Math Geol (2007) 39: 469–490
DOI 10.1007/s11004-007-9109-5

Benford's Law Applied to Hydrology Data—Results and Relevance to Other Geophysical Data

Mark J. Nigrini · Steven J. Miller

Received: 24 February 2006 / Accepted: 1 February 2007 / Published online: 29 August 2007

Abstract Benford's Law gives the expected frequencies of the digits in tabulated data and asserts that the lower digits (1, 2, and 3) are expected to occur more frequently than the higher digits. This study tested whether the law applied to two large earth science data sets. The first test analyzed streamflow statistics and the finding was a close conformity to Benford's Law. The second test analyzed the sizes of lakes and wetlands, and the finding was that the data did not conform to Benford's Law. Further analysis showed that the lake and wetland data followed a power law. The expected digit frequencies for data following a power law were derived, and the lake data had a close fit to these expected digit frequencies.

The use of Benford's Law could serve as a quality check for streamflow data subsets, perhaps related to time or geographical area. Also, with the importance of lakes as essential components of the water cycle, either Benford's Law or the expected digit frequencies of data following a power law could be used as an authenticity and validity check on future databases dealing with water bodies. We give several applications and avenues for future research, including an assessment of whether the digit frequencies of data could be used to derive the power law exponent, and whether the digit frequencies could be used to verify the range over which a power law applies. Our results indicate that data related to water bodies should conform to Benford's Law and that nonconformity could be indicators of (a) an incomplete data set, (b) the sample not being representative of the population, (c) excessive rounding of the data, (d) data errors, inconsistencies, or anomalies, and/or (e) conformity to a power law with a large exponent.

M.J. Nigrini (✉)
Department of Business Administration and Accounting, Saint Michael's College, Colchester, VT 05439, USA
e-mail: mnigrini@smcvt.edu

S.J. Miller
Department of Mathematics, Brown University, Providence, RI 02912, USA
e-mail: sjmiller@math.brown.edu

Keywords Data integrity · Hydrographic statistics · Hydrometric statistics · Streamflow analysis · Power law exponent

Introduction

In the 1930s, Frank Benford, a physicist, noted that the first few pages of logarithm tables appeared to be more worn than the later pages. From this he deduced that people were looking up the logarithms of numbers with low first digits (such as 1, 2, or 3) more often than numbers with high first digits (such as 7, 8, or 9). Based on the results of his study of the digits in lists of numbers, and his assumed mathematical properties of numbers, he developed the expected frequencies of the digits in lists of numbers. Under Benford's Law the base 10 probability of a first digit j is $\log_{10}(1 + j^{-1})$, which implies that the first digit is a 1 about 30 percent of the time. The objective of this paper is to (a) test the conformity of two large hydrology-related data sets to Benford's Law, and then to (b) consider the relevance and potential utility of using Benford's Law to assess the integrity and authenticity of earth science and other geological data.

The Benford's Law literature falls into two broad categories. These are papers that either (a) advance the mathematical and statistical theory underlying the law or (b) show a practical application in settings related to uncovering fabricated data. About ten published papers have analyzed data sets that ranged in size from less than 100 records ("small") to around 100,000 records. A recent study by Wallace (2002) used four data sets with only 67 observations each. The data sets analyzed in past papers were related to financial data at the micro level (for a single firm or organization) or at the macro level (for a collection of firms), or to publicly available government or capital markets (stock exchange) data (Nigrini and Mittermaier 1997; Nigrini 2005; Wallace 2002; Ley 1996). In contrast, this paper analyzes two large data sets related to surface hydrology.

The first data set relates to water flows at streamgage sites over an extended period of time (1874 to 2004). This large data set had a near-perfect conformity to Benford's Law. The second data set analyzed was the global lakes and wetlands database. The sizes of these water bodies did not conform to Benford's Law, but the systematic pattern of the digits suggested that these numbers were distributed according to a power law. The Appendix derives the expected digit frequencies for data that follows a power law and notes the special case when such data should follow Benford's Law.

The conclusions are that data related to many hydrological phenomenon should conform to Benford's Law, and nonconformity could be indications of either (a) incomplete data, (b) the sample not being representative of the population, (c) rounding of the data, (d) data errors, (e) systematic biases in the data (rounding up or down to create some effect), or (f) adherence of the data to a power law with an exponent not near 1. Given the importance of hydrologic research to the development, management, and control of water resources, the analysis of digit frequencies could assist in assessing the accuracy, authenticity, and integrity of such data and thereby assist in improving decisions based on archived data.

Benford's Law

Benford (1938) analyzed the digit patterns of 20 data sets with a total of 20,229 observations. His results showed that 30.6 percent of the numbers had a 1 as the first digit, 18.5 percent of the numbers had a 2 as the first digit, with 9 being the first digit only 4.7 percent of the time. The first digit of a number is the leftmost non-zero digit; any minus sign or decimal point is ignored. Thus, the first digit of both 2214 and 0.0025 is a 2. Benford then noticed the logarithmic pattern in the actual digit frequencies and derived the formulas for the expected frequencies of the digits in tabulated data. These are shown with J_1 representing the first digit, and $J_1 J_2$ representing the first-two digits of a number:

$$\text{Prob}(J_1 = j_1) = \text{Log}_{10}(1 + 1/j_1), \quad j_1 \in \{1, 2, \ldots, 9\}, \tag{1}$$

$$\text{Prob}(J_1 J_2 = j_1 j_2) = \text{Log}_{10}(1 + 1/j_1 j_2), \quad j_1 j_2 \in \{10, 11, 12, \ldots, 99\}. \tag{2}$$

Equations (1) and (2) give the Benford's Law formulas for the expected proportions for the first digit and first-two digits. The expected proportions for the first, second, third, and fourth digits are shown in Table 1. This study is concerned with Benford's Law to the base 10 only because the tabulated data that will be analyzed is in base 10. The equations above can be converted to other bases, and Hill (1995) expands upon this concept. In Table 1, from left to right, the digits tend towards being uniformly distributed; this can easily be proved using Poisson summation.

A mathematical basis of Benford's Law is that if the observations are ranked from smallest to largest, they often approximate a geometric sequence. A geometric sequence with N terms is a sequence of the form

$$S_n = ar^{n-1}, \quad n = 1, 2, 3, \ldots, N, \tag{3}$$

where a is the first element of the sequence, and r is the ratio of the $(n + 1)$st element divided by the nth element. The geometric basis of the law was recognized

Table 1 The table gives the expected digit proportions of Benford's Law for the digits in tabulated data for the digits in the first four positions (Nigrini 1996). For example, the table shows that 30.103 percent of the numbers are expected to have a first (leftmost) digit of 1

	Position in number			
Digit	1st	2nd	3rd	4th
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.10882	0.10097	0.10010
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.09940	0.09994
7	0.05799	0.09035	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

by Benford himself in the second part of his paper titled "Geometric Basis of the Law" (Benford 1938) and by Raimi (1976). Raimi (1976, p. 525) discusses the special case where r is a rational power of 10. The conformity of a geometric sequence to Benford's Law depends jointly on the range of the data, the number of observations, and r. Both Benford (1938) and Raimi (1976) discuss situations where conformity to Benford's Law is achieved by data that is asymptotically (approximately) geometric, or where the data consists of a mixture of geometric sequences (interleaving sequences). A recent proof of the geometric basis of the law by Leemis et al. (2000) states: "*Let W be a uniformly distributed random variable on the interval $[a, b]$. If the interval $(10^a, 10^b)$ covers an integer number of orders of magnitude, then the first significant digit of the random variable $T = 10^W$ satisfies Benford's Law exactly*."

The probability distribution of all the digits of the possible values of T follows Benford's Law. T is a random variable, and just one number cannot be "Benford." Therefore, if b–a is an integer and the logarithms base 10 are equidistributed, then the exponentiated numbers follow Benford's Law. Diaconis (1976) provides an early proof of this equivalence, whereas Kontorovich and Miller (2005) and Lagarias and Soundararajan (2006) have recent results using this technique.

