Evaluation and correction of electricity consumption statistics based on Benford-Zipf

Yongjiang Yang *
College of Electrical Engineering and Automation, Fuzhou University, Fujian, China

*Corresponding author e-mail: N190127078@fzu.edu.cn

Abstract. In the power system, the power consumption data are of guiding significance for the formulation of production deployment plans. However, due to various reasons, the record of electricity consumption data is often unable to be comprehensive and accurate. To solve this problem, this paper presents a method based on Benford-Zipf. Firstly, Benford’s law is used to evaluate the reliability of recorded electricity data. Then Zipf’s law is used for the data with low reliability. Finally estimate the total electricity consumption of the whole industry with the electricity data of individual enterprises. Taking the power consumption data of metal products industry in Fujian Province from 2015 to 2017 as an example, the evaluation and correction results show that the method is practical and effective.

1. Introduction
As an important basis for power load forecasting, electricity consumption data plays a very important role in power generation planning and power distribution. Due to the characteristics of non-storage and simultaneous production and consumption of electric energy, the relevant data of electric power market has high authenticity. But with the continuous improvement of China's economic situation, in order to echo the economic data, there may be human intervention [1]. At the same time, in the process of recording user power consumption, due to various human or unexpected reasons will also lead to data missing and other anomalies.

Benford's law was discovered by Simon Newcomb[2] in 1881, and then proposed again by Frank Benford [3] in 1938, which has attracted the attention of the academic community and is now widely used in data quality assessment in various fields, such as census[4], financial[5,6], fishery and livestock [7], etc. Numerous practices have shown that it can evaluate the quality of data sets quickly and effectively. In 2016, Wang used Benford’s law to test the quality of power data, and also achieved good results [1]. It should be noted that the Benford’s law has requirements for the number of data in the studied data set. For the analysis of the first digit, the amount of data can be less than 300. To analyze the first two digits, at least 1000 data are required [8].

Zipf’s law was first applied to the word frequency analysis in the literature. With the continuous exploration of scholars, it was applied to other fields. The first application of Zipf’s law to the coal mine exploration industry was in the literature [9], and since then, this law has been widely used in the coal mining industry[10,11,12], While [13] used the method in the analysis of debris flow disasters, which demonstrating the wide applicability of Zipf's law.
In order to find out the data that may have problems and predict the real electricity consumption, this paper proposes a method based on Benford-Zipf. The method identifies problematic data in time by the distribution of the first digit of the data itself. On this basis, the Zipf’s law is used to predict the whole real data only with a small amount of data, which solves the influence of the problem data on the whole to a certain extent, and has the advantages of high efficiency and convenience. This paper takes the power consumption data of the metal products industry in Fujian Province from 2015 to 2017 as an example for empirical analysis. The results show that the method is effective and can provide more reliable electricity consumption information for the power sector.

2. Benford’s law

Benford's law is also known as the “law of first digit”. In 1938, Frank Benford obtained the mathematical expression of the law through statistical analysis of a large number of data [3, 14]. This law points out that the statistical probability of the first nonzero valid integer in many data sets in nature satisfies the logarithmic distribution, and its probability is expressed as:

\[ P(d_1) = \lg \left( 1 + \frac{1}{d_1} \right), \quad d_1 = 1, 2, 3 \ldots 9 \]  

(1)

Where \( P(d_1) \) denotes the probability of \( d_1 \) appearing as the first digit, Table 1 is the Benford distribution obtained by expression:

| First digit | Appearance probability |
|-------------|------------------------|
| 1           | 0.301                  |
| 2           | 0.176                  |
| 3           | 0.125                  |
| 4           | 0.097                  |
| 5           | 0.079                  |
| 6           | 0.067                  |
| 7           | 0.058                  |
| 8           | 0.051                  |
| 9           | 0.046                  |

As the economy develops, in some cases, electricity consumption data is artificially adjusted to respond to economic changes, making electricity consumption data diverge from GDP, industrial value added and other data and reducing the credibility of electricity consumption data greatly [1]. At the same time, the lack of electricity data caused by various reasons such as abnormal electricity consumption behavior of users has led to the decline in the availability of electricity data.

