Published in final edited form as:
Spat Spatiotemporal Epidemiol. 2012 April ; 3(1): 39–54. doi:10.1016/j.sste.2012.02.005.

The Effect of Administrative Boundaries and Geocoding Error on Cancer Rates in California

Daniel W. Goldberg[1] and **Myles G. Cockburn**[2]

Daniel W. Goldberg: dwgoldbe@usc.edu; Myles G. Cockburn: myles@med.usc.edu

[1]University of Southern California, Spatial Sciences Institute, Los Angeles CA.

[2]University of Southern California, Department of Preventive Medicine, Los Angeles CA.

Abstract

Geocoding is often used to produce maps of disease rates from the diagnosis addresses of incident cases to assist with disease surveillance, prevention, and control. In this process, diagnosis addresses are converted into latitude/longitude pairs which are then aggregated to produce rates at varying geographic scales such as Census tracts, neighborhoods, cities, counties, and states. The specific techniques used within geocoding systems have an impact on where the output geocode is located and can therefore have an effect on the derivation of disease rates at different geographic aggregations. This paper investigates how county-level cancer rates are affected by the choice of interpolation method when case data are geocoded to the ZIP code level. Four commonly used areal unit interpolation techniques are applied and the output of each is used to compute crude county-level five-year incidence rates of all cancers in California. We found that the rates observed for 44 out of the 58 counties in California vary based on which interpolation method is used, with rates in some counties increasing by nearly 400% between interpolation methods.

1. Introduction

Geocoding, or the process of translating textual information most commonly in the form of postal addresses into geographic locations, is typically one of the first geocomputational processes applied to enable spatially-based health science investigations[1, 2]. This process provides health scientists with the ability to place individuals and groups within a spatiotemporal context from which questions related to aspects such as geographic barriers to health services can be posed and investigated [3, 4]. The use of geocoding within the health sciences has a long history ranging from disease surveillance and outbreak monitoring[5–10] to epidemiological investigations into the role and impacts that environmental exposures have on human health[11–18].

Throughout this long history of using geocoding tools and geocoded data for health science research, numerous researchers have identified many limitations in using these data in scientific studies. These limitations are typically broken along two axes: (1) spatial accuracy of the geographic location computed for any particular subject – the distance between the output computed and the true location[18–26]; and (2) match rate achieved when geocoding a large set of records– the number of records capable of being assigned output geocodes

[19]. Each of these issues has been investigated on numerous occasions with researchers consistently finding both to be non-randomly distributed across space, time, and population-specific characteristics. Ignoring either of these issues typically results in studies containing geographic bias, potentially invalidating the results[13, 23, 27–29]. The reason for this can be seen in the standard data pipeline utilized in scientific practices displayed in [Figure 1. The geocoding process and the geocoded data it produces are the underlying source upon which all subsequent analyses are performed and conclusions are drawn. Therefore, any errors in the production of these data are propagated throughout the rest of the scientific pipeline.

As indicated on several occasions by numerous authors, the internal intricacies of a geocoding system are known to greatly influence the quality of the results that can be obtained, affecting both spatial accuracy and match rates[9, 19, 21–23, 30–32]. Most geocoding systems in use today for health science applications can be considered black boxes, where information or details describing the internal processing algorithms and/or data sources used in the process are not described. This most often stems from commercial reasons – to protect one's product line, ensure long term utilization of the system, and maintain a competitive advantage, it behooves a geocoding service or software provider to release none but the most general of details about the platform. In practice, this means that few details other than the reference data used are typically available.

This lack of technical detail and/or transparency of a vendor's geocoding process would not be a problem if geocoding was a simple, straightforward, error free process. However, this is simply not the case as evidenced by the broad range of academic disciplines that have contributed to our understanding and the development of geocoding techniques including but not limited to geography and geographic information science[20, 31, 33, 34], computer science[24, 35–37], and mathematics and statistics[38–41]. Published research reports describe several competing geocoding techniques (see [42] for a review), and prior work has described how a one-size-fits-all geocoding approach is simply not appropriate[19, 21–23]. To begin, a single input data set may exhibit a variety of characteristics known to perform better with particular geocoding techniques due to the nature of geographic features and addressing systems in a region[26, 43, 44]. For example, if a data set simultaneously contains both urban and rural records, records for urban addresses where parcels are small will perform better using parcel reference layers, while rural records may perform better using street centerline reference files because parcels are large[23, 27–29]. Likewise, different techniques and/or reference data sources may be appropriate given the time period covered by an input data set to be geocoded, with historical records perhaps favoring historical reference data layers[11, 45–49]. Finally, some techniques may simply be non-applicable for a particular region because the required data sets just do not exist[50, 51]. For example, if a digital parcel file is not available in a specific county, parcel-level geocoding approaches are just not an option. Similarly, if a specific area of a rural county has not yet upgraded to be E-911 compliant, it may be the case that Rural Route addresses are the only street address-like information available, despite the well-known limitations of using these data for geocoding[18, 42, 52–54].

The point here is that despite nearly 30 years of efforts devoted by a large group of researchers from a broad range of disciplines starting from the earliest systems at the US Census Bureau[33, 34, 55, 56], it is a fact that the process of geocoding is still error prone and results in output with varying degrees of spatial inaccuracy and some amount of non-matching records that affect match rates. Further, every geocoding system in use today will produce a geocode output that is different than every other geocoding system. Research reports show that these differences can be extremely small, or extremely large, but in all cases depend on a myriad of intertwined aspects inherent to the geocoding algorithms and

reference data layers available to and used within any particular geocoding process[19–23, 26, 30, 50, 57–59].

The primary goals of this paper are two-fold. The first is to continue to reinforce the importance of understanding how one's geocoded data are computed so that one may intelligently reflect on whether one's data are suitable for the use at hand. The second is to provide an example of how the choice of geocoding tools and techniques can have impacts on subsequent analyses. To achieve these goals, this paper presents an investigation on the effects of administrative boundaries on the derivation of cancer rates in California. In particular, this paper examines how the use of different techniques for computing the centroid of an areal unit can have dramatic impacts on five year incidence rates of all cancers reported at the county-level. The variability in county-level cancer rates (for all cancers) is illustrated through the use of four different commonly-employed areal unit interpolation techniques.

The term "interpolation" is used herein to mean the process of computing a single output point from an input linear or areal unit geographic object. While the term interpolation has other meanings, most notably to estimate the value of an attribute at a location where observations are not available, our use herein follows that from the geocoding literature which describe the feature interpolation component of a geocoding process[42, 60–62]. In these definitions, the feature interpolation component is used across situs, linear, and areal unit features to describe processes that manipulate geographic reference features using zero, one, or more computations to produce a final geographic output for a geocode. In this same light, we consider the computation of an areal unit's centroid an application of an minterpolation method. Although one can construct a definition of "interpolation" for which computing a centroid would not be an application, the terminology used herein follows directly from existing literature on geocoding processes[42, 60–62].

The remainder of this paper is organized as follows. Section 2 provides background on cancer rates and reporting in California, a discussion of the interpolation options used in geocoding systems, and details related to the problems associated with computing geocodes using areal units, specifically ZIP codes. Section 3 describes the data sources and methods used. Sections 4 presents the results achieved. Section 5 offers a discussion of interesting findings and concludes the paper with final remarks.