Benford noted that his probability law was derived from "events" through the medium of their descriptive numbers, and that it was not a law of numbers in and of themselves. Hill (1995) reviews the relationship between Benford's Law for base 10 numbers and the application of the law to other bases. Pinkham (1961) shows that Benford's Law is scale invariant: starting with a Benford Set (a set of numbers that conforms to Benford's Law) and multiplying all the observations by a nonzero constant, the new data set also follows Benford's Law. Pinkham also showed that only the frequencies of Benford's Law have this property. This attribute of scale invariance was noted by Raimi (1969), who stated that if a data set has non-Benford digit frequencies, then multiplication by a constant never changes the data set to a Benford set. The practical implication of the Pinkham theorem is if a Benford set is calibrated in cubic feet per second and then restated in cubic meters per second, the restated data set is also Benford. In the theorem by Leemis et al. (2000) noted above, if the random variable T were multiplied by a nonzero constant 10^x then the data would cover the interval $(10^{a+x}, 10^{b+x})$, which would still be an integer order of magnitude.

Prior research has analyzed financial data sets for conformity to Benford's Law. Nigrini (1996) showed that the digit frequencies of the interest amounts received on 91,022 tax returns for 1985 and 78,640 tax returns for 1988 had a close conformity to Benford's Law. The dollar amounts of 30,084 invoices approved for payment by a NYSE-listed oil company (Nigrini and Mittermaier 1997) and the dollar amounts of 36,515 invoices approved for payment by a software company (Drake and Nigrini 2000) also conformed to Benford's Law. Nigrini (2005) analyzed the revenue numbers from 4792 quarterly earnings releases in 2001 and 4196 quarterly earnings releases in 2002. The first digits of the revenue numbers conformed to Benford's Law. However, the second digits showed a pattern (excess second digit 0's and a shortage of second digit 9's) that was consistent with rounding up of revenue numbers around psychological reference points, such as US $200 million. Financial data within and across firms conforms reasonably well to Benford's Law.

Data Description and Analysis

Two sets of hydrological data were analyzed. The first set was streamflow data obtained from the U.S. Geological Survey, and the second set was related to lakes and wetlands. The analysis of the lakes and wetlands data showed that a relationship between Benford's Law and the power law (often used to describe the relative sizes of geological phenomenon) exists in nature.

Streamflow Data

Streamflow data was obtained from the U.S. Geological Survey (USGS) website. The agency's website lists many programs, including the National Streamflow Information Program (NSIP). Under this program the USGS operates and maintains approximately 7300 streamgages which provide data for many diverse users. There are several reasons for the collection of accurate, regular, and dependable streamflow data:

- Interstate and international waters—interstate compacts, court decrees, and international treaties may require long-term, accurate, and unbiased streamflow data at key points in a river.
- Streamflow forecasts—upstream flow data is used for flood and drought forecasting by the National Weather Service for improved estimates of risk and impacts for better hazard response and mitigation.
- Sentinel watersheds—accurate streamflow data is needed to describe the changes in the watersheds due to changes in climate, land and water use.
- Water quality—streamflow data is a component of the water quality program of the USGS.
- Design of bridges and other structures—streamflow data is required for water level and discharge during flood conditions.
- Endangered species—data is required for an assessment of survival in times of low flows.

The methods employed for measuring flow at most streamgages are almost identical to those used 100 years ago. Acoustic Doppler technology can widen the range of conditions for which accurate flow measurements are possible, but is not yet seen as providing enhanced efficiency or accuracy at most locations. New technology has yet to be developed to provide more accurate data over a wide range of hydrologic conditions, and more cost-effective than the traditional current meter methods.

The data for this study was obtained from the Surface–Water Data for the Nation. The data used was the annual data Calendar Year Streamflow Statistics for the Nation. To obtain a large data set the only condition that was imposed was that the period of record included calendar year 1950 or later. The data consisted of all the annual average readings for any site that had an annual average recorded in any of the years from 1950 to 2005. The only sites that were excluded were sites that only had data for the pre-1950 period. The fields downloaded were: (a) agency code, (b) USGS site number, (c) calendar year for value, and (d) annual mean value in cubic feet per second. Summary statistics are shown in Table 2.

Table 2 The table describes the annual streamflow data used in the study. The data pertained to U.S. rivers and streams and was obtained from the U.S. Geological Survey's website (http://www.usgs.gov/)

Description	Amount	Units
Number of observations from download	459,778	Records
Number of observations equal to zero	1,706	Records
Number of observations with a negative flow	108	Records
Number of null (blank) observations	1	Records
Number of usable observations	457,963	Records
Statistics for usable observations:		
Number of duplicate records	523	Records
Number of sites with duplicate records	12	Sites
Number of observations after deletion of duplicates	457,440	Records
Statistics of used observations:		
Number of unique sites	17,822	Sites
Highest record count for a single site	130	Records
Lowest record count for a single site	1	Record
Average count for each listed site	25.7	Records
Latest year on record for any site	2004	Calendar year
Earliest year on record for any site	1874	Calendar year
Year with the highest record count	1967	Calendar year
Year with the lowest record count	1874	Calendar year
Minimum flow for any single site year	0.001	Cubic feet per second
Quartile 1	34.8	Cubic feet per second
Quartile 2 (median)	166.0	Cubic feet per second
Quartile 3	674.0	Cubic feet per second
Maximum flow for any single site year	980,900.000	Cubic feet per second
Average flow over all records	2,199.087	Cubic feet per second

The downloaded data included some duplicate records for calendar year and annual mean value. An inspection of the duplicates showed that only the agency code differed between the duplicates. For example, one duplicate showed the agency code to be USIBW and another identical record showed the agency code to be USGS. The deletion of the duplicates ensured that any site and year would be used only once in the analysis.

An analysis of the 1706 zero flows showed that there were 495 sites that had a zero recorded for 1 or more years and that zero flows occurred in 80 different years indicating that this phenomenon was not restricted to a certain period of time. There were 199 sites that had a zero recorded for only one year and 296 sites that had zeroes recorded for more than one year. The results suggested that the zeroes were not data errors, but that the rivers either dried up or were diverted around the location of the original streamgage. The zeroes were ignored in the data analysis because they are essentially a non-event. In contrast, the 108 negative numbers did seem to be data

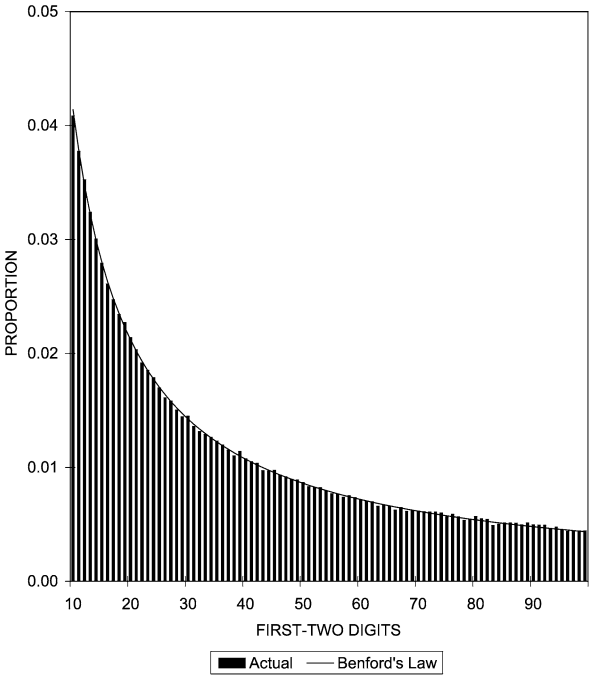

Fig. 1 The graph shows the first-two digit proportions of the streamflow data and the expected proportions of Benford's Law

errors thereby confirming the importance of data cleansing prior to analyzing the data.

