CrowdZIP: A System to Improve Reverse ZIP Code Geocoding using Spatial and Crowdsourced Data (Demo Paper)

Tunaggina Subrina Khan
George Mason University
tkhan10@gmu.edu

Anowarul Kabir
George Mason University
akabir4@gmu.edu

Dieter Pfoser
George Mason University
dpfoser@gmu.edu

Andreas Züfle
George Mason University
azufle@gmu.edu

ABSTRACT

Zoning Improvement Plan (ZIP) Codes provide a sub-division of space. Interestingly, the ZIP code area polygons for different data sources do not match, resulting in uncertainty for a range of services that rely on such data. This paper presents a system that employs traditional classification methods to map a given spatial coordinate to a distribution of ZIP-codes using various public available ZIP-code maps as predictors, and using the (not publicly available) United States Postal Service (USPS) map as an authoritative ground truth. We show that large sets of microblog data, from which we extract potential ZIP-codes, can significantly improve classification accuracy despite the noise of such data. The demonstrator allows users to select locations on a map of Orlando, FL, view the resulting distribution of ZIP-codes predicted for this location, compare the results to the ground-truth, and view the microblogs that have enriched the result. A focus will be on showing that the signal present in large, noisy, and 99.99% unrelated microblog data can indeed be used to improve reverse ZIP code geo-coding.

CCS CONCEPTS

• **Information systems** → **Geographic information systems**; *Specialized information retrieval.*

KEYWORDS

Geocoding, Reverse Geocoding, ZIP Codes, Location Based Services, Microblog Data, ZIP Code Classification

ACM Reference Format:
Tunaggina Subrina Khan, Anowarul Kabir, Dieter Pfoser, and Andreas Züfle. 2019. CrowdZIP: A System to Improve Reverse ZIP Code Geocoding using Spatial and Crowdsourced Data (Demo Paper). In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359362

1 INTRODUCTION

A Zoning Improvement Plan (ZIP) Code is a five digit number assigned to every address in the United States. In publicly available datasets, ZIP Codes are represented as polygons, although ZIP Codes correspond to address points assigned by the United States Postal Service (USPS). There has been an increasing understanding in the literature that these point-to-polygon interpolations can result in erroneous spatial analysis results. The USPS does not maintain or release the geographic boundaries of the ZIP Codes, though some USPS facilities create their own ZIP Code area maps for public interest. USPS is not obligated to report changes to ZIP Codes in any formal way. The fact that many different representations of ZIP Code polygons have been generated, none of which are from an authoritative source (USPS), potentially leads to critical problems for those conducting scientific spatial analysis [1], [2]. A thorough study of the impact of ZIP Code uncertainty on many different applications can be found in [3].

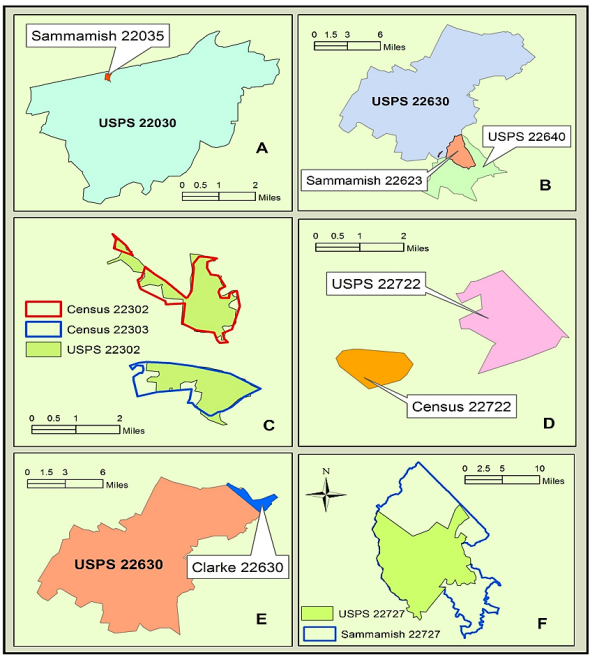


Figure 1: Deviation of ZIP Code boundaries across datasets

To illustrate these problems, Figure 1 shows examples of the deviations in size, shape and position of the ZIP Codes in four ZIP Code polygon datasets (USPS, Sammamish, Census, Clarke). Consider the example of Figure 1A, which has the Sammamish ZIP Code 22035 location contained in the 22030 USPS ZIP Code polygon. Moreover, there is no corresponding 22035 polygon in the

