Acta Geodaetica et Geophysica (2019) 54:315–331
https://doi.org/10.1007/s40328-019-00259-3

ORIGINAL STUDY

Usability of the Benford's law for the results of least square estimation

Nursu Tunalioglu[1] · Bahattin Erdogan[1]

Received: 14 January 2019 / Accepted: 23 May 2019 / Published online: 28 May 2019

Abstract

Benford's law (BL), also known as the first-digit or significant-digit law, is an intriguing pattern in data sets, considers the frequency of occurrence of the first digits, which are not uniformly distributed as might be expected, conversely follow a specified theoretical distribution. According to BL, the occurrence of first non-zero digit in a numerical data, which is generated or found in nature, depends on a logarithmic distribution. Least square estimation (LSE) method is mostly preferred for the estimation of the unknown parameters from different types of geodetic data. The residuals and the normalized residuals of the LSE method, which follow normal distribution and expected values of them are zero are used in outlier detection problem. In this study, BL is investigated for residuals and the normalized residuals estimated from LSE method. Three types of geodetic data are used: (1) simulated regression models, (2) global positioning system (GPS) data, (3) leveling network. The first group data sets are simulated based on linear regression and univariate models and each simulated group is generated for a number of 100, 1000, and 10,000 samples. To generate second group, an international global navigation satellite system (GNSS) service (IGS) station data (ISTA) is processed by kinematic PPP approach using GIPSY OASIS II v6.4 software. Here, the observation duration of GPS data is 4 days. For the last data, a leveling network with 55 points involving 110 observations of height differences is simulated. BL has been applied to the residuals (**v**) and normalized residuals (**w**) estimated from LSE method. Goodness-of-fit test has been implemented to determine whether a population has a specified BL distribution or not. This test is based on how good a fit we have between the frequency of occurrence of residuals and normalized residuals in an observed sample and the expected frequencies obtained from the hypothesized distribution. The results depending on the statistical test show that each data set (residuals and normalized residuals) used in this study follows BL.

Keywords Benford's law · Frequency of occurrence · Residuals and normalized residuals · Goodness-of-fit test · Least square estimation

✉ Nursu Tunalioglu
ntunali@yildiz.edu.tr

Bahattin Erdogan
berdogan@yildiz.edu.tr

[1] Department of Geomatic Engineering, Civil Engineering Faculty, Yildiz Technical University, Istanbul, Turkey

1 Introduction

Benford's law (BL), also known as the first-digit or significant-digit law, was first found in 1881 by Simon Newcomb. Newcomb asserts that the ten digits do not occur with an equal frequency that must be evident to anyone making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones (Newcomb 1881). BL states that many naturally occurring sets of observations follow a specified theoretical distribution and follow a monotonically decreasing logarithmic distribution, which is not uniformly distributed as might be expected (Berger and Hill 2015). In BL, the first leading (i.e., first non-zero) decimal digit is not equally likely to be any one of the nine possible digits. Instead of that, the occurrence of numbers starting with 1 and 2 is close to 30% and 18%, respectively, whereas the numbers starting with 8 or 9 are close to 5%.

In general, it is hard to say that there is a relation between the huge numerical data sets found in nature or produced by a human (Mir 2012). Benford (1938) tested the law on a wide range field such as the heights of the mountains, the lengths of the rivers, the surface areas of the rivers, physical constants, molecular weights, death rates or cost data, and figured out to follow the BL. In literature, there have been wide range of implementations of BL from social to numerical application fields. Mir (2014) referred to "Benford Online Bibliography (Berger et al. 2009)" for a comprehensive information on the numerous data sets, which obey BL and its applications across multiple disciplines. Moreover, some applications for figuring out the presence of BL in different fields may be given as follows. The early study was conducted by Becker (1982), who found that there was a logarithmic distribution when failure rates and mean time to failure (MTTF) values were read from left to right for the first non-zero digit. Nagasaka (1984) reviewed various sampling procedures by examining for the resulting sampled integers whether BL holds or not. It was proved that randomly sampled integers do not necessarily obey BL but their Banach limit does for polynomial sampling procedures. Hence, it was proved on BL for geometrical sampling procedures and for linear recurrence sampling procedures. Hill (1995) stated that many tables of numerical data do not follow logarithmic distribution such as lists of telephone numbers in a given region typically begin with the same few digits-and even "neutral" data such as square-root tables of integers are not good fits. However, a surprisingly diverse collection of empirical data does seem to obey the significant-digit law. The details of more empirical evidence may also be found in (Hill 1995). In the field of demographic data for world religion distribution, Mir (2012) investigated numerical data on the country-wise adherent distribution of seven major world religions i.e., Christianity, Islam, Buddhism, Hinduism, Sikhism, Judaism and Baha'ism to see if the proportion of the leading digits occurring in the distribution that conforms to BL. Mir (2012) exposed that the adherent data of all the religions, except Christianity, excellently does conform to BL. Ausloos et al. (2015) reviewed the long birth time series for Romania from BL point of view, distinguishing between families with a religious (Orthodox and Non-Orthodox) affiliation and figured out that there is a drastic breakdown of BL on results.

In this study, it is aimed to investigate the applicability of BL in residuals and normalized residuals estimated from Least Square Estimation (LSE) method. Moreover, BL is applied to the geodetic data set, which contain normal distribution to demonstrate the consistency of the law. In geodesy, the data observed at field contains random errors that are from a normal distribution ($N(0, \sigma)$) and the most probable values of the unknown parameters are generally estimated by LSE. To find out the residuals of observed values, differences between estimated value and each observation are taken.

Also, the normalized residuals of them are taken into consideration for the outlier detection problem. Here, three different data sets are studied as; (1) simulated data, (2) real data observed by global positioning system (GPS), (3) Leveling network data pertaining to the random error. According to the results, it is found that the residuals and normalized residuals of these data estimated from LSE follow BL.

