for population distribution. The LBS data is superior in locating users' residence, avoiding the omission or misstatement problem of migrants during the manual record process of the census.

Previous literature has consistently shown that population distribution of cities in the United States and many other countries fit Zipf's law, which states that city size distribution is well approximated by a Pareto distribution with coefficient 1 (Rosen & Resnick, 1980; Gabaix, 1999a, 1999b; Rossi-Hansberg & Wright, 2007). Previous studies also prove that if all cities exhibit the same expected growth rates and encounter idiosyncratic shocks, the city size distribution will converge to Zipf's law in the steady state (Gabaix, 1999a, 1999b). However, many studies find that the Chinese urban scale distribution largely fail to validate Zipf's law. Violating Zipf's Law is often regarded as evidence that Chinese population distribution deviates from steady state, which harms economic efficiency (Chauvin, Glaeser, Ma, & Tobio, 2017; Soo, 2014).

It's worth noting that earlier studies against Zipf's law mainly rely on traditional population statistics. However, previous studies often find it difficult to obtain high-quality measures of population distribution. For example, the census data of the United States is often criticized for its omission of immigration statistics (Siegel, 1974; Erickson & Kadane, 1985; Winkler, Johnson, Cheng, Voss, & Curtis, 2013). This problem is extremely severe in China during the mass urbanization process in past 40 years. We have observed evidence that official population statistics published by the National Bureau of Statistics, hereinafter called "the NBS population statistics", mismeasures the actual population number in many cities.

An interesting example emerges during the COVID-19 outbreak in Wuhan, China. In a press conference after the outbreak, the Wuhan Municipal Government announced that 5 million people have left the city before the lockdown and 9 million residents were still in the city. However, according to the NBS population statistics, the total amount of resident population of Wuhan is 11 million, which implies a gap of 3 million people, or nearly 30% of the total population.

Shenzhen is another interesting example. In the past 40 years, Shenzhen grows from small villages into a mega city, but its population size remains a mystery. According to the NBS population statistics, the total amount of Shenzhen's permanent residents in 2017 is 12.5 million. However, according to Shenzhen Municipal Social Security Bureau, the number of residents insured by social medical insurance is about 14 million. Furthermore, according to the China Mobile Data Center, the estimated population based on the number of mobile phone users is about 22 million in 2017, which almost doubles the population published by the NBS.¹

Wuhan and Shenzhen are typical epitomes of the difficulties in estimating the geographic population distribution in China. Numerous studies suggest three major defects that hurt the accuracy of NBS population statistics based on census data. First, the NBS statistics are prone to underestimate the floating population.² Migrants that often change their places of residence are much more likely to be missed by the census than local residents. In addition, landlords are unwilling to report the residence information of their renters, in order to hide rent income. Similarly, employers also tend to under-report migrant employees to save social security expenditure. As a result, large cities that attract a large crowd of migrants are prone to underestimate their population size (Lu & Duan, 2012; Niu, 2011; Wang et al., 2010).

Second, local governments also have incentives to over-report or under-report their city population. On one hand, local government officials of many small cities might tend to overstate their population in order to obtain fiscal transfers that are pegged to population (Feng, 2010; Pang, 2010). On the other hand, mega cities are confronted with population control targets and thus have incentives to under-report city population. As Fig. 1 shows, Beijing, Shanghai, Guangzhou, and Shenzhen all have defined their population control targets by 2020 in their 13th Five-Year Plans. Therefore, these local governments have the incentive to underestimate their population in order to reach their targets.

Third, population census is conducted only every ten years. During the ten years in between, 1% population sampling is conducted every five years and 1‰ population sampling is conducted every year. In non-census years, the statistical error could be very significant due to small sample problems.

Many previous studies have found circumstantial evidence of such statistical errors. For example, according to National Bureau of Statistics (NBS), the number of Beijing's floating population is 4.8 million in 2008. Meanwhile, based on field research, Beijing Floating Population Committee finds that the number is over 8 million (Lu & Duan, 2012). The inconsistency is obvious and significant. However, due to data limitations, these errors are very difficult to be corrected based on traditional data resources.

In this paper, we use mobile internet big data to construct an "Internet population" indicator to reduce biases in the traditional demographic statistics and to reestimate Chinese urban structure. This is made possible by the Location Based Service (LBS) data of the largest social media platform in China. The LBS data is aggregated to the city level and does not contain any individual user's privacy information. The mobile penetration rate of the social media platform reaches 96.4% among Chinese mobile Internet users by March 2020. Excluding users that are younger than 15 or elder than 64 years old, our data sample is almost the total sample of Chinese working-age population. The location based service (LBS) data is able to capture the accurate location of users' residence, avoiding the omission or misstatement of migrants during the manual record process of the census.

Based on the "Internet population" measurement, we have three main findings. First, the "Internet population" indicator has much better explanatory power for inelastic household consumption behaviors, such as water and electricity consumption and urban garbage disposal, than the NBS population statistics. In particular, "Internet population" explains 71.4% of variations of water consumption, which is about 50% higher than the explanatory power of NBS population statistics. These findings imply that "Internet population" is a better measure for actual population distribution in the sense of higher correlation with basic household consumption.

¹ According to data recently published by the 7th population census, Shenzhen has a population of 17.56 million, which is 4.12 million larger than the 13.44 million in 2019.

² Floating population are migrants who do not have local household registration (Hukou)

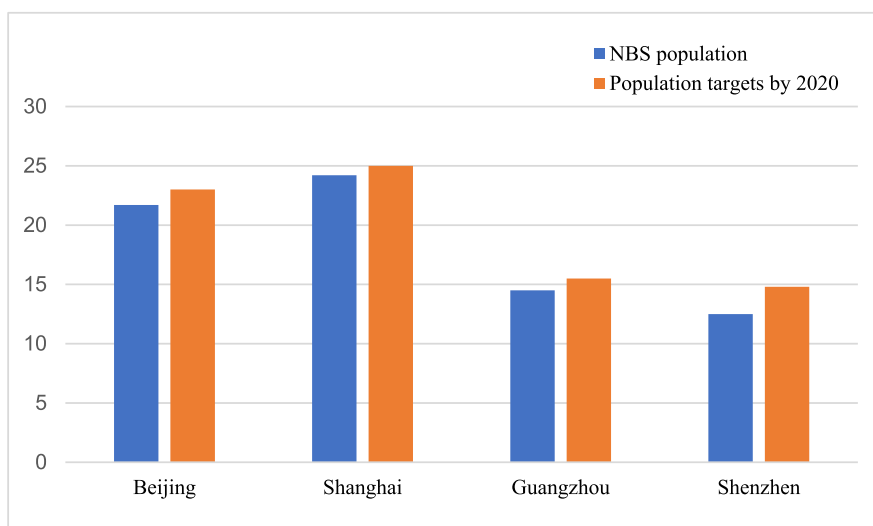


Fig. 1. Resident population in 2017 and control targets in 2020: first-tier cities.⁵¹

Second, contrary to previous studies based on the NBS population statistics, the “Internet population” distribution fits well into the Zipf’s Law with a R^2 of 90.7%. Further analysis shows that this is because the NBS statistics underestimate the population of large cities while overestimate the population of small cities. Compared with the NBS population statistics, the agglomeration of “Internet population” in large cities significantly increases.

Third, using the “Internet population” indicator, we find that the size of floating population is significantly underestimated in previous studies. The total floating population from the official census data is 224 million. In comparison, the total floating population calculated from the “Internet population” is about 336 million, which is about 50% higher from the NBS population statistics. The extra 112 million floating population is mostly crowded in large cities.

This paper pioneers the empirical study of Chinese urban structure by utilizing a new data source of mobile internet to reduce statistical biases of official demographic data and providing novel evidence of Chinese population distribution. This paper makes three main contributions.

First, this paper is the first to apply mobile internet data to estimate China population distribution, which makes a pioneering contribution to demographic and urban studies of China. In the past two decades, big data techniques have been introduced into urban economic studies. For example, Henderson, Storeygard, and Weil (2011) and Chen & Nordhaus (2010) prove the feasibility of using nighttime lighting data instead of GDP data. Glaeser, Kominers, Luca, and Naik (2018) use Google Street View to predict regional income levels. Nowadays, the popularity of smart phone devices and mobile internet has produced a wealth of internet user data with precise localization at high frequencies. As a pioneering study of constructing city-level population variables based on mobile internet data, this paper makes an explorative effort that might have some insights for future research in urban economics and policy design of urban planning.