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359362

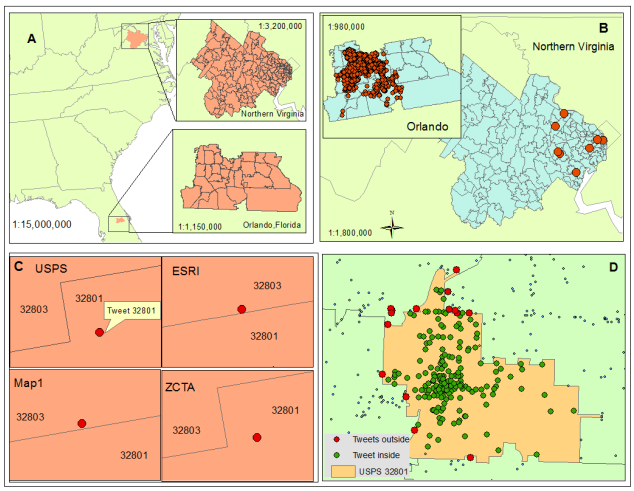


Figure 2: (A)Study area, (B)Tweet locations clipped with USPS map, (C)location of a tweet according to maps, (D)Tweet locations outside corresponding ZIP Code. (Best Viewed in Color)

USPS dataset. Similarly, the 22623 Sammamish polygon is actually a part of USPS 22630 and 22640 polygons (Figure 1B). Figure 1C shows that the USPS ZIP Code 22302 polygon is split into two different polygons (22302,22303) in the Census dataset. In Figure 1D, ZIP Code 22722 polygons have a different extent in the USPS and Census datasets, respectively. In Figure 1E, the 22630 USPS ZIP Code polygon is almost 40 times larger than the corresponding Clarke dataset polygon. Finally, Figure 1F shows that the 22727 Sammamish ZIP Code polygon is 300km larger than the respective USPS polygon.

In this demonstration, we showcase the result of our preliminary work ([4]) to combine multiple, publicly available maps to obtain a consensus answer. Our work additionally harnesses the wisdom of the crowd, by employing publicly available volunteered geographical information to improve the ZIP Code classification. Our proposed system scans millions of geo-tagged tweets from the public Twitter API for five-digit numbers. No further data cleaning and verification is performed, such that the semantics of many of these numbers do not correspond to ZIP Codes. Yet, using kNN-classification, the majority of such "CrowdZIPs" will frequently map to the correct ZIP Code, as shown in our initial experimental results [4]. Combined with existing map data, these TweetZIPs can successfully be used as a tie-breaker to help ZIP Code classification. This demonstration will allow users to use our system to see how unstructured and mostly unrelated Twitter data can be leveraged to improve GIS applications such as reverse ZIP Code geocoding.

2 RELATED WORK

Previous work [5] has studied the errors associated with commonly used US census ZIP Code Tabulation Areas (ZCTAs). Post Office Box addresses also cause problems when these employ ZIP Codes for geocoding rather than street addresses ([6]). The work of [7] showed problems with delineating service areas for broadband communication when combining ZIP code polygons with census block boundaries. Several studies mentioned the problems of inconsistency between ZCTA and ZIP Code boundaries ([8], [9]). For example, [9] showed that ZCTA boundaries were inconsistent with the scale of data collected at ZIP Code level. [10] compared different geographic units of observation such as ZIP Codes, census tracts and blocks and concluded that the ZIP Code is the least accurate geographic unit. While these previous studied have explored the effect and the magnitude of uncertainty in ZIP Code data, none of these works have proposed any solution to reducing the uncertainty. A first approach to combine multiple data sources to obtain consensus ZIP Code classification to alleviate the problem of this uncertainty was presented in [4]. The goal of this demonstration is to showcase the results of this work, and to show how data harvested from the crowd can support reverse geocoding.