In order to evaluate the quality of collected electricity data, this paper introduces Benford's rule to evaluate its credibility by the first digit distribution of data. Pearson correlation coefficient is used to characterize the deviation of digital distribution from Benford’s law in real data:

\[ r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \]  

(2)

The grading criteria are shown in Table 2 [15].

| Normal | To be observed | Suspicious |
|--------|----------------|------------|
| 0.99~1.0 | 0.97~0.99 | 0~0.97   |

In the electricity consumption data of Fujian metal industry customers recorded in 2015, the first digit distribution is shown in Figure 1. The correlation coefficient between the data set and the Benford distribution is 0.8403, which concludes that there is a problem with the accuracy of the data according to the determination criteria. By comparing the recorded power consumption of the metal industry with the full-caliber power consumption records of Fujian Province in 2015, it can be seen that there is a large gap, which shows that Benford’s law can find abnormal data.
3. Zipf's Law

Zipf's law was proposed by American scholar G.K. Zipf in 1948. The law states that if the data D1-Dn are arranged in order from largest to smallest, and each data rank number (from 1 to n) is given at the same time. These rank numbers are called rank, then the product of data size and rank is approximately satisfied formula (3) [16]:

\[ D_1 \times 1 = D_2 \times 2 = D_3 \times 3 = \ldots = D_n \times n = K \] (3)

Now Zipf's law is widely used in linguistics, geography, economics and information science. According to the relationship, if the data and their rank are regarded as variables, the two conform to an inverse proportional relationship. Figure 2 shows the results of sorting the power consumption of enterprises in a certain industry in Fujian Province from 2015 to 2018 according to the size relationship. It can be seen that this relationship is approximately consistent, indicating that Zipf’s law can be applied to the data analysis of the power industry.

At present, in the marketing department of the power grid, a large number of data do not match with the real data. This problem can be well solved by the Zipf’s law. For the data with defects, the electricity consumption of the whole industry can be predicted based on the reliable electricity consumption data of a few enterprises.

The application of Zipf's Law generally requires the following four steps:

1. Arrange the selected data from large to small;
2. The data except the maximum value is multiplied by the natural number and divided by the maximum value of the data, so that it is close to the set Zipf rank (the Zipf rank of the maximum value
is generally tried in turn at the beginning of 1). The results are denoted as $a_{ij}$, $i$ is the Zipf rank of the maximum value, $j$ is the $j$ largest number in the selected data, and a list is established based on this.

(3) For the value of the $i$th series, calculate its standard deviation and select the series with the smallest standard deviation, whose labelled rank is used as the final Zipf rank of each data.

(4) Based on the obtained Zipf rank, combined with its corresponding data, the total data volume can be derived.

However, the following problems exist in practical application:

(1) When the predicted total amount is finally determined, equation (4) is used for calculation, and the size of $N$ affects the final result, so how to determine the value of $N$?

$$S = F_{\text{max}} \times R_{\text{max}} \times \sum_{i=1}^{n} \frac{1}{i}$$

Where $F_{\text{max}}$ represents the selected maximum, $R_{\text{max}}$ represents the Zipf rank of the selected maximum.

(2) The small value produced by accident may be added to the calculation, which will make the $N$ value very large, resulting in high prediction results.

(3) For larger values, there may be two or more close values, which will greatly affect the results. The specific reason is that in the Zipf's law, there is a large gap between the higher rank data (the rank level refers to the rank of $N = 1$ is higher than that of $N = 2$). With the decrease of rank, the smaller number will appear approximate situation, and the emergence of two or more larger values will greatly reduce their own rank, so that the prediction results are too high.

(4) When selecting data, random selection in sample data may lead to a large fluctuation of the minimum selected value, which makes the final results quite different.

(5) Due to the randomness of the selection, the selected data does not necessarily strictly follow the Zipf distribution, which affects the results.

To solve these problems, this paper adopts the following methods:

(1) Select $N$ on the condition that the predicted minimum is slightly less than the selected minimum[11].

