The use of Zipf's law in the screening of analytical data: a step beyond Benford

Richard J. C. Brown*

Received 14th December 2006, Accepted 5th February 2007
First published as an Advance Article on the web 14th February 2007
DOI: 10.1039/b618255k

This study shows for the first time the effectiveness of Zipf's law in screening analytical data sets for outliers, data formatting and data transcription errors, particularly when the data sets are small. In the case of pollutant concentrations in ambient air, the multivariate nature of the measurement, and the relationship between the measured values of these multivariant quantities are the characteristics that allow a Zipf's law approach to data screening to be successful. Furthermore, it has been shown that Zipf's law has advantages over other novel data screening techniques, such as Benford's law, in terms of sensitivity and scope.

Introduction

The need to ensure the robustness of very large data sets produced by analytical measurement processes is increasing. This requires data screening techniques that can identify formatting or transcription errors in large data sets that have undergone multiple data-handling and manipulation procedures. Recently[1] Benford's law has been shown to provide a solution to this requirement for large data sets spanning several orders of magnitude. (Benford's law is the empirical observation that the digits 1 to 9 are not equally as likely to appear as the initial digit in multi-digit numbers resulting from the same phenomenon.[2,3]) However, the law does not apply to data sets that are too uniform or random (*e.g.* the rolling of a nine-faced die, or repeated pH measurements of a pH 6.80 buffer to a ± 0.02 accuracy), too non-random (*e.g.* the distribution of heights among humans), too small, or those subject to artificial limits. Since many of the useful application of data screening procedures in analytical chemistry require the examination of repeat measurements which do not span many orders of magnitude, or limited data sets with relatively few data points, Benford's law may not always be applicable for data screening applications.

An example of the requirement for the screening of complicated analytical data sets is the measured concentration of pollutants in ambient air. It has been shown previously[1] that Benford's law is useful for screening measurements of one pollutant at many measurement sites or of many pollutants at one measurement site provided that the measured quantities span a large range. However, Benford's law cannot effectively screen data sets which are small or do not span large ranges. Moreover, Benford's law does not specifically use the correlations between the multiple components measured at one site, which can provide added value to the screening process. This work proposes for the first time that Zipf's law (often considered to be a generalised form of Benford's law[4]) can be used as a data screening technique for analytical data. Moreover it is proposed that Zipf's law is able to take account of correlations between pollutants at one site, and is also effective for small data sets, which show a small range of values, and for mishandled data which cannot be effectively detected using Benford's law. Zipf's law also has the advantage of being able to be adapted to the particular data set under examination.

Analytical Science Group, National Physical Laboratory, Teddington, Middlesex, UK TW11 0LW. E-mail: richard.brown@npl.co.uk; Fax: +44 (0)20 8614 0423; Tel: +44 (0)20 8943 6409

Zipf's law

George Kingsley Zipf (1902–1950) was an American scholar, who was initially a philologist but came to describe himself as a statistical human ecologist.[5] He was a linguistics professor at Harvard for 20 years where he studied extensively the frequency of the occurrence of words in Chinese language. His studies of word-frequencies are regarded as classics in bibliometrics.[6–8] Only about 500 papers have ever been published relating to Zipf's law,[9] and its close relations:[10] the Lotka[11] and Bradford[12] laws. These have almost entirely concerned the study of language,[13] publishing,[14] population sizes,[15] and economics.[16] Only a few applications have been related to physical and biological sciences,[17–19] and none to possible uses in analytical chemistry and data screening. Generally stated, Zipf's law proposes that, the frequency of occurrence of some event is a function of its rank order, when the rank is determined by the frequency of occurrence, and is governed by the equation

$$f(r) = \frac{A}{r}, \tag{1}$$

where $f(r)$ is the expected relative frequency of occurrence of the r^{th} ranked term, and A is the frequency observed for $r = 1$. For comparative purposes, it is useful to normalise these observations so that $f(1) = 1$; that is to say when the relative frequencies are normalised so that the frequency of the first term is equal to 1, so that:

$$f'(r) = \frac{f(r)}{A} = \frac{1}{r} \tag{2}$$

In other words, the successive relative populations of a series of quantities are roughly proportional to 1, 1/2, 1/3, 1/4,

1/5…1/*n*, and so on. Examples of sets of data that approximately obey this law include:

- frequency of word usage in the English language;
- sizes of earthquakes;
- populations of cites;
- annual incomes of companies;
- frequency of keyword use in internet search engines;
- trophies won by English football clubs.

