

Benford's law applied to aerobiological data and its potential as a quality control tool

Silvia Docampo · María del Mar Trigo ·
María Jesús Aira · Baltasar Cabezudo ·
Antonio Flores-Moya

Received: 20 February 2009 / Accepted: 21 July 2009
© Springer Science+Business Media B.V. 2009

Abstract Benford's phenomenological law gives the expected frequencies of the first significant digits of any given series of numbers. According to this law, the frequency of 1 is higher than that of 2; this in turn appears more often 3, and so on. Similarly, Benford's law can also be applied to the first two significant digits (i.e., from 10 to 99), and so on. Here, we show that gross data sets of daily pollen counts from three aerobiological stations (located in European cities with different features regarding vegetation and climatology) fit Benford's law for the first significant digits, but this is not always true for the data transformed by a correction factor used in aerobiological studies. That is to say, the biases introduced by rounding and lower and upper built-in limits in pollen counts are detected by Benford's law analysis. The analysis of the first two significant digits from transformed data is better explained by a Power law than Benford's law. We propose that Benford's law

could be used as a quality control tool for numerical aerobiological data sets.

Keywords Aerobiology · Benford's law · Leading digit phenomenon · Power law · Quality control

1 Introduction

Because aerobiological studies are based on large series of numerical data, there is great concern about the quality of original data. To address these concerns, several slide sampling methods (Comtois et al. 1999; Sterling et al. 1999; Cariñanos et al. 2000), counts from colocated traps (Irdi et al. 2002), different adhesives for pollen sampling (Alcázar and Comtois 1999; Alcázar et al. 2003), and sampling methods (Pyrri and Kapsanaki-Gotsi 2007) have been compared. However, the human factor (pollen and spore identification, and counting procedures), with its nonquantifiable errors, can affect the final quality of aerobiological data.

A characteristic that usually is apparent in large sets of data is that the distribution of leading digits fit to a Power law. Between the sorts of scaling-laws it is emphasized Benford's law (Benford 1938), a phenomenological law also called the first digit law, leading digit phenomenon or law of anomalous

S. Docampo · M. del Mar Trigo · B. Cabezudo ·
A. Flores-Moya (✉)
Departamento de Biología Vegetal (Botánica), Facultad
de Ciencias, Universidad de Málaga, Campus de Teatinos
s/n, 29071 Málaga, Spain
e-mail: floresa@uma.es

M. J. Aira
Departamento de Botánica, Facultad de Farmacia,
Universidad de Santiago de Compostela, Campus sur,
15782 Santiago de Compostela, A Coruña, Spain

numbers. According to this law, the frequency (F) of the first significant digit (i.e., the leftmost nonzero digit, d) are not equidistributed (i.e., 11.1% for all the significant digits) just as might be expected but they can be computed as:

$$F_d = \log_{10}(1 + 1/d), \quad d \in \{1, \dots, 9\}. \quad (1)$$

Consequently, the observed frequency of the figures with 1 as first digit is higher than that whose first digit is 2 (30.1 and 17.6%, respectively) and so on until the figures with 9 as the leading digit with a frequency of 4.6%. This formula can also be used to compute the frequency for the first two significant digits (d_1d_2), and so on.

$$F_{d_1d_2} = \log_{10}(1 + (1/d_1d_2)), \quad d_1d_2 \in \{10, \dots, 99\}. \quad (2)$$

Many series of numbers that follow this law are found in fields such as atomic physics (Pain 2008), biology (Costas et al. 2008), hydrogeology (Nigrini and Miller 2007) and the stock market (Ley 1996). Moreover, several applications of Benford's law have been proposed, including detecting fraud in accounting (Nigrini 1994, 1996, 2005) and political election processes (Mebane 2006). Despite the many examples given by Benford (1938) and previously Newcomb (1881) no mathematically rigorous study was presented until that of Hill (1995a). In general, data sets governed by a multiplicative process obey Benford's law. Moreover, a property of this law is its scale-invariance, i.e., if we multiply all the observations by a nonzero constant the new data set also follows Benford's law (Pinkham 1961). Another aspect to take into account is that random samples from different distributions also conform to

Benford's law (Hill 1995b), even when distributions separately do not (Gent and Walsh 2001).

The aim of this work was to analyse the frequency distribution of the first and first two significant digits of data sets of atmospheric pollen concentrations from different European cities. We checked whether the distributions of aerobiological data obeyed Benford's law, and we discuss the possible implications that this study could have on quality control of data sets.

