User Demographics and Censorship on Sina Weibo

Wayne Kenney
Montclair State University
kenneyw1@montclair.edu

Christopher S. Leberknight
Montclair State University
leberknightc@montclair.edu

Abstract

This paper investigates the relationship between demographics and the frequency of censored posts (weibos) on Sina Weibo. Our results indicate that demographics such as location, gender and paid for features do not provide a good degree of predictive power but help explain how censorship is applied on social media. Using a dataset of 226 million weibos collected in 2012, we apply a binomial regression model to evaluate the predictive quality of user demographics to identify candidates that may be targeted for censorship. Our results suggest male users who are verified (pay for mobile and security features) are more likely to be censored than females or users who are not verified. In addition, users from provinces such as Hong Kong, Macao, and Beijing are more heavily censored compared to any other province in China over the same period.

1. Introduction

China has censored its internet since 1994 when it issued the first regulations that the internet could not be used to harm the interests of the state [1]. Censorship in China has grown in breadth and depth such that the Human Rights Watch declared in their 2020 World Report that, "China's government sees human rights as an existential threat."[2] Recent political events in China bring this into focus. Protests in Hong Kong are being stymied by the the Great Firewall, preventing effective coordination and communication [3]. China also exerts influence on international companies, such as forcing Apple to remove apps that it doesn't like from it's Chinese Apple Store [4], or spreading false information through Twitter [5]. These are just a few recent examples of China's influence on the internet.

China employs a widely distributed censorship policy that spans across social, political, and technological spheres. Recent political developments in China have sparked growing concerns regarding Internet freedom for citizens as many foreign tech companies withdraw from Hong Kong or refuse Chinese governmental requests to share their users data [6, 7]. The Chinese government expects servers hosting online content to censor users by whatever means they deem fit, as long as they accomplish the government's goals. As such, the specific mechanisms that are used vary across social media and newscasting platforms [8]. One effective method of censorship is to track heavy offenders with a large number of followers to facilitate the removal of objectionable content [9].

The increase and continuous expansion of Internet censorship practices by oppressive political agendas warrant further research to understand characteristics that trigger automated methods of information control.

2. Prior Work

Censorship in China is a well studied phenomenon because of the direct effect it has on freedom of expression, freedom from being oppressed by the government, and freedom of retribution by the government.

Prior research has sought to better understand what words are being censored. It was found that predictors for censorship were related to political scandals, the one child policy, housing policy, the pension system, political events, leaders of the communist party, officials names, content deletion itself, and profanity. [10]

Other studies found that negative sentiment, average number of idioms in each sentence, number of content word categories, number of idioms, number of complex semantic categories and verbs increase the probability of being censored. While positive sentiment, words related to leisure, reward and money decease the probability of being censored.[11] The theory being that posts with complex ideas have a higher idea of being censored.

It was also shown that the number of followers effect deletion rates and the dissemination of information. [9]

On a higher conceptual level, a seminal paper by King et al. showed that there was no statistical

URI: https://hdl.handle.net/10125/70944
978-0-9981331-4-0

difference between posts critical and supportive of the state (the **State Critique Theory**) when conditioning on high profile events (for instance Bo Xilai scandal). Rather posts that had the potential for citizens to meetup and form groups had a significantly higher probability of being censored than posts that did not, (the **Collective Action Potential Theory**). [12, 13].

The implication of King et al. is that the main aim of censorship is to prevent groups from forming. This makes sense in the broader context of Chinese history because it has a long history of increased censorship and tightened governmental controls immediately following riots or other incidents involving groups of people. The most famous example is the Tienanmen Square Incident, which arguably began the modern era of heavy Chinese censorship.

These papers constitute part of the core of the modern academic corpus of censorship in China. But surprisingly they fail to report basic demographics which is information that may be useful for a more robust understanding of the mechanisms behind censorship.

To our knowledge there is only one paper that tackles this issue. Bamman et al. computed deletion rates of users in each different province in order to estimate rates of censorship. A major limitation of their work is that it is impossible to tell if a post was deleted by the user for fear of governmental retribution, by governmental censors, or for other reasons.[14]

We overcome the limitations of Bamman et al. by using posts that are known to be censored. We also expand upon it by including gender, geographic location, and whether users have paid for premium features. We conjecture that in addition to having a large number of followers, user demographics may also be used to increase filtering relevancy.

