Does The Granary County Subsidy Policy Lead to Manipulation of Grain Production Data in China? – Evidence from a Natural Experiment

X. Yu[1]; X. Zhang[2]; L. You[3]

1: University of Goettingen, Agricultural economics and rural development, Germany, 2: Huazhong Agricultural University, College of Economics and Management, China, 3: International Food Policy Research Institute, , United States of America

Corresponding author email: xyu@gwdg.de

Abstract:

Manipulation of food production data could lead to catastrophic social and economic consequences. The accuracy of official agricultural statistics has long been questioned in China. As a natural experiment, this paper studies the linkage between agricultural production data manipulation and the Granary Country Subsidy Policy (GCSP). Chinese government gave subsidies to the counties with annual grain production more than 200 thousand tons to encourage these local governments to give priority on grain production from 2005. In order to obtain the subsidies, the prospective counties with food production slightly below the threshold may have incentives to over-report their grain production. Based on the McCrary (2008)'s density test, our empirical results confirm that the GCSP results in over-reporting of grain production in those countries. Furthermore, data manipulations are more likely to happen in major-grain-production, low-income and mid-western counties. The policy implication would be that the fiscal distribution rules of a central government should avoid data manipulation incentives in local governments, particularly should cut the linkage to the data which are self-reported by the local governments.

Acknowledegment: Thanks

JEL Codes: Q18, O23

#1578

Does The Granary County Subsidy Policy Lead to Manipulation of Grain Production Data in China? –Evidence from a Natural Experiment

Abstract: Manipulation of food production data could lead to catastrophic social and economic consequences. The accuracy of official agricultural statistics has long been questioned in China. As a natural experiment, this paper studies the linkage between agricultural production data manipulation and the Granary Country Subsidy Policy (GCSP). Chinese government gave subsidies to the counties with annual grain production more than 200 thousand tons to encourage these local governments to give priority on grain production from 2005. In order to obtain the subsidies, the prospective counties with food production slightly below the threshold may have incentives to over-report their grain production. Based on the McCrary (2008)'s density test, our empirical results confirm that the GCSP results in over-reporting of grain production in those countries. Furthermore, data manipulations are more likely to happen in major-grain-production, low-income and mid-western counties. The policy implication would be that the fiscal distribution rules of a central government should avoid data manipulation incentives in local governments, particularly should cut the linkage to the data which are self-reported by the local governments.

Keywords: The Granary County Subsidy Policy, Grain Production, Data Manipulation, McCrary (2008)'s Density Test, China

JEL: Q11, Q 18, O23

Introduction

Manipulation of food production data could lead to catastrophic social and economic consequences, as many policies are made based on the data in modern society. It is known that the big famine at the beginning of 1960s in China which caused more than 30 million unusual deaths was linked to manipulation of food production data in the period of the Great Leap Forward (Lin 1990; Bernstein 2006). Given such a history, the accuracy of China's major statistics data generally has long been questioned. Some literature finds that (1) China's GDP is often over-reported, though the evidence is not solid (Holz, 2014); (2) China's CPI is often slightly under-reported

(Chamon and de Carvalho Filho, 2013; Nakamura, Steinsson and Liu, 2016) ; (3) Food production is often over-reported, but the consumption under-reported (Fuller, Hayes & Smith, 2000; Yu and Abler 2014 & 2016); (4) Enviornmental pollution data are often under-reported (Ghanem and Zhang 2014). In China, most of the data are reported by local governments, and the statistical data are linked to their performance review and future possible promotion (Yu and Abler 2014). It is comprehensible that such an incentive mechanism leads to data manipulation.

However, the incentive mechanism of data manipulation could be more than individual rational (e.g. performance review or promotion), and the fiscal system in China also plays important roles. China reformed its fiscal system, and carried out the so-called revenue-sharing system in 1994, in which the central government controlled most of the revenues (Lee, 2000; Chen, 2004; Parker and Thornton, 2007). In order to get more fiscal distributions from the central government, it is rational for local governments to manipulate their statistical data to fawn the central government's distributional rules. Agricultural sector is no exception.

Given the sheer size of its population and limited land and water resources, China always put food security in a prior position of their policy agenda. In order to maintain high food self-sufficient rate, China has taken a series of policy measure to subside agriculture after 2000.

Some policy changes, as natural experiments, could help us observe how the accuracy of agricultural statistical data in a region reacts to these subsidy policies. One good experiment is the Granary Country Subsidy Policy (GCSP), which is announced in 2005 to subsidize county governments with annually average grain

production more than 200 thousand tons[1] between 1998 and2002, and the commodity grain more than 100 thousand tons. The key statistics for the policy is 200 thousand tons of grain output, as 100 thousand tons of commodity grains could easily reach each when the total production is over 200 thousand ton. The policy states that the subsidies were based on the agricultural production statistics between 1998 and 2002; it seems that local governments cannot change the past data, and have no incentives for data manipulation. However, the document also stated that the list of subsidizing counties would not change in 3 years (Clause 12). Implicitly, the county list would be adjusted 3 years later when the GCSP continued. In other words, counties which were slightly below and could potentially reach the threshold of 200 thousand tons would have incentives to cook their production data from 2005, while the countries with production more than 200 thousand toms did not have such incentives.

Three approaches can help the counties below the threshold to achieve the granary county subsidy: (1) Increasing yield, (2) increasing harvest area, (3) manipulating the data. The former two approaches cannot be easily realized. Grain yield is relatively very high in China, and further increase in yield will result in heavy environmental and economic costs. The arable land has been fully used and even shows a decrease trend in China (Chen, 2009). In comparison, effortless data manipulation is a relatively easy way (Ghanem and Zhang, 2014).