The number of observations remaining after deletion of the null values, zeroes, negative numbers and duplicates was 457,440 records. This data set was particularly interesting because (a) the period covered is 130 years and it is rare for any data set to cover such an extended period, (b) the data set was the largest analyzed in the Benford's Law literature to date, (c) the range in streamflows indicated that the sites covered everything from the smallest streams to the largest waterways, (d) the measurement technology has been unchanged over the entire period, which suggests that there are no distortions due to technological changes, and (e) the data set is used for a variety of important purposes.

Most of the prior Benford's Law studies analyzed the first or second digits of the data under scrutiny. In this study the first-two digits are analyzed (see (2)) because the first-two digits reveal data anomalies that would be missed with an analysis of only the first or second digits. For example, the 47, 48 and 49 might all be overstated by x percent and if the 41, 42, and 43 are also all understated by x percent, then the first digit 4 would have an actual proportion that closely matched the expected proportion. However, an analysis of first-two digits would highlight these deviations which would present a more accurate assessment of the digit frequencies. The first-two digits of the streamflow numbers are shown in Fig. 1.

The graph shows the expected proportions of Benford's Law as a smooth monotonically decreasing line from 0.41 to 0.044. The actual proportions are shown as vertical bars. There are 90 bins and on average each bin is about 0.011. The visual fit to Benford's Law is excellent with a Mean Absolute Deviation (average of |Actual

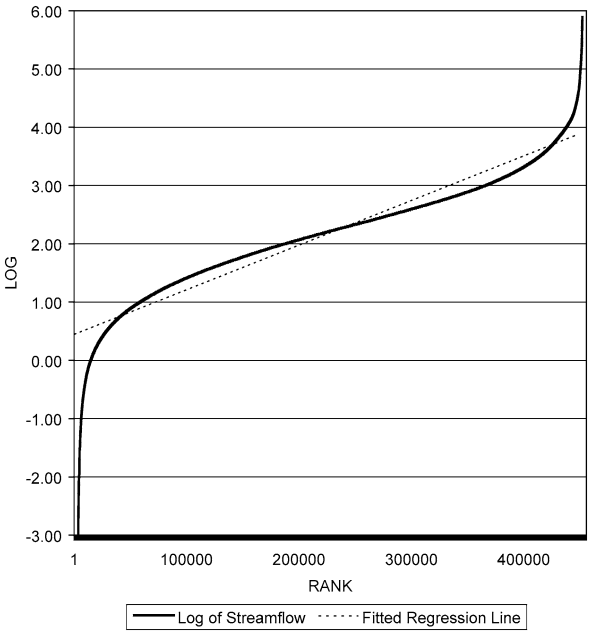

Fig. 2 The graph shows the ordered values of the logs (base 10) of the streamflow data with each of the 457,440 observations representing the average annual flow at a USGS monitoring station together with a fitted regression line

– Benford's Law|) of 0.00013. The low Mean Absolute Deviation means that, on average, the deviation of the actual percentage from that of Benford's Law was one-tenth of one percent. A visual review of the graph shows no sign of the "overs" or "unders" being clustered in certain parts of the graph, nor are any of the overs or unders systematic by occurring, for example, at multiples of 10 (10, 20, 30, ..., 90). The near-perfect visual fit to Benford's Law suggests that the data is consistent with the geometric pattern (or a combination of interweaving geometric series) assumed by Benford's Law. To further explore the anatomy and structure of the data, the base 10 logarithms of the ordered values were graphed. A regression line was fitted with the predictor variable (X) being Rank (1 to 457,440) and the response variable (Y) being the logarithm of the annual flow.

Figure 2 shows the graph of the logs of the annual flow data and the regression line obtained by regressing the logarithm of the streamflow on the Rank. The R-squared value is 0.918. The first intersection between the actual and the fitted line is at Rank = 36,139 and the last intersection between the two lines is at Rank = 428,359. This means that about 85.7 percent of the observations are very "close" to the fitted line. If all the observations were "close" to the fitted line then this would indicate that the data could be described as a single geometric series with a constant ratio r. The graph seems to be made up of three (connected) lines with three different slopes which suggests that the data comes from three successive geometric series, and that the "average" result is that the logs modulo 1 are equidistributed. A formal test of the mantissas (the fractional part of the logs) is presented in Fig. 3.

Figure 3 shows a plot of the ordered values of the mantissas. The graph also shows a plot of a regression line using the following equation

$$Y_i = -(1/N) + (1/N) \cdot \text{Rank}, \quad i = 1, 2, 3, \ldots, N, \tag{4}$$

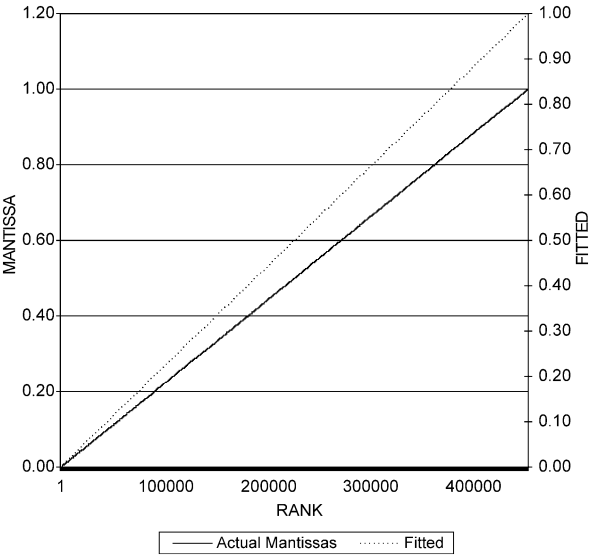

Fig. 3 The graph shows the ordered values of the mantissas of the streamflow data plotted as a *solid line* using the *left-hand side Y-axis* for the values. A fitted line showing a set of mantissas uniformly distributed over the [0,1) interval is shown as the *dotted line* using the *right-hand side Y-axis* for the values

where N equals the number of observations (457,440). The line described in (4) is the line that would result from the mantissas distributed at $0/N, 1/N, 2/N, \ldots, (N-1)/N$, which would be close enough to being equidistributed for all practical purposes. The two lines are both straight lines from 0 to 1, implying that the mantissas are equidistributed and that the data conforms closely to Benford's Law. The close level of conformity to Benford's Law is also clear from the digit frequencies presented in Fig. 1. The near-perfect fit of the streamflow data to Benford's Law is the closest fit of any set of natural data (as opposed to simulated data) to Benford's Law in the literature.

This data set is particularly interesting because while the fit is visually appealing, it is not a perfect fit to Benford's Law. We explore the results in more detail to set the stage for the analysis of the lake data and to offer some guidance to other researchers investigating archived earth science and other geological data. The first set of tests relates to the goodness of fit to Benford's Law, while the second set of tests relates to the internal structure of the data.