3. Methods

The dataset in this paper is from the Open Weiboscope Data Access [15] project that is funded by the University of Hong Kong Seed Funding Program for Basic Research. This project's aim is to make publicly available censored posts from selected microbloggers in order to promote a free internet and academic research. The authors of this paper have no association with this organization, and did not personally collect the data.

The data was collected in 2012 from Sina Weibo, a Chinese analogue to Twitter. We chose this dataset because Sina Weibo was the largest online micro-blogging website at that time and because its large size ensures a good representation of the population.

In order to construct this dataset, a list of users who have more than 1000 followers was collected. The original list contained 350,000 users. This list was added to by random sampling. The dataset contains roughly 14 million unique users.

Every day a sample of users timelines was downloaded and compared to previous saved timelines. If a previous post is not on the current timeline, the post was either deleted by the user or censor.

The post was then searched for by Message ID. If the result is "This weibo does not exist." That means it was deleted by the user. If the result is "Permission Denied" then the post was removed by the censor.

The dataset contains 226 million weibos, 86 thousand of which are known to be censored. Each row of the dataset contains the weibo posted, and information about each weibo, including an alpha numeric user identification code that has been pseudo randomly generated to protect the identity of individuals [10].

Along with the above dataset, an ancillary dataset with information about each pseudo random ID was available. This includes three self reported demographics: province, sex, and verified status. Province is a categorical variable with 36 distinct levels, one for each province code plus one for "Other". Sex is male or female, and verified status is whether the individual has paid for extra service, including status showoff, extra functions, mobile and security features. This data was retrieved through use of the Sina Weibo Open API and reflects self reported data.

The data came in the form of 52 comma separated files, each containing a weeks worth of weibos, and their corresponding information. Each file was initially processed in order to collect and condense the relevant variable for this analysis, then condensed into a single large file containing the entire years worth of weibos. This was accomplished by the use of a processing cluster with high memory availability.

For each user ID, the number of times that user was censored over the course of 2012 was tabulated. The censored count variable was combined with the ancillary dataset, to give a single dataset that has each pseudo random ID with demographics for that individual and the number of times they were censored. The province variable contained 47 codes that matched with 36 provinces codes, with 11 ambiguous levels [16]. Out of roughly 14 million rows, there were 12 thousand ambiguous responses that didn't code to any of the province codes. These were coded to "Other".

Due to the fact that the vast majority of users were never censored, the censored count variable was collapsed into the binary variable censored or not censored. The censored level of this variable represents

the number of people that were censored at least once over the year, and the not censored level represents the number of people that were never censored over the same period.

For efficient processing, identical rows of the three demographics were tabulated and two separate columns were created, one for the number of times censored and one for the number of times not censored. This scheme condensed roughly 14 million rows into 144 rows.

3.1. Descriptive Statistics

A summary of the censored variable is in Table 3.1. This shows the variable collapsed into two levels, users that were never censored and users that were censored at least once over the period.

Table 3.1. Probability Of Being Censored

Posts Censored	Users	p(Posts Censored)
Zero	13808789	0.99919
One or Greater	11214	0.00081

A preliminary analysis of user demographics against counts shows some expected trends. Province, summed across sex and verified for conciseness, are shown in Table 3.2. The three highest are Hong Kong, Beijing and Overseas. In Table 3.4 the sex of censored users is shown, with males being much more likely to be censored. Table 3.5 shows the probability of verified users vs non verified users to be censored. Verified users are much more likely to be censored than non verified. And lastly, the interaction of sex and verified is shown in Table 3.3. It is not clear if sex and verified status is an interaction, if verified male users are different than male non verified users, or if female verified users are different than female non verified users.

Table 3.2. By Province

Province	p(Censored)
Hong Kong	0.003749
Beijing	0.003557
Overseas	0.001236
Taiwan	0.000988
⋮	⋮
Ningxia	0.000166
Qinghai	0.000123

3.2. Binomial Model

The form of the model equation was then determined. The response variables are Censored and Not Censored, and the predictors are Province, Male, and Verified. The response variable has two levels which suggests logistic regression modeled as binomial counts. The form of the regression equation is suggested by the descriptive statistics in Section 3.1. Province is additive and Sex and Verified status is an interaction.

