Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Testing firm-level data quality in China against Benford's Law

Yasheng Huang [a], Zhiyong Niu [b], Clair Yang [c,*]

[a] Massachusetts Institute of Technology, 100 Main St, Cambridge, MA, USA
[b] Shanghai University of Finance and Economics, 777 Guoding Rd, Wu Jiao Chang, Shanghai, China
[c] University of Washington, King Ln NE, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 19 July 2019
Received in revised form 13 April 2020
Accepted 28 April 2020
Available online 4 May 2020

JEL classification:
C81
P20

Keywords:
Data quality
China
Benford's Law

ABSTRACT

The authenticity of China's economic data has long been questioned. We use a new statistical method, Benford's test, to evaluate data quality of the Chinese Industrial Census (CIC). We show that the method is effective to uncover data irregularities. Based on predicted industrial output by variables that are less manipulatable, such as employment and electricity, we further demonstrate that firms of different ownership types display different behavior in terms of the direction of data manipulation. We find no conclusive evidence of data manipulation by state-owned enterprises (SOEs), whereas private firms tend to under-report performance.

Published by Elsevier B.V.

1. Introduction

The accuracy of Chinese economic data has long been debated. The most prominent skeptic is none other than China's premier, Li Keqiang, who famously stated that China's GDP data are "man-made" and therefore he regards electricity and transport data to be better indicators of economic performance (WikiLeaks).

Previous studies on this topic have been inconclusive in terms of both the direction and the magnitude. In terms of direction, some argue that, similar to other developing countries, the underdeveloped reporting capacity of the Chinese state may lead to downward biases in growth rates (Holz, 2014).[1] Upward biases are also possible due to the political incentives of the authoritarian regime (Martinez, 2019). In terms of magnitude, some have uncovered evidence of substantial discrepancies (Rawski, 2001), whereas others believe that China's GDP data are largely accurate (Mehrotra and Pääkkönen, 2011; Chow, 2006; Holz, 2014).

Most existing studies on this topic rely on macroeconomic data and as a result, their estimations are based on a limited number of observations. In contrast, in this paper we utilize firm-level micro data. We apply a statistical test, known as Benford's Law or the First-Digit Law, to test the data quality of the Chinese Industrial Census (CIC).

Benford's Law was separately discovered by Frank Benford (1938) and Simon Newcomb (1881). It states that for many naturally occurring numerical sequences, the probability of observing a first digit of n should be approximately $log_{10}(1 + 1/n)$. This was initially applied to data series occurring in the natural environment, such as the length of rivers, molecular weights, or death rates (Hill, 1995; for a survey, see Miller, 2015). More recent advances in statistics and economics have extended Benford's Law beyond the natural environment to fraud detection in social activities, including stock prices (Ley, 1996), accounting (Nigrini, 2012; Fernandes and Guedes, 2010), elections (Pericchi and Torres, 2011; Deckert et al., 2011), international trade (Barabesi et al., 2018), and macro-economic data (Michalski and Stoltz, 2013; Rauch et al., 2011; Gonzalez-Garcia and Pastor, 2009). Benford's Law arises naturally for many types of data, including process of exponential growth (such as population and economic growth), products of independent random variables (such as sales and industrial output), and random data pooled from a large population of independent agents (such as population or industrial census) (Miller, 2015).[2]

In this paper, we applied two procedures. First, we use the CIC sample to calculate a city-level Benford's index as a measure of relative data quality. Because Benford's index does not convey information on the direction of data manipulations, we further use the city-level industrial output data to generate estimations

* Correspondence to: Thomson Hall, King Ln NE, Seattle, WA 98105, USA.
E-mail addresses: yshuang@mit.edu (Y. Huang), niu.zhiyong@mail.shufe.edu.cn (Z. Niu), clayang@uw.edu (C. Yang).

[1] For example, China's GDP data in the 1990s were frequently adjusted upward by the World Bank.