2. Background

2.1 Cancer Rates and Reporting in California and the US

As in many states[7], the State of California has legislation in place that requires the operation of a state-wide cancer registry system, the California Cancer Registry (CCR). The goal of this system is to provide a population-based cancer surveillance system for use in cancer prevention, treatment, and control. The CCR collects and provides data on cancer in California to facilitate demographic and geographic analysis of factors that affect cancer risk, early detection, and effective treatment of cancer patients as well as to help determine where early detection, educational, and other cancer-related programs should be directed[63].

This state-level entity is responsible for collecting, integrating, and assembling the county-level cancer incidence data from each of three regional cancer registries in California (Los Angeles, San Francisco/Bay Area, and Greater California)into a single state-wide database. Each of these three regional registries gathers county-specific data for each of the 58 California counties it is responsible for. To do so, each registry performs various tasks including finding, recording, abstracting, and reporting all cancer cases within their

jurisdiction to the CCR. These data are gathered from hospitals, clinics, doctor's offices and other diagnosis and treatment facilities. Upon discovery, these data are cleaned, abstracted (where a patient's record is analyzed and "coded" to associate specific characteristics of treatment, diagnosis, demographics, etc. with the record) and eventually submitted to the CCR. In this manner, the CCR and the regional-level registries work together to provide an accurate, complete, and up-to-date, picture of the burden of cancer in the State of California.

Before and after submission to the CCR by each of the regional registries, these data are processed for quality control and quality assurance using a series of automated and manual editing processes[64]. These checks are used to identify and correct known limitations in reporting procedures, inconsistencies in cancer abstracts (the resulting synthesized and coded report describing the cancer incidence case from all materials discovered about the case), and highlight unlikely combinations of abstract details that may indicate a high likelihood of an erroneous data value. At this stage, other processing techniques are applied including geocoding to associate a geographic coordinate with the diagnosis address of the cancer abstract. Geocoding for the CCR is performed by a third party vendor, the University of Southern California Spatial Sciences Institute, using the USC Geocoding System [65]. The details of this system are described elsewhere so only those relevant to the current experiments will be described in detail in Section 3. In addition to associating a latitude and longitude pair for the diagnosis address, the USC Geocoding System also computes and associates various US Census Bureau variables related to the diagnosis address including the Census block group, tract, city, and county within which the diagnosis address falls.

Once processed, these data are used for cancer surveillance, prevention, and control at many administrative levels including, among others, local-scale prevention efforts, state-wide policy initiatives, and national-level activities[1, 66]. To enable the last of these, portions of the CCR data are transferred to national-level entities. In order to maintain confidentiality of cancer incidence data, detailed geographical identifiers (such as latitude and longitude) are not provided to national-level identities. Instead, geographically aggregated data are provided. In most cases, the only geographical identifier provided to national-level entities is the census tract (CT). One national-level organization is the North American Association of Central Cancer Registries (NAACCR), a national body composed of institutional and individual cancer registry members spanning North America (including Canada)[67]. Among other roles, this organization is responsible for creating and implementing cancer data standards, coordinating large-scale multi-state research projects, and facilitating the derivation and dissemination of best practices with regard to data collection, organization, and analysis.

2.2 Interpolation Techniques and Areal Unit Geocoding

Geocoding techniques typically fall within one of three classes based on the type of geographic objects contained within the geographic reference layers used by the system: situs, linear, and areal unit features[42, 54, 62].

Situs Features—The first, those that use situs – single point – reference data features include pre-derived lists of building centroids or address points as would be contained in E-911 databases[1, 53]. Here, the geocoding process simply has to identify the most likely match between the input address and the geographic items in the reference data layers and return the matched geographic object directly; no form of interpolation is required to produce an output from the geographic footprint associated with the reference object matched to, simply because the geographic object is a point and can be returned directly. Unfortunately, although these types of situs reference layers are typically on the higher end of the spatial accuracy scale and should be used wherever and whenever possible, these data

sources often require enormous amounts of field work and/or manual labor to produce and/or maintain and are as such sparsely available[52, 68].Geocoding approaches that return "Rooftop Geocoding" commonly employ situs reference data layers representing pre-computed building centroids that are returned for an address.

The second two classes, linear and areal unit features are far more commonly encountered and both share the same characteristic that each requires some form of interpolation to be applied on a matched reference feature in order for a point output to be returned.

Linear Features—The first of these, linear features, include the most common source of geographic reference data used in geocoding – street centerline databases. These databases are composed of street segments with associated address ranges describing the set of valid house numbers along each side of the street in addition to the other street address components required to describe an address (street name, street suffix, street pre- and post-directional, city, ZIP code, etc.)[42, 60]. First developed at the US Census Bureau in the 1970s, the Dual Independent Map Encoding (DIME) format for representing street segments is still in use today and is the most common format of geographic data employed in geocoding systems[33, 34, 55, 56]. The main reason for this is that the US Census Bureau publishes the TIGER/Lines files, a free and regularly updated street segment database with near-complete coverage across the entirety of the US including the majority of the US territories and other outlying islands[69]. While many research reports have shown that there are definite errors and inconsistencies present in the TIGER/Line files that are known to impact geocoding spatial accuracy and match rate completeness[31, 58, 59, 70], these files nonetheless continue to form the foundation for numerous commercial and research geocoding systems. Enhanced versions of street centerline databases are available from numerous commercial vendors which offer increased spatial accuracy for the geometries of the underlying reference features most commonly in the form of additional interior points along a polyline that describe the curvature of the street, increased completeness in terms of non-spatial attributes such as alternative names for streets, and more precise and up-to-date inclusion of changes, removals, and/or additions to street networks as transportation routes in a region grow, shrink, or are otherwise altered. Unfortunately, the price tag associated with these commercial versions is often out of reach of many geocoding researchers and developers and thus are not included in freely available geocoding systems; hence the continued use of TIGER/Line files.

Regardless of what linear reference data file is included, all systems using these types of reference data layers must perform linear interpolation to produce a final point output drawn from the geography and non-spatial characteristics of the reference street segment matched to[54, 58]. Most commonly in this process, a proportional distance down the length of the reference feature is used based on the ratio of the input address to the number of addresses associated with the street segment. The output point is first computed at the corresponding proportional distance along the geometry of the reference feature and then an offset is applied to move the output orthogonally to one side of the street or the other depending on the parity (even/odd) of the input address and each side of the reference feature[54, 58]. This process is known as address-range interpolation [35] and has been shown on numerous occasions to be rife with assumptions whose resulting error magnitude depends on many aspects such as the rural/urban nature of the region[27–29], the size of the parcel and/or density of addresses[54], and the number, size, and orientation of lots along a street [35]. Because of this, additional data is often brought into the process to enhance the linear interpolation technique; however, the underlying approach of selecting (i.e., interpolating) a location somewhere along or nearby the reference street feature still remains[35].

Areal Unit Features—The final type of reference data layers used in geocoding systems are composed of areal unit features. Areal units can be described simply as a polygon geographic object, or any geographic object that has an interior *area*. Counted among these classes of geographic features would be digital parcel, ZIP code, city, county, and state boundary lines in addition to any other type of geographic area that can be uniquely identified such as a database of building footprints, or more specifically, a list of hospital or prison footprints. As should be obvious from the diverse and varied list of types of geographic objects that fall into this category, areal unit-based geocoding forms the basis of some of the most-(e.g., building footprint-level) and least-accurate (e.g., city-centroid) geocoding approaches available. At one end of the spectrum are extremely detailed and high-resolution (small-scale) geographic features such as digital parcel boundaries and exact building footprints, which, if available and capable of being matched to, would result in highly precise geocode output in many instances for urban addresses [54]. At the other end are low-resolution (large-scale) geographic features such as state and county boundaries that provide little in the way of high precision geocoded results. Unfortunately, as with linear reference feature data layers, the extremely detailed areal unit layers that would provide ideal data sources for use in geocoding systems are often hard to come by. Digital parcel layers are available on a commercial basis either directly from local County Assessor's Offices or through third party vendors who have contracts with these groups; however, reports indicate that these files are only available for 88% of the US at present[71], and prices for obtaining them on a national basis are often prohibitive. Building footprints are even harder to obtain and are only available for a handful of cities nationwide.