2 Materials and methods

The total daily pollen count data of three aerobiological stations were used in this analysis. The samples were obtained by using a Hirst type volumetric trap (Hirst 1952) and the counts were expressed as daily mean of the no. of pollen grains m^{-3} , following the methodology proposed by the Spanish aerobiology network (REA; Domínguez-Vilches et al. 1991; Galán et al. 2007). For this purpose, the raw count data are multiplied by a correction factor ranging from 0.486 to 0.54, depending on the field of the microscope used in the counting. Sampling stations were located in different geographic and bioclimatic areas, two located in Spain (Málaga and Ourense) and the third in Germany (Münster). The airborne pollen spectrum as well as the seasonal behavior and other aerobiological parameters are described in detail in Recio et al. (1998), Aira et al. (2005), Trigo et al. (2006) and Rodríguez-Rajo et al. (2007), but main characteristics of the sampling stations and the origin of the data are showed in Table 1.

Table 1 Sampling location and description of the aerobiological data used in the analyses

Sampling station	Data set	Period of sampling	No. of data for FSD	No. of data for FTSD	Range of the series (grains m^{-3})
Málaga (36°43'N, 004°25'W)	Daily mean values	1992–2007	–	4,230	10–4,054
	Daily mean values	1992–1999	2,918	–	1–4,054
	Daily counts	1992–1999	2,918	–	1–7,544
Orense (42°20'N, 007°51'W)	Daily mean values	1993–2007	5,055	3,162	1–1,558
Münster (52°57'N, 007°37'E)	Daily mean values	2002–2004	894	–	1–3,514
	Daily counts	2002–2004	894	–	1–7,231

FSD first significant digit, FTSD first two significant digits

We first studied the fit of the pollen data set to the frequencies described by Benford's law (Eqs. 1, 2). When statistically significant differences between actual and expected data were found, other analyses were carried out comparing the data set to the expected frequencies according to a Power law. In this case, the expected frequencies of the first two digits d_1d_2 were given by the Power law when $m \neq 0$:

$$F_{d_1d_2} = (d_1d_2/10)^{-m} (1 - ((d_1d_2 + 1)/d_1d_2)^{-m}) / (1 - 10^{-m}); \quad d_1d_2 \in \{10, \dots, 99\}. \quad (3)$$

Nigrini and Miller (2007) demonstrated that Benford's law for the first two significant digits (see Eq. 2) is a particular case of the Power law from Eq. 3 in which $m = 0$. Two different methodologies were used to calculate the value of m : by means of an interactive process that minimizes the χ^2 test statistic and, alternatively, following the method proposed by Newman (2005) and Clauset et al. (2007) using:

$$m + 1 = 1 + N \left[\sum_{i=1}^N \ln(x_i/x_{\min}) \right]^{-1} \quad (4)$$

where x_i , with $i = 1, \dots, N$, are the observed values of x , and $x_{\min}$ is the minimum value of x .

The goodness of fit of the data to the values predicted by the different models employed was checked by χ^2 and Kolmogorov–Smirnov tests. Moreover, the mean absolute deviation (MAD) defined by Drake and Nigrini (2000) was also calculated.

3 Results

3.1 Analysis of first significant digit frequency

The lower the first significant digit of pollen counts in the three sampling stations, the higher is its frequency. The frequency of the first significant digit conforms to Benford's law in the case of pollen from Ourense ($\chi^2 = 15.597$, $df = 8$, $\alpha = 0.01$) but not from Málaga and Münster ($\chi^2 = 35.561$, and 22.011 , respectively, $df = 8$, $\alpha < 0.005$ in both cases; Fig. 1). However, if the original counts are used (i.e., without applying any correction factor, instead the transformed values of the no. of pollen grains m^{-3} being

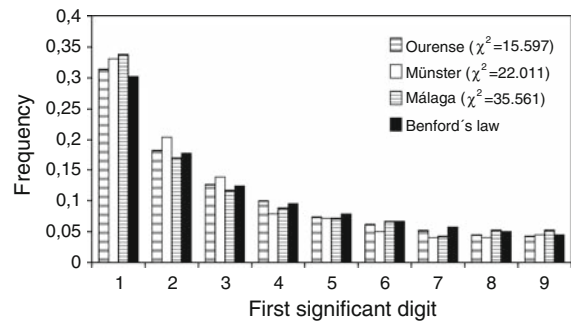


Fig. 1 Observed frequency of the first significant digits of the daily mean pollen grain concentrations recorded in Ourense, Münster and Málaga, and expected frequency according to Benford's law. The values of χ^2 -test ($df = 8$) for the goodness of fit of the data to the values predicted by Benford's law are shown

