The Stata Journal (2020)
20, Number 4, pp. 805–811 DOI: 10.1177/1536867X20976313

Extracting Chinese geographic data from Baidu Map API

Yuan Xue
School of Management
Huazhong University of Science and Technology
Wuhan, China
xueyuan19920310@163.com

Chuntao Li
School of Economics
Henan University
Kaifeng, China
chtl@henu.edu.cn

Abstract. In this article, we describe the two new commands `cngcode` and `cnaddress`, which can be used to link Chinese addresses to locations defined by their longitudes and latitudes through Baidu Map API v3.0 (http: // api.map. baidu.com), an online map and navigation system widely used in China. `cngcode` transfers Chinese addresses to locations, whereas `cnaddress` does the opposite. These two commands make it easier with Stata to deal with addresses and locations in China.

Keywords: dm0104, cngcode, cnaddress, China, location, Baidu Map

1 Introduction

Geographic information, because of its fundamental nature, has been playing an increasingly important role in scientific research. All statistical software programs need to address geographic data, especially when converting from human languages to geographic locations or vice versa. This is why there are several commands to address geographic location issues. Those commands include `gcode` (Ansari 2015), `geocode` (Ozimek and Miles 2011), `geocodehere` (Heß 2015), and `geocodeopen` (Anderson 2013). Those commands help users to transfer between addresses and locations using different electronic maps, including Google Map, Yahoo Map, HERE Map (Nokis), and MapQuest. However, most of the above maps are, by default, not designed for Chinese language users. Chinese addresses cannot be processed properly and updated timely. Also, some of those electronic maps are not accessible in China because of country security and ideology. And these map services are usually not displayed in Chinese even for Chinese addresses. Because China is the world's most populous country, there is an ever-increasing population in the field of scientific research who needs to process Chinese addresses or geographic locations, which calls for new commands in Stata. In this article, we present two new commands, `cngcode` and `cnaddress`, that can fill this gap with the application of Baidu Map API.

Similar to Google Map, Baidu Map is an online map provided by Baidu Co. Ltd., a company that focuses on search-engine services for Chinese language users. Baidu Map provides online map and navigation services. It covers more than 400 cities and thousands of counties and districts. Information can detail a newly built flyover, a one-way lane with little transportation, and a country road used only by pedestrians.

Most Chinese rely on its navigation services while driving or walking around to find a small restaurant in an unfamiliar place or even around their own neighborhood. With the help of Baidu Map, people can easily find the locations of a restaurant, a bank, a parking lot, a gas station, and the like.

On April 23, 2010, Baidu offered open map API to developers for free. Since then, it has provided both a JavaScript API and a web-service API. With those applications, we can extract the longitude and latitude of a Chinese address and convert a location into the corresponding address in Chinese. On June 18, 2019, Baidu updated the API to version 3.0 to optimize services and provided completely new geocoding, which is used by `cngcode` and `cnaddress`. Users who would like to stay on Baidu API 2.0 should use `chinagcode` and `chinaaddress` (Li and Xue 2016) instead of these two new commands.

2 Baidu Map API key

Before you use these two commands to access Baidu Map in Stata, a Chinese mobile number is required to apply for a Baidu Map API key. The API key is an official permission to use the Baidu APIs. To apply for this permission, users must first use the Chinese mobile number to register for a Baidu account.

After the registration, users can log on to Baidu Map open platform (http://lbsyun.baidu.com) to apply for the Baidu Map API key. In this step, an applicant needs to key in his or her name, mobile number, and email address and agree to a declaration on certain legal issues to use the platform. The procedure is straightforward, but it takes several days for users to finally get their API key after submitting the application form online.

A typical Baidu Map API key is an alphanumeric string. Both commands rely on this API key. Thus, users have to explicitly specify it. Suppose the user already has a Baidu Map API key, which is, say, `CH8eakl6UTlEb1OakeWYvofh`; then, the `baidukey()` option must be specified as `baidukey(CH8eakl6UTlEb1OakeWYvofh)`.[1]

3 The commands

3.1 Overview

Both commands rely on Baidu Map. Because the source code of the Baidu Map website is UTF-8 encoded, `cngcode` and `cnaddress` require Stata 14 or above.

