Application Research of Benford's Law in Testing Agrometeorological Data

Lang Qin, Shuqing Han, Liwei Xing, Jianhua Zhang, Jing Zhang, Ruyi Yang and Jianzhai Wu

Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural, Beijing 100081, China

*Corresponding author’s e-mail: wujianzhai@caas.cn

Abstract. Benford's law is an effective tool to test the quality of data sets. In order to strengthen the research on the quality evaluation of agrometeorological data, this paper empirically analyzed the precipitation data sets of China from 1951 to 2015 from the time, space and time staggered dimensions. It was found that the overall and local quality of precipitation data sets in China was high. Further improvement and expansion of Benford's law's application in the future can help to detect and control the quality of agricultural data efficiently and improve the utilization rate of agricultural data.

1. Introduction

Data is an important force driving the development of agricultural modernization and the key to link agricultural production, management, consumption, market and trade, etc[1]. The effective application of agricultural data, on the one hand, can reflect the whole process of agriculture in holographic 3-D, promote the relationship between the relevant elements. On the other hand, it can also predict the future through the correlation characteristics of data, prepare in advance in a bid to respond to industry changes [2-3]. However, the quality of agricultural data in China is not high, and there is still a dilemma of rich data but poor information[4-5]. In the process of agricultural modernization, a large amount of data has been accumulated in each process of production, circulation and consumption, etc. Agricultural modernization also has the characteristics of lagging information technology, wide data coverage, complex data sources, closely related to time and space, and long production cycle, all of which make the quality of agricultural data in China poor. There exist many abnormal data, moreover, low-quality data sets are not conducive to make scientific decisions[2,6].

This paper introduces Benford's law, which has wide applicability and advantages in testing large amount of sample data into the field of agricultural data testing. Taking agrometeorological data as an example, the research is aimed at evaluating the quality of agrometeorological data in China, improving the efficient exploration and analysis of data in the field of agriculture in an effort to promote agricultural development.

2. Benford's Law

2.1. Overview of Benford's Law
Benford's law, also known as the "first figure law" or "the first digit law", refers to the statistical probability of data sets with the first digit of 1-9 in nature, which is not equal to 1/9 in general cognition, but satisfies the logarithmic distribution [7]. Wherein, the first digit refers to the first non-zero effective integer in the number. It is a simple numerical description rule, but it has a significant effect on the anomaly test in the data set [8, 9]. According to Benford's law, the first-digit frequency distribution table is shown in Table 1:

| First-digit probability | 0.3010 | 0.1761 | 0.1250 | 0.0969 | 0.0792 | 0.0670 | 0.0580 | 0.0512 | 0.0458 |

2.2. Benford's Law Application

In the field of natural sciences, many scholars have conducted a lot of research in mathematics, biology and physics, etc., and proved that many data sets conform to Benford's law. In mathematics, Pinkham[10] pointed out that data sets conforming to Benford's law are still conforming to Benford's law when multiplied by any non-zero constant. Berger A[11,12] proves that Benford's law is the only digital theorem with scaling invariance in mathematics. This series of studies on the nature of Benford's law laid a foundation for the study of data sets with correlation between indicators, and greatly broadened its application breadth and depth. Leemis L M's [13] statistical results for several common survival distributions of biology, J Burke's[14] statistics of the first digits of a large number of physical constants are consistent with Benford's frequency.

In the field of social sciences, due to the influence of environment and other factors, the phenomenon that data sets deviate from Benford's law in an acceptable range is more common. The method is applied to the statistical data detection research in the fields of taxation and accounting [15,16], social security and finance[17], showing that Benford's law has a significant effect on detecting the reliability of social science data.

2.3. Data Conditions Applying Benford's Law

Benford's law has a wide range of applications and has been applied to many unrelated fields. Nigrini[16] has made a thorough analysis and collation of the data detection research on a large number of applications of Benford's law, and put forward four necessary data conditions to satisfy the law: sample size large enough to represent the data as a whole; data values are unlimited; data is naturally formed, unaffected by human factors or less affected; there is no high coupling or cohesion between data.

