The "Power" of Tourism in Portugal

Davide Provenzano

Dipartimento di Scienze Statistiche e Matematiche "Silvio Vianelli"

Facoltà di Economia, Università di Palermo

Viale delle Scienze, 90128, Palermo, ITALY

Abstract.

In this paper we analyze the upper tail of the distribution of tourism supply in Portugal, from 2002 to 2009, using data belonging to the Instituto Nacional de Estatística database.

Tourism supply is defined in terms of lodging capacity of hotel establishments in about 250 tourism destinations.

It is shown that the empirical distribution of tourism supply in Portugal is heavy-tailed and consistent with a power law behavior in its upper tail. Such behavior seems to be stable over the years, provided that for the time horizon covered by our data sets, the scaling parameter is always close to the value of 2.

The power law hypothesis is positively tested by making use of graphical and analytical methods.

Keywords: heavy-tailed distribution; power law behaviour; scaling parameter; tourism supply distribution.

1. Introduction.

Many of the empirical quantities scientists measure, as for instance the heights of human beings or the air pressure, have a typical value around which individual measurements are centered. For these quantities the typical value is therefore representative of most observations.

The sizes of cities, people's personal wealth, the occurrence of words in a English text, and many other quantities, instead, vary over an enormous dynamic range of values, sometimes many orders of magnitude, making their typical or average value not a characterizing measurement. For these quantities the probability of measuring a particular value varies inversely as a power of that value and, therefore, they are said to follow a power law, also known as Zipf's law or Pareto distribution.

Power laws appear widely in physics, biology, engineering, astronomy, economics, finance, computer science, demography, statistics and social sciences.

Mathematically, a quantity x is said to follow a power law if its probability density function (PDF), $f(x)$, is such that

$$f(x) = Cx^{-\alpha} \text{ for } x \geq x_{\min} \quad (1)$$

where C is a normalization constant, α is a constant known as the scaling parameter or tail index, and x_{min} is a lower bound that guarantees the power law behavior[1]. Very often, indeed, the power law only applies for values greater that some x_{min} related to the tail of the distribution. The scaling parameter typically lies in the range $2 < \alpha < 3$, although there are occasional exceptions[2].

The purpose of the study is to verify if the upper tail of the distribution of tourism supply in Portugal follows a power law and to give estimates of the tail index. Data belonging to the Instituto Nacional de Estatística database [3] have been used.

The size of tourism supply is defined in terms of lodging capacity of hotel establishments (hotels, boarding houses, inns, lodging houses, motels, apartment hotels, tourist villages and tourist apartments) in about 250 tourism destinations in Portugal.

The time horizon spans from 2002 to 2009 to study whether the distribution of tourism supply is persistent over time[4].

The power law hypothesis has been tested using both graphical and analytical methods[5].

This study is original in two key regards. In general, no prior research has focused on the empirical distribution of the lodging capacity in a tourism destination within the framework of the extreme value theory and, in particular, this is the only example we know of a study centered on the distribution of tourism supply in Portugal.

---

[1] (1) represents the probability density function of a continuous power law distribution. For mathematical convenience, the continuous form is commonly used to also approximate a discrete power law behaviour, whose formula is not as simple, by rounding the continuous power law value to the nearest integer. For more details see [3].

[2] The value of α is always assumed greater than 1 since distributions with $\alpha \leq 1$ are not normalizable and hence cannot occur in nature.

[3] http://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE.

[4] Actually, 2002 and 2009 are respectively the first and the last data set available for what concerns the lodging capacity of tourism destinations in Portugal.

[5] The statistical techniques used in the present study are discussed in [3].

The structure of the paper is as follows. Section 2 describes the data sets used for the analysis. Section 3 focuses on power law and tail index estimation. Section 4 describes several graphical methods for verifying the power law hypothesis. Section 5 presents the statistical techniques used to support the qualitative analysis. Section 6 concludes.

2. Data description and basic statistics

The unbalanced panel we use in our analysis is part of the database of the Instituto Nacional de Estatística. Data are available for periods since 2002 to 2009 and refer to tourism destinations where a lodging service can be provided to tourists.

