Deviations from Zipf's Law for American cities: an empirical examination

Rafael, González-Val

Universidad de Zaragoza

10 November 2008

Deviations from Zipf's Law for American cities: an empirical examination

Rafael González-Val

Department of Economic Analysis

Universidad de Zaragoza

Abstract: This work presents a simple method for calculating deviations regarding city size and the size which would correspond to it with a Pareto exponent equal to one unit (Zipf's Law). Recent works show that when considering the entire sample without size restrictions, the estimated Pareto exponent tends to be much lower than one. Our aim is to analyse the distribution element by element, taking data from all American cities in 2000, and explain the deviation of the size predicted by Zipf's Law and the real size of each city, using variables for each city of per capita income, distribution of employment among sectors, individuals by level of education, etc.; explicative variables which attempt to capture the influence of local externalities. To do this a Multinomial Logit Model is used.

Keywords: Cities, Zipf's Law, deviations, Pareto distribution, Multinomial logit.

JEL: C16, R00, R12.

Address: Rafael González-Val,

Dpto. de Análisis Económico, Universidad de Zaragoza

Facultad de CC. Económicas y Empresariales

Gran Vía, 2, 50005 Zaragoza (Spain)

E-mail: rafaelg@unizar.es

1. Introduction

One of the stylised facts in urban economics is that the city size distribution in many countries can be approximated by a Pareto distribution, and it is an extensively studied empirical regularity that the parameter of this distribution, the Pareto exponent, is close to one[1]. This has given rise to theoretical developments explaining the fulfilment of Zipf's Law[2], justifying it analytically, associating it directly with an equilibrium situation and relating it to parallel growth (Gibrat's Law[3]). A large part of this literature takes as a reference the case of the United States, assuming a Pareto exponent equal to 1.

Gabaix [1999] presents a model based on local random amenity shocks, independent and identically distributed, which through migrations between cities generate Zipf's Law. The main contribution of the work is to justify the fulfilment of Zipf's Law in that the cities in the upper tail of the distribution follow similar growth processes, that is, that the fulfilment of Gibrat's Law involves Zipf's Law. Córdoba [2008] concludes that, under certain plausible conditions, Zipf's Law is equivalent to Gibrat's Law.

Rossi-Hansberg and Wright [2007] develop a model of urban growth which generates Zipf's Law in two restrictive cases (when physical capital is not used in production and productivity shocks are permanent, or when production is linear in physical capital and human capital is not used in production), and identifies the typical deviation of industrial productivity shocks as the key parameter which determines dispersion in the city size distribution. Eeckhout [2004] presents a model which also relates the migration of individuals between cities with productive shocks, obtaining as a result a lognormal and non-Paretian distribution of cities, although satisfying Gibrat's Law. Duranton [2006, 2007] offers a model of urban economy with endogenous growth based on knowledge spillovers which in the stationary state reproduce Zipf's Law for cities in the upper tail of the distribution; it also introduces some extensions which give empirically observed results (for example, a concave relationship between the rank and population logarithms).

To summarise, these theoretical models rest on local externalities, whether amenities or shocks in production or tastes, which must be randomly distributed independently of size, and identify deviations from Zipf's Law with a distribution of these shocks which is not independent of size. Other works also show the empirical relevance of other variables distributed clearly heterogeneously, such as climate or geographical advantages (access to the sea, bridges, etc), in the growth rate of cities.

These theoretical developments arise in response to numerous empirical works which explore the relationship between the growth rate and Zipf's Law. For a dynamic analysis, Ioannides and Overman [2003] use data from metropolitan areas from 1900 to 1990 and arrive at the conclusion that Gibrat's Law is fulfilled in the urban growth processes and that Zipf's Law is also fulfilled approximately well for a wide range of

[1] However, the values of the Pareto exponent vary greatly between countries (Rosen and Resnick [1980], Soo [2005]). And recent works demonstrate its sensitivity to the geographical unit chosen and the sample size (Eeckhout [2004]).

[2] Zipf's Law is an empirical regularity which appears when Pareto's exponent of the distribution is equal to one. The term was coined after a work by Zipf [1949], which observed that the frequency of the words of any language is clearly defined in statistical terms by constant values.

[3] Gibrat [1931] observed that the distribution of size (measured by sales or the number of employees) of firms tends to be lognormal, and his explanation was that the growth process of firms can be multiplicative and independent of the size of the firm.

city sizes. However, their results suggest that local values of Zipf's exponent can vary considerably with the size of cities. Nevertheless, Black and Henderson [2003] arrive at different conclusions for the same period (perhaps because they use different metropolitan areas). Zipf's Law would be fulfilled only for cities in the upper third of the distribution, while Gibrat's would be rejected for any sample size. These results highlight the extreme sensitivity of conclusions to the geographical unit chosen and to sample size. To close the debate, Eeckhout [2004] demonstrates that if we consider all the cities for the period 1990 to 2000 the city size distribution follows a lognormal rather than a Pareto distribution, so that the value of Zipf's parameter is not one, as earlier works concluded, but is about 0.5, and Gibrat's Law is also fulfilled for the entire sample.

Thus, if we accept that Zipf's Law is not fulfilled when considering the distribution of American cities, we can ask what factors explain this deviation from an empirical point of view. That is, we can analyse the distribution element by element and explain the deviation between the size predicted by Zipf's Law (associated with a Pareto exponent equal to one) and the real size of each city, using data on per capita income, employment distribution among sectors, individuals by levels of education, etc; variables which attempt to capture the influence of local externalities. This is the objective of this work, and for this data from the year 2000 are used, the first census in which the US Census Bureau offers data on all cities (places) without size restrictions.

This question has already been dealt with in the literature, but indirectly. On one hand Ioannides and Overman [2003] contrast the relationship between Zipf's and Gibrat's Laws for the United States using graphic and non-parametric methods, confirming the theoretical results of Gabaix [1999]: the explanation for the smaller cities' having a smaller Pareto exponent is that the variance of their growth rate is larger (deviations from Zipf's Law appear due to deviations in Gibrat's Law). On the other hand there are also works which explore the factors influencing growth rates. For the US, Glaeser and Shapiro [2001] study what factors influence the growth rate of American cities (cities of over 100,000 inhabitants and MSAs) using a very wide range of explicative variables (per capita income, average age of the residents, variables in the education level of individuals, temperature, distribution of employment among sectors, public spending per capita, etc.). According to this work, the three most relevant variables would be human capital, climate and transport systems for individuals (public or private).

