Feasibility Study on Business Register Updating in China Based on Internet Data Sources

Ran Tao*
National Bureau of Statistics, Beijing, China – tomrand@sohu.com

Hengjun Huang
Lanzhou University of Finance and Economics, Lanzhou, China –hhj91@163.com

Abstract

Available data source for maintenance and updating of Business Register has some deficiencies on cost, timeliness and burden on data provider in China. This paper proposes an idea for the maintenance and updating of Business Register from the perspective of big data, which demonstrates the advantage of taking internet data sources as information of the maintenance and updating of Business Register from the point of view of participant behavior and data quality, then discusses the technical means to obtain basic information and geographic position of Business Register, and gives examples of applications.

Keywords: big data; business register; internet data sources; updating and maintenance.

1. Introduction

With the emergence of sensor technology and information dissemination methods including social media, data is accumulating and growing in an unprecedented form, speed and breadth. Big data phenomenon aroused great interest in academic organizations, government agencies and companies at home and abroad. The United Nations Economic Commission for Europe (2013) established a preliminary framework for big data definition, data sources and data integration of government statistics. Heerschap (2013) discussed Statistics Netherlands has used Internet data in price statistics and tourism statistics of traditional statistical process. In China, the National Bureau of Statistics (NBS) also gives full attention to and a positive response the phenomenon of big data.

Due to significant difference between big data processing mode and traditional statistical work flow, big data used in government statistics should include three progressive levels: firstly, using heterogeneous data sources to assist and supplement existing statistical workflow; secondly, using big data thinking and methods to transform existing statistical work processes; and thirdly, creating a new statistical method and indicator.

The construction of Business Register is the core and foundation of government statistics and one of Four Major Programs of statistics in China. This paper discusses the application of big data in updating and maintenance of Business Register.

2. Status quo of Business Register and data source

Business Register refers to a database including basic identification and main attribute information of all legal entities and industrial activity units and is the basis for economic and social management and use of statistical agencies in statistical investigation. Business Register has attracted attention from government statistics agencies at home and abroad. Liu & Fang (2012) mention that American, Japanese and French experience in establishment of Business Register is worth to draw lessons from. Since the mid 1990s, Chinese Business Register has achieved remarkable results in construction, especially has made substantial progress after the implementation of Four Major Programs.

The data sources used in updating of Business Register mainly include internal sources including economic census, statistical investigation of basic units and various professional statistical investigations and external sources including administrative records of various departments. The data from economic census provides comprehensive and accurate information for Business Register and updated every five years, and statistical investigation of basic units and various professional statistical investigations were used in updating of part of Business Register every month, quarter or six months.

As external sources to update Business Register, administrative records are the materials of added, changed and cancelled entities offered by relevant departments at and above the county level (including departments of institutional organization, civil affairs, taxation, industrial and commercial administration and quality supervision) to statistical department and updated every six months. From timeliness speaking, quinquennial census and semiannual updating of administrative records as well as monthly and quarterly partial updating of surveys are difficult to adapt to the changing social and economic environment. Especially for enterprises below designated size, these data sources can not immediately reflect the increase or decrease in basic units.

We draw a figure for the participants and dimension of data quality, mentioned in Fu & Tao (2007) and Huang & Fu (2009), in construction and updating of Business Register (see Figure 1). For participant behavior, the statistical agencies collect relevant information of basic units in various survey methods to update and maintain Business Register. The basic units are usually only data providers who are difficult to form the users of Business Register and not enthusiastic to provide information. Nonstatistical departments are both of data providers and potential users of Business Register, but Business Register established by statistical department may not meet their needs due to differences between departments. Although the statistical department initially established a system of updating and maintenance of Business Register, an additional burden will be largely imposed on the nonstatistical departments providing the statistical department with data because of recording standard and technology differences between departments and information exchange barriers in the actual work. From a cost perspective, statistical agencies need to invest more manpower and resources to obtain accurate information so as to enhance the enthusiasm of participants.

