

Collecting Location Data in Web-Based Surveys: Challenges and Lessons from Self-Reported Census Tracts

Survey Methods: Insights from the Field,

Zeynep Altinay  | Chrissy Klenke  | Carlos Ramirez-Reyes 

How to cite this article : Altinay, Z., Klenke C. & Ramirez-Reyes C. (2026). Collecting Location Data in Web-Based Surveys: Challenges and Lessons from Self-Reported Census Tracts. Survey Methods: Insights from the Field. Retrieved from <https://surveyinsights.org/?p=21909>

DOI : 10.13094/SMIF-2026-00007

Copyright : © the authors 2026. This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

Abstract : Geographic context shapes how people perceive environmental issues, health risks, and policy decisions, yet collecting location data in web-based surveys presents challenges related to privacy, accuracy, and analytic precision. Traditional methods, such as device tracking or collecting residential addresses, are intrusive, while broader units like states or counties are too coarse for many research questions. This study tested the feasibility of collecting census tract codes through participant self-report to balance methodological rigor with ethical concerns. Manual entry of census tract codes introduced substantial formatting inconsistencies — including omitted leading zeros and inconsistent decimal notation — relative to the official census tract format. A sequential standardisation procedure was required to reconcile these inconsistencies with the official reference data, after which 92.2% of tract responses were successfully matched to a valid census tract. Manually entered ZIP codes required comparatively minimal standardisation, yielding a 97.2% match rate against official U.S. Census Bureau ZIP Code Tabulation Areas (ZCTAs). These findings indicate that while both self-reported identifiers can achieve high validity, census tract data demand substantially greater processing effort to reconcile respondent formatting with official geographic standards. We discuss the methodological and ethical implications of collecting geographic information in online surveys, including the trade-offs among spatial precision, respondent burden, and the analytic overhead required to make finer-grained geographic self-report usable.

1 Introduction

Public opinion studies in the social sciences often benefit from collecting location data from survey participants, as geographic context plays a significant role in informing and shaping perceptions of social and environmental issues. Proximity data has gained momentum, particularly in risk perception studies examining environmental and technological hazards and it further supplements the broader environmental psychology and public opinion literature (e.g. Egan & Mullin, 2012; Feddersen et al., 2016). Studies in this area often link participants' location data to attitude and behaviour questions. Many studies also examine perceptions of proximity (e.g. Gadekar & Brandão, 2025; Moore & Kestens, 2011). Perceived proximity, however, does not always align closely with objective measures of distance (Lackey & Kaczynski, 2009). One method of collecting geographic information involves through self-report, which

requires respondents to be both willing and able to accurately report their location. Automated location collection offers an alternative by reducing respondent burden and allowing geographic information to be obtained even when respondents cannot provide a usable address or location identifier (Felderer & Blom, 2022). Automated approaches, such as IP-based or device-derived location data, however, raise important ethical and methodological considerations related to privacy, informed consent, and respondents' willingness to share digital information (Couper & Singer, 2013; Singer, 2011; Struminskaya et al., 2021). Consequently, identifying privacy-preserving methods for collecting geographically precise survey data remains an important methodological challenge (Vaske, 2011; Wardropper et al., 2021). These challenges are heightened by growing public concerns about the erosion of privacy on digital platforms, which can make participants more reluctant to share location information or lead them to provide incomplete or inaccurate data. One approach to addressing these challenges is to collect aggregate location data by asking for state, county, or city information, though these units are often too broad to capture the precision needed for social science analysis.

A common solution is to ask participants for their ZIP (Zone Improvement Plan) code, the standard postal code used in the United States, which is generally easy for respondents to recall and provides a finer level of geographic detail than city- or state-level information. Although ZIP codes do not provide neighbourhood-level precision, they are often sufficient for studies that do not require highly localized spatial analyses. For example, Wei et al., (2025) mapped survey respondents' ZIP codes to the nearest weather stations to examine how local weather conditions influence public perceptions of global warming, which demonstrates that broader geographic units can be appropriate when neighbourhood-level data are unnecessary. Although ZIP codes offer a less intrusive alternative to collecting precise addresses, they present several challenges for scientific research. ZIP code boundaries are designed for mail delivery rather than social, geographical, or demographic analysis, which limits their precision and alignment with meaningful community or neighbourhood boundaries (Grubestic, 2008; Grubestic & Matisziw, 2006). As a result, populations within a single ZIP code can be highly heterogeneous, obscuring important differences. ZIP codes also change over time, and participants may provide inaccurate or outdated information, further reducing data reliability. In addition, many demographic, and environmental and health datasets in the United States are organized by census tracts rather than ZIP codes, making it difficult to integrate ZIP code data with other sources.