Second, using Location-Based-Service data of mobile internet users, this paper reduces statistical biases of the NBS population statistics and contributes to the estimation of population distribution, especially floating population. We find that the floating population in China is heavily underestimated by official statistics. And we also find that “Internet population” is superior to the NBS population statistics in explaining inelastic household consumption variables such as water, electricity and garbage disposal, which implies that “Internet population” is a better description of actual population distribution in China.³

Finally, this paper finds novel empirical evidence of Chinese population distribution. Contrary to previous studies based on the NBS population statistics, the “Internet population” distribution fits well into the Zipf’s Law with a R^2 of 90.7%. This means that the basic economic forces play a more fundamental role in population mobility, despite multiple obstacles that hinder free migration. These results have important policy implication. The NBS population statistics systematically overestimate population in small cities and underestimate population in large cities. However, the allocation of public resources is largely relied on the flawed statistics, which will lead to spatial distortion of public resource distribution and severe social welfare loss.

The rest of this paper is organized as follows. Section 2 introduces data and the construction methodology of “Internet population”. Section 3 reports the data quality tests. This is mainly by examining whether inelastic household consumption indicators are better correlated with the NBS population statistics or “Internet population”. Section 4 discusses China’s population distribution and Zipf’s law. Section 5 further discusses the distortion of public resource allocation caused by the *Hukou* system. Section 6 summarizes our main

³ The seventh census data published in May 2021 support our argument. According to the seventh census data, there are 376 million floating population, which is much larger than the earlier 244 million. It is also larger than our 336 million. This difference is probably because our data is for 2017, while the census data is for 2020.

findings.

2. Data and methods

2.1. Data source

The original data used in this paper is mainly obtained from the largest social media company in China. It records location-based services (LBS) in connecting people at work, home, and play. All data included in this paper are anonymous encrypted and summarized to the city level, which does not involve any individual user's privacy information. The dataset covers 31 provinces and municipalities and 358 prefecture-level cities in mainland China (excluding Hong Kong, Macao and Taiwan regions) in 2017.

2.2. Data processing

Prior to the estimation of city population, we define two characteristic variables of each social media APP user. The first variable is the user's resident city, defined as the city with the highest frequency of logins in 2017. The second variable is user's age. We use the information of the registered age, the average age of the chatting group members, etc., and estimate it by the linear fitting model of xgboost.

When defining a user's resident city, we have to deal with two difficulties. First, although the penetration of the social media APPs in China is very high, there are still some people that have not yet been covered, especially the elderly and children. Second, some people might use several different social media accounts at the same time. To deal with these difficulties, we adopt the following three steps:

- (1) Match the social media account id with the device id, and deduplicate the user accounts on the same device.
- (2) Remove social media accounts with abnormally low activity, and accounts created for marketing or business purposes.
- (3) Remove social media users that are younger than 15 or older than 64 years old. The "Internet population" of each city is defined as the total number of deduplicated residents aged 15–64.

By December 2017, there are 1.46 billion social media accounts in China with information of age and residence. After the first two steps of deduplication, there are 0.96 billion accounts left, of which 0.88 billion are 15–64 years old, covering 88% of the total Chinese population aged 15–64 in 2017. Therefore, our construction of "Internet population" is comparable to the working-age NBS population. It is better described as "working-age social media population".

Although the "Internet population" cannot perfectly represent the total Chinese working-age population, it still has several advantages compared with the NBS population statistics. First, the location based service (LBS) data can accurately identify the residential address of migrants. Based on big data automatically collected by mobile internet, the "Internet population" measurement can avoid the omission or misstatement of migrants during the manual record process of the census. As discussed in the following sections, the errors and omission of "Internet population" are relatively minor, compared with the huge number of floating population missed by the NBS statistics. Second, the "Internet population" statistics also have very low collection costs and the advantage of real-time online. The "Internet population" can be updated monthly, and even daily in the future, which is of great significance to the causal analysis of policy shocks. Third, the "Internet population" has the advantage of finer geographic scales than ever before. The GPS positioning of mobile devices can be accurate to 10^{-4} latitudes and longitudes, or roughly 10 m. This means that future studies on urban structure can employ block level data to overcome restrictions of administrative regions. Based on the database with high precision and high frequency, the "Internet population" has good potential for future research on demographic geography, urban economics, and city management.

In this study, we compare "Internet population" with two official population measures that are frequently employed in the existing literature, "resident population" and "registered population". Resident population is published by the National Bureau of Statistics (NBS), which is also referred to "NBS population" in this paper. It is estimated from official statistical surveys. The national population census is conducted once every ten years. The national 1% sample survey is conducted once every five years. The national 1‰ sample survey is conducted in other years.

Registered population is collected by the Ministry of Public Security (MPS). It is the total amount of residents registered in household *Hukou* system. The registration is often used as the basis of public service allocation such as education, pension, and medical benefits. It is also employed in dwelling purchasing regulation.

Recent studies also use Worldpop data, Night-time Light data, and LandScan data to describe city population (Bennett & Smith, 2017; Lloyd, Sorichetta, & Tatem, 2017; Mellander, Lobo, Stolarick, & Matheson, 2015; Meng, Jiang, Wang, & Long, 2021; Zeng, Zhou, Wang, Yan, & Zhao, 2011). Compared with these data source, the "Internet population" based on LBS data still have advantages in many aspects.

First, the Worldpop database provides estimates of numbers of people per hectare in 1990, 2000, 2010 in China. In this database the Chinese population data were obtained from the National Bureau of Statistics of China via the Chinese Center for Disease Control and Prevention at the Quxian level (county level). LandScan is developed using demographic (Census) and geographic data, remote

⁵ This graph compares resident population in 2017 with governmental population control targets by 2020 of first-tier cities. The resident population data is obtained from the National Bureau of Statistics database. The control targets are collected from government websites.

sensing, and imagery analysis techniques within a multivariate dasymetric modeling framework to disaggregate census counts within an administrative boundary. It is obvious that both the Worldpop data and the LandScan data are derived from traditional demographic (Census) data from the National Bureau of Statistics of China (NBS). Therefore, they also suffer severe omissions of migrants. Our LBS data is superior in locating users' residence, avoiding the omission or misstatement of migrants during the manual record process of the census.

Second, night-time light data is an alternative source of information to indirectly characterize the human population distribution. However, because of the light saturation in high-density population areas, night-time lights tend to underestimate the population density in the urban centres and overestimate the density in suburban areas (Sutton et al., 1997). In addition, night-time light is significantly affected by economic development, which might imply that a large part of the resident population may be undetected in underdeveloped areas. And some studies also find a spatial mismatch between radiance and population distribution. Brightest lights are to be found in downtown areas and tourist areas where tend not to have the highest residential population (Sutton, 2003; Doll, 2008). By comparison, the LBS data directly characterizes the individual location, and consequently provides better description of population distribution.

Needless to say, all data are imperfect, including this new data. In the end, we need different data to complement each other in providing more information. It is clear that the "Internet population" provides new information about China's urbanization.

3. Comparison of "Internet population" with traditional population indicators

Theoretically, household demand for basic consumption needs, such as water, electricity and garbage disposal etc., is relatively inelastic. Consumption per capita for such necessities is supposed to approximately identical among similar cities, especially with similar climate and income level. In other words, the city population is expected to be highly correlated with these basic consumption indicators.

Every year, the National Bureau of Statistics (NBS) publishes the total amount of household electricity consumption, household water usage, and urban domestic garbage in each city. These statistics are collected by professional companies providing these living facilities, and therefore have high accuracy.

In this section, we test data quality of the "Internet population" by examining whether these household consumption indicators are better predicted by NBS population or "Internet population". If "Internet population" has a higher explanatory power than NBS population, it gives us confidence that it might be a good measure of city population.