Table 1 summarizes the basic descriptive statistics of our panel. For any year it shows the number of tourism destination taken into consideration, the total number of bed-places in hotel establishments (hotels, boarding houses, inns, lodging houses, motels, apartment hotels, tourist villages, and tourist apartments), the average lodging capacity, its standard deviation, and the maximum and minimum lodging capacity.

Year	Tourism destinations	Lodging capacity	Mean	St. Dev	Max	Min
2002	245	239903	979,1959	3538,593	35853	8
2003	245	245778	1003,176	3607,172	37210	12
2004	245	253852	1036,131	3754,773	37906	8
2005	250	263814	1055,256	3850,631	40294	12
2006	251	264037	1051,94	3788,449	39852	13
2007	252	264747	1050,583	3763,972	39712	16
2008	255	273975	1074,412	3827,751	40575	16
2009	254	273804	1077,969	3839,241	40227	12

Table 1. Basic statistics of the Portuguese tourism supply data sets.

3. Power laws and tail index estimation

The most common approach for testing empirical data against a hypothesized power law distribution is to observe that (1) implies the linear form

$$\log f(x) = -\alpha \log x + \log C \tag{2}$$

showing a straight line on a doubly logarithmic plot (the so called Zipf's plot).

Therefore, the presence of a linear relationship on the log-log axes can be seen as a first signal of power law behavior in the distribution, possibly for $x \geq x_{\min}$, with the scaling parameter α given by the absolute slope of the straight line.

There are several techniques to estimate the parameter α and the lower bound $x_{\min}$. In this paper we have used the maximum likelihood estimators (MLEs) for the scaling parameter and the Kolmogorov-Smirnov (KS) statistic for estimating the lower bound.

The MLE for the discrete case is [3]

$$\hat{a} \simeq 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min} - \frac{1}{2}}\right]^{-1} \tag{3}$$

whose average error decays as $O(x_{\min}^{-2})$ and become smaller than 1% of the value of α as $x_{\min} \gtrsim 6$.

Table 2 shows the estimates of $x_{\min}$ and α along with the number of tourism destinations $N_{x \geq \hat{x}_{\min}}$ whose total lodging capacity is greater than $\hat{x}_{\min}$. The last column therefore shows the number of firms for which the power law fitting may be more correct.

Year	$\hat{x}_{\min}$	$\hat{a}$	$N_{x \geq \hat{x}_{\min}}$
2002	456	1.9500	74
2003	489	1.9500	72
2004	529	1.9700	70
2005	596	2.0200	69
2006	585	1.9900	69
2007	668	2.0000	61
2008	590	2.0200	74
2009	620	2.0100	70

Table 2. Power law fits of the Portuguese tourism supply data sets.

All the values of $\hat{a}$ are close to 2.0 and are, therefore, consistent with the above mentioned range[6]. Moreover, the persistence of the value for eight years suggests that the upper tail of the distribution of Portuguese tourism supply is pretty stable over time. This result, together with the increase of $\hat{x}_{\min}$ over time, can suggest a certain translation invariance of the lodging capacity distribution.

Fig. 1 shows the distributions of our data sets with the solid line representing best fits to the data using the MLE. In those and all subsequent plots, we show the complementary cumulative distribution function (CDF), 1-*F(x)*, where *F(x)* is given by

$$F(x) = \left(\frac{x}{x_{\min}} \right)^{-\alpha+1} \tag{4}$$

provided that the visual form of the CDF is more robust than that of the PDF against fluctuations due to finite sample sizes, particularly in the tail of the distribution.

The plots definitely show a clear linearity in the right tail of the CDF with linearity being persistent over time.

4. Graphical testing of the power laws hypothesis

The Zipf's plots shown in Fig 1 give us just a first clue that actual data are effectively drawn by some power law distribution[7]. Yet, since a log-normal distribution or an exponential distribution can look roughly straight on a log-log plot as well, a further graphical analysis of data turns out to be necessary for supporting the hypothesis of a power law behavior. Therefore, we study the tail behavior of our distributions within the framework of the extreme value theory and present the results of two specific tools: the mean excess function (MEF), and the quantile-quantile plot (QQ plot).

The MEF of a sample $X_1, X_2, \ldots, X_n$ is defined as

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)}{\sum_{i=1}^{n} I_{(X_i > u)}} \tag{5}$$

[6] Unfortunately, in the literature there are not other empirical studies of the same type to compare our results to.