The subsidy did increase the fiscal coffer of the granary counties. In 2005, approximate 800 counties shared a total amount of RMB5.5 billion subsidies. The average subsidy amount for each county is about RMB 6.7 million, accounting for 3 per cent of their fiscal revenues. The amount of subsidy increased to RMB39.3 billion

[1] The orginal document from the Ministry of Finance of China can be found at: http://www.mof.gov.cn/zhengwuxinxi/caizhengwengao/caizhengbuwengao2005/caizhengbuwengao20056/200805/t20080525_42774.html

in 2016, with an annual growth rate of 20%. Unfortunately, the central government changed the subsidy rules in 2008, and perhaps realized that the subsidy rules in 2005 might have incentives to local government to over-reporting their production data. Long-time systematically over-reporting of grain production could distort China's agricultural policy. In addition, data manipulation undermines the credibility of the governments (Ghanem and Zhang, 2014)

Three years later, in 2008, the Ministry of Finance declared that the subsidizing counties remained the same as those in 2007[2], which are based on the 5 year average production between 2002 and 2006. As a principle, the subsidizing counties would remain unchanged after 2008. The increased subsidies later mainly went to the top 100 super granary counties. As the list of granary counties were fixed since then, the incentives of data manipulation disappeared.

The changes in subsidizing rules are a perfect nature experiment for us to conduct a research to see if local counties (particularly those are slightly below the threshold of 200 thousand tons) cooked their grain production data in reaction to the GCSP between 2005 and 2007, while the data after 2008 could be used as a robust check.

Background and Economic Mechanism

Data manipulation is widely studied in many sub-fields of economics. Driven by a variety of motivations, agents are likely to hide some information to meet with the particular standards due to asymmetric information. Zitzewitz (2012) refers to this research topic as "forensic economics". Though the questions differ in different sub-fields, the detection techniques are quite similar. Comparing the reported data with

[2] The Granary County Subsidy Policy in 2008 from the Ministry of Finance could be seen in: http://www.mof.gov.cn/preview/gp/jingjijianshesi/200806/t20080625_52829.html

other source is the simplest approach to detect manipulation. By comparing the self-reported data to official data, Zinman and Zitzewitz (2009) found that ski resorts over-report substantially more snowfalls on weekends for greater benefits. Fisman and Wei (2004) compared Chinese Mainland's official reported imports from Hong Kong with Hong Kong's official reported exports to Chinese Mainland, and found that higher tariff products were misclassified into lower tariff categories for tax break.

Econometric and statistical tests can be used to detect the manipulation when there is no alternative data source. Burgstahler and Dichev (1997) developed a pooled cross-sectional distribution approach and revealed that firms are likely to manipulate their reported earnings to avoid earnings decreases and losses. In order to avoid tax, Saez (2010) found substantial evidence of discontinuity in the density of income. McCrary (2008) developed a density test and found strong evidence of manipulation of the roll call votes in the House through representatives' repeated game, but no evidence of manipulation in the popular elections to the United States House of Representatives. Chen et al. (2012) and Ghanem and Zhang (2014) adopted Burgstahler and Dichev (1997)'s distribution test and McCrary (2008)'s density test to detect the manipulation of the self-reported daily air pollution concentrations, and they found that some Chinese cities under-report the data to response to the requirement of air pollution abatement and increase the number of "blue-sky days".

The simplest way to detect manipulation is to use independent statistics to validate the self-reported data. Unfortunately, the grain production data from the alternative sources is unavailable. Therefore, this paper turns to econometric methods to uncover the evidence of manipulation.

We will adopt the density test developed by McCary (2008) to detect whether or not the prospective counties over-report their grain production at 200 thousand tons to

achieve subsidy. In the absence of manipulation, the distribution of grain production for all of the prospective counties are expected to be continuous because the distribution of yield can be assumed to be continuous, though being skewed to the left (Kim and Chavas, 2003; Wang et al., 2017). The increase in grain yield will only shift the distribution rather than leading to a discontinuity of grain production. In this context, the mechanism here is that if the grain production of a county doesn't reach the cut-off of 200 thousand tons for a small amount, the county is likely to over-report its grain production. This behavior will result in a discontinuity of grain production around the cut-off if it occurs for most of prospective counties. Theoretically, it is possible that manipulation would not lead to a discontinuity, but the manipulators must have knowledge of the distribution of all the prospective counties which is unlikely. Therefore, we will conduct the density test at different intervals and different breakpoints to confirm the existence of discontinuity at 200 thousand tons.

However, this discontinuity may also attribute to the expanding of grain harvest areas because prospective counties approximately know to what extent they can exceed the threshold by expanding their sown areas based on both grain production and yield of last year. While it is a very difficult to precisely expand grain harvest area to achieve the grain production of 200 thousand tons, we cannot take it as manipulation. We will test whether the grain harvest areas (rice, wheat and corn harvest areas) of prospective counties experience a significant increase after introducing the granary county subsidy. If the grain harvest areas increase insignificantly, we will more confidently attribute the discontinuity to data manipulation. Thereafter, we will furtherly classify the prospective counties according to both economic development level and location to investigate the characteristics of which the county is likely to manipulate data.

This paper has two contributions. First, although "forensic economics" has been applied to many fields, this paper firstly apply the detection techniques on Chinese grain production data to investigate whether the granary county subsidy result in data manipulation. Given the central role of food security in Chinese agricultural policy, this paper will provide valuable implications for policymakers. Second, the discontinuity at cut-off can attribute to both data manipulation and grain harvest area expanding. Therefore, we furtherly use the grain harvest area data to confirm that the discontinuity of grain production is due to data manipulation.

Data

The data used in this paper are collected from the county level statistical yearbooks from 2000-2010. There are more than 2000 counties in our dataset. Figure 1 shows that the average grain production per county increases from 210 thousand tons in 2000 to 270 thousand tons in 2010, with an annual growth rate of 2.9 per cent. In 2005, the average grain production is 0.23 million tons, which is higher than the subsidizing threshold. Table 1 shows that the grain production of 1220 counties, accounting for 59.8 per cent of the number of counties, is less than 200 thousand tons, which is the threshold of GCSP in 2005.

We are interested in whether the granary county subsidy results in data manipulation. Table 1 also presents the number and proportion of counties with grain production above and below the threshold. After the granary county subsidy being introduced in 2005, the proportion of counties with grain production between 180-200 thousand tons remains the same in 2005, and decrease sharply to 3.3 per cent in 2006. At the same time, on the contrary, the proportion of counties with grain production between 200-220 million tons had a moderate increase from 3.1 per cent in 2004 to

3.2 per cent in 2005, and jumped to 4.5 per cent in 2006. Furthermore, the proportion of counties with grain production between 180-200 thousand tons reached its lowest level in 2006 and 2007, but the proportion of counties with grain production between 200-220 thousand tons reached its peak level in 2006. The descriptive statistics have shown some evidence of data manipulation, particularly for the counties with grain production between 180-200 thousand tons. In this context, it is thus necessary for us to detect the data manipulation with the statistical tests of McCary (2008).