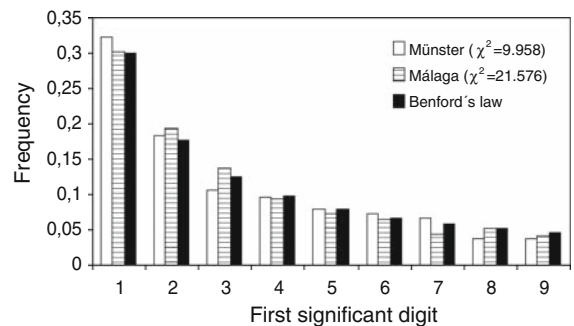


Fig. 2 Observed frequency of the first significant digits of the original daily pollen counts (i.e., raw data to which no correction factor was applied) in Münster and Málaga, and expected frequency of Benford's law. The values of the χ^2 -test ($df = 8$) for the goodness of fit of the data to the values predicted by Benford's law are shown

computed as raw values multiplied by the correction factor), both Málaga and Münster data set patterns obey Benford's law ($\chi^2 = 21.576$ and 9.958 , $df = 8$ in both cases, $\alpha = 0.005$ and 0.1 , respectively; Fig. 2).

3.2 Analysis of first two significant digits frequency

In this case, Málaga and Ourense data sets were used because the number of observations from Münster was too small (see Table 1). The first two significant digits from both sampling stations did not conform to Benford's law (Fig. 3). However, a similar tendency

could be observed at both sampling stations: the frequency of the first digits is higher than that of the following ones.

To better understand this phenomenon, further analyses were carried out. Firstly, the internal structures of the data sets were studied. For this, the logarithms of the pollen concentration and the ranking of frequencies were plotted (Fig. 4), showing similar patterns in both cases. The graphs show a sigmoid line with three different slopes that indicate that the data are described by three successive geometric series (when a data set conforms to Benford's law, it usually presents a nearly geometric pattern). The three slopes could not always be detected because of the inferior limit; since the minimum concentration value included was 1 grain m^{-3} , for which the logarithm is zero, no negative values exist. The horizontal steps show that there are equal concentration values, so lines with a slope of zero can be observed (for details see Nigrini and Miller 2007).

We also constructed a plot of the pollen concentration frequencies and their rankings on a logarithmic scale (cumulative distribution function). A negative slope with a noisy increase on the left side as the frequencies decrease can be seen (Fig. 5). This is a sign that the data distribution follows a Power law (see Newman 2005). Figure 6 presents additional

support; the plot shows the pollen concentrations and their frequencies on a logarithmic scale. The line with a negative slope and the noisy results on the right side are characteristic of data that conform to a Power law (see Newman 2005). This is based on the fact that a great number of measurements of naturally occurring phenomena are distributed logarithmically, and thus the logarithm of these data usually tends to be uniformly distributed.

In addition, the ordered values of the mantissas were analyzed. If the data set fits the frequencies described by Benford's law, then it will fit to a straight line representing a uniform distribution over the [0,1] interval with the equation:

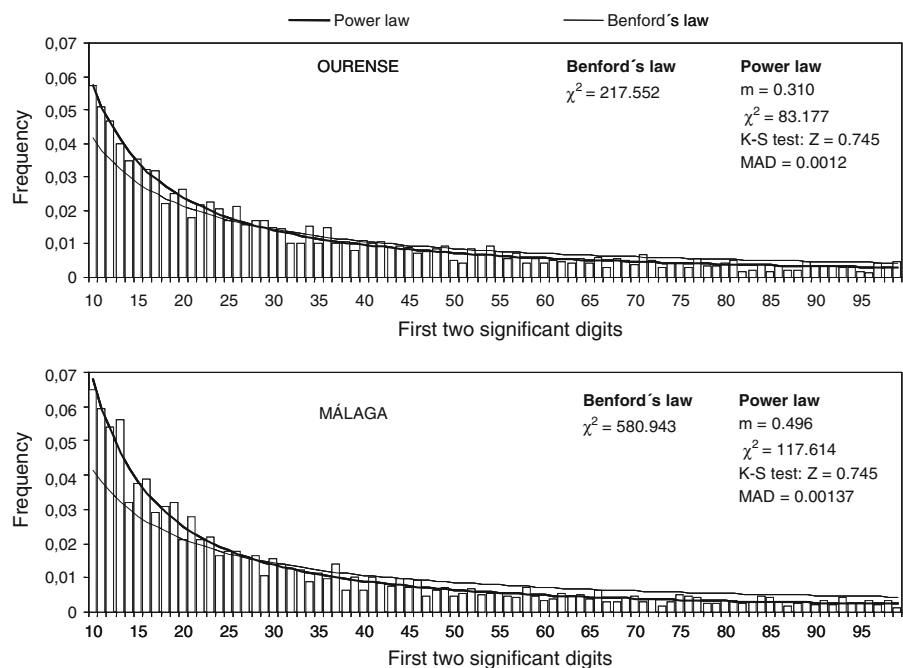
$$Y_i = -(1/N) + (1/N) \text{Rank}, \quad i \in \{1, 2, \dots, N\} \quad (5)$$