[7] On the contrary, if the CDF were non linear in the right tail, the power law hypothesis would be ruled out without the need of any other analysis.

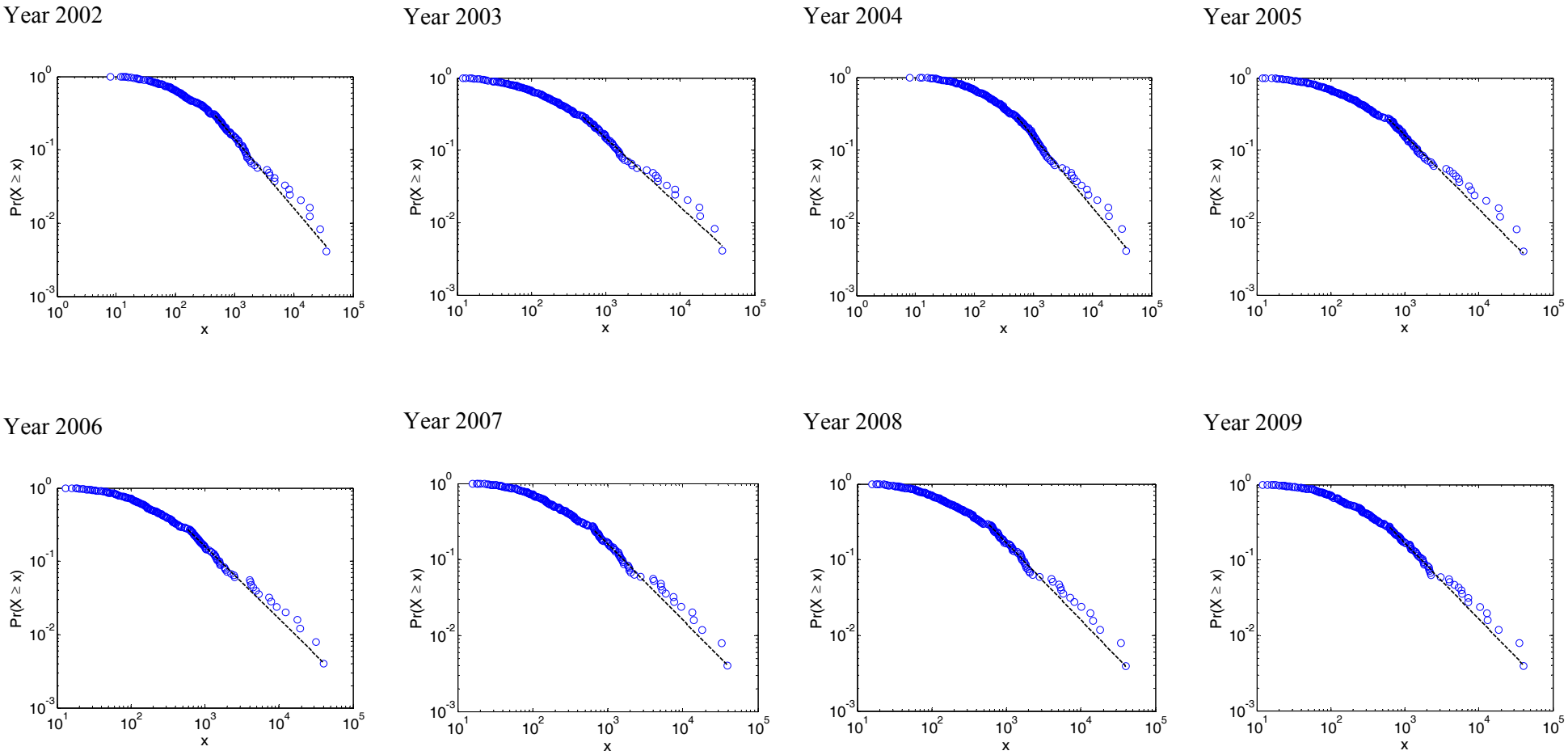


Fig. 1. The cumulative distribution functions and their maximum likelihood power law fits for the Portuguese tourism supply data sets.

that is the sum of the excesses over the threshold u divided by the number of data points exceeding u ($I = 1$ if $X_i > u$ and 0 otherwise).