In addition, our dataset also includes the harvest areas of rice, wheat and corn in research period. These data come from the Ministry of Agriculture and the Chinese Academy of Agriculture Science (CAAS).

[Table 1 here]

[Figure 1 here]

Methodology

Due to lack of an alternative data source, an empirical method with the density test of McCary (2008) will be adopted in this paper. The prospective counties are likely to over-report their grain production and cause bunching of grain production counties above the cut-off. Estimating a density function is thus a simple and straightforward method to identify the discontinuity. Both traditional histogram techniques and kernel density estimates have been used in previous studies (DiNardo and Lee, 2004; Jacob and Lefgren, 2004). However, it is well known that the kernel density estimator is badly biased at the boundary (McCrary, 2008). The local linear density estimator developed by Cheng et al. (1993) and Cheng et al. (1997) can overcome boundary bias and has theoretical and practical advantages (McCrary,

2008). As a simply extension of Cheng et al. (1997)'s local linear density estimator, McCrary (2008) developed a more general density test to employ all the data on either side of the cut-off and make the results visual (Chen et al., 2012).

The density test developed by McCrary (2008) is informative when the existence of the program induces agents to adjust the data in one direction only. The granary county subsidy policy creates an incentive for the prospective counties slightly below the threshold to draw their grain production data just above the cut-off of 200 thousand tons. Therefore, the manipulation of grain production is expected to be monotonic and the McCrary (2008)'s density test is appropriate.

The McCrary (2008)'s density test is a Wald test and the null hypothesis is that the discontinuity is zero. It includes two steps to detect the discontinuity on grain production P_i.

Step 1: obtaining a finely gridded histogram. The bins of the histogram should be undersmoothed enough to guarantee that all of the histogram bins don't include points both to the left and right of the cut-off. The histogram of grain production can be written as the frequency table of its discretized version.

$$g(P_i)=\left\lfloor \frac{P_i-c}{b} \right\rfloor b+\frac{b}{2}+c \in \left\{ ...,c-5\frac{b}{2},c-3\frac{b}{2},c-\frac{b}{2},c+\frac{b}{2},c+3\frac{b}{2},c+5\frac{b}{2},... \right\} \quad (1)$$

where $\left\lfloor \frac{P_i-c}{b} \right\rfloor$ is the floor function and denotes the greatest integer in $\frac{P_i-c}{b}$. b is bin size, and c is the cut-off or break point. More specifically, there is an equi-spaced grid $\{X_1, X_2, X_3, ..., X_J\}$ with width b covering the support of $g(P_i)$. The cellsize of j th bin is $Y_j=\frac{1}{nb}\sum_{i=1}^{n} 1\left(g(P_i)=X_j\right)$. The scatterplot $\left(X_j, Y_j\right)$ is the histogram of grain production.

Step 2: conducting the local linear regression to smooth the histogram separately on either side of the cut-off. Define the estimator as the log difference in height between the left and right limit of the density of the grain production at the cut-off.

$$\theta = \ln \lim_{r \downarrow c} f(p) - \ln \lim_{r \uparrow c} f(p) \equiv \ln f^{+} - \ln f^{-} \tag{2}$$

where $f(p)$ is the density function denoting the height of bin at point p. f^{+} and f^{-} denote the right and left limit. We can estimate f^{+} and f^{-}, respectively. However, it is more precise to estimate two separate local linear regressions on either side of cut-off with $X_j - c$ as regressor (McCrary, 2008). The local linear regression uses the bin midpoints to explain the height of the bins to smooth the histogram. The local linear estimators of θ is as follows:

$$\begin{aligned}\hat{\theta} &\equiv \ln \hat{f}^{+} - \ln \hat{f}^{-} \\ &= \ln \left\{ \sum_{X_j > c} K\left(\frac{X_j - c}{h} \right) \frac{S_{n,2}^{+} - S_{n,1}^{+}\left(X_j - c \right)}{S_{n,2}^{+} S_{n,0}^{+} - \left(S_{n,1}^{+} \right)^2} Y_j \right\} - \ln \left\{ \sum_{X_j < c} K\left(\frac{X_j - c}{h} \right) \frac{S_{n,2}^{-} - S_{n,1}^{-}\left(X_j - c \right)}{S_{n,2}^{-} S_{n,0}^{-} - \left(S_{n,1}^{-} \right)^2} Y_j \right\}\end{aligned} \tag{3}$$

Where $S_{n,k}^{+} = \sum_{X_j > c} K\left(\frac{X_j - c}{h} \right) \left(X_j - c \right)^k$ and $S_{n,k}^{-} = \sum_{X_j < c} K\left(\frac{X_j - c}{h} \right) \left(X_j - c \right)^k$. $K(\mathrm{g})$ is the kernel function, and $K(t) = \max\{0, 1 - |t|\}$ is defined as triangle kernel in McCrary (2008)'s density test. The estimator $\hat{\theta}$ is asymptotically normal:

$$\sqrt{nh}\left(\hat{\theta} - \theta \right) \xrightarrow{d} N\left(B, \frac{24}{5}\left(\frac{1}{f^{+}} + \frac{1}{f^{-}} \right) \right) \tag{4}$$

Where h is the bandwidth and $B = \frac{H}{20}\left(\frac{-f^{+"}}{f^{+}} - \frac{-f^{-"}}{f^{-}} \right)$, $H = \lim_{n \to \infty, h \to 0} h^2 \sqrt{nh}$, and $H \in [0, \infty)$. Bandwidth h is very important to achieve good performance of $\hat{\theta}$. McCrary (2008) demonstrated that the estimator $\hat{\theta}$ is robust to different choices of bin

size b for a fixed bandwidth and $h/b > 10$. To be practical, McCrary (2008) suggested a subjective choice of bandwidth h based on an automatic procedure.