3.1. The statistical anomalies

We start with descriptive statistics of the relationship between household consumption and different measures of population. Surprisingly, we find the NBS population statistics correlate poorly with basic consumption variables, which we label as statistical anomalies. Even more interestingly, we find the statistical anomalies disappear when we use the "Internet population" to correlate with the basic consumption variables.

In Fig. 2, we compare two measurements of household consumption per capita in major cities in 2017. The NBS measurements of consumption per capita are defined as "total amount of household consumption/NBS population". And the Internet measurements are defined as "total amount of household consumption /Internet population".

As shown in Fig. 2.A and Fig. 2.C, anomalies emerge if we use NBS population measures to calculate the per capita household water and electricity consumption. For example, Guangzhou's per capita household water usage is four times that of Chongqing. Meanwhile, its per capita electricity consumption is three times that of Chongqing. This is highly unlikely because both Guangzhou and Chongqing are high income cities in southern China. Besides these two cities, the per capita household consumption also varies widely among other large cities with similar income level.

Furthermore, we find that cities with abnormally high household consumption of water and electricity are mainly areas with large migrant inflow, such as Guangzhou, Shenzhen and Dongguan. In contrast, cities with large migrant outflow are accompanied by abnormally low per capita consumption, such as Tianjin and Chongqing. One reasonable explanation is that NBS population of migrant-importing cities is underestimated due to the omission of floating population. The underestimation of population results in the overrated consumption per capita in Guangzhou, Shenzhen and Dongguan. Meanwhile, NBS population of Chongqing and Tianjin is overestimated. As a result, the per capita consumption is relatively underestimated.

The patterns presented in Fig. 2.A and Fig. 2.C may be disturbed by climatic factors. Guangdong Province is the sub-tropical and tropical region with long and muggy summers, which might lead to relatively larger demand for water and electricity per capita than cities in Northern China. In order to eliminate the influence of climatic factors, we focus on the cities in Guangdong Province, as shown in Fig. 2.B and Fig. 2.D. The divergences of the NBS indicators within cities of Guangdong Province are even more significant. The per capita water consumption in Guangzhou is 15 times that of Shanwei. And the per capita electricity consumption in Zhongshan is nearly 30 times that of Jieyang. This evidence strongly implies the existence of systematic biases in the official population data. In fact, Guangdong is a province with a large amount of inner-province migrants. The urbanized Pearl River Delta area⁴ has attracted a large inflow of workers from the rest parts of Guangdong province (Shen, Feng, & Wong, 2006), which also exacerbates the omission of

⁴ The Pearl River Delta area is the low-lying area surrounding the Pearl River estuary in Guangdong province, which include metropolises such as Shenzhen, Guangzhou, Dongguan, Foshan, Zhongshan, Zhuhai.

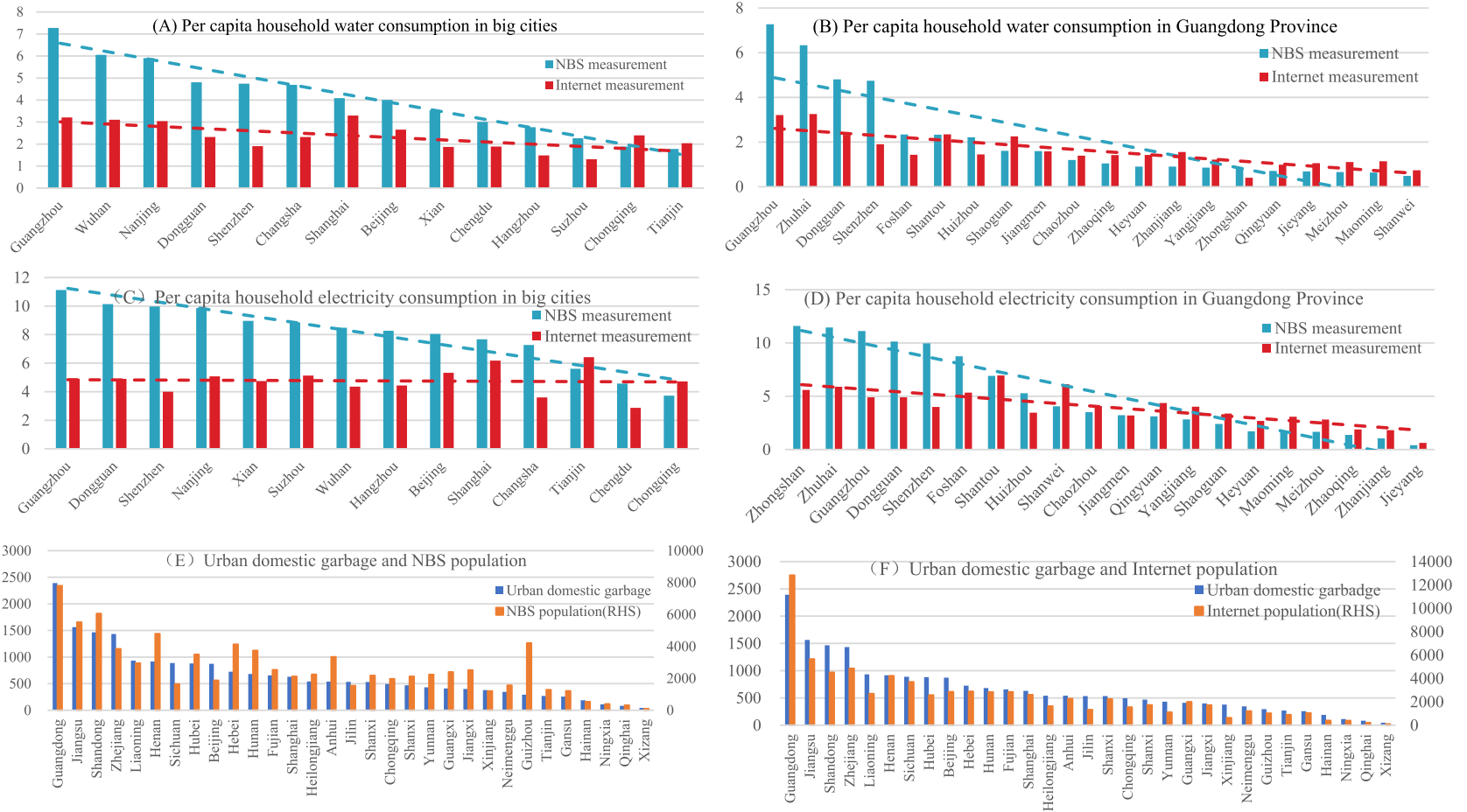


Fig. 2. The correlation of population indicators with household consumption.⁶¹

floating population. Consequently, it is not surprising that the miscalculations of population cause such important differences.

In contrast to the NBS measurements, from Fig. 2.A to Fig. 2.D, statistical anomalies vanish when we apply the “Internet population” to calculate the per capita household water and electricity usage. For example, Guangzhou’s Internet indicator of per capita water consumption is only one-third higher than that of Chongqing. And Guangzhou’s Internet measure of per capita electricity consumption is roughly equal to that of Chongqing. In other words, the “Internet population” accords closely with household consumption for such necessities. These results are consistent with our prediction. In this sense, the “Internet population” is a better measurement of actual population size for most cities.

We obtain similar findings when we use urban domestic garbage and the two population indicators at the provincial level. Fig. 2.E shows the relationship between the total urban garbage and the NBS population in each province. Fig. 2.F shows the total urban garbage disposal and the “Internet population”. In Fig. 2.E, there is only a weak correlation between the garbage disposal and the NBS population, while the trends of the domestic garbage and the “Internet population” in Fig. 2.F are highly correlated. This result also shows that Internet population has much better explanatory power for inelastic household consumption.

3.2. Explanatory power

The previous subsection describes the bivariate correlations between household consumption and different population indicators. These tests have an obvious drawback that household garbage, water usage and electricity usage might be affected by other macro variables. In this section, we proceed to quantify the relationship between household consumption and population indicators using multivariate OLS regressions. Considering the comparability of the samples, we use the sample of 255 cities with officially published resident population data. We find that, overall, the “Internet population” has a much stronger explanatory power for household consumption than the NBS population does.

