Zipf's law applicability in geographical language variation and empirical data clustering.

Salvatore Raia, David Sánchez, Luis Seoane
Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB)
Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain

Abstract

It is well known that the frequency of use for words in a particular language decreases with the work rank. This behavior can be approximated by a Zipf distribution. Claims for the universality of the Zipf law that have appeared in the literature, are based on general corpora that do not take into account language variation. In this work, we investigate the variation of the statistical distribution for English word usage across space. For that purpose, we consider a large dataset built from geolocalized Twitter messages. Using [here comes a summary the methods]... We find [here comes a summary of the results]...

1 Introduction

A Zipfian behaviour is s

[1, **?**]. A discussion on the power law we want to verify and the two important parameter, α, x_{min}.

2 Methods

All the analysis will be performed by working on the frequencies of words, and considering separately all the counties. This means that each county has its own list of (sorted) frequencies for all the words used in that county.

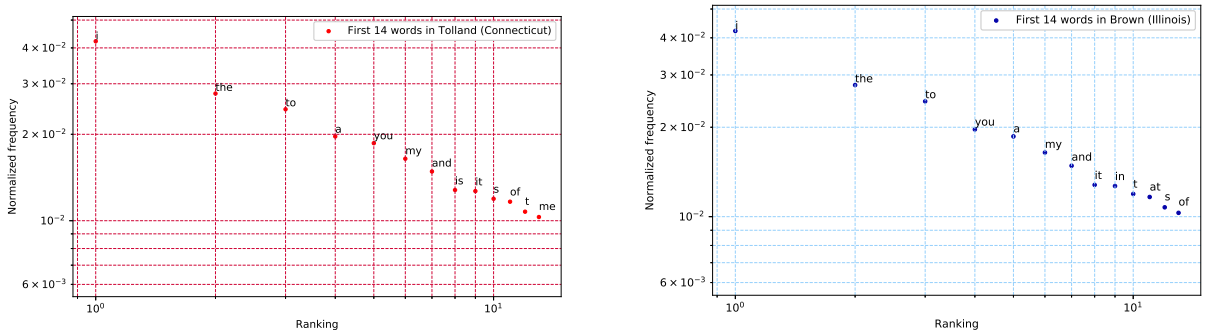


Figure 1:

2.1 Part 1

As we expect a power-law behavior [citation] here we give the expression for the probability density function, the PDF that in countinuos case is [cit]:

$$p(x) = cx^{-\alpha} \tag{1}$$

$$c = \frac{(\alpha - 1)}{x_{min}^{1-\alpha}} \tag{2}$$

Where α is a constant parameter specific to that distribution called *scaling parameter* [cit] and c instead represents the normalization factor that dependendt on x_{min}. x_{min} is a lower bound for the power-law behavior in order to avoid the PDF diverging when x approaches to 0. This means that the data are power-law distributed for x values greater than x_{min}[cit].
A continuos approximation for the PDF function is consistent with the large sample of data and really usefull as it is less computational expensive. [cit]

2.1.1 Linear regression

The frequencies will be plotted in the typical frequencies-ranking graph [citation?] and after setting a log-log scale, we expect to find linearity. This will be reasulting in $-\alpha$ to be the angular coefficient of that straight line and $\log_{10} c$ its intercept.

$$\log_{10} p(x) = -\alpha \log_{10} x + \log_{10} c \tag{3}$$

Calling Y the left side of the equation and $q = \log_{10} c$ we have a straight line function

$$Y = -\alpha X + q \tag{4}$$

In order to first extimate the magnitude of α a first and less accurate analysis will be performed. Doing a simple linear regression on X and Y, excluding all the x values below x_{min}, we can estimate α. x_{min} instead is chosen graphically.
The results are then compared to a set of points, Zipf distributed and generated with the same α and x_{min}.
(some results in graphics)
comments

2.1.2 MLEs estimator and KS statistics

With a first idea of α and x_{min} we can now move to a more accurate extimation of the two parameters, following the recepie suggested in [Newmann ecc..]. [cit?]