Table 3.3. By Sex and Verified

Male	Verified	p(Censored)
False	False	0.000098
True	False	0.000408
False	True	0.030360
True	True	0.061107

Table 3.4. By Sex

Sex	p(Censored)
Male	0.001576
Female	0.000313

Binomial regression with a logit link function was chosen. We set C_i to be the number of censored users on row i, and T_i for the total number of users for row i. Also, M is categorical for male or not male (female). V is for verified or not verified. There is also categorical P_j for $j = 1, 2, \ldots, 36$ for each province/administrative district/other. The demographics for row i have β_m for Male, β_v for Verified, and β_{mv} for the interaction Male:Verified. For each Province, there is a β_{P_j} C is Binomial, with size T and probability λ, where λ is the logit transformed linear transformation of all our predictors.

$$
\begin{aligned}
C_i &\sim \text{Binomial}(T_i, \lambda_i) \\
\lambda_i &= \frac{1}{1 + e^{-X}} \\
X_i &= \beta_I + \beta_m M_i + \beta_v V_i + \beta_{mv} M_i V_i \\
&\quad + \beta_{P_1} P_{1i} + \beta_{P_2} P_{2i} \\
&\quad + \cdots + \beta_{P_{36}} P_{36i}
\end{aligned}
$$

This model is fit with Huber robust errors to account for heteroskedasticity through the R package 'robustbase'[17]. The city of Beijing was set to be the reference level or the categorical variable Province.

Logistic regression was chosen for two reasons. First, a statistical method that returns standard errors

Table 3.5. By Verified Status

Status	p(Censored)
Verified	0.050269
Not Verified	0.000223

was desired in order to establish the significance of the predictors. Second, the meaning of logistic regression coefficients are widely known which enhances the interpretability and communication of results.

4. Results

Table A.1 shows the regression summary. The main effects for Verified and Sex are highly significant while the interaction term for Verified and Sex is also significant. This means that while Sex and Verified have different values for each combination of the values of their levels.

The marginal effects for the interaction between Male and Verified are in Figure 4.1, and the marginal effect of Province is in Figure 4.2. The marginal effects summaries are in Table A.3 and A.2. While the regression summary shows the effect on the logit transformed Censored variable, the marginal effects show the probability of being censored at least once for a random member of the given category.

These marginal tables are different from the descriptive statistics in Section 3.1 because they show the probability of being censored after the effect of the other categories are removed.

Verified males have the highest probability of being censored at least once ($.04 \pm .002$, 95% CI). Females who are verified have roughly half that amount ($.018 \pm .001$, 95% CI). Unverified members have a much lower probability of being censored (Males $.0003 \pm .00002$, Females 0.00008 ± 0.000006, 95% CI). The highest location to be censored is Overseas ($.0061 \pm .0006$, 95% CI) followed by Hong Kong, Beijing and Macao ($.006 \pm .0006$, $.005 \pm .0002$, and $.004 \pm .002$, 95% CI's). The least censored province is Qinghai (0.0009 ± 0.0008, 95% CI).

