Routledge
Taylor & Francis Group

Police Practice and Research
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gppr20

A two-step process to increase successful geocoding in publicly available police stop data

Danielle Wallace, Edward Helderop, Anthony Grubesic, Jason Walker, Xiaoyue Cathy Liu, Ran Wei, Yirong Zhou & Connor Stewart

To cite this article: Danielle Wallace, Edward Helderop, Anthony Grubesic, Jason Walker, Xiaoyue Cathy Liu, Ran Wei, Yirong Zhou & Connor Stewart (2023): A two-step process to increase successful geocoding in publicly available police stop data, Police Practice and Research, DOI: 10.1080/15614263.2023.2181169

To link to this article: https://doi.org/10.1080/15614263.2023.2181169

Published online: 21 Feb 2023.

Routledge
Taylor & Francis Group

A two-step process to increase successful geocoding in publicly available police stop data

Danielle Wallace[a], Edward Helderop[b], Anthony Grubesic[b], Jason Walker[a], Xiaoyue Cathy Liu[c], Ran Wei[b], Yirong Zhou[c] and Connor Stewart[a]

[a]School of Criminology and Criminal Justice, Arizona State University, Phoenix, USA; [b]School of Public Policy, University of California Riverside, Riverside, USA; [c]Department of Civil & Environmental Engineering, University of Utah, USA

ABSTRACT

Many police departments are meeting calls for transparency by releasing publicly accessible data. High-quality address locations are critical for successful and accurate geocoding, though the content and quality of that data can drastically vary across datasets. In this study, we showcase a two-step geocoding process that helps convert low-quality address locations into geo-locatable addresses using traditional geocoding and Jaro-Winkler edit distance methods with police stop data from the San Diego Police Department. For reference, only 83% of stops were geocoded when using traditional geocoding methods. By employing the Jaro-Winkler edit distance to clean the stop address strings, we were able to geocode 99% of stops. We further discuss data creation practices and solutions for data quality-related issues for police departments and researchers when using publicly available policing data.

ARTICLE HISTORY
Received 29 November 2022
Accepted 10 February 2023

KEYWORDS
Policing data; geocoding; publicly available data; data quality; Jaro-Winkler edit distance

Introduction

Police departments are meeting calls for increased transparency through publicly accessible data. For example, many major metropolitan police departments now release policing data on open data portals. Moreover, a handful of projects and initiatives collate policing data from many agencies; for example, the Stanford Open Policing Project creates standardized traffic and pedestrian stops from a host of agencies across the US. Currently, their data repository has data on over 100 million police stops available to researchers (see https://openpolicing.stanford.edu/data/; Pierson et al., 2020).

While a boon for researchers, the sorts of data police departments release publicly and the content of that data is often entirely up to the police department. Given the decentralized nature of policing in the US, government requirements – at any level – about the content and quality of the data police departments release to the public rarely exist. Consequently, police departments control the nature of their public data releases, with departmental culture often shaping its content. While some states have reporting requirements (of S, 2018), without legislation surrounding data content, quality, and documentation or large data warehouses such as the Stanford Open Policing Project, researchers have no guarantee that the data would contain their needed content and/or have adequate documentation for appropriate usage.

Without regulations, data quality may also be an issue. Departments may release data with crucial missing information or without variables that officers are mandated to collect in the field.

CONTACT Danielle Wallace ✉ danielle.wallace@asu.edu School of Criminology and Criminal Justice, Arizona State University, 411 N. Central Ave. Room 600, Phoenix, AZ 85004

Additionally, these data releases often contain spelling errors and a lack of detail concerning abbreviations or acronyms within the data. Given that many policing researchers seek to solve problems in policing and provide potential solutions, high-quality data are essential from public-facing sources (Helderop et al., 2023). Inaccurate policing data undermines the validity of policing research. Unless researchers download data from public sources that engage in quality control and/or standardization, researchers must either trust that the data is accurate or find methodological solutions to use low-quality data.

In this study, we detail a geographic data quality problem we encountered using the newly released Racial Identity and Profiling Act (RIPA) data from the San Diego Police Department (SDPD). While the SDPD RIPA data includes the geographic location of the incident, officers report this information in an unstructured, open-text entry. While the data is unique in that Open Justice, the data aggregator for the RIPA data within California, demands quality control for missing information, there is no guidance (that we can find) given to the departments on whether and how they should clean their address data prior to release, only guidance to the officer on how to report cities (not address locations). Providing address data that is unstructured leads to a host of problems, ranging from spelling errors to addresses that do not make sense geographically, which researchers must overcome when trying to make sense of the locational information. Our experience using the SDPD RIPA data was no different.