Zipf's law relationships are best represented by a logarithmic plot of frequency against rank. Indeed, a word frequency analysis of the previous data screening paper on Benford's law[1] (Fig. 1) shows a good agreement with this law, but with a weaker power law dependence as shall be discussed later.

This relationship can be used for data screening in a similar way as has been proposed for Benford's law,[1] but with an improved sensitivity to smaller, but multi-component data sets. In fact a generalised form of Zipf's law can be even more powerful for data screening purposes

$$f'(r) = \frac{1}{r^s}, \tag{3}$$

where s is an exponent characterising the distribution. (In the case of Fig. 1 the best fit for the distribution is found when $s = 0.86$.) For a given exponent the proportion of each rank is defined by eqn (3). Variation of the exponent s allows data to be fitted to the expected distribution of relative probabilities; this characteristic makes Zipf's law more flexible than Benford's law in data screening applications. The exponent may be characterised by the use of a series of valid data sets, in a similar way to a calibration procedure. Variations from Zipf's law may then be easily identified in very small data sets.

The use of Zipf's law for data screening can be clearly illustrated by examining concentrations of a set of pollutants measured at a given location. This type of data set would be difficult for Benford's law to screen since the data set is small and may span only a small range of values. Zipf's law may be especially useful at industrial monitoring locations where the polluting process is expected to emit similar pollutants on a regular basis. In this way Zipf's law, like other chemometrics techniques, is able to spot deviation from expected relationships within a multivariate set of data in a situation where the screening of data from a concentration series of the same analyte is not possible. That is to say, whilst we do not know what to expect in terms of the absolute concentrations of any analyte of a set of analytes at a particular location (because of changing process output and meteorological conditions), we expect the relationship between the concentrations to remain within expected limits (because of the nature of the process being monitored). Similar methods of data screening by chemometrics techniques will be the subject of a later work.[20]

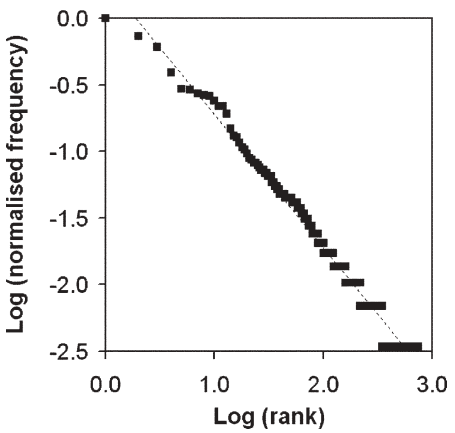

Fig. 1 Analysis of the logarithm of normalised frequency of the use of individual words in ref. 1 against the logarithm of their rank.

Experimental

The UK government has established many air quality measurement networks in order to monitor the ambient concentrations of the most important pollutants, in particular: sulfur dioxide, particulate matter, carbon monoxide, nitrogen dioxide, ozone, benzene and other hydrocarbons, polycyclic aromatic hydrocarbons and heavy metals (currently operated by the National Physical Laboratory). Individual monitoring networks are operated by contractors on behalf of the Department of the Environment, Food and Rural Affairs (Defra) and are responsible for reporting measured pollutant concentrations to Defra and the public Air Quality Archive.[21] It is advantageous, both to Defra, and to contractors of the monitoring networks, to have a simple and robust procedure to screen data sets. The application of Zipf's law to perform this data screening procedure is a novel and exciting possibility.

The UK Heavy Metals Monitoring Network consists of 17 sites situated around the UK at roadside, industrial, rural and urban background locations.[22] Particulate samples are taken at all sites using Partisol 2000 instruments, fitted with PM_{10} [23] size-selective heads, operating at a calibrated flow rate of approximately 1 m^3 h^{-1}. Samples are collected for a period of one week onto 0.8 μm pore size GN Metricel membrane filters. After sampling the filters are digested in acid and analysed for their As, Cd, Cr, Cu, Fe, Mn, Ni, Pb, Pt, V, Zn and Hg content using a PerkinElmer Elan DRC II ICP-MS and the standard procedure detailed in EN 14902.[24] (The analysis of Pt and Hg requires a slightly different digestion procedure.) This produces a maximum of 52 sets of results for 12 metals at each site every year. Usually, fewer than 52 sets of data are produced because of instrument servicing, or breakdown, or obviously contaminated filters that have to be discarded. The data capture rate in 2005 was 93%. For the purposes of this data analysis, results for Pt and Hg at all sites have been discarded since these are commonly below the detection limit and would bias the results.