The approach proposed in this work is simpler and empirically more direct. The only precedent would be the work of Soo [2005], insofar as it explains the differences in the Pareto exponent between different countries using such explicative variables as per capita income, area, population, transport costs, public spending, political variables, etc., with the important difference that as it uses Pareto's exponent per country as a dependent variable, it is comparing entire distributions, while we propose studying the deviations of each of the elements within a single distribution.

The next section sets out the method used to calculate deviations from Zipf's Law. Section 3 presents the variables used to try to explain the deviations. Section 4 shows the empirical model used, a Multinomial Logit Model (MNLM), and analyses the results obtained. The work ends with our conclusions.

2. Calculating the deviations

Let S be the city size, distributed according to a Pareto distribution. Then, following Eeckhout [2004], the density function $p(S)$ and the accumulated probability function $P(S)$ will be:

$$p(S) = \frac{a\underline{S}^a}{S^{a+1}}, \qquad \forall\, S \geq \underline{S}$$

$$P(S) = 1 - \left(\frac{\underline{S}}{S}\right)^a, \qquad \forall\, S \geq \underline{S}$$

where $a > 0$ is the Pareto exponent, $\underline{N}$ is the number of cities above the truncation point and $\underline{S}$ is the population of the city at the truncation point. The relationship with the rank R empirically observed is:

$$R = \underline{N} \cdot (1 - P(S)) = \underline{N} \cdot \left(\frac{\underline{S}}{S}\right)^a$$

Taking logarithms we obtain the linear specification usually estimated:

$$\ln R = \ln \underline{N} + a \ln \underline{S} - a \ln S + u = K - a \ln S + u\,,$$

where u represents a random error which we suppose to meet the standard conditions, $E(u) = 0$ and $Var(u) = \sigma^2$.

If Zipf's Law, $a = 1$, is strictly fulfilled, the above expression can be formulated in deterministic terms:

$$\ln R = \ln \underline{N} + \ln \underline{S} - \ln S^Z \rightarrow \ln S^Z = \ln \underline{N} + \ln \underline{S} - \ln R \qquad (1)$$

This expression can be directly brought back to the rank-size rule $\left(S = \frac{\overline{S}}{R}\right)$, where $\overline{S}$ is the population of the largest city:

$$\ln S^Z = \ln \underline{N} + \ln \underline{S} - \ln R \rightarrow \ln S^Z = \ln(\underline{S} \cdot \underline{N}) - \ln R = \ln \overline{S} - \ln R = \ln\left(\frac{\overline{S}}{R}\right),$$

although it is preferable to leave it in terms of the size of the smallest city, as the most populous is always bigger than predicted by Zipf's Law, for various reasons (especially political).

However, if the estimated parameter is other than 1 and the errors are not white noises, we would obtain an estimated size for each city:

$$\ln R = \ln \underline{N} + \hat{a} \ln \underline{S} - \hat{a} \ln S + \hat{u} \rightarrow \ln S = \frac{1}{\hat{a}} \cdot \ln \underline{N} + \ln \underline{S} - \frac{1}{\hat{a}} \cdot \ln R + \frac{1}{\hat{a}} \cdot \hat{u}\,. \quad (2)$$

Subtracting (2) from (1) we obtain a relationship between the size which fulfils Zipf's Law ($a = 1$ and $\ln S^Z$) with the real size of the city and the estimated value of the Pareto exponent ($\ln S$ and $\hat{a}$):

$$\ln S^Z = \ln S + \left(\frac{1}{\hat{a}} - 1\right)\left(\ln R - \ln \underline{N}\right) - \frac{1}{\hat{a}} \cdot \hat{u} ,$$

$$\ln\left(\frac{S^Z}{S}\right) = \left(\frac{1}{\hat{a}} - 1\right)\left(\ln R - \ln \underline{N}\right) - \frac{1}{\hat{a}} \cdot \hat{u} . \qquad (3)$$

Graphically $\ln\left(S^Z/S\right)$ represents the distance between the two distributions, the real one ($\ln S$) and the Paretian distribution corresponding to Zipf's Law ($\ln S^Z$). The upper part of Figure 1 represents the estimation for 2000 of both density functions through an adaptive kernel. The lower part shows the sample values of $\ln\left(S^Z/S\right)$. The calculation is done by applying (3) and using the entire distribution, 25,000 cities, from New York City with a population of 8,008,278 inhabitants to Paoli town with 42 inhabitants. The estimated[4] Pareto exponent is $\hat{a} = 0.534$.

As Eeckhout [2004] shows, the real city size distribution comes close to being a lognormal when all the sample is considered, and is found above the Pareto density function for almost all sample sizes. In fact, Wilcoxon's Rank-sum test offers a p-value of 0.4168 when considering the entire distribution, offering evidence in favour of the null hypothesis of lognormality[5].

However, for very small cities the behaviour is the opposite. This indicates that in general, cities will have a larger size than would guarantee the fulfilment of Zipf's Law, except for the smallest cities, whose size is much smaller than would correspond to a Pareto distribution. This can be seen in Table 1, showing the values of S^Z/S for the largest and smallest cities. It is also notable that for larger cities, the upper-tail distribution, deviations are reduced until they almost[6] disappear, agreeing with the general consensus that Zipf's Law is a phenomenon which mainly appears when considering larger cities. Recently Eeckhout [2008] shows that in the upper tail both distributions, Pareto and lognormal, can be valid.

Figure 1 also shows that there is a point where both density functions cross, after which real size is always larger than the size which would fulfil Zipf's Law, although in the upper tail of the distribution both density functions again become closer. In the sample, this point corresponds to cities with 310 inhabitants.

3. Data description

We use data for all cities in the Unites States, without imposing any minimum population truncation point, as our proposal is to cover the entire distribution. The source of data is the 2000 census[7]. We identify cities as what the US Census Bureau calls places. This generic denomination, since the 2000 census, includes all incorporated and unincorporated places.

---

[4] This value coincides with that obtained by Eeckhout [2004].