Goodness of Fit Tests

The chi-square test was used to measure the goodness-of-fit to Benford's Law. For the first-two digits (with i from 10 to 99) the computed value of the chi-squared statistic was 122.595. The critical point of the chi-square distribution with 89 degrees of freedom and a right-hand tail area of $\alpha = 0.05$ is 112.02, and the test therefore calls for a rejection of the null hypothesis and the data conforms to Benford's Law; however, the critical point for a right-hand tail area of $\alpha = 0.01$ is 122.94, and thus the test would not call for a rejection of the null hypothesis at the 99% confidence level. A second goodness-of-fit test employed was the Kolmogorov–Smirnov test. The calculated D-statistic (maximum difference between the actual and expected distribution functions) was 0.0017, which was compared to the critical value at $\alpha =$

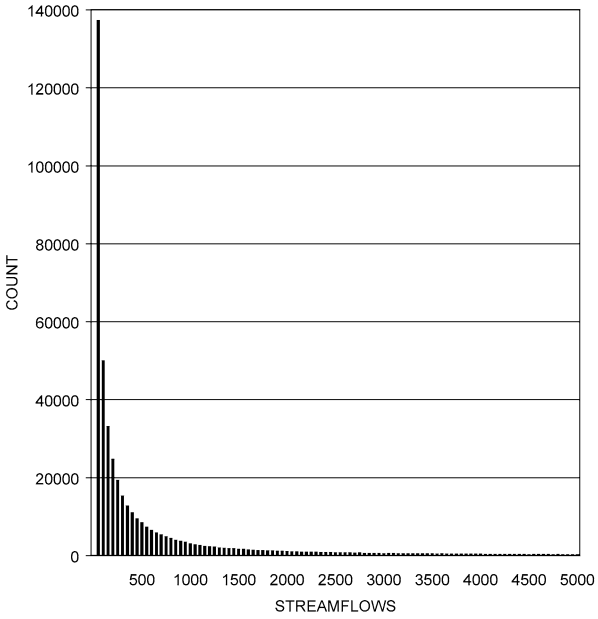

Fig. 4 The figure shows the counts of the streamflow data in the form of a histogram. Each *bar* covers a range of 50 cubic feet per second. The counts for values above 5000 cubic feet per second are small and are not shown on the figure

0.05 of $1.36 \cdot \sqrt{N} = 0.0020$. At $\alpha = 0.05$ the evidence is not persuasive enough to reject the null hypothesis that the data conforms to Benford's Law. The goodness-of-fit tests therefore indicates that at an α of 0.05, the null hypothesis of conformity is narrowly rejected by the chi-square test and narrowly accepted by the Kolmogorov–Smirnov test (and not rejected by a chi-square test at an α of 0.01).

Given the narrow margins for the reject/accept goodness-of-fit decisions, we performed a runs test to investigate whether the overs and unders were randomly distributed for the first-two digits. For each of the 90 bins, an over occurs when the actual proportion exceeds that of Benford's Law, and an under represents the converse. Letting n_1 denote the number of overs, n_2 the number of unders, and u the number of runs of overs and unders (for example, the sequence 'over over under under under over under over' has 5 runs), there were $u = 38$ runs with $n_1 = 48$ and $n_2 = 42$. As n_1 and n_2 are larger than 30, u should be approximately normally distributed. The computed value of the Z-test statistic was -1.661 which is less than the cutoff of 1.96 (at $\alpha = 0.05$) indicating that the overs and unders do not have a systematic pattern.

Given (a) the narrow margins for the goodness-of-fit tests, (b) the results of the lake data tests, and (c) the observation by DeGroot and Schervish (2002) that prior to summarily rejecting the null hypothesis in cases where the sample size is large (due to small differences having a high impact on the calculated statistics), the statistician should consider other plausible distribution functions with which the sample provides a closer agreement. The final test was whether the data follows a power law, and whether this could be the cause of the (admittedly small) deviations from Benford's Law. A histogram was plotted to see whether it had the properties expected for a power law.

Figure 4 shows a histogram of the counts of the streamflow values in bins with a range of 50 cubic feet per second up to 5000 cubic feet per second. The histogram

shows a pattern consistent with data following a power law. The next step in the analysis was to calculate the power law exponent. This was done following the methodology in Newman (2005), which provided an excellent review of power laws and the Pareto distribution. A power law has the density function

$$f_{a,b,m}(x) = C(a,b,m)x^{-(m+1)} \quad \text{for } x \text{ in } [a,b] \text{ and 0 otherwise,} \tag{5}$$

where the range $[a,b]$ is restricted such that $[a,b] = [10^k, 10^n]$, with k and n integral, and $m+1$ the exponent. We use this form of the exponent as it simplifies future formulas (see the Appendix); $m = 0$ corresponds to Benford behaviour.

Given the range of the streamflow data (0.001 to 980,900), the values of k and n are -3 and 6, respectively. Newman (2005) gives a simple and reliable method for extracting the exponent using

$$m + 1 = 1 + N\left[\sum_{i=1}^{N} \ln(x_i/x_{\min})\right], \tag{6}$$

where the quantities x_i, $i = 1, \ldots, N$, are the measured (observed) values of x, and $x_{\min}$ is the minimum value of x. Using (6), the power law exponent was calculated to be 1.084 (giving $m = 0.084$). The error estimate for $m+1$ is difficult to calculate since this needs to be done using a standard bootstrap or jackknife resampling method, and also because the equation is really only valid for the range over which the power law is expected to hold. For example, the density in (5) cannot hold for arbitrarily large values of x if $m \leq 0$. The calculated exponent is close to, but not exactly equal to, 1; the digit bias is Benford if and only if the exponent is 1. If the exponent is not 1, then this is a possible explanation for the small differences between the actual and expected digit distributions of the streamflow data, as evidenced by the narrow reject/accept goodness-of-fit test results. The relationship between data following a power law with an exponent not equal to 1 and the expected frequencies of Benford's Law is explored further in the next sections and the Appendix. A second explanation for the differences could be that even though the sample size is large, the data set might not be a perfect representation of streamflow statistics. The sample might be biased because the measuring stations are not perfectly randomly dispersed throughout the waterways of the United States. Although the conformity of this large streamflow data set is not perfect, the conclusion is that the fit is excellent for all practical purposes.

Lake and Wetlands Data

The data was obtained from the global lakes and wetlands database (GLWD) developed by the Center for Environmental Systems Research at the University of Kassel. This database is described in detail in Lehner and Döll (2004). The data analyzed was for large lakes and reservoirs, and smaller water bodies (GLWD-1 appended to GLWD-2). The data provided statistics on 248,613 water bodies.

Lehner and Döll (2004) include a review of the importance of knowledge about water bodies. There are issues in defining exactly what constitutes a lake. For example, for lakes adjacent to the sea (also called lagoons) the distinction between slow-moving rivers and lakes may be ambiguous. There may also be a continuum between

Table 3 The table describes the lake, river, and reservoir data used in the study. The data was obtained from the global lakes and wetlands database (GLWD) developed by the Center for Environmental Systems Research at the University of Kassel

Description	Amount	Units
Number of observations from download	248,613	Records
Number of observations used	248,613	Records
Classifications:		
Lakes	246,135	Units
Rivers	1656	Units
Reservoirs	822	Units
Perimeter Statistics		
Minimum perimeter	1.0	Kilometers
Quartile 1	4.0	Kilometers
Quartile 2 (median)	5.6	Kilometers
Quartile 3	9.3	Kilometers
Maximum perimeter	36,641.2	Kilometers
Average perimeter over all records	14.6	Kilometers
Area Statistics		
Minimum area	0.1	Square kilometers
Quartile 1	0.9	Square kilometers
Quartile 2 (median)	1.5	Square kilometers
Quartile 3	3.1	Square kilometers
Maximum area	378,119.3	Square kilometers
Average area over all records	12.2	Square kilometers

lakes and wetlands. The authors define lakes to be permanent still water bodies (lentic water bodies) without a direct connection to the sea, but they accepted saline lakes and lagoons (but not lagoon areas) as lakes, and also manmade reservoirs. Their database excluded intermittent or ephemeral water bodies. The database was restricted to lakes with an area greater than 0.1 km^2 (1 hectare).

Table 3 shows that 99 percent of the water bodies are lakes. All the records in the database were included in the analysis to keep the data set consistent with prior studies that have used this data, and for consistency with possible future studies. The perimeter field had values ranging from 1 km to 36,641 km. For the perimeter data, the first and third quartile values of 4.0 and 9.3, respectively, show that close to 50 percent of the lakes had perimeters from 4.0 to 9.3 km. The data is strongly positively skewed. The area of the lakes is also strongly positively skewed with 29.6 percent of the lakes having areas under 1 km^2 and 31.4 percent of lakes having areas in the 1.0 to 1.9 km^2 range.