2 Benford's law

Like Newcomb before him, Benford (1938) observed that the first pages in logarithmic tables were more referred to than the last pages, which presented an empirical law that shows the distribution of the leading digits are not equal among the naturally occurring phenomena such as area of the river, population of cities, addresses, death rate etc. (Jamain 2001). Although its' basic form introduced by Benford (1938) was based on empirical observations, several mathematical series i.e., binomial coefficients and factorial (Sarkar 1973), Fibonacci and Lucas numbers (Wlodarski 1971) are related with this law. The BL states that the smaller digits are found more frequently than larger digits, which can be expressed as an expected frequency distribution of the first digits for many numerical data sets (Chandra Das et al. 2017). In its most common formulation, the special case of the first significant (i.e., first non-zero) decimal digit, Benford's law asserts that the leading digit is not equally likely to be any one of the nine possible digits 1, 2,…, 9. The first digit of the number is the leftmost digit and differs from 0. According to this law, the frequency of the first significant digit (i.e., nonzero digit, k) can be computed as (Benford 1938):

$$P(k) = \log_{10}(k+1) - \log_{10}(k) = \log_{10}\left(1 + \frac{1}{k}\right) \tag{1}$$

where P is the probability of the number k, and k is the any number in the set $\{1, 2, 3, \ldots, 9\}$.

Considering the BL distribution on the smaller digits for 1 and 2, respectively, if the first leading digit is 1, the probability of occurrence is $P(k = 1) = \log_{10}(2) = 0.3010$, which is more than 30% of the time and if the first leading digit is 2, the probability of the occurrence is $P(k = 2) = \log_{10}\left(\frac{3}{2}\right) = 0.1760$, which is more than 18% of the time. However, if the larger two digits are computed according to the BL, i.e., $P(k = 8) = \log_{10}(9/8) = 0.0512$, which is about 5% of the time and for 9, the probability is $P(k = 9) = \log_{10}(10/9) = 0.0458$, which is less than 5%. Table 1 shows the distributions of first digits as derived from BL.

Table 1 The first digit distribution of Benford's law

First digit (k)	0	1	2	3	4	5	6	7	8	9
Proportion	–	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0458

3 Data

Surveying is an important field of application of geodesy. In a general manner, surveying is the measuring horizontal and vertical distances between objects, angles between lines, determining the direction of lines, and establishing points by predetermined angular and linear measurements. Apart from classical surveys, satellite-based positioning is the determination of positions of observing sites on land or at sea, in the air and in space by means of artificial satellites (Hofmann-Wellenhof et al. 2001). Either classical surveying or satellite-based surveying generates measurements for certain physical quantities. However, in both cases due to some kind of influences, measurements conducted from the same quantity will differ in general. An error is a difference between a measured value and its true value of a quantity. The sources of the errors in measurements can be listed as humans, instruments and circumstances. These errors raise uncertainty in measurements. In order to eliminate the uncertainty, to improve instruments used during the measure, to train the person or to provide better circumstance will help, however, do not provide error-free measurements and all measurements contain errors. Typically, errors can be classified into three groups: systematical errors, random errors and gross errors. Among them, gross errors can be detected by repeated measurements. Systematical errors that generally pertain to survey instruments or external influences (circumstance) occur in the same direction and magnitude on measurements. They can be eliminated by calibration of instruments and implementation of the proper corrections on measurements for circumstances influences (Berber 2008; Teunissen 2003). However, random errors cannot be eliminated in such a procedure explained for the others and follow normal distribution on measurements. At the stage of the LSE, these errors reflect on the residuals.

BL proportions are compared with the proportions of residuals and normalized residuals. The data used in this study has normal distribution and equations given below for LSE are used to compute the residual and normalized residuals. It is supposed that there are n independent equally weighted measurements, denoted as vector **l**. The differences between the most probable value of measurements and measured values are the residuals, denoted as **v**. Accordingly, to compute the normalized residual (**w**), which is the specific form of residuals, Eq. 9 is implemented.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T_{n\times 1} \quad \text{(for univariate model)} \tag{2a}$$

$$\mathbf{A} = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix}^T_{n\times 1} \quad \text{(for linear regression model)} \tag{2b}$$

$$\hat{\boldsymbol{x}} = \left(\mathbf{A^T P A}\right)^{-1}\left(\mathbf{A^T P l}\right) \tag{3}$$

$$\mathbf{v} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{l} \tag{4}$$

$$\mathbf{Q_{xx}} = \left(\mathbf{A^T P A}\right)^{-1} \tag{5}$$

$$\mathbf{Q_{ll}} = \mathbf{A Q_{xx} A^T} \tag{6}$$

$$\mathbf{Q_{vv}} = \mathbf{P}^{-1} - \mathbf{Q_{ll}} \tag{7}$$

$$\sigma_o = \sqrt{\frac{\mathbf{v^T P v}}{n - u}} \tag{8}$$

$$w_i = \frac{|v_i|}{\sigma_o \sqrt{Q_{vivi}}} \tag{9}$$

Here **l** is the observation vector, $\hat{\mathbf{x}}$ is the unknown vector, **A** is the coefficient matrix, **P** is the weight matrix that is identity matrix, **v** is the residual vector, **w** is the normalized residuals, $\hat{\sigma}_0^2$ is the variance of unit weight, n is the number of data and u is the number of unknown parameters.