[5] Wilcoxon's test (rank-sum test) is a non-parametric test for assessing whether two samples of observations come from the same distribution. The null hypothesis is that the two samples are drawn from a single population, and therefore that their probability distributions are equal, in our case, the lognormal distribution. Wilcoxon's test has the advantage of being appropriate for any sample size.

[6] The estimated Pareto exponent is close to the value 1, but does not equal to unity. For the 100 largest cities, $\hat{a} = 1.32$.

[7] The US Census Bureau offers information on a wide range of variables for different geographical levels, available through its website: www.census.gov/main/www/cen2000.html.

The US Census Bureau uses the generic term incorporated place to refer to a type of governmental unit incorporated under state law as a city, town (except the New England states, New York, and Wisconsin), borough (except in Alaska and New York), or village and having legally prescribed limits, powers, and functions. On the other hand there are the unincorporated places (which were renamed Census Designated Places, CDPs, in 1980), which designate a statistical entity, defined for each decennial census according to Census Bureau guidelines, comprising a densely settled concentration of population that is not within an incorporated place, but is locally identified by a name. Evidently, the geographical boundaries of unincorporated places may change if settlements move, so that the same unincorporated place may have different boundaries in different census. They are the statistical counterpart of the incorporated places. The difference between them in most cases is merely political and/or administrative. Thus for example, due to a state law of Hawaii there are no incorporated places there; they are all unincorporated.

The US Census Bureau established size restrictions for the inclusion of unincorporated places, with the main criterion being that they have more than 1,000 inhabitants. The 2000 census is the first to include them all without size restrictions, and this is why we take only this year. However, there are no data for some of the explicative variables for all cities, slightly reducing the sample size to 23,519 cities. However, the range of city sizes is as wide as possible, from cities of 76 inhabitants to the largest, New York City, with a population of 8,008,278 inhabitants.

The chosen explicative variables coincide with those of other studies on urban growth in the United States and city size. These are variables whose influence on city size has been tested empirically by other works (see Glaeser and Shapiro [2001]), although our endogenous variable is completely different. We can group them in three types of variables: local external effects variables, human capital variables and productive structure variables. Table 2 presents the variables and gives some descriptive statistics. It is notable that in general the typical deviations are fairly high, showing great heterogeneity among the variables chosen when considering all places.

The variables in local external effects basically aim to gather some of the costs of urban congestion. In the first place we monitor the economic size of the city using Per capita income in 1999; it would make no sense to include the population again, as it was already used to calculate deviation. We also include two variables which reflect the age of the city: the variable “Percent housing units built 1939 or earlier” which we as a proxy for the physical age of cities, and the variable “Total population: median age”, which reflects the age of the city’s inhabitants.

One of the most typical congestion costs is the increased cost of housing as the city size increases (taking into account that the supply of housing tends to be fairly rigid and responds slowly to increased demand). Glaeser et al. [2006] analyses the role of the housing supply in urban and productivity growth in the USA. We attempt to capture this effect through the variable “Percent owner-occupied housing units with a mortgage; contract to purchase; or similar debt”, as it is to be expected that as housing prices rise more individuals will be obliged to incur mortgages or other debts. Commuting costs are another characteristic congestion cost of urban growth and are explicitly included in some theoretical models. That is, the idea that as the population of a city grows, so do the costs in terms of time for individuals to get from their homes to their places of work. To capture this effect we use the variable “Workers 16 years and over who did not work at home: Median travel time to work (in minutes)”.

The last two variables in this group refer to the division produced in United States cities depending on whether they are built around public transport or private cars. As Glaeser and Kahn show [2003, 2001], in the last few decades the model of United States cities has been characterised by being built around private cars, while public transport loses importance.

Regarding the human capital variables, there are many works demonstrating the influence of human capital on city size, as cities with individuals with higher levels of human capital tend to grow more. We take two human capital variables: "Percent population 25 years and over: High school graduate (includes equivalency) or higher degree" and "Percent population 25 years and over: Some college or higher degree". The former represents a wide concept of human capital, while the second centres on high educational levels (some college, Associate degree, Bachelor's degree, Master's degree, Professional school degree, and Doctorate degree).

The third group of variables, referring to productive structure, contains the sectorial distribution of employment. The distribution of work among the different productive activities provides valuable information on other aspects of the city. Thus, the level of employment in the primary sector (agriculture; forestry; fishing and hunting; and mining) also represents by proxy the natural physical resources of the city (farming land, sea, etc.). This is also a sector which, like construction, is characterised by constant or even decreasing returns to scale.

Employment in manufacturing informs us of the level of local economies of scale in production, as this is a sector which normally presents increasing returns to scale. The level of pecuniary externalities also depends on the size of the industrial sector. Marshall put forward that (i) the concentration of companies of a single sector in a single place creates a joint market of qualified workers, benefiting both workers and firms; (ii) an industrial centre enables a larger variety at a lower cost of concrete factors needed for the sector which are not traded, and (iii) an industrial centre generates knowledge spillovers. This approach forms part of the basis of economic geography models, along with circular causation: workers go to cities with strong industrial sectors, and firms prefer to locate nearer larger cities with bigger markets. Thus, industrial employment also represents a measurement of the size of the local market. Another proxy for the market size of the city is the employment in commerce, whether retail or wholesale.

4. Empirical model and results

4.1 Empirical model

Unfortunately we have data for a single period only, the year 2000, as the census for 2000 is the first to include all places without size restrictions, and we wanted to consider the entire sample. This involves possible endogeneity and simultaneity problems for any regression we might attempt.

Also, our endogenous variable, deviation from the size which satisfies Zipf's Law, presents two clearly differentiated behaviours, so that the interpretation of the influence of either of the explicative variables cannot be unequivocal, as happens with standard regressions. We define deviations from the size of Zipf's Law as $\ln\left(S^Z/S\right)$, and they are calculated applying (3). This specification implies that deviation for cities with a larger size than would fulfil Zipf's Law will have a negative value. This is the majority case, as shown in Figure 1. Concretely, of 23,519 cities in the sample, 18,874 present a negative deviation (80.25 %). Meanwhile, for the remaining 4,645 cities

(19.75 % of the sample) the value of the deviation is positive, as their size is less than would fulfil Zipf's Law.