3. Application of Agrometeorological Data Quality Detection

Firstly, the original data set is preprocessed to solve the problems of data missing, data redundancy, and data unavailability, etc. Then Benford's law is applied to detect data quality. If the statistical results of the first digit in time, space and time staggered dimensions, all the first digit statistical results pass significance test and obey Benford's law, the quality of data set is high, and the effective information can be extracted based on this part of data. If fails, the data set violates Benford's law, the anomaly data pool is screened for further analysis. According to the different characteristics of the test statistic, when the sample size is small, $\chi^2, VN^*, m^*, d^*$ can be applied to conduct significance test upon the results. When the sample size is large, the correlation coefficient $r$ is used to determine the result [18]. The statistics used in this paper are as follows.
Table 3. Grading Criteria based on Correlation Coefficient.

| Grading   | Grading Criteria of Correlation Coefficient |
|-----------|---------------------------------------------|
| Regular   | $0.99 < r \leq 1$                          |
| Attention | $0.97 < r \leq 0.99$                       |
| Suspicious| $r \leq 0.97$                              |

3.1. Data Sources
The agrometeorological data selected in this paper are obtained from China Meteorological Statistical Bureau. The data set includes the 65-year meteorological data of 2,481 stations in China from 1951 to 2015, mainly including the maximum temperature, minimum temperature, sunshine duration, daily precipitation and other indicators of each station. In agrometeorological data, precipitation, temperature, sunshine and other data reflect the weather conditions of different regions, which can accurately reflect the severity of natural disasters such as drought, flood and so on. It is significant to study these data for agricultural production. According to the four application limitations of Benford's law, highly cohesive data sets are not suitable for Benford's law test. Therefore, this paper chooses precipitation data from 31 provinces as the main research objects.

3.2. Data Preprocessing
Cleaning the selected data sources can reduce the interference of invalid data and improve the processing efficiency of Benford's law. After observing the agrometeorological data, we find that some precipitation data are distributed between 3,000mm and 3,070mm. However, according to the grading of precipitation levels in China, it is rainy if the daily precipitation is more than 100mm. This part of data deviates obviously from the precipitation records of normal data sets and shall be ignored directly in the statistical data. In addition, Benford's law counts the first digits from 1 to 9 and does not include the number “0”, so the record of daily precipitation of 0 shall also be ignored. Because of the particularity of Benford's law, ignoring this part of data will not have a negative impact on the test results.

3.3. Applying Benford's Law to Detect Data Quality

3.3.1. Time Dimension Test.
(1) The precipitation data of 31 provinces in China from 1951 to 2015 are divided into 7 groups based on 10 years' interval. To respectively study the distribution frequencies of the first digits of data sets, statistical results are shown in table 4.

| Benford freq. | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    |
|---------------|------|------|------|------|------|------|------|------|------|
| 1951-1959     | 30.103 | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| 1951-1959     | 32.656 | 18.420 | 12.523 | 9.214 | 7.339 | 6.090 | 5.221 | 4.526 | 4.011 |
| 1960-1969     | 32.601 | 18.675 | 12.504 | 9.199 | 7.349 | 6.084 | 5.164 | 4.491 | 3.932 |
| 1970-1979     | 32.463 | 18.820 | 12.524 | 9.177 | 7.330 | 6.044 | 5.167 | 4.519 | 3.955 |
| 1980-1989     | 32.700 | 18.694 | 12.519 | 9.214 | 7.296 | 6.040 | 5.140 | 4.482 | 3.915 |
| 1990-1999     | 32.905 | 18.598 | 12.432 | 9.178 | 7.310 | 6.023 | 5.135 | 4.499 | 3.919 |
From Table 4, it can be seen that the first digital frequency distribution of seven group’s data sets shows a monotonic decreasing trend, which is in good agreement with Benford's law. The frequency of which the first digit is "1" and "2" of precipitation data sets in China is about 32.7% and 18.6%, which is quite different from Benford's corresponding frequencies. Significance test is needed to determine whether these seven groups’ data sets obey Benford's law. Because these seven groups’ data are large sample data sets, according to the applicable conditions of Benford's law significance test method, distance m, d and correlation coefficient r pairs are selected as statistics of significance test.