The first-two digit patterns of the perimeter numbers are shown in Fig. 5; the data does not conform to Benford's Law. For the first-two digits, the computed value of the chi-squared statistic was 88,120, exceeding the $\alpha = 0.05$ critical point of the chi-square distribution with 89 degrees of freedom by a large margin. The test calls for

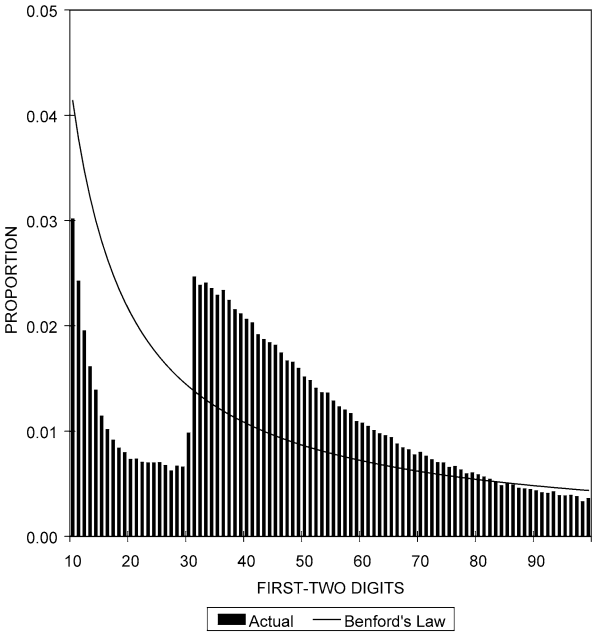

Fig. 5 The graph shows the first-two digit proportions of the lake perimeter data and the expected proportions of Benford's Law

a rejection of the null hypothesis that the data conforms to Benford's Law. Using the Kolmogorov–Smirnov test, the calculated D-statistic (the largest difference between the actual and expected distribution functions) was 0.2619, exceeding the $\alpha = 0.05$ critical value of $1.36 \cdot \sqrt{N} = 0.0027$ by a wide margin. Finally, as expected with the overs (n_1) and unders (n_2) clustered into clear groups, the runs test showed that there were 3 runs with $n_1 = 52$ and $n_2 = 38$. The computed value of the Z-statistic was -9.109, far greater than the cutoff of 1.96 (at $\alpha = 0.05$). This indicates that the overs and unders had a systematic pattern.

The distribution of the perimeter values in Table 4 suggests that the data set starts with lakes that are 3.0 km or larger and indicates that the range of 3.0 to 4.9 km dominates the data. Slightly over one-third of the lakes have perimeters in the 3.0 to 4.9 km range. The non-Benford digit patterns confirm that the data set has a minimum value of 3. There are some lakes with perimeters <3, but these are relatively few in number. One possible reason for the non-Benford behaviour could be the lack of a clear definition of a lake. The fractal nature of the perimeters of lakes might play some part in accurately measuring the perimeters. Another possible explanation is that perimeter is not a correct measurement for the size of a lake. The next step was to analyze the digit patterns of the surface areas of lakes since these might provide a better measurement of size.

Figure 6 shows the first-two digit patterns of the surface areas of the 248,613 lakes. The data does not conform to Benford's Law given a Mean Absolute Deviation (average of |Actual − Expected|) of 0.0071. On average, the actual proportion differed from the expected proportion by seven-tenths of one percent. The largest deviations occurred for the high round value combinations (50, 60, 70, 80 and 90). The chi-square test produced an even larger test statistic (at 544,735) than for the perimeters,

Table 4 The table shows the number and percentages for the lake perimeter values in the GLWD data set in increments of 1 km, from 1 km to 10 km, with the final row showing the count and percentage for lakes greater than or equal to 10 km. Perimeter values were rounded to one-tenth of one km

From (km)	To (km)	Count	Percentage
1.0	1.9	1,055	0.42
2.0	2.9	7,792	3.13
3.0	3.9	49,887	20.07
4.0	4.9	42,911	17.26
5.0	5.9	31,140	12.53
6.0	6.9	22,108	8.89
7.0	7.9	16,212	6.52
8.0	8.9	11,926	4.80
9.0	9.9	9,266	3.73
10.0	36,641.2	56,316	22.65
Total		248,613	100.00

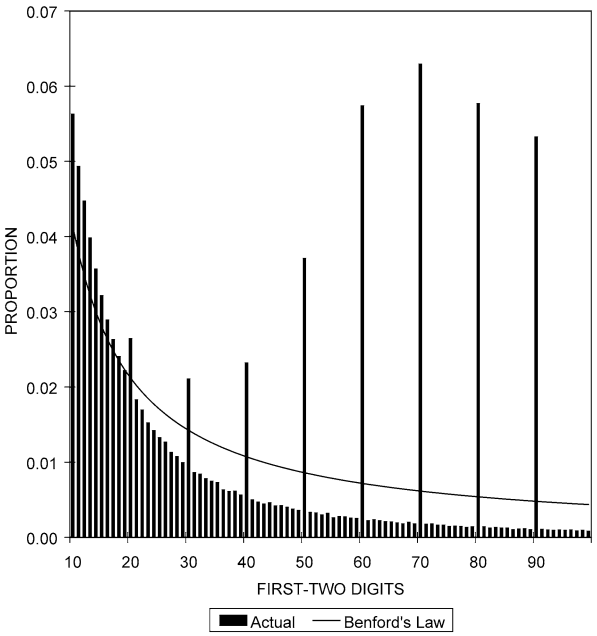

Fig. 6 The graph shows the first-two digit proportions of the lake surface areas and the expected proportions of Benford's Law

but the Kolmogorov–Smirnov test statistic was lower at 0.0635. The null hypothesis of conformity was still soundly rejected by both tests. The runs test also rejected the null hypothesis of a random distribution since the overs (n_1) and unders (n_2) were clustered into clear groups. The test showed that there were 18 runs with $n_1 = 17$ and $n_2 = 73$ giving a computed value of the Z-statistic of -3.686, which was again greater than the cutoff of 1.96 (at $\alpha = 0.05$).

A review of the data showed that 29.61 percent of the values were less than 1.0. These values were recorded to one decimal place only and were therefore recorded as

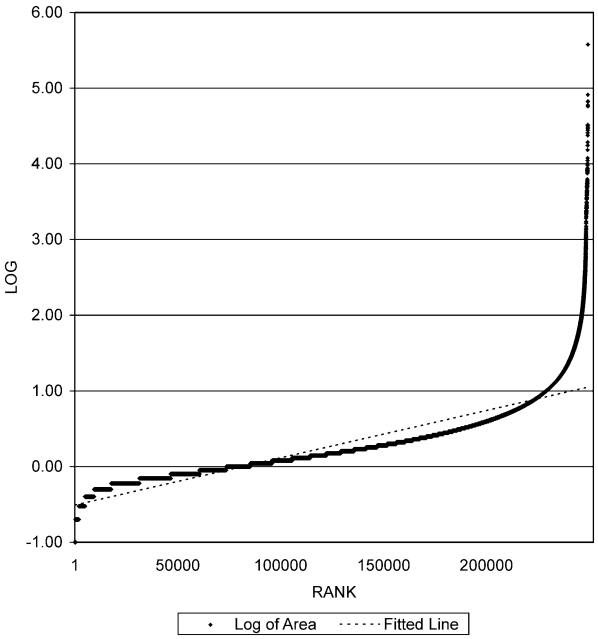

Fig. 7 The figure shows the ordered values of the logs (base 10) of the 248,613 lake surface areas from the GLWD data. The *straight line* is the fitted regression line. The horizontal "steps" evident in the first 100,000 data points show that there are many observations with the same numeric values

0.1, 0.2, 0.3, . . . , 0.9. The values from 0.1 to 0.9 were given imputed first-two digit values of 10, 20, 30, . . . , 90 since 0.1 can be written as 0.10 and 0.2 can be written as 0.20. These <1 values were so numerous that they distorted the digit patterns. The round first-two digit values (10 through 90) were not true 10s, 20s, 30s, . . . , 90s, but occurred because the data was rounded. For example, the area could have been calculated to be 0.481957 km^2 and then rounded to 0.50 km^2. This small amount of rounding would not only change the second digit to a 0 but would also change the first digit. The <1 values were deleted to explore the patterns of the remaining values. Prior to this (to further explore the structure of the data), the logarithms (base 10) of the numbers were calculated and graphed similar to what was done for the streamflow data.