(2) To avoid selecting a smaller value so that the predicted value is too high, the selected data should be satisfied that the minimum value of the selected data is at least 1% of the maximum value. For the value of the $i$th series, calculate its standard deviation and select the series with the smallest standard deviation, whose labelled rank is used as the final Zipf rank of each data.

(3) If $n$ data are selected to participate in the calculation, the data set is divided into $n$ intervals in order, and a data is randomly selected in each interval.

(4) Fixed the minimum data satisfying $\min/\max \geq 1\%$ into the selected group to make the result relatively stable.

(5) Multiple selections are made and the optimal Zipf rank is finally chosen according to the conditions set.

4. Benford-Zipf electricity consumption evaluation and correction algorithm

First, Benford’s law is used to evaluate the quality of data sets:

(1) Record the distribution of the first digits in the dataset.

(2) Calculation of Pearson correlation coefficient with Benford distribution.

(3) When the obtained correlation coefficient falls into the “normal” interval, it is considered that there is no artificial defect in the dataset; When the correlation coefficient falls into the range of “to be observed” and “suspicious”, the next calculation and analysis of power consumption data are carried out. Based on the obtained Zipf rank, combined with its corresponding data, the total data volume can be derived.

After finding the data with problems through Benford’s law, Zipf’s law is used to estimate the total amount. The specific calculation process is optimized as follows:
(1) For the users of the record, select some of the data to satisfy min/max \( \geq 1\% \), choose the smallest value that satisfies the requirement as the smallest selected value, and try to avoid selecting larger data that are similar to each other, then sorting the selected data.

(2) The data except the selected maximum are multiplied by the natural number divided by the maximum to make it close to the set Zipf rank.

(3) Calculating the standard deviation of each dataset, then the set with the smallest standard deviation, whose corresponding Zipf rank is used as the calculated rank, is then selected.

(4) Because the selected data set does not necessarily strictly comply with the Zipf distribution, so take multiple selection calculations (this paper carried out 100000 calculations in the example), and set the correlation coefficient threshold of the data set and the Zipf rank to remove the low correlation data set. Then we use the formula (5) to get the selected trend value, and use the Zipf rank which has the largest selected trend value to estimate.

\[
ST_i = \frac{n_i^2}{100000} \sum \delta
\]  

Where \( ST_i \) is the selection trend value for calculating rank \( i \), \( n_i \) is the number of cases for calculating rank \( i \), and \( \sum \delta \) is the sum of standard deviation for calculating rank \( i \).

(5) The final prediction value is obtained by formula (4). If the prediction value is smaller than the total amount of existing data, the step (4) is returned, and the second largest Zipf is selected to calculate the rank.

The algorithm flow is as follows:

**Figure 3.** Power consumption data evaluation and correction process based on Benford-Zipf method

5. **Case analysis**

In order to test the effectiveness of the algorithm, the data of customer electricity consumption records of the metal products industry in Fujian Province from 2015 to 2017 are selected as the object of assessment and correction. Firstly, Benford’s law is used to analyze its credibility. Then Zipf’s law is
used to estimate the real total amount of low-credibility data. Finally the accuracy of the results is
analyzed by taking the full aperture power consumption of Fujian Province from 2015 to 2017 as the
reference data.

The data set recorded a electricity consumption data of 282 relevant enterprises from 2015 to 2017,
with 243 valid data for 2015, 269 for 2016 and 282 for 2017. The frequency of the first digit for each
year is shown in Table 3.

| First digit | 2015  | 2016  | 2017  |
|-------------|-------|-------|-------|
| 1           | 50    | 50    | 41    |
| 2           | 34    | 36    | 17    |
| 3           | 42    | 48    | 50    |
| 4           | 36    | 43    | 54    |
| 5           | 25    | 31    | 49    |
| 6           | 22    | 29    | 31    |
| 7           | 15    | 12    | 22    |
| 8           | 12    | 12    | 11    |
| 9           | 7     | 8     | 7     |

A comparison with the Benford distribution is shown in Table 4.

