GeoHealth

RESEARCH ARTICLE

10.1029/2021GH000430

Area-Based Geocoding: An Approach to Exposure Assessment Incorporating Positional Uncertainty

Laura K. Thompson[1], Bryan Langholz[1], Daniel W. Goldberg[2,3], John P. Wilson[4], Beate Ritz[5], Carrie Tayour[6], and Myles Cockburn[1,4]

[1]Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [2]Department of Geography, College of Geosciences, Texas A&M University, College Station, TX, USA, [3]Department of Computer Science and Engineering, College of Geosciences, Texas A&M University, College Station, TX, USA, [4]Spatial Sciences Institute, University of Southern California, Los Angeles, CA, USA, [5]Department of Epidemiology and Environmental Sciences, Fielding School of Public Health, University of California, Los Angeles, CA, USA, [6]Los Angeles County Department of Public Health, Los Angeles, CA, USA

Key Points:

- Area-based geocoding includes the precise area of uncertainty on an address-by-address basis, which is not possible using current methods
- Geocode positional error can bias exposure and associated risk estimates in unpredictable ways
- Area-based geocoding incorporates heterogeneity of the exposure surface that likely plays a role in the magnitude and direction of biases

Correspondence to:

L. K. Thompson,
lkthomps@usc.edu

Citation:

Thompson, L. K., Langholz, B., Goldberg, D. W., Wilson, J. P., Ritz, B., Tayour, C., & Cockburn, M. (2021). Area-based geocoding: An approach to exposure assessment incorporating positional uncertainty. *GeoHealth*, *5*, e2021GH000430. https://doi.org/10.1029/2021GH000430

Received 24 MAR 2021
Accepted 5 NOV 2021

Abstract

While the spatial resolution of exposure surfaces has greatly improved, our ability to locate people in space remains a limiting factor in accurate exposure assessment. In this case-control study, two approaches to geocoding participant locations were used to study the impact of geocoding uncertainty on the estimation of ambient pesticide exposure and breast cancer risk among women living in California's Central Valley. Residential and occupational histories were collected and geocoded using a traditional point-based method along with a novel area-based method. The standard approach to geocoding uses centroid points to represent all geocoded locations, and is unable to adapt exposure areas based on geocode quality, except through the exclusion of low-certainty locations. In contrast, area-based geocoding retains the complete area to which an address matched (the same area from which the centroid is returned), and therefore maintains the appropriate level of precision when it comes to assessing exposure by geography. Incorporating the total potential exposure area for each geocoded location resulted in different exposure classifications and resulting odds ratio estimates than estimates derived from the centroids of those same areas (using a traditional point-based geocoder). The direction and magnitude of these differences varied by pesticide, but in all cases odds ratios differed by at least 6% and up to 35%. These findings demonstrate the importance of geocoding in exposure estimation and suggest it is important to consider geocode certainty and quality throughout exposure assessment, rather than simply using the best available point geocodes.

Plain Language Summary

Understanding the relationship between environmental exposures and cancer development is limited by how precisely we can locate people. While ideally all estimates would be based on building-level precision, epidemiologic research must accommodate varying levels of locational accuracy, and is dependent on input address data quality (often patient addresses). This study uses traditional point-based geocoding and a novel method of geocoding (area-based) to estimate the relationship between ambient pesticide exposure and breast cancer. Although a "point" representing a geocoded location implies precision, point coordinates can be based on anything from an exact building centroid to an entire city and may miss relevant exposure for larger areas. Using area-based geocoding, exposure estimation for an address resolved only to its ZIP Code is based on the entire ZIP Code area. We identified more individuals with potential pesticide exposure using area-based geocoding. Importantly, the proportion of exposed cases and controls was inconsistent across geocoding methods and varied by pesticide, resulting in changes in the estimated exposure-disease relationship. Geocoding quality plays a critical role in environmental exposure research, and misclassification may not be consistent or readily predictable. Methods incorporating spatial uncertainty (e.g., area-based geocoding), may shed more light on this issue and support improvements.

1. Introduction

In environmental exposure research, geocoding is typically used to convert text-based location information (i.e., addresses) into spatial data that can be overlaid onto exposure surfaces to estimate individual exposure potential. Spatially referenced data have been developed for a variety of exposures and have been effectively used to estimate exposures that would otherwise be impractical or impossible to assess on a population basis, such as traffic-related air pollutants (Ghosh et al., 2012, 2013; Wilhelm et al., 2012), lifetime ultraviolet exposure (Linos et al., 2017; Wojcik et al., 2019), ambient pesticide exposures (Cockburn et al., 2011; Costello et al., 2009; Fitzmaurice et al., 2014; Narayan et al., 2017), and even the impacts of the local built environment and greenspace on disease outcomes (Clarke et al., 2010; Gatto et al., 2009, 2010; Gomez et al., 2010; Jia et al., 2019; Manthripragada et al., 2010; Shariff-Marco et al., 2014; Twohig-Bennett & Jones, 2018). Such approaches not only allow for estimation of both current and past exposures, but they also offer efficient alternatives to traditional methods of collecting exposure data longitudinally, enhancing the utility of existing large cohorts by updating exposures when improved exposure surfaces become available (Chang et al., 2011; Heck et al., 2013; Henry et al., 2013). Despite the increased availability of high-resolution exposure data, the ability to locate individuals on these exposure surfaces continues to pose a challenge to exposure assessment research.

The impacts geocoding can have on the analysis itself have been shown to not be insignificant (Ganguly et al., 2015; Goldberg & Cockburn, 2012; Zandbergen, 2009), however geocoding accuracy is still often overlooked (Jacquez, 2012). Geocoding systems typically return a pair of latitude/longitude coordinates representing the centroid of the smallest unambiguous area derived from each input address. Depending on the completeness of an address, the accuracy of these coordinates can range from as precise as the centroid of a rooftop to as coarse as the centroid of a county (Zandbergen, 2008, 2009). Geocoding systems often return some measure of the precision associated with each geocoded address, such as a certainty score, and provide a categorical description of the matched area used as a spatial reference, such as parcel, ZIP Code tabulation area (ZCTA) or city (Sahar et al., 2019). These approaches fail to recognize variability within each of the aforementioned geographical units, geolocation, exposure surface, and/or covariate estimates.

Controlling for geocoding uncertainty by removing locations associated with less precise geographic boundaries is undesirable for several reasons. First, each subgroup is composed of a wide range of area sizes, and elimination based on boundary type fails to account for variations in coarseness. ZIP Codes in the US have more than 100-fold variability in area; segments of streets can vary in size from a city block of 100 m to rural street sections of many kilometers. Further, substantial variations in geocode quality have been shown to bias estimates (Bichler & Balchak, 2007; Goldberg & Cockburn, 2012; Oliver et al., 2005; Schootman et al., 2007; Zandbergen & Green, 2007). Prior work has found geocode quality to be spatially autocorrelated (Zimmerman & Ji, 2010), and removing or failing to account for this variation has the potential to introduce bias related to either geographic characteristics (i.e., rurality) (Krieger et al., 2002; Oliver et al., 2005) or individual-level characteristics (i.e., ethnicity or age) (Gilboa et al., 2006). If locational accuracy is not related to the exposure or outcome this will result in bias towards the null; if, as is almost always the case, locational accuracy is related to exposure or outcome, bias becomes unpredictable in both direction and magnitude (Cockburn et al., 2011).

One approach to addressing uncertainty is to manually resolve geocoded addresses using supplemental locational information (such as landmarks or street intersections) collected along with address histories (Cockburn et al., 2011; Tayour et al., 2019). While this approach represents an effective way to eliminate much of the geocoding uncertainty, it is not always possible or practical to manually resolve geocodes, especially for data sets where participant contact is not possible. Many existing data sets may not have the information needed to manually resolve locations, and newer data sources being leveraged for geolocation (e.g., credit reports indicating residential addresses) do not provide opportunities for patient contact to obtain additional spatial descriptors. Furthermore, manual resolution is human-intensive and not practical for large data sets (Goldberg, Wilson, et al., 2008).

To accommodate a wide variety of data and applications, we have developed a geocoding approach that records the exact spatial extent of each geocoded location and fully contains the spatial range in which the address (and therefore the study participant) is known to be located. Instead of limiting analysis to only the most precise data (e.g., records resolved to address points or tax assessor parcel level), which will likely introduce either selection bias, measurement error, or both, this approach allows every record to be used in analysis, accompanied by the appropriate information on variability for both exposure and covariates. Instead of relying on a single point to represent all geocoded locations regardless of the quality and size of the area, this approach returns the complete linear (in the case of street segments) or areal feature matched to each location. These features can then be compared against the exposure surface in a GIS to calculate exposure based on known location. In the case of rooftop-level accuracy, exposure measurements for the centroid point and the parcel area may not differ substantially, but as geocode quality worsens, differences in exposure potential may become large enough to impact results, particularly if the exposure surface is not uniform across these relatively large areas.