[2] For a mathematical treatment of the properties and further discussion about the suitability of Benford's Law for our industrial census data, please see the online appendix.

https://doi.org/10.1016/j.econlet.2020.109182
0165-1765/Published by Elsevier B.V.

on the direction and the magnitude of data manipulations. Two findings are particularly noteworthy. First, firm-level industrial output data in the CIC sample deviate substantially from the predictions of Benford's law. Second and more interestingly, the deviations from Benford's Law are not uniform across the universe of Chinese industrial firms. We find significant evidence of private firms' under-reporting, but only suggestive and inconclusive evidence of state-owned enterprises' (SOEs) over-reporting. The average impact on industrial output is about 1.3 percent under-reporting by the private firms over the sample period.

2. The data and Benford's index

Our paper relies on two sources of data. The firm-level data are from the Chinese Industrial Census (CIC). The CIC, compiled by the Chinese National Bureau of Statistics (NBS), includes the entire population of industrial firms with sales in excess of 5 million yuan (roughly US $600,000) for each of the census years.[3] The CIC contains information on industry classification, assets, sales, profits, employment, and registration type (such as a SOE or a private firm). We supplemented the CIC with data from the annual *Chinese City Statistical Yearbook* (CCSY). Our sample covers 195 cities in thirty provinces for the period from 1999 to 2011. On average, there are about 300,000 firms per year and a total of 1,029,642 unique firms over the entire sample period. Industrial firms covered by our data account for a majority of China's industrial output, ranging from 58 percent (1999) to 82 percent (2004).[4]

2.1. Benford's index

Benford's Law states that for many natural data sequences, the probability of observing a first digit of i should be approximately equal to $\log_{10}\left(\frac{i+1}{i}\right)$. To test an empirical distribution against Benford's Law, we use the Pearson's Chi-square test, with χ^2 statistics defined as

$$\chi^2 = \sum_{i=1}^{9} \frac{(e_i - b_i)^2}{Nb_i}$$

where N is the sample size, e_i is the observed frequency of digit i appearing as the first digit, and b_i is the expected frequency of i according to Benford's Law.[5]

We use the relative deviation from Benford's Law across cities and years as a measure of data quality. We calculate χ^2 statistics for a group of firms and cities in a specific year based on the CIC sample. This is referred to as the Benford's index. Unfortunately, we do not have a set of firms that we know do not manipulate. Hence, the Benford's indices we provide should only be interpreted in relative and comparative sense.

Fig. 1 shows the distribution of the Benford's indices across prefectural cities. The colored regions are those for which we have CIC data—the deeper the color the higher the index.

3. Benford's indices and abnormal output

To assess the extent of data manipulation, two issues must be considered: the frequency and the magnitude of the manipulations. The Benford's Index gives information about the occurrence (frequency) of data manipulation, but not the direction nor the magnitude.

To further investigate the direction and magnitude, we use electricity and employment to predict city-level industrial output and generate city-level industrial output residuals. For the purpose of this paper, we define the residuals not explained by electricity and employment as abnormal industrial output.[6] Let t denotes the year, and j the city. To allow for heterogeneous firm behavior based on different ownership types, we use I to distinguish SOEs from private firms, $I \in \{SOE, Private\}$. The following specification controls for the year and city two-way fixed effect:

$$\mathit{lnOutput}^I_{jt} = \beta^I_1 \mathit{lnElectricity}^I_{jt} + \beta^I_2 \mathit{lnEmployment}^I_{jt} + \alpha^I_t + \gamma^I_j + \epsilon^I_{jt}$$

$$\mathit{Res}^I_{jt} = \mathit{lnOutput}^I_{jt} - \widehat{\mathit{lnOutput}}^I_{jt}$$

Descriptive statistics in the online appendix show that there is no systematic difference in the output residuals between different ownership types.

Next, we use panel regression model to test whether the time-varying abnormal output correlates with the changes in data manipulation, as measured by Benford's index, and if so, in what direction.

$$\mathit{Res}^I_{jt} = \beta^I_3 \mathit{Benford}^I_{jt} + \alpha^I_t + \gamma^I_j + \epsilon^I_{jt}$$

Here, the sign of β^I signals the direction of data manipulation, and the absolute value of β^I measures the magnitude.