Data analysis was rapid and straightforward, and used widely available PC software such as Microsoft Excel. The weekly concentration data (expressed as a volume concentration, ng m^{-3}, in ambient air) acquired by the UK Heavy Metals Monitoring Network in 2005 has been used to demonstrate the applicability of Zipf's law. The average yearly concentration data have been ranked at each site, and the results of this analysis are shown in Table 1. The 2005 annual average concentrations across the entire 17 sites are shown in Table 2.

Table 1 The annual average concentration ranking of each element at each monitoring site for 2005

	Rank									
Site	1	2	3	4	5	6	7	8	9	10
Avonmouth 1	Fe	Zn	Pb	Mn	Cu	Ni	V	Cr	Cd	As
Avonmouth 2	Fe	Zn	Pb	Mn	Cu	Ni	V	Cr	Cd	As
Cardiff	Fe	Zn	Pb	Cu	Mn	V	Ni	Cr	As	Cd
Eskdalemuir	Fe	Zn	Pb	Ni	Cu	Mn	V	Cr	As	Cd
Glasgow	Fe	Pb	Zn	Cu	Ni	Mn	Cr	V	As	Cd
Leeds	Fe	Zn	Pb	Cu	Mn	Ni	V	Cr	As	Cd
London 1	Fe	Cu	Zn	Pb	Ni	Mn	V	Cr	As	Cd
London 2	Fe	Cu	Zn	Pb	Mn	Ni	V	Cr	As	Cd
London 3	Fe	Zn	Cu	Pb	Mn	V	Ni	Cr	As	Cd
Manchester	Fe	Cu	Zn	Pb	Ni	Mn	Cr	V	As	Cd
Motherwell	Fe	Zn	Cu	Pb	Mn	Ni	Cr	V	As	Cd
Newcastle	Fe	Zn	Cu	Pb	Mn	V	Ni	Cr	As	Cd
Runcorn	Fe	Zn	Pb	Cu	V	Mn	Ni	Cr	As	Cd
Sheffield	Fe	Ni	Zn	Pb	Mn	Cr	Cu	V	As	Cd
Swansea	Fe	Ni	Zn	Pb	V	Cu	Mn	Cr	As	Cd
Walsall 1	Zn	Fe	Pb	Cu	Mn	Ni	Cr	V	Cd	As
Walsall 2	Fe	Zn	Pb	Cu	Mn	V	Ni	Cr	As	Cd

Table 2 The 2005 annual average concentration, c, for each element across the entire 17 sites of the Network

Analyte	c/ng m^{-3}
Fe	386
Zn	77.8
Cu	19.6
Pb	18.8
Mn	7.73
Ni	4.75
V	4.21
Cr	3.94
As	1.03
Cd	0.59

Results and discussion

We can define a term Δ_{zl} as the normalised deviation from Zipf's law thus:

$$\Delta_{zl} = \frac{1}{N}\sum_{r=1}^{r=N}\left|\frac{f'(r) - f'_{obs}(r)}{f'(r)}\right|, \quad (4)$$

where N is the number of terms in the ranking list (here ten metals at each site) and where $f'_{obs}(r)$ is the observed relative frequency of occurrence of the r^{th} ranked term. The parameter Δ_{zl} represents the average normalised deviation of each data point from the expected optimised Zipf's law relationship. In order to define a quantity-expressing deviation from Zipf's law, a least modulus approach is preferable to fitting the data to a power law using a least squares approach since the former is less sensitive to outliers that usually occur at higher ranks.