4 Regression models

BL has been applied to two geodetic survey data sets simulated depending on regression models with normal distribution explained below that whether or not follow a specific BL distribution due to residuals and normalized residuals estimated from LSE. For generating the data based on regression models, the standard deviation of the models is taken 0.02 m with expected value 0. Formann (2010) reviewed some common distributions to explore their relations with BL. One of the experienced distribution examined by Formann (2010) is the normal distribution, $N(\mu, \sigma)$. In Formann (2010), the simulated random data is generated with normal distribution ($\mu = 1.1, \sigma = 0.25$) and results are concluded that normal distribution does not follow BL distribution. However, the first digit of the expected value dominates the results, thus the majority of the data will be involved into the first digit, in which the expected value starts. This case can be seen in Formann (2010). Then, in this study, we chose the expected value as 0 to eliminate the effect of first digit of the expected value on results. Moreover, the expected values of the residuals and normalized residuals estimated from LSE will be 0 that will not affect the results.

4.1 Univariate model

The first model is the univariate model. The random errors e_i are generated by a random generator, and y (the measurement) is obtained with $y = a + e_i$. The random errors follow normal distribution (N(0;0.02 m)). Three different cases have been simulated for different numbers of observations (100, 1000 and 10,000). For the simulation a was chosen 5 m and observation was obtained by adding the random errors to the 5 m. LSE is applied to data to estimate unknown parameter, residuals and normalized residuals.

4.2 Linear regression model

The second simulation model is the linear regression model, that was obtained with $y = a + bx + e_i$. For the simulation, a and b were chosen as 1 m and the random errors **e** follow normal distribution (N(0;0.02 m)). For the analysis, LSE is applied to estimate unknown parameters, residuals and normalized residuals.

4.3 GPS data

To generate the second group, an IGS station data (ISTA) is processed by kinematic Precise Point Positioning (PPP) approach using GIPSY OASIS II v6.4 software. Here, the observation duration of GPS data is 4 days (on 1st, 2nd, 3rd, 4th of January 2016) (Fig. 1). LSE is applied to estimate mean values of coordinates for the 4 days. Also, the residuals and normalized residuals of the three-dimensional coordinates was estimated for the BL analysis.

4.4 Leveling network

The third data group is simulated for a leveling network, which involves 55 points and 110 observations. For the leveling network, the height differences Δh_k are computed from the fixed points. They are free of random errors, and then the random errors are generated from a normal distribution. They are added to the height differences. The precision is taken as $\sigma_h = \sigma_0\sqrt{S}\left(\sigma_0 = 1\frac{\text{mm}}{\sqrt{1\,\text{km}}}\right)$ for the leveling network where S is the total length of the leveling lines in km. The height differences were considered as observations and unknown parameters (heights of the points) were estimated by the LSE method (Koch 1999). The residuals and normalized residuals for height differences were used for BL analysis.

4.5 Data analysis

The Null-Hypothesis (Ho) is written to test for compliance with BL between the observed and expected first digit distributions. According to the Ho hypothesis that the frequencies obtained from observations are the same as expected frequencies basis of BL. If the observed frequencies are close to corresponding expected frequencies, the χ^2 value will be small, indicating a good fit. If the observed frequencies differ considerably from the expected frequencies, the χ^2 value will be large and the fit will be poor. A good fit leads to acceptance of Ho, whereas a poor fit leads to its rejection. To test whether Ho hypothesis is accepted or not, the Chi-square goodness-of-fit statistic is calculated using (Walpole et al. 1998);

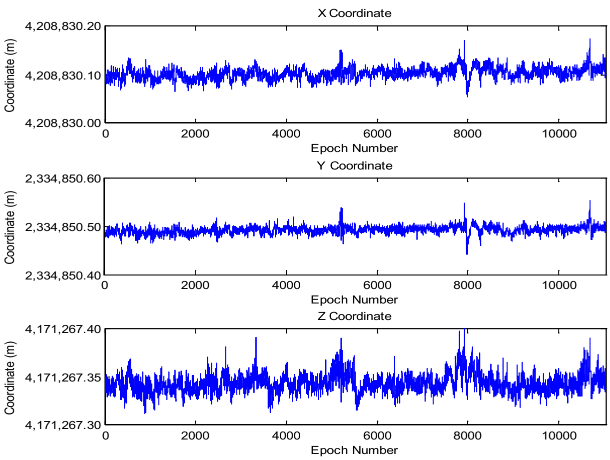

Fig. 1 X, Y, Z coordinates of real data

$$\chi^2 = N \sum_{k=1}^{9} \frac{(P(k) - B(k))^2}{B(k)} \tag{10}$$

where χ^2 is a value of a random variable whose sampling distribution is approximated very closely by the Chi-squared distribution with $v = 9 - 1$ degrees of freedom. N is the sum of the frequencies, and P(k) is the proportion from the data and B(k) is the proportion from the BL. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution (Walpole et al. 1998).

In addition to this, the numerical range of the data set in terms of order of magnitude (OOM) may be used as a strong indicator for compliance with BL (Brown 2005; Kossovsky 2014). Brown (2005) emphasized that the data sets which vary with a large numerical range can have an expectation to show a good correlation with BL. The OOM value equal or bigger than 4 would be suitable. Kossovsky (2014) and Whyman et al. (2016) state that requirement for data configuration with regards to compliance with BL should be sufficient, also Kossovsky (2014) declares that the value of the OOM of the data set should be approximately over 3. The numerical range of the data set used to commit conformity or non-conformity of the BL can be estimated by OOM:

$$\mathrm{OOM} = \log\left(\frac{x_{\max}}{x_{\min}}\right) \tag{11}$$

where x_{min} and x_{max} are the minimum and maximum values of the data set, respectively. Here, the OOM is computed from the absolute values of max and min values of data set excluding zeroes.

5 Results

To test the compliance with BL, goodness-of-fit test were applied to the proportions. As we have 9 first digits, the degrees of freedom is taken $9 - 1 = 8$. Considering the 95% confidence level, the decision value (χ^2) is computed 15.507. If this value is exceeded in any case, the compliance with BL will be rejected, otherwise, it will be accepted.