The aim of this study was to compare exposure to a select group of pesticides using: (a) locations geocoded to points; and (b) locations geocoded to areal or linear features. Both approaches to geocoding were applied to data used in a prior case-control study evaluating the role of pesticides on the development of breast cancer among women living in California's Central Valley (Tayour et al., 2019). The availability of long-term Pesticide Use Reports (PUR; CalEPA, 2013) in California makes it possible to quantitatively measure exposure to etiologically relevant chemicals that may be present in the air surrounding an individual's home or workplace as a result of pesticide drift (Goldberg, Zhang, et al., 2007; Rull & Ritz, 2003; Tayour et al., 2019). Pesticide exposure is spatially heterogenous, and as a result represents a useful test case for the evaluation of these geocoding methods.

2. Materials and Methods

2.1. Study Setting

California is the most agriculturally productive state in the US, and accounts for about one quarter of all pesticides used in the country (CalEPA, 2011). Kern, Tulare, and Fresno counties rank as the top three counties for agricultural density and commercial pesticide use in the US (CalEPA, 2011), and are relevant to this study due to the potential for pesticides applied in these counties to drift into residential areas, causing ambient residential pesticide exposures (Cockburn et al., 2011). In highly agricultural regions, pesticide drift from neighboring application sites presents a major source of non-occupational exposure (Deziel et al., 2015; Harnly et al., 2009; Lu et al., 2000; Quirós-Alcalá et al., 2011).

2.2. Study Population

This case-control study utilized a study population recruited for a previous study on breast cancer and ambient pesticide exposure and has been previously described in detail (Tayour et al., 2019). The study population includes 155 cases and 150 controls living in Kern, Tulare, and Fresno counties in the Central Valley of California at the time of diagnosis (cases) or interview (controls). Cases were recruited in 2011 through 2013 from the California Central Cancer Registry, aged 55–74 years, of non-Hispanic white ethnicity, and with postmenopausal histologically confirmed breast cancer diagnosed in 2007 and 2008.

Controls were originally recruited for a study of Parkinson's disease in the same geographic area from 2001 through 2011. Controls were recruited from Medicare listings, randomized mailings to tax assessor parcel addresses using Internet searches, and marketing companies. Beginning in 2009, additional controls were recruited during randomized home visits. More information on control recruitment can be found elsewhere (Narayan et al., 2013; Wang et al., 2014). From this population, we identified postmenopausal women, aged 55–74, of non-Hispanic white ethnicity. Women who had been diagnosed with breast cancer or other female reproductive cancers were excluded, along with those who opted to complete a shortened questionnaire without lifetime residential and occupational histories. Both cases and controls were excluded if they did not live in California for at least five years prior to recruitment. To determine exposure, cases and controls were asked to recall their complete residential and occupational histories up to diagnosis or interview, to complete a job history questionnaire, and to participate in a comprehensive risk factor phone interview. No P.O. Boxes were collected as part of residential or occupational histories.

2.3. Geocoding

Address histories for cases and controls were geocoded using the Texas A&M geocoding service (http://geoservices.tamu.edu/Services/Geocode), which follows the traditional organization of a geocoding system detailed many places in the literature (e.g., Goldberg, Wilson, & Knoblock, 2007; Sahar et al., 2019). Input postal addresses are standardized to the USPS Publication 28 format using a deterministic approach wherein the input address is split into tokens based on whitespace. Each token is tagged with a set of candidate address fields (street name, pre/post directional, suffix, etc.). A rule-based approach is used to assign the most likely address field to each address token based on the tags associated to itself and the tags for tokens which precede and follow it. A deterministic feature matching algorithm generates a series of SQL queries which are issued to the reference data sources by combining the raw and Soundex value for the street name series of tokens with combinations of the USPS Zip Code and City (along with alias names for each city, where applicable). The NAACCR-recommended ordering of GIS Coordinate Quality Codes (Havener & Thornton, 2008, p. 169, Data Item #366) is used to prioritize queries to the set of reference layers—authoritative address point reference sources collected by the team over a number of years from commercial and public sources would be queried first. If no suitable match is found, commercial and public parcel and street centerline layers collected by the team are queried (in that order) to attempt to locate an address level match. If a street level match cannot be found, USPS ZIP Codes and US Census ZCTA reference sources are used, followed by US Census Places, Minor Civil Divisions, Counties, and State layers (each in this respective order). The addresses contained in the reference feature layers are standardized using the same approach described above.

This geocoding platform uses the deterministic match scoring approach (Goldberg, 2008, p. 71) whereby each candidate match from each reference source found by the system is assigned a match score using a deterministic weight-based approach. In this model, the values for each component of an address are assigned a weight. A non-perfect agreement between the components of an input address and reference feature will undergo a result penalty computation determined via string similarity measures including edit distance, word length, synonyms, and others. The penalties for each component are multiplied by the weight for that component and aggregated to compute an overall match score for each candidate. An 88% match score threshold is used as a cutoff to define a potential valid candidate as a "match." Valid point level matches are returned directly as candidate results; address-range interpolation is used to compute linear-interpolation based results for street centerline matches; and center-of-mass calculations are used to produce an output for areal unit-based matches (ZIP Codes, parcels, etc.). All candidates from a reference source are scored simultaneously, and the candidate with the highest score at or above the acceptable match score (88%) is returned as the result if one exists. If one does not exist, the system continues searching through the rest of the reference layer hierarchy in the order described above. In cases where the input address is matched to multiple features with the same geographic resolution, the geocoding service returns the most precise geographic feature shared by all possible matches (Goldberg, Swift, & Wilson, 2008). For example, if an address was matched to two street segment features within a common ZCTA, the ZCTA was be returned, because this is the smallest known geography that unambiguously describes the location. If these street segments were located in separate ZCTAs but were within the same city, the city is returned (Goldberg, Swift, & Wilson, 2008).

In additional to returning a pair of coordinates associated with each matched feature's centroid or point location (point-based), users can also request the underlying spatial data (areas or lines) for each matched feature from which the coordinates are derived (area-based). The matched feature remains the same regardless of the output type selected. As described above, possible feature matches include address point, parcel, street segment, US Postal Service (USPS) ZIP Code, ZCTA, city, county sub-region, and county and are dependent on the accuracy or completeness of each input address. For the point-based method, all these matches are represented using the centroid of the matched geographic boundary. For the area-based version, the entire extent of the matched geographic boundary is returned as either a linear (street segments) or areal feature. Because an address point is not associated with an area, address points were manually joined to their respective parcels for area-based analyses.

We evaluated exposure using all available addresses for point-based and area-based geocodes. We also compared these two approaches to a selective point-based geocoding approach, a commonly used method to control for geocode quality by evaluating exposure only for address-years associated with high-certainty geocoded locations (exact address points, parcels, and street segments). Because all individuals had at least one address with a high-certainty geocoded location, no individuals were removed from this selective point-based analysis.

2.4. Selection of Pesticides

We evaluated a group of organochlorine pesticides with known estrogenic effects that are most likely related to breast carcinogenesis (aldrin, chlordane, dicofol, dieldrin, endosulfan, lindane, methoxychlor, and toxaphene) (Coumoul et al., 2001; Kojima et al., 2004; Lemaire et al., 2006; Louis et al., 2017; Soto & Sonnenschein, 2010; Valerón et al., 2009) as well as three pesticides (chlorpyrifos, diazinon and 1,3-dichloropropene) detected at levels of concern to human health in air monitoring conducted over one year in a farming community in Fresno County (Wofford et al., 2014). Endosulfan and Dicofol are responsible for 82.6% of all organochlorine exposures in our study. In addition, several pesticides with highly similar chemical profiles and estrogenic properties were also included in our organochlorine group because they are used too infrequently to be examined for exposure independently (Kaushik & Kaushik, 2007).

2.5. Exposure Surface Generation

Exposures were determined using GIS-Based Residential Ambient Pesticide Estimation System (GRAPES) software, version 4.2 (University of Southern California, Spatial Sciences Institute, Los Angeles, CA). GRAPES uses a combination of California land use data and pesticide use reports (PUR), which document the active ingredient of the pesticide applied, the amount of pesticide applied in pounds, the acreage and crop type of the field, and the date and method of application (California Department of Pesticide Regulation, 2011; Rull & Ritz, 2003). Each pesticide application is associated with the Public Land Survey System (PLSS) section in which it was applied. This PLSS grid, which has a spatial resolution of about one square mile, is then associated with crop field locations identified by the land use data to improve the spatial resolution of the exposure surface (Rull & Ritz, 2003).

To determine exposure, GRAPES overlays each geocoded address onto the data described above and exposure is calculated within a user-defined buffer around the point, line or polygon associated with the address. A buffer distance of 500-meters was chosen to maintain consistency with previously published work (Tayour et al., 2019). The GRAPES output was used to categorize individuals as either unexposed or exposed (at either residences or workplaces). Participants were classified as exposed if they had any exposure for the time period 1984 to interview or diagnosis.

2.6. Statistical Analysis

The relationship between ambient pesticide exposure and breast cancer in postmenopausal women was estimated using unconditional logistic regression. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated for study participants exposed to specific pesticides compared to those not exposed. An individual was considered exposed to a pesticide when the pounds per acre of applied pesticide within the buffer area was greater than zero at either their residence, workplace, or both during the period from 1984 until the year of diagnosis for cases and the year of interview for controls. All analyses were conducted using SAS, version 9.3 software (SAS Institute, Inc., Cary, NC).

