International Cartographic Association
Association Cartographique Internationale

# Visualizing and Evaluating the Geocoding Uncertainty in Social Sensing for Disaster Management

Debayan Mandal [a, *], Binbin Lin [a], Bing Zhou [a], Mingzheng Yang [a]

[a] *Department of Geography, Texas A&M University, rohan_debayan@tamu.edu, Binbin Lin, linbinbin_gis@163.com, Bing Zhou, bing_zhou_barrett@hotmail.com, Mingzheng Yang, ymz2020@tamu.edu*

* Corresponding author

**Keywords:** Geocoding Uncertainty, Social Sensing, Twitter, Disaster Management, Geo-visualization

**Abstract:**

As the usage of GPS-enabled portable devices has become an indispensable part of people's daily life, big user-generated geospatial data collected by apps on those devices, e.g., location-based social media, provide a new avenue to observe human behaviours and reveal socioeconomic characteristics. These geospatial big data, referred to as social sensing data play an increasingly important role in understanding human-space-environment interactions, which can be applied in different fields, such as crisis management, disaster relief, political sway, religious and economic trends, and most recently, combating the Covid-19 crisis.

To understand and visualize the raw data obtained, these are attached to geographical factors like location of origin, user details, subject of such information – to evaluate and glean information. Most social sensing research determines the location of each record through enabled GPS (reported by the device), address mentioned in the body of the data, or the user-filled profile location. For procuring this type of data from social media sites, many researchers turn to Twitter, given its large and ever-growing user base compared to other platforms as the general populace indulges a lot in social media. Recently in 2019, however, due to less and less usage of the precise geotagging feature, the popular social media site like Twitter has decided to call off the feature to make for a more effortless experience for the users while protecting their privacy. In the absence of precise geotagged tweets, the next approach to geolocate the information is to check if there is any kind of addresses mentioned in the tweet or use the user profile location as the origin of the tweet – leading to some algorithmic uncertainties in analysing those data, producing biased or even misleading results. Hence, it needs to be known how accurate it is to replace the precise geotags with the user profile location in text format.

The purpose of this study is to quantify the accuracy of such geocoding method- to check if using user profile locations to locate the tweet is accurate – based on different cartographic boundaries, such as country, state, county, and examine if the geocoding uncertainty would exacerbate the modifiable area unit problem (MAUP). This research studies the tweets during Hurricane Harvey, the most devastating coastal hazard in the U.S.'s history, as there were a lot of geotagged tweets during its occurrence, providing a larger sample space to obtain more robust results. There are three objectives of this study: (1) to obtain Twitter data relevant to Hurricane Harvey and process it for geotagged tweets, alongside developing a visualization framework for accuracy discrepancies at multiple spatial-temporal scales; (2) to visualize the distribution of error distance for country boundaries from the geotagged location to the user profile location; (3) to visualize percentage agreement of locations of these two locators based on smaller administrative boundaries (e.g., state, counties).

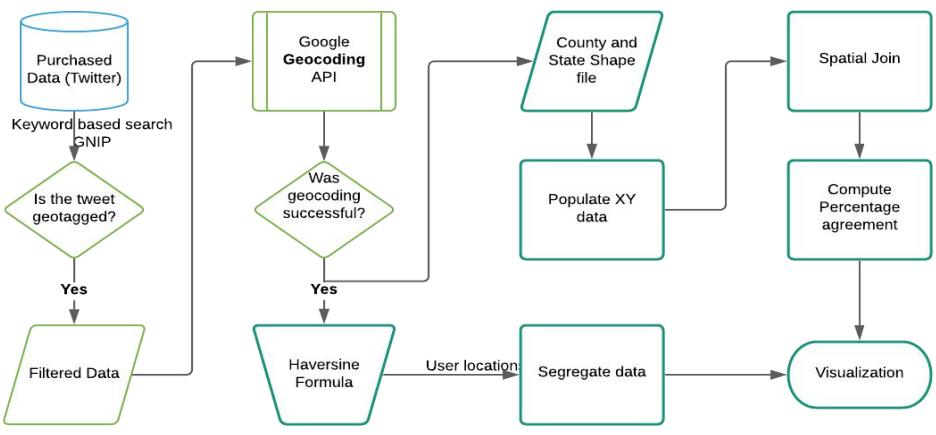


Figure 1. Methodology of data collection, processing, and analysis.