Census tracts provide more precise geographic information than ZIP codes because they represent smaller, more uniform areas. Unlike ZIP codes, which vary widely in size and often lack geographic consistency, census tracts are standardised and designed to capture more precise, neighbourhood-level data. For example, in the United States, ZIP codes in urban and suburban areas may include 4,000 people, whereas a census tracts or block groups typically include 1,500 individuals (Henry & Boscoe, 2008). States such as Nevada illustrate this issue particularly well, as ZIP code boundaries often span vast, sparsely populated regions that bear little resemblance to standardised census geographies, making spatial analyses of distance and exposure prone to error (Figure 1). Even though census tract codes are more appropriate for scientific research, these are generally not information that individuals know from memory or routinely use, and many may not know how to retrieve them. For example, the U.S. Postal Service provides a user-friendly ZIP code lookup web tool that allows individuals to enter a residential or business address and immediately obtain the corresponding ZIP code (USPS, 2026). In contrast, locating a census tract typically requires navigating specialized government databases or mapping tools that are less familiar to the general public and may require additional geographic knowledge to use effectively. As a result, using ZIP codes, which are routinely used by individuals, often remains the more practical approach for data collection in social sciences.

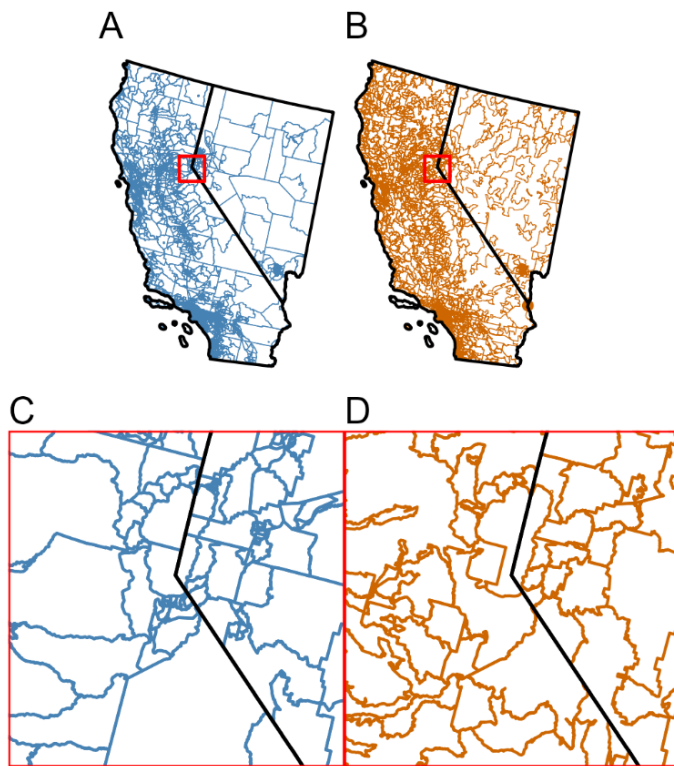


Figure 1. Comparison of U.S. Census tract and ZIP Code Tabulation Area (ZCTA) geographic extents across California and Nevada. (A) Census tract boundaries for the full study area, with state boundaries shown in black. (B) ZCTA boundaries for the same extent. (C) Detailed image of census tracts corresponding to the red box shown in panel A. (D) ZCTAs within the same extent, corresponding to the red box shown in panel B.

Several studies have primarily focused on respondents' willingness to share geographic information in digital settings with limited attention to the methodological challenges associated with collecting and validating self-reported location identifiers (e.g. Chan et al., 2017; Di Matteo et al., 2018; Toepoel & Lugtig, 2014). However, even if survey respondents are willing to share more precise geographic location, there might be barriers to obtain this information with researchers. In this study, we examined the extent to which individuals could successfully retrieve and report their census tract numbers, a method designed to enhance location accuracy while maintaining respondent privacy. Our goal is to quantitatively evaluate the feasibility of collecting self-reported census tract information in survey research and to compare its performance with more commonly used geographic identifiers such as ZIP codes.

This study addresses that gap by systematically evaluating respondents' ability to provide census tract information, documenting the practical challenges associated with data collection and processing, and comparing census tract and ZIP code-based approaches with respect to geographic precision, data quality, respondent burden, and privacy protection. In doing so, we provide empirical evidence regarding the trade-offs researchers face when selecting geographic identifiers for survey-based research. In the sections that follow, we provide a practical roadmap for collecting and processing self-reported census tract information in survey research. We first describe our procedures for collecting location data, including the instructions provided to respondents to guide them through the process. We then outline our data-cleaning steps, including identifying and correcting human errors, organizing responses, and

conducting validity checks. Next, we compare self-reported ZIP codes and self-reported census tracts to evaluate the accuracy of self-reported census tract and ZIP code identifiers. Finally, we discuss the implications of these findings in light of collecting geographic information while balancing privacy, precision, and respondent burden in spatial analysis.

2 Methods

2.1 Survey design and data collection

This study was conducted as part of a broader project examining public perceptions of environmental and technological risks associated with lithium mining in California and Nevada, United States. The broader project consisted of a survey of 62 questions (including the informed consent and eligibility screening questions) designed to measure respondents' perceptions of lithium mining developments, including perceived risks, benefits, environmental impacts, and environmental justice concerns associated with the industry in the region. While the full survey addressed these topics, the present study focuses exclusively on analysing geographic questions requesting census tract and ZIP code information.

Ethics approval

The study was approved by the Institutional Review Board, reference number [2314583-2], of the authors' institution. Participation was voluntary, and respondents could skip any questions they did not wish to answer.