Years with missing residential addresses due to incomplete recall or unresolved geocodes were imputed with the average exposure across all known years for each person (Weinberg et al., 1996). Gaps in workplace histories where women did not report addresses because they were unemployed, worked from home, cared for children at home, or were retired or disabled were imputed with that participant's residential exposure for the respective time frame. Pesticide exposure could not be identified for addresses outside of California since the exposure model includes only California PUR data; therefore, these locations were considered unexposed. The institutional review boards at the California Health and Human Services Agency and the University of Southern California approved the study protocol for cases participating in this study. The Institutional Review Board at the University of California, Los Angeles approved the study protocol for controls used in this study. Informed consent was obtained for all participants.

3. Results

3.1. Address Completeness, Geocoding, and Geocode Precision

Geocode feature matches for each person's address-years are found in Table 1. In this table, there are 4,079 yearly residential locations for cases and 4,011 for controls, and yearly occupational location totals of 4,221 for cases and 4,154 for controls. These counts include instances of multiple addresses per year during years where a residential or occupational move occurred (exposure was averaged in these instances). Cases had an average of 2.4 residences and 3.5 workplaces and controls had an average of 3.2 residences and 4.2 workplaces from 1984 to year of diagnosis or interview. Because geocode quality is closely related to area size, differences in geocode quality across cases and controls may impact exposure potential. More than 70% of residential case and control addresses matched to address points/parcels or street segments, the two most precise feature types. Overall address quality was lower for occupational addresses, with roughly 60% of addresses matching to address points/parcels or street segments. Average area sizes for all addresses differed between cases and controls by approximately 8.1 km^2 for residential addresses and 9.3 km^2 for occupational addresses, with controls less precise in both cases (Mann-Whitney p-value <0.001 and <0.01, respectively).

Table 1
Area Size Statistics for Each Geocoded Address (Multiplied by the Number of Years it was Used to Determine Individual Exposure), Grouped by Geography Type

Geocoded address	Area[a] (km^2)						
Geography type	*n*	%	Mean	Median	Std. Dev.	Min	Max
Cases, Residential							
Address point/parcel	2,650	64.97	0.03	0.0009	0.14	0.0001	0.79
Street segment	287	7.04	0.01	0.0044	0.01	0.0023	0.04
ZCTA/USPS ZIP	466	11.42	250.78	194.73	285.52	0.32	992.21
City	650	15.94	116.35	36.00	151.84	0.79	1,362.45
County	12	0.29	3,667.53	2,856.30	2,192.64	92.33	7,063.88
Unmatchable	14	0.34	–	–	–	–	–
Overall	4,079	100.00	58.20	0.0032	267.82	0.0001	7,063.88
Controls, residential							
Address point/parcel	2,733	68.14	0.0069	0.0008	0.04	0.0001	0.79
Street segment	280	6.98	0.0095	0.0043	0.02	0.0004	0.08
ZCTA/USPS ZIP	342	8.53	278.89	84.20	322.11	0.84	1,229.54
City	509	12.69	171.42	60.29	238.12	0.12	1,362.45
County	33	0.82	2,290.66	1,572.61	2,583.45	1,406.97	16,037.94
Unmatchable	114	2.84	–	–	–	–	–
Overall	4,011	100.00	66.27	0.0012	349.30	0.0001	16,037.94
Cases, occupational							
Address point/parcel	2,134	50.56	0.05	0.0021	0.15	0.0001	1.07
Street segment	321	7.60	0.01	0.0043	0.03	0.0004	0.13
ZCTA/USPS ZIP	551	13.05	308.82	204.61	344.23	0.32	1,229.54
City	1,150	27.24	174.53	100.23	192.44	0.79	1,362.45
County	7	0.17	2,777.22	2,335.57	1,986.22	698.83	7,063.88
Unmatchable	58	1.37	–	–	–	–	–
Overall	4,221	100.00	93.78	0.06	238.30	0.0001	7,063.88
Controls, occupational							
Address point/parcel	2,006	48.29	0.04	0.0014	0.17	0.0001	2.62
Street segment	421	10.13	0.0049	0.0032	0.0043	0.0004	0.02
ZCTA/USPS ZIP	473	11.39	218.84	39.41	284.83	1.18	1,229.54
City	1,074	25.85	230.85	290.88	243.61	0.12	1,362.45
County	41	0.99	1,518.49	1,127.62	2,379.34	445.43	16,037.94
Unmatchable	139	3.35	–	–	–	–	--
Overall	4,154	100.00	103.06	0.02	338.60	0.0001	16,037.94

[a] All areas presented in this table were calculated prior to adding the 500 m buffer used in exposure assessment. The area of a 500 m buffered point is 0.79 km^2. 500 m buffers added to larger features will result in a greater area growth than smaller features.

3.2. Exposure Classification Differences Between Area- and Point-Based Approaches

As expected, inclusive point-based geocoded locations classified fewer individuals as exposed than area-based geocoded locations. However, the magnitude of this difference varied by pesticide (Table 2). Across all participants ($n = 305$), chlorpyrifos experienced the smallest percent change in number of individuals exposed (−63 people, −25.4%), while 1,3-dichloropropene experienced the greatest percent change in number of individuals exposed (−121 people, −59.6%). For organochlorines and diazinon, we found slightly greater changes in exposure prevalence than for chlorpyrifos (−84 people, 29.1% and −73 people, −33.3%, respectively). Selective

Table 2
A Comparison of Area-Based Geocoding to Inclusive and Selective Point-Based Geocoding Methods to Estimate Breast Cancer Risk From Ambient Exposure to Selected Pesticides Based on Residential and Occupational Address Histories Among Women in Kern, Tulare, and Fresno Counties Using Linked Pesticide Use Reports Data for 1984–2011

	Cases		Controls		Crude OR		
Exposure	No. (%)	Change in exposure	No. (%)	Change in exposure	Estimate	95% CI	Change in OR
Organochlorines[a]							
Area-based							
Unexposed	23 (14.8)		31 (20.7)		1.00		
Exposed	132 (85.2)		119 (79.4)		1.50	0.83, 2.71	
Point-based, inclusive							
Unexposed	56 (36.1)	143.5%	71 (47.3)	129.0%	1.00		
Exposed	99 (63.9)	−25.0%	79 (52.6)	−33.6%	1.59	1.01, 2.51	6.0%
Point-based, selective[b]							
Unexposed	80 (51.6)	247.8%	86 (57.3)	177.4%	1.00		
Exposed	75 (48.4)	−43.2%	64 (42.7)	−46.2%	1.26	0.80, 1.98	−16.0%
Chlorpyrifos							
Area-based							
Unexposed	26 (16.8)		31 (20.7)		1.00		
Exposed	129 (83.2)		119 (79.4)		1.29	0.73, 2.30	
Point-based, inclusive							
Unexposed	51 (32.9)	96.2%	69 (46.0)	122.6%	1.00		
Exposed	104 (67.1)	−19.4%	81 (54.0)	−31.9%	1.74	1.09, 2.76	34.9%
Point-based, selective[b]							
Unexposed	71 (45.8)	173.1%	84 (56.0)	171.0%	1.00		
Exposed	84 (54.2)	−34.9%	66 (44.0)	−44.5%	1.51	0.96, 2.37	17.1%
Diazinon							
Area-based							
Unexposed	24 (15.5)		29 (19.3)		1.00		
Exposed	131 (84.5)		121 (80.7)		1.31	0.72, 2.37	
Point-based, inclusive							
Unexposed	66 (42.6)	175.0%	71 (47.3)	144.8%	1.00		
Exposed	89 (57.4)	−32.1%	79 (52.7)	−34.7%	1.21	0.77, 1.90	−7.6%
Point-based, selective[b]							
Unexposed	92 (59.4)	283.3%	84 (56.0)	189.7%	1.00		
Exposed	63 (40.7)	−51.9%	66 (44.0)	−45.5%	0.87	0.55, 1.37	−33.6%
1,3-Dichloropropene							
Area-based							
Unexposed	53 (34.2)		49 (32.7)		1.00		
Exposed	102 (65.8)		101 (65.2)		0.93	0.58, 1.50	
Point-based, inclusive							
Unexposed	113 (72.9)	113.2%	110 (73.3)	124.5%	1.00		
Exposed	42 (27.1)	−58.8%	40 (26.7)	−60.4%	1.02	0.62, 1.70	9.7%

Table 2
Continued

	Cases		Controls		Crude OR		
Exposure	No. (%)	Change in exposure	No. (%)	Change in exposure	Estimate	95% CI	Change in OR
Point-based, selective[b]							
Unexposed	129 (83.2)	143.4%	122 (81.3)	149.0%	1.00		
Exposed	26 (16.8)	−74.5%	28 (18.7)	−72.3%	0.88	0.49, 1.58	−5.4%

[a]Aldrin, chlordane, dicofol, dieldrin, endosulfan, lindane, methoxychlor, and toxaphene. [b]Selective point-based geocoding refers to the standard approach of retaining only high-certainty geocodes (parcels and street segments) and excluding low-certainty geocodes.

point-based geocoding reduced the number of exposed individuals even further since it excluded the exposures for all years associated with low-certainty geocoded locations.