By conducting significance tests on the 7 groups’ data, the results show that the correlation coefficient r of precipitation data from 1970 to 1979 is the lowest 0.9997, and that of the other six groups is 0.9998. According to the grading criteria of correlation coefficient, the precipitation data set is considered to be reliable as a whole. In addition, amongst the 7 data sets, the nearest distance from M is 0.025, the farthest distance is 0.03, the nearest distance from D is 0.027 and the farthest distance is 0.033. It can be considered that the frequency difference between data sets and Benford freq. is very small. The above test results show that the precipitation data sets in China conform to Benford's law, and the quality of each data set is high.

(2) Taking 5-year as group interval, the precipitation data sets from 1951 to 2015 in China are divided into 13 groups. Benford's law is tested. The statistical results of the test are shown in Table 5. Table 5. 1951-2015 national statistics on the first digital distribution of precipitation data (5 years).

| Year Interval | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|--------------|----|----|----|----|----|----|----|----|----|
| Standard Freq.| 30.103 | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| 1951-1955    | 32.652 | 18.286 | 12.325 | 9.180 | 7.336 | 6.200 | 5.308 | 4.670 | 4.042 |
| 1956-1959    | 32.660 | 18.706 | 12.525 | 9.195 | 7.282 | 6.047 | 5.147 | 4.503 | 3.935 |
| 1960-1964    | 32.543 | 18.655 | 12.518 | 9.228 | 7.348 | 6.065 | 5.188 | 4.506 | 3.949 |
| 1965-1969    | 32.658 | 18.695 | 12.490 | 9.171 | 7.350 | 6.103 | 5.140 | 4.476 | 3.916 |
| 1970-1974    | 32.466 | 18.875 | 12.532 | 9.172 | 7.324 | 6.032 | 5.143 | 4.514 | 3.943 |
| 1975-1979    | 32.461 | 18.767 | 12.516 | 9.182 | 7.336 | 6.056 | 5.192 | 4.524 | 3.967 |
| 1980-1984    | 32.660 | 18.706 | 12.525 | 9.195 | 7.282 | 6.047 | 5.147 | 4.503 | 3.935 |
| 1985-1989    | 32.740 | 18.682 | 12.513 | 9.234 | 7.311 | 6.033 | 5.132 | 4.461 | 3.896 |
| 1990-1994    | 32.985 | 18.605 | 12.421 | 9.146 | 7.268 | 6.016 | 5.137 | 4.515 | 3.908 |
| 1995-1999    | 32.820 | 18.591 | 12.443 | 9.212 | 7.355 | 6.031 | 5.133 | 4.483 | 3.931 |
| 2000-2004    | 32.809 | 18.524 | 12.441 | 9.198 | 7.379 | 6.049 | 5.149 | 4.507 | 3.943 |
| 2005-2009    | 32.915 | 18.416 | 12.382 | 9.182 | 7.363 | 6.074 | 5.169 | 4.548 | 3.951 |
| 2010-2015    | 32.831 | 18.464 | 12.459 | 9.222 | 7.349 | 6.070 | 5.125 | 4.546 | 3.935 |

Table 5 shows that the frequency of "1" and "2" of China’s precipitation data sets is about 32.7% or 18.5%, which is quite different from Benford's frequency of 30.103% and 17.609%. It needs to be judged whether the error can be approved by significance test. Since precipitation data sets belong to large sample data, according to the applicability of Benford's law test method, distance m, d and correlation coefficient r pairs are selected as statistics for significance test. Significance test results show that the correlation coefficient r of precipitation data in 1970-1979 is the lowest 0.9996, and the correlation coefficient r of precipitation data in 2000-2009 is the highest 0.9998. According to the grading criterion of correlation coefficient, 13 sets of precipitation data conform to Benford's law, and the overall quality of data sets is high.