3 CROWD-SUPPORTED ZIP CODE REVERSE GEOCODING

In the following, we want to address the reverse geocoding ZIP Code problem, i.e., given a point location, what is its associated ZIP code. For the scope of this work, we focus on the regions of Orlando, FL and Northern Virginia as shown in Figure 2A.

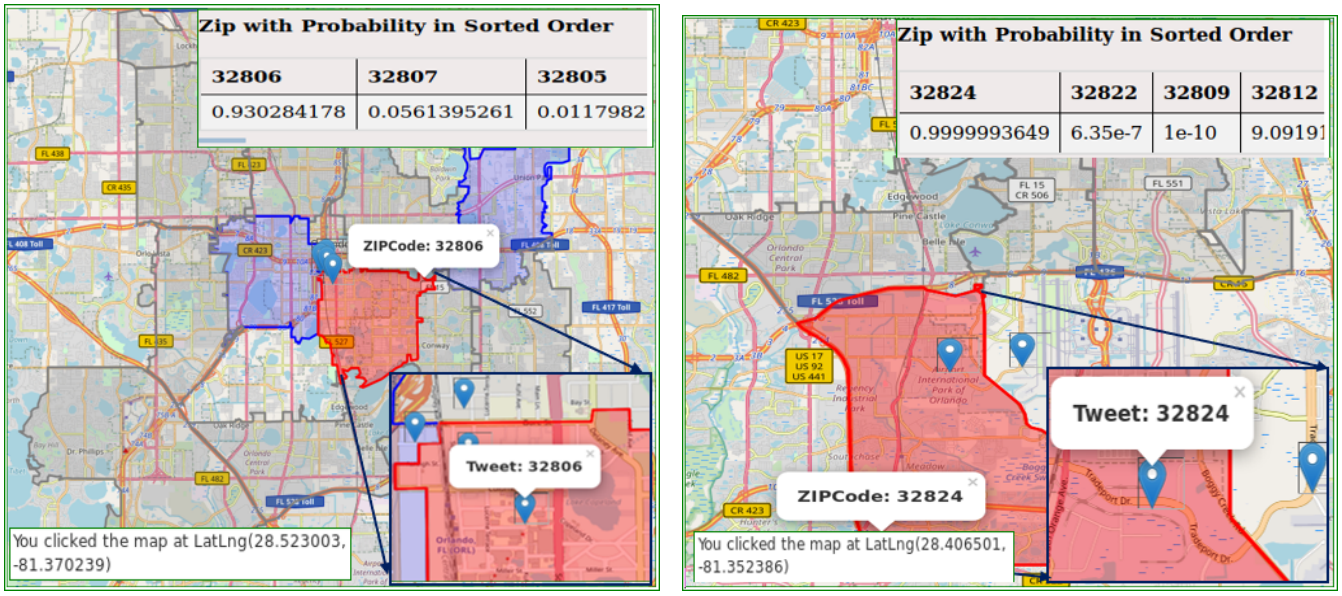
Definition 3.1 (ZIP Code Reverse Geocoding). Let $L = (long, lat)$ be a location defined by longitude and latitude. The problem of ZIP Code reverse geocoding is to map any location L in the united states to a ZIP Code according to the USPS ground-truth data set.

To solve this problem, a variety of ZIP Code polygon maps have been designed and published, e.g.,[11–13]. However, the datasets do not agree on the spatial extent of each and every ZIP Code area (cf. Section 1). Our approach to solving this problem and being able to utilize a range of dataset is to train a model that considers all datasets and learns which one to "trust" in different situations. In addition, we want to enrich our ZIP Code classification using the wisdom of the crowd. Users living in an area can be considered "local experts" and are likely know the true ZIP Code of a specific location.