Similar to linear features which use a single linear interpolation method regardless of if the underlying reference features are high- or low-resolution, areal unit features also traditionally have a finite set of interpolation techniques that can be applied to produce a point output from a reference feature. These methods all compute and return the "centroid" of a geographic feature, but the means by which the centroid is computed can vary. The first and most common of these is the Bounding Box (BB) method([Figure 2a), also known as the Minimum Bounding Rectangle (MBR) method, in which the minimum and maximum latitude and longitude values associated with any point on the boundary of the object are used to compute a value halfway between the top, bottom, left, and right sides of the polygon. The BB method has the advantage that it is extremely fast to compute and is easily implemented within geocoding systems.

The second method, the Geometric Centroid ([Figure 2b), also known as the Center of Mass (CoM) method differs from the BB method because it uses one or more physics-based calculations to determine where the centroid of the geographic object would be[72, 73]. According to Worboys the "centroid of an areal object is the point at which it would balance if it were cut out of a sheet of material of uniform density" [74, pp 214]. This method of computation is also straightforward for regular polygons, but is more time consuming and computationally complex than a simple BB operation (see [74, pp 214]).

The third method is the weighted centroid method. This method is used to bias the geometric centroid of the object toward an area that contains a higher concentration of a value of interest[75]. In geocoding research, the centroid of an areal unit is often weighted using population density to influence the output location to be closer to where a higher percentage of the population lives. The intuition behind this approach is that if the output is closer to where more of the people in the region live, then there is a higher chance that the person (and/or location) of interest will be closer to the weighted centroid then farther away. This method is by far the most computationally complex and time consuming; however, it is often reported in research as providing more accurate geocode output than either the BB or

the CoM methods, and is thus used frequently across many disciplines that utilize geocoded data[75–78].

2.3 Limitations of Areal Unit Geocoding in Disease Rates

Areal unit-based geocoding has several shortcomings that are both common to all geocoding techniques as well as specific to the characteristics of areal units. Many of these limitations have direct effects on the computation of disease rates when these data are included and become especially pronounced with the output of areal unit geocoding is used in a point-inpolygon intersection approach to associate the geocode output with areal units from other data layers such as county boundaries or census tracts. Our discussion herein does not address implications of boundary error on polygon-in-polygon (polygon intersection) operations because the current research is intended to address one-to-one associations (cancer case to a single administrative unit), and polygon intersection approaches can result in one-to-many (one case to many administrative units). Although we would argue that a many-to-many association could be beneficial in many disease reporting, surveillance, and research scenarios, we know of few where this has been put into practice.

False Clusters—The majority of areal unit interpolation approaches utilized in geocoding techniques are deterministic. That is, they always produce the same result given the same input. The alternative to this would be a probabilistic interpolation approach that could result in two different outputs given the same two inputs. By virtue of being deterministic, areal unit interpolation approaches at the ZIP code or city level often result in false clusters of high disease incidences[27–29]. The reason for this is that all cases with the same city or ZIP code are all placed at the exact same location, stacking up on top of each other and giving the appearance that a cluster exists when in reality one may not if better geocodes could be determined from the street address associated with the case. However in many instances, ZIP code or city level geocodes are the best that can be computed. PO Boxes are one example of this as are Rural Route addresses although the latter are being phased out of practice as E-911 databases are developed throughout the US[28, 43, 79].

Spatial Uncertainty—The fundamental limitation of utilizing areal units in a geocoding process is that one must compute a single geographic location as output from the whole of the area associated with the areal unit. No matter what approach is taken (BB, CoM, or Weighted Centroids), there will always be some uncertainty as to whether the location chosen was the correct one[24]. This problem exists in linear interpolation methods and to some extent in situs reference data layers as well because even for building centroids, one single point must be selected from all possible locations along the rooftop or street centerline. However, because areal unit reference data layers such as boundaries for ZIP codes and cities generally contain large geographic features, the spatial error (or uncertainty) associated with the output location is often larger than that associated with street centerline geocoding[28, 80]. When parcels are used in urban areas, the spatial uncertainty associated with an output geocode can often be lower than a corresponding linear interpolation-based output[54]; however, the same cannot be said as confidently in rural areas[81, 82].The spatial uncertainty associated with each geocode is typically not included in the computation of disease rates, although recent research reports have made calls for it to be included [24]. The proposed quantitative approaches to describing uncertainty as equally distributed across the full area of the areal unit would some what overcome the false clustering described above because although they would still output the same geographic location at the ZIP code level given two addresses in the same ZIP code, each would have a low confidence score of being in the correct location which could be used to weight a level of confidence for disease rates based on them[24, 37].

Inside/Outside Centroids—A further limitation of areal unit geocoding is that the centroid computed may lie outside of the boundary of the polygon if it has an irregular shape [83]. Here, assuming that the reference data layer comprises a surface of polygons without gaps or holes (aside from the boundary extent of the reference layer), the computed centroid of the areal unit may fall within the boundary of another neighboring areal unit. If the geocode resulting from this process were to be used in a point-in-polygon intersection to associate some other value with the point to compute disease rates by an aggregate real unit (such as rates by county or census tract), this artifact of the areal unit geocoding process would result in misclassification of the geocode.

Non-Applicable Centroids—Additional problems can occur when the centroid of an areal unit is placed within the boundary of the areal unit but at a location which would be considered invalid due to non-applicability. This affects processes that are applied to these geocoded data such as point-in-polygon intersections with other data layers. This type of error commonly occurs when the centroid of an areal unit is placed within a water body when the query of interest is to associate census tract values with the output. Here, it may be the case that no census tract covers the water region (in the case of a coast line) and thus it would be impossible to find a census tract to intersect with the geocoded location. Here, the nearest census tract is often chosen, which may or may not be an appropriate decision given the requirements of a particular study. When computing disease rates in particular, this practice can result in the unrealistic appearance of clusters at these locations which were only chosen because they happen to be the closest to the centroid of the areal unit.

Cross-Boundary Centroids—A final problem with the use of areal units for geocoding relates to reference features that cross the boundaries of areal units in other reference data layers such as a ZIP code crossing two or more county boundaries. This also suffers from the same problems listed above in that misclassification of a value is assigned to an output geocode following a point-in-polygon intersection. This problem is particularly troubling when computing county-level disease rates because it may not always be clear which county the centroid of a ZIP code should rightfully be placed in. Should it be the county with a greater proportion of the area of the ZIP code? Should it be in the county with a greater amount of the population? An example of this is shown in [Figure 3 for the ZIP code 95469 which crosses the administrative boundaries of the two counties in California –Mendocino and Lake Counties.