A MEF with an increasing linear trend is a symptom of the presence of a power law in the right tail of the distribution.

Fig. 2 shows the mean excess function plot (the so-called meplot) for years 2005, 2006, 2007, 2008, 2009[8]. The upward trend shown by the plot confirms a heavy-tailed distribution of data and its persistence over time.

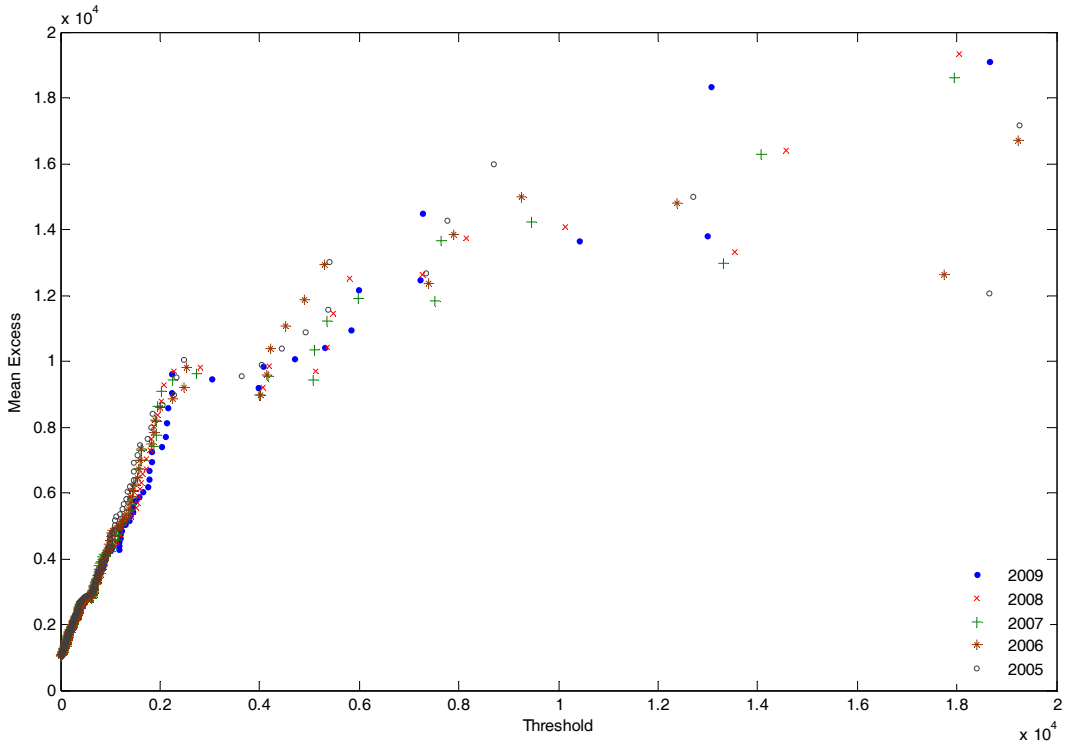


Fig. 2. The MEF of the Portuguese tourism supply for the last five years of our data sets.

In statistics, a QQ-plot (quantile-quantile plot) is a convenient visual tool to examine whether a sample comes from a specific distribution. Specifically, the quantiles of an empirical distribution are plotted against the quantiles of a hypothesized distribution. If the sample comes from the hypothesized distribution or a linear transformation of it, the QQ-plot is linear.

In the extreme value theory, the QQ-plot is typically plotted against the exponential distribution (i.e, a distribution with a medium-sized tail) to measure the fat-tailness of a distribution. If the sample comes from an exponential distribution, the points on the graph would lie along a straight line. A concave presence in the

[8] The other years in the panel are not shown in the plot just to make it more readable.

plot would indicate a fat-tailed distribution, whereas a convex departure is an indication of a short-tailed distribution.

Fig. 3 shows the QQ plots for our data sets. The concave form can be easily recognized in any of the plots and, therefore, these findings are in agreement with previous qualitative results.

5. The goodness-of-fit tests.

Methods described until now are very useful tools for studying the tail behavior of an empirical distribution, but they are not sufficient.