Table 1 reports the regressions of total household water usage on two population indicators. The first column is the OLS regression of total household water usage on the “Internet population”, both in log values. The second column is the regression of the log value of total household water usage on the log resident population. Although the coefficients of the “Internet population” and NBS population are both significantly positive at the 99% confidence level, the R^2 of the two regressions is quite different. “Internet population” explains 71.4% of the variation in household water usage. In contrast, the NBS population indicator only explains 45.2% of the variation in household water consumption, which is 26.2 percentage points lower than that of “Internet population”.

In order to attenuate the impact of economic, climatic and geographical factors on household water consumption, we add control variables such as GDP, city tiers and province-level fixed effects into the regression (Columns (3) and (4) in Table 1). Under the same economic and climatic conditions, city size is expected to be positively correlated with total household water usage. We find that after adding the control variables, the coefficient of “Internet population” in Column (3) is still significantly positive at the confidence level of 95%. However, the coefficient of the resident population in Column (4) becomes significantly negative, which obviously violates our prediction.

We observe a similar pattern in total household electricity consumption in Table 2. The first and second columns are OLS regressions of total household electricity consumption on “Internet population” and NBS population, respectively. We also find that the explanatory power of “Internet population” is far larger than that of NBS population. The R^2 in the first column is 73.9%, while the R^2 in the second column is only 50.8%. The third and fourth columns show that after including control variables, the coefficient of “Internet population” is still significantly positive at the 99% confidence level, while the coefficient of NBS population is no longer significant. From the perspective of predicting the inelastic household consumption for basic necessities, the “Internet population” is a much better description of the Chinese population distribution than the official population data.

Table 3 presents the regression of the household garbage disposal on population indicators. Since the official household garbage data is published at the provincial level, we conduct regression analysis on 31 provinces and municipalities in mainland China. We also observed a similar pattern as in Tables 1 and 2. The “Internet population” explains 93.4% of the variation in household garbage disposal, while NBS population explains 83.3% of the variation. After adding control variables, the coefficient of “Internet population” is still significantly positive at 99% confidence level. However, the coefficient of NBS population is no longer significant.

Overall, we find that the “Internet population” has stronger explanatory power for household consumption variables. The difference in R^2 is economically significant at around 20 to 30 percentage points. This evidence supports that the “Internet population” might have high precision in describing Chinese actual population distribution.

4. Chinese population distribution and Zipf’s law

4.1. Chinese population distribution estimated by “internet population”

The previous section shows that “Internet population” is a better measure of city population than NBS population statistics. In this section, we employ this indicator to describe Chinese population distribution and compare it with traditional population indicators. Our main finding is that labor mobility has been greatly underestimated by official statistics. The population in Chinese labor-importing megacities is underestimated by the NBS statistical indicators. On the other hand, the population in labor-exporting cities is overestimated by official indicators.

Before going into the details, it’s worth noting that “Internet population” is defined as the total number of deduplicated residents of working age, namely, aged 15–64. Naturally, we will compare “Internet population” with official resident population and registered population of the same age interval. Labor-importing areas are defined as cities where “Internet population” exceeds NBS population of

Table 1

Regression of city water consumption on different population measures.

	<i>Ln(water_consumption_i)</i>			
	(1)	(2)	(3)	(4)
<i>Ln(Internet_Population_i)</i>	1.053*** (24.19)		0.281** (2.30)	
<i>Ln(NBS_Population_i)</i>		1.028*** (13.89)		−0.167** (−2.07)
<i>Ln(GDP_i)</i>			0.757*** (6.37)	0.899*** (10.90)
<i>Dummy_{2ndtier}</i>			−0.131 (−0.40)	−0.081 (−0.25)
<i>Dummy_{3rdtier}</i>			−1.017*** (−2.95)	−0.942*** (−2.75)
Constant	−2.438*** (−9.80)	−2.257*** (−5.39)	0.598 (0.92)	0.938 (1.45)
Province Fixed Effect	N	N	Y	Y
Observations	236	236	236	236
R ²	0.714	0.452	0.867	0.869

This table reports OLS regressions of city water consumption on different population measures. The dependent variable, *Ln(water_consumption_i)* is the natural logarithm of water consumption in city *i*. *Ln(Internet_Population_i)* is the natural logarithm of “Internet Population” in city *i*. *Ln(NBS_Population_i)* is the natural logarithm of the population of city *i* published by the NBS (National Bureau of Statistics). *Ln(GDP_i)* is the natural logarithm of GDP of city *i*. All data are for 2017. *Dummy_{2ndtier}* and *Dummy_{3rdtier}* are dummy variables that are equal to one if city *i* is included in 2nd tier cities and 3rd tier cities, respectively. We include province fixed effects in the regressions of column (3) and column (4). The *t*-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

Table 2

Regression of city electricity consumption on different population measures.

	<i>Ln(electricity_consumption_i)</i>			
	(1)	(2)	(3)	(4)
<i>Ln(Internet_Population_i)</i>	1.094*** (25.74)		0.530*** (4.26)	
<i>Ln(NBS_Population_i)</i>		1.113*** (15.54)		0.116 (1.31)
<i>Ln(GDP_i)</i>			0.275** (2.20)	0.650*** (7.16)
<i>Dummy_{2ndtier}</i>			−0.000 (−0.00)	0.033 (0.09)
<i>Dummy_{3rdtier}</i>			−0.698* (−1.93)	−0.841** (−2.23)
Constant	−6.430*** (−26.47)	−6.494*** (−16.03)	−3.529*** (−5.18)	−3.173*** (−4.45)
Province Fixed Effect	N	N	Y	Y
Observations	236	236	236	236
R ²	0.739	0.508	0.860	0.848

This table reports OLS regressions of city electricity consumption on different population measures. The dependent variable, *Ln(electricity_consumption_i)* is the natural logarithm of household electricity consumption in city *i*. *Ln(Internet_Population_i)* is the natural logarithm of “Internet Population” in city *i*. *Ln(NBS_Population_i)* is the natural logarithm of the population of city *i* published by the NBS (National Bureau of Statistics). *Ln(GDP_i)* is the natural logarithm of GDP of city *i*. All data are for 2017. *Dummy_{2ndtier}* and *Dummy_{3rdtier}* are dummy variables that are equal to one if city *i* is included in 2nd tier cities and 3rd tier cities, respectively. We include province fixed effects in the regressions of column (3) and column (4). The *t*-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

working age. Labor-exporting areas are cities where “Internet population” is smaller than working-age NBS population. Fig. 3 shows an overview of China’s population migration. The cities with huge population inflow are concentrated in the Pearl River Delta and the Yangtze River Delta. In addition, regional central cities such as Beijing, Chengdu, Xi’an, Zhengzhou, and Changsha are also very attractive to migrants. In contrast, the third- and fourth-tier cities in Central China suffer the most severe population outflow. People born in these areas are more likely to leave their hometown because of relatively convenient transportation facilities and the huge attraction of neighboring megacities.

Figs. 4 and 5 report population indicators in major labor-importing cities and labor-exporting cities. Labor-importing cities include first-tier cities and most of second-tier cities. As illustrated in Fig. 4, the “Internet population” is substantially larger than the working-age resident population published by the NBS. In the four first-tier cities (Beijing, Shanghai, Guangzhou, Shenzhen), “Internet population” has exceeded 20 million, approximately 5 to 13 million higher than official resident population. The gap between registered population and “Internet population” is even larger.