The ZIP Code maps used for this experiment, were created by USPS, ESRI (both, for 2017), the US Census Bureau (ZCTA - for 2016) and from another vendor (Map1). The USPS dataset is considered ground truth. Figure 2A shows the study area.

3.1 Overview

Our demonstrator approaches the problem of classifying ZIP Codes for a given location in three steps (cf. [4] for details). Step 1 simply retrieves polygons from multiple data sources to obtain *"Map-ZIPs"*. Next, Step 2 uses Twitter data as a social media data source, consisting of geocoded locations along with ZIP Code information to obtain a *"Crowd-ZIPs"*. Using both Map-ZIPs and Crowd-ZIPs, our demonstrator uses a Bayesian classification approach to learn the reliability of these sources in different scenarios in Step 3. Data provided by the United States Postal Service (USPS) for the regions of Northern Virginia, and Orlando, FL is used as ground truth for training the model. The rationale of using a Bayesian learning approach is that for each class (ZIP Code), the model can fit a different model. Thus, for one ZIP code, the model may learn that a specific Map-ZIP is most reliable, while Crowd-ZIP is not, but for another ZIP Code, the model may learn that the Crowd-ZIP improves the predictive power of the model. The following section briefly describes our simple approach of obtaining ZIP code candidates from Twitter microblogs.



(a) High Uncertainty (b) Low Uncertainty

Figure 3: Demonstration of ZIP Code Reverse Geocoding.

3.2 Crowd-ZIPs

By using the public Twitter API [14], we obtained more than 35 millions of geocoded micro blogs (from February 2014 to July 2016).

Users cannot provide ZIP Code information in their tweets. Thus, to relate ZIP Codes to tweets, we scan the tweet text for 5-digit numbers using a regular expression. This approach yields a total of 18,000 tweets containing a five digit number related to our two study areas. Intuitively, we expect that this naïve approach of detecting ZIP code candidates will incur a large number of false positives, i.e., tweets that contain a five digit number that does not actually correspond to a ZIP Code. For example, a tweet may contain the message "@neu52285 i'm step up my game i'v been slackin", which is clearly not a ZIP Code (but rather a user name). In contrast, a tweet such as "#trespasser at king st sanitarium av 32803 #orlando" may very well refer to the local ZIP Code. To address this problem and to reduce the number of false positives, we remove tweets containing five digit numbers that do not correspond to any real ZIP Code, i.e., compare them to the USPS dataset. For any other five digit number, we assume that it corresponds to a ZIP Code. This approach reduces the number of valid ZIP Codes to 1,800 in the two studied regions. The distribution of these "Crowd-ZIPs" is shown in Figure 2B. Since we have many more Crowd-ZIPs for Orlando, we limit our demo to this area.

To motivate our problem, consider Figure 2C, showing different maps, and a ZIP-annotated tweet. We see that for this particular location, Twitter is able to identify the ground-truth ZIP 32801 (provided by the USPS map), while disagreeing with the ESRI map, and the "Map1" map. Figure 2D shows all Crowd-ZIPs for this area. Here, green dots correspond to tweets that contain the correct 32801 ZIP Code in their text corpus, while also having a location inside the ground-truth area. Red dots corresponds to instances of 32801 Crowd ZIPs that are outside of the ground-truth area. Small dots indicate Crowd-ZIPs that contain numbers other than 32801. This example shows that the wisdom of the Twitter crowd, despite being extremely noisy, is able to capture the ground-truth ZIP code regions very well.

A more detailed formalization of our approach to extract Crowd-ZIPs from Twitter microblogs, and further details on how to train a Bayesian model that considers Crowd-ZIPs and Map-ZIPs can be found in [4].