Therefore, the goodness of our power law hypothesis is further tested using a goodness-of-fit test, which allows us to determine a *p*-value that quantifies the extent to which a given distribution represents a good model for actual data.

Such test is based on the measurement of the distance between the distribution of the empirical data and the hypothesized model. This distance is compared with distance measurements for comparable synthetic data sets drawn from the same model, and the *p*-value is defined to be the fraction of the synthetic distances that are larger than the empirical distance. If *p* is large (close to 1), then the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if it is small, the model is not a plausible fit to the data.

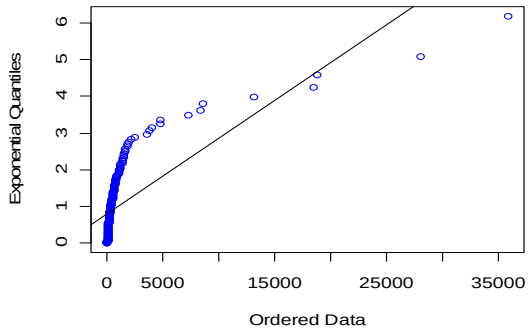
There are a variety of measures for quantifying the distance between two probability distributions, but for nonnormal data the commonest is the Kolmogorov-Smirnov (KS) test.

Table 3 reports the KS statistics for the fit to the power-law model for any year in the panel. *p*-values are statistically significant for each of the years under study, indicating that our data sets are consistent with a power-law distribution.

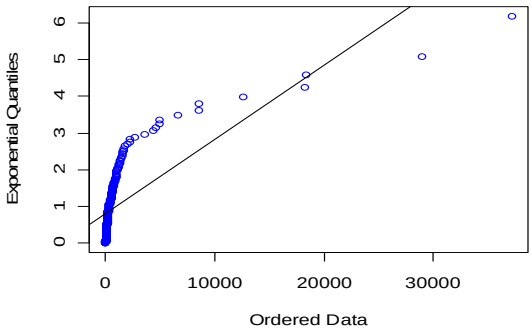
Year	2002	2003	2004	2005	2006	2007	2008	2009
p-value	0.79	0,58	0.62	0.72	0.67	0.54	0.58	0.58

Table 3. *p*-values for the fit to the power-law model for Portuguese tourism supply data.