For the air quality monitoring site at Avonmouth 2 a Zipf's law analysis was performed on the data acquired during 2005. The average concentration of each of these metals, measured weekly, was compared with the generalised Zipf's law in eqn (3). The exponent s was varied to optimise the agreement and to minimise Δ_{zl}. The optimisation curve for the data is shown in Fig. 2. The analysis in Fig. 2 shows that a value for s of 2.32 minimises the Δ_{zl} value. Fig. 3 shows the best-fit Zipf's law relationship for the annual average data at Avonmouth 2, as optimised in Fig. 2.

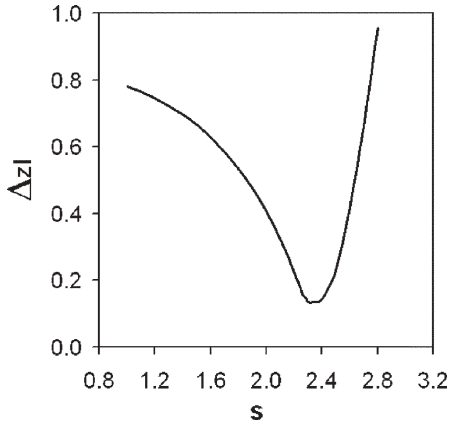

Fig. 2 Dependence of Δ_{zl} on s for annual average data at Avonmouth 2 in 2005.

As can be seen, the agreement of the experimental data with the optimised relationship is very good. Indeed, in all cases the experimental data at all sites agree well with the optimised Zipf's law. Clearly, good agreement of the data sets with Zipf's law is essential if examining deviations from the optimised Zipf's law is then to be used as a method to screen data. Four further optimised relationships for other monitoring sites are shown in Fig. 4. The appearance of an apparent higher power law at higher ranks, in particular for the Eskdalemuir plot in Fig. 4, seems to be a common feature of Zipf's law studies, and has been frequently observed elsewhere.[8,13,18]

The optimised exponent s has been calculated for a Zipf's law distribution at all sites, along with the average normalised deviation from Zipf's law that this yields. The results of this analysis are shown in Fig. 5. Fig. 5 shows that a range of optimised exponents is observed between about 1.6 and 2.8. In general, lower exponents are seen at industrial monitoring locations (except Eskdalemuir) and higher exponents are seen

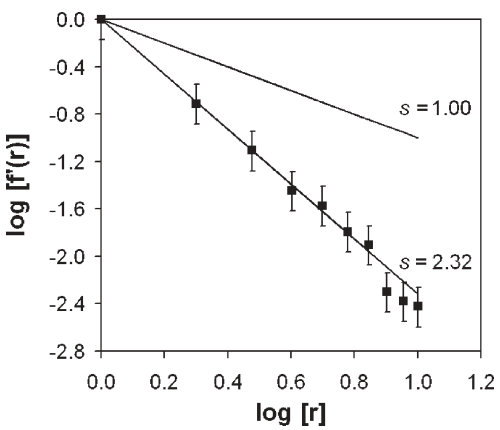

Fig. 3 Relationship between relative concentration and rank for the annual average metals concentrations at Avonmouth 2 (■), with the corresponding relationship describing the minimised average normalised deviation from Zipf's law when s = 2.32. The s = 1 case is included for comparison. The error bars represent the expanded measurement uncertainty expressed at the 95% confidence interval.

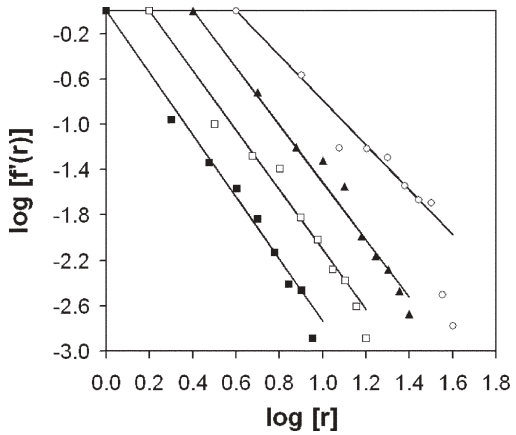

Fig. 4 Relationship between relative concentration and rank for the annual average metals concentrations at Cardiff (■), London 3 (□), Walsall 2 (▲), and Eskdalemuir (○), with the corresponding relationships describing the minimised average normalised deviation from Zipf's law (black lines) when s = 2.74, 2.63, 2.53 and 1.97 respectively. The data for London 3 (□), Walsall 2 (▲), and Eskdalemuir (○) have been offset by +0.2, +0.4 and +0.6 units on the x-axis, respectively, for clarity.