(3) To test data sets with relatively large fluctuations in the first digit in time dimension:

In order to show and study the differences and characteristics of data in different time periods more clearly, the error maps of different groups’ data frequency statistics and Benford frequency are drawn. Among them, because the errors of the digit "3-9" are small and the gap is not big, the digit "3" is
chosen as the representative. The error maps of the first digital distribution frequency of national precipitation from 1951 to 2015 are shown in Figure 1 and Figure 2.

Figure 1. The first-digit distribution frequency error of national precipitation volume from 1951 to 2015 (10-year intervals).

Figure 2. The first-digit distribution frequency error of national precipitation volume from 1951 to 2015 (5-year intervals).

Figure 1 and Figure 2 show the errors between the first-digit freq. of each group set and Benford freq. Overall, the first digit "1" error fluctuates at 2.5% frequency, and the first digit "2" error fluctuates at 1.0% freq. Assuming that the data set with large error fluctuations may exist anomalies, the years of the most pointed and extreme points in the above figures are selected to do detailed analysis.

Figure 1 shows that the first digit "1" is the lowest in 1970-1979 and the highest in 1990-1999. The first digit "2" is the smallest in 1951-1954, the largest in 1970-1979. As shown in Figure 2, the first digit "1" is the smallest in 1975-1979, the largest in 1990-1994, and the local maximum in 1965-1969; The first digit "2" is the lowest in 1951-1955 and the highest in 1970-1974. When studying these years’ data, in order to ensure the accuracy of the test, by taking major event scope as standard, we select the precipitation volume data of 1951-1954, 1970-1979 and 1990-1999 to do analysis year by year.

(4) By taking one-year as group interval, Benford's law is tested on three data sets of precipitation volume of 1951-1954, 1970-1979 and 1990-1999. The results are shown in Tables 6, 7 and 8.

| Year   | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    |
|--------|------|------|------|------|------|------|------|------|------|
| 1951   | 30.103 | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| 1952   | 31.697 | 18.086 | 12.959 | 9.339 | 7.623 | 6.137 | 5.287 | 4.848 | 4.025 |
| 1953   | 32.890 | 18.364 | 11.970 | 8.993 | 7.370 | 6.007 | 5.381 | 4.843 | 4.181 |
| 1954   | 33.017 | 18.237 | 12.302 | 9.142 | 7.268 | 6.174 | 5.197 | 4.719 | 3.945 |
Table 6 shows that the distribution of the first digit of national precipitation volume data from 1951 to 1954 is consistent with that of Benford on the whole, but there is slight fluctuation. Whether the difference is obvious or not requires a significant test. The results of significance test show that the correlation coefficient \( r \) of precipitation volume data in 1951 is the lowest 0.999, and that in 1952 is the highest 0.9998. According to the grading criteria of correlation coefficient, these five data sets are not doubtful.

Table 7 shows that the distribution of the first digit of national precipitation data from 1970 to 1979 is consistent with that of Benford on the whole, but there is slight fluctuation. Whether the difference is obvious or not requires a significant test, and these 10 sets of data sets are not doubtful.

Table 8 shows that the distribution of the first digit of national precipitation data from 1990 to 1999 is consistent with that of Benford on the whole, but there is slight fluctuation. Whether the difference is obvious or not requires a significant test. The results of significance test show that the correlation coefficient \( r \) of precipitation data in 1993 is the lowest 0.9995, and the correlation coefficient in 1995 is the highest 0.9998. The data sets pass the significance test, and these 10 sets of data sets are not doubtful.

From the above test, we can see that the data sets of precipitation in China conform to Benford's law in terms of time dimension, and the quality of data sets is high.
3.3.2. Spatial Dimension Test

(1) According to the spatial division of the eastern, central and western regions, the precipitation data sets of 31 provinces in China from 1951 to 2015 are divided into three groups as shown in Table 9. The first-digit distribution frequencies of data sets are studied respectively, and the statistical results are as shown in Table 10.