Unfortunately, the abnormal output variable will not be able to pick up consistent amount of manipulation in a city since that would be captured by the city fixed effect. But luckily, what we are interested in is not the absolute amount of manipulation but the direction of manipulation. By correlating the time-variant part of the abnormal output with that of the Benford's indices, we can get an estimate about the direction and the magnitude.

The results are shown in Table 1. Column (1) and (2) show that Benford's index does not correlate significantly with abnormal residuals for SOEs. Whereas in Column (3) and (4), private firms with a higher Benford's index are associated with lower (i.e., more negative) residuals. This suggests that private firms manipulate the reported data downward. This is plausible since managers of private firms have an incentive to reduce tax liabilities by under-reporting while managers of SOEs do not (Cai and Liu, 2009).[7]

As shown in Table 1, a one-unit increase in Benford's index for private firms is associated with a decrease in the city-level total output of private firms by about 0.023 percent. Based on the regression results, we estimate that in the CIC sample there is on average about 1.3 percent under-reporting for private firms over the sample period.

The online appendix supplies the robustness checks. We experimented with a bootstrap matching of SOEs and private firms based on size and industry to rule out the impact of confounding factors. Our results remain largely unchanged, except that the coefficient of SOEs turns out significant and positive for some specification, suggesting mild over-reporting for SOEs potentially due to managerial promotional incentives (Cao et al., 2018).

[3] The dataset does not include the service sector. Otherwise, its coverage of industry is comprehensive.

[4] Our data for national industrial output come from the *China Statistical Yearbook*.

[5] Here, the null hypothesis is that "the sample is drawn from a distribution satisfying Benford's Law".

[6] To calculate the predicted levels of industrial output, we assume that the value of industrial output to be a function of electricity usage and of employment. Arguably, such a prediction model omits many important factors that might help improve the prediction accuracy, most notably, industrial output in the previous years. We retain this omission because we want our prediction model to be free of data manipulation, hence we exclude variables that are of ambiguous quality. Moreover, our variable of interest, the Benford's index, is calculated based on the frequency that first digits appear in the firm—level data. This measures a pure statistical property and is very unlikely to correlate with any variable that might affect industrial output or its predicted value.

[7] Private firms also manipulate data more frequently than SOEs. The summary statistics table in the Appendix shows that the average Benford's index for SOEs is 8.84, whereas that for private firms is 61.5.

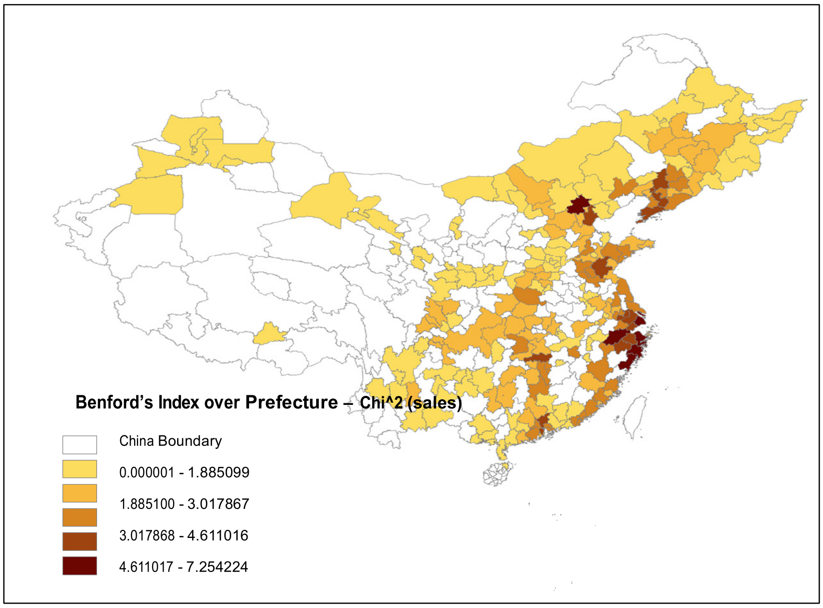

Fig. 1. The distribution of Benford's indices at prefecture level.

Table 1
Output residuals on Benford's indices: SOE vs. private.