where N is the number of data and the Rank is used in decreasing order.

However, in this case we could see that mantissas are not equidistributed, instead fitting a curve described by a cubic equation. The steps shown are due to the high index of repeated numbers that therefore have the same mantissa (Fig. 7).

Finally, the observed data were represented together with the Power law distribution obtained by the methodology described earlier. In both cases (Ourense and Málaga), a fit to a Power law

Fig. 3 The frequency of the first two digits of the daily mean pollen counts in Ourense and Málaga and the expected frequencies according to Benford's law and to a power law, respectively. The MAD values as well as the χ^2 -test ($df = 89$) and the Kolmogorov–Smirnov test for the goodness of fit of the data are included. The exponent of the power law (m) is also shown



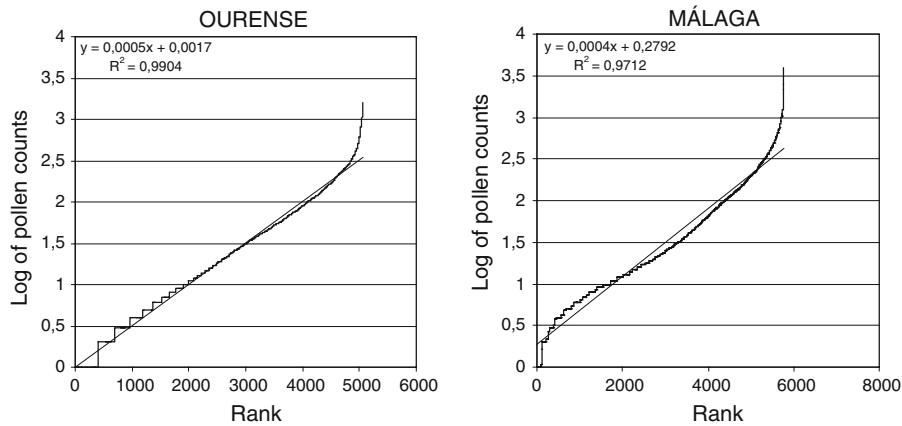


Fig. 4 Ordered values of logarithms of the pollen concentration data sets recorded in Ourense and Málaga. A fitted *regression line* and the equation obtained are also shown

Fig. 5 Ordered values of the pollen concentration frequencies in Ourense and Málaga on a logarithmic scale

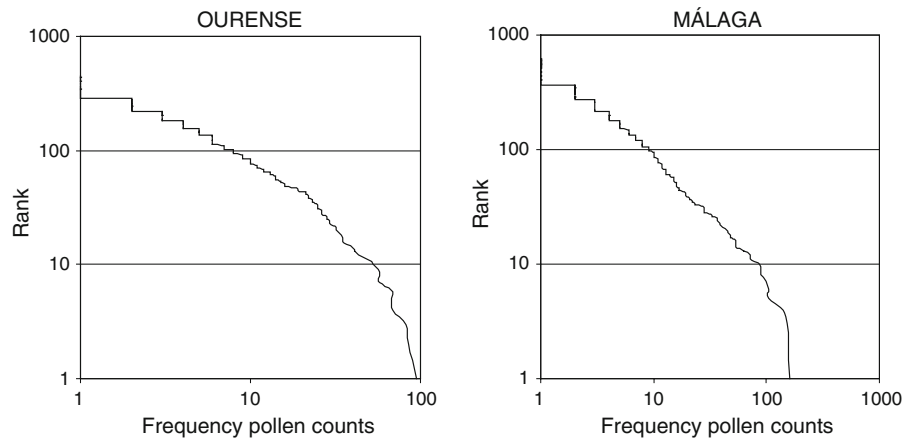
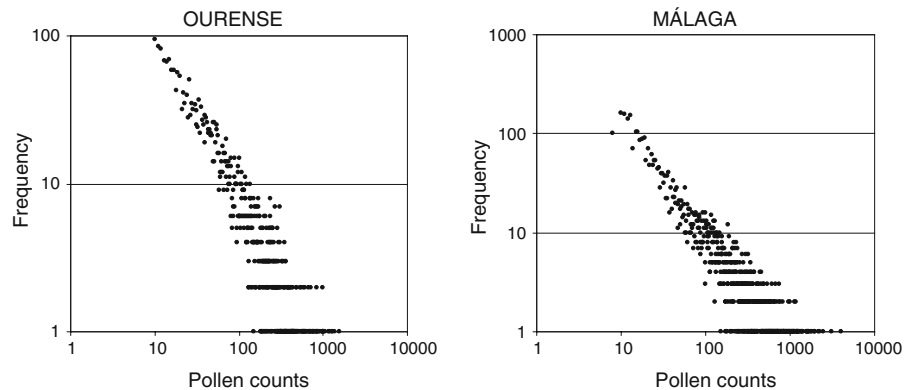