Table 9. Eastern, Central and Western Cities Distributions.

| Region    | Cities                                                                 |
|-----------|-------------------------------------------------------------------------|
| Eastern   | Beijing Municipality, Tianjin Municipality, Hebei Province, Fujian Province, Guangdong Province, Zhejiang Province, Shandong Province, Jiangsu Province, Liaoning Province, Shanghai Municipality, Guangxi Zhuang Autonomous Region and Hainan Province |
| Central   | Anhui Province, Henan Province, Heilongjiang Province, Hubei Province, Hunan Province, Jilin Province, Jiangxi Province, Inner Mongolia Autonomous Region and Shanxi Province |
| Western   | Gansu Province, Guizhou Province, Ningxia Hui Autonomous Region, Qinghai Province, Sichuan Province, Shaanxi Province, Tibet Autonomous Region, Xinjiang Uygur Autonomous Region, Yunnan Province, Chongqing Municipality |

Table 10. The first-digit distribution frequency of precipitation volume in the eastern, central and western regions.

| Region | Benford freq. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--------|---------------|---|---|---|---|---|---|---|---|---|
| Eastern| 30.103        | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| Central| 32.779        | 18.702 | 12.547 | 9.234 | 7.365 | 6.048 | 5.146 | 4.489 | 3.936 |
| Western| 32.832        | 18.498 | 12.428 | 9.173 | 7.357 | 6.077 | 5.179 | 4.522 | 3.934 |

From Table 10, we can see that the results of the first digital frequency distribution of data sets in accordance with geographical distribution also show a monotonous decreasing trend, which is in good agreement with Benford's law. This shows that the quality of precipitation data in all regions of China is better and is not affected by the degree of regional development. The frequency of which the first-digit is "1" or "2" is about 32.7% and 18.5% in accordance with the geographical distribution of precipitation data sets, which basically fits the frequency of the first digital distribution of the data sets analyzed from the time dimension, but it is still quite different from Benford's frequency of 30.103% and 17.609%, so it is necessary to do significance test on the data sets.

According to the applicability of Benford's law test method, the correlation coefficient $r$ pair is selected as the statistical significance test. The results show that, among the data sets distributed according to geography, the correlation coefficient $r$ of precipitation data in the eastern region is the lowest 0.9983, the correlation coefficient $r$ of precipitation data in the central region is 0.9986, and the correlation coefficient $r$ of precipitation data in the western region is the highest 0.9988. According to the grading standard of correlation coefficient, the overall quality of precipitation data sets is very high.

(2) According to the geographical division of East China, South China, Central China, North China, Northwest China, Southwest China and Northeast China, the precipitation data of 31 provinces from 1951 to 2015 are further subdivided into seven groups, as shown in Table 11. Benford's law is tested on 7 sets of data sets. The statistical results are shown in Table 12.

Table 11. City Distributions in Seven Major Regions.

| Regions   | Cities                                     |
|-----------|--------------------------------------------|
| East China| Shandong, Jiangsu, Anhui, Zhejiang, Fujian and Shanghai |
| South China| Guangdong, Guangxi and Hainan              |
| North China| Beijing, Tianjin, Hebei, Shanxi and Inner Mongolia |
| Central China| Hubei, Hunan, Henan and Jiangxi            |
| NW China | Ningxia, Xinjiang, Qinghai, Shaanxi and Gansu |
From Table 12, it can be seen that the results of the first-digit frequency distribution of the data sets are in good agreement with Benford's law, which also shows a monotonic decreasing trend. The frequency of which the first digit is "1" or "2" of precipitation data sets in seven major regions is about 32.6% and 18.5%, which is quite different from Benford's frequency of 30.103% and 17.609%. Whether this difference is significant still needs to be tested.