The method used to estimate the real value of α is called *Maximum likelihood estimator* MLE. This method gives the most likely value for α supposing that the data follow exactly a power-law distribution for all $x \geq x_{min}$. [cit]

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} ln \left(\frac{x_i}{x_{min}} \right) \right]^{-1} \tag{5}$$

Where x_i, with $i = 1...n$, is the observed values of x, rankings in our case, and such that $x_i \geq x_{min}$ and $\hat{\alpha}$ is our best estimation of the real value α. Since this method require the value of x_{min}, it is used combined with another equation.

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \tag{6}$$

This is actually the Kolmogorov-Smirnov or KS statistic [47] where the best estimation for x_{min} is then the value of x_{min} that minimizes D. $P(x)$ is the comulative density function, CDF, for the power-law model that best fits the data in the region $x \geq x_{min}$. $P(x)$ depends on the x_{min} and α. $S(x)$ instead, it is the CDF of the observed data with a value greater than x_{min}.

The procedure used is the following:

Given a county with n different words, $\hat{\alpha}_i$ is calculated using (4) for every possibile value of x_{min}. Then D is calculated for every couple $(\hat{\alpha}_i, x_{min_i})$ chosing eventually the couple corresponding to the lowest value of D. The results for $\hat{\alpha}$ and x_{min} are then plotted in a log-log graphic and compared with the previous ones.

2.2 Part 2

An analysis on empirical distribution is then performed, but this time without assuming that the data exactly follow a power-law behavior. The idea here is to set a metric and try to group

together the counties that share similar distributions. Two metrics are used in order to verify similarity between counties.

2.2.1 Metrics

The two metrics used are the **Kolmogorov-Smirnov**, the KS and the **Jensen–Shannon**, JS.

KS
As explained in the last section, KS distance is just a difference between two CDF.

$$D = \max_{x_{max} \geq x \geq x_{min}} |S(x) - P(x)| \tag{7}$$

$S(x)$ and $P(x)$ are two CDF from two different counties. The two distribution may have a different lenght, therefore the longest need to be cut. In fact an x_{max} is defined as the maximum x of the shortest distribution of the two.

JS
Jensen–Shannon metric is a symmetrized version of a more general metric, the Kullback-Leibler divergence [cit] that is defined as:

$$D_{KL}(P||M) = \sum_{i=1}^{n} P(x_i) log_{10}\left(\frac{P(x_i)}{M(x_i)}\right) \tag{8}$$

In order to symmetrize we need to take the square root.

$$JSD(P||S) = \sqrt{\frac{D_{KL}(P||M) + D_{KL}(S||M)}{2}} \tag{9}$$

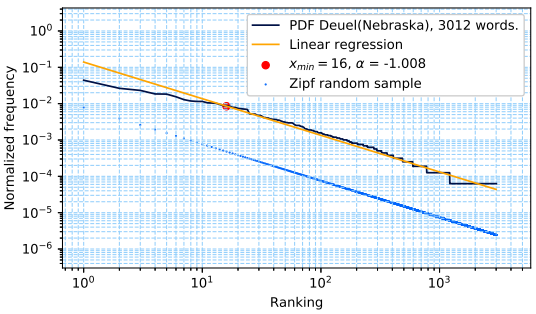
This time $S(x)$ and $P(x)$ are two PDF from two different counties. $M(x)$ instead is defined as: $M(x) = (P+S)/2$. To avoid the problem of P and S having different lenght, some zeros are added to the shortest distribution.
Let $P(x)$ be the shortest distribution, then we modify it adding zeros unitl we have the same length for both distributions $P(x) = \{P(x_1), P(x_2)...P(x_{n_p}), 0, 0...0\}$. In this way P gives zero contribution to D_{KL} in the points whose value is zero, while M is always well defined.

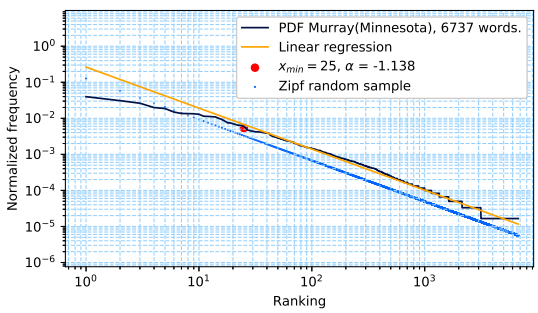
For each of the possible couple of counties KS and JS are calculated and then represented in two upper triangular matrix. A ratio of the two metrics is also calculated in order to check if their values are consistent with each other.