Table 3 illustrates the impact selective point-based geocoding had on the residential history data sets by case/control status. Selective point-based geocoding disproportionately removed addresses dated more than a decade prior to diagnosis (or recruitment). Selective point-based geocoding also tended to remove more addresses among women who were older at diagnosis; Cases and controls over the age of 65 (the median age at diagnosis or recruitment across all participants) had 1 and 2 more low-certainty person-years, respectively, than those diagnosed at age 65 or younger. Further, despite cases and controls having similar proportions of low-certainty person-years (32.2% and 28.2%, respectively), 31.3% of low-certainty person-years among cases were located in micropolitan and rural areas, while only 20.5% of low-certainty person-years among controls were located in micropolitan and rural areas (USDA Rural-Urban Commuting Area Codes, 1990, 2000, and 2010 [United States Department of Agriculture, 2021]).

Examples of exposure estimate changes for select low-certainty residential geocoding results near the city of Fresno are shown in Figure 1. These five example geocoding results matched to either ZCTAs (Z1, Z2, Z3) or cities (C1, C2), two relatively low-certainty geocoding categories. By considering the entire ZCTA or city associated with each address (the smallest/most precise feature matched to the input address), we observe substantial opportunities for exposure to organochlorines in 2003, particularly within Z1 and Z2. However, exposure classifications using buffered centroid points would classify the geocoded locations matched to Z1 and Z2 as unexposed. Regardless of the geocoding type used (point or area), we know only that the address falls somewhere within the ZCTA. Point-based geocoding leaves spatial information (and thus exposure potential) on the table that can be incorporated into exposure assessment using area-based geocoding.

In contrast, because chlorpyrifos exposure is relatively uniform across Z1 and Z2, the point estimate did correctly classify exposure to chlorpyrifos. Z1 and Z2 were classified as exposed to chlorpyrifos using both point- and area-based geocoding results. Population-weighted centroids, when available, may reduce the likelihood of exposure misclassification for point-based geocodes, but still only represent one of many possible points within often large areas of uncertainty. Note that applying a selective point-based approach (excluding low-certainty geocodes from analysis) would exclude all the example locations in Figure 1, which has the same effect as classifying all the geocoded locations in Figure 1 as unexposed.

Figure 1 also demonstrates the potential for exposure misclassification due to the range of area sizes within a geographic boundary type: The area of Z3 is substantially smaller than Z1 and Z2 because it is associated with a relatively population-dense area of Fresno, while Z1 and Z2 refer to less densely populated areas. Though all three addresses resolved to ZCTAs, the likelihood of exposure misclassification is substantially higher for Z1 and Z2.

3.3. Exposure Misclassification and Odds Ratio Estimates

In addition to geocoding method influencing the numbers of individuals classified as exposed, the proportions of individuals exposed also differed by case status across each of the three geocoding strategies, to varying degrees depending on the pesticide. Using entire matched feature areas to determine exposure estimation, we find elevated breast cancer risk estimates associated with exposure to some pesticides, but none of the odds ratio estimates were statistically significant. More than 75% of cases and controls were found to have potential exposure

Table 3
Count of Person-Years With Low-Certainty Geocodes by Case-Control Status, Location Type, and Time to Diagnosis (Under or Over 10 Years)

	Total person-years		Unmatchable geocodes[a]		Low-certainty geocodes[b]	
	Cases	Controls	Cases	Controls	Cases	Controls
Residential						
0–10 years before dx	1,550	1,500	0 (0%)	10 (0.7%)	247 (15.9%)	181 (12.1%)
>10 years before dx	2,252	2,123	8 (0.4%)	35 (1.7%)	686 (30.5%)	542 (25.5%)
Occupational						
0–10 years before dx	1,550	1,500	10 (0.7%)	9 (0.6%)	476 (30.7%)	360 (24.0%)
>10 years before dx	2,252	2,123	44 (2.0%)	58 (2.7%)	1,040 (46.2%)	957 (45.1%)

Note. Low-certainty geocoded locations do not include unmatchable geocodes.

[a] Person-year has no addresses which could be geocoded (address was "unmatchable" in geocoder). [b] Person-year has no addresses which could be geocoded to a parcel or street segment.

to organochlorines, chlorpyrifos, and diazinon at their residences, workplaces or both locations during the study period (Table 2). Breast cancer was more likely to occur among women exposed to organochlorines (OR = 1.50; 95% CI: 0.83, 2.71), chlorpyrifos (OR = 1.29: 95% CI: 0.73, 2.30) and diazinon (OR = 1.31; 95% CI: 0.72, 2.37), but confidence intervals were fairly wide and none of these estimated effects were statistically significant. In contrast, breast cancer was slightly less likely to occur among individuals exposed to 1,3-dichloropropene, though this effect estimate was very close to null and also characterized by wide confidence intervals (OR = 0.93; 95% CI: 0.58, 1.50).

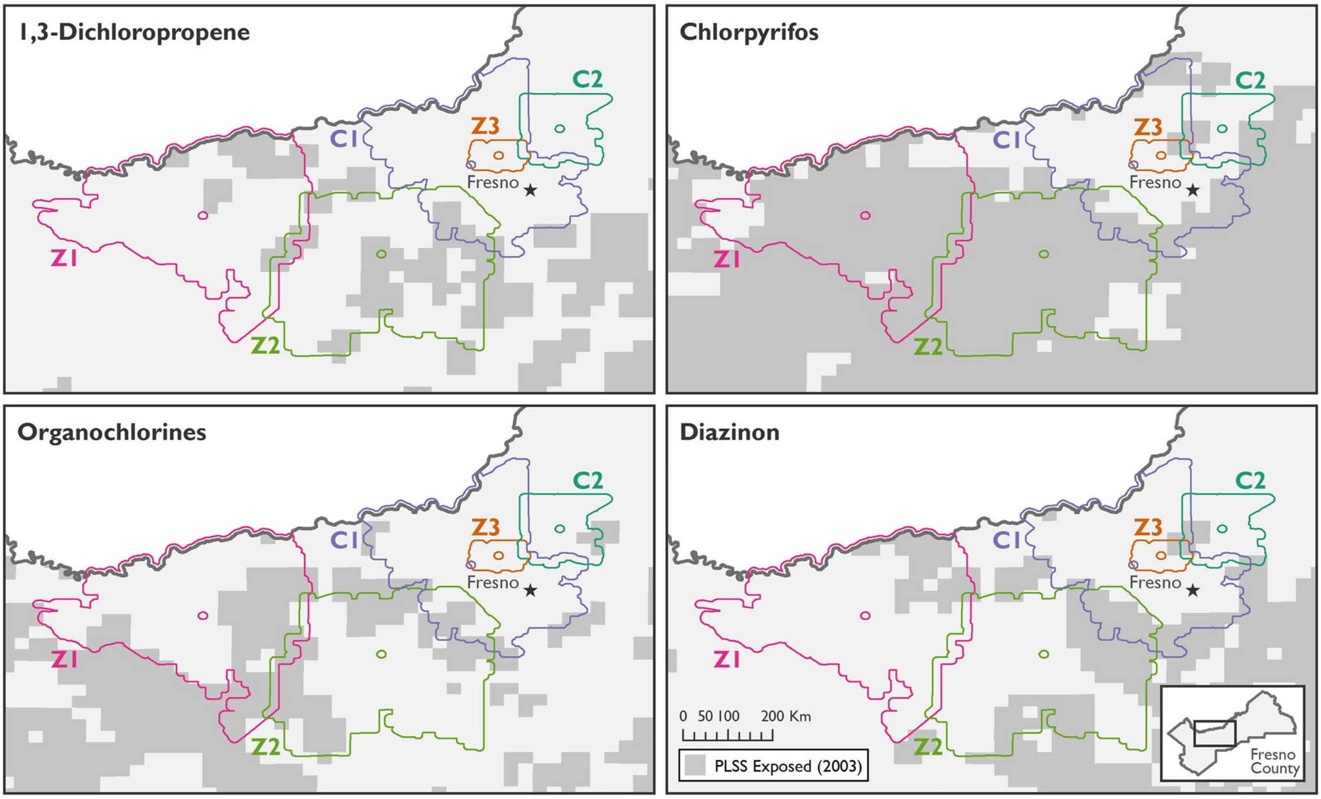

Figure 1. Examples of low-certainty point and area-based geocoding results near the city of Fresno, overlaid on Public Land Survey System (PLSS) based pesticide exposure for 1 year (2003). Geocoded location boundaries labeled with a "Z" represent ZCTA matches, those labeled with a "C" represent city matches. The smallest ZCTA area shown here is 27 km^2, and the largest ZCTA area shown is 482 km^2.