Fig. 6 Daily mean pollen concentrations in Ourense and Málaga and their frequencies on a logarithmic scale



distribution was found (Fig. 3). The χ^2 and Kolmogorov–Smirnov values supported this and the MAD calculated is classified as acceptable

(0.0006–0.0012) and marginally acceptable (0.0012–0.0018) conformity for data from Ourense and Málaga, respectively.

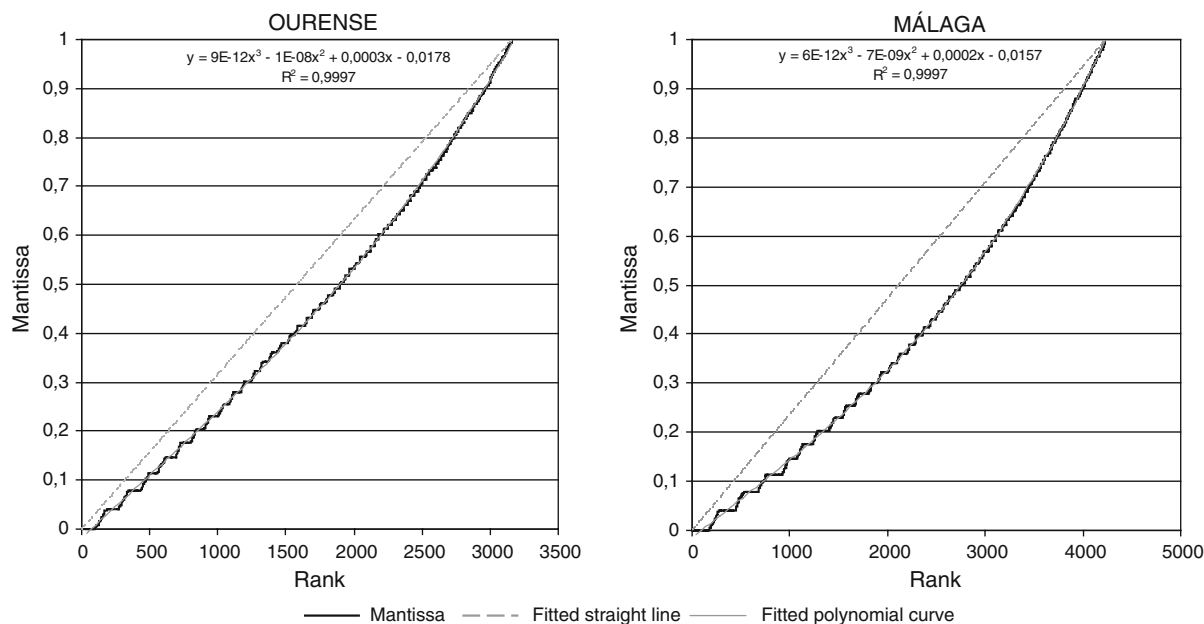


Fig. 7 Ordered values of the mantissas of the pollen concentrations in Ourense and Málaga. A fitted *straight line* following a uniform distribution and the *line* of a cubic equation fitted to the mantissas values are also plotted

4 Discussion

These analyses show that sometimes aerobiological data follow a Benford's law for the first significant digit, and although a tendency for the higher frequencies of the first digits compared to later ones is always apparent, data do not conform to this law in all cases. Whereas the original counts follow a Benford's distribution, the concentrations expressed as no. of pollen grains m^{-3} do not.