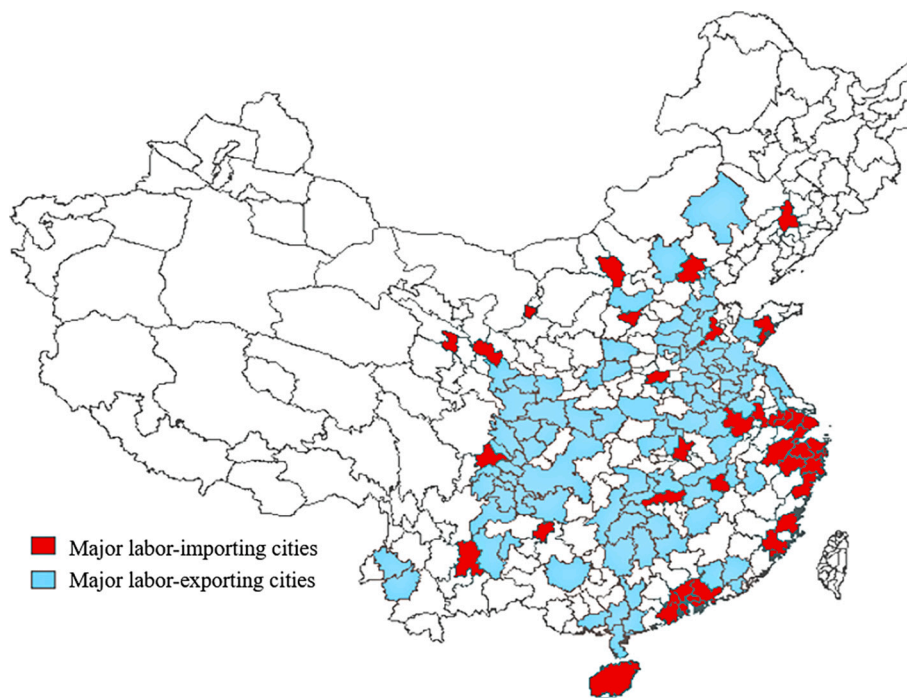
The pattern of labor-exporting cities is exactly the opposite (Fig. 5). In these areas of huge migrant outflow, such as southern Shandong, northern Jiangsu and Henan province, “Internet population” is much lower than official resident population and registered

Table 3

Regression of household waste disposal on different population measures.

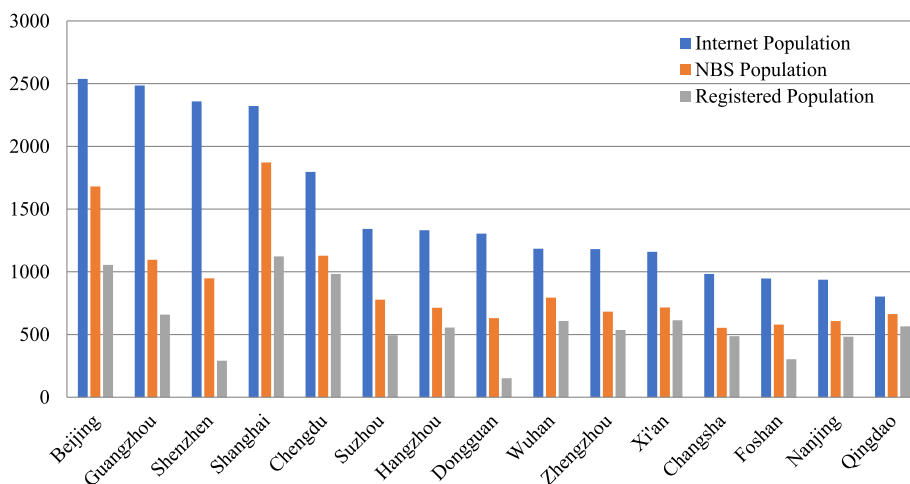
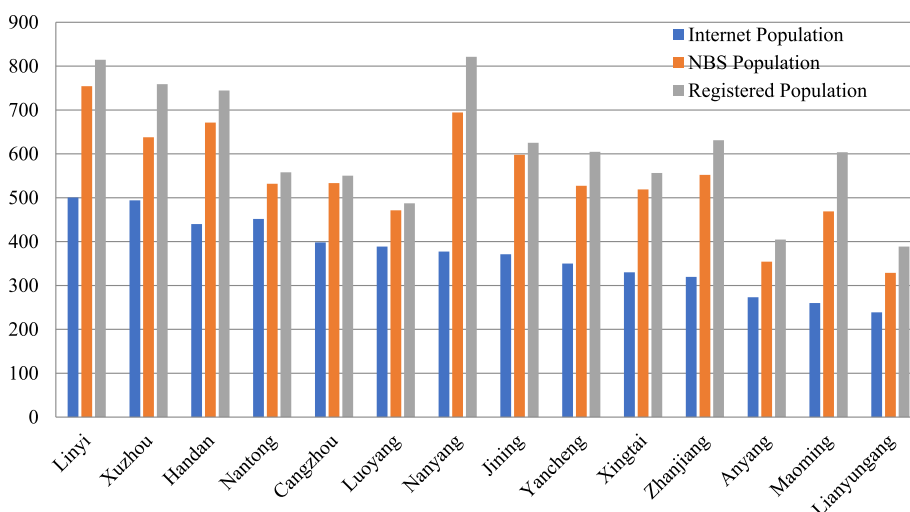
	$\ln(waste_i)$			
	(1)	(2)	(3)	(4)
$\ln(Interne_Population_i)$	0.895*** (20.33)		0.361** (2.30)	
$\ln(NBS_Population_i)$		0.850*** (12.01)		−0.009 (−0.08)
$\ln(GDP_i)$			0.576*** (3.41)	0.954*** (7.82)
$GDP_industryratio_i$			0.126** (2.16)	0.094 (1.41)
$\ln(Sales_i)$			0.192 (0.69)	0.381 (1.23)
Constant	−0.645* (−1.90)	0.012 (0.02)	−1.561 (−1.20)	−2.259 (−1.58)
Regional Fixed Effect	N	N	Y	Y
Observations	31	31	31	31
R^2	0.934	0.833	0.975	0.968

This table reports OLS regressions of household waste disposal on different population measures across 31 provinces. The dependent variable, $\ln(waste_i)$ is the natural logarithm value of total amount of waste disposal in province i . $\ln(Interne_Population_i)$ is the natural logarithm of “Internet Population” in province i . $\ln(NBS_Population_i)$ is the natural logarithm of the population of province i published by the NBS (National Bureau of Statistics). $\ln(GDP_i)$ is the natural logarithm value of GDP of province i . $GDP_industryratio_i$ is the GDP ratio of services to manufacture industry in province i . $\ln(Sales_i)$ is the total retail sales of consumer goods in province i . All data are for 2017. We include regional effects in the regressions of column (3) and column (4). The t -statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

**Fig. 3.** Major labor-importing cities and labor-exporting cities.⁷¹

population. That is to say, labor mobility has been greatly underestimated by official population data.

To summarize, both the attractiveness of the labor-importing megacities and the population outflow in less developed areas are much higher than in traditional statistics. In other words, the polarization of cities in China is much larger than reported by traditional population measures.

Fig. 4. Population of labor-importing cities.⁸¹Fig. 5. Population of labor-exporting cities.⁹¹

4.2. The Zipf's law in China

Previous literature shows that size distribution of cities in many countries demonstrates a striking pattern: city size distribution is well approximated by a Pareto distribution with coefficient 1 (Gabaix, 1999a, 1999b; Rozenfeld, Rybski, Gabaix, & Makse, 2011). To put differently, a linear regression of log-rank on log-size is supposed to yield a very high R^2 . The rule is named after Zipf (1949) who first popularized this empirical finding. It appears to hold in virtually all countries and at all dates, even in the United States in 1790 and India in 1911 (Zipf, 1949; Rosen & Resnick, 1980; Gabaix, 1999a, 1999b). Rosen & Resnick (1980) use data from 44 countries and found that the estimated coefficient was above 0.95 for 36 countries. Dobkins & Ioannides (2001) analyse the city growth processes of metro areas in the United States over the period 1900–1990 and also validate the Zipf's law.

Many researchers have provided explanations as to why and how the Zipf's Law has emerged. This literature has developed along two lines: first, some studies find that the assumption of Gibrat's law of proportional growth can explain Zipf's distributions (Champernowne, 1953; Simon, 1955; Gabaix, 1999a, 1999b). Gibrat's law means that the city growth rate is independent of size and related to some idiosyncratic shocks. When economic shocks are randomly distributed across cities, there would only be a few large

⁷ Fig. 3 shows an overview of China's population mobility. Major Labor-importing cities are top 50 cities with largest positive gap between the "Internet population" and the NBS population. Labor-exporting cities are top 50 cities with largest negative gap between the "Internet population" and the NBS population. The NBS population is the official resident population published by the National Bureau of Statistics (NBS). And the "Internet population" is constructed based on mobile internet big data.

cities due to the low probability of a city being consistently hit by positive economic shocks. Gabaix (1999a, 1999b) proves that if all cities exhibit the same expected growth rates and the same standard deviations, the city size distribution will converge to Zipf's law in the steady state.