In contrast to area-based geocoding, point-based geocoding relies on a uniform feature (i.e., a centroid point buffered by 500 m) to estimate exposure across all locations, regardless of the matched feature and its associated spatial precision (or lack thereof). The percentage of individuals exposed to a given pesticide were lower for point-based geocoded locations, with a 19% reduction in the percentage of cases exposed to chlorpyrifos on the low end, and a 60% reduction in the percentage of controls exposed to 1,3-dichloropropene on the hight end (Table 2). Breast cancer was statistically significantly more likely to occur among women exposed to organochlorines (OR = 1.59; 95% CI: 1.01, 2.51) and chlorpyrifos (OR = 1.74; 95% CI: 1.09, 2.76) using exposure determined from point locations. Across both of the aforementioned pesticides, odds ratio estimates were elevated and confidence intervals were narrower than those associated with area-based estimates. Contrastly, the odds ratio estimate for diazinon was slightly attenuated compared to the area-based estimate (OR = 1.21; 95% CI: 0.77, 1.90). The marginally protective effect estimate associated with area-based exposure to 1,3-dichloropropene was absent using point-based geocoded locations. Exposure to 1,3-dichloropropene was not found to be associated with breast cancer (OR = 1.02; 95% CI: 0.62, 1.70).

Odds ratio estimates for breast cancer risk were consistently attenuated when based on selective point-based geocoded locations versus an inclusive data set of point-based geocoded locations. Across three of the four pesticides, odds ratios based on selective point-based geocoded locations were also lower than area-based estimates. Breast cancer was more likely to occur among women exposed to chlorpyrifos (OR = 1.51; 95% CI: 0.96, 2.37), though this effect estimate was not statistically significant and fell between inclusive point- and area-based estimates. Breast cancer risk had a slight association with exposure to Organochlorines (OR = 1.26; 95% CI: 0.80, 1.98), but again this effect was not significant and was attenuated compared to estimates based on the other two geocoding methods. A slight protective effect was found for exposure to both Diazinon (OR = 0.87; 95% CI: 0.55, 1.37) and 1,3-dichloropropene (OR = 0.88; 95% CI: 0.49, 1.58).

4. Discussion

This case-control study evaluated the impact of geocoding and geocode uncertainty on the measurement of ambient pesticide exposure and its role in breast cancer development. Though prior work has identified the need to develop approaches to incorporate geocoded data of varying qualities into analyses, this is the first study to incorporate the spatial resolution of geocoding results into exposure assessment, accounting for the true variability in geocoding certainty by allowing the spatial unit used to determine exposure to vary according to geocoding quality. In contrast to limited prior work focused on the development of statistical tools aimed at reducing error introduced by low-certainty geocoding results (Zimmerman, 2008), the approach described here addresses uncertainty on a case-by-case basis for each address, incorporating the precise area of uncertainty generated by each geocoding result. We elected to use a 500-m buffer for points and areas in the present study, but acknowledge that different buffer sizes would result in different exposure estimates.

The traditional approach of using a single buffered point to represent geocoded locations is inherently subject to misclassification, implying that all addresses are geocoded to an exact point location. Except under perfect conditions, point locations are derived from larger areas of varying sizes, and analyses based on geocoded point locations fail to account for the geocoding precision gradient. Though slightly more informative, categorical descriptions of geocode precision fail to recognize and account for the wide range of area sizes within each category. Locations resolving to cities or ZCTAs tend to be larger than those resolving to streets or exact parcels or address points, but each of these categories encompasses a wide range of sizes.

Treating each geocoded location equally (using a point-based approach) or selectively removing low-certainty categories of geocoded locations (e.g., ZCTA and city matches) can not only result in the removal of informative data, it it also likely to generate selection bias (Gilboa et al., 2006; Oliver et al., 2005). Variations in the spatial precision of geocoded locations tend to be unevenly distributed across population characteristics (i.e., covariates). If, for example, individuals representing a particular demographic (age, racial/ethnic group, socioeconomic status) disproportionately reside in less population-dense areas, their geocoded locations are likely to be represented by larger geographic units (particularly those which prioritize population over area, such as ZCTAs). Using a single point to represent those areas is more likely to result in exposure misclassification for some categories of the covariate.

If the highest precision achieved by a geocoding service for an address is the ZCTA of that address, it is unlikely that the (unknown) true location of that address is sufficiently near the centroid of that ZCTA to be adequately represented by that interpolated centroid point and its buffer, resulting in spatial aggregation error (Hewko et al., 2002; Rushton et al., 2006). Though ZCTAs are often proportional to population density, their relationship to exposure patterns is likely to vary, generating variability within areas that is not captured using a single point. When using a point-based approach, the probability that the centroid point represents where a person actually resides, and therefore the chance that their exposure status was correctly classified using that centroid, is lower for larger areas (i.e., exposure classification for individuals residing in larger rural ZCTAs vs. smaller, urban ZCTAs). This is especially problematic when the area is substantially larger than the size of the exposure buffer applied around the point, increasing the likelihood that the true location falls entirely outside of the buffer from which exposure is measured. In the case of point-based geocoding, it is possible to misclassify exposure for an address in either direction (classifying unexposed as exposed or vice versa). In contrast, area-based geocoded locations guarantee the inclusion the true location of each address through the retention of the entire area in which an address is known to be located. Thus, we can confirm that an individual classified as unexposed was truly unexposed and that this exposure attribution was not made due to locational error when placing the buffer around the centroid point. However, since this area-based approach maximizes the sensitivity of exposure at the expense of specificity it tends to systematically move the effect estimate toward the null.

Adapting the geocoding service to return the complete areas from which these points are derived provides a more appropriate representation of the known precision of a geocoded location, which is reflected in the wider confidence intervals seen when estimating risk using area-based locations. Although it may seem counter-intuitive to suggest that a larger area is more accurate than a single point, in each case the same knowledge about location is known. If a single, interpolated point is used to represent locations regardless of whether it is derived from a parcel or a ZCTA, we fail to incorporate the appropriate level of certainty about geolocation, even though both methods result in exposure misclassification.

4.1. Variability in Geographic Units and Sizes

Whether using a point or area-based approach to geocoding, geocoding certainty should be similar across cases and controls. Regarding the point-based approach, if cases and controls have distinct distributions of geocoded location quality and one group is resolving to centroids of larger areas, it is more likely that the estimates of risk will be a product of geocoding quality differences rather than true exposure differences (Bichler & Balchak, 2007; Goldberg & Cockburn, 2012; Oliver et al., 2005; Schootman et al., 2007; Zandbergen & Green, 2007). Using either the point- or area-based approach, individuals with low-certainty addresses are more likely to be misclassified as exposed, though only point-based geocoding can incorrectly classify individuals as unexposed. In the present study, controls have larger area sizes than cases on average for both residential (66.27 and 58.20 km^2, respectively) and occupational locations (103.06 and 93.78 km^2, respectively). This is partly influenced by the presence of one large county-matched area greater than 10,000 km^2. Removing this county match from residential and occupational control areas reduced the area mean to 62.17 and 99.09 km^2. Because this creates more opportunity for exposure among controls, this difference in the location of controls may have resulted in slightly attenuated odds ratios. While both approaches have limitations and produce answers which are influenced by geocode quality, area-based geocoding is able to capture these uncertainties while point-based geocoding is not.

The ability to quantify the degree of uncertainty associated with each geocoded locations and the study populations in aggregate allows for an examination of potential bias that would not be possible if a geocoding service provided only centroid points and categorical descriptions of the matched feature type. The sizes of areal units matched by a geocoding system vary both within and between classes of geographic features. ZIP codes in urban areas are small and in rural areas are large. In contrast, cities are large in urban areas and small in rural areas. In agricultural regions, even parcels can be of large size. So, the feature type (ZIP, parcel, city, etc.) alone cannot be used to assess geocode quality since no single class of geographic objects have a minimum or maximum size or relation to the size of other classes. As demonstrated by Table 1, using the frequencies of categorical descriptions to estimate locational certainty would suggest controls have higher-quality geocoded locations because they have a slightly higher proportion of address point, parcel and street segment matches.

4.2. Exclusion of Low-Certainty Geocodes

In the present study, we observe some examples of the bias that may be introduced by excluding low-certainty geocodes from exposure assessment (selective point-based geocoding). Low-certainty geocodes disproportionately occurred among older addresses (those lived and worked in more than 10 years prior to diagnosis). Although more likely to have incomplete information due to recall and changes to the built environment, more temporally distant exposures (occurring decades prior to diagnosis) may be the most etiologically relevant to breast cancer (Verner et al., 2008) and it is important to incorporate historical data into exposure research (Alavanja et al., 2013; Rodgers et al., 2018). Selective point-based geocoding also excluded slightly more addresses among women who were older at diagnosis (or recruitment), and similar bias associated with individual-level characteristics has also been observed in prior work (Gilboa et al., 2006).