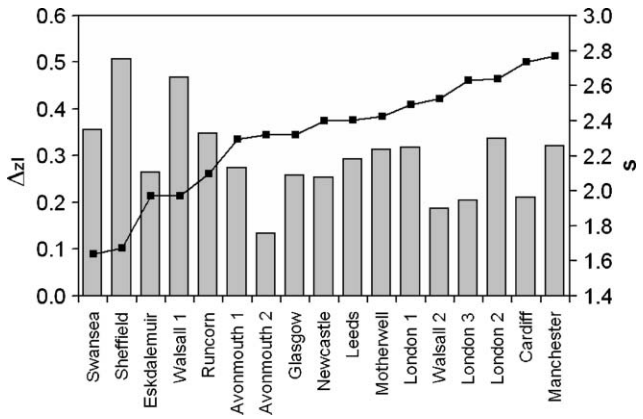

Fig. 5 Calculated values of the optimised exponent s at each monitoring site (■) and the average normalised deviation from Zipf's law, Δ_{zl}, which this value yields (grey bars).

at urban and roadside sites (except Walsall 2). This may be because several metals show similarly high concentrations at industrial sites, whereas at urban and roadside sites, concentrations are dominated by one or two metals. There is a weak negative correlation between s and Δ_{zl}. All these data show a good fit to an optimised Zipf's law distribution. Furthermore, since these relationships are optimised to the annual average data at each site, calculation of the average normalised deviation from Zipf's law of individual weekly data is a highly effective method of outlier detection. Such outliers may occur as a result of: an abnormally high blank filter value, contamination at any stage of the measurement procedure, or instrument malfunction or data corruption during data entry or data transmission. This procedure has been performed for the individual weekly data from Avonmouth 2 and Cardiff. These results are displayed in Fig. 6.

From the plots in Fig. 6 individual filters that might be considered outliers are easy to detect. Fig. 6 has used one standard deviation from the mean to detect outliers, but it is clear that once the analysis has reached this stage, any outlier statistics may be used to formulate criteria for rejection or acceptance of data. The sensitivity of this technique should also be noted. Rather like a multivariate principal component analysis the average deviation from Zipf's law relies on the relative values of all the metals and not just the variations in one value. In this way, 'real' high values caused by unusual metrological conditions or an incident at the industrial location being monitored would not be rejected as outliers. The corollary to this is that the variations in any individual metal concentration may be as large as those seen in Fig. 6 but are not necessarily sensitive to outliers. To detect outliers using variations in individual metal concentrations a plot for each metal would be required followed by a comparison of all the plots. This is illustrated clearly in Fig. 6(b) where the Fe concentrations show less obvious outliers, and any suspiciously high or low values which are seen do not