According to the applicability of Benford's law test method, the correlation coefficient $r$ pair is selected as the statistical significance test. The results show that the correlation coefficient $r$ of precipitation data in South China is the lowest 0.9981, and the correlation coefficient $r$ of precipitation data in Southwest China is the lowest 0.9988, which is very close to 1. According to the grading criteria of correlation coefficient, the correlation between precipitation data sets and Benford's law is very high.

(3) To test data sets with relatively large fluctuations of the first digit in spatial dimensions:

In order to display and study the differences and characteristics of data in different regions more clearly, we make the error maps of statistical results of each group data freq. and Benford freq. Among them, because the error of the digit "3-9" is small and the gap is not large, the digit "3" is chosen as the representative. The error maps of the first-digit distribution frequency of precipitation volume in seven major regions from 1951 to 2015 are shown in figure 3.

Figure 3 shows the errors between the first digital frequency of each group data set and Benford frequency. Overall, the first digit "1" error fluctuates at a frequency of 2.5%, while the first digit "2" fluctuates at an error of 1.0%, but the fluctuation range is very small. Assuming that the data set with large error fluctuation may have anomalies, the regions where the most pointed and extreme points in the above figure are selected for detailed analysis. The first digit "1" is the maximum in Central and Southwest China, and the first digit "2" is the minimum in Northwest China. Next, for the next step, the three regions are analyzed and studied.
(4) Taking each province as the research object, Benford's law is tested upon three sets of precipitation data sets in Central, Southwest and Northwest China. The results are shown in Table 13.

Table 13. Errors in the first digital frequency distribution of precipitation in various provinces from 1951 to 2015.

|       | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Benford freq. | 30.103 | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| Hubei  | 32.745 | 18.692 | 12.451 | 9.144 | 7.326 | 6.036 | 5.094 | 4.544 | 3.969 |
| Hunan  | 32.050 | 18.711 | 12.656 | 9.407 | 7.374 | 6.095 | 5.203 | 4.587 | 3.918 |
| Jiangxi| 32.624 | 13.389 | 12.435 | 9.257 | 7.439 | 6.135 | 5.232 | 4.527 | 3.945 |
| Sichuan| 32.913 | 18.799 | 12.461 | 9.037 | 7.239 | 5.968 | 5.101 | 4.529 | 3.962 |
| Guizhou| 33.247 | 18.571 | 12.322 | 9.020 | 7.306 | 5.971 | 5.137 | 4.498 | 3.927 |
| Chongqing| 32.858 | 18.532 | 12.442 | 9.259 | 7.317 | 6.055 | 5.142 | 4.473 | 3.922 |
| Tibet  | 32.194 | 17.840 | 12.136 | 9.281 | 7.681 | 6.458 | 5.456 | 4.801 | 4.152 |
| Ningxia| 32.789 | 12.444 | 9.301 | 7.407 | 6.132 | 5.123 | 4.565 | 3.947 |
| Xinjiang| 32.728 | 18.879 | 12.828 | 9.387 | 7.340 | 5.953 | 4.996 | 4.258 | 3.631 |
| Qinghai| 31.927 | 17.528 | 12.406 | 9.402 | 7.752 | 6.507 | 5.543 | 4.772 | 4.163 |
| Shaanxi| 32.906 | 18.850 | 12.483 | 9.099 | 7.192 | 6.007 | 5.099 | 4.482 | 3.882 |
| Gansu  | 33.094 | 18.111 | 12.326 | 9.170 | 7.365 | 6.103 | 5.265 | 4.602 | 3.964 |

From Table 13, it can be seen that the distribution of the first digit of precipitation data in 1951-2015 in all provinces is consistent with Benford on the whole, but with slight fluctuations. The results of significance test show that the correlation coefficient r of precipitation data in Hunan and Tibet, etc. is the lowest, which is 0.9995. The correlation coefficient of precipitation in Xinjiang is the highest, which is 0.9999. Significance test results show that the quality of precipitation data sets in different provinces is high, and there is no significant regional difference in the quality of precipitation data in China.