3. Experimental Design

Stated again, the main goal of this article is to evaluate the magnitude and effects of utilizing areal unit geocoding in the derivation of disease rates in order to understand how sensitive these rates are to the methods used by the geocoding process. In our experiments, we chose the ZIP code level areal unit as the target geographic output level for two reasons. First, there is a plethora of existing literature on the applicability and use of ZIP code data in the geocoding process as applied to health-related research which provides a solid context within which we can place our work (see [80] for a detailed review). Second, due to privacy and confidentially concerns of transmitting health-related data with personally-identifiable attributes, ZIP code level data are often computed, transferred, and utilized by researchers and policy-makers as a safe alternative to address-level data. This second reason should by no means be seen as an endorsement of this practice – there is still little if any consensus among researchers in the geographic confidentiality and privacy domain that this is an appropriate level of obfuscation[1, 84–86]. Nonetheless, the ZIP code level of aggregation is commonly used in health-related research so the analysis of this level of geographic data presented herein should serve to offer guidance to those who partake in this practice[87–89].

One important aspect to note is that the experiments presented herein compute crude, five year aggregate incident rates. These are different than standardized rates typically seen in annual reporting of cancer incidence, so the results described here will differ from other published reports such as those from the California Cancer Registry.

3.1 Data Sources

Case Data—ZIP code level data on all cancer cases diagnosed in the State of California for the five year period January 1, 2006 through December 31, 2010 were obtained from the CCR (n=730,124). These data included cases originally geocoded at all "levels" (street match, ZIP centroid, etc.). These data also included records that originally included standard street level addresses such as the hypothetical "123 Main Street" in addition to more challenging records such as those that originally listed PO Box addresses as well as other errors or omissions such as the street address being listed as "UNKNOWN". Only the ZIP code associated with these records were provided by CCR; street addresses, PO Boxes and cities were not included. Each unique ZIP code was listed once along with an attribute that described the count of cases within that ZIP code.

Geographic Data—The 2010 US Census Bureau Zip Code Tabulation Area (ZCTA) boundary file for California was used to compute ZIP code centroids as described below. The 2010 US Census Bureau County boundary file for California was used to assign each geocode output to a county as described below. Both files are available as part of the 2010 US Census Bureau TIGER/Lines set of files [69] and are commonly used throughout geocoding research, development, and practice as well as spatially-based health related research due to their free price tag, their nationwide coverage, and their routine updating.

It should be noted here that there is a difference between USPS ZIP codes and the US Census Bureau's ZCTAs. The first represents a postal delivery route without a true geographic boundary, whereas the second is an approximation made by the US Census Bureau of the former. Numerous research reports discuss the pros and cons of using both of these within geocoding and health science research; the interested reader should see [28, 80, 90] and the references within for a full discussion of the strengths, weaknesses, and limitations of using either in health-based analyses. For the purposes of this study – to examine the impact of varying interpolation methods on disease rates – the use of ZCTAs is sufficient to illustrate our point. Further, these data are readily available at no cost, so other researchers may easily replicate, validate, or employ our methods to achieve similar results. From this point forward, the words ZIP code and ZCTA will be used interchangeably.

3.2 Methods

Interpolation Methods—Four methods were used to compute ZIP code centroids for each of the case ZIP codes and the output from each method was assigned to the county within which the output fell. Three of the four rely on the Esri ArcGIS platform [91], whereas the fourth utilizes the Microsoft .Net programming framework[92]. The first system, ArcGIS, is an industry standard GIS platform with a reported million plus users worldwide that is in common use in spatially-based health investigations. ArcGIS has over 700 built in spatial analysis, modeling, and visualization tools that can be applied to any number of research questions; this project makes use of two of them (described below).

The second system, the .Net framework, is one of the underlying components of the Microsoft Windows operating system and is by many accounts the platform of choice for developing custom built software for Windows machines. As part of the libraries accessible to a developer using .Net, Microsoft provides a suite of geographic data types upon which geographic operators can be applied. These geographic operators include standards-

compliant instantiations of many of the Open Geospatial Consortium (OGC) set of geographic operators, including ST Centroid() which computes the centroid of an areal unit.

1. *ArcGIS "Polygon to point", within "inside" option selected*. The approach computes the geometric centroid of the areal unit and enforces that the output point must be contained within the input polygon.
2. *ArcGIS "Polygon to point", within "inside" option not selected*. This approach computes the geometric centroid of the areal unit, but does not enforce that the output point must be contained within the input polygon.
3. *ArcGIS "Feature envelope" and ArcGIS "Polygon to point"*. This approach first computes the bounding box of the areal unit and then computes the geometric centroid using the same ArcGIS operator as the first two approaches. The "inside" option was selected but would not make a difference in this case because a bounding box is a regular polygon and therefore the centroid will always be within the bounding box by definition.
4. .Net "ST Centroid()". This approach creates an in-memory geospatial object from the boundary of the areal unit and calls the .Net built-in ST Centroid() method which computes the geometric centroid of the areal unit. This method does not enforce that the output point will be within the input polygon.

Although Census data describing where populations reside within regions are available, the use of weighted centroid techniques is not common in geocoding systems due to the computational and time complexities described earlier. Hence, population weighted-centroid methods were not included in the present research because the choice was made to evaluate the most commonly used methods applicable to a wide breadth of geocoding systems.

County Intersection Methods—County codes were associated with the output geocode for each of the above four methods within ArcGIS using the "Spatial Join" operator. This approach uses a point-in-polygon intersection method to associate the county code from the county areal unit within which a geocode output falls.

4. Results

4.1 Geocoding Match Rates at the ZIP Code-Level

A first step in the ZIP code-level geocoding approach described here is to analyze the characteristics of the input data. To do so, we first note that the input case data for the five year period used here contained 2,515 unique ZIP codes. Descriptive statistics of the characteristics of incidences per ZIP code over the five year period are shown in[Table 1. The cumulative distribution of record incidences per ZIP code over the five year period shown in [Figure 4 reveals that 25% of the ZIP codes have ten or fewer cases reported and 50% of the data have 60 or fewer.

A next informative step is to investigate the match rates achievable using the US Census Bureau ZCTA files as a reference data source. 1,667 (66%) of the 2,515 input ZIP codes successful matched to a ZIP code in the ZCTA reference data layer. This means that 34% of the ZIP codes associated with ZIP codes listed on case records were not locatable in the ZCTA reference data file. The case counts associated with these ZIP codes account for 3.3% (n=24,136) of the total records for the five-year period. The cumulative distribution frequencies shown in [Figure 5 for the counts of cases ZIP codes that were unmatchable reveals that 50% all unmatchable ZIP codes have only one cases located within them. The fictitious ZIP code 99999 used by CCR to indicate that the ZIP code for case records is not known accounted for 2,065 of these records that could not be matched.

4.2 ZIP Code Centroid County Value Changes

2% (41) of the unique ZIP codes associated with the case data for the five-year period were affected by the difference in areal unit centroid derivation method. The affected ZIP codes along with the counties between which the computed centroids changed and the number of cases impacted over the five-year period are listed in [Table 2. The geographic distribution of these ZIP codes is displayed in [Figure 6 which also indicates the number of records affected.

A specific example of a ZIP code 93514 which resulted in the assignment of three different counties is shown in [Figure 7. Here, the .Net centroid method results in Mono County, the ArcGIS MBR method results in Fresno County, and both the ArcGIS Within and ArcGIS Not-Within result in Inyo County.

4.3 County-Level Cumulative Five Year Incidence Rate Variation

44 out of the 58 counties in California (76%) have cumulative five year incidence rates that changed based on the method used to compute the areal unit centroid. The 44 counties for which rates changed are listed in [Table 1. These results show that Mono and Inyo counties have the highest levels of changes between each of the four methods at ~3% and 2% changes, respectively.