3 Results and discussion

3.1 Part 1

Using the linear regression methods, described before in part 1, we estimated a first rough value for the scaling parameter alpha. Here some representative results for counties with different numbers of unique words.





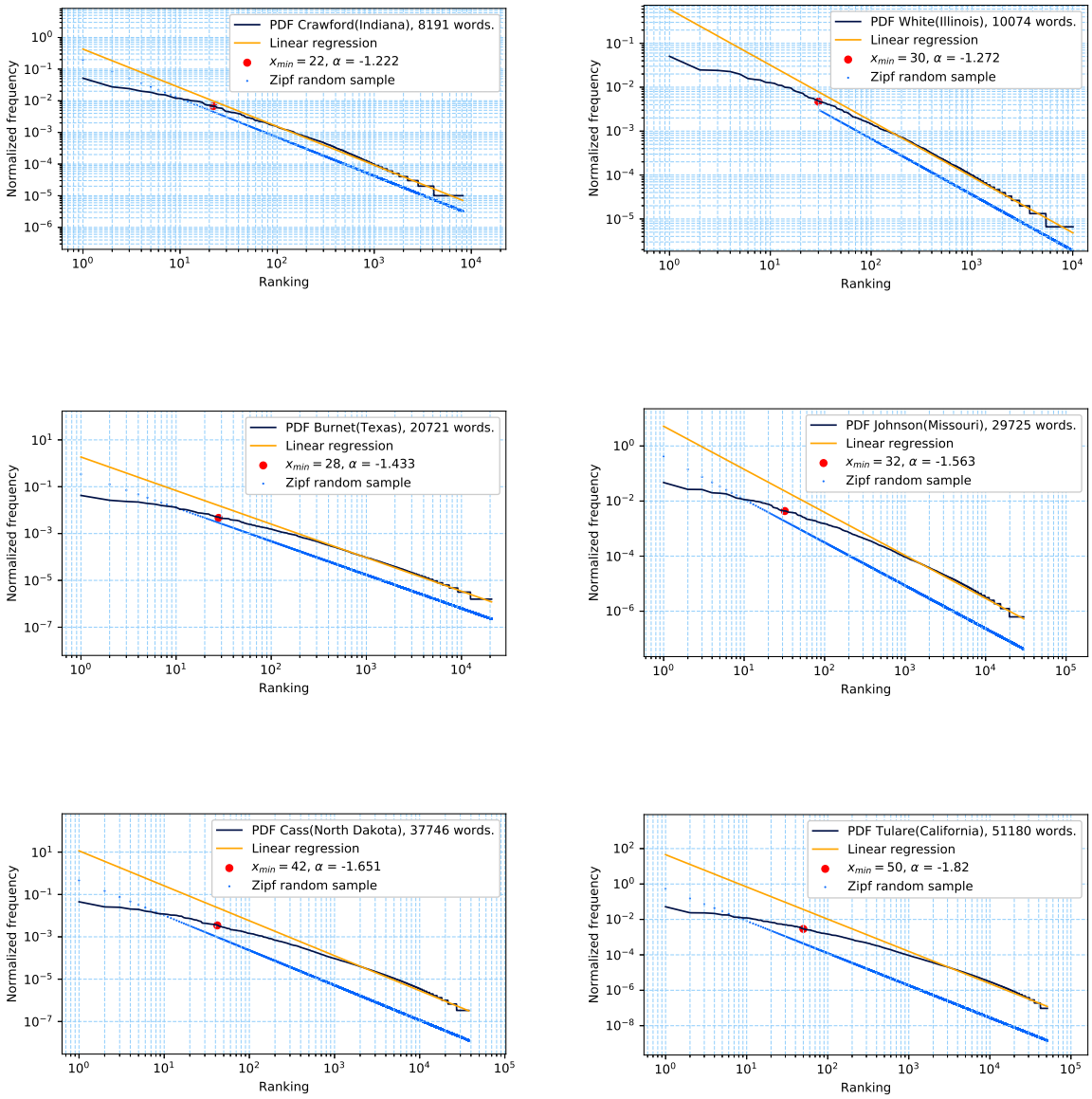


Figure 2: Here probability density functions and linear regressions (calculated from x_{min} to the last x). The estimated alpha is then used to generate a random ZIpf sample.