3.3.3. Time and Space Staggered Dimension Test

(1) Table 13 shows that the first digit "1" and "2" of Hunan Province are significantly higher than Benford's law. In order to narrow the scope of the study, Hunan 2010-2015 data sets are divided into six groups and conduct Benford's law tests; the distribution frequency of the first digit is as shown in Table 14.

Table 14. The first digital distribution freq of precipitation in Hunan from 2010 to 2015.

|       | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Benford freq. | 30.103 | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| 2010  | 34.096 | 18.707 | 11.878 | 8.810 | 6.939 | 5.781 | 5.093 | 4.598 | 4.096 |
| 2011  | 32.759 | 18.102 | 12.476 | 9.553 | 7.605 | 6.133 | 5.135 | 4.465 | 3.772 |
| 2012  | 32.997 | 18.814 | 12.777 | 8.996 | 7.222 | 5.697 | 4.813 | 4.750 | 3.934 |
| 2013  | 33.478 | 18.869 | 13.075 | 9.335 | 6.801 | 5.923 | 4.755 | 4.183 | 3.580 |
| 2014  | 33.131 | 18.593 | 12.349 | 9.076 | 7.221 | 6.064 | 5.113 | 4.536 | 3.917 |
| 2015  | 33.505 | 17.576 | 12.581 | 9.299 | 7.567 | 6.102 | 4.982 | 4.487 | 3.901 |

From Table 14, we can see that the distribution of the first digit of precipitation data in Hunan Province in 2010-2015 is consistent with Benford on the whole, but the fluctuation is larger than other data sets, so it is necessary to test whether the difference is significant. Significance test results show that the correlation coefficient r of precipitation data in Hunan in 2010 is the lowest, 0.9984; the correlation coefficient in 2011 is the highest, 0.9998. Therefore, according to the criterion for judgment of correlation coefficient r, the data set conforms to Benford's law.
To further narrow the scope of the study, data of 12 regions in Hunan in 2010 are selected and divided into three groups, including Anhua, Anren, Anxiang and Baoqing; Chaling, Changning, Changde and Binzhou; Chenxi, Chengbu, Cili and Dao County, we conduct Benford's law tests upon them, the results of Benford's law tests are as shown in Table 15.

Table 15. The first digital distribution frequency of precipitation data in parts of Hunan in 2010.

| Region     | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  |
|------------|----|----|----|----|----|----|----|----|----|
| Benford freq. | 30.103 | 17.609 | 12.494 | 9.691 | 7.918 | 6.695 | 5.799 | 5.115 | 4.576 |
| Anhua, etc. | 31.429 | 18.195 | 12.632 | 9.323 | 9.173 | 4.211 | 6.015 | 4.662 | 4.361 |
| Chaling, etc. | 33.385 | 21.429 | 10.870 | 8.540 | 6.832 | 4.348 | 5.745 | 4.658 | 4.193 |
| Chenxi, etc. | 35.216 | 18.937 | 12.957 | 7.807 | 6.811 | 5.648 | 5.150 | 4.319 | 3.156 |

Table 15 shows that the precipitation data sets in some areas of Hunan are slightly different from Benford's law, and whether the difference is significant still needs to be tested. Because the sample sizes of each group are 665, 644 and 602 respectively, which are small sample data sets, the significance test can adopt statistics $\chi^2$, $V_{N^*}$, $m^*$, $d^*$, and the regional statistics of Anhua, etc. are 11.91, 1.29, 0.025 and 0.03 successively. The statistics of Chaling and other areas are 18.813, 1.951, 0.038 and 0.06 successively. The statistics of Chenxi and other areas are 15.334, 2.526, 0.05 and 0.061 in turn, which are less than the critical value of each statistic at confidence level of 0.01, and the correlation coefficient $r$ is greater than 0.99. Therefore, at confidence level of 0.01, these three groups of data sets also conform to Benford's law.