It is expected that prospective counties are likely to draw their grain production just above the cut-off of 200 thousand tons. It will result in the right limit to be higher than the left limit.

Empirical results

Based on McCrary (2008)'s density test and above dataset, this section will uncover the suggestive evidence of data manipulation that attribute to the implementation of granary county subsidy.

In order to detect the data manipulation, we will first conduct McCrary (2008)'s density test at different break points and different intervals to confirm the existence of discontinuity at 200 thousand tons. As mentioned above, both manipulation and expanding harvest areas can lead to discontinuity. Therefore, based on regression discontinuity model, we will investigate whether the granary county subsidy cause significant increase in grain harvest areas. If the grain harvest area increases insignificantly in 2005, 2006 and 2007, we can rule out the possibility that the discontinuity of grain production is caused by harvest area expanding. Therefore, we can attribute the discontinuity of grain production to manipulation. In addition, the Ministry of Finance changed the subsidizing rules that the subsidy candidates remained unchanged after 2008. This change provides us with a perfect nature experiment to conduct a robustness check. We will also conduct McCrary (2008)'s density test for 2008, 2009 and 2010, and our expectation is that there is no evidence of discontinuity in these years.

There are no incentives for counties with true grain production far more than 200 thousand tons to over-report their data, only the counties with grain production m slightly below 200 thousand tons are likely to manipulate the data. If the data are drawn by large amounts, the central government may doubt the grain production reported by local governments. Therefore, this paper will conduct McCrary (2008)'s density test around the cut-off. Three different intervals are defined, and they are grain production between 0.17-0.23 million tons, 0.18-0.22 million tons, and 0.19-0.21 million tons. Figure 2 shows that the grain production has discontinuity at 200 thousand tons in 2005 and 2006 for interval of 0.18-0.22, but it is continue for other years.

[Figure 2 here]

Although the graph is more visual, t-statistic is more precise because it is normalized by its variance. This paper uses a more strict 1 per cent critical t-statistic because our motivation is to detect data manipulation. This paper focuses on the behavior of over-reporting the grain production, thus the test is one-side and t-statistic is expected to be more than 2.36[3]. However, a larger t-statistic just implies a higher level of confidence to reject the null hypotheses with the existence of manipulation. It does not mean a higher level of manipulation.

Table 2 presents the results of McCrary (2008)'s density test for different intervals. We reported two results for each interval because the estimator is robust to different bin size for a fixed bandwidth if the ratio of bandwidth and bin size is greater than 10. The bandwidth and bin size in first row of each interval is from the automatic

[3] Due to different intervals, the test has different degree of freedom. For interval of 0.19-0.21 million tons, the t-statistic shall be greater than 2.37. For interval 0.18-0.22 million tons, the t-statistic shall be greater than 2.36. For intervals 0.17-0.23 and 0.16-0.24 million tons, the t-statistic shall be greater than 2.35.

procedure recommended by McCrary (2008). Bandwidth in third row of each interval is also from the automatic procedure, but the bin size is subjectively adjusted to guarantee the value of h/b greater than 10. The t-statistics indicate that the null hypothesis of the discontinuity being zero cannot be rejected at the 1 per cent significance level for all of the intervals and different h/b values before the granary county subsidy policy being implemented. However, the all of the null hypotheses are significantly rejected at the 1 per cent significance level (the t-statistic is significant at the 2.5 per cent significance level for interval of 0.19-0.21 million tons) when granary county subsidy introduced in 2005. The results imply that grain production has a strong evidence of discontinuity at 200 thousand tons. The results in 2006 are similar with the results in 2005, but all of the t-statistics become insignificant at the 1 per cent significance level in 2007. This result indicates that the implementation of granary county subsidy may result in a short term discontinuity on grain production. In addition, we have found that the results are similar for the different h/b values (smaller and greater than 10), thus, in the following test, we will only report the results using the bandwidth and bin size recommended by the automatic procedure.

[Table 2 here]

Based on the results conducted on different intervals, this paper also investigates whether the grain production is continuous or not at different break points. Table 3 presents the t-statistics for the break points at 0.19, 0.195, 0.2, 0.205 and 0.21 million tons, and the McCrary (2008) density test was conducted in the interval of 0.18-0.22 million tons. The results show that only the null hypothesis of the discontinuity being zero at 200 thousand tons in 2006 is rejected at 1 per cent significance level. The other t-statistics are not statistically significant. The results are consistent with our

expectation and furtherly demonstrate that the grain production has and only has one discontinuity at 200 thousand tons. In addition, it is worth noting that the null hypotheses of discontinuity being zero at 0.19 and 0.20 million tons are significantly rejected at 2.5 per cent significance level in 2005. One possible explanation is that data manipulation may occur step by step. The county with grain production smaller than 0.19 million tons may firstly draw their grain production above 0.19 million tons and then draw the data above 200 thousand tons in next year because small adjustment is difficult to discernible.

[Table 3 here]

Based on the results in Table 2 and 3, we have strong evidences that county level grain production has a discontinuity at 200 thousand tons in 2005 and 2006. However, as mentioned above, we cannot simply attribute the discontinuity to manipulation. The prospective counties can encourage farmers to expand their harvest areas to exceed the grain production threshold, and they know to what extent the grain production should be increased according to the grain yield and grain production of last year. Although it is very difficult to precisely achieve this target, we should take this possibility into consideration. Therefore, we will furtherly investigate whether or not the grain harvest areas of prospective counties increase significantly in 2005 or 2006.

Fortunately, our dataset includes rice, wheat and corn harvest areas. We assume that prospective counties mainly expand their rice, wheat or corn harvest areas to increase their grain production to exceed the threshold because the yield of rice, wheat and maize are higher than that of beans and tubers (NBSC, 2011). Based on

regression discontinuity method, this paper will furtherly estimate the impact of granary county subsidy on rice, wheat and corn harvest areas of prospective counties.

Both regression discontinuity method and McCrary (2008) density test are used to estimate the discontinuity of a target variable. The regression discontinuity model focuses on evaluating the impact of a specific variable on the discontinuity of target variable. The McCrary (2008) density test is used to test the discontinuity of the target variable itself by using its own density function.