We also found that a disproportionate amount of the excluded low-certainty addresses among cases were located in micropolitan (15.8% vs. 9.0% for high-certainty locations) and rural (15.5% vs. 8.6%) census tracts, with fewer low-certainty addresses occurring in metropolitan locations. The exclusion of addresses based on geocode quality was more likely to remove rural locations among cases than among controls (or than metropolitan locations among cases). Pesticide use is more prevalent in rural areas, so if there is a true association between pesticides and breast cancer, this rural-urban bias in exclusion likely results in the underestimation of exposure in cases but not controls, and may be a contributing factor to the attenuation of odds ratios associated with selective point-based geocoding in this analysis. This rural-urban bias in geocode quality has been noted elsewhere (Krieger et al., 2002; Oliver et al., 2005).

4.3. Variability in Exposure Misclassification by Case Status and Pesticide

The magnitude of exposure misclassification between point- and area-based geocoding varied by pesticide, as shown in Table 2 and demonstrated in Figure 1. Because the points and areas used to determine exposure consist of the same locations, the only differences between point- and area-based exposure classifications are due to the use of "area" as a proxy for geolocation certainty in the latter. Area-based geocoding will classify an individual as exposed if there is any exposure potential in the geocoded area, whether that area is a residential parcel or an entire county. Point-based geocoding will classify an individual as exposed (or unexposed) based on exposure at the arbitrary centroid point of the area. Within either geocoding method, all participants resolved to the same ZIP Code, County, etc., will be assigned the same exposure status for a given year.

Area-based geocoding (with low-certainty geocodes) is likely to overestimate exposure when compared to the true (unknown) point locations, while point-based geocoding (with low-certainty geocodes) can over or underestimate exposure. However, area-based geocoding does not overestimate exposure in comparison to point-based geocoding unless the arbitrary centroid (point-based) is unexposed, but there is exposure in the geocoded area—in which case a classification of exposed is the best answer based on what is known about the location (e.g., exposure to organochlorines or diazinon in Z1 and Z2 in Figure 1).

If the change in exposure classification within each pesticide was consistent across case and control groups, it would suggest that point-based geocoding and ignoring geocoding uncertainty does not impact our ability to estimate the effect of pesticide exposure on breast cancer risk (aside from a predictable attenuation of risk estimates due to non-differential misclassification). Instead, once we consider the complete area location for each address, we identify several more individuals with exposure potential that was missed using centroid point locations. The degree of misclassification varied by case status. For example, compared to area-based geocoded locations (inclusive), point-based estimates of exposure to chlorpyrifos were 31.9% lower for controls, and only 19.4% lower for cases.

Across all pesticides, using precise points to represent low-certainty locations resulted in odds ratios which changed by at least 6% and up to 35%. The pesticide with the greatest difference in number of individuals exposed (1,3-dichloropropene) experienced a relatively moderate change in its breast cancer odds ratio estimate (9.7%). For organochlorines and chlorpyrifos, point-based geocoding resulted in higher and statistically significant odds ratio estimates, with increases in point estimates of 6.0% and 34.9%, respectively. While the estimated effect for both of these pesticides remained higher, area-based geocoding identified additional exposure opportunities (slightly more so among controls) that were previously not detected, and this increase in sensitivity resulted in attenuated odds ratios and widened confidence intervals, suggesting we lack the spatial precision needed to confirm a statistically significant difference between cohorts. In contrast with the aforementioned pesticides, the estimated effect for breast cancer risk associated with diazinon exposure became stronger using area-based geocoded locations (vs. points). Differences in magnitude and direction of this change in breast cancer risk estimates occurred for all four pesticides, suggesting that exposure misclassification due to geocoding uncertainty affects the odds ratios for the outcome in an inconsistent manner that cannot easily be predicted, modeled or adjusted for.

Though not a perfect solution, using area-based geocoding we have removed some of the opportunity for misclassification to occur by relaxing the requirement that all geocoded locations need to be represented with a singular point regardless of quality. Truly unexposed locations will be correctly classified as such, and all exposure potential will be captured regardless geocode certainty. While this does result in widened confidence intervals, these confidence intervals represent the true spatial precision of our geocoded locations.

We have established that point-based geocoded locations do not capture potential locational variability when used to assign exposures, predominantly because the location of the point is unlikely to fall at the true location of the address when geocoding is incomplete (i.e., when an exact address match cannot be found by the geocoding service). While area-based exposure geocoding addresses this issue through the retention of complete geocoded feature areas, it does not provide the high level of spatial precision that characterizes high-resolution exposure surfaces. For more coarse exposure pathways, or those with a relatively uniform distribution across the study area, differences between geocoders may be smaller. However, in the case of pesticides (and likely many other exposures), geocoding uncertainty, when unaccounted for by centroid points, leads to exposure misclassification and affects risk estimates in an unpredictable manner.

4.4. Implications for Ongoing Work

The development of an approach to geocoding that incorporates the true known area for each address ensures our ability to capture complete exposure potential for each geocoded location and to correctly assign those who are unexposed. Compared to point-based geocoding, confidence intervals accurately reflect the level of specificity we are able to obtain based on address quality. Area-based geocoding also provides an opportunity to investigate the relationship between the exposure surface and geocoding certainty, since the amount of pesticide applied (in pounds of chemical) can be determined for an area. A limitation of this approach is that, if the underlying data used are characterized by a high degree of locational uncertainty (thus relying on large areal features for exposure assessment), our ability to detect differences in exposure between study populations may be reduced. The next step for refining this approach is to improve the resolution of these areas. This can be achieved through the use of spatially referenced population data (such as census blocks and/or building footprints) to refine geocoded areas to portions of those areas where the population is found. This approach has been described in Langholz et al. (2020). If high-resolution population data is unavailable for a study area, a distance decay weighting function could be used to assign exposure values based on distance from the centroid of a feature. The centroid point minimizes positional error and receives the highest weighting, while the exposure values of areas closer to the edge are given incrementally lower influence.

5. Conclusions

We introduce an alternate approach to geocoding which retains the complete area range for each geocoded location. Point-based geocoding is unable to effectively incorporate information about the spatial uncertainty of geocoded locations and produces exposure estimates based on false spatial precision and characterized by narrow confidence intervals that cannot adapt to varying geocode certainty conditions. Selective point-based geocoding, often used in an attempt to address geocode uncertainty, is likely to introduce bias due to individual and area-level characteristics which are more likely to result in low-certainty geocoded locations. Area-based geocoding reduces or eliminates the need to exclude geocoded data with low-certainty, and reduces the potential for bias to be introduced by the removal of lower-quality geocoded locations. This approach may be especially important for data that cannot be manually resolved, does not contain complete addresses information, or requires historic address information that may be subject to poor recall. Area-based geocoding is able to capture complete exposure potential and does not rely on centroid points that are, especially for larger geographic boundaries, unlikely to represent actual location. The use of area-based geocodes often resulted in wider confidence intervals, which represent an appropriate reduction in certainty. While area-based estimates lie within the confidence intervals of our original point-based estimates, the difference between risk estimates varies by pesticide and differs by case status. Prior work using point-based geocoded locations do provide estimates of the appropriate magnitude, but the use of area-based geocoded locations adds additional understanding of variability, particularly in this relatively small data set. For studies with the ability to manually correct geocodes or with exact address matches across all locations, the geocode feature type used (points or areas) would not affect exposure estimates.

Though more research is needed in this area, our work suggests that the magnitude and direction of exposure misclassification resulting from the use of point-based geocoded locations varies by exposure (pesticides in this case) and can vary by case status, favoring a solution that accounts for geocoding uncertainty at the individual address level and which does not rely on a generalized model (e.g., categorial removal of geocoded locations as seen with selective point-based geocoding). To further improve the area-based approach, Langholz et al. (2020) has developed a method for population-weighted exposure assessment within larger geographic boundaries. This approach uses census block population data to identify populated locations within these areas and calculate exposure for these populated areas, effectively refining area-based exposure assessment (Langholz et al., 2020). Next steps for this work include applying this approach to a case-control data set such as the breast cancer data. Continued work aimed at improving and evaluating methods accounting for locational uncertainty is important to further understanding of the impacts of geocoding method on exposure assessment and to inform ongoing research on the role pesticides in the development of certain cancers.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Patient data supporting this research were provided by the California Cancer Registry. Controls were recruited as part of a separate study. More information on control recruitment can be found elsewhere (Narayan et al., 2013; Wang et al., 2014). Data are not publicly available but can be requested by researchers for approved studies. Requests should be made by the principal investigator on behalf of the institution with which the investigator is affiliated. Pesticide use data are available from the California Department of Pesticide Regulation (https://www.cdpr.ca.gov/docs/pur/purmain.htm), and land use data are available from the California Department of Water Resources (https://gis.water.ca.gov/app/CADWRLandUseViewer/).

Acknowledgments
The authors would like to acknowledge the University of Southern California Norris Comprehensive Cancer Center's Population Research Core for their support of this research. This work was supported by the U.S. National Institute of Environmental Health Sciences (grant numbers 5P30ES007048, ES019986, ES018960) and the U.S. National Cancer Institute (grant number CA195218) on behalf of the investigator, Myles Cockburn.