Recruitment and eligibility

Two screening questions were used to verify eligibility: participants were required to be at least 18 years old and reside in California or Nevada. Participants received \$3.00 for completing the survey (equivalent to an hourly rate of approximately \$15.00), consistent with Prolific's fair-pay guidelines (Prolific, 2025). Data were collected between May 8 and June 24, 2025, using the Qualtrics survey platform (Qualtrics, 2026), with participants recruited and screened through the Prolific crowdsourcing platform.

Geographic information collection

Two questions in the survey required participants to self-report their geographic location. The first question (Appendix Figure A1) required participants to provide their census tract, and included visual instructions on how to retrieve it from the US Federal Financial Institutions Examination Council (FFIEC) Geocoding/Mapping System (Federal Financial Institutions Examination Council, 2024), a government-affiliated website that generates multiple geographic identifiers, including state, county, and census tract codes, from a full residential mailing address. The FFIEC platform was chosen over the U.S. Census Bureau's lookup tool (United States Census Bureau, 2026) due to its simpler user interface.

The second question related to geographic location asked participants to provide their ZIP codes with no additional instructions as that information is usually known to adults. To reduce nonresponse, participants were reminded that their responses would be analysed and reported only in aggregate form. Census tract codes were collected as the primary geographic measure of a participant location, offering more localized

geographic information that could be linked to demographic and environmental datasets used in the main study. In contrast, ZIP codes were collected as a secondary data point to capture participants' general geographic location and to cross-check the accuracy of their self-reported census tracts. Comparing these two measures allowed us to evaluate the trade-offs between analytical precision and ease of reporting in the context of proximity-based risk perception research.

2.2 Data preparation

Initial data cleaning focused on ensuring response quality and consistency. Duplicate submissions were identified using Prolific Participant IDs and Qualtrics metadata and subsequently removed. Two attention-check items were embedded within the survey to assess respondent attentiveness and data quality. To maximize data quality, respondents who failed either attention check were excluded from further analysis. A major challenge in preparing the dataset was standardising the formatting of census tract numbers in survey data. Several formatting and data-entry issues emerged during data preparation. Responses varied considerably in format, possibly because participants retrieved census tract information from an external website and either manually entered or copied and pasted the results into the survey. Common problems included invalid entries (e.g., geographic information other than the requested census tract identifier, missing responses, entries including leading spaces, formatting inconsistencies, and typographical errors such as multiple decimal points or the substitution of the letter "O" for the numeral zero). In addition, some respondents provided partial or full mailing addresses rather than census tract identifiers. Because the nature of these errors varied substantially across respondents, individual responses were manually reviewed primarily to identify and correct obvious formatting and transcription errors. The formatting issues, mentioned here, were specific to the self-reported census tract data and did not occur with ZIP codes, likely because participants could recall their postal codes from memory.

Exclusions

Of the 784 survey responses, 80 were excluded during preliminary data screening because they did not meet the study's inclusion or data quality criteria. Most exclusions resulted from missing geographic information, including blank census tract and ZIP code fields ($n = 59$). Additional exclusions included low-quality responses characterized by unusually short completion times (under approximately 4.5 minutes) combined with consecutive identical or missing responses across multiple matrix items ($n = 4$), survey abandonment or excessive completion times exceeding 2.5 hours ($n = 4$), invalid geographic entries that initially appeared plausible but were determined during manual review to contain formatting errors, incomplete information, or nonstandard geographic identifiers ($n = 4$), address-based responses that did not conform to the study protocol for self-reported census tract identifiers ($n = 6$), and other exclusions resulting from participant-reported errors encountered while retrieving census tract information ($n = 3$). Following preliminary data screening, 704 responses were retained for subsequent geospatial analyses. Figure 2 illustrates the geographic identifier validation workflow, including data collection, preliminary data screening, and geospatial validation.

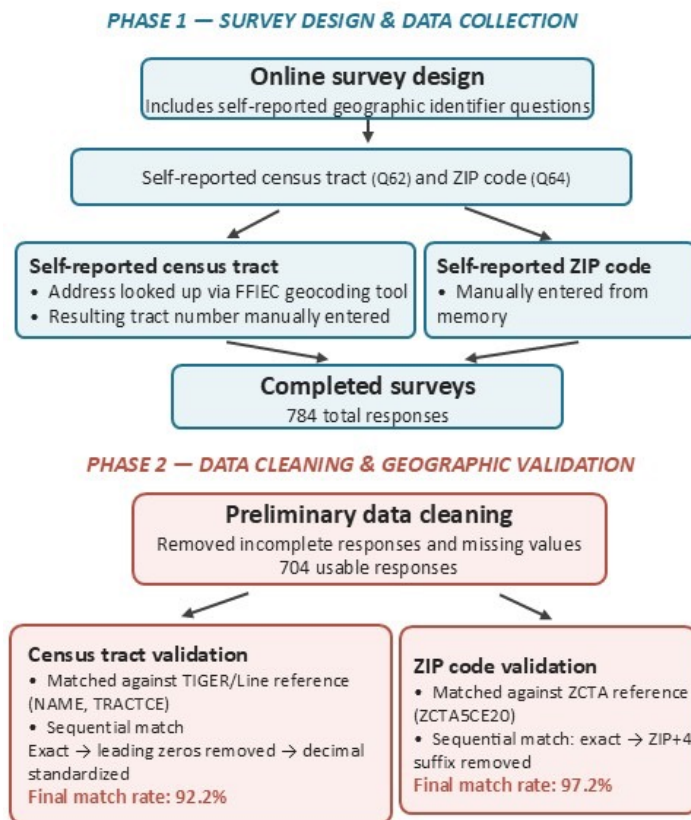


Figure 2. Geographic identifier validation workflow. Overview of the collection, cleaning, and validation procedures for self-reported census tract identifiers and ZIP codes.