Second, some studies use the theory of agglomeration economies to explain Zipf's law. Rossi-Hansberg & Wright (2007) propose a theory that the equilibrium size of cities is determined by the trade-off between agglomeration effects and congestion costs. And the equilibrium state of their model is a city size distribution that is well approximated by the Zipf's law. Behrens, Duranton, and Robert-Nicoud (2014) argue that workers move to larger cities for agglomeration economies. However, the higher living costs in big cities let only talented and highly productive individuals stay. They demonstrate that city size distribution in equilibrium will follow the Zipf's law if a Pareto distribution is assumed for the talent across cities.

In contrast to the almost uniform support for Zipf's Law, many previous studies find that the distribution of Chinese city size seriously deviates from Zipf's Law (Anderson & Ge, 2005; Chauvin et al., 2017; Soo, 2014). Deviation from Zipf's Law is often regarded as the evidence that Chinese population distribution is far from the steady state (Chauvin et al., 2017; Soo, 2014). Explanations for the deviation are mainly focused on administrative interventions, such as *Hukou* system. *Hukou* system requires residents to register in local government to obtain better access to various welfare benefits, such as health care, education and the eligibility to purchase properties. The *hukou* location specifies where one is entitled to receive social benefits. Without a local *Hukou*, it is difficult for migrants to obtain access to the necessities of life, such as child education and public apartment renting in large cities. Previous studies argue that *Hukou* system increases migration costs and leads to the distorted distribution of city population. Au & Henderson (2006) find that due to strict *Hukou* restrictions, most of China's cities are too small, causing about 30% of welfare losses. Desmet & Rossi-Hansberg (2013) find that owing to severe restrictions on population movements, China's welfare losses are as high as 47%, while the welfare losses are only 1.2% for the US. Chauvin et al. (2017) argue that China's population still deviates from the ideal distribution.

However, Chinese government has gradually eased the restrictions of *Hukou* system by two stages since 1980. The first stage is from the 1980s to the end of the 20th century. The *Hukou* policy changed from "strictly restricting urban development" to "strictly controlling the scale of large cities, rationally developing small and medium-sized cities, and actively developing small towns." In this period, the Chinese urban population increased from 190 million to 460 million, and the urbanization rate increased from 19% to 36%.

The second stage is from 2001 to the present, the central government announced a new *Hukou* policy of "coordinating development of large, medium and small cities and small towns" and promoted the protection of legal rights of migrant workers. By now, only the first-tier cities, including Beijing, Shanghai, Guangzhou and Shenzhen, have specific targets to control population size. Except for the first-tier cities and second-tier cities, other cities have almost eliminated their *Hukou* restrictions.

The *Hukou* system reform might accelerate the labor migration in China. And as shown in previous sections, traditional population indicators suffer from the problem of underestimating labor mobility. And it's worth noting that almost all the previous studies use the official population statistics to describe the population distribution in China. Since the NBS population itself is biased, the results of these studies need to be re-examined.

In this section, we re-examine the Zipf's Law in China based on the "Internet population" indicator proposed in the previous sections. Consistent with the methods in the literatures such as Krugman (2004) and Gabaix (1999a, 1999b), we generate the logarithm of city population size $\ln(\text{Size})$ and the logarithm of city ranking $\ln(\text{Rank})$ and run the OLS regression of $\ln(\text{Rank})$ on $\ln(\text{Size})$. If Zipf's Law holds, the relationship between the two variables should be linear and the coefficient should be close to -1 .

The first column in Table 4 is the regression using "Internet population" as population size. The coefficient of the $\ln(\text{Size})$ is -1.085 , which is very close to -1 . The R^2 of the regression is 0.907. And we draw the graph where the y-axis represents the log of the city rank ($\ln(\text{Rank})$) and the x-axis represents the log of the "Internet population" of the corresponding city. The resulting graph in Fig. 6 shows almost a perfect straight line with a slope of -1 , except for a few of the largest cities.

The result suggests that the Zipf's law is a good description of the city size distribution presented by the "Internet population" indicator. Except for the first-tier cities, almost all the other cities have reached the steady state predicted by Zipf's Law. The exceptional cases may be related to the fact that the four first-tier cities are still constrained by size restrictions. As we discussed earlier, *Hukou* restrictions still apply in these mega cities.

In comparison, the city size distribution estimated by the NBS population indicators shows systematic and significant deviations from Zipf's Law. As shown in Fig. 7, most cities are deviated from the straight line with a slope of -1 . In the second column of Table 4, R^2 is 0.775, which is 13% lower than that of "Internet population". And the coefficient of $\ln(\text{Size})$ is -1.223 , with a 22.3% deviation

⁹ Fig. 5 reports population indicators in major labor-exporting cities. "Internet population" is defined as the total number of deduplicated residents of working age, namely, aged 15–64. Naturally, we will compare "Internet population" with official resident population and registered population of working age. Labor-exporting cities are defined as cities where "Internet population" is smaller than NBS population of working age. "Internet population" is constructed based on mobile internet big data. NBS population is the official resident population published by the National Bureau of Statistics (NBS). Registered population is the total amount of residents registered in household *Hukou*, collected by the Ministry of Public Security (MPS).

⁹ Fig. 5 reports population indicators in major labor-exporting cities. "Internet population" is defined as the total number of deduplicated residents of working age, namely, aged 15–64. Naturally, we will compare "Internet population" with official resident population and registered population of working age. Labor-exporting cities are defined as cities where "Internet population" is smaller than NBS population of working age. "Internet population" is constructed based on mobile internet big data. NBS population is the official resident population published by the National Bureau of Statistics (NBS). Registered population is the total amount of residents registered in household *Hukou*, collected by the Ministry of Public Security (MPS).

Table 4
Testing Zipf's Law.

	$\ln(\text{Rank_Internet}_i)$	$\ln(\text{Rank_NBS}_i)$
$\ln(\text{Internet_Population}_i)$	-1.085*** (-23.10)	
$\ln(\text{NBS_Population}_i)$		-1.223*** (-12.88)
Constant	10.092*** (85.85)	10.885*** (48.32)
Observations	256	256
R^2	0.907	0.775

This table reports OLS regressions testing Zipf's Law. The dependent variable, $\ln(\text{Rank_Internet}_i)$ is the natural logarithm of city size ranking based on internet population of city i . $\ln(\text{Rank_NBS}_i)$ is the natural logarithm of city size ranking based on the NBS population of city i . $\ln(\text{Internet_Population}_i)$ is the natural logarithm of "Internet Population" in city i . $\ln(\text{NBS_Population}_i)$ is the natural logarithm of the population of city i published by the NBS (National Bureau of Statistics). All data are for 2017. The t -statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

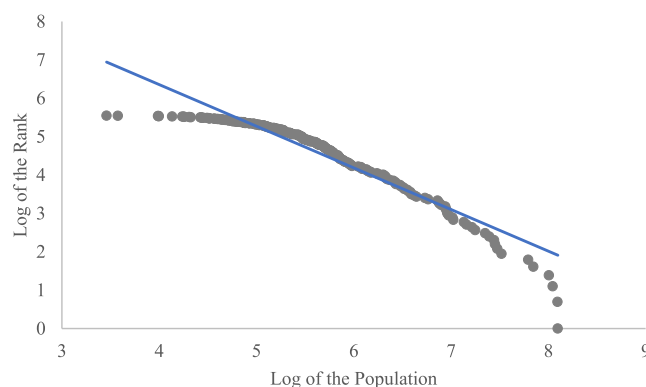


Fig. 6. "Internet population" and Zipf's Law.¹⁰¹

from -1. This evidence also supports our argument that the deviation of the Chinese population distribution from Zipf's Law observed in past empirical studies is likely due to statistical bias.

5. The distortion of public resources

The distortion caused by the Hukou policy consists of two parts. First, the Hukou system inhibits household migration. Second, the government supply of public facilities, such as public schools, public hospitals and fiscal transfer payment, is also attached to the official population statistics (Zhang & Martinez-Vazquez, 2003), which might deviate from the actual population distribution.