References

Alavanja, M. C., Ross, M. K., & Bonner, M. R. (2013). Increased cancer burden among pesticide applicators and others due to pesticide exposure. *CA: A Cancer Journal for Clinicians*, *63*, 120–142. https://doi.org/10.3322/caac.21170

Bichler, G., & Balchak, S. (2007). Address matching bias: Ignorance is not bliss. *Policing: An International Journal of Police Strategies & Management*, *30*(1), 32–60. https://doi.org/10.1108/13639510710725613

California Environmental Protection Agency (CalEPA), California Department of Pesticide Regulation. (2011). *Pesticide use reporting—2010 Summary data*. Retrieved from https://www.cdpr.ca.gov/docs/pur/purmain.htm

California Environmental Protection Agency (CalEPA), California Department of Pesticide Regulation. (2013). *Pesticide use reporting pesticide information portal (CalPIP)*. Retrieved from http://www.cdpr.ca.gov/docs/pur/purmain.htm

Chang, E. T., Canchola, A. J., Cockburn, M., Lu, Y., Wang, S. S., Bernstein, L., et al. (2011). Adulthood residential ultraviolet radiation, sun sensitivity, dietary vitamin D, and risk of lymphoid malignancies in the California Teachers Study. *Blood*, *118*(6), 1591–1599. https://doi.org/10.1182/blood-2011-02-336065

Clarke, C., Moy, L., Swetter, S., Zadnick, J., & Cockburn, M. (2010). Interaction of area-level socioeconomic status and ultraviolet radiation on melanoma occurrence in California. *Cancer Epidemiology, Biomarkers & Prevention*, *19*(11), 2727–2733. https://doi.org/10.1158/1055-9965.EPI-10-0692

Cockburn, M., Mills, P., Zhang, X., Zadnick, J., Goldberg, D., & Ritz, B. (2011). Prostate cancer and ambient pesticide exposure in agriculturally intensive areas in California. *American Journal of Epidemiology*, *173*(11), 1280–1288. https://doi.org/10.1093/aje/kwr003

Costello, S., Cockburn, M., Bronstein, J., Zhang, X., & Ritz, B. (2009). Parkinson's disease and residential exposure to maneb and paraquat from agricultural applications in the central valley of California. *American Journal of Epidemiology*, *169*(8), 919–926. https://doi.org/10.1093/aje/kwp006

Coumoul, X., Diry, M., Robillot, C., & Barouki, R. (2001). Differential regulation of cytochrome P450 1A1 and 1B1 by a combination of dioxin and pesticides in the breast tumor cell line MCF-7. *Cancer Research*, *61*(10), 3942–3948.

Deziel, N., Friesen, M., Hoppin, J., Hines, C., Thomas, K., & Freeman, L. (2015). A review of nonoccupational pathways for pesticide exposure in women living in agricultural areas. *Environmental Health Perspectives*, *123*(6), 515–524. https://doi.org/10.1289/ehp.1408273

Fitzmaurice, A. G., Rhodes, S. L., Cockburn, M., Ritz, B., & Bronstein, J. M. (2014). Aldehyde dehydrogenase variation enhances effect of pesticides associated with Parkinson disease. *Neurology*, *82*(5), 419–426. https://doi.org/10.1212/WNL.0000000000000083

Ganguly, R., Batterman, S., Isakov, V., Snyder, M., Breen, M., & Brakefield-Caldwell, W. (2015). Effect of geocoding errors on traffic-related air pollutant exposure and concentration estimates. *Journal of Exposure Science & Environmental Epidemiology*, *25*(5), 490–498. https://doi.org/10.1038/jes.2015.1

Gatto, N. M., Cockburn, M., Bronstein, J., Manthripragada, A. D., & Ritz, B. (2009). Well-water consumption and Parkinson's disease in rural California. *Environmental Health Perspectives*, *117*(12), 1912–1918. https://doi.org/10.1289/ehp.0900852

Gatto, N. M., Rhodes, S. L., Manthripragada, A. D., Bronstein, J., Cockburn, M., Farrer, M., & Ritz, B. (2010). α-Synuclein gene may interact with environmental factors in increasing risk of Parkinson's disease. *Neuroepidemiology*, *35*(3), 191–195. https://doi.org/10.1159/000315157

Ghosh, J. K., Heck, J. E., Cockburn, M., Su, J., Jerret, M., & Ritz, B. (2013). Prenatal exposure to traffic-related air pollution and risk of early childhood cancers. *American Journal of Epidemiology*, *178*(8), 1233–1239. https://doi.org/10.1093/aje/kwt129

Ghosh, J. K., Wilhelm, M., Su, J., Goldberg, D., Cockburn, M., Jerrett, M., & Ritz, B. (2012). Assessing the influence of traffic-related air pollution on risk of term low birth weight on the basis of land-use-based regression models and measures of air toxics. *American Journal of Epidemiology*, *175*(12), 1262–1274. https://doi.org/10.1093/aje/kwr469

Gilboa, S. M., Mendola, P., Olshan, A. F., Harness, C., Loomis, D., Langlois, P. H., et al. (2006). Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research*, *101*(2), 256–262. https://doi.org/10.1016/j.envres.2006.01.004

Goldberg, D. W. (2008). *A geocoding best practices guide*. North American Association of Central Cancer Registries.

Goldberg, D. W., & Cockburn, M. G. (2012). The effect of administrative boundaries and geocoding error on cancer rates in California. *Spatial and Spatio-temporal Epidemiology*, *3*(1), 39–54. https://doi.org/10.1016/j.sste.2012.02.005

Goldberg, D. W., Swift, J. N., & Wilson, J. P. (2008). *Geocoding best practices: Reference data, input data, and feature matching*. *USC GIS Research Laboratory Technical Report No. 8*. Retrieved from https://spatial.usc.edu/wp-content/uploads/2014/03/gislabtr8.pdf

Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From text to geographic coordinates: The current state of geocoding. *URISA Journal*, *19*(1), 33–46.

Goldberg, D. W., Wilson, J. P., Knoblock, C. A., Ritz, B., & Cockburn, M. G. (2008). An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics*, *7*, 60. https://doi.org/10.1186/1476-072X-7-60

Goldberg, D. W., Zhang, X., Wilson, J. P., Ritz, B., & Cockburn, M. G. (2007). Development of an automated pesticide exposure analyst for the California's central valley. In *Proceedings of the URISA GIS in Public Health Conference, May 20–23, 2007, New Orleans, Louisiana*.

Gomez, S. L., Quach, T., Horn-Ross, P. L., Pham, J. T., Cockburn, M., Chang, E. T., et al. (2010). Hidden breast cancer disparities in Asian women: Disaggregating incidence rates by ethnicity and migrant status. *American Journal of Public Health*, *100*(S1), S125–S131. https://doi.org/10.2105/AJPH.2009.163931

Harnly, M. E., Bradman, A., Nishioka, M., McKone, T. E., Smith, D., McLaughlin, R., et al. (2009). Pesticides in dust from homes in an agricultural area. *Environmental Science and Technology*, *43*, 8767–8774. https://doi.org/10.1021/es9020958

Havener, L., & Thornton, M. (Eds.). (2008). *Standards for Cancer Registries Volume II: Data Standards and Data Dictionary* (13th ed., Version 11.3). North American Association of Central Cancer Registries.

Heck, J. E., Wu, J., Lombardi, C., Qiu, J., Meyers, T. J., Wilhelm, M., et al. (2013). Childhood cancer and traffic-related air pollution exposure in pregnancy and early life. *Environmental Health Perspectives*, *121*(11–12), 1385–1239. https://doi.org/10.1093/aje/kwt129

Henry, K. A., Sherman, R., Farber, S., Cockburn, M., Goldberg, D. W., & Stroup, A. M. (2013). The joint effects of census tract poverty and geographic access on late-stage breast cancer diagnosis in 10 US States. *Health and Place*, *21*, 110–121. https://doi.org/10.1016/j.healthplace.2013.01.007

Hewko, J., Smoyer-Tomic, K. E., & Hodgson, M. J. (2002). Measuring neighbourhood spatial accessibility to urban amenities: Does aggregation error matter? *Environment and Planning A*, *34*(7), 1185–1206. https://doi.org/10.1068/a34171

Jacquez, G. M. (2012). A research agenda: Does geocoding positional error matter in health GIS studies? *Spatial and Spatio-temporal Epidemiology*, *3*(1), 7–16. https://doi.org/10.1016/j.sste.2012.02.002

Jia, P., Xue, H., Cheng, X., Wang, Y., & Wang, Y. (2019). Association of neighborhood built environments with childhood obesity: Evidence from a 9-year longitudinal, nationally representative survey in the US. *Environment International*, *128*, 158–164. https://doi.org/10.1016/j.envint.2019.03.067

Kaushik, P., & Kaushik, G. (2007). An assessment of structure and toxicity correlation in organochlorine pesticides. *Journal of Hazardous Materials*, *143*, 102–111. https://doi.org/10.1016/j.jhazmat.2006.08.073

Kojima, H., Katsura, E., Takeuchi, S., Niiyama, K., & Kobayashi, K. (2004). Screening for estrogen and androgen receptor activities in 200 pesticides by in vitro reporter gene assays using Chinese hamster ovary cells. *Environmental Health Perspectives*, *112*, 524–531. https://doi.org/10.1289/ehp.6649