The graphics show different plots for different counties covering almost all the range of possibile numbers of words. The probability density functions are plotted together with their linear regression (calculated from x_{min} to the last x) and a Zipf random sample with a scaling parameter given by the linear regression.

Looking at the graphics above we can clearly see how the linear regressions get worse when the number of words increases. This is due to the noticeable concavity shown by the PDF. In addition, most of the contribution in the linear regression is due to the last decades which contein most of the point in the graph. This gives a main role to the tail of the distributions.

With some doubts about the aplicability of Zipf to these samples, we move to the next and more sophisticated estimation.

Here it goes a graph that the xmin estimated with (6) always gives the last possible x for a given county. This is probably due to the concavity and the sensitivity of the method that suggest to reject the Zipf hypothesis.

3.2 Part 2

Using the two metrics here we show the resulting distances for each couple of county in two different upper triangular matrices.

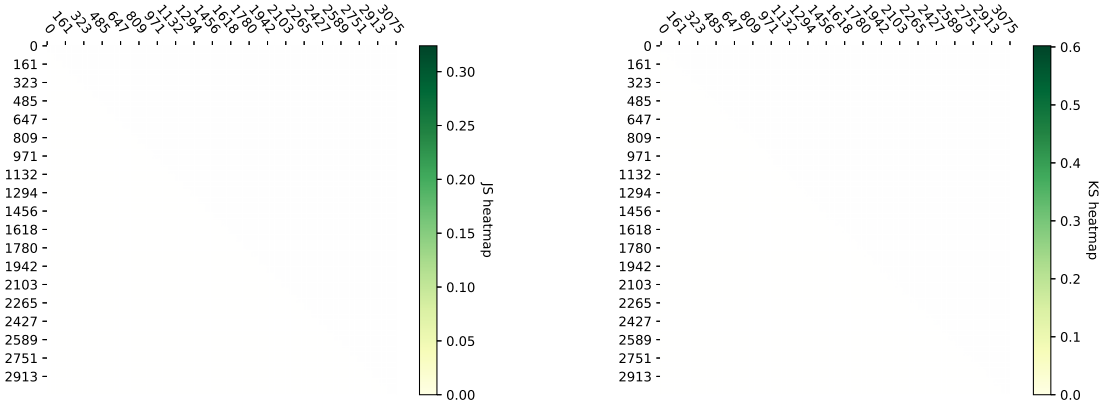


Figure 3: The two distance matrix, one for KS and one for JS. Each little square represent the distance between two different counties.

In these two graphics, each spot quantify how similar two different distributions are. Each county is represented by a number and their order is the same in both of the graphics. At a first glace, the two distance matrix seem to agree with each other in fact they show the same patterns. To be sure it is worth to check the ratio between the values of the two distances.