5.1 Regression models

The distribution of residuals and normalized residuals calculated from the univariate model for the number of 100, 1000 and 10,000 data are shown in Fig. 2 and the statistical considerations are given in Tables 2 and 3. For each data group simulated by the univariate model, v and w proportions approximate BL proportions. According to the first digits from 1 to 9, BL proposes the occurrence of first digits from 30.10 to 4.58%, respectively. When we consider the data for 100 measurements, the proportions of v and w are obtained from 37 to 3% and from 34 to 7%, respectively. The similar results are also obtained for 1.000 and 10.000 measurements (See, column 5 and 7 of Tables 2, 3).

In the same manner, when we compare the residuals and normalized residuals' first digit occurrence estimated by linear regression model and BL (Fig. 3), it can be seen that the graphical representations resemble the expected BL proportions. Tables 4 and 5

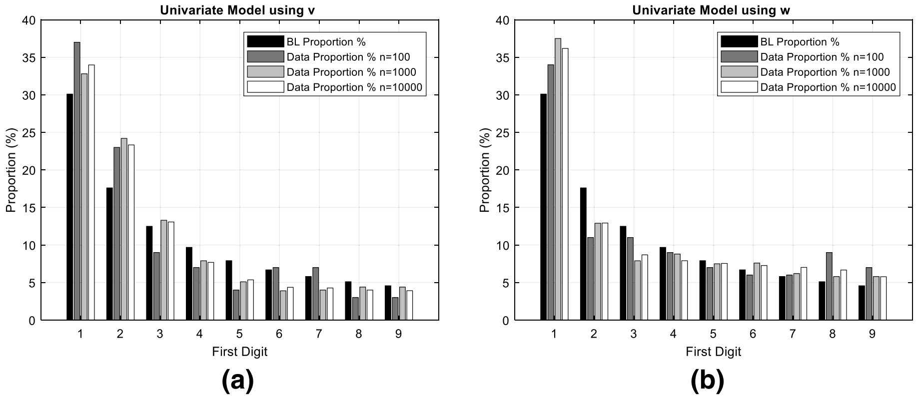

Fig. 2 Comparisons of residuals (**a**) and normalized residuals (**b**) first digit frequency estimated by univariate model and expected frequencies according to BL

Table 2 Frequencies and proportions for univariate model using residuals

First digit k	Frequency f	Proportion	Frequency f	Proportion	Frequency f	Proportion	BL proportion
1	37	0.3700	328	0.3280	3398	0.3398	0.3010
2	23	0.2300	242	0.2420	2333	0.2333	0.1761
3	9	0.0900	133	0.1330	1307	0.1307	0.1249
4	7	0.0700	79	0.0790	769	0.0769	0.0969
5	4	0.0400	51	0.0510	536	0.0536	0.0792
6	7	0.0700	39	0.0390	437	0.0437	0.0669
7	7	0.0700	40	0.0400	428	0.0428	0.0580
8	3	0.0300	44	0.0440	400	0.0400	0.0512
9	3	0.0300	44	0.0440	392	0.0392	0.0458
Sum	100	1.0000	1000	1.0000	10,000	1.0000	1.0000
χ^2	8.5736		5.9261		5.1665		15.507

Table 3 Frequencies and proportions for univariate model using normalized residuals

First digit k	Frequency f	Proportion	Frequency f	Proportion	Frequency f	Proportion	BL proportion
1	34	0.3400	375	0.3750	3618	0.3618	0.3010
2	11	0.1100	129	0.1290	1293	0.1293	0.1761
3	11	0.1100	79	0.0790	869	0.0869	0.1249
4	9	0.0900	88	0.0880	791	0.0791	0.0969
5	7	0.0700	75	0.0750	755	0.0755	0.0792
6	6	0.0600	76	0.0760	727	0.0727	0.0669
7	6	0.0600	62	0.0620	702	0.0702	0.0580
8	9	0.0900	58	0.0580	667	0.0667	0.0512
9	7	0.0700	58	0.0580	578	0.0578	0.0458
Sum	100	1.0000	1000	1.0000	10,000	1.0000	1.0000
χ^2	7.6331		5.4394		5.0686		15.507

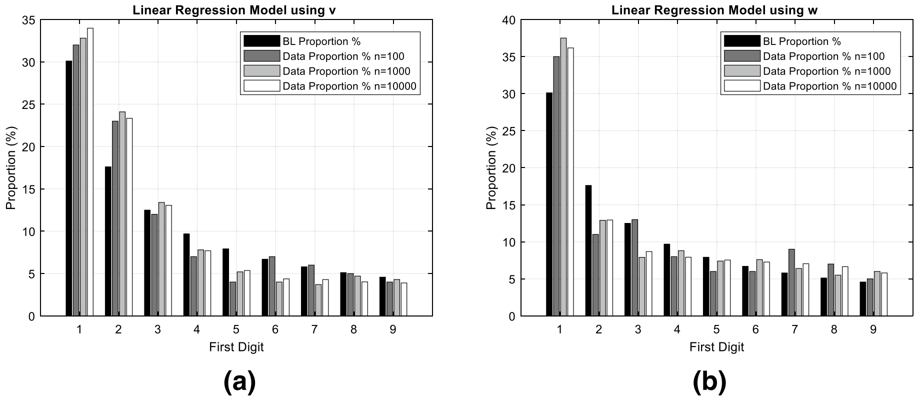

Fig. 3 Comparisons of residuals (**a**) and normalized residuals (**b**) first digit frequency estimated by linear regression model and expected frequencies according to BL