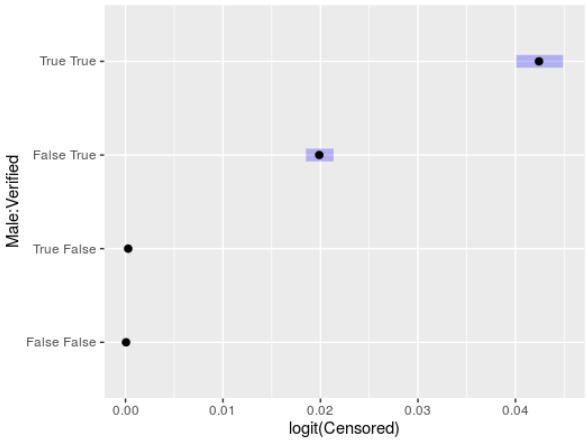


Figure 4.1. Male:Verified

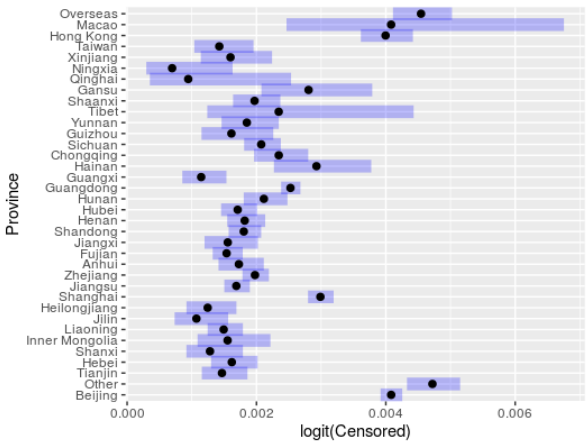


Figure 4.2. Province

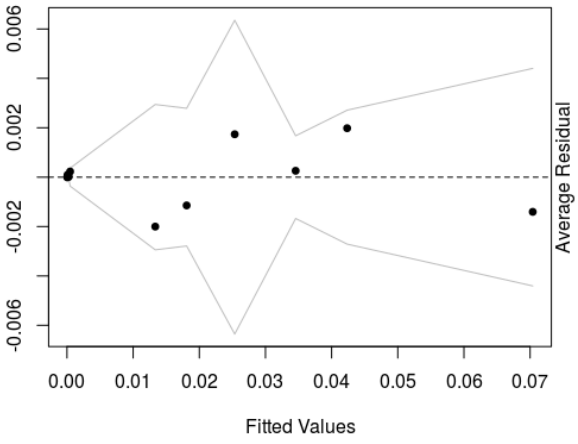
4.1. Model Validation

To validate these models for inference purposes, what is required is to ensure that there is no systemic bias or patterns in the residuals. To do this we follow the procedure given by Gelman and Hill.[18] We divide the fitted values into categories with an equal number of points. The particular number of values is somewhat arbitrary but Gelmen recommends the square root of the number of values. Rounded down, this is 12 for our dataset. These bins are then plotted as the average residual value by the average fitted for that bin. The expected bounds are $\pm 2\sqrt{p(1-p)/n}$, as p ranges as from 0 to 1. If more than 5% of the average residual by average fitted value for each bin lies outside the expected bounds, the model is suspect. We expect 95% of the points to lay inside the bounds because the above

formula for the bounds is a confidence interval normal estimate of a proportion at a 95% confidence level. In our case, we should expect 12*.05=.6, or no more than one point outside the bounds, preferably none.[18]

In Figure 4.3 we show the average binned residual by binned average fitted. For this model there are a large number of points on top of each other at zero fitted and zero residual. This was to be expected as that is true to the underlying data. There are no points outside the bounds, as evidenced by Figure 4.3

Figure 4.3. Residuals by Fitted



We also performed a leave one out cross validation. The decision boundary was chosen such that it maximized the product of precision and recall. The decision boundary that maximizes this product is .022. This low value is to be expected because an overwhelming majority of the underlying dataset were not censored. The confusion matrix is in Table 4.1. Precision is 0.056 and recall is .69 for the positive label Censored.

Table 4.1. Confusion Matrix for Boundary .022

Actual:Predicted	Not Censored	Censored
Not Censored	13465940	129092
Censored	3509	7705

5. Discussion

The precision and recall of the model is very low for the label Censored. This is to be expected with the very low predictive power of the categories, as can be seen in the descriptive statistics in Tables 3.2 and 3.3. Because of the very high imbalance, few categories, and low power of those categories, no classifier can produce a high precision value, given the predictors. We therefore reject this model for prediction, not for lack of fit or a better possible model, but because better predictive ability is not possible with this dataset. Even an overfit model would have poor predictive ability. For instance, the row with the highest probability of censorship is 11 %, making this an unobtainable upper bound for precision.

This model is still valid for inference, as Figure 4.3 shows no systematic bias. The interaction between sex and verification status is significant by Table A.1, so looking at sex and verified independently is not justified. Referring to the margins Figure 4.1, male verified users are most likely to be censored over the year 2012. It is possible that male verified users are more likely to post offensive things to the government. Female verified users have the marginal probability just under half as much as male verified users. Unverified users both male and female have a much lower marginal probability. It may be possible that more active users become verified and as such are more likely to be censored.

In Figure 4.2, it is shown that the locations most likely to be censored are Oversees, Macao, Hong Kong, Other, and Beijing. This makes sense as these areas are much more politically sensitive. Macao and Hong Kong has a very long history of protests and political divisiveness from the rest of China. On the other hand the province with the lowest probabilities of a user being censored is Qinghai and Ningxia, which has a history of Tibetan monk immolation, and Hui and Uighurs ethnic group tensions respectively.