Krieger, N., Waterman, P., Chen, J. T., Soobader, M. J., Subramanian, S. V., & Carson, R. (2002). Zip code caveat: Bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas–the Public Health Disparities Geocoding Project. *American Journal of Public Health*, *92*(7), 1100–1102. https://doi.org/10.2105/ajph.92.7.1100

Langholz, B., Escobedo, L. A., Goldberg, D. W., Heck, J. E., Thompson, L. K., Ritz, B., & Cockburn, M. (2020). Analysis of case-control data when there is geolocation uncertainty. *Spatial Statistics*, 100486. https://doi.org/10.1016/j.spasta.2020.100486

Lemaire, G., Mnif, W., Mauvais, P., Balaguer, P., & Rahmani, R. (2006). Activation of alpha- and beta-estrogen receptors by persistent pesticides in reporter cell lines. *Life Sciences*, *79*(12), 1160–1169. https://doi.org/10.1016/j.lfs.2006.03.023

Linos, E., Li, W., Han, J., Li, T., Cho, E., & Qureshi, A. (2017). Lifetime ultraviolet radiation exposure and lentigo maligna melanoma. *British Journal of Dermatology*, *176*, 1666–1668. https://doi.org/10.1111/bjd.15218

Louis, L. M., Lerro, C. C., Friesen, M. C., Andreotti, G., Koutros, S., Sandler, D. P., et al. (2017). A prospective study of cancer risk among Agricultural Health Study farm spouses associated with personal use of organochlorine insecticides. *Environmental Health*, *16*(1), 95. https://doi.org/10.1186/s12940-017-0298-1

Lu, C., Fenske, R. A., Simcox, N. J., & Kalman, D. (2000). Pesticide exposure of children in an agricultural community: Evidence of household proximity to farmland and take home exposure pathways. *Environmental Research*, *84*, 290–302. https://doi.org/10.1006/enrs.2000.4076

Manthripragada, A. D., Costello, S., Cockburn, M. G., Bronstein, J. M., & Ritz, B. (2010). Paraoxonase 1, agricultural organophosphate exposure, and Parkinson disease. *Epidemiology*, *21*(1), 87–94. https://doi.org/10.1097/EDE.0b013e3181c15ec6

Narayan, S., Liew, Z., Bronstein, J. M., & Ritz, B. (2017). Occupational pesticide use and Parkinson's disease in the Parkinson Environment Gene (PEG) study. *Environment International*, *107*, 226–273. https://doi.org/10.1016/j.envint.2017.04.010

Narayan, S., Liew, Z., Paul, K., Lee, P. C., Sinsheimer, J. S., Bronstein, J. M., & Ritz, B. (2013). Household organophosphorus pesticide use and Parkinson's disease. *International Journal of Epidemiology*, *42*, 1476–1485. https://doi.org/10.1093/ije/dyt170

Oliver, M. N., Matthews, K. A., Siadaty, M., Hauck, F. R., & Pickle, L. W. (2005). Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics*, *4*, 29. https://doi.org/10.1186/1476-072X-4-29

Quirós-Alcalá, L., Bradman, A., Nishioka, M., Harnly, M. E., Hubbard, A., McKone, T. E., et al. (2011). Pesticides in house dust from urban and farmworker households in California: An observational measurement study. *Environmental Health*, *10*, 19. https://doi.org/10.1186/1476-069X-10-19

Rodgers, K. M., Udesky, J. O., Rudel, R. A., & Brody, J. G. (2018). Environmental chemicals and breast cancer: An updated review of epidemiological literature informed by biological mechanisms. *Environmental Research*, *160*, 152–182. https://doi.org/10.1016/j.envres.2017.08.045

Rull, R. P., & Ritz, B. (2003). Historical pesticide exposure in California using pesticide use reports and land-use surveys: An assessment of misclassification error and bias. *Environmental Health Perspectives*, *111*(13), 1582–1589. https://doi.org/10.1289/ehp.6118

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmerman, D. L. (2006). Geocoding in cancer research: A review. *American Journal of Preventive Medicine*, *30*(S2), S16–S24. https://doi.org/10.1016/j.amepre.2005.09.011

Sahar, L., Foster, S. L., Sherman, R. L., Henry, K. A., Goldberg, D. W., Stinchcomb, D. G., & Bauer, J. E. (2019). GIScience and cancer: State of the art and trends for cancer surveillance and epidemiology. *Cancer*, *125*, 2544–2560. https://doi.org/10.1002/cncr.32052

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology*, *17*(6), 379–470. https://doi.org/10.1016/j.annepidem.2006.10.015

Shariff-Marco, S., Yang, J., John, E. M., Sangaramoorthy, M., Hertz, A., Koo, J., et al. (2014). Impact of neighborhood and individual socioeconomic status on survival after breast cancer varies by race/ethnicity: The neighborhood and breast cancer study. *Cancer Epidemiology, Biomarkers & Prevention*, *23*(5), 793–811. https://doi.org/10.1158/1055-9965.EPI-13-0924

Soto, A. M., & Sonnenschein, C. (2010). Environmental causes of cancer: Endocrine disruptors as carcinogens. *Nature Reviews, Endocrinology*, *6*(7), 363–370. https://doi.org/10.1038/nrendo.2010.87

Tayour, C., Ritz, B., Langholz, B., Mills, P. K., Wu, A., Wilson, J. P., et al. (2019). A case–control study of breast cancer risk and ambient exposure to pesticides. *Environmental Epidemiology*, *3*(5), e070. https://doi.org/10.1097/EE9.0000000000000070

Twohig-Bennett, C., & Jones, A. (2018). The health benefits of the great outdoors: A systematic review and meta-analysis of greenspace exposure and health outcomes. *Environmental Research*, *166*, 628–637. https://doi.org/10.1016/j.envres.2018.06.030

United States Department of Agriculture (USDA), Economic Research Service. (2021). *Rural-urban commuting area codes*. Retrieved from https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/

Valerón, P. F., Pestano, J. J., Luzardo, O. P., Zumbado, M. L., Almeida, M., & Boada, L. D. (2009). Differential effects exerted on human mammary epithelial cells by environmentally relevant organochlorine pesticides either individually or in combination. *Chemico-Biological Interactions*, *180*(3), 485–491. https://doi.org/10.1016/j.cbi.2009.04.010

Verner, M. A., Charbonneau, M., López-Carrillo, L., & Haddad, S. (2008). Physiologically based pharmacokinetic modeling of persistent organic pollutants for lifetime exposure assessment: A new tool in breast cancer epidemiologic studies. *Environmental Health Perspectives*, *116*, 886–892. https://doi.org/10.1289/ehp.10917

Wang, A., Cockburn, M., Ly, T. T., Bronstein, J. M., & Ritz, B. (2014). The association between ambient exposure to organophosphates and Parkinson's disease risk. *Occupational and Environmental Medicine*, *71*, 275–281. https://doi.org/10.1136/oemed-2013-101394

Weinberg, C. R., Moledor, E. S., Umbach, D. M., & Sandler, D. P. (1996). Imputation for exposure histories with gaps, under an excess relative risk model. *Epidemiology*, *7*(5), 490–497. https://doi.org/10.1097/00001648-199609000-00007

Wilhelm, M., Ghosh, J. K., Su, J., Cockburn, M., Jerrett, M., & Ritz, B. (2012). Traffic-related air toxics and term low birth weight in Los Angeles County, California. *Environmental Health Perspectives*, *120*(1), 132–138. https://doi.org/10.1289/ehp.1103408

Wofford, P., Segawa, R., Schreider, J., Federighi, V., Neal, R., & Brattesani, M. (2014). Community air monitoring for pesticides. Part 3: Using health-based screening levels to evaluate results collected for a year. *Environmental Monitoring and Assessment*, *186*(3), 1355–1370. https://doi.org/10.1007/s10661-013-3394-x

Wojcik, K. Y., Escobedo, L. A., Wysong, A., Heck, J. E., Ritz, B., Hamilton, A. S., et al. (2019). High birth weight, early UV exposure, and melanoma risk in children, adolescents, and young adults. *Epidemiology*, *30*(2), 278–284. https://doi.org/10.1097/EDE.0000000000000963

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, *32*(3), 214–232. https://doi.org/10.1016/j.compenvurbsys.2007.11.006

Zandbergen, P. A. (2009). Geocoding quality and implications for spatial analysis. *Geography Compass*, *3*(2), 647–680. https://doi.org/10.1111/j.1749-8198.2008.00205.x

Zandbergen, P. A., & Green, J. W. (2007). Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives*, *115*(9), 1363–1370. https://doi.org/10.1289/ehp.9668

Zimmerman, D., Ji, J., & Fang, X. (2010). Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in Medicine*, *29*, 1025–1036. https://doi.org/10.1002/sim.3836

Zimmerman, D. L. (2008). Estimating the intensity of a spatial point process from locations coarsened by incomplete geocoding. *Biometrics*, *64*(1), 262–270. https://doi.org/10.1111/j.1541-0420.2007.00870.x