The problems associated with the geographic location quality in the data were compounded by the intended use of the data: stop data would later be employed as a part of a racial profiling benchmarking study. In police departments examining the practice of racialized policing, whether voluntarily or under a court order or consent decree, officers are often subject to new standards of evaluation, monitoring, and data collection surrounding the race/ethnicity of the individuals with whom they contact while on duty. Researchers have documented officers' perspectives surrounding these interventions, with few officers feeling positive about the new practices and procedures (Chanin & Welsh, 2021; Davis et al., 2006). Organizational justice is a critical component of departments' being able to address police misconduct (Wolfe & Piquero, 2011), and likely critical for the effective construction and implementation of racialized policing interventions. Our concern was that a low threshold for successful geocodes, where 15% or even 10% of stops were excluded due to failed geocoding, would give officers pause over departmental practice that is already likely to be viewed skeptically. While excluding stops that are unsuccessfully geocoded may not affect statistics (this has yet to be tested), excluding stops may shape officers' perceptions of the validity of the benchmark. Thus, we decided that for our purposes, the 85% standard for successful geocoding was too low.

Together with our need for a higher threshold of geocoding accuracy and the raw nature of the stop location information, we employed a more complex, multistep geocoding strategy that involved traditional geocoding coupled with computer science and statistical techniques surrounding string cleaning. In the coming sections, we detail our stop locations geocoding approach and its associated results. In addition, we aim to highlight alternative geocoding methods when researchers face geographic data quality issues. As more policing and criminological data becomes available, scholars and researchers need a more extensive and specialized toolkit for using and manipulating publicly available data that does not come with the data quality assurances in data warehouses like ICSPR.

Methods and materials

The State of California passed the Racial and Identity Profiling Act (RIPA; AB 953, 2015-2016) in 2015. The RIPA mandated that all state and local agencies employing police officers report yearly information on all police incidents using a standardized data reporting format, which delineates all of the included information and the nature of the stop data. Agencies that employed over 1,000 police officers, such as SDPD, were the first to collect and report RIPA data. SDPD began collecting

RIPA data on 1 July 2018. While all RIPA data from large agencies in California is available at Open Justice (https://openjustice.doj.ca.gov/data), SDPD released a more inclusive RIPA dataset (see https://data.sandiego.gov/datasets/police-ripa-stops/) on their open data portal. We employ this data in our study. Our analytic dataset includes the complete SDPD RIPA data of 404,107 police stops from 1 July 2018, to 31 March 2021, without exclusion.

Analysis plan

Within the SDPD RIPA data, recorded stop locations were categorized as either an address or an intersection. SDPD did not provide latitude-longitude coordinates that corresponded with a stop, nor was the address data cleaned before release. We understand that the address variables in the SDPD RIPA data are the raw, uncleaned version of what officers' type into the stop contact form. As a result, the quality of the location information within these data was highly variable; while many entries contained a coherent and correctly-spelled address, a significant number did not. These irregular entries contained misspellings, missing data, and ill-formatted entries. Before engaging in geocoding, we tried to clean and correct any obvious spelling errors in the address fields and homogenize street abbreviations. As an example, 'Misson' was changed to 'MISSION' and 'street' or 'st' were changed to 'st', while all highway and freeway references were also standardized. For example, I-15, 15N, 15S, and Interstate 15 all refer to the same road, but we converted them to I15. We performed the same process for state routes (for example, SR56) with the added consideration that many state routes have names (SR56 is also known as the Ted Williams Freeway, for example).

Additionally, we excluded 1,432 (0.35%) stops entirely; these entries were missing critical information, such as a street name or a second road for an intersection stop, or were too garbled to interpret reliably. Examples include 'Ski Beach', 'Clairmont Dr & 4700', and 'Ronson Rd.' Finally, we geocoded the remaining 402,675 stops using our multistep process. We performed all data cleaning, reformatting, and geocoding processes with a custom R script.

Consistently formatted, correctly spelled location data are easier to geocode. However, many of the stop locations in this dataset contain misspellings, irregular formatting, or other data quality issues. Geocoding misspelled addresses or alternative address formats without correction would have introduced significant spatial error in our geocoded results. Instead, we used bespoke string cleaning scripts and manual entry reviews to generate high-quality geocodes. The geocoding process contained multiple filters designed to ensure high-accuracy geocodes while minimizing the number of entries flagged for manual review.

We began with the generation of two master location datasets. The first was a complete address point shapefile for San Diego generated using a SITUS address point dataset maintained by the county assessor office and distributed by Data San Diego (Data SD, 2022a). The second was a street intersection dataset generated using a complete San Diego County road shapefile (Data SD, 2022b). Each intersection was identified as a point and associated with spatial coordinates.