While every province has some history of unrest, this analysis only shows the relationship between the self-reported demographics and the probability of a user being censored at least once. It only shows how any unrest is reflected in the number of people censored, and so is a better measure of how ubiquitous censorship is among these demographics, as opposed to the depth of or intensity of censorship.

Also, it may be possible that instead of these demographics correlating with qualities that make users more likely to make posts that will be censored, the censors filter by gender, verified status, and location to reduce the computational effort and sharpen the focus of who to censor. It is impossible to tell which of these two hypothesises are true by this dataset.

The results of data analysis on a sample can only be generalized to the population if the sample is large enough and the data generating process did not bias the sample. Out of the roughly 14 million users in the study, 350,000 were selected due to high follower count. The rest was a simple random sample. This may slightly bias the results towards users with more than 1000 followers.

6. Limitations of this Work

The main limitation of this work is the age of the dataset. The dataset was collected over 2012 and constitutes the censorship at that time. The paramount leader at that time was Hu Jintao, who many considered to be reserved. The nature and intensity of censorship took a turn in 2013 when Xi Jingping became paramount leader. This dataset does not reflect all the changes in censorship that was to happen over the next decade under Xi Jinping. But may serve as a way-point in understanding the demographics of censorship at that time. Future research will focus on investigating the evolution of censorship when a similar more recent dataset becomes available.

Acknowledgement

The work is supported by the National Science Foundation under Grant No.: 1704113, Division of Computer and Networked Systems, Secure Trustworthy Cyberspace (SaTC).The data processing cluster was supported by the National Science Foundation under Grant No. CNS 1625636.

References

[1] S. Council, "Regulations of the people's republic of china for safety protection of computer information systems," Feb. 1994.

[2] K. Roth, "China's global threat to human rights." https://www.hrw.org/sites/default/files/world_report_download/hrw_world_report_2020_0.pdf, 2020.

[3] J. Tarabay, "China's great firewall looms over hong kong as surveillance grows." https://www.bloomberg.com/news/articles/2020-06-16/china-s-great-firewall-looms-over-hong-kong-as-surveillance-grows, June 2020.

[4] A. Hern, "Apple removes two podcast apps from china store after censorship demands." https://www.theguardian.com/technology/2020/jun/12/apple-removes-two-podcast-apps-from-china-store-after-censorship-demands, June 2020.

[5] S. Dayaram, "Twitter fact-checks china official's post claiming coronavirus originated in us," May 2020.

[6] S. Pham, "Tiktok is leaving hong kong following controversial national security law." https://www.cnn.com/2020/07/07/tech/tiktok-leaving-hong-kong-intl-hnk/index.html, July 2020.

[7] J. Yeug, "China has passed a controversial national security law in hong kong.." https://www.cnn.com/2020/06/25/asia/hong-kong-national-security-law-explainer-intl-hnk-scli/index.html, July 2020.

[8] R. MacKinnon, "China's censorship 2.0: How companies censor bloggers," *First Monday*, vol. 14, Jan. 2009.

[9] P. D. P. A. C. J. R. Zhu, T. and D. S. Wallach, "The velocity of censorship: High-fidelity detection of microblog post deletions," August 2013.

[10] M. C. King-wa Fu, CH Chan, "Assessing censorship on microblogs in china: Discriminatory keyword analysis and impact evaluation of the 'real name registration' policy," *IEEE Internet Computing*, vol. 17(3), pp. 42–50, 2013.

[11] K. Y. Ng, A. Feldman, J. Peng, and C. Leberknight, "Linguistic fingerprints of internet censorship: The case of sina weibo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. 2, pp. 446–453, Association for the Advancement of Artificial Intelligence, April 2020.

[12] G. King, J. Pan, and M. E. Roberts, "How censorship in china allows government criticism but silences collective expression," *American Political Science Review*, vol. 107, no. 2 (May), pp. 1–18, 2013.

[13] G. King, J. Pan, and M. E. Roberts, "Reverse-engineering censorship in china: Randomized experimentation and participant observation," *Science*, vol. 345, no. 6199, pp. 1–10, 2014.