Data sets that follow this law are scale-invariant (Pinkham 1961), so if we multiply all the observed values by a nonzero constant, these data sets will continue conforming to Benford's law. However, in two of the three sampling stations studied this did not happen because we used a constant multiplicative factor and the first digit frequencies changed. A rounding phenomenon could be a possible cause of this result. Due to the indivisible nature of pollen grains, decimal digits are not used to express concentrations, so values are rounded up or down. In this way, in Málaga counts 0.54 would be computable as first digit 1 instead 5, 1.62 would change from 1 to 2 and so on. However, another influential factor seems to be the limit of the data set. A requirement for data that follow Benford's law is

the lack of upper or lower built-in limits (Drake and Nigrini 2000). The series of data included in this work have an upper limit (4,054 pollen grains m^{-3}) and a lower limit (1 pollen grain m^{-3} ; see Table 1) because each is a series of integer and positive numbers and they are restricted to an artificially bounded space (1 m^3 of air). In the case of original counts, this data set presents a better fit to Benford's distribution although a built-in limit also exists. This might be caused by the less restricted volume of air (14.4 m^3 for the daily total), which permits higher values.

However, we observed a tendency of the frequencies of the first digits to increase and the frequencies of the later ones to decrease. If we construct a graph with a series of correlative numbers (i.e. $n = 999$), it shows an equidistribution of the frequencies for the first and first two digits. But if we multiply this series by a constant (0.54, for instance), the frequency of the first digits (up to 4 and 54, respectively) is higher than that where no factor was used, sharply decreasing in the consecutive digits (Fig. 8). In this way it is probable that something similar may happen with pollen data since the first two significant digits have a higher frequency than the one predicted by Benford's law and the later ones a lower frequency. This pattern

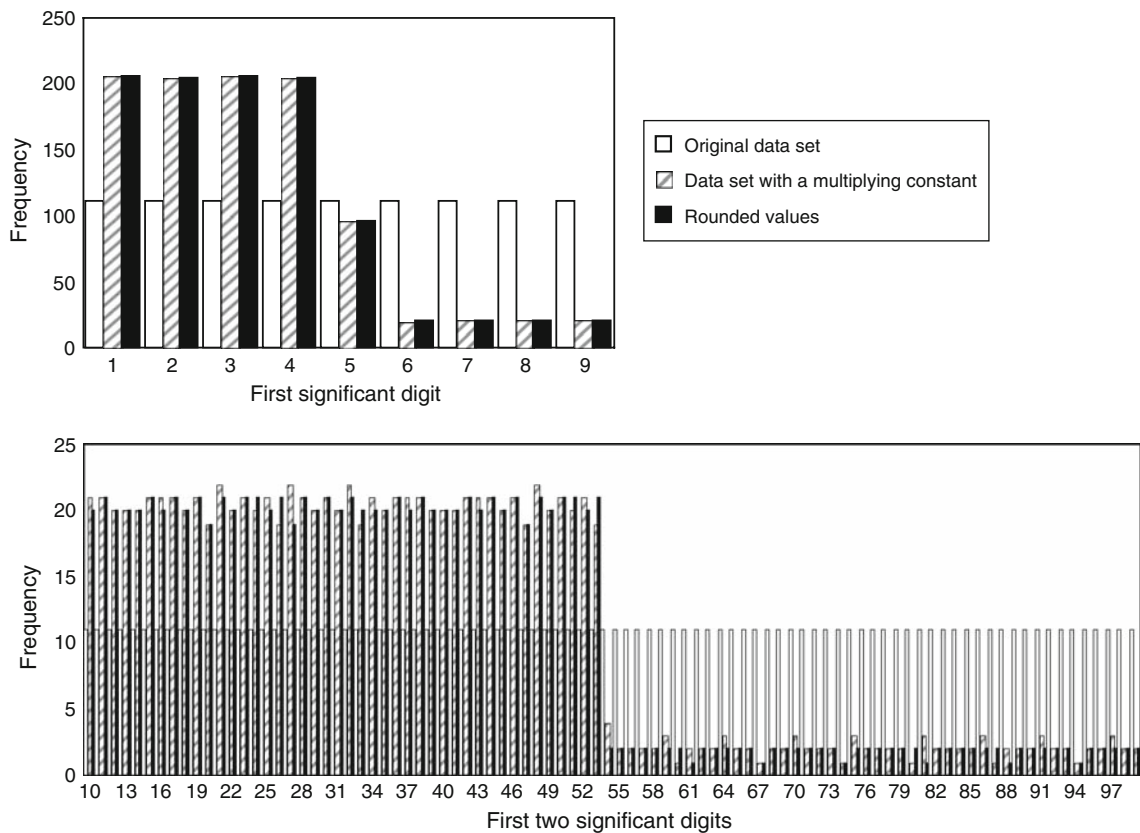


Fig. 8 First digit and first two digit frequencies of a data set of ordered numbers over the [1–999] interval (*left*). The same data set after multiplying the number series by a constant (0.54), and the values after rounding them up (*right*)