2.3 Geospatial validation

Reference datasets

Geospatial validation of survey responses consisted of matching respondents' self-reported census tract and ZIP code to official, government-provided geographic reference datasets. Official census tract boundaries for California and Nevada were obtained from the U.S. Census Bureau's 2024 Topologically Integrated Geographic Encoding and Referencing (TIGER/Line) Census Tracts shapefiles, the Bureau's official digital geographic boundary files (tl_2024_06_tract.shp for California and tl_2024_32_tract.shp for Nevada). State boundaries were obtained from the 2024 TIGER/Line States shapefile (tl_2024_us_state.shp), and ZIP Code Tabulation Area (ZCTA) boundaries — census-defined geographic areas that approximate U.S. Postal Service ZIP code service areas — were obtained from the 2024 TIGER/Line ZCTA shapefile (tl_2024_us_zcta520.shp) (U.S. Census Bureau, 2024c, 2024b, 2024a).

Census tract and ZIP code matching and validation

We used R software (R Core Team, 2024; version 4.2), with the *tidyverse* (Wickham et al., 2019), *sf* (Pebesma, 2018), and *spatialEco* (Evans & Murphy, 2025) packages, to import, manipulate, and compare self-reported geographic identifiers against the official reference datasets described above. To facilitate data manipulation, the California and Nevada census tract shapefiles were merged into a single layer using the Merge tool in ArcGIS Pro (version 3.6) (ESRI, 2026); the national ZCTA shapefile was retained in full, without clipping to California and Nevada, because a self-reported ZIP code could correspond to a

ZCTA outside the two-state study area (e.g., due to respondent relocation or data-entry error).

Self-reported census tract values were validated using a sequential, three-step matching procedure designed to account for common, predictable deviations from the official tract-number format. At each step, only responses that had not already matched in a preceding, stricter step were carried forward for evaluation. First, reported tract values were compared directly to the Census-formatted tract name field (NAME; e.g., “35.19”) in the TIGER/Line reference dataset using exact string matching (R’s `%in%` operator). Second, for responses that did not match exactly, leading zeros were removed from the reported value (e.g., “0032.08” → “32.08”) and the result was compared to NAME. Third, for responses still unmatched, the decimal point was removed from the reported value to reconstruct the six-digit format used in the official tract code field (e.g., “0032.08” → “003208”) and compared to TRACTCE. Responses that did not match at any of the three steps were classified as invalid. For each respondent, we recorded which step (if any) produced a match in a categorical variable (exact match, leading zeros removed, decimal removed, or no match) and appended the corresponding official tract identifiers from the matched reference record — the TIGER tract name and the six-digit TRACTCE code — to the survey dataset.

Self-reported ZIP codes were validated using an analogous sequential procedure against the ZCTA5CE20 field of the national ZCTA reference dataset. Reported ZIP codes were first compared directly using exact string matching. For responses that did not match, a leading zero was restored where applicable (e.g., “2138” → “02138”) and the response was re-compared. For responses still unmatched, an appended ZIP+4 suffix, if present, was removed (e.g., “89701-1234” → “89701”) and the response was compared a final time. As with tract validation, a categorical match-type variable and the standardised official ZCTA identifier from the matched record were retained for each respondent.

We calculated the overall proportion of valid tract and ZIP responses at each stage of the matching procedure and summarized final match rates across respondent demographic characteristics. Because overall match rates were high, and some demographic categories included few respondents, we report subgroup match rates alongside sample sizes and interpret them descriptively rather than through formal significance testing. The complete R script is available in Appendix A2.

3 Results

Sample characteristics

The final sample consisted of 704 respondents, including 373 men (53%), 326 women (46.3%), and 4 respondents who identified as non-binary or a third gender (0.6%). Educational attainment was relatively high, with 310 respondents (44%) holding a bachelor’s degree and 129 (18.3%) holding a graduate or professional degree. The sample was predominantly White/Caucasian (59.8%), with 18.8% identifying as Hispanic. Most respondents were employed full-time (53.1%), or part-time (16.9%) with the remaining participants representing a range of employment statuses. Annual household income was broadly distributed across income categories, with no single category accounting for more than 21% of respondents. Political affiliation was relatively balanced across Democratic, Republican, and Independent affiliations, with Lean Democratic (25.3%) and Independent (23.7%) representing the largest groups.

Census tract validation

Geographic validation of self-reported census tracts proceeded through a sequential, three-step matching procedure. Exact string matching against the official TIGER tract name field identified 198 valid matches (28.1%). Removing leading zeros from unmatched responses and re-comparing against the same field increased the cumulative match rate to 571 (81.1%). Removing the decimal point from remaining unmatched responses and comparing against the six-digit official tract code field increased the cumulative match rate to 649 (92.2%) (Table1). The remaining 55 responses (7.8%) did not match a valid census tract identifier under any step and were classified as invalid.