All of this leads us to use a Multinomial Logit Model (MNLM), which solves all the problems described above. It consists of transforming our dependent variable into categories, enabling us to differentiate specifically between the two behaviours observed (positive and negative deviations) and at the same time solves the possible problems of endogeneity and simultaneity which could arise when considering a single period. With the consequence that the results of the estimations will give us information about the probability (while not causality) with which each variable affects each category.

Based on the deviations, $\ln\left(S^Z/S\right)$, calculated based on (3), we construct four categories $\left(K=1,2,3,4\right)$ applying the following criterion, taking into account that as shown in Figure 1, $\ln\left(S^Z/S\right)$ ranges from -2 to 2:

$$\begin{cases} \text{Strong negative deviation} \rightarrow \quad K=1 \text{ if } -2<\ln\left(S^Z/S\right)<-1.2 & \text{(7,039 places)} \\ \text{Medium negative deviation} \rightarrow K=2 \text{ if } -1.2<\ln\left(S^Z/S\right)<-0.6 & \text{(7,007 places)} \\ \text{Weak negative deviation} \rightarrow \quad K=3 \text{ if } -0.6<\ln\left(S^Z/S\right)<0 & \text{(4,828 places)} \\ \text{Positive deviation} \rightarrow \quad K=4 \text{ if } 0<\ln\left(S^Z/S\right)<2 & \text{(4,645 places)} \end{cases}$$

This criterion enables us to differentiate between the cities presenting a negative deviation (80.25 %), whose size is greater than that predicted by Zipf's Law –grouped in categories 1, 2 and 3– and the cities (19.75 %) for which the deviation value is positive, as their size is less than that which would fulfil Zipf's Law. These particular grouping also ensures that the groups will be as homogeneous as possible in size.

With the MNLM we estimate a separate binary logit for each pair of categories of the dependent variable. Formally, the MNLM can be written as:

$$\ln \phi_{m|b} = \ln \frac{\Pr\left(K=m|\mathbf{x}\right)}{\Pr\left(K=b|\mathbf{x}\right)} = \mathbf{x}'\beta_{m|b} \qquad \text{for} \quad m=1 \text{ a } J\text{,} \qquad (4)$$

where b is the base category (in our case this will be category 1, as it contains more cities), $J=4$ and $\mathbf{x}$ is the vector of the explicative variables, reflecting the local external effects, human capital or productive structure[8]. We propose to study how these explicative variables affect the odds of a city being in one or the other category, that is, presenting a positive or negative deviation (greater or smaller). For example, if the percentage of the population with higher education increases (Percent population 25 years and over: Some college or higher degree), does this increase the probability of the city size being larger than the size it would have if Zipf's Law were fulfilled? And if so,

[8] The MNLM makes the assumption known as the independence of irrelevant alternatives (IIA). In this model: $\ln \frac{\Pr\left(K=m|\mathbf{x}\right)}{\Pr\left(K=n|\mathbf{x}\right)} = e^{\mathbf{x}'\left(\beta_{m|\mathbf{b}}-\beta_{n|\mathbf{b}}\right)}$, where the odds between each pair of alternatives do not depend on other available alternatives. Thus, adding or deleting alternatives does not affect the odds between the remaining alternatives. The assumption of independence follows from the initial assumptions that the disturbances are independent and homoscedastic. We have considered one of the most common tests developed for testing the validity of the assumption, the Small-Hsiao test [1985], and we could not reject the null hypothesis, that is, the odds are independent of other alternatives, indicating that the MNLM is appropriate.

will we be able to know if this is a strong, medium or weak deviation (which of the three categories with a negative deviation will be most likely)?

To deal with these questions we use odds ratios (also known as factor change coefficients). Maintaining the other variables constant, the change in the odds of the outcome m against outcome n, when x_i increases by δ, equals:

$$\frac{\phi_{m|b}(\mathbf{x}, x_i + \delta)}{\phi_{n|b}(\mathbf{x}, x_i)} = e^{\beta_{i,m|n}\delta}. \qquad (5)$$

Thus, if $\delta = 1$ the odds ratio can be interpreted as follows: for each unitary change of x_i it is expected that the odds of m versus n change by a factor $e^{\beta_{i,m|n}}$, maintaining the other variables constant.

4.2 Results

This model includes many coefficients, making it difficult to interpret the effects for all pairs of categories. To simplify the analysis odds-ratio plots were developed, shown in Figures 2, 3 and 4 for each of the three groups of variables. To analyse the effect of each variable in the change in probability of a city being in one category or another, Table 3 shows the marginal effects for each category and the absolute average change in probability.

In an odds-ratio plot, the independent variables are each represented on a separate row, and the horizontal axis indicates the relative magnitude of the β coefficients associated with each outcome. The numbers which appear (1, 2, 3 or 4) are the four possible outcomes, the categories which we previously constructed.

These graphs reveal a great deal of information (for more details see Long and Freese [2006]). To begin, if a category is to the right of another, it indicates that increases in the independent variable make the outcome to the right more likely. Also, the distance between each pair of numbers indicates the magnitude of the effect. And when a line connects a pair of categories this indicates a lack of statistical significance for this particular coefficient, suggesting that these two outcomes are tied together. The three graphs take as a base category outcome 1 (as this has most cities).

External local effects variables

Table 3 shows that the variable presenting the greatest absolute average change in probability (0.0428) is per capita income in 1999. Also, Figure 2 shows how given an increase of 1 unit in the logarithm of per capita income the most likely category is, markedly, 1 (strong negative deviation). This means that increases in the per capita income of the city increase the probability of a strong negative deviation, that is, that larger cities in economic terms will probably be cities with a much larger population than predicted by Zipf's Law.

Regarding the variable which we use to try to reflect commuting costs, "Workers 16 years and over who did not work at home: Median travel time to work (in minutes)", at first glance the effect is the opposite of what we expected. In principle, the bigger the size of the city, the longer the median travel time which workers must bear. However, Figure 2 shows category 4 as more likely, which would indicate that given an increase of one unit of the median travel time the most likely outcome is that the size of this city will be less than would correspond with a Pareto exponent equal to one. Therefore, this probability must be interpreted the other way around: the probability of the median

travel time to work increasing is greater in smaller cities, as in very big cities it is very possible that commuting costs are close to their maximum value.

It should also be noted that the two variables used as proxies for the age of the cities present very similar behaviour. In both cases, the greater the average age of the total population or "Percent housing units: Built 1939 or earlier" the greater the probability of the city presenting a positive deviation (category 4). That is, the cities with the oldest individuals or which were founded before "1939 or earlier" have a greater probability of having a population lower than would correspond with a Pareto exponent equal to one.