The 13 counties for which no rate change were observed are listed in [Table 4. These include some of the largest (Los Angeles, San Diego) and smallest (Alpine, Del Norte) counties in California. The levels of change of incidence rates over the five year period are shown in [Figure 8.

5. Discussion & Conclusions

This article has presented an evaluation of how county-based cumulative five year incidence rates of cancer are affected by assignment of ZIP code-level incidence counts. Using five years of data on all cancer cases diagnosed in the State of California from the California Cancer Registry, crude cancer incidence rates were derived using four different methods of area unit interpolation at the ZIP code-level. From these experiments, it is interesting to observe that 76% of the counties in California have incidence rates that change depending on the method of areal unit centroid computation; only 13 out of the 58 counties in California have rates that remain stable regardless of the method used.

Although the actual amount by which the incident rates change across all counties was small, ranging from 0 to only 2.8%, these numbers represent dramatic increases in overall rates when considering the difference between the set of cancer rates produced using each method. For example, Mono County five year incident rates increase from 1.3 to 4.1% of the population, close to a 400% increase in disease rates. Yolo County saw similarly high five year rate increases, from 0.95 to 2.95% incidence rate. However, across all of the counties for which variations in rates were observed, the average rate change was only 0.28%. This indicates that although five year incidence rates in specific counties are greatly affected when the number of incidences (numerator) and size of the underlying population (denominator) are both small, most counties in California have rates that are relatively stable across available interpolation methods.

The analysis performed here revealed no clear geographic patterns of where these rate changes occur or how large the magnitude of those changes could be. The 41 ZIP codes that were affected were distributed across the entire state and the number of records affected by each ZIP code did not exhibit any clear geographic pattern.One thing which is clear from this analysis is that the shape of the area to be geocoded (e.g. ZIP code) and its proximity to

the target geography border (e.g. county) will be important determinants of sensitivity to centroid calculation and assignment to target geography. Further research should be performed to shed additional light on why and how these ZIP codes behave in this manner so that researchers can predict areas where these types of problems may occur and prompt closer inspection of the data to ensure correctness.

The experiments and analysis presented here have two important limitations. The first limitation is that we used all cancers instead of site-specific cancers. The reason for this is that the goal of the current research was to demonstrate the magnitude of problems stemming from different interpolation techniques, and all cancers suffice for this aim. Future work should examine how the use site-specific cancers impacts misclassification and rates of disease. We anticipate that when a site-specific cancer is uniformly distributed across California, the amount of misclassification as a proportion of all site-specific records would decrease, simply because the number of cases would be lower. However, if a site-specific cancer exhibits a clustering behavior at specific locations, misclassification could account for a large proportion of records when the location of the cluster happens to be in one of the areas subject to the boundary errors found here. Our recommendation in the case of site specific clusters would be to investigate the sensitivity of the cluster location to these boundary errors.

The second limitation of the current work is that we compute crude five year estimates of disease rates. Epidemiological analysis is typically conducted on rates that have been filtered and/or smoothed in some fashion to account for the small number problem[93]. The results we present here represent a worst-case scenario for misclassification errors due to varying interpolation techniques. It is conceivable that the differences between interpolation methods could be greatly reduced after noise filtering. Further research should be performed to assess the scope, magnitude, and implications of using smoothing approaches in conjunction with each of the interpolation methods.

Despite these limitations, the findings of our experiments confirm historical, recent, and recurring conclusions of other researchers who have found that the specific inner-workings of any particular geocoding system can have large impacts on the results of research projects based on geocoded data[18, 19, 21–23, 26, 30, 44, 57]. In this particular case, we investigated crude five year cancer incidence rates as an example of disease rate computations which is used to inform policy decisions at the local, regional, and national levels. Based on our findings, it is conceivable that two different researchers or agencies applying two different geocoding techniques could determine cancer incidence rates nearly four times the magnitude of each other. Our results show that depending on the geocoding method used, it could appear that either a 400% reduction or increase of cancer incidence is observed. This could have major impacts on the allocation of funding, outreach, or prevention activities as well as analyses that asses the effectiveness of these programs.

The unfortunate reality is that every geocoding process used by a researcher, agency, or policy maker is likely to utilize different data sources, interpolation methods, or other internal algorithms that can affect the quality of the results. As we have shown here, even when a single reference data layers (US Census Bureau ZCTA boundaries) along with allegedly comparable interpolation techniques – ArcGIS non-within and .Net ST Centroid should both produce the same output because they claim to perform the same task – large differences in output can be and are often observed. In many instances it is difficult to say which method should be used or should produce the most accurate results, so in practice few users care or know which approach is being utilized. However, the findings presented in this article should serve to motivate geocoding users to peek under the hood of their geocoding systems to see what is truly going on if they are to understand the opportunities and

limitations present in their data. These same users need to make loud and frequent calls to geocoding system developers to expose metadata about the inner-workings of geocoding systems as well as provide flexibility in what geocoding algorithms are utilized under what circumstances.

Acknowledgments

This work was supported in part by award number 5P30ES007048 from the National Institute of Environmental Health Sciences, contract number N01-PC-35139 from the National Cancer Institute, and by cooperative agreement number 1H13EH000793-01 from the Centers for Disease Control and Prevention. The contents of this work are solely the responsibility of the authors and do not necessarily reflect the official views of any of these sponsors.

References

1. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. Geocoding in cancer research: A review. American Journal of Preventive Medicine. 2006; 30(2):S16–S24. [PubMed: 16458786]
2. Nuckols JR, Ward MH, Jarup L. Jarup: Using geographic information systems for exposure assessment in environmental epidemiology studies. Environmental Health Perspectives. 2004; 112(9):1007–1015. [PubMed: 15198921]
3. Henry K, Boscoe F, Johnson C, Goldberg D, Sherman R, Cockburn M. Breast Cancer Stage at Diagnosis: Is Travel Time Important? Journal of Community Health. 2011 in press.
4. Boscoe F, Johnson C, Henry K, Goldberg D, Shahabi K, Elkin E, Ballas L, Cockburn M. Geographic Proximity to Treatment for Early Stage Breast Cancer and Likelihood of Mastectomy. The Breast. 2011
5. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? American Journal of Epidemiology. 2002; 156(5): 471–482. [PubMed: 12196317]
6. Gumpertz ML, Pickle LW, Miller BA, Bell BS. Geographic patterns of advanced breast cancer in Los Angeles: Associations with biological and sociodemographic factors (United States). Cancer Causes and Control. 2006; 17(3):325–339. [PubMed: 16489540]
7. Abe, T.; Stinchcomb, DG. Geocoding Practices in Cancer Registries. In: Rushton, G.; Armstrong, MP.; Gittler, J.; Greene, BR.; Pavlik, CE.; West, MM.; Zimmerman, DL., editors. Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice. Boca Raton, FL: CRC Press; 2008. p. 195-223.
8. Bell BS, Hoskins RE, Pickle LW, Wartenberg D. Current practices in spatial analysis of cancer data: Mapping health statistics to inform policymakers and the public. International Journal of Health Geographics. 2006; 5(49)
9. Boscoe FP, Ward MH, Reynolds P. Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. International Journal of Health Geographics. 2004; 3(28)
10. Boulos MNK. Towards evidence-based, GIS-driven national spatial health information infrastructure and surveillance services in the United Kingdom. International Journal of Health Geographics. 2004; 3(1)
11. Brody JG, Aschengrau A, McKelvey W, Rudel R, Swartz C, Kennedy T. Breast cancer risk and historical exposure to pesticides from wide-area applications assessed with GIS. Environmental Health Perspectives. 2004; 112:889–897. [PubMed: 15175178]
12. McConnell R, Berhane K, Yao L, Jerrett M, Lurmann F, Gilliland F, Künzli N, Gauderman J, Avol E, Thomas D, et al. Traffic, susceptibility, and childhood asthma. Environmental Health Perspectives. 2006; 114(5):766–772. [PubMed: 16675435]
13. Zandbergen PA, Green JW. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. Environmental Health Perspectives. 2007; 115(9):1363–1370. [PubMed: 17805429]