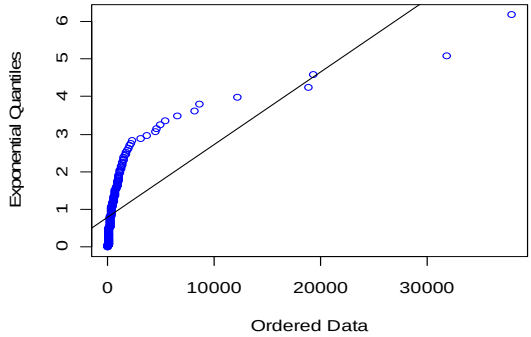
Year 2002



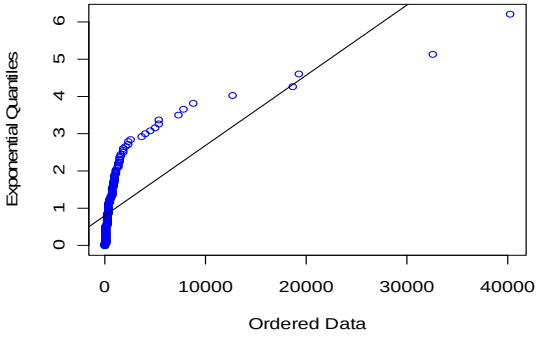
Year 2003



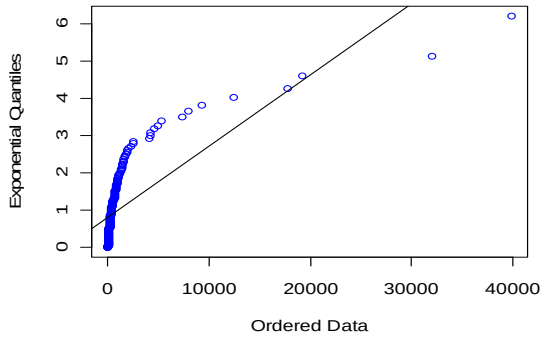
Year 2004



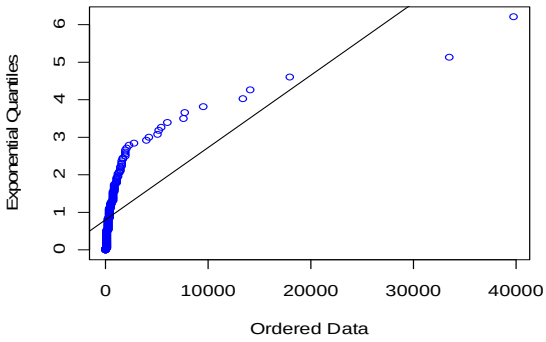
Year 2005



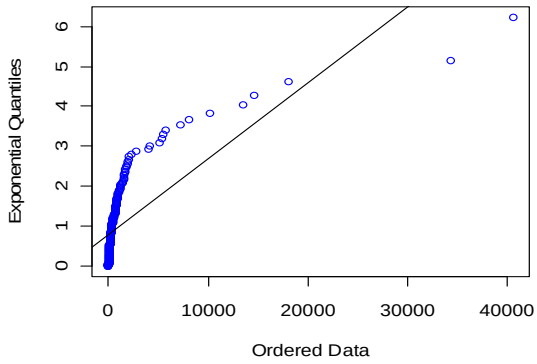
Year 2006



Year 2007



Year 2008



Year 2009

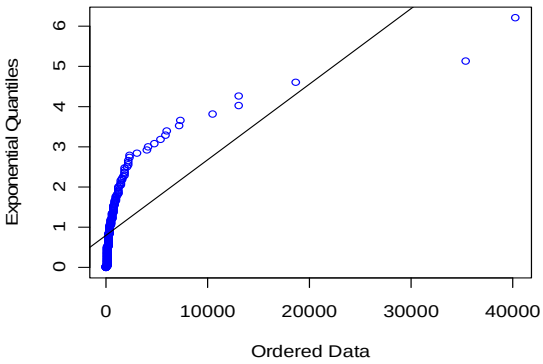


Fig. 3. QQ plots of the Portuguese tourism supply data against standard exponential quantiles.

Conclusions.

In this paper we have analyzed the distribution of tourism supply in Portugal, measured by the lodging capacity of hotel establishments in about 250 tourism destinations from 2002 to 2009, by making use of both qualitative and quantitative methods.

We have firstly followed the common practice of identifying power-law distributions by the approximately straight-line behavior of a histogram on a doubly logarithmic plot.

Yet, since such straight-line behavior is a necessary but by no means sufficient condition for true power-law behavior, we have also used a set of techniques that allow us for validation and qualification of power laws.

Our qualitative analysis shows that there is objective evidence for the heavy-tailed distribution of tourism supply in Portugal, indicating that the number of tourism destinations with a large lodging capacity is noticeably greater than what one would expect with a simple Gaussian distribution.

Moreover, supported by our tests results, we do also claim that the power law hypothesis for the distribution of Portuguese tourism supply turns out to be, statistically speaking, a reasonable description of the data.

These two characteristics have shown to be robust over time.

For planners involved in developing tourism destinations (international tourism development organizations, government agencies, as well as private tourism companies and investors) quantifying the heavy tail can be important to address questions concerning future infrastructure needs. The upper tail is indeed related to the largest lodging capacities and to their frequencies.

References

1. Cirillo, P., Hüsler, J. (2009), 'On the upper tail of Italian firms' size distribution', *Physica A*, Vol 388, pp 1546-1554

2. Clementi, F., Gallegati, M. (2005), 'Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States', http://arxiv.org/pdf/physics/0504217.

3. Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009), 'Power-law distributions in empirical data', *SIAM Review*, Vol 51, No 4, pp 661-703.

4. Gençay, R., Selçuk, F., Ulugülyağci, A. (2001), 'EVIM: A Software Package for Extreme Value Analysis in MATHLAB', *Studies in Nonlinear Dynamics and Econometrics*, Vol 5, No 3, pp 213-239.

5. Newman M. E. J. (2006), 'Power laws, Pareto distributions and Zipf's law', http://arxiv.org/abs/cond-mat/0412004v3.