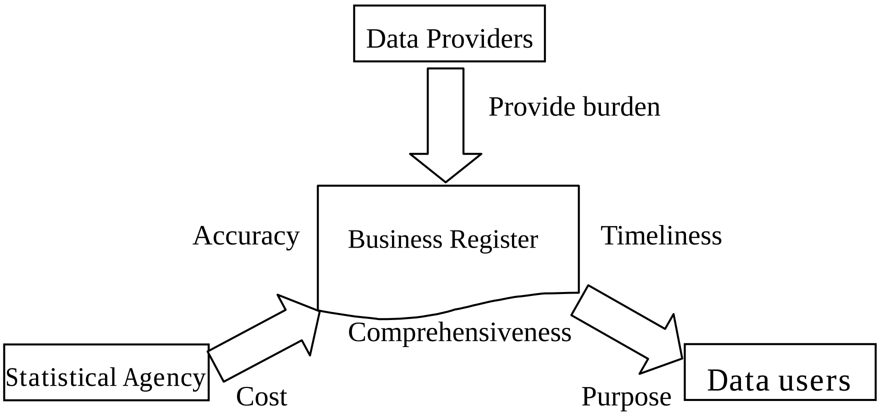

Figure 1 Participants and Dimension of Data Quality in Updating of Business Register

In summary, available data sources have some deficiencies on cost, timeliness and burden on data providers, thereby affecting the accuracy and comprehensiveness of Business Register to have a great impact on its maintenance and updating. In the era of big data, it is necessary to expand the data source and take heterogeneous Internet data as supplements of Business Register to alleviate the shortage of available data sources.

3. Exploration on updating of Business Register based on Internet data sources

With the rapid development of information technology, the Internet has become a carrier of a large number of information, and how to effectively extract and use this information has become an important issue. For ease of discussion, we only discuss catering companies and information required for Business Register includes basic information, attribute information and geographic information.

3.1 Determination of information scope

Because the basic units of Business Register are aimed at legal entities and industrial activity units, we first need to find relevant sites related to reliable information of these basic units. For catering industry, commercial websites of Location Based Service (LBS) can often provide the information to update of Business Register, such as FourSquare.com at abroad, and dianping.com, koubei.com, lashou.com and nuomi.com at home.

In order to extract the information used to updated Business Register, we need to build a Uniform Resource Locator (URL) List. LBS commercial website usually provides text list pages about business classification (e.g. regional classification, main business classification, etc), and the hyperlinks of its HTML source code include identification codes of these classification information, and the combination of these identification codes often constitutes “specific pages” directly or indirectly used in updating of Business Register.

3.2 Information extraction

On the basis of analysis on all “specific pages” containing updated information of Business Register, we need to extract useful information to form readable traditional bivariate table. This information is often included in the HTML code behind "specific pages", and the code is formed by some programming mechanisms and follows a certain pattern. We can use regular expression to match the data conforming to mode definition so as to filter out the remaining content and extract the desired information. Table 1 gives an example of a simple HTML code, which the basic information required for Business Register is listed in bold text.

Table 1 Example of HTML Code Containing Basic Information of Business Register

```
<html>
 ……
   <body>
       ……
     <p><a href="http://……">xx restaurant </a></p>
     <p> No. xx, Dazhong Lane, Chengguan District </p>
            <p> Telephone：09318888888 </p>
     ……
   </body>
</html>
```

In order to extract bold text from Table 1, we only need the following regular expression[1]:

. *> ([^> <] +?) <. *   (1)

(1) represents a logical rule that we need information in the bracket "()" and this information represents the shortest matching content between label ">" and "<" but excluding ">" and "<". In order to obtain " Telephone: 09318888888", we may only need to a series of numbers representing the phone number to be extracted by regular expression:

[^ 0-9] * ([0-9] +) [^ 0-9] *   (2)

If the rule mentioned above is used for all pages containing the updated information of Business Register, all updated information of Business Register will be extracted. Of course, the practical situation would be much more complex than that of Table 1. However, for using the regular expression for pattern matching, the basic approach is the same.