14. Reynolds P, Hurley SE, Goldberg DE, Yerabati S, Gunier RB, Hertz A, Anton-Culver H, Bernstein L, Deapen D, Horn-Ross PL, et al. Residential proximity to agricultural pesticide use and incidence of breast cancer in the California Teachers Study cohort. Environmental Research. 2004; 96:206–218. [PubMed: 15325881]
15. Rull RP, Gunier R, Behren JV, Hertz A, Crouse V, Buffler PA, Reynolds P. Residential proximity to agricultural pesticide applications and childhood acute lymphoblastic leukemia. Environmental Research. 2009; 109(7):891–899. [PubMed: 19700145]
16. van Wiechen C, Houthuijs D, Ameling C, Lebret E. Exposure Assessment of Environmental Noise for Use in Small Area Health Studies. Epidemiology. 2004; 15(4):S205.
17. Ferguson EC, Maheswaran R, Daly M. Road-traffic pollution and asthma - using modelled exposure assessment for routine public health surveillance. International Journal of Health Geographics. 2004; 3(1)
18. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. International Journal of Health Geographics. 2008; 7(13)
19. Zhan FB, Brender JD, De Lima I, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Annals of Epidemiology. 2006; 16(11):842–849. [PubMed: 17027286]
20. Zandbergen PA. Positional accuracy of spatial data: Non-normal distributions and a critique of the National Standard for Data Accuracy. Transactions in GIS. 2008; 12(1):103–130.
21. Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, Heiss G. Accuracy of commercial geocoding: Assessment and implications. Epidemiologic Perspectives & Innovations. 2006; 3(8)
22. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, Mix W, Colt JS, Hartge P. Positional accuracy of two methods of geocoding. Epidemiology. 2005; 16(4):542–547. [PubMed: 15951673]
23. Schootman M, Sterling DA, Struthers J, Yan Y, Laboube T, Emo B, Higgs G. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. Annals of Epidemiology. 2007; 17(6):379–387.
24. Goldberg, D.; Cockburn, M. Accuracy 2010. Leicester, UK: 2010. Toward quantitative geocode accuracy metrics; p. 329-332.
25. Fulcomer, MC.; Bastardi, MM.; Raza, H.; Duffy, M.; Dufficy, E.; Sass, MM. Assessing the accuracy of geocoding using address data from birth certificates: New Jersey, 1989 to 1996. In: Williams, RC.; Howie, MM.; Lee, CV.; Henriques, WD., editors. Proceedings of the 1998 Geographic Information Systems in Public Health Conference. San Diego, CA: 1998. p. 547-560.
26. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003; 14(4):408–411. [PubMed: 12843763]
27. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. Geographic bias related to geocoding in epidemiologic studies. International Journal of Health Geographics. 2005; 4(29)
28. Krieger N, Waterman PD, Chen JT, Soobader MJ, Subramanian SV, Carson R. ZIP code caveat: Bias due to spatiotemporal mismatches between ZIP codes and US census-defined areas: The public health disparities geocoding project. American Journal of Public Health. 2002; 92(7):1100–1102. [PubMed: 12084688]
29. Bichler G, Balchak S. Address matching bias: Ignorance is not bliss. PIJPSM. 2007; 30(1):32–60.
30. Zandbergen PA. Geocoding accuracy considerations in determining residency restrictions for sex offenders. Criminal Justice Policy Review. 2009; 20(1):62–90.
31. Wu J, Funk TH, Lurmann FW, Winer AM. Improving spatial accuracy of roadway networks and geocoded addresses. Transactions in GIS. 2005; 9(4):585–601.
32. Gatrell AC. On the spatial representation and accuracy of address-based data in the United Kingdom. International Journal of Geographical Information Science. 1989; 3(4):335–348.
33. Tobler, W. Proceedings of the National Geocoding Conference. Washington DC: U.S. Department of Transportation; 1972. Geocoding theory.

34. O'Reagan, RT.; Saalfeld, A. Statistical Research Report. Washington, DC: United States Bureau of Census; 1987. Geocoding theory and practice at the Bureau of the Census.
35. Bakshi, R.; Knoblock, CA.; Thakkar, S. Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems. Washington, DC: ACM Press; 2004. Exploiting online sources to accurately geocode addresses; p. 194-203.
36. Goldberg D, Wilson J, Knoblock C. Using Spatially Varying Block Metrics to Improve the Geocoding Process. Computers, Environment and Urban Systems. 2010 Under review.
37. Goldberg D, Cockburn M. Improving Geocode Accuracy with Candidate Selection Criteria. Transactions in GIS. 2010; 14(s1):149–176.
38. Christen, P.; Churches, T.; Willmore, A. Proceedings of the Australasian Data Mining Conference. Cairns, AU: 2004. A probabilistic geocoding system based on a national address file.
39. Christen, P.; Churches, T. Proceedings of the Australian Research Council Health Data Mining Workshop. Canberra, AU: 2005. A probabilistic deduplication, record linkage and geocoding system.
40. Jaro M. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association. 1989; 89:414–420.
41. Jaro, M. Statistical Research Division Report Series. Washington, DC: United States Census Bureau; 1984. Record linkage research and the calibration of record linkage algorithms.
42. Goldberg DW, Wilson JP, Knoblock CA. From text to geographic coordinates: The current state of geocoding. Urisa Journal. 2007; 19(1):33–47.
43. McElroy JA, Remington PL, Trentham-Dietz A, Roberts SA, Newcomber PA. Geocoding addresses from a large population based study: Lessons learned. Epidemiology. 2003; 14(4):399–407. [PubMed: 12843762]
44. Kravets N, Hadden WC. The accuracy of address coding and the effects of coding errors. Health & Place. 2007; 13(1)
45. Smith, D.; Crane, G. Research and Advanced Technology for Digital Libraries: Fifth European Conference (ECDL 2001). Darmstadt, Germany: 2001. Disambiguating geographic names in a historical digital library; p. 127-136.
46. Rull RP, Ritz B. Historical pesticide exposure in California using pesticide use reports and land-use surveys: An assessment of misclassification error and bias. Environmental Health Perspectives. 2003; 111(13):1582–1589. [PubMed: 14527836]
47. Rose KM, Wood JL, Knowles S, Pollitt RA, Whitsel EA, Diez Roux AV, Yoon D, Heiss G. Historical measures of social context in life course studies: retrospective linkage of addresses to decennial censuses. International Journal of Health Geographics. 2004; 3(27)
48. Kennedy, TC.; Brody, JG.; Gardner, JN. Proceedings of the 2003 ESRI Health GIS Conference. Arlington, Virginia: 2003. Modeling historical environmental exposures using GIS: implications for disease surveillance.
49. Brody JG, Vorhees DJ, Melly SJ, Swedis SR, Drivas PJ, Rudel RA. Using GIS and historical records to reconstruct residential exposure to large-scale pesticide application. Journal of Exposure Analysis and Environmental Epidemiology. 2002; 12(1):64–80. [PubMed: 11859434]
50. Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. Computers, Environment and Urban Systems. 2008; 32:214–232.
51. Stage, D.; von Meyer, N. Federal Geographic Data Committee Subcommittee on Cadastral Data. 2005. An assessment of parcel data in the United States 2005 Survey results.
52. Zimmerman DL, Fang X, Mazumdar S, Rushton G. Modeling the probability distribution of positional errors incurred by residential address geocoding. International Journal of Health Geographics. 2007; 6(1)
53. Vieira V, Fraser A, Webster T, Howard G, Bartell S. Accuracy of Automated and E911 Geocoding Methods for Rural Addresses. Epidemiology. 2008; 19(6):S352.
54. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics. 2003; 2(10)
55. Dueker KJ. Urban geocoding. Annals of the Association of American Geographers. 1974; 64(2): 318–325.