Table 4 Frequencies and proportions for linear regression model using residuals

First digit k	Frequency f	Proportion	Frequency f	Proportion	Frequency f	Proportion	BL proportion
1	32	0.3200	328	0.3280	3398	0.3398	0.3010
2	23	0.2300	241	0.2410	2334	0.2334	0.1761
3	12	0.1200	134	0.1340	1306	0.1306	0.1249
4	7	0.0700	78	0.0780	769	0.0769	0.0969
5	4	0.0400	52	0.0520	536	0.0536	0.0792
6	7	0.0700	40	0.0400	437	0.0437	0.0669
7	6	0.0600	37	0.0370	429	0.0429	0.0580
8	5	0.0500	47	0.0470	402	0.0402	0.0512
9	4	0.0400	43	0.0430	389	0.0389	0.0458
Sum	100	1.0000	1000	1.0000	10,000	1.0000	1.0000
χ^2	4.5714		5.8968		5.1670		15.507

Table 5 Frequencies and proportions for linear regression model using normalized residuals

First digit k	Frequency f	Proportion	Frequency f	Proportion	Frequency f	Proportion	BL proportion
1	35	0.3500	375	0.3750	3617	0.3617	0.3010
2	11	0.1100	129	0.1290	1294	0.1294	0.1761
3	13	0.1300	79	0.0790	868	0.0868	0.1249
4	8	0.0800	88	0.0880	792	0.0792	0.0969
5	6	0.0600	74	0.0740	754	0.0754	0.0792
6	6	0.0600	76	0.0760	727	0.0727	0.0669
7	9	0.0900	64	0.0640	704	0.0704	0.0580
8	7	0.0700	55	0.0550	664	0.0664	0.0512
9	5	0.0500	60	0.0600	580	0.0580	0.0458
Sum:	100	1.0000	1000	1.0000	10,000	1.0000	1.0000
χ^2	6.6299		5.5388		5.0637		15.507

show the proportions of the observation number 100, 1000 and 10,000 in columns 3, 5 and 7, which approximate the BL proportions.

Since the computed χ^2 values are less than $\chi^2_{0.05} = 15.507$ for 1 degrees of freedom, we can conclude that if the LSE method is applied to the data used in this study as specified geodetic data that have normal distribution with expected value 0 and standard deviation of 0.02 provide a good fit for the distribution of BL in terms of computed residuals and normalized residuals that their expected values are 0.

Moreover, we changed the standard deviations for simulated data to explore its effect on the results. To simulate the data, the standard deviation is altered from 0.001 to 1 with taking the expected value 0. Tables 6 and 7 represent the expected and observed proportions provided by the univariate model with calculated χ^2, respectively. Since the calculated χ^2 values (the last column of Tables 6, 7) are compared with the decision value ($\chi^2_{0.05} = 15.507$), it is seen that the calculated χ^2 values are smaller than the decision value in all cases. Changing the standard deviations does not bring any meaningful results on the outcomes. The similar results are also obtained for normalized residuals (see, Tables 8, 9). However, in all cases of changing the standard deviation, it may be

Table 6 Frequencies of univariate model for residuals

Univariate model	First digit frequencies									
	1	2	3	4	5	6	7	8	9	χ^2
Residual										
Normal distribution (μ = 0, σ = 0.001)										
n = 100	32	11	14	6	8	8	8	7	6	6.4158
n = 1000	375	130	79	88	72	77	64	57	58	5.4678
n = 10,000	3640	1305	865	792	742	731	689	673	563	5.0496
Normal distribution (μ = 0, σ = 0.01)										
n = 100	32	11	14	6	8	8	8	7	6	6.4158
n = 1000	375	130	79	88	72	77	64	57	58	5.4678
n = 10,000	3640	1305	865	792	742	731	689	673	563	5.0496
Normal distribution (μ = 0, σ = 0.05)										
n = 100	25	14	16	13	9	10	7	4	2	7.4400
n = 1000	209	160	141	115	106	95	77	61	36	6.6093
n = 10,000	2170	1534	1420	1236	1091	869	694	568	418	5.6523
Normal distribution (μ = 0, σ = 0.25)										
n = 100	30	22	17	6	6	5	4	3	7	7.7371
n = 1000	301	221	172	97	66	49	33	27	34	6.1381
n = 10,000	2954	2327	1563	986	648	454	382	338	348	5.1017
Normal distribution (μ = 0, σ = 0.50)										
n = 100	25	14	16	13	9	10	7	4	2	7.4400
n = 1000	209	160	141	115	106	95	77	61	36	6.6093
n = 10,000	2170	1534	1420	1236	1091	869	694	568	418	5.6523
Normal distribution (μ = 0, σ = 1.00)										
n = 100	32	11	14	6	8	8	8	7	6	6.4158
n = 1000	375	130	79	88	72	77	64	57	58	5.4678
n = 10,000	3640	1305	865	792	742	731	689	673	563	5.0496

Table 7 Frequencies of univariate model for normalized residuals

Univariate model	First digit frequencies									
	1	2	3	4	5	6	7	8	9	χ^2
Normalized residual										
Normal distribution (μ = 0, σ = 0.001)										
n = 100	34	11	11	9	7	6	6	9	7	7.6331
n = 1000	375	129	79	88	75	76	62	58	58	5.4394
n = 10,000	3618	1293	869	791	755	727	702	667	578	5.0686
Normal distribution (μ = 0, σ = 0.01)										
n = 100	34	11	11	9	7	6	6	9	7	7.6331
n = 1000	375	129	79	88	75	76	62	58	58	5.4394
n = 10,000	3618	1293	869	791	755	727	702	667	578	5.0686
Normal distribution (μ = 0, σ = 0.05)										
n = 100	34	11	11	9	7	6	6	9	7	7.6331
n = 1000	375	129	79	88	75	76	62	58	58	5.4394
n = 10,000	3618	1293	869	791	755	727	702	667	578	5.0686
Normal distribution (μ = 0, σ = 0.25)										
n = 100	34	11	11	9	7	6	6	9	7	7.6331
n = 1000	375	129	79	88	75	76	62	58	58	5.4394
n = 10,000	3618	1293	869	791	755	727	702	667	578	5.0686
Normal distribution (μ = 0, σ = 0.50)										
n = 100	34	11	11	9	7	6	6	9	7	7.6331
n = 1000	375	129	79	88	75	76	62	58	58	5.4394
n = 10,000	3618	1293	869	791	755	727	702	667	578	5.0686
Normal distribution (μ = 0, σ = 1.00)										
n = 100	34	11	11	9	7	6	6	9	7	7.6331
n = 1000	375	129	79	88	75	76	62	58	58	5.4394
n = 10,000	3618	1293	869	791	755	727	702	667	578	5.0686

concluded that the larger data size gives the lower χ^2 values for v and w values of both models simulated.