https://doi.org/10.5194/ica-abs-3-192-2021 

A total of 47 million tweets were purchased, which were already filtered by disaster-related keyword-based search. Each collected tweet was checked for geotags (500k tweets) and only those having it were kept. These were then geocoded to a location using Google geocoding Application Programming Interface (API). The error distance between the geotags and the geocoded user location was calculated using the Haversine formula. The error distances were visualized per country if the number of tweets was sufficient to avoid the small number problem. Focusing on the USA, these tweets were then checked if both locators lied within same administrative boundaries to find the accuracy for studies geared to such (Figure 1).

Results show that on a country level, all countries having considerable number of tweets showed similar patterns. Tweets are highly reliable to have its geotagged locations be replaced with user profile locations. As such, for brevity, the error distribution of locators in USA is presented with y-axis in logarithmic scale (Figure 2). It is clear that majority of the data lies in the first two bins – 0km (exact match) and 0 – 240 km, respectively (~27% of the total data each).

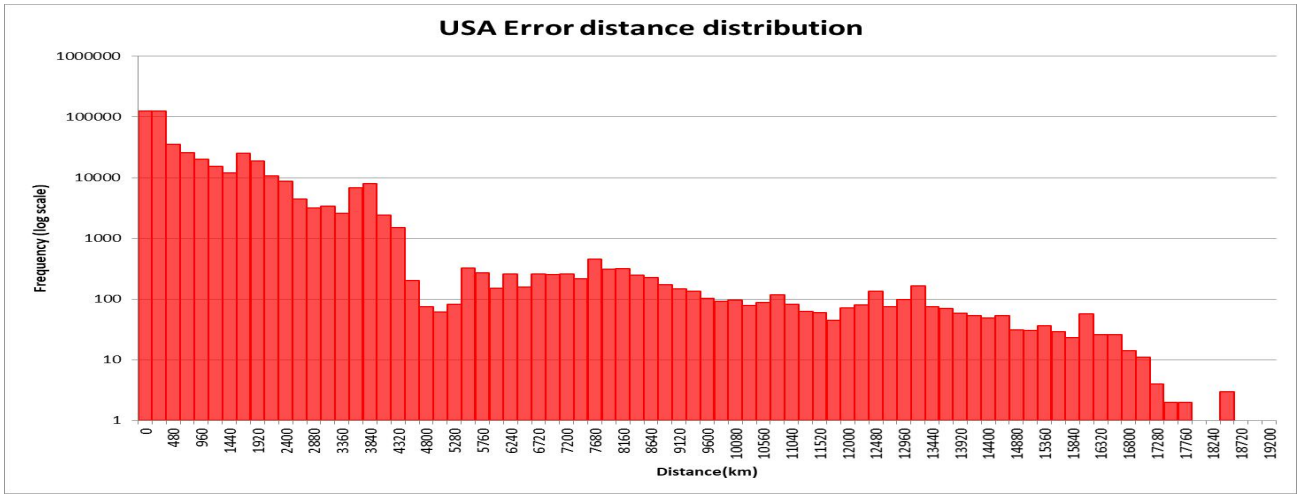


Figure 2. Error Distance Distribution for USA, along with enlarged view of first two bins. (Y-axis: Log scale)

For state and county levels, as Figure 3 shows, accuracy levels drop sharply. The state of Texas has the highest sample space since it is the prime affected zone and shows the highest accuracy. This goes on to show that closer to impact zones user locations would be interchangeable – however if one moves further away, the accuracies vary (Figure 3). The county level accuracies, however, severely drop and the distribution is lop-sided to below fifty percentage accuracy. This highlights a severe problem in interchanging the locators on a county level study or effort. For instance, highly time sensitive efforts like rescue operations or disaster relief distribution operations would get displaced and would present a false image of severity if frameworks and tools for analysis are based on this.