56. Werner PA. National geocoding. Annals of the Association of American Geographers. 1974; 64(2):310–317.
57. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. American Journal of Public Health. 2001; 91(7):1114–1116. [PubMed: 11441740]
58. Ratcliffe JH. On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. International Journal of Geographical Information Science. 2001; 15(5):473–485.
59. Lee B. Spatial Pattern of Uncertainties: An Accuracy Assessment of the TIGER Files. Journal of Geography and Geology. 2009; 1(2):2–12.
60. Boscoe, FP. Rushton, G.; Armstrong, MP.; Gittler, J.; Greene, BR.; Pavlik, CE.; West, MM.; Zimmerman, DL. Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice. Boca Raton, FL: CRC Press; 2008. The science and art of geocoding; p. 95-109.edn.
61. Goldberg, DW. A Geocoding Best Practices Guide. Springfield, IL: North American Association of Central Cancer Registries; 2008.
62. Zandbergen PA. Geocoding quality and implications for spatial analysis. Geography Compass. 2009; 3(2):647–680.
63. Welcome to the California Cancer Regsitry. [http://www.ccrcal.org/]
64. Goldberg DW, Wilson JP, Knoblock CA, Ritz B, Cockburn MG. An effective and efficient approach for manually improving geocoded data. International Journal of Health Geographics. 2008; 7(60)
65. Goldberg, DW. The USC WebGIS Geocoding Platform. Los Angeles CA: University of Southern California GIS Research Laboratory; 2011.
66. Vine MF, Degnan D, C H. Geographic information systems: Their use in environmental epidemiologic research. Journal of Environmental Health. 1998; 61:7–16.
67. North American Association of Central Cancer Registries. [http://www.naaccr.org]
68. Rushton G, Peleg I, Banerjee A, Smith G, West MM. Analyzing geographic patterns of disease incidence: Rates of late-stage colorectal cancer in Iowa. Journal of Medical Systems. 2004; 28(3): 223–236. [PubMed: 15446614]
69. U.S. Census Bureau TIGER/Line. [http://www.census.gov/geo/www/tiger]
70. Block, R., editor. Geocoding of crime incidents using the 1990 TIGER file: The Chicago example. Washington, DC: Police Executive Research Forum; 1995.
71. National Parcelmap Data Portal Coverage. [http://www.boundarysolutions.com/BSI/coverage.php]
72. Lafond D, Duarte M, Prince F. Comparison of three methods to estimate the center of mass during balance assessment. Journal of biomechanics. 2004; 37(9):1421–1426. [PubMed: 15275850]
73. Thomas S, Fusco T, Tokovinin A, Nicolle M, Michau V, Rousset G. Comparison of centroid computation algorithms in a Shack--Hartmann sensor. Monthly Notices of the Royal Astronomical Society. 2006; 371(1):323–336.
74. Worboys, M. GIS: A Computing Perspective. Bristol, PA: Taylor & Francis; 1997.
75. Martin DJ. Mapping population data from zone centroid locations. Transactions of the Institute of British Geographers. 1989; 14(1):90–97. [PubMed: 12281689]
76. Bracken I, Martin D. The generation of spatial population distributions from census centroid data. Environment and Planning A. 1989; 21(4):537–543. [PubMed: 12315638]
77. Sharkey JR, Horel S. Neighborhood socioeconomic deprivation and minority composition are associated with better potential spatial access to the ground-truthed food environment in a large rural area. The Journal of nutrition. 2008; 138(3):620–627. [PubMed: 18287376]
78. Pearce J, Witten K, Bartie P. Neighbourhoods and health: a GIS approach to measuring community resource accessibility. Journal of Epidemiology and Community Health. 2006; 60(5):389–395. [PubMed: 16614327]
79. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post office box addresses: A challenge for geographic information system-based studies. Epidemiology. 2003; 14:386–391. [PubMed: 12843760]

80. Beyer, KMM.; Schultz, AF.; Rushton, G. Using ZIP codes as geocodes in cancer research. In: Rushton, G.; Armstrong, MP.; Gittler, J.; Greene, BR.; Pavlik, CE.; West, MM.; Zimmerman, DL., editors. Geocoding Health Data - The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice. Boca Raton, FL: CRC Press; 2008. p. 37-68.
81. Durr PA, Froggatt AEA. How best to georeference farms? A case study from Cornwall, England. Preventive Veterinary Medicine. 2002; 56:51–62. [PubMed: 12419599]
82. Stevenson MA, Wilesmith J, Ryan J, Morris R, Lawson A, Pfeiffer D, Lin D. Descriptive spatial analysis of the epidemic of bovine spongiform encephalopathy in Great Britain to June 1997. The Veterinary Record. 2000; 147(14):379–384. [PubMed: 11072999]
83. van Roessel JW. An algorithm for locating candidate labeling boxes within a polygon. Cartography and Geographic Information Science. 1989; 16(3):201–209.
84. Curtis AJ, Mills JW, Agustin L, Cockburn MG. Confidentiality risks in fine scale aggregations of health data. Computers, Environment and Urban Systems. 2010 Corrected proof published online September 6, 2010.
85. Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. International Journal of Health Geographics. 2006; 5(44)
86. Curtis AJ, Mills JW, Leitner M. Keeping an eye on privacy issues with geospatial data. Nature. 2006; 441:150. [PubMed: 16688146]
87. Pappas G, Hadden WC, Kozak LJ, Fisher GF. Potentially avoidable hospitalizations: inequalities in rates between US socioeconomic groups. American Journal of Public Health. 1997; 87(5):811–886. [PubMed: 9184511]
88. Thomas AJ, Eberly LE, Davey Smith G, Neaton JD. ZIP-code-based versus tract-based income measures as long-term risk-adjusted mortality predictors. American journal of epidemiology. 2006; 164(6):586–590. [PubMed: 16893922]
89. Drewnowski A, D Rehm C, Solet D. Disparities in obesity rates: analysis by ZIP code area. Social Science & Medicine. 2007; 65(12):2458–2463. [PubMed: 17761378]
90. Grubesic T, Matisziw T. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. International journal of health geographics. 2006; 5(1)
91. Environmental Systems Research Institute. ArcGIS: A Complete Integrated System. Vol. vol. 2010. Redlands, CA: Environmental Systems Research Institute; 2011.
92. OGC Methods on Geography Instances. [http://msdn.microsoft.com/en-us/library/bb933917.aspx]
93. MacEachren AM, Brewer CA, Pickle LW. Visualizing georeferenced data: Representing reliability of health statistics. Environment and Planning A. 1998; 30:1547–1561.