The first geocoding step was to search for exact matches, wherein the street names present in the observation were matched definitively to street names in the master location datasets. For the intersection stops, this was enough to generate a high-confidence geocoded result. For the address-based stops, if we found an exact match for the street number (i.e., the listed street number corresponded to an existing address in San Diego), we assigned that location as the address for geocoding. When street numbers did not correspond to an existing address, we used the closest match as the geocode. For example, we assigned a stop listed as occurring at '100 Main St.' with the coordinate pair of '110 Main St.' if that was the closest existing address. After these exact matches, we performed a second geocoding process devoid of street name suffixes to isolate instances with errors (e.g., '1st St.' when the correct name is actually '1st Ave.').

To successfully match basic misspellings, we identified the most similar existing street names to the stop data using the Jaro-Winkler edit distance (Jaro, 1989; Wang et al., 2017; Winkler, 1999). The Jaro-Winkler distance between two strings provides a metric of their similarity based on the

Table 1. Example of the jaro-Winkler edit distance cleaned address using the word avenue.

Original Address	Address Corrected with Jaro-Winkler Technique
600 07TH AVEUE	600 7TH AVE
800 07TH SVEBUE	800 7TH AVE
300 08TH AVEVEVEE	300 8TH AVE
700 05TH AVENEU	700 5TH AVE
1600 04THAVENUE	1600 4TH AVE

number of correct characters, the string length, and the number of transpositions, with a bias towards correct characters at the beginning of the string. The resulting Jaro-Winkler distance ranges from 0–1, where a 0 indicates a perfect match (i.e., the two strings are identical) and a 1 indicates a complete difference between the strings. Thus, for each remaining observation, the nearest match street name was identified along with a metric that denoted how similar the match was. Based on a manual review of these matches, we set a threshold at 0.2. Any matches with a Jaro-Winkler distance below or equal to 0.2 (recall that a score of 0 indicates perfect similarity) were considered high-confidence matches and geocoded accordingly. We manually checked all observations with scores above 0.2. In all, we checked 8,880 observations for accuracy. Using the example of the word 'avenue', Table 1 shows the original address and the finalized address using the Jaro-Winkler edit distance string cleaning technique.

Results

Table 2 shows the total number and percentage of stops by how they were processed in the study. Remember that 1,432 stops had address information that was not able to be geocoded and were ultimately excluded from the analysis. Next, once we cleaned the addresses, we performed a first-pass geocode, looking for identical matches. Using this method, we were able to geocode 83.3% of all stops ($n = 335{,}414$ out of 402,675 eligible stops), which is below the threshold of 85% commonly used in crime mapping and analysis (Ratcliffe, 2004) and well below the accuracy we need to create an external benchmark we intend to employ in our larger study. To improve this outcome, we used the Jaro-Winkler edit distance metric to identify the likeliest match for the remaining stops ($n = 67{,}311$).

Of the remaining 67,311 stops without an exact match, we were able to geocode 66,227 of them using a combination of the Jaro-Winkler edit distance and manual review. Only 1,034 stops remained without coordinates after using Jaro-Winkler edit distance technique. These stops typically had street names that were too garbled for identifying a successful match (e.g., '2200 N/A IMPERIAL AVE,' 'CCCCCCC,' and 'bldg 7 jemma avenue'). In other instances, assigning entries to more than one correct street name was possible.

In all, using the combination of these two methods, from the original 404,107 stops in the data, there were only 2,466 stops that we could not geocode due to missing address information or unsuccessful string cleaning ($n = 1{,}432$ and $n = 1{,}034$, respectively). Our successful geocoding rate was 99.4% (i.e., 401,107 successfully geocoded stops out of 404,107 total stops).

Table 2. Percentage of stops geocoded by geocoding method.

Geocoding Method	Frequency	Percent
Traditional Geocoding	335,414	83.00
Successful Jaro-Winkler Edit Distance Geocoding	66,227	16.39
Unsuccessful Jaro-Winkler Edit Distance Geocoding	1,034	0.26
Unable to be Geocoded	1,432	0.35
Total	404,107	100.00

Finally, we took steps to validate the geocoding results coming from both methods. For each stop, the data contained information on the beat where the stop occurred. We used this information to validate the geocoded location of the stop. It is important to note that the beat information is not officer-recorded, unlike the address data, which is hand-entered by the officer. The beat comes from another data generating process within the San Diego Police Department, such as dispatch. Unfortunately, we have no information about the process through which beats are assigned to stops. Thus, we expect some differences between the geocoded location of stops and the beat given the different mechanisms of data gathering for these two points of information.