When using `cngcode` to translate a Chinese address to a longitude- and latitude-defined location, users can specify the address in two ways, either separate mode or full-address mode. In the separate mode, address information is broken into several variables, including province, city, district, and the street name and number. In the full-address mode, there is only one string variable, which encompasses all the above information.

1. The API key used here is for illustration only.

Sometimes, we have both a separated address and an all-in-one address. In this case, users can even specify both. However, when both are used but only one of the addresses can return a meaningful location, `cngcode` will pick the one which works.[2] If both work but return slightly different locations, the default choice for `cngcode` is to use the location yielded from the separated address. Users can change this priority with the `ffirst` option, which will return the location from the full address if the location of the full address and separate address is different. When you use `cnaddress`, you will get the province, city, district, street, full address, and a semantic description of the location defined by longitude and latitude.

The accuracy of the commands depends on the accuracy of Baidu Map, which also depends on the way you specify the addresses. If the address is not specific—say, we specify only a university's name as an address for a university with several campuses in a certain city or province—Baidu Map may give the location of only one of the campuses, which may not be the one we are interested in.

3.2 Syntax of cngcode

We have a sample dataset, `example.dta`, with addresses in Chinese. Variable names are as follows:

- `Prov`: A string in Chinese, this refers to the name of the province where a company is located.
- `City`: A string in Chinese, this refers to the name of the city where a company is located.
- `Dis`: A string in Chinese, this refers to the name of the district or county where the company is located. A typical Chinese city is divided into districts and counties.
- `Address`: A string in Chinese, this address may include address components besides the province, city, and district names. It normally includes information about the street name, street number, building name, floor of the building, and zip code. In instances where some parts are missing, `cngcode` still works, but the location may not be that accurate.
- `FullA`: A string in Chinese, this may encompass all the information in `Prov`, `City`, `Dis`, and `Address` into one single long string. However, the full address may not be exactly as the separated address because of the different descriptions of one given address.

Let's now look at our dataset:

```
. use example.dta
```

2. This happens when the data are not uniformly complete. For example, for some observations, the full address is available, whereas for other observations, separated addresses are available.

Given the above information, `cngcode` can get the location in three ways:

1. Separate address, where users supply address parts with different variables. The syntax is as follows:

```
. cngcode, baidukey(CH8eakl6UTlEb1OakeWYvofh) province(Prov) city(City)
> district(Dis) address(Address) longitude(lon1) latitude(lat1)
Address in Obs 4 is missing or incorrect, no location extracted
. list lon1 lat1
```

	lon1	lat1
1.	114.38794	30.47953
2.	114.42009	30.518951
3.	114.37184	30.543798
4.	.	.

2. Full address, where users supply address with an all-in-one address line, here the variable `FullA`. The syntax is as follows:

```
. cngcode, baidukey(CH8eakl6UTlEb1OakeWYvofh) fulladdress(FullA)
> longitude(lon2) latitude(lat2)
Address in Obs 3 is missing or incorrect, no location extracted
. list lon1 - lat2
```

	lon1	lat1	lon2	lat2
1.	114.38794	30.47953	114.38794	30.47953
2.	114.42009	30.518951	114.4199	30.513473
3.	114.37184	30.543798	.	.
4.	.	.	114.41643	30.406685

3. As we can see in the above, different descriptions of one location may cause different results; missing address information will cause missing values of longitude and latitude. Users could combine 1 and 2 in a single line of the command. The syntax is as follows:

```
. cngcode, baidukey(CH8eakl6UTlEb1OakeWYvofh) province(Prov) city(City)
> district(Dis) address(Address) fulladdress(FullA) longitude(lon3)
> latitude(lat3)
. list lon1 - lat3
```

	lon1	lat1	lon2	lat2	lon3	lat3
1.	114.38794	30.47953	114.38794	30.47953	114.38794	30.47953
2.	114.42009	30.518951	114.4199	30.513473	114.42009	30.518951
3.	114.37184	30.543798	.	.	114.37184	30.543798
4.	.	.	114.41643	30.406685	114.41643	30.406685

In the combined address mode, as specified in both the separated address and the all-in-one address, `cngcode` will try to calculate two locations for both the separated

address and the full address. If one of them does not yield a meaningful location, it will be ignored automatically, and `cnaddress` will report the location from the one that has yielded the meaningful address.