4. Conclusion
This paper studied the first-digit statistics of precipitation data sets of 65 years in China from the time, space and time-space interlaced dimensions. Benford's law was applied to test data quality. In order to eliminate the possibility that a small amount of abnormal data hidden in large data sets, the scope of testing was continuously narrowed down in the process of data quality testing. Finally, the scope of testing was reduced to 12 regions in Hunan Province in 2010. The first digit distribution frequency of data sets passes the significance tests and obeys Benford's law well. The empirical results showed that from the point of view of Benford's law, the quality of national precipitation data sets is good, and there are basically no suspicious data. This part of agrometeorological data can effectively guide agricultural production.

Acknowledgements
This work was supported by the Fundamental Research Funds for Central Non-profit Scientific Institution (Grant Number Y2018ZK46, JBYW-AII-2019-08, JBYW-AII-2019-41). *Corresponding author: WU Jianzhai (wujianzhai@caas.cn)

References
[1] Xu Shiwei, Wang Dongjie and Li Zhemin, Application Research on Big Data promoting Agricultural Modernization [J]. Agricultural Science in China, 2015, 48(17):3429-3438.
[2] Xu Shiwei, Agricultural Information Analytics [M]. Beijing: Higher Education Press, 2013.
[3] Chen Wei, Guo Shupu. Current situation and Existing Problems of Agricultural Informatization Technology in China [J]. Journal of Agricultural Engineering, 2013, 29(22): 196-205.
[4] Xu Shiwei. Agricultural Big Data and Early Warning of Agricultural Products Monitoring [J]. Review of China Agricultural Science and Technology, 2014, 16(05):14-20.
[5] Sun Zhongfu, Du Keming, Zheng Feixiang and Yin Shouyi, Research and Application Prospect of Big Data in Intelligent Agriculture [J]. Journal of China Agricultural Science and Technology, 2013, 15(06):63-71.
[6] Yu Yufeng, Wan Dingsheng. Application Research of Benford's Law in Hydrological Data Quality Exploration [J]. Microelectronics and Computers, 2011, 28(08):180-183+186.
[7] Benford, F. The Law of Anomalous Numbers [J]. Proceedings of the American Philosophical Society, 1938, 78(4):551-572.
[8] Hill, T. P. Base-invariance implies Benford's Law [J]. Proceedings of the American Mathematical Society, 1995, 123: 887-895.
[9] Hill, T. P. The significant-digit phenomenon [J]. American Mathematical Monthly, 1995, 102: 322-327.
[10] Pinkham, R. S. On the distribution of first significant digits [J]. The Annals of Mathematical Statistics, 1961, 32: 1223-1230.
[11] Berger, A., Bunimovich, L. A., Hill, T. P. One-Dimensional Dynamical Systems and Benford's Law [J]. Transactions of the American Mathematical Society, 2005, 357(1):197-219.
[12] Berger, A., Alberta, U. O. A basic theory of Benford’s Law [J]. Probability Surveys, 2011, 8(2011):1-126.
[13] Leemis, L. M., Schmeiser, B. W., Evans, D. L. Survival distributions satisfying Benford's law [J]. The American Statistician, 2000, 54: 236-241.
[14] Burke, J., Kincanon, E. Benford's law and physical constants: The distribution of initial digits [J]. American Journal of Physics, 1991, 59(10):952-952.
[15] Nigrini, M. J. A taxpayer compliance application of Benford's law [J]. The Journal of the American Taxation Association, 1996, 18: 72-91.
[16] Nigrini, M. J., Mittermaier, L. J. The use of Benford's law as an aid in analytical procedures [J]. Auditing, 1997, 16: 52-67.
[17] Rauch, B. Deficit versus social statistics: empirical evidence for the effectiveness of Benford's law [J]. Applied Economics Letters, 2014, 21(3):147-151.
[18] Liu Yunxia, Zeng Wuyi. Further study on the comprehensive application of Benford's rule and other methods to assess the quality of statistical data [J]. Statistical research, 2013 (8)30: 3-9.