The graph in Fig. 7 shows the ordered values of the lake areas and shows a similar pattern to that of Fig. 2. The left side of the curve has more of a curve to it and several horizontal steps can be seen. These horizontal steps indicate that there are runs with equal values causing sections of the line to have a slope of zero. The basic shape of the lake area graph is similar to the streamflow graph and yet the streamflow data conformed more closely to Benford's Law. The next step was to separate the rounded values (<1) from the remainder and then to investigate separately the values greater than 1. The digit frequencies of the areas ≥ 1 are shown in Fig. 8.

The digit frequencies of the lake areas ≥ 1.0 km^2 in Fig. 8 show a smooth pattern. There is a pronounced skewness, and the downward sloping curve of the actual proportions is more extreme than that of Benford's Law. The Mean Absolute Deviation of 0.0052 is less than that for Fig. 6. As compared to the complete set of lake areas, the chi-square test statistic was smaller at 49,165, but the calculated test statistic for the Kolmogorov–Smirnov test was about four times larger at 0.2341. Again the null

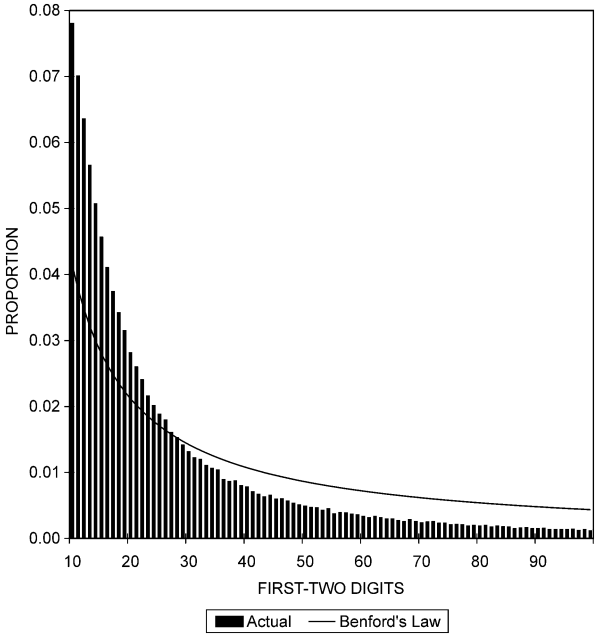

Fig. 8 The graph shows the first-two digit proportions of the lake surface areas for those areas ≥1.0 and the expected proportions of Benford's Law

hypothesis of conformity was soundly rejected by both tests. The runs gave an interesting result when testing the null hypothesis of a random distribution, given the overs (n_1) and unders (n_2) were again clustered into clear groups. There were 2 runs with $n_1 = 19$ and $n_2 = 71$, giving a computed value of the Z-statistic of -9.275, which was again far greater than the cutoff of 1.96 (at $\alpha = 0.05$) and about three times as large as the test statistic for the full lakes data set.

Power Law Association

To further investigate the internal structure of the data, a histogram was constructed and the results are shown in Fig. 9.

Figure 9 is a histogram of the lake areas with the area (1 km^2 ≤ area < 1,000 km^2) plotted on the X-axis and the count on the Y-axis. There were only 236 lakes with an area >1,000 km^2. The axes in Fig. 9 were plotted on logarithmic scales. The negative slope of the line coupled with the "noisy" results on the right hand side is an almost perfect representation of data that follows a power law. The value of the exponent $(m + 1)$ was calculated using (6) with an $x_{\min} = 1$ giving a calculated value of 1.941. The Appendix shows that it is only with an exponent approximately equal to 1 ($m = 0$) that the data will conform to Benford's law. The Appendix also includes an expectation for the first and first-two digits for data that follows a power law for exponents $(m + 1) > 1$.

The penultimate test was to test for a systematic pattern to the ordered values of the mantissas of the power law data and these results are shown in Fig. 10. If the data conformed to Benford's Law, then the mantissas would follow the path of the regression line plotted using (4) with $N = 175{,}009$ (the count for areas greater

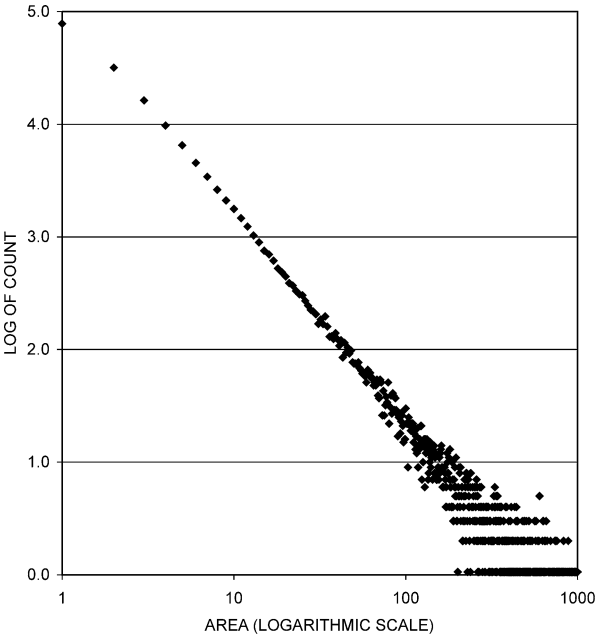

Fig. 9 The figure shows a plot of the lake areas (from 1 to 1000 km^2) and the counts for these values on a logarithmic scale

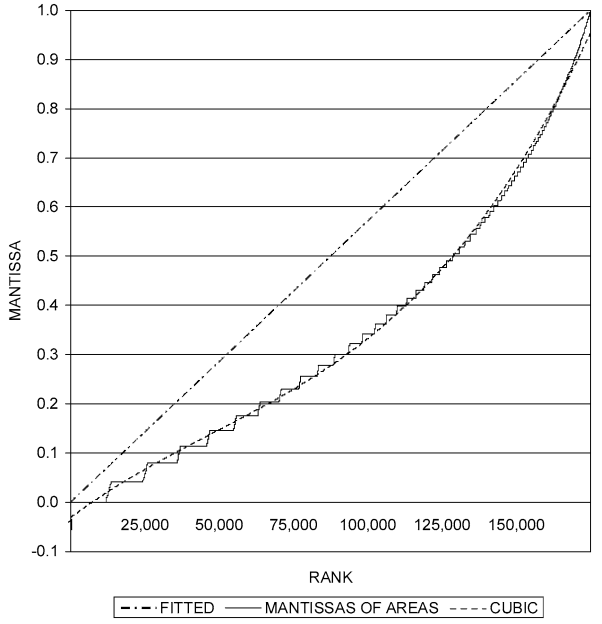

Fig. 10 The figure shows the ordered values of the mantissas of the lake surface areas for areas greater than or equal to 1.0 km^2. Also shown is a *straight line* representing a uniform distribution over the [0,1) interval and a fitted line of a cubic equation with $Y = a + bX + cX^2 + dX^3$ fitted to the actual mantissa values

than 1). The results show that the actual plot of the mantissas follows a curved path from 0 to 1, indicating that the mantissas are not equidistributed. The visible "steps" in the graph occurred because there were many numbers that were repeated (e.g. 1.0 and 1.1) and the repeating numbers have equal mantissas. The mantissas seem