[14] D. Bamman, B. O'Connor, and N. Smith, "Censorship and deletion practices in chinese social media," *First Monday*, vol. 17, Mar. 2012.

[15] K. W. Fu, "Weiboscope open data (dataset)." https://hub.hku.hk/cris/dataset/dataset107483, 2017.

[16] "Province city code table." https://open.weibo.com/wiki/.

[17] S. Müller and A. H. Welsh, "Robust model selection in generalized linear models," *Statistica Sinica*, vol. 19, no. 3, pp. 1155–1170, 2009.

[18] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2017.

Appendices

A. Results

Table A.1. Logistic Regression

	logit(Censored)	SE
(Intercept)	-9.05	0.05
MaleTrue	1.62	0.05
VerifiedTrue	5.91	0.05
MaleTrue:VerifiedTrue	-0.83	0.06
ProvinceOther	0.15	0.05
ProvinceTianjin	-1.03	0.12
ProvinceHebei	-0.93	0.11
ProvinceShanxi	-1.16	0.17
ProvinceInner Mongolia	-0.97	0.18
ProvinceLiaoning	-1.01	0.09
ProvinceJilin	-1.34	0.19
ProvinceHeilongjiang	-1.19	0.16
ProvinceShanghai	-0.31	0.04
ProvinceJiangsu	-0.89	0.06
ProvinceZhejiang	-0.73	0.05
ProvinceAnhui	-0.86	0.10
ProvinceFujian	-0.98	0.08
ProvinceJiangxi	-0.97	0.13
ProvinceShandong	-0.82	0.07
ProvinceHenan	-0.81	0.08
ProvinceHubei	-0.87	0.08
ProvinceHunan	-0.66	0.08
ProvinceGuangdong	-0.48	0.03
ProvinceGuangxi	-1.27	0.15
ProvinceHainan	-0.33	0.13
ProvinceChongqing	-0.56	0.09
ProvinceSichuan	-0.68	0.07
ProvinceGuizhou	-0.93	0.17
ProvinceYunnan	-0.79	0.12
ProvinceTibet	-0.56	0.33
ProvinceShaanxi	-0.73	0.10
ProvinceGansu	-0.38	0.15
ProvinceQinghai	-1.47	0.50
ProvinceNingxia	-1.77	0.44
ProvinceXinjiang	-0.94	0.17
ProvinceTaiwan	-1.06	0.16
ProvinceHong Kong	-0.02	0.05
ProvinceMacao	-0.00	0.26
ProvinceOverseas	0.11	0.05

Table A.2. Marginal Effect of Sex and Verified

Male	Verified	p(Censored)	SE
False	False	0.0000766	0.0000033
True	False	0.0003144	0.0000102
False	True	0.0184836	0.0006523
True	True	0.0401515	0.0011171

Table A.3. Marginal Effect of Province

Province	p(Censored)	SE
Overseas	0.0061	0.0003
Hong Kong	0.0060	0.0003
Beijing	0.0050	0.0001
Macao	0.0041	0.0010
Other	0.0041	0.0002
Shanghai	0.0032	0.0001
Hainan	0.0031	0.0004
Gansu	0.0029	0.0004
Chongqing	0.0028	0.0002
Guangdong	0.0023	0.0001
Tibet	0.0023	0.0007
Hunan	0.0023	0.0002
Taiwan	0.0022	0.0003
Sichuan	0.0022	0.0001
Shaanxi	0.0021	0.0002
Zhejiang	0.0021	0.0001
Yunnan	0.0019	0.0002
Shandong	0.0019	0.0001
Henan	0.0019	0.0002
Jiangsu	0.0019	0.0001
Hubei	0.0018	0.0001
Anhui	0.0018	0.0002
Xinjiang	0.0017	0.0003
Guizhou	0.0017	0.0003
Hebei	0.0016	0.0002
Fujian	0.0016	0.0001
Jiangxi	0.0016	0.0002
Liaoning	0.0016	0.0001
Inner Mongolia	0.0016	0.0003
Shanxi	0.0015	0.0002
Tianjin	0.0015	0.0002
Heilongjiang	0.0013	0.0002
Guangxi	0.0012	0.0002
Jilin	0.0011	0.0002
Ningxia	0.0009	0.0003
Qinghai	0.0009	0.0004