5.2 GPS data

Apart from simulated data, the residuals and normalized residuals of GPS data were used as the real data set. The epoch number of GPS measurements is 11.036. LSE is applied to estimate the mean values of the coordinates for 4 days. Also, the residuals and normalized residuals are separately estimated for three coordinate components (denoted in tables and figures as X, Y, Z). The graphical representations of comparisons of residuals and normalized residuals for X, Y, Z coordinates with BL proportions are shown in Fig. 4 and the statistical considerations are given in Tables 10 and 11. The proportions of v and w for X coordinate component are obtained from 36.29 to 5.62% and from 33.65 to 6.14%, respectively. The similar results are also obtained for Y and Z coordinates (see, columns 5 and 7 of Tables 10, 11).

Table 8 Frequencies of linear regression model for residuals

Linear regression model	First digit frequencies									
	1	2	3	4	5	6	7	8	9	χ^2
Residual										
Normal distribution (μ = 0, σ = 0.001)										
n = 100	35	11	13	9	6	7	8	9	2	9.0609
n = 1000	375	130	77	90	74	75	65	54	60	5.5873
n = 10,000	3640	1305	866	791	743	730	690	672	563	5.0415
Normal distribution (μ = 0, σ = 0.01)										
n = 100	35	11	13	9	6	7	8	9	2	9.0609
n = 1000	375	130	77	90	74	75	65	54	60	5.5873
n = 10,000	3640	1305	866	791	743	730	690	672	563	5.0415
Normal distribution (μ = 0, σ = 0.05)										
n = 100	24	15	15	11	6	12	9	5	3	9.2843
n = 1000	207	164	140	114	106	94	78	61	36	6.5926
n = 10,000	2171	1534	1420	1235	1092	867	696	568	417	5.6466
Normal distribution (μ = 0, σ = 0.25)										
n = 100	30	17	21	8	6	5	4	5	4	7.6345
n = 1000	304	220	172	97	66	49	34	28	30	6.1542
n = 10,000	2954	2327	1563	985	647	456	383	337	348	5.0922
Normal distribution (μ = 0, σ = 0.50)										
n = 100	24	15	15	11	6	12	9	5	3	9.2843
n = 1000	207	164	140	114	106	94	78	61	36	6.5926
n = 10,000	2171	1534	1420	1235	1092	867	696	568	417	5.6466
Normal distribution (μ = 0, σ = 1.00)										
n = 100	35	11	13	9	6	7	8	9	2	9.0609
n = 1000	375	130	77	90	74	75	65	54	60	5.5873
n = 10,000	3640	1305	866	791	743	730	690	672	563	5.0415

5.3 Leveling network

As a simulated data, we generate leveling network contains 55 points and LSE is applied to height differences of 110 observations. The residuals and normalized residuals are obtained using LSE. Here, the height differences were considered as observations and the coefficient matrix was formed due to the network configuration. Figure 5 represents the graphical forms of BL proportions for residuals and normalized residuals, respectively. The statistical results are given in Table 12. According to the results, the proportions of v and w are obtained from 33.64 to 4.55% and from 30 to 6.76%, respectively. These findings fit the BL proportions.

According to the results given in Table 13, the χ^2 values of each data set were computed and it is seen that the decision value is greater than them in all cases, which figure out the compliance with BL. Moreover, the larger data size gives the lower χ^2 values for v and w values of both models simulated and varying standard deviation on generating the random data does not change the distribution significantly. When we compare the decision value of χ^2 of the real data (see Table 13), the χ^2 values of the data are under the decision value

Table 9 Frequencies of linear regression model for normalized residuals

Linear regression model	First digit frequencies									
	1	2	3	4	5	6	7	8	9	χ^2
Normalized residual										
Normal distribution (μ = 0, σ = 0.001)										
n = 100	35	11	13	8	6	6	9	7	5	6.6299
n = 1000	375	129	79	88	74	76	64	55	60	5.5388
n = 10,000	3617	1294	868	792	754	727	704	664	580	5.0637
Normal distribution (μ = 0, σ = 0.01)										
n = 100	35	11	13	8	6	6	9	7	5	6.6299
n = 1000	375	129	79	88	74	76	64	55	60	5.5388
n = 10,000	3617	1294	868	792	754	727	704	664	580	5.0637
Normal distribution (μ = 0, σ = 0.05)										
n = 100	35	11	13	8	6	6	9	7	5	6.6299
n = 1000	375	129	79	88	74	76	64	55	60	5.5388
n = 10,000	3617	1294	868	792	754	727	704	664	580	5.0637
Normal distribution (μ = 0, σ = 0.25)										
n = 100	35	11	13	8	6	6	9	7	5	6.6299
n = 1000	375	129	79	88	74	76	64	55	60	5.5388
n = 10,000	3617	1294	868	792	754	727	704	664	580	5.0637
Normal distribution (μ = 0, σ = 0.50)										
n = 100	35	11	13	8	6	6	9	7	5	6.6299
n = 1000	375	129	79	88	74	76	64	55	60	5.5388
n = 10,000	3617	1294	868	792	754	727	704	664	580	5.0637
Normal distribution (μ = 0, σ = 1.00)										
n = 100	35	11	13	8	6	6	9	7	5	6.6299
n = 1000	375	129	79	88	74	76	64	55	60	5.5388
n = 10,000	3617	1294	868	792	754	727	704	664	580	5.0637