In the previous section, we find that the spatial distribution of "Internet population" in China is well predicted by the Zipf's law. It means that the restriction of the Hukou policy on population migration is gradually weakening. But does the distortion of public resources resulting from the Hukou system still persist? In this subsection, we examine whether the distribution of public resources is consistent with "Internet population", a good description of actual population distribution.

First, we examine the correlation between educational facilities and three population indicators, "Internet population", the registered population, and the NBS population. Table 5 presents the regression of the number of primary schools and teachers on these three population variables. We find that the registered population has the strongest explanatory power for the distribution of educational facilities, with R^2 of 0.480 and 0.804 for primary schools and school teachers, respectively. And the regression coefficient between the registered population and the number of primary schools is 1.623, statistically significant at 99% level. If the registered population increases by 1% and the other variables remain the same, the number of primary schools increases by about 1.623% on average. The R^2 of the regression of the number of primary schools on NBS population is 0.266, much smaller than that of registered population. The explanatory power of "Internet population" is the weakest, with R^2 of 0.049 and 0.424 respectively. And the regression coefficient of primary schools on "Internet population" is 0.312, much smaller than that of registered population.

This result shows that the allocation of primary schools is determined by the registered population, rather than "Internet population", a better measure of Chinese population distribution. The distribution distortion has caused the well-known problem of left-behind children in China. In labor-importing cities, educational facilities are in short supply because of the underestimated registered population and undersupply of public resources. Without local *Hukou*, children of migrant workers hardly have access to primary

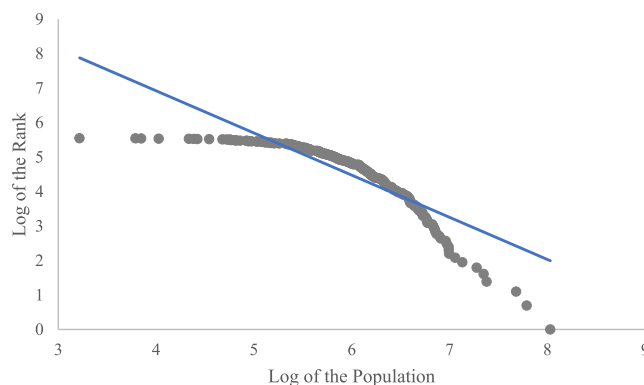
Fig. 7. NBS population and Zipf's Law.¹¹¹

Table 5

Population distribution and allocation of educational resources.

	<i>School_i</i>			<i>Teachers_i</i>		
<i>Internet_Population_i</i>	0.312*** (3.49)			0.041*** (13.14)		
<i>NBS_Population_i</i>		1.065*** (9.22)			0.080*** (24.57)	
<i>Registered_Population_i</i>			1.623*** (14.72)			0.095*** (31.08)
<i>Constant</i>	521.824*** (11.52)	253.768*** (5.03)	70.606 (1.55)	20.507*** (12.80)	6.265*** (4.42)	1.345 (1.06)
Observations	237	237	237	237	237	237
<i>R</i> ²	0.049	0.266	0.480	0.424	0.720	0.804

This table reports OLS regressions of number of schools and school teachers on different population measures, including “Internet Population”, NBS population and registered population. The dependent variable, *School_i* is the number of primary schools of city *i*. *Teachers_i* is the number of primary school teachers of city *i*. *Internet_Population_i* is “Internet Population” in city *i*. *NBS_Population_i* is the resident population of city *i* published by the NBS (National Bureau of Statistics). *Registered_Population_i* is the registered population of city *i*. The *t*-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

schools in these megacities. Most of these children remain in hometown while their parents leave to work in megacities. In many cases, these children are taken care of by their grandparents or family friends in hometown. According to Zhang, Li, and Xue (2015), these left-behind children are far behind their peers in terms of cognitive ability and mental health.

Second, the relationship between medical facilities and three population indicators shows a similar pattern. Table 6 reports the regression of the number of hospitals and beds on three population variables. Registered population still has the strongest explanatory power for the number of hospitals, with the *R*² of 0.558. And the number of hospital beds is best explained by NBS population, with the *R*² of 0.892. “Internet population” has the lowest explanatory power for both medical resource variables. In addition, the regression coefficients on “Internet population” are also much smaller than those of registered population and NBS population. These results also show that the allocation of medical facilities is mainly determined by the registered population and the resident population. The deviation in the estimation of the actual resident population will directly distort the spatial allocation of medical resources.

Finally, we investigate the correlation between fiscal revenue and expenditure and population measures. As Table 7 shows, the regression of fiscal revenue on “Internet population” has the highest *R*². However, the fiscal expenditure is more strongly correlated with NBS population than “Internet population”. This suggests that the fiscal revenue is mainly determined by the “Internet population”, a good measure to the actual population distribution. It also implies that the central government shifts substantial public resources across jurisdictions to mitigate spatial economic disparities. The amount of transfer payment is correlated with official population statistics. After the redistribution process, the fiscal expenditure is mainly determined by the NBS population, rather than actual population, which leads to the distortion in spatial distribution of public utility.

To summarize, the distortion of population distribution is gradually eased in the urbanization process. However, the distribution of public resources is still determined by biased population indicators, including registered population and NBS resident population. This distortion not only leads to a loss in economic efficiency, but also causes various social problems. On the one hand, there is a large number of underutilized classrooms and hospitals in remote rural areas. On the other hand, educational and medical resources are in

¹⁰ Fig. 6 shows the relationship between “Internet population” and city rank. The y-axis represents the log of the city rank ($\ln(Rank)$). The x-axis represents the log of the “Internet population”.

Table 6

Population distribution and allocation of medical resources.

	<i>Hospitals_i</i>			<i>Beds_i</i>		
<i>Internet_Population_i</i>	0.193*** (8.02)			40.389*** (22.36)		
<i>NBS_Population_i</i>		0.426*** (14.86)			68.143*** (44.11)	
<i>Registered_Population_i</i>			0.518*** (17.16)			72.185*** (28.73)
<i>Constant</i>	156.636*** (12.80)	70.793*** (5.63)	41.972*** (3.35)	9607.192*** (10.48)	−1013.348 (−1.50)	−1996.915* (−1.92)
Observations	235	235	235	237	237	237
R ²	0.216	0.486	0.558	0.680	0.892	0.778

This table reports OLS regressions of hospitals and hospital beds on different population measures, including “Internet Population”, NBS population and registered population. The dependent variable, *Hospitals_i* is the number of hospitals of city *i*. *Beds_i* is the number of hospital beds of city *i*. *Internet_Population_i* is “Internet Population” in city *i*. *NBS_Population_i* is the residents population of city *i* published by the NBS (National Bureau of Statistics). *Registered_Population_i* is the registered population of city *i*. The *t*-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

Table 7

Population distribution and allocation of fiscal resources.

	<i>Revenue_i</i>			<i>Expenditure_i</i>		
<i>Internet_Population_i</i>	101.49*** (27.49)			126.82*** (23.39)		
<i>NBS_Population_i</i>		110.97*** (18.01)			166.47*** (28.13)	
<i>Registered_Population_i</i>			83.25*** (10.79)			143.05*** (17.96)
<i>Constant</i>	−5474.83*** (−3.73)	−13,245.29*** (−5.42)	−4562.41 (−1.47)	8093.05*** (3.76)	−10,732.52*** (−4.57)	−3743.57 (−1.17)
Observations	232	232	232	232	232	232
R ²	0.767	0.585	0.336	0.704	0.775	0.584

This table reports OLS regressions of fiscal revenue and fiscal expenditure on different population measures, including “Internet Population”, NBS population and registered population. The dependent variable, *Revenue_i* is fiscal revenue of city *i*. *Expenditure_i* is fiscal expenditure of city *i*. *Internet_Population_i* is “Internet Population” in city *i*. *NBS_Population_i* is the residents population of city *i* published by the NBS (National Bureau of Statistics). *Registered_Population_i* is the registered population of city *i*. The *t*-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 99%, 95%, and 90% levels, respectively.

short supply in large cities. Millions of migrants have no access to public services in their cities of residence. In short, the distortion of population distribution is disappearing, but the distortion of public resources still persists.