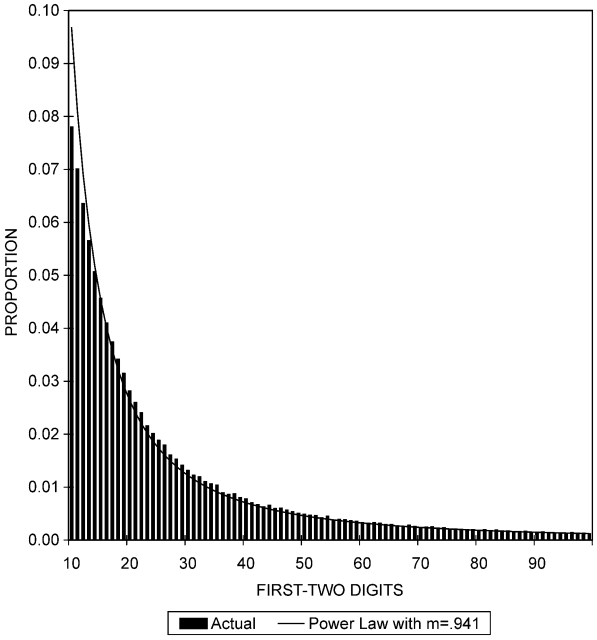

Fig. 11 The graph shows the first-two digit proportions of the lake surface areas for those areas ≥ 1.0 and the expected proportions of data distributed according to a power law with an exponent of 1.941

to follow a pattern that could be approximated with a cubic equation of the form $Y = a + bX + cX^2 + dX^3$. A cubic line was fitted using regression and the fitted line is shown in Fig. 10. The fact that there was a systematic pattern to the ordered mantissas and to the digit patterns of the power law data suggested that a functional form for the expected digit patterns could be derived for $m > 0$. Results are shown in the Appendix.

The final test for the lake areas ≥ 1 was to test the actual first-two digit frequencies against the expected first-two digit frequencies for a power law data using the theorem in the Appendix. The results are shown in Fig. 11.

Figure 11 shows the actual first-two digit frequencies of the lake areas ≥ 1 and the expected first-two digit frequencies of data distributed according to a power law with an exponent of 1.941 using the theorem in the Appendix. The fit is a visually close fit. The differences could be due to a number of issues. The exponent of 1.941 was calculated for data in the [1,1000] range whereas the graph shows the digit frequencies for all lakes ≥ 1 km^2. The analysis was done for all lakes ≥ 1 to allow for comparisons between future research studies on lake areas using the GLWD database and for researchers using the same data source for other purposes. Also, there is no perfect method for calculating the exponent for the data that follows a power law. Finally, the differences could also result from the data not following a power law exactly over the entire [1,1000] range, or it could signal some issues in measuring the areas of lakes. The following section reviews and discusses the findings related to the lake data.

Discussion of Power Laws and Digit Frequencies

The Appendix shows that a data set characterized by a power law with exponent $m+1$ will tend towards having the digit frequencies of Benford's Law if m is small. As m increases, provided that the range of the numbers approximates $a = 10^k$ and $b = 10^n$ for integers k and n, the digit patterns are more skewed than for Benford's Law. The Appendix concludes with a theorem giving the explicit expected digit frequencies for data following a power law for any m.

Power laws are known to describe the relative sizes of many natural phenomena and Newman (2005) reviews many diverse instances of data conforming to power laws including earthquake magnitudes, crater diameters, peak gamma ray intensities of solar flares, and the numbers of species in biological taxa. Newman's calculated exponents α generally ranged from 2.0 to 3.5.

Results from Fig. 11 suggests future research could investigate the formal question as to whether digit frequencies could be used (a) to confirm that data follows a power law, (b) to estimate the exponent of the power law, and (c) to assist researchers in confirming the range over which the power law holds true. Researchers could also analyze the digit patterns of data known to follow a power law to assess what types of data integrity issues (e.g. such as incomplete data sets or errors in the measuring apparatus used) could be detected. Finally, if a coherent and extensive body of knowledge is developed, the digit frequencies might even give rise to researchers being able to suggest correction factors (data subset N_1 appears to be under- or overstated by y percent) for data that is inherently difficult or costly to measure precisely.

The practical uses of such research could be that researchers evaluating earth science and other geological data might be able to use the digit frequencies to assess data authenticity issues. This topic is highly relevant given the recent issues that have surfaced in stem cell research (The Economist 2006). If the digit frequencies do not follow the expected patterns then such analyses might support conclusions that the data is possibly (a) highly rounded, (b) incomplete, (c) biased (evidenced by excessive rounding upwards or downwards, usually found by an examination of second or later digits), or (d) subject to intentional or unintentional errors. Furthermore, the analysis of digit frequencies and the distribution of the logarithms could also shed additional light on the internal structure of the data by providing insights that are not apparent from summary statistics such as the mean, median, and standard deviation.

Conclusions

Benford's Law provides the expected digit frequencies for data sets describing many diverse natural phenomena. To date there have been relatively few papers that have analyzed the digit patterns of actual data. Most of these papers have dealt with financial data. The objective of this paper was to analyze the digit frequencies of two large data sets related to surface hydrology and thereafter to comment on the possible utility of the results for researchers analyzing earth sciences and other geological data.

The first data set analyzed was annual average flows at streamgage sites throughout the U.S. over an extended period of time (1874 to 2004). This large data set of

457,440 records had a near-perfect conformity to Benford's Law. The second data set analyzed was the global lakes and wetlands database (GLWD) with 248,613 records on lakes, rivers, and reservoirs. The perimeters of the water bodies did not conform to Benford's Law. The areas of the lakes also deviated from the expected patterns of Benford's Law, but the systematic pattern of the digits indicated that these numbers were distributed according to a power law. This paper showed that, under certain circumstances related both to the power law exponent and the range of the data, there was a close conformity to Benford's Law. Under other circumstances the fit would be weak. Several avenues for future research related to the relationship between data conforming to a power law and the expected digit patterns for such data exists. These avenues include using the digit frequencies to ascertain the range for which the power law is valid and using the frequencies to examine data authenticity and data integrity issues.

From a practical perspective the streamflow results suggest that data related to water bodies should conform to Benford's Law. Nonconformity could be indicators of either (a) an incomplete data set, (b) excessive rounding of the data, (c) data errors, or (d) adherence of the data to a power law with a high value for the exponent. Given the importance of the streamflow data for flood prediction, adherence to interstate covenants, bridge construction, and the preservation of endangered species, the use of Benford's Law can serve as a quality check for subsets (perhaps related to time or geographical area) of the data. Lakes are essential components of the hydrological and biogeochemical water cycles with influences on many aspects of ecology, the economy, and human welfare. Either Benford's Law or the expected digit frequencies of data following a power law could be used as a validity check on future generations of databases containing data related to water bodies.

Acknowledgements We wish to thank George Ashline, Alain Brizard, Darlene Olsen, Michael Popowits, and the editor and reviewers of an earlier version of this paper for their careful and constructive comments. Steven Miller was partly supported by NSF grant DMS0600848.

Appendix Relationship Between Benford's Law and Power Laws

Consider a power law with density $f_{a,b,m}(x) = C(a,b,m)x^{-(m+1)}$ for x in $[a,b]$ and 0 otherwise. The admissible values of a (which can be zero) and b (which can be ∞) depend on the exponent m. To ensure that the integral of $f_{a,b,m}(x)$ is finite, if $m \leq 0$ then $b < \infty$, while if $m \geq 0$ then $a > 0$. The normalization constant $C(a,b,m)$ is easily computed; it is $1/\log(b/a)$ if $m = 0, m/(a^{-m} - b^{-m})$ if $m > 0$, and $|m|/(b^{|m|} - a^{|m|})$ if $m < 0$.