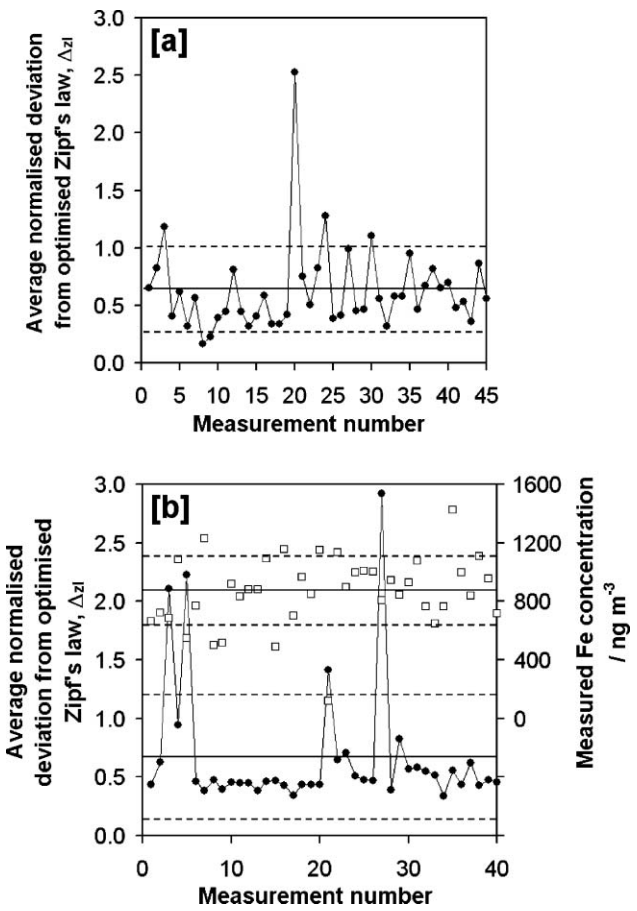

Fig. 6 Average normalised deviations from the optimised Zipf's law relationship for weekly measurements (●) in chronological order at (a) Avonmouth 2 and (b) Cardiff. The solid black lines indicate the average value of the deviations, and of the average Fe concentration in Fig. 6(b), whilst the dashed lines indicate one standard deviation above and below these averages. Fig. 6(b) also includes the weekly measured Fe concentrations (□) for comparison.

necessarily correspond to the outliers determined by the Zipf's law test. If we review the individual data from the possible outliers highlighted in all cases it can be seen that there is indeed further investigation of the data required (for the reason given in parentheses) – Fig. 6(a): measurement numbers 3 (high Cr), 20 (high Ni, possibly low Fe), 24 (possibly high Cd), 30 (high Ni); and Fig. 6(b): measurement numbers 3 (high Ni, possibly high As), 5 (high Ni and As), 22 (high V) and 27 (high Ni). The measurements in Fig. 6 are numbered from left to right starting with measurement number 1 (and also happen to be in chronological order, although this does not affect the data analysis). These data may then be corrected, discarded (if an error is found but it is not correctable), or kept, as appropriate. This procedure is also useful in identifying potentially high filter blanks. It has previously been observed[22] that the filters generally used for ambient air sampling have relatively high blank levels of metals. Few low metal content filters exist for ambient sampling, and those that do are not cost effective for use on a large scale such as in a national network. The problem is not so much that the metals levels in these types of filter are high but that they can vary unpredictably, in a non-Gaussian fashion. Analysis of large numbers of filters has shown that in offending filters only one, or occasionally two, metals are abnormally high at any one time. Unfortunately, it is not possible to analyse filters before they are dispatched for sampling, and once they are returned post-sampling it is not possible to deconvolute the measured metals concentration from the filter blank concentration. Because the high filter blank values are only expected in one, or two, elements on a contaminated filter, in the same way that analytical problems may be detected using Zipf's law as described above, this technique is also suitable for detecting abnormally high filter blank levels, and dealing with them appropriately (either discarding results, or increasing uncertainties, where appropriate).

Having shown that these types of data set may be expected to follow Zipf's law closely, it is instructive to examine the potential effect on these data sets of errors during data processing and manipulation – 'data mishandling'. As a working model, it is assumed that data mishandling results in two types of error: (1) the omission of the initial digit from a certain percentage of the 'real' data set; or, (2) incorrect placing of a decimal point in the data. To study these effects we deliberately introduce both types at random into the selected data set under examination, which was the weekly metals concentrations at the Walsall 2 monitoring site during 2005. For the percentage of data where the initial digit is omitted during data manipulation, the second digit will then become the initial digit (unless the second digit is a zero, in which case the third digit becomes the initial digit, and so on). For the percentage of data where the decimal point was incorrectly placed, it was chosen at random whether or not the mishandling increased or decreased the datum by a factor of ten. Fig. 7 shows the effect of omitting the initial digit from a certain percentage of the test data, and the effect this has, when compared with the situation where no data are mishandled, on the data's: average, standard deviation, average normalised deviation from Zipf's law, and sum of the normalised