Table 4 presents the results of the regression discontinuity method for rice, wheat and corn harvest areas. The regression discontinuity method is conducted in the interval of 0.18-0.22 million tons. The estimates are sensitive to the choice of bandwidth. This paper mainly focuses on the results using the default bandwidth of 100, and we will also take the results using the twice bandwidth of 200 as references.

The results show that for 2005, the average corn harvest areas of prospective counties increase significantly, but the average rice harvest areas and the average wheat harvest areas have no significant increase. In 2006, we don't find any significant increase for rice, wheat and corn harvest areas. Furthermore, we add some control variables including employment in agriculture, total power of agricultural machinery, agriculture value added, cotton harvest areas and oil harvest areas in the regression discontinuity model. The results in Table 5 are consistent with the results in Table 4. We find that the average harvest areas of rice and corn increase significantly in 2005, but the average harvest areas of rice, wheat and corn have no significant increase in 2006.

Based on the results in Table 4 and 5, we may conclude that we have no sufficient evidences to attribute the grain production discontinuity in 2005 to prospective counties' manipulation because their grain harvest areas (corn and rice)

increased significantly at the same time. It is important to note that the increase in grain harvest areas may be also due to over-report. In addition, the increase in grain harvest areas will not necessarily lead to discontinuity. The increase in grain harvest areas may only shift the distribution of grain production due to the continuous distribution of yield. However, we don't find significant increase in grain harvest areas of prospective counties in 2006. Therefore we can attribute the grain production discontinuity in 2006 to prospective counties' manipulation. Furthermore, based on the result of 2006, it has high possibility that the grain production discontinuity in 2005 may attribute to manipulation.

[Table 4 here]

[Table 5 here]

Robustness checks

In 2008, Ministry of Finance changed the subsidizing rules that the subsidy candidates remained the same as before in 2008. There is thus a lack of incentive for prospective counties to over-report their grain production. This change provides us with a perfect nature experiment to conduct a robustness check. This paper conducts McCrary (2008)'s density test for 2008, 2009 and 2010 in different intervals and at different break points. Table 6 presents the t-statistics for different intervals, and the results show that none of t-statistics is significant at 1 per cent significance level. In addition, Table 7 shows that all of the t-statistics are also insignificant at different break points. The results of Table 6 and Table 7 imply that there is no evidence to support the existence of discontinuity in grain production in 2008-2010. On the contrary, the results of regression discontinuity method indicate that the grain harvest

areas experience significant change in the same period. This result confirms that the variation in grain harvest areas will not necessarily lead to discontinuity.

Based on all of the above results, we may more confidently conclude that the grain production discontinuity in 2005 and 2006 may attribute to prospective counties' manipulation.

[Table 6 here]

[Table 7 here]

[Table 8 here]

The characteristics of manipulator

After confirming the existence of manipulation in grain production, we now shed some light on the characteristics of manipulators. First, we classify the counties as mid-western counties and eastern counties, and the results of McCrary (2008) density test in Table 9 suggest strong evidence consistent with manipulation for mid-western counties, but no evidence for eastern counties. Second, we classify the counties as the major grain production regions and non-major grain production regions. Based on the results in Table 9, we find evidence consistent with manipulation for the major grain production regions, but no evidence for non-major grain production regions. Last but not least, according to the criterion set by World Bank, we classify the counties with GDP per capital less than USD906 as low income counties and the others as middle and high income counties[4]. The *t*-statistics indicate that the null hypothesis is rejected at 5 per cent significance level for low income counties. However, the *t*-statistics is

[4] China is a developing country. GDP per capital of most of counties are less than USD 906. Therefore, we only classify the counties as two groups. In addition, we only have the data of GDP and population, thus we use GDP instead of GNI.

statistically insignificant for middle and high income counties. One possible explanation is that there is more incentive for low income counties and mid-western counties to get the subsidy because the fiscal revenue is very low for these counties.

[Table 9 here]

Conclusions

Manipulation of food production data could lead to catastrophic social and economic consequences, as many policies are made based on the data in modern society. The data manipulation of local officials could be driven by better performance reviews and promotion. However, the fiscal system in China also plays important rules for local governmental data manipulation. In order to ensure food security in China, China's central government started to heavily subsidize grain production. In 2005, China made a policy to subsidize the granary counties with average annual grain production more than 200 thousand tons between 1998-2002. The list of subsidizing counties would be adjusted in three years. This created an incentive for the counties with production slightly below the threshold to manipulate the data. The central government in 2008 changed the rules and fixed the subsidizing counties based on the data between 2002 and2006. The incentive for the data manipulation disappeared. This creates a natural experiment for us to test if the accuracy of statistical data in local governments reacts to the central government's fiscal distribution rules.

In this context, we propose two research questions. First, whether the granary county subsidy policy causes the prospective county to manipulate the grain production data to get subsidy. Second, we shed light on the characteristics of the counties which are likely to manipulate data. Based on McCrary (2008)'s density test,

we have strong evidences that county level grain production has a discontinuity at 200 thousand tons in 2005 and 2006. Furthermore, we use the regression discontinuity method to rule out the possibility that the grain harvest area expanding contributes to the grain production discontinuity. Finally, based on the results of McCrary (2008)'s density test and the regression discontinuity method, as well as the robustness checks, we attribute the grain production discontinuity to data manipulation, and we conclude that the granary county subsidy result in the prospective counties over-reporting their grain production. In addition, we found strong evidence consistent with manipulation in the major grain production regions, low income counties and mid-western counties.

The policy implication would be that the fiscal distribution rules of a central government should avoid data manipulation incentives in local governments, particularly should cut the linkage to the data which are self-reported by the local governments.

References:

Bernstein T. P. (2006) Mao Zedong and the Famine of 1959–1960: A Study in Wilfulness, China Quarterly, Vol. 186:421-445.

Burgstahler D., I. Dichev. Earnings management to avoid earnings decreases and losses. Journal of Accounting and Economics, 1997, 24: 99-126.

Chamon M. and I. de Carvalho Filho (2013) "Consumption Based Estimates of Urban Chinese Growth", IMF Working Paper WP/13/265. IMF, Washington D.C.Chen K. (2004) Fiscal centralization and the form of corruption in China. European Journal of Political Economy, Vol. 20(4):1001-1009.