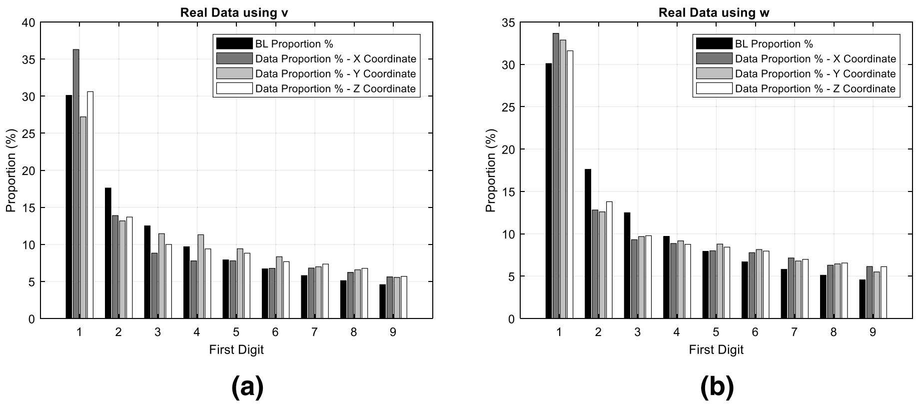

Fig. 4 Comparisons of residuals (**a**) and normalized residuals (**b**) first digit frequency computed for X, Y, Z and expected frequencies according to BL

Table 10 Frequencies and proportions for real data using residuals (X, Y, Z)

First digit k	X		Y		Z		BL proportion
	Frequency f	Proportion	Frequency f	Proportion	Frequency f	Proportion	
1	4005	0.3629	3003	0.2721	3378	0.3061	0.3010
2	1533	0.1389	1454	0.1318	1511	0.1369	0.1761
3	974	0.0883	1263	0.1144	1103	0.0999	0.1249
4	858	0.0777	1248	0.1131	1037	0.0940	0.0969
5	860	0.0779	1039	0.0941	973	0.0882	0.0792
6	748	0.0678	920	0.0834	847	0.0767	0.0669
7	752	0.0681	771	0.0699	811	0.0735	0.0580
8	686	0.0622	726	0.0658	748	0.0678	0.0512
9	620	0.0562	612	0.0555	628	0.0569	0.0458
Sum	11,036	1.0000	11,036	1.0000	11,036	1.0000	1.0000
χ^2	4.1677		3.3049		2.8604		15.507

Table 11 Frequencies and proportions for real data using normalized residuals (X, Y, Z)

First digit k	X		Y		Z		BL proportion
	Frequency f	Proportion	Frequency f	Proportion	Frequency f	Proportion	
1	3714	0.3365	3628	0.3287	3488	0.3161	0.3010
2	1414	0.1281	1390	0.1260	1522	0.1379	0.1761
3	1027	0.0931	1068	0.0968	1079	0.0978	0.1249
4	978	0.0886	1012	0.0917	967	0.0876	0.0969
5	883	0.0800	970	0.0879	930	0.0843	0.0792
6	858	0.0777	899	0.0815	879	0.0796	0.0669
7	789	0.0715	750	0.0680	771	0.0699	0.0580
8	695	0.0630	712	0.0645	724	0.0656	0.0512
9	678	0.0614	607	0.0550	676	0.0613	0.0458
Sum	11,036	1.0000	11,036	1.0000	11,036	1.0000	1.0000
χ^2	3.9095		3.4634		3.0323		15.507

(15.507) in all cases that is satisfied the compliance with BL. The χ^2 values computed for real data are lower than the simulated models and the larger data size provides lower χ^2.

Table 14 shows the OOM values of the data sets computed from the absolute values of residuals and normalized residuals by Eq. (11). For all data sets, the OOM values are larger than 3 (or 4). For the simulated data set depending on the linear regression model, the OOM values increase, while the numerical range of the data sets increase. Although there is a slightly decrease between the number of data 100–1000, the OOM values are still over 3 (or 4).

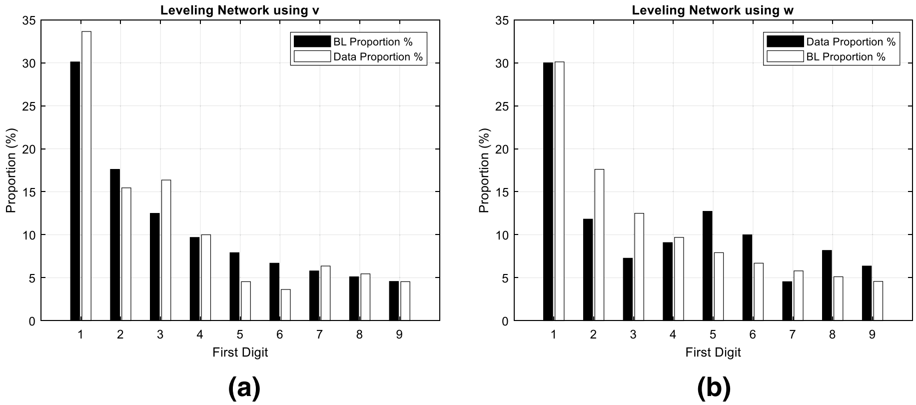

Fig. 5 Comparisons of residuals (**a**) and normalized residuals (**b**) first digit frequency estimated by leveling network and expected frequencies according to BL