Table 1. Validation of self-reported census tract identifiers against Census TIGER/Line reference data

Validation step	Description	Number of matches	Percentage of responses (%)
Exact match	Self-reported tract identifier matched directly with the TIGER/Line NAME field (e.g., reported 32.02 matched reference tract 32.02).	198	28.1
Remove leading zeros	Leading zeros were removed from reported tract identifiers before matching (e.g., reported 0032.02 standardised to 32.02, matching reference tract 32.02). Cumulative total.	571	81.1
Standardise tract code format	Decimal points were removed and values were standardised to the official Census tract code format (TRACTCE) (e.g., reported 32.02 or 0032.02 standardised to tract code 003202). Cumulative total.	649	92.2

Note. Matches are cumulative: each step was applied only to responses not already matched in a preceding step. Percentages are of total responses compared (N = 704). The remaining 55 responses (7.8%) did not match a valid tract identifier under any step.

ZIP code validation

Validation of self-reported ZIP codes followed an analogous procedure. Exact matching against the official ZCTA identifier field identified 678 valid matches (96.3%) (Table2). No additional responses were recovered by restoring a dropped leading zero. Removing an appended ZIP+4 suffix from remaining unmatched responses recovered 6 additional matches, increasing the cumulative match rate to 684 (97.2%). The remaining 20 responses (2.8%) did not match a valid ZCTA under any step.

Table 2. Validation of self-reported ZIP codes against Census ZIP Code Tabulation Area (ZCTA) reference data

Validation step	Description	Number of matches	Percentage of responses (%)
Exact match	Self-reported ZIP code matched directly with the ZCTA5CE20 field in the national TIGER/Line ZCTA reference dataset (e.g., reported 89503 matched reference ZCTA 89503).	678	96.3
ZIP+4 suffix removed	An appended ZIP+4 suffix was removed from the reported value before matching (e.g., reported 89701-1234 standardised to 89701, matching reference ZCTA 89701). Cumulative total.	684	97.2

Note. Categories are mutually exclusive and sum to total responses compared (N = 704). A leading-zero restoration step (e.g., reported 2138 standardised to 02138) was also attempted but did not recover any additional matches.

Demographic patterns

Match rates were generally high and comparatively stable across demographic subgroups, though some patterns emerged (Appendix Table A1). ZIP code validation exceeded 98% for nearly all respondent characteristics. In terms of Race, rates were slightly lower for those who identified Black or African American (89%, $n=82$) and Other (94%, $n= 34$). Census tract match rates showed a similar trend, where most respondents groups achieved similar validation rates. Census tract match rates were highest among respondents with some college or an associate degree (95.2%, $n = 189$) and lowest among those with a high school diploma or equivalent (87.5%, $n = 72$). No consistent patterns were observed across income, political affiliation, or ideological groups. Several demographic categories with very small sample sizes (e.g., Native Hawaiian or Pacific Islander, $n = 3$; less than high school education, $n = 1$) produced match rates of 0% or 100% that are not interpretable given the limited number of respondents and are reported for completeness rather than as evidence of a subgroup effect. Given the overall high validity of self-reported geographic identifiers and the small size of several subgroups, we interpreted demographic patterns descriptively rather than through formal statistical testing.

Among the 59 respondents who did not provide usable location information because they left the geographic fields blank or otherwise chose not to share their location (excluding invalid geographic entries), we examined demographic patterns in non-response. Because this group was small, we are cautious about drawing conclusions for demographic subgroups with few respondents. Gender distributions, however, remained sufficiently represented to permit a descriptive comparison. Women accounted for 56% of non-responses, compared with 31.7% among men and 11.7% among respondents with missing or unspecified gender.

4 Discussion

In this study, we examined whether asking respondents to locate and manually enter their census tract codes could serve as a privacy-preserving method for collecting precise geographic information without requiring respondents to disclose their residential addresses or relying on device-derived paradata. Results from this cross-sectional study suggest that collecting self-reported census tract codes can serve as a viable privacy-preserving method for obtaining geographic information. However, successful implementation depends on addressing challenges faced by both researchers and survey participants. An important methodological takeaway is that geographic validation discrepancies may not always reflect

respondent error. During the exploratory analyses, we noticed that many initially unmatched census tract identifiers became valid matches after accounting for differences in identifier formatting, particularly the presence or absence of leading zeros. A substantial portion of the data wrangling required in this study stemmed from a mismatch between the format displayed by the geocoding tool respondents used and the format needed to match Census reference data. Respondents looked up their tract using the Federal Financial Institutions Examination Council (FFIEC) Geocoding/Mapping System, which displays tract identifiers with a decimal point and inconsistent leading zeros — closer to the TIGER/Line NAME field (e.g., “218.14,” no leading zeros, decimal retained) than to the six-digit, zero-padded TRACTCE field (e.g., “021814,” no decimal) used elsewhere in the reference data. This produced varied respondent entries: “218.14” matched NAME directly; “0032.08” required only that its leading zero be dropped to match NAME (“32.08”); and “0106.00” happened to reproduce the correct TRACTCE value once its decimal was removed (“010600”), since its leading zero was, in this case, actually part of the official code rather than a formatting issue. Had the FFIEC tool displayed the standardised six-digit TRACTCE code directly, we could simply have instructed respondents to enter six digits including leading zeros, avoiding this ambiguity entirely. Researchers should be aware that apparent validation failures may sometimes reflect inconsistencies across official data systems rather than inaccuracies in participant responses. Even when respondents accurately report their census tract as displayed by one official source, an exact match may fail against another official reference that uses a different formatting convention. These findings highlight the importance of implementing standardised formatting and preprocessing procedures when validating self-reported geographic identifiers.