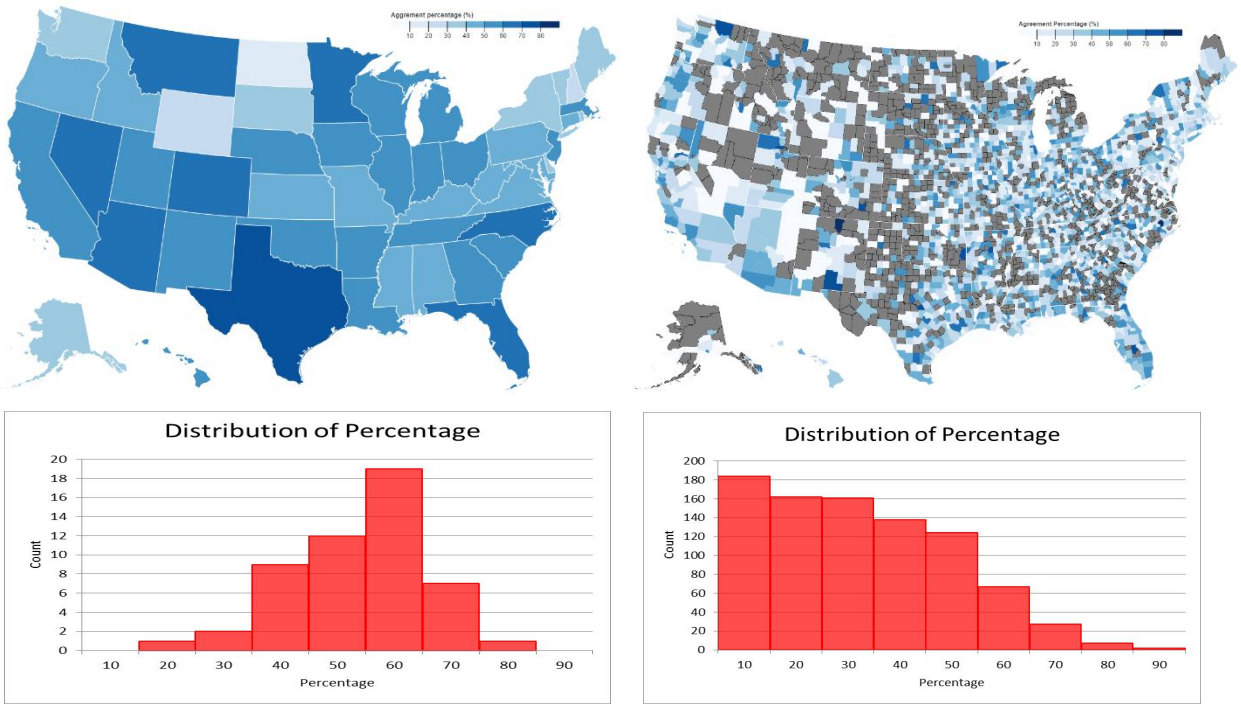


Figure 3. State and County accuracy for USA (above: Choropleths; below: Histograms)

Abstracts of the International Cartographic Association, 3, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy.
https://doi.org/10.5194/ica-abs-3-192-2021 | © Author(s) 2021. CC BY 4.0 License.

Findings from this project show that data from social media used during disasters can be used depending on the locational constraints and threshold limit of errors would be subject to each unique study type– while this study shows a keystone that can be referenced to and based upon while making such decisions.

Abstracts of the International Cartographic Association, 3, 2021.
30th International Cartographic Conference (ICC 2021), 14–18 December 2021, Florence, Italy.
https://doi.org/10.5194/ica-abs-3-192-2021 | © Author(s) 2021. CC BY 4.0 License.