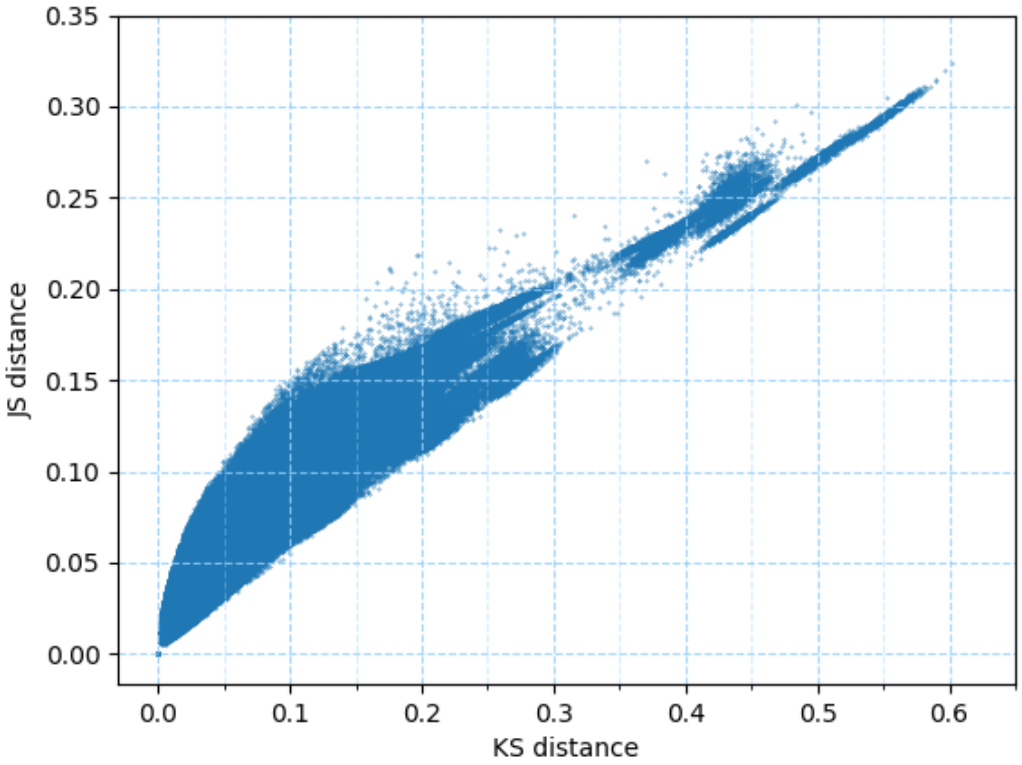


Figure 4: Each dot is a representation of JS vs KS for a given couple of counties.

For every couple of counties we have already calculate KS and JS, here we plot JS vs KS for every given couple. What we obtain has a clear linear trend and it tells us that the two ways of

calculating distances agree with each other.

What we are doing next is to group together counties with a similar distribution. In order to do this we use two different clustering method: Kmeans and the Hierarchical Clustering. Both methods require to give the number of clusters *a priori*. In the following figures, some representative geo-plots varying the number of clusters, for both clustering method.

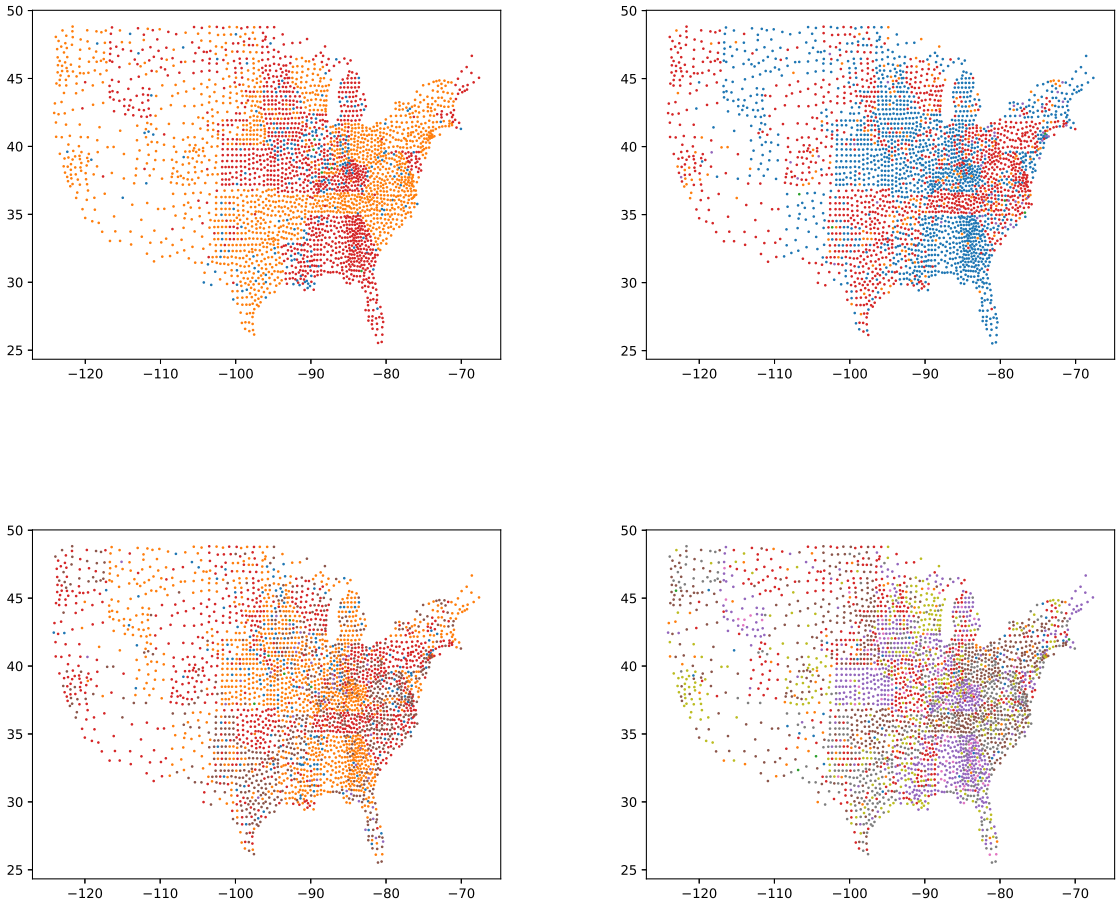
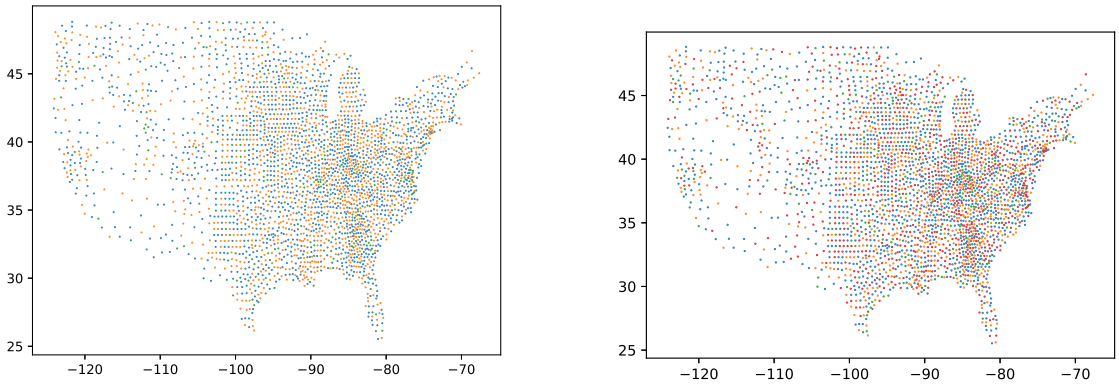