6. Conclusion

This paper constructs new indicators of the city-level population, the “Internet population”, based on the mobile internet big data of social media users. Based on the Location Based Service (LBS) data of an almost total sample of Chinese working-age population, the “Internet population” is able to quantify and correct some biases of Chinese census data, especially the omission or misstatement of migrants.

Employing the “Internet population” indicator, we re-estimate the geographic population distribution in China and obtain three main findings. First, Chinese population mobility has far exceeded earlier estimations. Actually, the population distribution is very close to the prediction of Zipf’s Law, gradually approaching the theoretical steady state.

Second, the new indicator eliminates many statistical anomalies coexisting with the NBS population data. The “Internet population” is proved to be highly correlated with inelastic household consumption, such as water, electricity and garbage disposal. Meanwhile, NBS population has a much weaker explanatory power for household consumption than “Internet population”. It suggests that “Internet population” might be a better description of actual population distribution in China.

Third, the distribution of public resources is still determined by the biased population indicators, namely the registered population and the NBS population, which causes severe distortions. This result has very important policy implications. Based on the “Internet population”, policies of public resource allocation need to be reconsidered, such as the Hukou policy and fiscal policy. These are also important directions for future academic research.

¹¹ Fig. 7 shows the relationship between NBS population and city rank. The y-axis represents the log of the city rank ($\ln(\text{Rank})$). The x-axis represents the log of the NBS population.

This paper is a pioneer in constructing macro population variables by summing micro big data of billions of social media users. It has important implications for future research in this field. In the age of big data, there will be more and more interdisciplinary studies in economics and data science.

Declaration of Competing Interest

None.

References

- Anderson, G., & Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics*, 35(6), 756–776.
- Au, C. C., & Henderson, J. V. (2006). Are Chinese cities too small? *The Review of Economic Studies*, 73(3), 549–576.
- Behrens, K., Duranton, G., & Robert-Nicoud, F. (2014). Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy*, 122(3), 507–553.
- Bennett, M. M., & Smith, L. C. (2017). Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sensing of Environment*, 192, 176–197.
- Bertinelli, L., & Black, D. (2004). Urbanization and growth. *Journal of Urban Economics*, 56(1), 80–96.
- Brandt, L., & Zhu, X. (2010). "Accounting for China's Growth." Working Paper.
- Champernowne, D. G. (1953). A model of Income distribution. *The Economic Journal*, 63(250), 318–351.
- Chauvin, J. P., Glaeser, E., Ma, Y., & Tobio, K. (2017). What is different about urbanization in rich and poor countries? Cities in Brazil, China, India and the United States. *Journal of Urban Economics*, 98, 17–49.
- Chen, X., & Nordhaus, W. (2010). The value of luminosity data as a proxy for Economic statistics. *National Bureau of Economic Research, Working paper*, No. w16317.
- Desmet, K., & Rossi-Hansberg, E. (2013). Urban accounting and welfare. *American Economic Review*, 103(6), 2296–2327.
- Dobkins, L. H., & Ioannides, Y. M. (2001). Spatial interactions among US cities: 1900–1990. *Regional Science and Urban Economics*, 31(6), 701–731.
- Doll, C. (2008). CIESIN thematic guide to night-time light remote sensing and its applications.. *CIESIN Thematic Guide*. Center for International Earth Science Information Network of Columbia University, Palisades, NY.
- Erickson, E. P., & Kadane, J. B. (1985). Estimating the population in a census year 1980 and beyond. *Journal of the American Statistical Association*, 80(389), 98–109.
- Feng, N. (2010). Talk about demographics. *Statistical Education (China)*, 2(1).
- Gabaix, X. (1999a). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, 114(3), 739–767.
- Gabaix, X. (1999b). Zipf's law and the growth of cities. *American Economic Review*, 89(2), 129–132.
- Glaeser, E. L., Komnins, S. D., Luca, M., & Naik, N. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1), 114–137.
- Glaeser, E. L., & Gottlieb, J. D. (2009). The wealth of cities: Agglomeration economies and spatial equilibrium in the United States. *Journal of Economic Literature*, 47(4), 983–1028.
- Henderson, V., Storeygard, A., & Weil, D. N. (2011). A bright idea for Measuring Economic growth. *American Economic Review*, 101(3), 194–199.
- Krugman, P. (2004). Confronting the mystery of urban hierarchy. *Journal of the Japanese & International Economies*, 10(4), 399–418.
- Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. *Scientific Data*, 4(1), 1–17.
- Lu, L., & Duan, C. (2012). Summary and thoughts on the statistical survey of floating population in China. *Southern Population (China)*, 3, 73–80.
- Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-time light data: A good proxy measure for Economic activity? *PLoS One*, 10(10), Article e0139779.
- Meng, X., Jiang, Z., Wang, X., & Long, Y. (2021). Shrinking cities on the globe: Evidence from LandScan 2000–2019. *Environment and Planning A: Economy and Space*, 53(6), 1244–1248.
- Mera, K. (1973). On the urban agglomeration and Economic efficiency. *Economic Development and Cultural Change*, 21(2), 309–324.
- Niu, H. (2011). Analysis on the administrative system factors of the population census data error. *Legal System and Economy (China)*, 8, 192–193.
- Pang. (2010). Solving the difficult problem of the sixth National Census. *Population Research (China)*, 34(6), 19–31.
- Rosen, K., & Resnick, M. (1980). The size distribution of cities: an examination of the Pareto law and primacy. *Journal of Urban Economics*, 8(2), 165–186.
- Rossi-Hansberg, E., & Wright, M. L. (2007). Urban structure and growth. *The Review of Economic Studies*, 74(2), 597–624.
- Rozenfeld, H. D., Rybski, D., Gabaix, X., & Makse, H. (2011). The area and population of cities: New insights from a different perspective on cities. *American Economic Review*, 101(5), 2205–2225.
- Shen, J., Feng, Z., & Wong, K. Y. (2006). Dual-track urbanization in a transitional economy: The case of Pearl River Delta in South China. *Habitat International*, 30(3), 690–705.
- Siegel, J. S. (1974). Estimates of coverage of the population by sex, race, and age in the 1970 census. *Demography*, 11(1), 1–23.
- Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, 42(3/4), 425–440.
- Soo, K. (2014). Zipf, Gibrat and geography: Evidence from China, India and Brazil. *Papers in Regional Science*, 93(1), 159–181.
- Sutton, P., et al. (1997). A Comparison of Nighttime Satellite Imagery. *Photogrammetric Engineering & Remote Sensing*, 63(11), 1303–1313.
- Sutton, P. (2003). *14 Estimation of Human Population Parameters Using Night-Time Satellite Imagery*, 3. *Remotely-Sensed Cities*. USA: CRC Press.
- Wang, Q., et al. (2010). China's sixth population census: Experience and enlightenment. *Population Research (China)*, 34(6), 19–31.
- Winkler, R., Johnson, K. M., Cheng, C., Voss, P. R., & Curtis, K. J. (2013). County-specific net migration by five-year age groups, Hispanic origin. In *Race and sex 2000–2010*.
- Young, A. (2003). Gold into base metals: Productivity growth in the People's Republic of China during the Reform period. *Journal of Political Economy*, 111(6), 1220–1261.
- Zeng, C., Zhou, Y., Wang, S., Yan, F., & Zhao, Q. (2011). Population Spatialization in China based on night-time imagery and land use data. *International Journal of Remote Sensing*, 32(24), 9599–9620.
- Zhang, D., Li, X., & Xue, J. (2015). Education Inequality between Rural and urban areas of the People's Republic of China, Migrants' children education, and some implications. *Asian Development Review*, 32(1), 196–224.
- Zhang, Z., & Martinez-Vazquez, J. (2003). "The System of Equalization Transfers in China." No. paper0312. *International Center for Public Policy*. Andrew Young School of Policy Studies, Georgia State University.
- Zhu, X. (2012). Understanding China's growth: Past, present, and future. *Journal of Economic Perspectives*, 26(4), 103–124.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*.