However, if both addresses yield meaningful locations, the default choice is to report the location generated from the separated address, which is normally more accurate than the all-in-one full address. Users can use the option `ffirst` to report the location generated from the all-in-one full address instead. The full command is specified as follows:

```
. cngcode, baidukey(CH8eakl6UTlEb1OakeWYvofh) province(Prov) city(City)
> district(Dis) address(Address) fulladdress(FullA) ffirst
> longitude(lon4) latitude(lat4)
. list lon1 - lat2 lon4 lat4
```

	lon1	lat1	lon2	lat2	lon4	lat4
1.	114.38794	30.47953	114.38794	30.47953	114.38794	30.47953
2.	114.42009	30.518951	114.4199	30.513473	114.4199	30.513473
3.	114.37184	30.543798	.	.	114.37184	30.543798
4.	.	.	114.41643	30.406685	114.41643	30.406685

3.3 Syntax of cnaddress

1. Compared with the syntax for `cngcode`, that for `cnaddress` is more straightforward because the input is simple. Here we use the two variables we previously obtained, `lat4` and `lon4`, which refer to the latitude and longitude of locations. A simple command to get the corresponding Chinese address is as follows:

   ```
   . keep lon4 lat4
   . cnaddress, baidukey(CH8eakl6UTlEb1OakeWYvofh) latitude(lat4) longitude(lon4)
   ```

   The output from the above command includes a separate address flavor, including province name, city, district, street, and a full all-in-one address.

2. Different types of coordinates, such as WGS-84, GCJ-02, and BD-09, have different latitude and longitude to a given location. We can specify the type of coordinate in the option `coordtype()`, which is `bd09ll` by default.

   ```
   . keep lon4 lat4
   . cnaddress, baidukey(CH8eakl6UTlEb1OakeWYvofh) latitude(lat4)
   > longitude(lon4) coordtype(wgs84ll)
   ```

   Options for `cnaddress` help to specify the names of the output variables, and they are self-explanatory. Thus, we will ignore this part to save space.

4 Conclusion

Before the release of these two commands, it was difficult for Chinese Stata users to process Chinese geographic information. `cngcode` and `cnaddress` fill this gap, thanks to the Baidu Map API. These two commands provide users with the convenience for accessing Chinese geographic information and, without any doubt, will attract more researchers to the Stata-user community.

Using longitude and latitude, we can also use the Baidu Map's navigation system to determine the traveling time between two locations by air, train, or car (Li, Xue, and Zhang 2019). Such information could be useful for researchers in transportation, education, the environment, and economics. Besides, there is still a lot to do in the future. We hope to see more researchers devote their time and effort to this area to promote more in-depth research in certain areas with the use of Stata software.

5 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 20-4
. net install dm0104     (to install program files, if available)
. net get dm0104         (to install ancillary files, if available)
```

6 References

Anderson, M. L. 2013. geocodeopen: Stata module to geocode addresses using MapQuest Open Geocoding Services and Open Street Maps. Statistical Software Components S457733, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457733.html.

Ansari, M. R. 2015. gcode: Stata module to download Google geocode data. Statistical Software Components S457969, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457969.html.

Heß, S. 2015. geocodehere: Stata module to provide geocoding relying on Nokia's Here Maps API. Statistical Software Components S457969, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458048.html.

Li, C., and Y. Xue. 2016. chinagcode: Stata module to geocode Chinese addresses. Statistical Software Components S458242, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458242.html.

Li, C., Y. Xue, and X. Zhang. 2019. cntraveltime: Stata module to extract the time needed for traveling between two locations from Baidu Map. Statistical Software Components S458603, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458603.html.

Ozimek, A., and D. Miles. 2011. Stata utilities for geocoding and generating travel time and travel distance information. *Stata Journal* 11: 106–119. https://doi.org/10.1177/1536867X1101100107.

About the authors

Yuan Xue is a PhD student in accounting at the Huazhong University of Science and Technology in Wuhan, China.

Chuntao Li is a professor of finance at the Henan University in Kaifeng, as well as a professor of finance at Zhongnan University of Economics and Law in Wuhan, China.