Neither do the two variables introduced to reflect the division produced in US cities depending on whether they were built around public transport or private cars show clearly differentiated behaviour. The signs of the marginal effects by category coincide in both variables, although the variable "Percent workers 16 years and over: Public transportation" presents a higher absolute average change in probability (0.0056 versus 0.0016). And Figure 2 shows how for both variables the most likely outcome is a medium negative deviation (category 2)**.**

Finally, the variable "Percent owner-occupied housing units with a mortgage; contract to purchase; or similar debt", which we use as a proxy for urban congestion costs through housing prices, presents the expected behaviour. As the price of housing increases, more individuals are obliged to resort to mortgages or similar debts. Figure 2 shows category 1 (strong negative deviation) as the most likely outcome; this indicates that the cities with very high housing prices, and thus a high congestion cost, are a long way from what would be their size as predicted by Zipf's Law.

Human capital variables

The results show opposing behaviour for the two human capital variables we introduced. Thus, the signs of the marginal effects by category (Table 3) are the opposite, and Figure 3 shows how the most likely outcomes are the opposite categories, a strong negative deviation and a positive deviation (1 and 4). Thus, increases in the more educated percentage of the population makes it more likely that the city size is much larger than the size which would fulfil Zipf's Law, while if there is a higher percentage with further education in its wider human capital sense (Percent population 25 years and over: High school graduate (includes equivalency) or higher degree) the most likely outcome is that the city will be smaller than would correspond with a Pareto exponent equal to one.

This result must be seen in relation to that obtained for per capita income, as education is usually closely related to per capita income. We have seen how with increases in both variables the most likely outcome is that city size will be much higher than predicted by Zipf's Law, in agreement with the results of other studies. Simon and Nardinelli [2002] analyse the period 1900-1990 for the USA and conclude that cities with individuals with greater levels of human capital tend to grow more, and Glaeser and Saiz [2003] analyse the period 1970-2000 and show that this is due to skilled cities being more productive economically.

Productive structure variables

Table 3 shows that the sector of activity presenting the greatest absolute average change in probability (0.0131) is the primary sector (agriculture; forestry; fishing and hunting; and mining). If we interpret this variable as a proxy for the natural physical resources available to the city (farmland, sea, etc.), Figure 4 shows category 4 (positive

deviation) as the most likely outcome by a large margin. That is, more natural resources and higher employment in the primary sector mean a higher probability of city size being lower than would fulfil Zipf's Law. This result coincides with the traditional interpretation of employment in the agricultural sector in theoretical models as a force for dispersion of economic activity, since the pioneering work of Krugman [1991].

The other employment sector usually identified as a dispersing force is construction. The results show that the variable "Percent employed civilian population 16 years and over: Construction" has a similar effect. Figure 4 shows category 4 (positive deviation) as the most likely outcome. Thus, the larger the percentage of labour employed in construction, the greater the probability that city size will be less than would correspond to a Pareto exponent equal to one. Although in Figure 4 categories 3 and 4 are joined by a line, indicating that an increase of 1% in "Percent employed civilian population 16 years and over: Construction" makes outcome 4 more likely than categories 1 and 2, regarding category 3 (weak negative deviation) the effect is not significant.

However, in the case of employment in manufacturing, a sector which usually presents economies of scale, Figure 4 shows category 3 as the most likely outcome, a weak negative deviation (although the effect on category 4 is not significant). Thus, an increase in industrial employment increases the probability of the city being larger than would correspond with a Pareto exponent equal to one, although the deviation from Zipf's Law will be small.

In services, we can see differentiated behaviour. Increases in the percentage of employment dedicated to finance; insurance; real estate, rental and leasing; and wholesale and retail trade increase the probability of the city size being much bigger than the size which would fulfil Zipf's Law (strong negative deviation), while if employment increases in educational, health and social services or in Public administration the most likely outcome is a weak negative deviation (category 3). Again, this result must be seen in relation to that obtained for per capita income, as the activities finance; insurance; real estate, rental and leasing; and wholesale and retail trade depend directly on the size of the local market, so that the percentage of employment in these activities will be higher in cities with higher per capita income. In contrast, cities with lower per capita income will have a higher employment percentage in social services.

5. Conclusions

Eeckhout [2004] demonstrates that, considering the entire sample, in 2000 the distribution of size of US cities follows a lognormal, and not a Paretian, distribution. In this work we present a simple method for calculating deviations city by city in relation to their size and the size which would correspond with a Pareto exponent equal to one (Zipf's Law). Our objective is to analyse the distribution element by element and explain the deviation from Zipf's Law using data for each city of per capita income, distribution of employment among sectors, individuals by level of education, etc.; variables which try to capture the influence of local externalities. For this a Multinomial Logit Model is used, enabling us to know the influence of each of these variables in terms of probability.

The results show two differentiated behaviours. Of the 23,519 cities of the sample, 18,874 present a negative deviation (80.25 %), meaning they present a greater size than that which would fulfil Zipf's Law. The variables increasing the probability of cities presenting this type of deviation are Per capita income in 1999, Percent owner-

occupied housing units with a mortgage; contract to purchase; or similar debt (which we use as a proxy for the cost of urban congestion through housing cost), higher levels of human capital (percent population 25 years and over: Some college or higher degree), and employment in certain services (finance; insurance; real estate, rental and leasing and Wholesale and Retail trade).

Meanwhile, the size of the remaining 4,645 cities (19.75 % of the sample) is lower than would fulfil Zipf's Law (which we define as a positive deviation). In this case the variables raising the probability of presenting a positive deviation are the variables measuring the age of the city (whether of the inhabitants, Total population: Median age, or the physical age of the buildings, Percent housing units: Built 1939 or earlier), the percentage of the population educated from a wider human capital point of view (Percent population 25 years and over: High school graduate (includes equivalency) or higher degree), and employment in productive sectors with constant or decreasing returns to scale (agriculture; forestry; fishing and hunting; mining, and construction).

References

[1] Black, D., and V. Henderson, [2003]. Urban evolution in the USA. Journal of Economic Geography 3, pp. 343-372.