	(1) Res_SOE	(2) Res_SOE	(5) Res_Private	(6) Res_Private
Benford's Index_Sales	0.00129 (0.00144)		−0.000226*** (5.77e−05)	
Benford's Index _Cost		0.000876 (0.00190)		−0.000313** (0.000129)
Sample	SOE	SOE	Private	Private
Fixed effect	Yr-City	Yr-City	Yr-City	Yr-City
Error	Clu Prov	Clu Prov	Clu Prov	Clu Prov
Obs	1489	1489	1491	1491
R^2	0.013	0.006	0.022	0.015
#city	283	283	283	283

Note: GDP per capita is included in the regression as a control for inequalities in regional development.

4. Conclusion

We contribute to the literature and the debates on Chinese statistics by using firm-level data, rather than relying on a limited number of macroeconomic data as is common in existing studies. We introduce a new statistical method, Benford's Law. The method imposes no onerous requirements on the underlying data other than the availability of a large number of observations, a requirement that can easily be satisfied by utilizing census data. We improve on the literature by allowing for heterogeneous firm behavior and we obtain results that are consistent with the imputed incentives of Chinese firms.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2020.109182.

References

Barabesi, Lucio, Cerasa, Andrea, Cerioli, Andrea, Perrotta, Domenico, 2018. Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud. J. Bus. Econom. Statist. 36 (2), 346–358.

Benford, Frank, 1938. The law of anomalous numbers. Proc. Am. Phil. Soc. 78 (4), 551–572.

Cai, Hongbin, Liu, Qiao, 2009. Competition and corporate tax avoidance: Evidence from Chinese industrial firms. Econ. J. 119 (537), 764–795.

Cao, X., Lemmon, M., Pan, X., Qian, M., Tian, G., 2018. Political promotion, CEO incentives, and the relationship between pay and performance. Manage. Sci..

Chow, Gregory C., 2006. Are Chinese official statistics reliable? CESifo Econ. Stud. 52 (2), 396–414.

Deckert, Joseph, Myagkov, Mikhail, Ordeshook, Peter C., 2011. Benford's Law and the detection of election fraud. Political Anal. 19 (3), 245–268.

Fernandes, Nuno, Guedes, José, 2010. Keeping up with the Joneses: A model and a test of collective accounting fraud. Eur. Financial Manag. 16 (1), 72–93.

Gonzalez-Garcia, Jesus, Pastor, Gonzalo C., 2009. Benford's Law and Macroeconomic Data Quality. International Monetary Fund, Washington, DC, WP/09/10.

Hill, Theodore P., 1995. A statistical derivation of the significant-digit law. Statist. Sci. 10 (4), 354–363.

Holz, Carsten A., 2014. The quality of China's GDP statistics. China Econ. Rev. 30, 309–338.

Ley, Eduardo, 1996. On the peculiar distribution of the US stock indexes' digits. Amer. Statist. 50 (4), 311–313.

Martinez, Luis, 2019. How much should we trust the dictator's GDP growth estimates? Available at SSRN: https://ssrn.com/abstract=3093296 or http://dx.doi.org/10.2139/ssrn.3093296.

Mehrotra, Aaron, Pääkkönen, Jenni, 2011. Comparing China's GDP statistics with coincident indicators. J. Comp. Econ. 39 (3), 406–411.

Michalski, Tomas, Stoltz, Gilles, 2013. Do countries falsify economic data strategically? some evidence that they might. Rev. Econ. Stat. 95 (2), 591–616.

Miller, Steve J., 2015. Benford's Law: Theory and Applications. Princeton University Press.

Newcomb, Simon, 1881. Note on the frequency of use of the different digits in natural numbers. Amer. J. Math. 4 (1), 39–40.

Nigrini, Mark J., 2012. Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. John Wiley & Sons.

Pericchi, Luis, Torres, David, 2011. Quick anomaly detection by the Newcomb—Benford Law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. Statist. Sci. 26 (4), 502–516.

Rauch, Bernhard, Gōttsche, Max, Brähler, Gernot, Engel, Stefan, 2011. Fact and fiction in EU governmental economic data. Ger. Econ. Rev. 12 (3), 243–255.

Rawski, Thomas G., 2001. What is happening to China's GDP statistics? Econ. Rev. 12 (4), 347–354.