is in accordance with a Power law. The exponent value varied depending on the station and the method used to calculate it. In this case we think the value obtained by means of an interactive process that minimizes the χ^2 test statistic is more reliable, owing to the difficulty in establishing $x_{\min}$ in the Newman equation (Newman 2005). The Ourense data distribution model had a lower exponent m than the Málaga data set (0.310 against 0.496), which is a sign of the closer fit to Benford's law, whose exponent m value is 0 (Nigrini and Miller 2007).

Although further study of this phenomenon is necessary, there is clear evidence that frequencies of the first significant digits from aerobiological data obey Benford's law or, failing that, another Power law, being independent of the geographic origin. It has been recently reported that aerobiological data do not conform to a normal distribution, fitting much better the patterns described for other sorts of data distributions, such as lognormal and exponential

distributions (Limpert et al. 2008). It is another indication of the probable behavior of the leading digit frequencies in aerobiological data, since these types of distributions (lognormal and exponential distributions) conform to Benford's law to a high degree (Lolbert 2008).

Moreover, similar analyses to those carried out in our study have also been applied to the study of large data sets corresponding to pollutant concentrations in ambient air. Brown (2005, 2007) has used Benford's law and Zipf's law (Power law) in the screening of analytical data finding advantages in the last one over Benford's law, specially when data sets are small or have few orders of magnitude as it occurs in our case.

Finally, these distributional properties could be used as quality control tools for detecting anomalies in aerobiological data sets, such as rounding and error counts or transcription errors during data processing, the first digit analysis giving general information and the first two digits providing a more detailed analysis

that would otherwise be overlooked (Nigrini and Miller 2007).

Acknowledgments This work was financially supported by the Science and Education Ministry (project CGL2006-12648-C03-03/BOS) and the Andalusian Regional Government (projects P06-RNM 02034 and P05-RNM 00935). The aerobiological study of Ourense was supported by the Regional Government of Galicia (Health and Environmental Departments). Dr Eric C. Henry (Herbarium, Department of Botany and Plant Pathology, Oregon State University) kindly revised the English style and usage. Dr. Marta Recio and Marta Melgar helped to collect data from Málaga and Münster.