[2] Córdoba, J. C., [2008]. A Generalized Gibrat's Law. International Economic Review, Forthcoming, November 2008.

[3] Duranton, G., [2006]. Some Foundations for Zipf´s Law: Product Proliferation and Local Spillovers. Regional Science and Urban Economics, vol. 36, pages 542-563.

[4] Duranton, G., [2007]. Urban Evolutions: The Fast, the Slow, and the Still. American Economic Review 97(1), pp. 197-221.

[5] Eeckhout, J., [2004]. Gibrat's Law for (All) Cities. American Economic Review, American Economic Association, vol. 94(5), pages 1429-1451.

[6] Eeckhout, J., [2008]. Gibrat's Law for (All) Cities: Reply. American Economic Review, forthcoming.

[7] Gabaix, X., [1999]. Zipf's Law for cities: An explanation. Quarterly Journal of Economics, 114(3):739-767.

[8] Glaeser, E. L., J. Gyourko, and R. E. Saks, [2006]. Urban growth and housing supply. Journal of Economic Geography 6 (2006) pp. 71–89.

[9] Glaeser, E. L., and M. E. Kahn, [2001]. Decentralized Employment and the Transformation of the American City. Brookings-Wharton Papers on Urban Affairs: 1-47.

[10] Glaeser, E. L., and M. E. Kahn, [2003]. Sprawl and Urban Growth. Harvard Institute of Economic Research, Discussion Paper number 2004.

[11] Glaeser, E. L., and A. Saiz, [2003]. The Rise of the Skilled City. Harvard Institute of Economic Research, Discussion Paper number 2025.

[12] Glaeser, E. L., and J. Shapiro, [2001]. Is there a new urbanism? The growth of US cities in the 1990's. Harvard Institute of Economic Research, Discussion Paper number 1925.

[13] Gibrat, R., [1931]. Les inégalités économiques. Paris: Librairie du Recueil Sirey.

[14] Iaonnides, Y. M., and H. G. Overman, [2003]. Zipf's Law for cities: an empirical examination. Regional Science and Urban Economics 33, 127-137.

[15] Krugman, P., [1991]. Increasing returns and economic geography. Journal of Political Economy 99, 483-499.

[16] Long, J. S., and J. Freese, [2006]. Regression Models for Categorical Dependent Variables Using Stata. Stata Press Publication, 2nd ed.

[17] Rosen, K. T., and M. Resnick, [1980]. The Size Distribution of Cities: An Examination of the Pareto Law and Primacy. Journal of Urban Economics, vol. 8, pages 165-186.

[18] Rossi-Hansberg, E., and M. L. J. Wright, [2007]. Urban structure and growth. Review of Economic Studies 74, 597–624.

[19] Simon, C. J., and C. Nardinelli, [2002]. Human capital and the rise of American cities, 1900-1990. Regional Science and Urban Economics, 32: 59-96.

[20] Small, K.A., and C. Hsiao, [1985]. Multinomial logit specification tests. International Economic Review 26, 619-627.

[21] Soo, K. T., [2005]. Zipf's Law for cities: a cross-country investigation. Regional Science and Urban Economics, 35: 239-263.

[22] Zipf, G., [1949]. Human Behaviour and the Principle of Least Effort. Cambridge, MA: Addison-Wesley.

Tables

Table 1.- Deviations for the ten biggest and smallest cities

Ranking	City	S	$\left(S^Z/S\right)$
1	New York City	8,008,278	0.787
2	Los Angeles	3,694,820	0.853
3	Chicago	2,896,016	0.726
4	Houston	1,953,631	0.807
5	Philadelphia	1,517,550	0.831
6	Phoenix	1,321,045	0.795
7	San Diego	1,223,400	0.736
8	Dallas	1,188,580	0.663
9	San Antonio	1,144,646	0.612
10	Detroit	951,270	0.663
24,991	Stotesbury city	43	5.870
24,992	Antelope CDP	43	5.870
24,993	Saltaire village	43	5.870
24,994	Braddock city	43	5.869
24,995	Regan city	43	5.869
24,996	Atlantic CDP	43	5.869
24,997	Hetland city	43	5.869
24,998	Washam CDP	43	5.868
24,999	McCarthy CDP	42	6.008
25,000	Montezuma city	42	6.008

Note:

S : City Population in 2000 (Source: US Census Bureau),

S^Z : Population which would correspond with a Pareto exponent equal to 1.

Table 2.- Descriptive statistics

Variables	Average	Typ. Dev.	Minimum	Maximum
External local effects variables				
Per capita income in 1999	18947.70	9713.34	1539	200087
Total population: Median age	37.32	6.60	10.80	79.20
Percent housing units: Built 1939 or earlier	22.42	19.00	0	97.88
Percent owner-occupied housing units with a mortgage; contract to purchase; or similar debt	47.96	17.39	0	100
Workers 16 years and over who did not work at home: Median travel time to work (in minutes)	24,45	6,94	2,59	109,05
Percent workers 16 years and over: Car; truck; or van; Drove alone	76.93	10.37	0	100
Percent workers 16 years and over: Public transportation	1.36	3.25	0	57.16
Human capital variables				
Percent population 25 years and over: Some college or higher degree	43.79	16.89	0	99.57
Percent population 25 years and over: High school graduate (includes equivalency) or higher degree	78.30	12.32	5.11	100
Productive structure variables				
Percent employed civilian population 16 years and over:				
Agriculture; forestry; fishing and hunting; and mining	3.52	5.40	0	72.75
Construction	7.62	4.15	0	40.32
Manufacturing	16.31	10.32	0	70.63
Wholesale and Retail trade	15.27	4.69	0	67.86
Finance; insurance; real estate and rental and leasing	5.16	3.54	0	46.67
Educational; health; and social services	20.32	7.20	0	87.18
Public administration	5.21	4.19	0	60.71

Nota: All the variables correspond to 2000, except the per capita income in 1999. Source: US Census Bureau.