3.3 Information integration

In the 44th Session of the UN Statistical Commission in February 2013, Report of the Australian Bureau of Statistics on developing a statistical-geospatial framework proposed that government statistics departments should establish the ability to link socio-economic information with elements of spatial position to enhance the value of statistical information. For the construction of Business

[1] Most programming languages support regular expression, but vary in expression mode. The contents discussed in this paper are in line with the Perl language rule. For detailed description of the expression mode, see http://en.wikipedia.org/wiki/Regular_expression.

Register, it is necessary to include latitude and longitude coordinates reflecting exact geographic information in Business Register. The latitude and longitude coordinates are more consistent geographical representation requirements than text messages. From the follow-up application of Business Register, latitude and longitude coordinates can be the basis for the precise positioning of statistical work (such as GDP accounting).

In China, Baidu and Tencent provide geographic information services and free development interface. We can integrate the basic information from LBS commercial websites with the development interfaces of geographic information providers to get the latitude and longitude of units in Business Register. Table 2 shows an example of Baidu LBS open platform interface call, and interface call of other geographic information provider is made in a similar manner.

Table 2 Example of Baidu LBS Open Platform Interface Call Mode

```
  http://api.map.baidu.com/geocoder/v2/?
     ak= password
     &output=json
     &address= xx restaurant (or No. xx, Dazhong
Lane, Chengguan District)
     &city= Lanzhou
```

Because Baidu interface obtains the latitude and longitude information in the form of GET, Table 2 shows the interface call mode called URL (In fact, URL address should be a character string, but Table 2 branches it for clarity). The left of "?" shows the essential elements of URL, including protocol type, host name and specific resources directory. The right of "?" shows the parameters to obtain latitude and longitude coordinates of basic units in Business Register. The parameters exist in the form of key assignments and are separated by "&", in which "ak" is password key and need to apply to Baidu open platform, "output" is an optional output parameter that specifies the output as XML (default) or JSON structure, "address" is the name or address of basic units in basic information of Business Register, and "city" is the city where the basic unit locates. If these parameters are replaced with the information of basic units in Business Register and the URL in Table 2 is submitted to Baidu server, the JSON result is as follows:

Table 3 Longitude and Latitude Query Information (JSON)

```
{"status":0,
   "result":
   {"location":

    {"lng":103.7835750948,"lat":36.088498400678},
     "precise":1,
     "confidence":90,
     "level":""
     }
}
```

In Table 3, the bold figure is the latitude and longitude coordinates of "xx Restaurant" and Baidu also gives the reliable information of the address, in which "precise" of 1 refers to exact search and "confidence" represents credibility. Analysis of JSON format in Table 3 (such as using rjson package in R software) can integrate the latitude and longitude coordinates into Business Register. We can get the positioning information of many basic units by replacing the parameter contents of "address" and "city".

3.4 Information screening

Through the steps mentioned above, we can get a lot of data that can be used to update Business Register. Of course, we also have to screen and verify data on the basis of access to information .

One idea is mutual corroboration of data sources. Because of messy nature of the Internet, lack and error of the data itself and in the processing procedure is normal, which requires using a variety of data sources to complement and verify each other. For example, we can collect data from LBS websites to find and make up the deficiencies in basic information and attribute information used in updating of Business Register, and can use the development interfaces of several location service providers for mutual authentication of geographic information, and compare Internet data with the existing information of Business Register on this basis.

Another idea is mutual corroboration of business information and consumer information. For example, birth and death of basic units can be judged according to the time of the release of information by businesses and the time of the evaluation by consumer, if a business is not clicked or reviewed by consumer in a very long time, the business may cease to exist. The main business and operating scale of a business can also be judged according to the label and consumption price and even review added by consumers.