Figure 5:



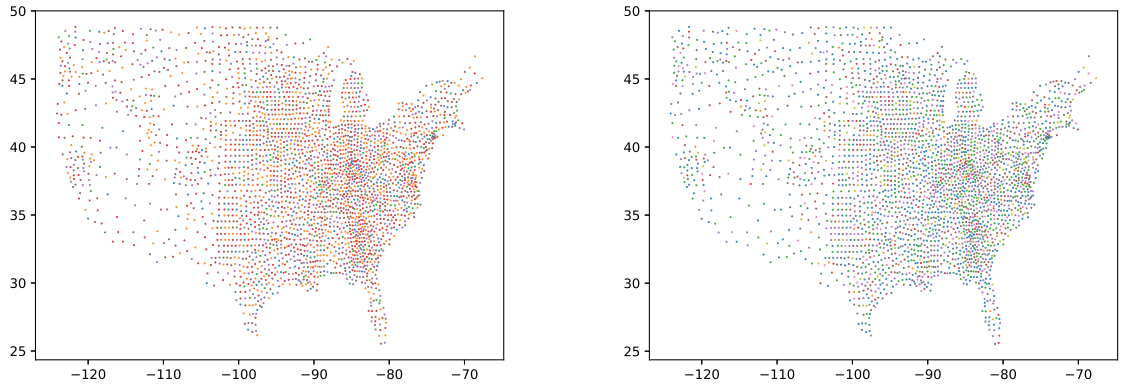


Figure 6:

4 Conclusions

What we have found so far is that Zipf seems not holding for counties. It might be due to the fact that taking sub-sample of a set of elements that follow Zipf's law, doesn't guarantee that the sub-sample follow ZIpf's law [cit]. .

References

[1] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-Law distributions in empirical data.

[2] Matthieu Cristelli, Michael Batty, Luciano Pietronero. There is More than a Power Law in Zipf