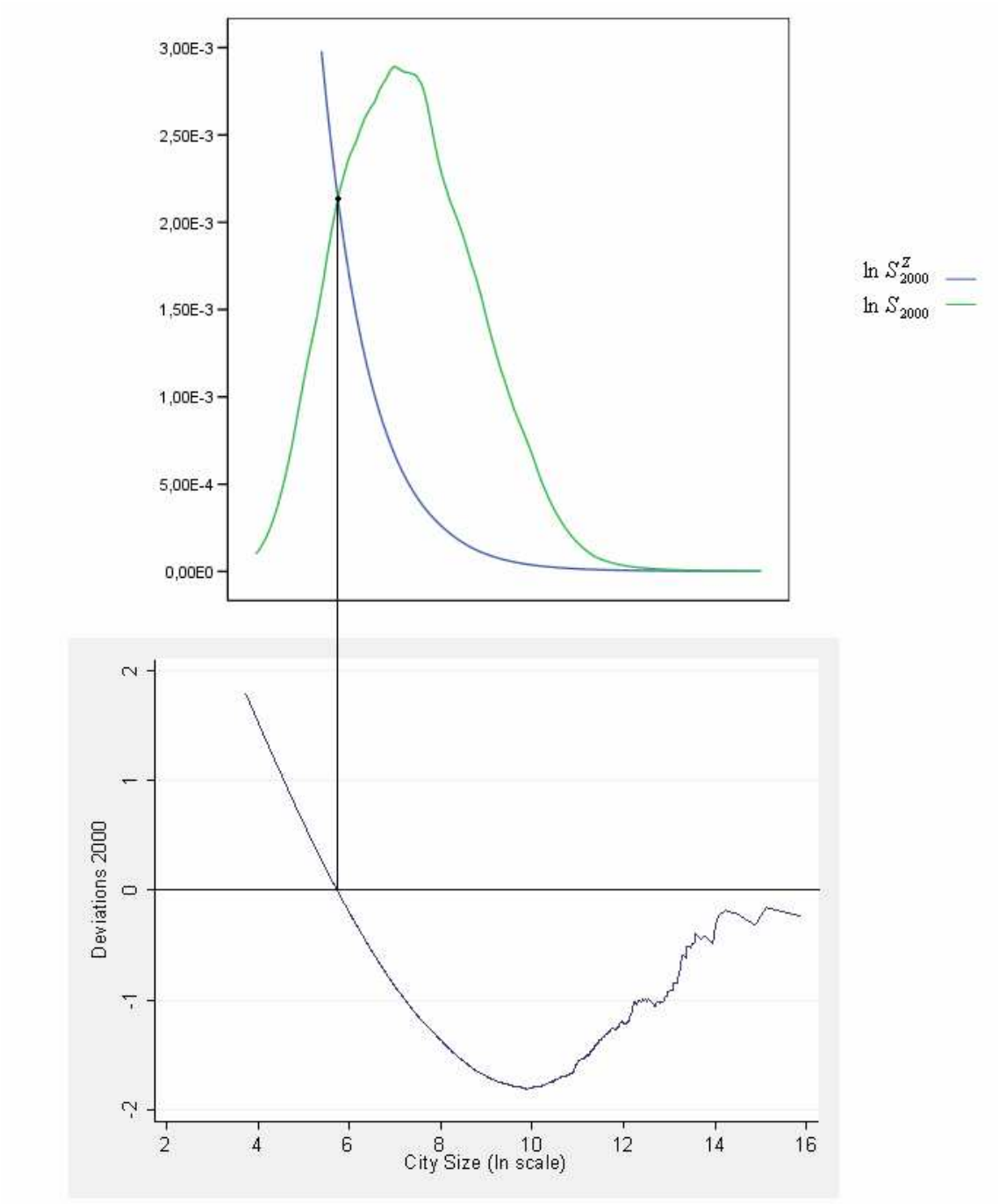
Table 3.- Marginal effects for each category and the average absolute change in the probability

	Categories				
	1	2	3	4	Total average
External local effects variables					
Log (Per capita income in 1999)	0.0856***	-0.0231***	-0.0301***	-0.0324***	0.0428***
Total population: Median age	-0.0070***	0.0033***	0.0017***	0.0020***	0.0035***
Percent housing units: Built 1939 or earlier	-0.0050***	-0.0021***	0.0028***	0.0043***	0.0035***
Percent owner-occupied housing units with a mortgage; contract to purchase; or similar debt	0.0082***	0.0030***	-0.0047***	-0.0065***	0.0056***
Workers 16 years and over who did not work at home: Median travel time to work (in minutes)	-0.0104***	-0.0025***	0.0069***	0.0061***	0.0065***
Percent workers 16 years and over: Car; truck; or van; Drove alone	0.0003	0.0028**	-0.0017***	-0.0014***	0.0016***
Percent workers 16 years and over: Public transportation	0.0001	0.0110***	-0.0014	-0.0097***	0.0056***
Human capital variables					
Percent population 25 years and over: Some college or higher degree	0.0046***	0.0025***	-0.0041***	-0.0030***	0.0035***
Percent population 25 years and over: High school graduate (includes equivalency) or higher degree	-0.0114***	-0.0046***	0.0088***	0.0071***	0.0080***
Productive structure variables					
Percent employed civilian population 16 years and over:					
Agriculture; forestry; fishing and hunting; and mining	-0.0262***	0.0020***	0.0139***	0.0103***	0.1310***
Construction	-0.0197***	0.0034***	0.0105***	0.0057***	0.0098***
Manufacturing	-0.0078***	0.0008***	0.0049***	0.0021***	0.0039***
Wholesale and Retail trade	0.0033***	-0.0006***	0.0001**	-0.0028***	0.0017***
Finance; insurance; real estate and rental and leasing	0.0060***	0.0029**	-0.0021***	-0.0068***	0.0045***
Educational; health; and social services	-0.0035***	0.0033***	0.0031***	-0.0030*	0.0032***
Public administration	-0.0061***	0.0015***	0.0033***	0.0013***	0.0031***