Table 12 Frequencies and proportions for leveling network

First digit k	v		w		BL proportion
	Frequency f	Proportion	Frequency f	Proportion	
1	37	0.3364	33	0.3000	0.3010
2	17	0.1545	13	0.1182	0.1761
3	18	0.1636	8	0.0727	0.1249
4	11	0.1000	10	0.0909	0.0969
5	5	0.0455	14	0.1273	0.0792
6	4	0.0364	11	0.1000	0.0669
7	7	0.0636	5	0.0455	0.0580
8	6	0.0545	9	0.0818	0.0512
9	5	0.0455	7	0.0636	0.0458
Sum	110	1.0000	110	1.0000	1.0000
χ^2	4.7981		11.4846		15.507

6 Conclusion

In this study, we have shown, for the first time that the residuals and normalized residuals of geodetic data set estimated from LSE follow BL. The first data set was simulated according to the regression models and randomly distributed errors were added. Without giving outliers, the data, which we assume that it should involve these types of errors that come from the nature of the survey, conform to this law with high statistical accuracy. To prove it with real data applications, we have chosen GPS data and leveling network. In the same manner, the statistical results support the same outcomes with higher accuracy. One point common in the data set is that the larger number of data provides the lower χ^2. This situation is also provided by the OOM values. The numerical

Table 13 χ^2 values of data sets for residuals and normalized residuals

	χ^2 n = 100	χ^2 n = 1000	χ^2 n = 10,000	χ^2 (95%)
Regression models				
Univariate model				
v	8.5736	5.9261	5.1665	15.507
w	7.6331	5.4394	5.0686	15.507
Linear regression model				
v	4.5714	5.8968	5.1670	15.507
w	6.6299	5.5388	5.0637	15.507
	X	Y	Z	
GPS data				
v	4.1677	3.3049	2.8604	15.507
w	3.9095	3.4634	3.0323	15.507
Leveling network				
v		4.7981		15.507
w		11.4846		15.507

Table 14 OOM values of data sets for residuals and normalized residuals

	n = 100	n = 1000	n = 10,000
Regression models			
Univariate model			
v	7.7	6.8	9.4
w	7.7	6.8	9.4
Linear regression model			
v	5.0	7.0	9.4
w	5.0	7.0	9.4
	X	Y	Z
GPS data			
v	11.3	10.6	10.9
w	11.3	10.6	10.9
Leveling network			
v		4.9	
w		4.4	

range of the data sets increase when the number of data sets are increased. Meanwhile, the OOM values computed are over 3 (or 4) that they can be suggested as showing a good correlation with BL. BL applied as a useful tool for fraud detection in financial fields. Accordingly, this paper presents the first outcomes of BL implementation in residuals and normalized residuals estimated from the LSE method in geodetic studies.

References

Ausloos M, Herteliu C, Ileanu B (2015) Breakdown of Benford's law for birth data. Physica A 419:736–745. https://doi.org/10.1016/j.physa.2014.10.041

Becker PW (1982) Patterns in listings of failure-rate & MTTF values and listings of other data. IEEE Trans Reliab 31:132–134

Benford F (1938) The law of anomalous numbers. Proc Am Philos Soc 78(4):551–573

Berber M (2008) Error analysis of geodetic networks. Shaker Publishing, Maastricht

Berger A, Hill TP (2015) An introduction to Benford's law. Princeton University Press, Princeton

Berger A, Hill TP, Rogers E (2009) Benford online bibliography. http://www.benfordonline.net. Last accessed 25 Mar 2019

Brown RJC (2005) Benford's Law and the screening of analytical data: the case of pollutant concentrations in ambient air. Analyst 130:1280–1285. https://doi.org/10.1039/b504462f

Chandra Das R, Sekhar Mishra C, Rajib P (2017) Detection of anomalies in accounting data using Benford's law: evidence from India. J Soc Sci Stud 4(1):123–139

Formann AK (2010) The Newcomb–Benford law in its relation to some common distributions. PLoS ONE 5(5):e10541. https://doi.org/10.1371/journal.pone.0010541

Hill TP (1995) A statistical derivation of the significant-digit law. Stat Sci 10(4):354–363

Hofmann-Wellenhof B, Lichtenegger H, Collins J (2001) Global positioning system: theory and practice, 5th edn. Springer, New York

Jamain A (2001) Benford's law. Master's thesis, Department of Mathematics, Imperial College of London and ENSIMAG, London, UK

Koch KR (1999) Parameter estimation and hypothesis testing in linear models, 2nd edn. Springer, Berlin

Kossovsky AE (2014) Benford's law theory, the general law of relative quantities and forensic fraud detection applications. World Scientific Publishing Co.Pte.Ltd, Singapore

Mir TA (2012) The law of the leading digits and the world religions. Physica A 391:792–798. https://doi.org/10.1016/j.physa.2011.09.001

Mir TA (2014) The Benford law behavior of the religious activity data. Physica A 408:1–9. https://doi.org/10.1016/j.physa.2014.03.074

Nagasaka K (1984) On Benford's law. Ann Inst Stat Math 36(Part A):337–352

Newcomb S (1881) Note on the frequency of use of the different digits in natural numbers. Am J Math 4(1):39–40

Sarkar BP (1973) An observation on the significant digits of binomial coefficients and factorials. Sankhyā Indian J Stat Ser B (1960–2002) 35(3):363–364

Teunissen PJG (2003) Adjustment theory an introduction. Delft University Press, Delft

Walpole RE, Myers RH, Myers SL (1998) Probability and statistics for engineers and scientists, 6th edn. Prentice Hall, Upper Saddle River

Whyman G, Ohtori N, Shulzinger E, Bormashenko E (2016) Revisiting the Benford law: when the Benford-like distribution of leading digits in sets of numerical data is expectable? Physica A 461:595–601

Wlodarski J (1971) Fibonacci and Lucas numbers tend to obey Benford's law. Fibonacci Q 9:87–88