References

- Aira, M. J., Jato, V., Iglesias, I., & Rodríguez-Rajo, F. J. (2005). *Calidad del aire. Polen y esporas en la comunidad gallega* (p. 237). Santiago de Compostela: Xunta de Galicia.
- Alcázar, P., & Comtois, P. (1999). A new adhesive for airborne pollen sampling. *Aerobiologia*, 15, 105–108.
- Alcázar, P., Galán, C., Cariñanos, P., & Domínguez-Vilches, E. (2003). A new adhesive for airborne pollen sampling in Spain. *Aerobiologia*, 19, 57–61.
- Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78, 551–572.
- Brown, R. J. C. (2005). Benford's law and the screening of analytical data: The case of pollutant concentrations in ambient air. *The Analyst*, 130, 1280–1285.
- Brown, R. J. C. (2007). The use of Zipf's law in the screening of analytical data: A step beyond Benford. *The Analyst*, 132, 344–349.
- Cariñanos, P., Emberlin, J., Galán, C., & Domínguez-Vilches, E. (2000). Comparison of two pollen counting methods of slides from Hirst type volumetric trap. *Aerobiologia*, 16, 339–346.
- Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2007). Power-law distributions in empirical data. *arXiv:0706.1062v1*. [physics.data-an]: *Physics. Data-analysis, statistics and probability*. <http://arxiv.org/abs/0706.1062v1>. Accessed 1 June 2009.
- Comtois, P., Alcázar, P., & Nerón, D. (1999). Pollen counts statistics and its relevance to precision. *Aerobiologia*, 15, 19–28.
- Costas, E., López-Rodas, V., Toro, F. J., & Flores-Moya, A. (2008). The number of cells in colonies of the cyanobacterium *Microcystis aeruginosa* satisfies Benford's law. *Aquatic Botany*, 89, 341–343.
- Domínguez-Vilches, E., Galán, C., Villamandos, F., & Infante, F. (1991). Manejo y evaluación de los datos obtenidos en los muestreos aerobiológicos. *Monografías Rea/EAN*, 1, 1–18.
- Drake, P. D., & Nigrini, M. J. (2000). Computer assisted analytical procedures using Benford's Law. *Journal of Accounting Education*, 18, 127–146.
- Galán, C., Cariñanos, P., Alcázar, P., & Domínguez-Vilches, E. (2007). *Manual de calidad y gestión de la Red Española de Aerobiología*. Córdoba: Servicio de Publicaciones de la Universidad de Córdoba.
- Gent, I. & Walsh, T. (2001). *Benford's law*. APES-25-2001. Research Reports of the APES Group. United Kingdom. <http://www.dcs.st-and.ac.uk/~apes/apesreports.html>. Accessed 1 June 2009.
- Hill, T. (1995a). Base-invariance implies Benford's Law. *Proceedings of the American Mathematical Society*, 123(3), 887–895.
- Hill, T. (1995b). A statistical derivation of the significant-digit law. *Statistical Science*, 10(4), 354–363.
- Hirst, J. M. (1952). An automatic volumetric spore trap. *The Annals of Applied Biology*, 39(2), 257–265.
- Irdi, G. A., Jones, J. R., & White, C. M. (2002). Pollen and fungal spore sampling and analysis. Statistical evaluations. *Grana*, 41, 44–47.
- Ley, E. (1996). On the peculiar distribution of the U.S. stock indices digits. *The American Statistician*, 50(31), 1–313.
- Limpert, E., Burke, J., Galán, C., Trigo, M. M., West, J. S., & Stahel, W. A. (2008). Data, not only in aerobiology: How normal is the normal distribution? *Aerobiologia*, 24, 121–124.
- Lolbert, T. (2008). On the non-existence of a general Benford's law. *Mathematical Social Sciences*, 55, 103–106.
- Mebane, W. (2006). *Election forensics: The second-digit Benford's law test and recent American presidential election*, Election Fraud Conference, Salt Lake City, Utah, September 29–30, 2006. <http://www-personal.umich.edu/~wmebane/pm06.pdf>.
- Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4, 39–40.
- Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46, 323–351.
- Nigrini, M. J. (1994). Using digital frequencies to detect fraud. *The White Paper*, (8), 3–6.
- Nigrini, M. J. (1996). Digital Analysis and the Reduction of Auditor Litigation Risk. In M. Ettredge (Ed.), *Proceedings of the 1996 Deloitte & Touche/University of Kansas symposium on auditing problems* (pp. 69–81). Lawrence, KS: University of Kansas.
- Nigrini, M. J. (2005). An assessment of the change in the incidence of earnings management around the Enron-Andersen episode. *Review of Accounting and Finance*, 4, 92–110.
- Nigrini, M. J., & Miller, S. J. (2007). Benford's Law applied to hydrology data—Results and relevance to other geophysical data. *Mathematical Geology*, 39, 469–490.
- Pain, J. C. (2008). Benford's law and complex atomic spectra. *Physical Review E*, 77, 012102.
- Pinkham, R. S. (1961). On the distribution of first significant digits. *The Annals of Mathematical Statistics*, 32, 1223–1230.
- Pyrri, I., & Kapsanaki-Gotsi, E. (2007). A comparative study on the airborne fungi in Athens, Greece, By viable and non-viable sampling methods. *Aerobiologia*, 23, 3–15.
- Recio, M., Cabezudo, B., Trigo, M. M., & Toro, F. J. (1998). Pollen calendar of Malaga (Southern Spain), 1991–1995. *Aerobiologia*, 14, 101–107.

- Rodríguez-Rajo, J., Jato, V., Iglesias, I., & Aira, M. J. (2007). *Estudio aerobiológico. Galicia 2006* (p. 142). Santiago de Compostela: Xunta de Galicia.
- Sterling, M., Rogers, C., & Levetin, E. (1999). An evaluation of two methods used for microscopic analysis of airborne fungal spore concentrations from the Burkard Spore Trap. *Aerobiologia*, 15, 9–18.
- Trigo, M. M., Melgar, M., Sinclair, N. & Meinardus-Hager, G. (2006). Two years of aerobiological study of the atmosphere of Münster (northwestern Germany). In: *Abstracts of the 8th International Congress on Aerobiology* (pp. 208). Neuchâtel, Switzerland.