Additionally, the process itself could present a labour-intensive process of data cleaning and validation. Unlike ZIP codes, census tract identifiers are not information that most individuals routinely know or use. As a result, when respondents are required to retrieve and manually enter the identifier, the process can introduce an added respondent burden and increase human error in retrieval and transcription errors. Prompting participants to retrieve unfamiliar census tract identifiers may also create confusion or heightened anxiety compared to more familiar indicators such as ZIP codes (Couper et al., 2008; Kunz & Gummer, 2025). In contrast, ZIP codes, which are also collected through self-report in this study, may continue to provide a simpler and more familiar geographic reference for participants, with more than 96% of responses successfully validated in this study.

We intentionally did not rely on automatically collected location-related paradata, such as IP addresses or GPS coordinates, to assign respondents’ geographic locations. Although paradata can reduce certain types of human error, they may introduce other sources of measurement error. For example, IP-based geolocation can produce inaccurate location estimates because of VPNs, proxy servers, mobile networks, and internet routing (Godinho et al., 2020; Prolific, 2026). Popular online survey platforms also acknowledge that IP-derived location data are often insufficient for identifying the neighbourhood-level residential location needed to determine a participant’s census tract. For example, Qualtrics estimates respondent location using IP addresses, which, according to the company, are generally accurate only to the city level within the United States and, in many cases, only to the country level internationally (Qualtrics, n.d.). Although some survey platforms offer offline applications capable of collecting precise GPS coordinates from GPS-enabled devices, this functionality is unavailable when respondents complete surveys on desktop or laptop computers.

Implications for Future Research

Despite the challenges identified in this study, future work should focus on developing standardised collection and validation procedures, along with more intuitive lookup interfaces, to improve the accuracy of self-reported census tract data while preserving respondent privacy. This recommendation is also

supported by prior research demonstrating that respondents are more willing to voluntarily provide geographic information than to permit its automatic collection (Felderer & Blom, 2022). Consistent with this perspective, the findings of the present study suggest that respondents are willing and able to provide geographic information that is familiar to them, as evidenced by the high validation rate for self-reported ZIP codes. Although validation of self-reported census tract codes was more challenging, exploratory analyses indicated that many apparent mismatches reflected formatting inconsistencies rather than inaccurate reporting, highlighting the importance of standardised collection and validation procedures.

The present study did not comprehensively examine why some respondents successfully reported valid census tract identifiers while others did not. Previous research examining respondents' willingness to share geographic information has found that both methodological and respondent characteristics influence location disclosure. For example, greater willingness to disclose geographic information has been associated with higher educational attainment, nonminority status (Olmsted-Hawala & Nichols, 2019), male gender, larger household size, and greater survey experience (Ochoa Gómez, 2022). Methodological factors, such as placing location questions toward the end of a survey, have also been shown to increase respondents' willingness to disclose geographic information (Olmsted-Hawala & Nichols, 2019). Because the location nonresponse group in this study was small, we limited descriptive comparisons to gender, where the observed pattern was consistent with previous research. Women accounted for a larger share of respondents who did not provide location information than men. Future research should build on these findings by examining the underlying causes of both location nonresponse and invalid location reporting. Such work should investigate respondent-level factors (e.g., educational attainment, digital literacy, and privacy concerns) as well as methodological factors (e.g., lookup tools, survey instructions, interface design, and data formatting conventions) that may influence the successful reporting and validation of census tract identifiers. Improved understanding of these factors could increase the accuracy of self-reported geographic information while reducing reliance on automated location data collection.

5 Limitations

Current study employed a cross-sectional survey design; therefore, the findings reflect the responses of the study sample and should not be interpreted as representative of the broader population. The survey captured respondents' behaviour and perceptions at a single point in time. Future studies employing repeated measures could examine the consistency of self-reported census tract identifiers and determine whether respondents' ability to retrieve and report their census tract improves with repeated attempts. In addition, the study sample was more highly educated than the general U.S. population.

Another limitation is that the study did not determine why some self-reported census tract identifiers failed validation after the sequence of operations we performed. Invalid entries may have resulted from retrieval errors, transcription or formatting errors, misunderstanding of the survey instructions, or intentionally inaccurate reporting due to privacy concerns. Consequently, the specific reasons underlying unsuccessful validation could not be identified. Finally, the present study evaluated only one approach to self-reported census tract retrieval, in which participants used the FFIEC Geocoding/Mapping System to locate and report their census tract. Consequently, the findings reflect the feasibility of this specific retrieval method and may not generalize to alternative lookup interfaces. Future research should directly compare these methods to determine their relative accuracy, usability, and data quality.