Chen Xiwen. Review of China's agricultural and rural development: policy changes and current issues. China Agricultural Economic Review, 2009, 1(2):121-135.

Chen Yuyu, Ginger Zhe Jin, Naresh Kumar, Guang Shi. Gaming in air pollution data? Lessons from China. The B.E. Journal of Economic Analysis & Policy, 2012, Vol. 12, Iss. 3 (advances), Article 2.

Cheng, M.Y.. Boundary aware estimators of integrated density products. Journal of the Royal Statistical Society, 1997, 59(1):191-203.

Cheng, M.Y.. On boundary effects of smooth curve estimators (dissertation). Unpublished manuscript Series #2319, Institute for Statistics, University of North Carolina.

DiNardo, J.E., Lee, D.S.. Economic impacts of new unionization on private sector employers: 1984-2011. Quarterly Journal of Economics, 2004, 119(4): 1001-1044.

Fisman Raymond, Shangjin Wei. Tax rates and tax evasion: Evidence from missing imports in China. Journal of Political Economy, 2004, 112(2):471-496.

Fuller, F., Hayes, D., & Smith, D. (2000). Reconciling Chinese meat production and consumption data. *Economic Development and Cultural Change*, 49, 23–43.

Ghanem, D. and J. Zhang, 'Effortless Perfection:' Do Chinese cities manipulate air pollution data? Journal of Environmental Economics and Management, 2014, 68(2): 203-225.

Holz C. A. (2014) The Quality of China's GDP Statistics, China Economic Review, Vol. 30:309-338Huang Jikun, Xiaobing Wang, Huayong Zhi, Zhurong Huang, and Scott Rozelle. Subsidies and distortions in China's agriculture: evidence from producer-level data. The Australian Journal of Agricultural and Resource Economics, 2011,55(1): 53-71.

Huang Jikun, Yang Guolei. Understanding recent challenges and new food policy in China, Global Food Security, 2017,12: 119-126.

Jacob, B.A., Lefgren, L.. Remedial education and student achievement a regression-discontinuity analysis. Review of Economics and Statistics, 2004, 86(1): 226-244.

Kim Kwansoo, Jean-Paul Chavas. Technological change and risk management: an application to the economics of corn production. Agricultural Economics, 2003,29:125-142.

Lee P. K. (2000), Into the Trap of Strengthening State Capacity: China's Tax-Assignment Reform. China Quarterly, Vol. 164:1007-1024.

Lin J Y (1990) Collectivization and China's Agricultural Crisis in 1959-1961.Journal of Political Economy 1990 98:6, 1228-1252.

Ma, H., Huang, J., & Rozelle, S. (2004). Reassessing China's livestock statistics: an analysis of discrepancies and the creation of new data series. Economic Development and Cultural Change, 52, 445–473.McCrary Justin. Manipulation of the running variable in the regression discontinuity design: a density test. *Journal of Econometrics*,2008, 142（2）: 698-714.

Nakamura E., J. Steinsson and M. Liu (2016) "Are Chinese Growth and Inflation Too Smooth? Evidence from Engel Curves". *American Economic Journal: Macroeconomics* 2016, Vol.8(3): 113–144. NBSC, National Bureau of Statistics of China, China Statistical Yearbook, 2011. China Statistical Press, Beijing.

Parker E. and J. Thornton (2007) Fiscal Centralisation and Decentralisation in Russia and China. Comparative Economic Studies, Vol. 49(4)：514-542.

Saez Emmanuel. Do taxpayers bunch at Kink points? American Economic Journal: Economic Policy, 2010, 2(3): 180-212.

Smith, L., 2015. The great Indian calorie debate: explaining rising undernourishment during India's rapid economic growth. Food Policy 50, 53–67.

Wang, X., & Woo, W. T. (2011). The size and distribution of hidden household income in China. Asian Economic Papers, 10, 1–26

Wang Yangjie, Jikun Huang, Jinxia Wang, Christopher Findlay. Mitigating rice production risks from drought through improving irrigation infrastructure and management in China. Australian Journal of Agricultural and Resource Economics, 2017, 59:1-16.

Yi Fujin, Dingqiang Sun, Yingheng Zhou. Grain subsidy, liquidity constraints and food security—Impact of the grain subsidy program on the grain-harvest areas in China. Food Policy, 2015, 50: 114-124.
Yu Xiaohua, Abler David. Where have all the pigs gone? Inconsistencies in pork statistics in China. China Economic Review, 2014, 30:469-484.
Yu, X. and D. Abler, Matching food with mouths: A statistical explanation to the abnormal decline of per capita food consumption in rural China. Food Policy, 2016, 63:36-43.
Zinman Jonathan, Eric Zitzewitz. Wintertime for deceptive advertising? American Economic Journal: Applied Economics, 2016,8(1):177-192.
Zitzewitz Eric. Forensic Economics. Journal of Economic Literature, 2012,50(3): 731-769.

Table 1 the number and proportion of counties with grain production above and below the threshold

Year	Total number	<0.20 million tons		0.16-0.20 million tons		>0.20 million tons		0.20-0.24 million tons	
		Number	proportion	Number	proportion	Number	proportion	Number	proportion
2000	2045	1224	59.9	72	3.5	821	40.1	80	3.9
2001	1966	1210	61.5	77	3.9	756	38.5	75	3.8
2002	2045	1286	62.9	84	4.1	759	37.1	82	4.0
2003	2046	1345	65.7	71	3.5	701	34.3	66	3.2
2004	2042	1245	61.0	82	4.0	797	39.0	63	3.1
2005	2041	1220	59.8	81	4.0	821	40.2	66	3.2
2006	1960	1146	58.5	65	3.3	814	41.5	89	4.5
2007	2043	1185	58.0	67	3.3	858	42.0	81	4.0
2008	2040	1157	56.7	74	3.6	883	43.3	94	4.6
2009	2037	1141	56.0	73	3.6	896	44.0	77	3.8
2010	2041	1125	55.1	77	3.8	916	44.9	67	3.3