| First digit | 2015  | 2016  | 2017  | Benford Distribution |
|-------------|-------|-------|-------|-----------------------|
| 1           | 0.206 | 0.186 | 0.145 | 0.301                 |
| 2           | 0.140 | 0.133 | 0.060 | 0.176                 |
| 3           | 0.173 | 0.178 | 0.177 | 0.125                 |
| 4           | 0.148 | 0.160 | 0.191 | 0.097                 |
| 5           | 0.103 | 0.115 | 0.174 | 0.079                 |
| 6           | 0.091 | 0.108 | 0.110 | 0.067                 |
| 7           | 0.062 | 0.045 | 0.078 | 0.058                 |
| 8           | 0.049 | 0.045 | 0.039 | 0.051                 |
| 9           | 0.029 | 0.030 | 0.025 | 0.046                 |

Calculate the Pearson correlation coefficient between Benford and annual data, the results are shown
in Table 5.

| Year   | Correlation coefficient | 2015 | 2016 | 2017 |
|--------|-------------------------|------|------|------|
|        | 0.8403                  | 0.731| 0.3024|
| Ratings| Suspicious              | Suspicious | Suspicious |

It can be seen that the data recorded for 2015-2017 falls into the "suspicious" range and has a high
level of error, thus suggesting that there are problems with the electricity consumption data for these
three years, and the next step is to use Zipf's law to estimate the overall true electricity consumption
data.

Talking 2017 as an example, the maximum data $F_{\text{max}}=22515.14$ and the minimum data $F_{10}=345.9687$
were selected firstly. Under the above constraints, 10 groups of data were randomly selected as follows:
Table 6. A group of experimental data

|   |   |
|---|---|
| F1 | 1234.072 |
| F2 | 949.6471 |
| F3 | 770.117 |
| F4 | 735.834 |
| F5 | 582.8482 |
| F6 | 514.7257 |
| F7 | 511.5613 |
| F8 | 447.818 |
| F9 | 439.7448 |
| F10 | 345.9867 |

Multiplying each data except the maximum by a natural number and then dividing by the maximum to make them close to the set Zipf rank. Then get a series of data as follows:

Table 7. Related data for calculated rank

| Calculated rank | I  | II | III | IV  |
|----------------|----|----|-----|-----|
| F1             | 0.987 | 1.973 | 3.015 | 4.001 |
| F2             | 1.012 | 1.982 | 2.995 | 4.007 |
| F3             | 0.992 | 1.984 | 3.010 | 4.002 |
| F4             | 1.013 | 1.994 | 3.007 | 3.987 |
| F5             | 1.010 | 1.993 | 3.003 | 4.012 |
| F6             | 1.006 | 1.989 | 2.995 | 4.001 |
| F7             | 1.022 | 1.999 | 2.999 | 3.999 |
| F8             | 0.994 | 2.009 | 3.003 | 3.999 |
| F9             | 0.996 | 1.992 | 3.008 | 4.004 |
| F10            | 0.999 | 1.998 | 2.997 | 3.995 |

(Rank)

Mean value

Mean value

Standard deviation

According to the above steps, taking the data of 2017 as an example, 20 data are selected for calculation each time, and a total of 100000 calculations are carried out. The relevant parameters are as follows:

Table 8. Relevant parameters obtained after 100000 calculations

| Rank | I  | II | III | IV  |
|------|----|----|-----|-----|
| n    | 36135 | 22881 | 22838 | 18146 |
| δ    | 0.026 | 0.019 | 0.023 | 0.014 |
| ST   | 14.152 | 11.809 | 10.013 | 12.851 |

Where n is the number of choices, δ is the average standard deviation, and ST is the selection trend value. According to the relevant parameters obtained from the calculation, the Zipf rank of the maximum Fmax should be selected as 1. However, the result of formula (4) does not meet the constraints, therefore choose 4 as Zipf rank to calculate.

Finally, the predicted results of electricity consumption data in 2017 are shown in Table 9. The deviation rate of 3.77% is a good result.