6 Conclusions

Online surveys offer distinct advantages over other modes of data collection, including greater speed, convenience, and cost-effectiveness. Random-digit-dial phone surveys now face sharply declining response rates alongside the reduced prevalence of landlines, while mail surveys are both slow and resource-intensive. At the same time, growing concerns about digital privacy make participants less willing to share sensitive information such as home addresses in online settings, which raises the risk of nonresponse. Passive location data collection through digital trace data also raises important methodological and ethical concerns related to respondent privacy and informed consent (Couper & Singer, 2013; Felderer & Blom, 2022). In this study, we sought to balance these challenges by collecting census tract and ZIP code data rather than full addresses, or passive data collection, providing researchers with spatial detail while reducing the intrusiveness of the request.

Taken together, our findings show that self-reported census tract data can achieve high validity, 92.2% of responses matched an official tract, but only after a systematic standardisation procedure was applied to reconcile inconsistent respondent formatting with the official census tract code. ZIP codes, by contrast, required minimal standardisation to reach an even higher match rate (97.2%), pointing to a clear trade-off between the finer spatial resolution that census tracts can offer and the additional processing burden required to make that resolution usable. Previous research has focused largely on respondents' willingness to share location data rather than the methodological challenges faced by researchers. Future research should investigate the factors contributing to invalid census tract reporting and identify strategies to improve reporting accuracy and validation.

Moving forward, we encourage the field to develop clearer guidance, more intuitive mapping interfaces, and standardised spatial workflows that support both novice and experienced researchers. Our results suggest that even a small mismatch between the format displayed by a commonly used, publicly available geocoding tool and the format used in official census reference files can meaningfully increase the processing burden required to validate self-reported geographic data. Thus, an ideal interface to retrieve codes should provide the standardised six-digit tract code directly, rather than a decimal-formatted approximation, would likely reduce both respondent confusion and downstream analytic effort.

Bringing these issues into broader scholarly conversation will help strengthen the integration of geographic context in public opinion research. Expanding beyond ZIP codes to incorporate scalable solutions can offer richer geographic context for analysing public sentiment. As online surveys continue to evolve, methodological transparency and privacy-protective practices will remain essential for building trust and generating robust spatial insights.

Author Contribution: The authors confirm sole responsibility for the study conception, data analysis, interpretation of results, and writing of the manuscript.

Competing Interests

The authors report there are no competing interests to declare.

Funding

This research was supported by the Reynolds School of Journalism, Center for Advanced Media Studies.

Ethics Approval and Consent to Participate

This study was approved by the host university's Institutional Review Board (IRB), and informed consent was obtained from all participants.

Consent for Publication

Not applicable

Data availability

De-identified or aggregated data may be made available from the corresponding author upon reasonable request.

Appendix A

Figure A1. Survey instructions for collecting self-reported geographic information

Table A1. Census tract and ZIP code match rates by respondent characteristics (N = 704)

Appendix A2

R script used for geographic validation and spatial analyses

References

1. Chan, Y.-F. Y., Wang, P., Rogers, L., Tignor, N., Zweig, M., Hershman, S. G., Genes, N., Scott, E. R., Krock, E., Badgeley, M., Edgar, R., Violante, S., Wright, R., Powell, C. A., Dudley, J. T., & Schadt, E. E. (2017). The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. *Nature Biotechnology*, 35(4), 354–362. <https://doi.org/10.1038/nbt.3826>
2. Couper, M. P., & Singer, E. (2013). Informed Consent for Web Paradata Use. *Survey Research Methods*, 7(1), 57–67.
3. Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of Official Statistics*, 24(2), 255–275.
4. Di Matteo, D., Fine, A., Fotinos, K., Rose, J., & Katzman, M. (2018). Patient Willingness to Consent to Mobile Phone Data Collection for Mental Health Apps: Structured Questionnaire. *JMIR Mental Health*, 5(3), e56. <https://doi.org/10.2196/mental.9539>
5. Egan, P. J., & Mullin, M. (2012). Turning Personal Experience into Political Attitudes: The Effect of Local Weather on Americans' Perceptions about Global Warming. *The Journal of Politics*, 74(3), 796–809. <https://doi.org/10.1017/S0022381612000448>
6. ESRI. (2026). *Add Join (Data Management Tools)*. Retrieved June 23, 2026, from <https://doc.esri.com/en/arcgis-pro/latest/tool-reference/data-management/add-join.html?tabs=dialog>
7. Evans, J. S., & Murphy, M. A. (2025). *spatialEco* (2.0-3). Retrieved October 15, 2025, from