To assess the geocoded results for beat-stop location concordance, we took a random sample of 20,000 addresses from the group of stops geocoded through the first-pass and another random sample of 20,000 addresses cleaned and geocoded using the Jaro-Winkler technique. Of the 20,000 addresses that were easily geocoded through the first pass, 15426 addresses matched the beat (77.13%). Next, of the 20,000 addresses cleaned and geocoded using the Jaro-Winkler technique, 15374 matched the beat (76.87%). The similarity in beat correspondence between the stops geocoded in the first pass and those cleaned and geocoded cleaned using the Jaro-Winkler technique suggests that the Jaro-Winkler string cleaning technique does not introduce systematic bias into the stop locations. As such, we are confident in using the Jaro-Winkler string cleaning technique to increase addresses that can be successfully geocoded.

Discussion

This study details our approach to geocoding raw, uncleaned address locations in a publicly available police stop dataset. Using a combination of traditional geocoding and the Jaro-Winkler edit distance method for cleaning string values and subsequent geocoding, we ultimately had a 99.4% success rate for geocoding stops. This study has implications for researchers and police departments, which both use and produce policing data, which we close with below.

For researchers, we first recommend not giving up on publicly available data with data quality issues, particularly issues that research teams can address. For example, the address information in SDPD data contained scores of spelling errors and address formatting differences, including abbreviations, which are incredibly common in address records. While there was nothing we could do about nonsensical address locations, we used computational and statistical techniques for string alteration to clean the address locations. The return on investment is clear: if we kept with traditional geocoding methods and had not engaged in more rigorous systematic string cleaning, our geocoding success rate would be only 83.3%, below the commonly used standard of 85% for quality geocoding (Ratcliffe, 2004).

Rather than avoid data with problems, we encourage researchers to explore and engage with interdisciplinary ways to combat those problems, particularly when immediate solutions are not available with in our field. To this end, it would benefit criminologists to learn open science techniques and platforms, like GitHub, that house publicly available code and data for other scientists to use. There is often not a need to reinvent the wheel every time analysts need to turn to other methods; its rather common that the issue is not unique and someone has developed code to handle it. To that end, we make a version of our Jaro-Winkler edit distance code available here: https://github.com/Ehelderop/SDPD-stop-geocoding. That said, one of the difficulties of doing interdisciplinary is skill sets. Often times, solutions are available, but we may not have to skills or experience needed to do them (i.e., the need to learn a new analysis technique or a new software package). This is when cultivating working relationships with interdisciplinary scholars is critical. Interdisciplinary work is more than just engaging the literature, it is also engaging those with different perspectives and capacities.

For police departments and agencies that house criminal justice data, we have a few recommendations surrounding data quality, many of which focus on easing data-entry burdens faced by police officers. First, we recommend standardizing data entry methods on stop forms. Officers should have a homogenous way of entering addresses, intersections, and street names and their abbreviations; ideally, this includes some automation for the officer rather than forcing them to rely on hand-typed data entry. Better still is automating XY coordinates to stop forms that can be easily obtained from dispatch rather than requiring officers to fill out this complex information that is likely duplicative. For example, in the SDPD RIPA data, and all RIPA data for that matter, officers are required to fill out a stop contact form for each person involved in the stop. While this information is precious, this is a considerable data entry 'lift' for officers. As a result, it is unsurprising that address locations had spelling errors and other formatting problems: officers simply do not have time, and perhaps even willingness (see Chanin & Welsh, 2021 for a discussion on data entry burdens at SDPD), to correct their entries. Automation in forms is critical for high-quality data and keeping officers willing to provide that data.

Acknowledgments

We want to thank our anonymous reviewers for their valuable insights and suggestions. Additionally, this work was supported by the National Science Foundation under Award #2051226, "Constructing Race-Specific Driving Patterns to Address Racial Profiling." Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the NSF [2051226].

Notes on contributors

Danielle Wallace is an Associate Professor in the School of Criminology and Criminal Justice and an Associate Director at the Center for Violence Prevention and Community Safety at Arizona State University. Her research agenda includes neighborhoods and crime, policing, racial/ethnic and disability-related disparities in policing, and issues related to incarceration, re-entry and health.

Edward Helderop is a systems analyst and the associate director of the Center for Geospatial Sciences at the University of California, Riverside. His main interests include GIScience, big data, and network analytics (particularly as applied to urban infrastructure systems). His previous research explored turnover and resiliency in plant-pollinator networks and urban transportation modeling during disasters. Eddie received his B.S. in Biology from Hope College, his M.S. in Geography from Oregon State University, and his Ph.D. in Geography from Arizona State University.