Table 9. Predicted results of electricity consumption

| predicted results | Total enterprise data | full aperture records | gap | deviation rate |
|-------------------|-----------------------|-----------------------|-----|----------------|
| 552955            | 332194                | 574612                | 242417 | 3.77%          |
In summary, the proposed algorithm can effectively find the problem data and estimate the overall electricity consumption through limited electricity consumption data. Deviation rate is within acceptable range, with the increase of the number of selected data, the effect of prediction will improve to a certain extent.

6. Conclusion

Aiming at the problem that the collected power consumption data may have a certain deviation from the real power consumption data, this paper proposes a correction method for power consumption data based on Benford-Zipf. Firstly, Benford’s law is used to evaluate the data quality, and then Zipf’s law is used to estimate the total amount of electricity data for the data with problems. The example analysis shows that the method has good prediction effect which can provide more effective data information for the power sector. However, at the same time, using Benford's law is still unable to identify the location of the specific data which is abnormal, and a certain amount of data is required. While using Zipf’s law is also difficult to determine the appropriate number of selected values and the search range of the maximum rank. Further research and improvement are needed.

References

[1] X. Wang. “Comparative analysis of the quality of electricity data and economic data,” The World of Survey and Research, no.10, pp.34-37, 2016.
[2] Newcomb, Simon. “Note on the Frequency of Use of the Different Digits in Natural Numbers,” American Journal of Mathematics, vol.4, 1881.
[3] Benford, F. “The Law of Anomalous Numbers,” Proceedings of the American Philosophical Society, no.78, 1938.
[4] J. Meng, X. Wang, R. Zhang. “Research of Corrected Benford's Distribution and Its Statistical Simulation,” Statistics&Information Forum, vol.32, no.9, pp.9-16, 2017.
[5] Z. Mi, X. Yang. “Benford’s Law: A New Paradigm of China’s Macroeconomic Statistical Data Quality Evaluation,” Mathematics in Practice and Theory, vol.44, no.24, pp.10-18, 2014.
[6] J. Zhang. “Benford Evaluation and Empirical Study on the Quality of Macroeconomic Statistics in China,” M.S. thesis, Shanxi University of Finance and Economics, Taiyuan, China, 2011.
[7] Z. Duan. “Study the Reliability about Data Quality of Farming, Forestry, Animal Husbandry and Fishery by Using Benford Law,” M.S. thesis, Yanshan University, Qinhuangdao, China, 2015.
[8] Nigrini M, Wells J T. “Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection,” 2012.
[9] N.J.Rowlands, D.Sampey. “Zipf's Law--An Aid to Resource Inventory Prediction in Partially Explored Areas,” Mathematical Geology, vol.9, no.4, pp.383-391, 1997.
[10] Y. He. “Application of Zipf's Law in Predicting Mineral Resources of Baijiazuizi Copper-Nickel,” Copper Engineering, no.6, pp.23-26, 2018.
[11] Q. Liu, S. Yan, S. Zhao. “Application of Qibofu Method: Prediction of Gold Resources,” Geology and Exploration, vol.35, no.4, pp.33-35, 1999.
[12] T. Zhou. “Application of Zipov’s Law and Bradford’s Law in Reserve Prediction of Laojinchang Gold Deposit,” Geology and Exploration, no.2, pp.23-26, 1989.
[13] Z. Yuan, Y. Zhou, W. Tang. “Analysis of the landslide and debris flow in China Based on the Law of Zipf,” The Chinese Journal of Geological Hazard and Control, vol.24, no.4, pp.34-39, 2013.
[14] Theodore P. Hill. “A Statistical Derivation of the Significant-Digit Law,” Statistical Science, vol.10, no.4, pp.354-363, 1995.
[15] C. Xu, D. Wang, F. Zhang. “The Empirical Analysis of Audit Opinion of Listed Company Detecting Fraud Using Benfords Law,” Journal of Nanjing University of Finance and Economics, no.4, pp.56-60, 2009.
[16] F. Ma. “A probabilistic model of the Bradford-Zippoff distribution,” Intelligence Science, no.2 pp.22-33, 1982.