<https://github.com/jeffreyevans/spatialEco>

8. Feddersen, J., Metcalfe, R., & Wooden, M. (2016). Subjective Wellbeing: Why Weather Matters. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 179(1), 203–228.
<https://doi.org/10.1111/rssa.12118>
9. Federal Financial Institutions Examination Council. (2024). *The FFIEC Geocoding/Mapping System*. Retrieved April 8, 2025, from <https://geomap.ffiec.gov/ffiecgeomap/>
10. Felderer, B., & Blom, A. G. (2022). Acceptance of the Automated Online Collection of Geographical Information. *Sociological Methods & Research*, 51(2), 866–886.
<https://doi.org/10.1177/0049124119882480>
11. Gadekar, M., & Brandão, A. (2025). Consumers' perceptions of proximity to plastic pollution and support for environmentally friendly retail brands: An emerging market perspective. *Journal of Retailing and Consumer Services*, 83, 104126. <https://doi.org/10.1016/j.jretconser.2024.104126>
12. Godinho, A., Schell, C., & Cunningham, J. A. (2020). Out Damn Bot, Out: Recruiting Real People into Substance Use Studies on the Internet. *Substance Abuse*, 41(1), 3–5.
<https://doi.org/10.1080/08897077.2019.1691131>
13. Grubestic, T. H. (2008). Zip codes and spatial analysis: Problems and prospects. *Socio-Economic Planning Sciences*, 42(2), 129–149. <https://doi.org/10.1016/j.seps.2006.09.001>
14. Grubestic, T. H., & Matisziw, T. C. (2006). On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics*, 5(1), 58. <https://doi.org/10.1186/1476-072X-5-58>
15. Henry, K. A., & Boscoe, F. P. (2008). Estimating the accuracy of geographical imputation. *International Journal of Health Geographics*, 7(1), 3. <https://doi.org/10.1186/1476-072X-7-3>
16. Kunz, T., & Gummer, T. (2025). Effects of objective and perceived burden on response quality in web surveys. *International Journal of Social Research Methodology*, 28(4), 385–395.
<https://doi.org/10.1080/13645579.2024.2393795>
17. Lackey, K. J., & Kaczynski, A. T. (2009). Correspondence of perceived vs. objective proximity to parks and their relationship to park-based physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, 6(1), 53. <https://doi.org/10.1186/1479-5868-6-53>
18. Moore, S., & Kestens, Y. (2011). Neighbourhood Environmental Correlates of Perceived Park Proximity in Montreal. *Canadian Journal of Public Health*, 102(3), 176–179. <https://doi.org/10.1007/BF03404890>
19. Ochoa Gómez, C. (2022). Willingness to participate in geolocation-based research. *PLOS ONE*, 17(12), e0278416. <https://doi.org/10.1371/journal.pone.0278416>
20. Olmsted-Hawala, E., & Nichols, E. (2019). Willingness of the Public to Share Geolocation Information in a U.S. Census Bureau Survey. *Social Science Computer Review*, 37(4), 568–588.
<https://doi.org/10.1177/0894439318781022>
21. Pebesma, E. (2018). Simple Features for R: Standardised Support for Spatial Vector Data. *The R Journal*, 10(1), 439. <https://doi.org/10.32614/RJ-2018-009>
22. Prolific. (2025, October 16). *Prolific's payment principles*.
<https://researcher-help.prolific.com/en/articles/445129-study-cost>
23. Prolific. (2026, February 9). *VPN guidance*. Prolific.
<https://participant-help.prolific.com/en/articles/445054-vpn-guidance>
24. Qualtrics. (n.d.). *Understanding Your Dataset*. Retrieved July 10, 2026, from
<https://www.qualtrics.com/support/survey-platform/data-and-analysis-module/data/download-data/understanding-your-dataset/>
25. Qualtrics. (2026). *Qualtrics*. Retrieved April 8, 2025 from <https://www.qualtrics.com/>
26. Singer, E. (2011). Ethical Considerations in Internet Surveys 1. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (1st ed., pp. 133–162). Routledge.
<https://doi.org/10.4324/9780203844922-6>
27. Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2021). Understanding Willingness to Share Smartphone-Sensor Data. *Public Opinion Quarterly*, 84(3), 725–759.

<https://doi.org/10.1093/poq/nfaa044>

28. Toepoel, V., & Lugtig, P. (2014). What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review*, 32(4), 544-560. <https://doi.org/10.1177/0894439313510482>
29. United States Census Bureau. (2026). *Welcome to the Geocoder!* Retrieved April 4, 2025 from <https://geocoding.geo.census.gov/geocoder/>
30. U.S. Census Bureau. (2024a). *Cartographic Boundary Files: States*. Retrieved September 8, 2025 from <https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html>
31. U.S. Census Bureau. (2024b). *Geography Program: ZIP Code Tabulation Areas (ZCTAs)*. Retrieved September 8, 2025 from <https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html>
32. U.S. Census Bureau. (2024c). *TIGER/Line Shapefiles*. Retrieved September 8, 2025 from <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html>
33. USPS. (2026). *Look Up a ZIP Code*. Retrieved July 15, 2026 from <https://tools.usps.com/zip-code-lookup.htm>
34. Vaske, J. J. (2011). Advantages and Disadvantages of Internet Surveys: Introduction to the Special Issue. *Human Dimensions of Wildlife*, 16(3), 149-153. <https://doi.org/10.1080/10871209.2011.572143>
35. Wardropper, C. B., Dayer, A. A., Goebel, M. S., & Martin, V. Y. (2021). Conducting conservation social science surveys online. *Conservation Biology*, 35(5), 1650-1658. <https://doi.org/10.1111/cobi.13747>
36. Wei, X., Bohnett, E., & An, L. (2025). Assessing U.S. public perceptions of global warming using social survey and climate data. *MethodsX*, 14, 103081. <https://doi.org/10.1016/j.mex.2024.103081>
37. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>