Anthony H. Grubesic is a professor in the School of Public Policy at the University of California, Riverside. His research and teaching interests are in geocomputation, spatial analysis, regional development, and public policy evaluation.

Jason Walker is a doctoral student in the School of Criminology and Criminal Justice at Arizona State University. His primary research interests focus on neighborhood crime and disorder, police, public health in the criminal justice system, and sentencing outcomes. Prior to attending the doctoral program at Arizona State University, Jason worked as an analyst for the United States Sentencing Commission.

Xiaoyue Cathy Liu received her B.S. in electrical engineering from Beijing Jiaotong University, and Ph.D. from the University of Washington. She is currently an associate professor in the Civil & Environmental Engineering at the University of Utah. Her research interests include multimodal transportation system, electrified mobility, equity in transportation, and big data applications.

Dr. Ran Wei is currently an Associate Professor in the School of Public Policy and a founding member of the Center for Geospatial Sciences at the University of California, Riverside. Her areas of emphasis include GIScience, urban and regional analysis, spatial analysis, optimization, geovisualization, high performance computing and location analysis.

Substantively, she has focused on a range of national and international issues, including urban/regional growth, transportation, public health, crime, housing mobility, energy infrastructure, and environmental sustainability.

Yirong Zhou (M'94) received B.S. in Statistics from the University of Science and Technology of China in 2017 and an M.S. in Statistics from George Washington University in 2019. He is a current Ph.D. Student in Civil & Environmental Engineering at the University of Utah under the supervision of Prof Xiaoyue Cathy Liu. His research focuses on data-driven transportation system modeling.

Connor Stewart is a graduate student at Arizona State University. His interests include the use of data science in criminology, networks, and domestic terrorism.

ORCID

Danielle Wallace http://orcid.org/0000-0001-6648-9986
Edward Helderop http://orcid.org/0000-0003-0590-5258
Anthony Grubesic http://orcid.org/0000-0003-4517-586X
Jason Walker http://orcid.org/0000-0003-1530-8924
Xiaoyue Cathy Liu http://orcid.org/0000-0002-5162-891X
Ran Wei http://orcid.org/0000-0002-2737-1712
Yirong Zhou http://orcid.org/0000-0002-0524-3437
Connor Stewart http://orcid.org/0000-0002-9834-5546

References

A.B. 953, 2015-2016, Reg. Sess. (Cal. 2015). https://perma.cc/VS8V-KYSL.

Chanin, J., & Welsh, M. (2021). Examining the validity of traffic stop data: A mixed-methods analysis of police officer compliance. *Police Quarterly*, *24*(1), 3–30. https://doi.org/10.1177/1098611120933644

Data SD. 2022a. Address Points to APN. https://data.sandiego.gov/datasets/address-points-apn/.

Data SD. 2022b. Roads Lines. https://data.sandiego.gov/datasets/roads-lines/

Davis, R. C., Henderson, N. J., Mandelstam, J., Ortiz, C. W., & Miller, J. (2006). Federal intervention in local policing: Pittsburgh's experience with a consent decree. *US Department of Justice.*

Helderop, E., Nelson, J. R., & Grubesic, T. H. (2023). 'Unmasking' masked address data: A medoid geocoding solution. *MethodsX.*

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785

Legislators, N. C. of S. (2018). *State trends in law enforcement legislation, 2014-2017*. https://www.ncsl.org/research/civil-and-criminal-justice/state-trends-in-law-enforcement-legislation-2014-2017.aspx

Pierson, E., Simoiu, C., Overgoor, J., Corbett Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, *4*(7), 736–745. https://doi.org/10.1038/s41562-020-0858-1

Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, *18*(1), 61–72. https://doi.org/10.1080/13658810310001596076

Wang, Y., Qin, J., & Wang, W. (2017). *Efficient approximate entity matching using Jaro-Winkler distance BT - Web Information Systems Engineering – WISE 2017* (Bouguettaya, A., Gao, Y., Klimenko, A., Chen, L., Zhang, X., Dzerzhinskiy, F., Jia, W. S. V. Klimenko & Q. Li (Eds.), Springer International Publishing.

Winkler, W. E. (1999). The state of record linkage and current research problems. In *Statistical society of Canada, proceedings of the survey methods section* (pp. 73–90).

Wolfe, S. E., & Piquero, A. R. (2011). Organizational justice and police misconduct. *Criminal Justice and Behavior*, *38*(4), 332–353. https://doi.org/10.1177/0093854810397739