4 DEMONSTRATION DESCRIPTION

Our framework to be presented at ACM SIGSPATIAL 2019 will allow conference attendees to use a graphical interface to select locations, view probabilities of ZIP Codes for the given location, and explore the microblogs that support this decision. Figure 3 shows screenshots of this framework.[1] Initially, the demonstrator will allow users to select a location in Orlanda, Florida, USA (as this is the only area for which we have authoritative USPS data and a large number of Crowd-ZIPs). Upon selecting a location in Orlando, the geo-coordinates of the selected location are fed to our reverse ZIP Code geocoding algorithm and the resulting probability distribution of ZIP Code polygons (of the ground truth USPS ZIP Code boundary map) is shown (in the top right corner in Figures 3(a) and 3(b)). The results are color coded as polygons on the map to illustrate the location of different polygons and their probabilities. For example, in Figure 3(a), ZIP Code 32806 has a probability of more than 93% to be the correct result, indicated by the red color. In this example, two more ZIP Codes have a probability of more than

[1]Screenshots were altered for better paper presentation. ZIP Code probabilities are shown in a table outside of the map. Only one tooltip label is depicted at any time and depends on mouse pointer location. Zoomed in areas are added as inset maps.

1%, and a number of ZIP Codes polygons are shown having non-zero probabilities of less than 1%. In contrast, the location selected in 3(b) yields a single ZIP-code with more than 99.99% probability, showing that the system is highly confident in the correctness of this result. Hovering the mouse of a polygon shows the user the corresponding ZIP Code.

In addition to ZIP Code polygons and their corresponding result probabilities, our system also return information about the microblogs that enriched this result. Crowd-ZIPs that were considered in the result are shown on the map (blue location markers in Figures 3(a) and 3(b)). Users can click on these markers to show the candidate ZIP Code found in this tweet, as shown in the inset maps at the bottom right of Figures 3(a) and 3(b). Finally, the text corpus of each such tweet is shown in a separate view. This is to illustrate that, in most cases, the corresponding five-digit number found in a tweet indeed corresponds to a ZIP Code, and if not, the false ZIP Code is likely to be far from the current location, thus having minor effect on the Bayesian inference

The system utilizes a Web-based client-server model. Following a location selection on the map, a POST request is sent to the server. The server runs the prediction procedure and the output, a set of probability distributions with corresponding ZIP Codes, is sent back to the client as a key-value pair in JSON format. For each (ZIP Code, Probability) pair, a colored polygon is created on the map representing probabilities. The software is available on GitHub [15].

5 CONCLUSIONS AND FUTURE DIRECTIONS

The demo is proof of concept towards a framework to automatically label the ZIP code of a given location in the United States, a problem that is highly important for applications that employ reverse geocoding. Traditionally, applications use publicly available datasets to solve this problem. Yet, as has been shown in [3], these datasets differ significantly from each other.

This work is an effort towards addressing this problem by considering a consensus model which is training based on multiple such publicly available datasets, while employing a Bayesian learning approach to learn, given a ground-truth data set for supervision, the reliability of each source. We in addition utilize the wisdom of the crowd for ZIP Code identification by using nearby ZIP Code mentions in microblog data. Furthermore, the proposed approach allows to assess the reliability of ZIP Codes returned by the system. In cases where different ZIP Code polygon maps disagree, and microblogs disagree further, the resulting uncertainty is quantified in the result. This is particular important, as the true domain of a ZIP Code, which we assumed to be defined by the USPS in this work, may depend on applications and opinions of users. Thus, it is paramount that the uncertainty, is quantified in the reverse geocoding process and reported to the user.

Yet, there are many open research directions. First, our crowd-sourced ZIP-data set is very small, containing only 16,000 geo-tagged and ZIP Code enriched microblogs. We want to expand this study to other sources of data. Yet, our first studies have shown that this is not as easy as it might seem. Most of the geotagged ZIP Code data on the web, e.g., www.TripAdvisor.com, only contains geocoded ZIP-Codes using one of our data sources (e.g., data form ESRI). Thus, using such data will not yield any additional information, as these data sets will always yield the same independent variable as directly using the source data. Thus, we need to look into data sets where the ZIP-Codes are provided by local individuals (rather than being geocoded by a system). We plan to look into textual comments annotated with images published on Flickr. Furthermore, we will also look at Open-Street-Map (OSM). In OSM, the challenge will be to identify ZIP-Codes that are crowd-sourced (annotated by users), rather than automatically geocoded.

