MJ Health Research Foundation

MJ Health Resource Center

Technical Report

Geocoding

Jackie Yuan-Chieh Chuang

MJHRF-TR-05

2016/05/18

Geocoding
Jackie Yuan-Chieh Chuang

I. Foreword

The development of Geographic Information Systems (GIS) allows us to view research data not only in terms of planar dimensions, but also the geospatial dimensions. In order to allow MJ health data to be used in geospatial research, the MJ Health Resource Center established a geocoding database. The goal of setting up such database is to promote geospatial health research projects by combining the geographic information with health data on behalf of the future data users.

II. Information Source

Every time MJ members received physical examinations, they provided their address information for mailing purposes. MJ member's address information were the kept in the database. Between 1994 and 2014, over 1,536,964 addresses were stored in the database, including some missing values and very small amount of PO boxes in the early days of data storage (0.32%).

III. Geocoding Tools

Currently, the most widely used platform for geographic information is Google Map. Thus, Google Map API was used to convert all MJ member's addresses into longitude and latitude. The longitude and latitude coordinates were produced by the World Geodetic System 1984 (WGS84) used by Google Map, which can be directly or indirectly used on various geographic information platforms.

IV. The Geocoding Process

Google provides its geocoding service on a web page, which allows users to enter a single address string and returns a HTML formatted coordinate data, thus it is difficult to convert multiple addresses at once. To make batch conversion possible, a VB module on excel was develop. If the system failed to acquire the longitude or latitude coordinates for a given address, part of texts that may hinder geocoding process such as floor level, apartment number, neighborhood, town, etc., were manually removed from the address string, then conversion was repeated.

V. Geocoding Results and Verification

A total of 1,529,476 addresses were successfully converted to longitude and latitude via Google Map API, which accounts for approximately 99.5% of the total number of addresses. In order to verify the accuracy of the geocoding process, reverse geocoding was performed on 1000 randomly selected coordinates. Existing online resources was used for batch-processing purposes.

After receiving the addresses via reverse geocoding, they were split into three categories – counties, townships and streets, and compared with the original based on each category. For the county comparison, out of 1000 addresses, 6 addresses did not match the original, and all 6 were from Kinmen. Further examination on Google Map revealed that the new coordinates were in fact close to Kinmen, but for the purpose of verification, cases like these were still considered mismatch. Out of the last 1000 checked addresses, 970 of them were consistent with the original, meaning the percentage of accurate geocoding on the county level is 97%.

For the township comparison, out of 1000 coordinates, 958 townships generated via reverse geocoding remained consistent with the original, which means geocoding accuracy is 95.8%. Some inconsistencies occurred as a result of addresses being located right at the junction of two townships, and a slight shift in the geocoding process led to change of township. Even so, the reverse geocoding still generated addresses that were quite close to the original, therefore the actual accuracy rate should be better than 95.8%.

For the street comparison, out of 1000 checked addresses, 880 street or road names matched the original, meaning accuracy rate is at 88%. However, some addresses were located on intersections, for example, between Zhongxiao East Road and Fuxing South Road. In this case, original road name is Zhongxiao East Road, but after the geocoding and reverse-geocoding, however, the road was changed to Fuxing South Road. Though the actual coordinates generated by geocoding were not far off from each other.

VI. Personal Privacy Protection

In order to protect the personal privacy of MJ members, the original longitude and latitude coordinates generated by geocoding will not be directly released to the data user. Instead, coordinates will be processed to prevent it from being accurately reverse-geocoded back to the actual address. The difference between raw and processed coordinates will still be within acceptable range for the purposes of research and analysis.

As a result of increasing awareness on the importance of personal privacy, developed countries have created statistical areas as the basic unit for data collection, analysis and publication. In recent years, the Ministry of the Interior has also completed the construction of “The Minimum Statistical Area” (1), taking into account socioeconomic indicators, terrain, administrative boundaries, etc., in order to design a stable unit boundary for time series analysis after time conversion. In the future, MJ Health Database will organize the geographic information into minimum statistical area to not only protect MJ member’s personal privacy, but also provide researchers with better resources for geospatial analysis.

VII. Conclusion

After accumulating 20 years’ worth of addresses of MJ members, incomplete or inaccurate information remained as a result of typographical errors, format conversions or changes in definitions for certain administrative areas. In the future, MJ Health Resource Center will work to enhance the accuracy of address information stored in the database. On the other hand, geocoding will also be done on an annual basis to equip MJ Health Database with the most complete and up-to-date geospatial information.

References

1. Statistical Department, Ministry of the Interior, 2016, Statistical Area Access Services. Retrieved from http://moisagis.moi.gov.tw/moiap/match/system_common.cfm. Accessed on 2016/5/18.