Power law distributions are related to Benford's Law. Let Y be a random variable with density given by a power law with $m = 0, a = 10^k$ and $b = 10^n$ (for integers k and n). The distribution of the digits of Y base 10 are Benford. This immediately follows from studying the density on intervals $[10^j, 10^{j+1})$. As each such interval has probability $\log(10)/C(10^k, 10^n, 0) = 1/(n-k)$, it is sufficient to consider the special case when $n = k+1$. In this case, the normalization constant is $\log(10)$ and the probability of observing a first digit of d is $(\log(d \cdot 10^k) - \log((d+1) \cdot 10^k))/\log(10) = \log_{10}(1+d^{-1})$, which is the Benford probability. If $[a,b]$ is not of the form $[10^k, 10^n]$

for some integers k and n, then while there will be digit bias, it obviously cannot be Benford. For example, if $[a, b] = [2, 4]$ then the first digit can only be a 2, 3 or a 4! While the distribution is not Benford, in a restricted sense it will have Benford behaviour. In this case, the ratio of the probability of a first digit of 2 versus a first digit of 3 is $\log(3/2)/\log(4/3)$, exactly the same answer for Benford's Law. Unfortunately, such a property will not hold for all intervals (consider $[a, b] = [1, 30]$).

Let Y be a random variable with density given by the power law $f_{a,b,m}(x)$. Assume $m \neq 0$ and $[a, b] = [10^k, 10^n]$ (with k and n integers, k possibly equal to $-\infty$ or n equal to ∞). If m is small, then the distribution of digits of Y will be close to Benford's Law. This follows from continuity of integration of continuous functions. For example, consider $[a, b] = [10^j, 10^{j+1}]$ with j an integer, and let m be a small positive number; the case of negative m is handled similarly (and yields the same formula). The difference in the probability of a first digit of d is given by integrating $f_{10^j,10^{j+1},m}(x) - f_{10^j,10^{j+1},0}(x)$ from $d \cdot 10^j$ to $(d+1) \cdot 10^j$. Simple algebra shows the integral of $f_{10^j,10^{j+1},m}(x)$ is

$$d^{-m} \cdot \left(1 - \left(\frac{d+1}{d}\right)^{-m}\right) \Big/ \left(1 - 10^{-m}\right).$$

As $m \to 0$ through positive values, d^{-m} tends to 1, and by L'Hospital's rule (remember we differentiate with respect to m) and the change of base theorem for logarithms

$$\begin{aligned}
&\lim_{m\to 0}\left(1 - \left(\frac{d+1}{d}\right)^{-m}\right) \Big/ \left(1 - 10^{-m}\right) \\
&\quad = \lim_{m\to 0}\left(m \log\left(\frac{d+1}{d}\right) \cdot \left(\frac{d+1}{d}\right)^{-m} \Big/ m \log(10) \cdot 10^{-m}\right) \\
&\quad = \mathrm{Log}_{10}\left(\frac{d+1}{d}\right) \lim_{m\to 0}\left(\frac{d+1}{10d}\right)^{-m} = \mathrm{Log}_{10}\left(1 + d^{-1}\right).
\end{aligned}$$

This is the Benford probability. We find from integrating $f_{10^j,10^{j+1},0}(x)$, a similar result holds if m is negative and small. We therefore conclude that, if $m > 0$ is small and $[a, b] = [10^j, 10^{j+1}]$, then the difference from Benford's Law for observing a first digit of d is

$$d^{-m} \cdot \left(1 - \left(\frac{d+1}{d}\right)^{-m}\right) \Big/ \left(1 - 10^{-m}\right) - \log_{10}\left(\frac{d+1}{d}\right). \tag{7}$$

As remarked earlier, this tends to zero as m tends to 0. Further, note the above quantification of the deviation from Benford's Law is independent of j. Thus writing

$$\left[10^k, 10^n\right] = \left[10^k, 10^{k+1}\right] \cup \left[10^{k+1}, 10^{k+2}\right] \cup \cdots \cup \left[10^{n-1}, 10^n\right],$$

we see (7) also holds for the difference from Benford's Law for the interval $[a, b] = [10^k, 10^n]$, and gives the base 10 digit bias for a power law with positive exponent m covering an integral number of orders of magnitude. Similar integration and algebra yields analogues of (7) for negative m. Results indicate the following assertion.

Theorem 1 *Let T be a random variable with the density given by a power law. The density is $f_{a,b,m}(x) = C(a,b,m)x^{-(m+1)}$ for x in $[a,b]$ and 0 otherwise, where $C(a,b,m)$ is $1/\log(b/a)$ if $m = 0$, $m/(a^{-m} - b^{-m})$ if $m > 0$, and $|m|/(b^{|m|} - a^{|m|})$ if $m < 0$. If $[a,b] = [10^k, 10^n]$, then:*

- *If m does not equal 0, then the probability of the first digit of T (base 10) equaling d in $\{1,\ldots,9\}$ is $d^{-m} \cdot (1 - (d + 1/d)^{-m})/(1 - 10^{-m})$; if m equals 0, then the probability of the first digit of T (base 10) equaling d is $\log_{10}(1 + 1/d)$.*
- *If m does not equal 0, then the probability of the first-two digits of T (base 10) equaling d_1d_2 in $\{10,\ldots,99\}$ is $(d_1d_2/10)^{-m}(1 - ((d_1d_2 + 1)/d_1d_2)^{-m})/(1 - 10^{-m})$; if m equals 0, then the probability of the first-two digits of T equaling d_1d_2 is $\log_{10}(1 + 1/d_1d_2)$.*

References

Benford F (1938) The law of anomalous numbers. Proc Am Philos Soc 78(4):551–572

DeGroot M, Schervish M (2002) Probability and statistics, 3rd edn. Addison-Wesley, Reading

Diaconis P (1976) The distribution of leading digits and uniform distribution mod 1. Ann Probab 5(1):72–81

Drake PD, Nigrini MJ (2000) Computer assisted analytical procedures using Benford's Law. J Account Educ 18(2):127–146

Hill TP (1995) Base-invariance implies Benford's Law. Proc Am Math Soc 123(3):887–895

Kontorovich AV, Miller SJ (2005) Benford's Law, values of L-functions, and the $3x + 1$ problem. Acta Arith 120(3):269–297

Lagarias J, Soundararajan K (2006) Benford's Law for the $3x + 1$ function. J Lond Math Soc 74(2):273–288

Leemis LM, Schmeiser BW, Evans DL (2000) Survival distributions satisfying Benford's Law. Am Stat 54(3):1–6

Lehner B, Döll P (2004) Development and validation of a global database of lakes, reservoirs and wetlands. J Hydrol 296(1–4):1–22

Ley E (1996) On the peculiar distribution of the US Stock Indices first digits. Am Stat 50(4):311–313

Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46(5):323–351

Nigrini MJ (1996) A taxpayer compliance application of Benford's Law. J Am Tax Assoc 18:72–91

Nigrini MJ (2005) An assessment of the change in the incidence of earnings management around the Enron–Andersen episode. Rev Account Financ 4(1):92–110

Nigrini MJ, Mittermaier LJ (1997) The use of Benford's Law as an aid in analytical procedures: Auditing. J Pract Theory 16(2):52–67

Pinkham RS (1961) On the distribution of first significant digits. Ann Math Stat 32(4):1223–1230

Raimi R (1969) The peculiar distribution of first digits. Sci Am 221(6):109–120

Raimi R (1976) The first digit problem. Am Math Mon 83(7):521–538

The Economist (2006) Scientific fraud: Egg on his face, 5 January 2006

Wallace WA (2002) Assessing the quality of data used for benchmarking and decision-making. J Gov Financ Manag 51(3):16–22