We want to extend our work to the entire contiguous United States. We hope to obtain a larger ground-truth data set from USPS after publishing these first results. We expect similar results for other parts of the US, as our study areas already include an urban area with a high tweet density (Orlando), as well as a suburban area with a rather low tweet density (Northern Virginia). Yet, a thorough experimental evaluation with more ground-truth data will answer this question with more confidence.

Finally, another open problem is the run time of our solution. Currently, our approach employs no indexing support for searching nearest ZIP-enriched tweets for a location. Once our database of ZIP-enriched tweets becomes larger, we will employ a spatial index such as an R*-tree [16] and use efficient kNN retrieval algorithms ([17]) to achieve scalable run-times.

REFERENCES

[1] K. M. Beyer, A. F. Saftlas, A. B. Wallis, C. Peek-Asa, and G. Rushton, "A probabilistic sampling method (psm) for estimating geographic distance to health services when only the region of residence is known," *International journal of health geographics*, vol. 10, no. 1, p. 4, 2011.

[2] J. A. McElroy, P. L. Remington, A. Trentham-Dietz, S. A. Robert, and P. A. Newcomb, "Geocoding addresses from a large population-based study: lessons learned," *Epidemiology*, vol. 14, no. 4, pp. 399–407, 2003.

[3] T. Khan, "Evaluating the errors associated with zip code polygon when employed for spatial analyses," 2012, Master's Thesis. [Online]. Available: http://ebot.gmu.edu/handle/1920/8023

[4] T. S. Khan, "Zip-code classification using spatial and crowdsourced data," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 1694–1698.

[5] X. Shi, "Evaluating the uncertainty caused by post office box addresses in environmental health studies: A restricted monte carlo approach," *International Journal of Geographical Information Science*, vol. 21, no. 3, pp. 325–340, 2007.

[6] S. E. Hurley, T. M. Saunders, R. Nivas, A. Hertz, and P. Reynolds, "Post office box addresses: a challenge for geographic information system-based studies," *Epidemiology*, vol. 14, no. 4, pp. 386–391, 2003.

[7] T. H. Grubesic, "Spatial data constraints: Implications for measuring broadband," *Telecommunications Policy*, vol. 32, no. 7, pp. 490–502, 2008.

[8] T. H. Grubesic and T. C. Matisziw, "On the use of zip codes and zip code tabulation areas (zctas) for the spatial analysis of epidemiological data," *International journal of health geographics*, vol. 5, no. 1, p. 58, 2006.

[9] D. Dai, "Black residential segregation, disparities in spatial access to health care facilities, and late-stage breast cancer diagnosis in metropolitan detroit," *Health & place*, vol. 16, no. 5, pp. 1038–1052, 2010.

[10] N. Krieger, J. T. Chen, P. D. Waterman, M.-J. Soobader, S. Subramanian, and R. Carson, "Geocoding and monitoring of us socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? the public health disparities geocoding project," *American journal of epidemiology*, vol. 156, no. 5, pp. 471–482, 2002.

[11] "Usa zip code areas," http://www.arcgis.com/home/item.html?id=8d2012a2016e484dafaac0451f9aea24, accessed: 2017-12-6.

[12] "Map features," http://www.sammdata.com/, accessed: 2017-12-6.

[13] "https://www.census.gov/geo/reference/zctas.html."

[14] "docs," https://developer.twitter.com/en/docs, accessed: 2017-12-6.

[15] T. Khan and A. Kabir, "Zipcode-classification-demo," https://github.com/subrina0013/ZIPCode-Classification-Demo, 2019.

[16] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, *The R*-tree: an efficient and robust access method for points and rectangles*. ACM, 1990, vol. 19, no. 2.

[17] G. R. Hjaltason and H. Samet, "Ranking in spatial databases," in *International Symposium on Spatial Databases*. Springer, 1995, pp. 83–95.