Table 2 t-statistics of McCrary density test for different intervals: break point 0.2million tons

Year		2003	2004	2005	2006	2007
0.17-0.23 million tons	h/b	4.55	5.56	4.90	5.22	4.24
	t-statistics	1.07	0.76	2.46***	2.36***	0.54
	h/b	22.39	23.28	16.50	15.42	20.42
	t-statistics	1.23	0.59	2.67***	2.56***	1.09
0.18-0.22 million tons	h/b	5.60	3.47	4.27	4.48	4.27
	t-statistics	1.44	0.69	2.21	2.45***	1.22
	h/b	14.64	11.62	18.17	9.50	14.60
	t-statistics	0.96	0.60	2.58***	2.55***	1.01
0.19-0.21 million tons	h/b	1.27	2.12	3.15	2.43	4.65
	t-statistics	0.46	0.45	1.55	1.89	0.84
	h/b	9.72	17.91	14.99	13.84	14.78
	t-statistics	0.06	0.88	1.65	2.44***	1.24

Notes: (1) ***indicates statistically significant at the 1 per cent for one-side critical t-statistic; (2) h denotes bandwidth, and b is bin size. Bandwidth and bin size in first row of each interval are from the automatic procedure suggested by McCrary (2008). Bandwidth in third row of each interval is also from the automatic procedure, but the bin size is subjectively adjusted to guarantee the value of h/b greater than 10.

Table 3 t-statistics of McCrary density test for different break points

Break points	0.19	0.195	0.2	0.205	0.21
2003	-0.01	0.90	1.44	0.78	-0.45
2004	0.34	-0.76	0.69	-0.23	0.06
2005	2.24	0.67	2.21	-0.21	-1.32
2006	1.86	-0.13	2.45***	0.69	0.44
2007	-0.14	-0.65	1.22	0.56	-1.29

Notes: (1) ***indicates statistically significant at the 1 per cent for one-side critical t -statistic; (2) Bandwidth and bin size are from the automatic procedure suggested by McCrary (2008); (3) The McCrary (2008) density test was conducted in the interval of 0.18-0.22 million tons.

Table 4 the results of regression discontinuity for grain harvest area

Variables	Rice harvest area: hectare		Wheat harvest area: hectare		Corn harvest area: hectare	
	2005	2006	2005	2006	2005	2006
Lwald100	6620	NA	NA	NA	7842*	NA
	(4676)				(4165)	
Lwald200	-1847	6590	66.69	497.3	-1826	2564
	(2165)	(4030)	(1078)	(2850)	(1662)	(2306)
Observations	1304	1304	1318	1318	1356	1356

Notes: (1) standard errors in parentheses, *, **, *** indicate statistically significant at the 10%, 5%, and 1%, respectively; (2) Lwald100 means Local Wald Estimator with the bandwidth of 100; (3) 2005 and 2006 mean the cut-off in regression discontinuity model; (4) the regression discontinuity model was conducted in the interval of 0.18-0.22 million tons; (5) The variation in observations mainly attribute to the different grain planting structure for counties.

Table 5 the results of regression discontinuity for grain harvest area with some control variables

Variables	Rice harvest area: hectare		Wheat harvest area: hectare		Corn harvest area: hectare	
	2005	2006	2005	2006	2005	2006
Lwald100	13500**	NA	NA	NA	8994*	NA
	(6300)				(5381)	
Lwald200	438.3	5139	-603.0	2040	-3228*	4551
	(2,474)	(3397)	(1286)	(3360)	(1908)	(2834)
Observations	1304	1304	1318	1318	1356	1356

Notes: (1) standard errors in parentheses, *, **, *** indicate statistically significant at the 10%, 5%, and 1%, respectively; (2) Lwald100 means Local Wald Estimator with the bandwidth of 100; (3) 2005 and 2006 mean the cut-off in regression discontinuity model; (4) the control variables include employment in agriculture, total power of agricultural machinery, agriculture value added, cotton harvest area and oil harvest area; (5) the regression discontinuity model was conducted in the interval of 0.18-0.22 million tons; (6) The variation in observations mainly attribute to the different grain planting structure for counties.

Table 6 t-statistics of McCrary density test for different intervals: a robustness check

Intervals		2008	2009	2010
0.17-0.23 million tons	h/b	4.59	7.45	6.42
	t-statistics	1.42	-0.04	0.51
	h/b	16.10	22.34	21.15
	t-statistics	1.88	0.08	0.78
0.18-0.22 million tons	h/b	5.53	4.23	3.02
	t-statistics	1.74	0.33	1.80
	h/b	15.29	14.38	10.11
	t-statistics	1.59	0.58	1.98
0.19-0.21 million tons	h/b	3.07	2.61	NA
	t-statistics	1.68	0.56	
	h/b	12.45	14.13	15.96
	t-statistics	1.54	0.43	1.69

Notes: (1) ***indicates statistically significant at the 1 per cent for one-side critical t-statistic; (2) h denotes bandwidth, and b is bin size. Bandwidth and bin size in first row of each interval are from the automatic procedure suggested by McCrary (2008). Bandwidth in third row of each interval is also from the automatic procedure, but the bin size is subjectively adjusted to guarantee the value of h/b greater than 10.

Table 7 t-statistics of McCrary density test for different break points: a robustness check

Break points	0.19	0.195	0.2	0.205	0.21
2008	-0.04	0.15	1.74	0.92	1.37
2009	-0.46	-0.30	0.33	1.26	-0.49
2010	-0.62	1.44	1.80	-0.23	0.60

Notes: (1) ***indicates statistically significant at the 1 per cent for one-side critical t -statistic; (2) Bandwidth and bin size are from the automatic procedure suggested by McCrary (2008); (3) The McCrary (2008) density test was conducted in the interval of 0.18-0.22 million tons.