4. Example

To illustrate the technical feasibility of updating idea mentioned above of Business Register, we carry out preliminary integration of data collection for a few LBS commercial websites to integrate the geographic information system of the location service provider. Data collection involves four districts in Lanzhou (Chengguan District, Qilihe District, Anning District and Xigu District) and includes basic information and geographic information. The data collection gets 14,725 catering enterprises from four districts in Lanzhou after excluding missing, abnormal and repeated data and those that cannot obtain a valid latitude and longitude. The result is a regional distribution map of catering industry (Figure 2).

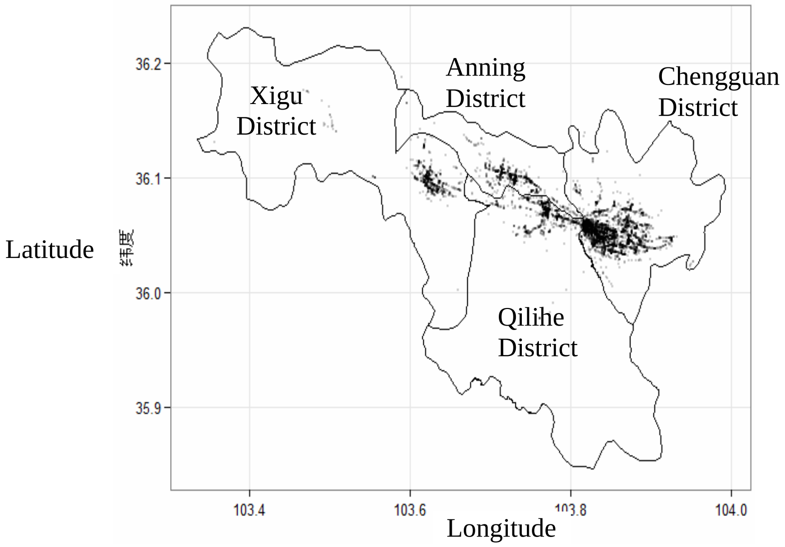

Figure 2 Spatial Distribution of Catering Industry in Four District of Lanzhou

In Figure 2, we have mapped the four districts of Lanzhou. Because of centralized distribution of catering enterprises in Lanzhou, the black dots are the location corresponding to catering enterprises, which is mapped by the latitude and longitude data.

5. Conclusion

Of course, implementation of this approach need discuss many details. Firstly, on a technical level, there is a problem in data interface because of difference between the Internet data and the data source view of traditional Business Register and difficulty of indicators in direct correspondence. Their effective integration needs some means of models and repeated practice. Secondly, in terms of the normative level, a discussion is required for whether the data are assets, whether the approach of data collection would violate the rights of LBS commercial websites and pose a threat to privacy of businesses and consumers, and how statistical agencies regulate the use of such information. Finally, we believe that Internet information as updating channels of Business Register is of great value. Statistical agencies cooperating with LBS commercial websites and location service provider to develop and share this valuable asset may well be a practical way.

References

UNECE (2013). What Does "Big Data" Mean for Official Statistics?. http://www1.unece.org/stat/platform/display/hlgbas.

Heerschap, N. (2013). Internet as a New Source of Information for the Production of Official Statistics, Experiences of Statistics Netherlands. Hong Kong.

Ma, J. T. (2013). Exploration and application of big data in government statistics. Beijing, China Statistics Press.

Xu, Y. F. (2014). Retrospective and Prospective View on the Construction of Basic Unit Database in China. Statistical research. 31(2).

Liu, J. P., &Fang Y. L. (2012). Comparative Study on the Construction and Development of Basic Units(Enterprises Registration) among China, the U.S. and Japan. Statistical research.29(4).

Fu, D. Y., &Tao, R. (2007). Discussion on Quality Cost of Statistical Data of the Government. Statistical research. 24(8).

Huang, H. J., &Fu, D. Y. (2009). Discussion of Survey Quality Dimensions. Statistical research. 26(11).

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No.71301033) and the National Social Science Foundation of China (No.14CTJ009).