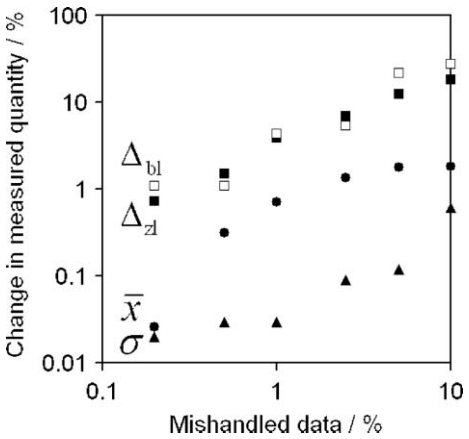

Fig. 7 Percentage change in: the sum of the normalised deviations from Benford's law, Δ_{bl} (□); the average normalised deviation from Zipf's law, Δ_{zl} (■); the arithmetic mean of the data set, $\bar{x}$ (●); and the standard deviation of the data set, σ (▲); as a function of the percentage of simulated data mishandling, involving initial digit omission for weekly metals concentration data collected during 2005 at the monitoring site Walsall 2.

deviations from Benford's law. The sum of the normalised deviations from Benford's law, Δ_{bl}, is given by[1]

$$\Delta_{bl} = \sum_{d=1}^{d=9} \left| \frac{P(d_1) - P_{obs}(d_1)}{P(d_1)} \right|, \quad (5)$$

where $P(d_1)$ is the expected normalised probability of the initial digit d from Benford's law and $P_{obs}(d_1)$ is the normalised observed probability of initial digit d in the experimental data set. $P(d_1)$ is given by:

$$P(d_1) = \log\left[1 + \left(\frac{1}{d_1}\right)\right] \quad (6)$$

For this type of data mishandling it has previously been shown[1] that Δ_{bl} is a much more sensitive quantity in terms of detecting data mishandling than either the average or the standard deviation of the data, and this is proved again in Fig. 7. Moreover, it is shown for the first time in Fig. 7 that the quantity Δ_{zl} is also sensitive to this type of data mishandling and compares excellently with the performance of Benford's law for the data presented here.

Another common form of data mishandling error is incorrect positioning of the decimal point. Although this can happen during data transmission processes, it is most likely to occur because of data formatting, data transcription and data entry errors. This type of error is undetectable by a Benford's law analysis approach since it does not result in any change in the initial digit of any datum. The sensitivity of the average, standard deviation and average normalised deviation from Zipf's law of the weekly metals concentrations at Walsall 2 during 2005 to this type of data mishandling is shown in Fig. 8.

Fig. 8 clearly shows that Δ_{zl} is extremely sensitive to this type of data handling error that would not be detected at all by a Benford's law analysis, *i.e* Δ_{bl} for this graph would remain zero, regardless of the percentage of mishandled data. It can also be seen that Δ_{zl} is vastly more sensitive than the average

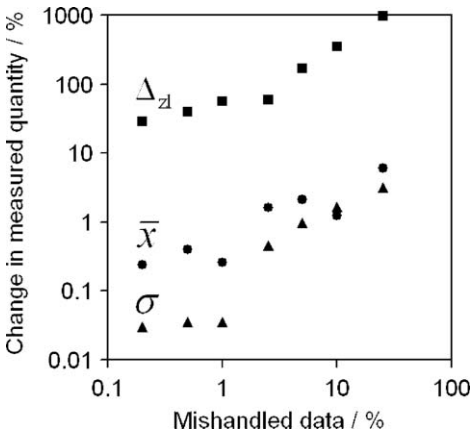

Fig. 8 Percentage change in: the average normalised deviation from Zipf's law, Δ_{zl} (■); the arithmetic mean of the data set, $\bar{x}$ (●); and the standard deviation of the data set, σ (▲); as a function of the percentage of simulated data mishandling, involving incorrect placing of the decimal point, for weekly metals concentration data collected during 2005 at monitoring site Walsall 2.

and standard deviation of the data. Moreover, a Zipf's law analysis has the advantage over a Benford's law analysis in that it can be used successfully on very small data sets. It has the additional strength that it uses the multivariate proprieties of data sets to perform the analysis, in a similar way to a principal component analysis. A Zipf's law analysis has the additional advantage over principal component analysis in that it uses the data from all variables to identify outliers. Usually, principal component analysis only uses the two axes of maximum variation to describe data and can therefore miss potential outliers that a Zipf's law analysis would detect.

Conclusions

Zipf's law has been reported previously as a method for analysing the frequency of word usage, the population of cities and the income of companies. This paper has assessed, for the first time, the potential of Zipf's law to screen analytical data sets; in this case, ambient air pollutant concentrations.