Table 8 the results of regression discontinuity for grain harvest area

Variables	Rice harvest area: hectare		Wheat harvest area: hectare		Corn harvest area: hectare	
	2008	2009	2008	2009	2008	2009
Lwald100	8873	-25,679	-3037**	4676**	-2942**	4039**
	(7971)	(16353)	(1217)	(2173)	(1153)	(1947)
Lwald200	15954**	-13045	-9671***	4007	-5373***	5839**
	(7462)	(9044)	(1922)	(2470)	(1607)	(2465)
Observations	1443	1443	1457	1457	1495	1495

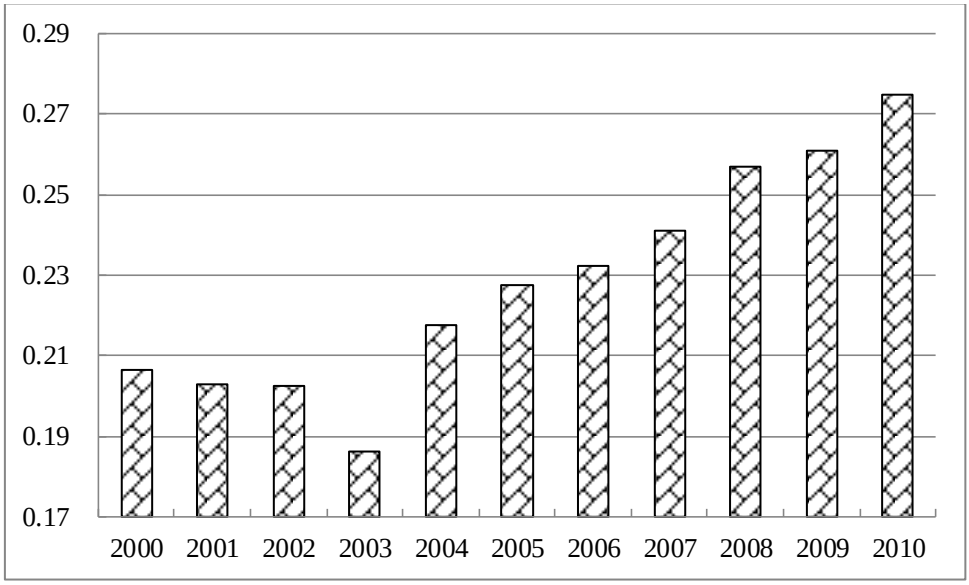
Notes: (1) standard errors in parentheses, *, **, *** indicate statistically significant at the 10%, 5%, and 1%, respectively; (2) Lwald100 means Local Wald Estimator with the bandwidth of 100; (3) 2008 and 2009 mean the cut-off in regression discontinuity model; (4) the regression discontinuity model was conducted in the interval of 0.18-0.22 million tons; (5) The variation in observations mainly attribute to the different grain planting structure for counties.

Table 9 t-statistics of McCrary density test for different characteristics

Characteristics		t-statistics
Regions	Central and western regions	3.010***
	Eastern regions	1.512
Major grain production region or not	Non-major grain production region	1.586
	Major grain production region	2.366***
Income	Low income	2.238[1]
	Middle and high income	1.890

Note: (1) ***indicates statistically significant at the 1 per cent for one-side critical t -statistic; (2) Bandwidth and bin size are from the automatic procedure suggested by McCrary (2008); (3) The McCrary (2008) density test was conducted in the interval of 0.18-0.22 million tons for both 2005 and 2006.

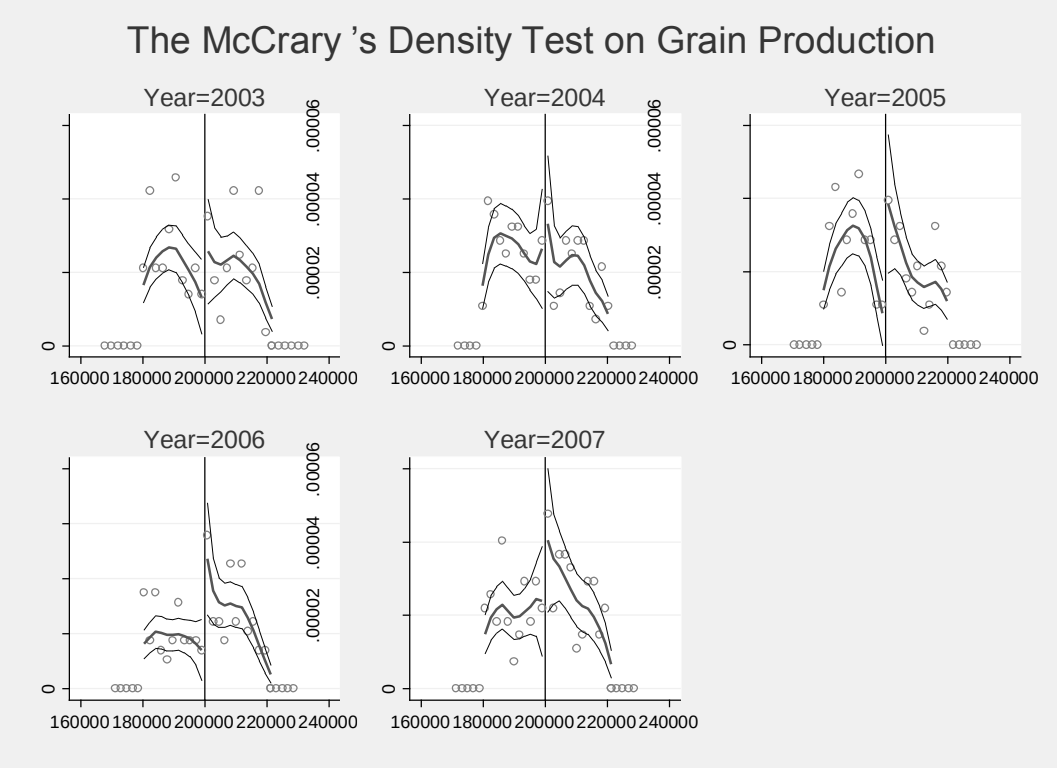
[1] t-statistic of low income counties is significant at 5 per cent significance level.



Note: the unit is million tons

Source: county level statistical yearbooks in each year

Figure 1 the average grain production from 2000-2010



Note: (1) X-axis denotes grain production; the unit is tons; (2) The McCrary (2008) density test was conducted in the interval of 0.18-0.22 million tons.

Figure 2 The McCrary (2008)'s Density Test on Grain Production for 2003-2007