***Significant at the 1% level, **Significant at the 5% level, *Significant at the 10% level

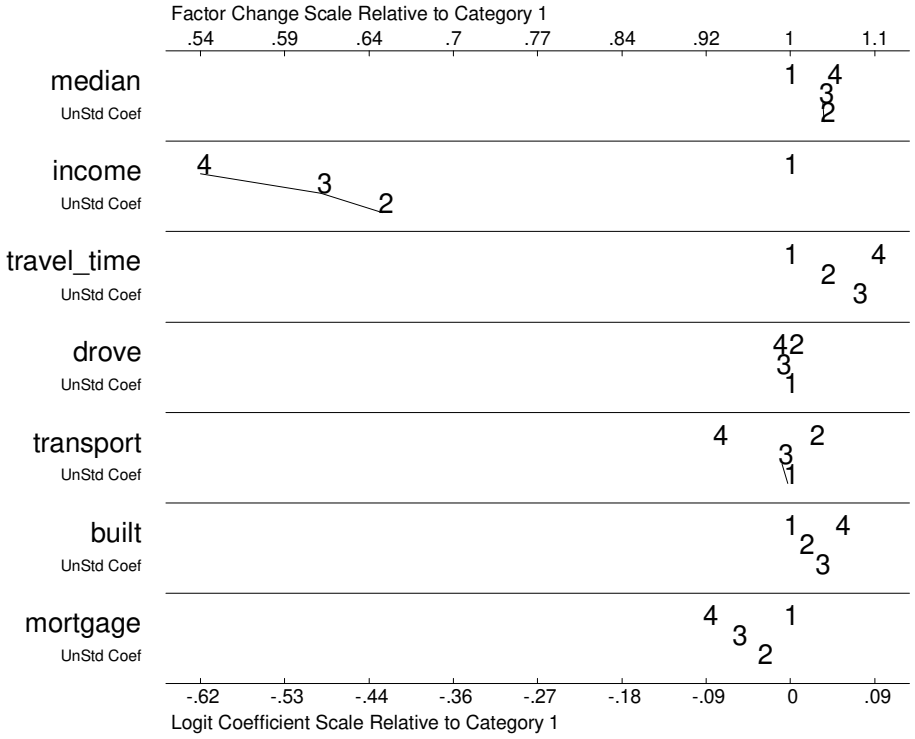
Figures

Figure 1.- Relationship between Size (scale ln), Size fulfilling Zipf's Law (scale ln) and S^Z/S (scale ln)



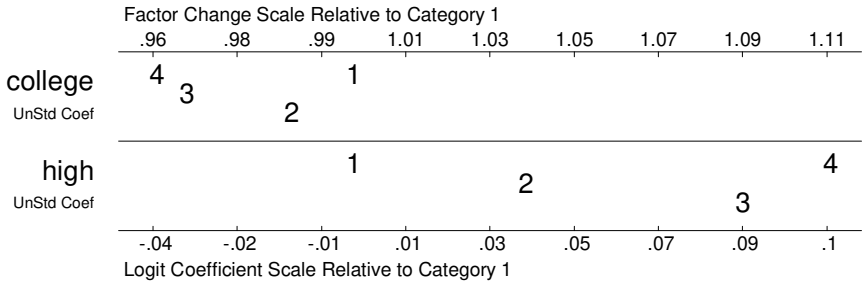
Note: The upper figure corresponds to the adaptive kernels estimated for $\ln S^Z$ and $\ln S$ in the year 2000, while the lower figure represents the sample values of $\ln\left(S^Z/S\right)$ calculated applying (3).

Figure 2.- Odds-ratio plot of external local effects variables



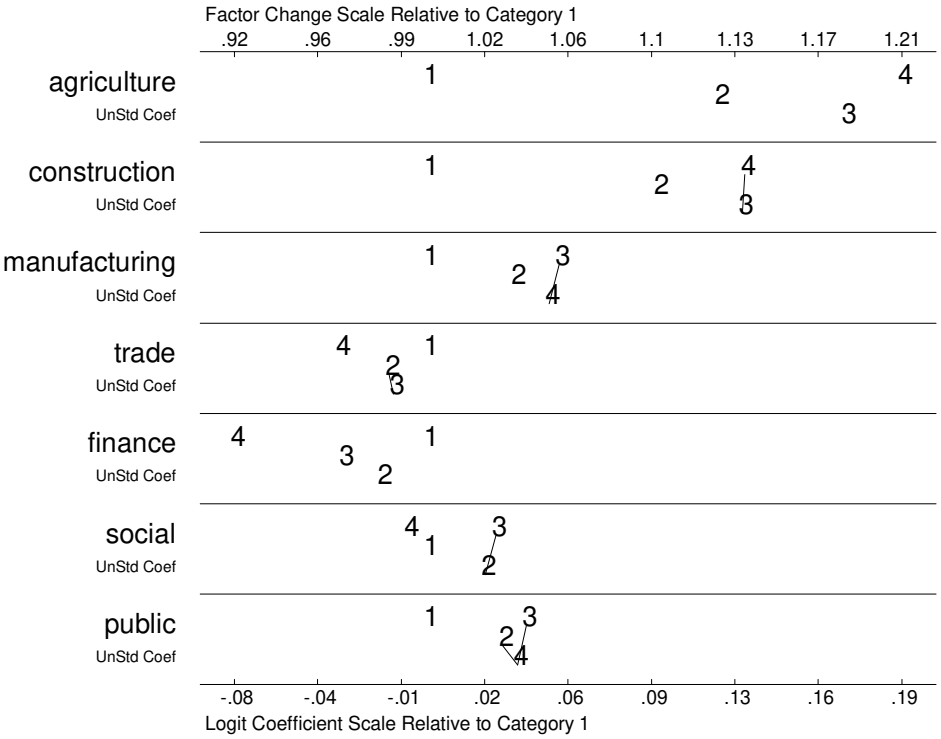
Key:
Income: Log (Per capita income in 1999)
Media: Total population: Median age
Built: Percent housing units: Built 1939 or earlier
Mortgage: Percent owner-occupied housing units with a mortgage; contract to purchase; or similar debt
Travel_time: Workers 16 years and over who did not work at home: Median travel time to work (in minutes)
Drove: Percent workers 16 years and over: Car; truck; or van; Drove alone
Transport: Percent workers 16 years and over: Public transportation

Figure 3.- Odds-ratio plot of human capital variables



Key:
College: Percent population 25 years and over: Some college or higher degree
High: Percent population 25 years and over: High school graduate (includes equivalency) or higher degree

Figure 4.- Odds-ratio plot of productive structure variables



Key:
Percent employed civilian population 16 years and over:
Agriculture: Agriculture; forestry; fishing and hunting; and mining
Construction: Construction
Manufacturing: Manufacturing
Trade: Wholesale and Retail trade
Finance: Finance; insurance; real estate and rental and leasing
Social: Educational; health; and social services
Public: Public administration