Data sets consisting of weekly ambient metals concentrations at 17 locations around the UK during 2005 were analysed for their adherence to Zipf's law. It was found that the data fitted a generalised form of Zipf's law, in which an exponent is used, very well. The value of the exponent was optimised until the deviation from Zipf's law was minimised. In general, optimisation required lower exponents at industrial monitoring locations and higher exponents at urban and roadside sites. The use of a Zipf's law analysis to identify filters with potentially high blank levels was also discussed. It was also shown that, much like Benford's law, Zipf's law is very sensitive to data mishandling events. However, unlike Benford's law, it has been shown that Zipf's law can be used to detect mishandling in very small data sets, such as the individual weekly data examined, and can deal with data mishandling errors that involve the incorrect positioning of the decimal point. Clearly the better the initial agreement of the data set with Zipf's law, the more sensitive the technique will be to outliers and data mishandling events.

This paper has shown for the first time that Zipf's law may be added to Benford's law as a promising new tool for screening, and checking the authenticity of, sets of analytical data. It has been shown that Zipf's law can be used not only as a general, and top-level, data screening technique for analytical managers but also as a detailed screening tool for small data sets by the analysts themselves. The technique is particularly well suited to any data sets that have multiple components in a similar, but simpler, fashion to principal component analysis.

Acknowledgements

The UK Department of the Environment, Food and Rural Affairs' funding of NPL's management and operation of the UK Heavy Metals Monitoring Network is acknowledged.

References and notes

1 R. J. C. Brown, *Analyst*, 2005, **130**, 1280.
2 R. Matthews, *New Scientist*, 1999, issue no. 2194, p. 27.
3 F. Benford, *Proc. Am. Phil. Soc.*, 1938, **78**, 551.
4 L. Pietronero, E. Tosatti, V. Tosatti and A. Vespignani, *Phys. A (Amsterdam, Neth.)*, 2001, **293**, 297.
5 R. Rousseau, *Glottometrics*, 2002, **3**, 11.
6 G. K. Zipf, *Harvard Studies in Classical Philology*, 1929, **15**, 1.
7 G. K. Zipf, *Psycho-biology of Languages*, Houghton-Miller, Boston, 1935.
8 G. K. Zipf, *Human Behaviour and the Principle of Least Effort*, Addison-Wesley Press, Cambridge, 1949.
9 http://www.nslij-genetics.org/wli/zipf (accessed November 2006).
10 Y.-S. Chen and F. F. Leimkuhler, *J. Am. Soc. Inf. Sci.*, 1986, **37**, 307.
11 A. J. Lotka, *J. Wash. Acad. Sci.*, 1926, **16**, 317.
12 S. C. Bradford, *Engineering*, 1934, **137**, 85.
13 R. Ferrer and I. Cancho, *Eur. Phys. J. B*, 2005, **44**, 249.
14 C. Wang and Z. Wang, *Scientometrics*, 1998, **42**, 89.
15 L. Gana, D. Li and S. Song, *Econ. Lett.*, 2006, **92**, 256.
16 K. Okuyamaa, M. Takayasu and H. Takayasu, *Phys. A (Amsterdam, Neth.)*, 1999, **269**, 125.
17 W. Li and Y. Yang, *J. Theor. Biol.*, 2002, **219**, 539.
18 T. Lu, C. M. Costello, P. J. P. Croucher, R. Hasler, G. Deuschl and S. Schreiber, *BMC Bioinformatics*, 2005, **6**, 37.
19 N. E. Israeloff, M. Kagalenko and K. Chan, *Phys. Rev. Lett.*, 1996, **76**, 1976.
20 R. J. C. Brown, in preparation.
21 http://www.airquality.co.uk (accessed November 2006).
22 R. J. C. Brown, N. J. Harrison, A. S. Brown, R. E. Yardley and D. Velummylum, Report to the Department of Environment, Food and Rural Affairs by the National Physical Laboratory: annual report for 2004 on the UK Heavy Metals Network, NPL Report DQL-AS 024, January 2006.
23 PM_{10} is defined as: air pollutants consisting of particles with an aerodynamic diameter less than or equal to a 10 micrometers.
24 EN 14902:2005, Ambient air quality – Standard method for the measurement of Pb, Cd, As and Ni in the PM_{10} fraction of suspended particulate matter, CEN, Brussels, 2005.