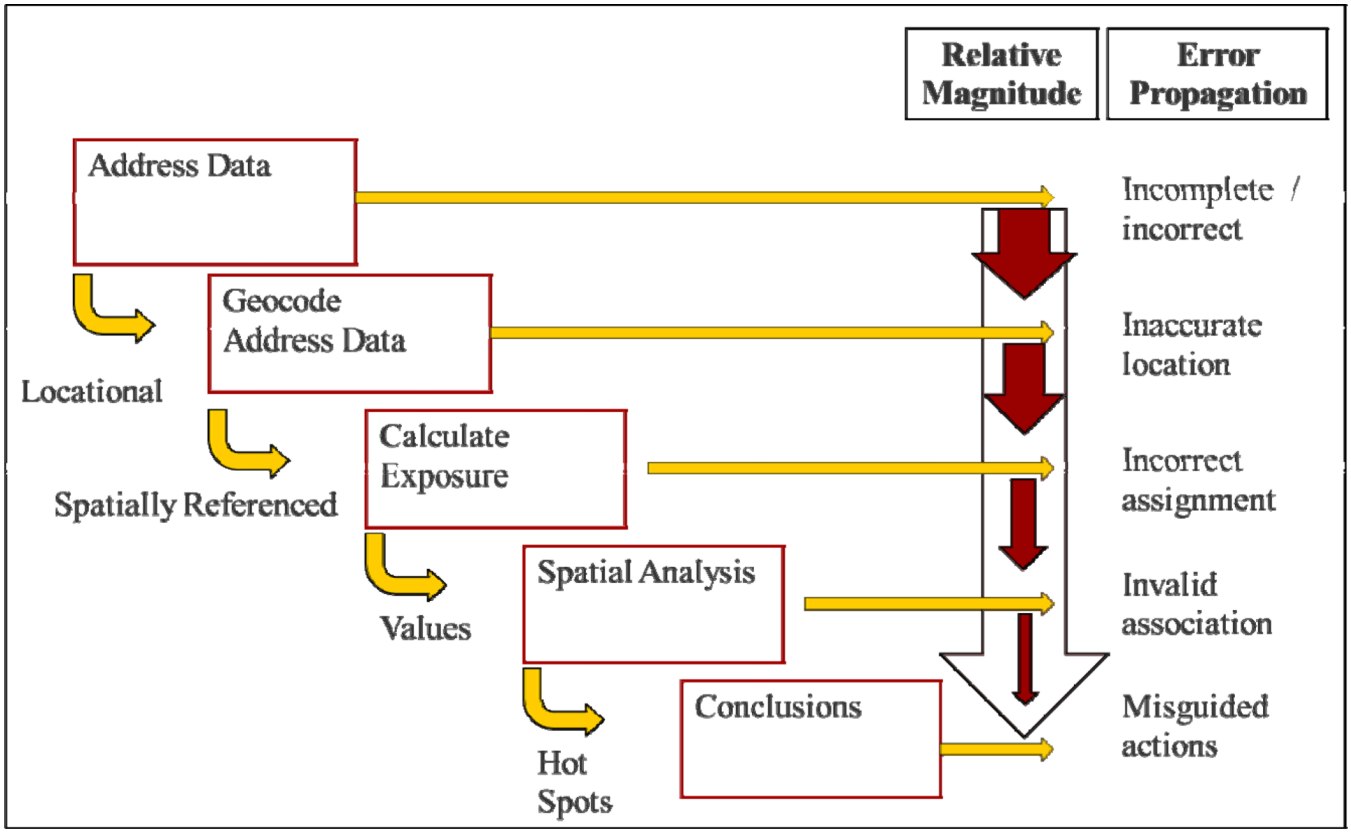

Figure 1.
Geocoded data pipeline and error propagation in environmental exposure studies

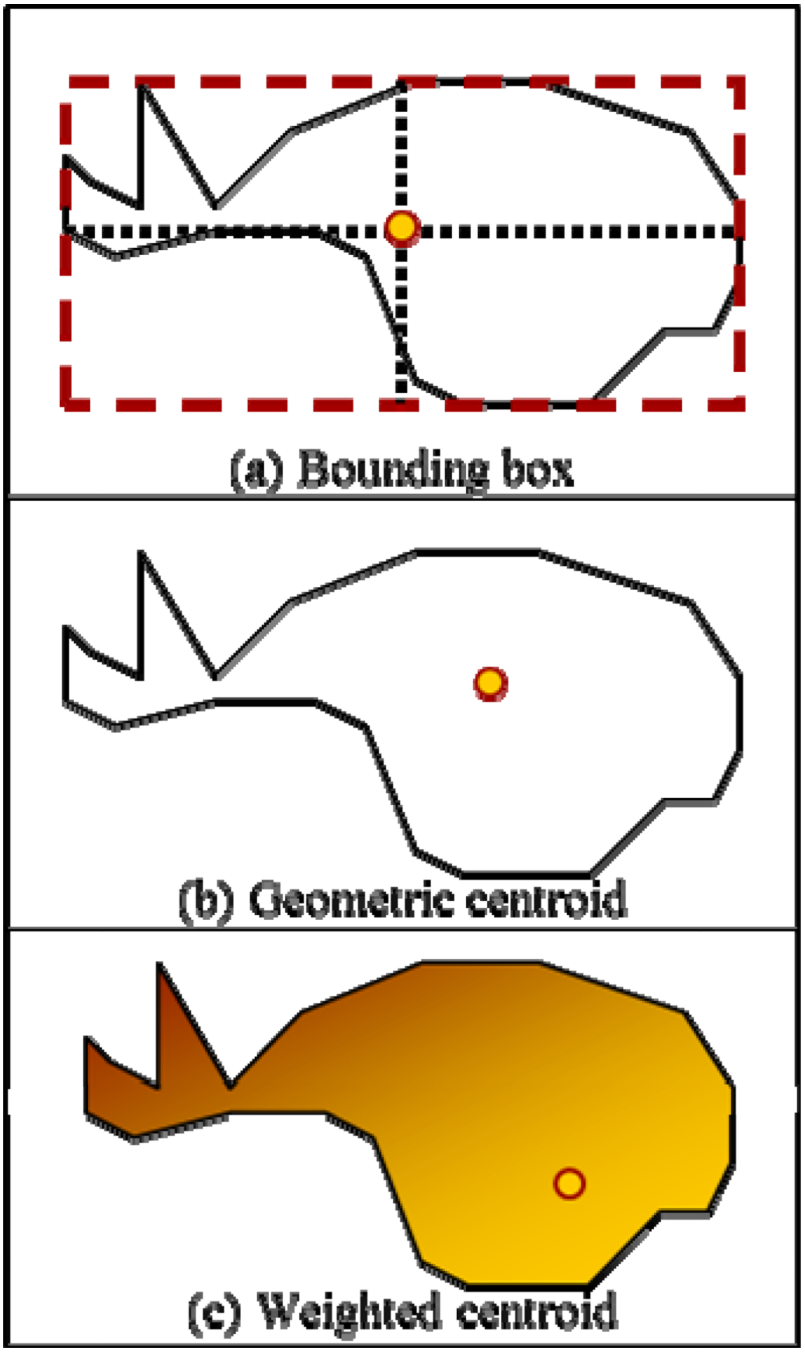

Figure 2.
Three different methods of computing a centroid for a hypothetical areal unit: a) bounding box; b) geometric centroid; and c) weighted centroid

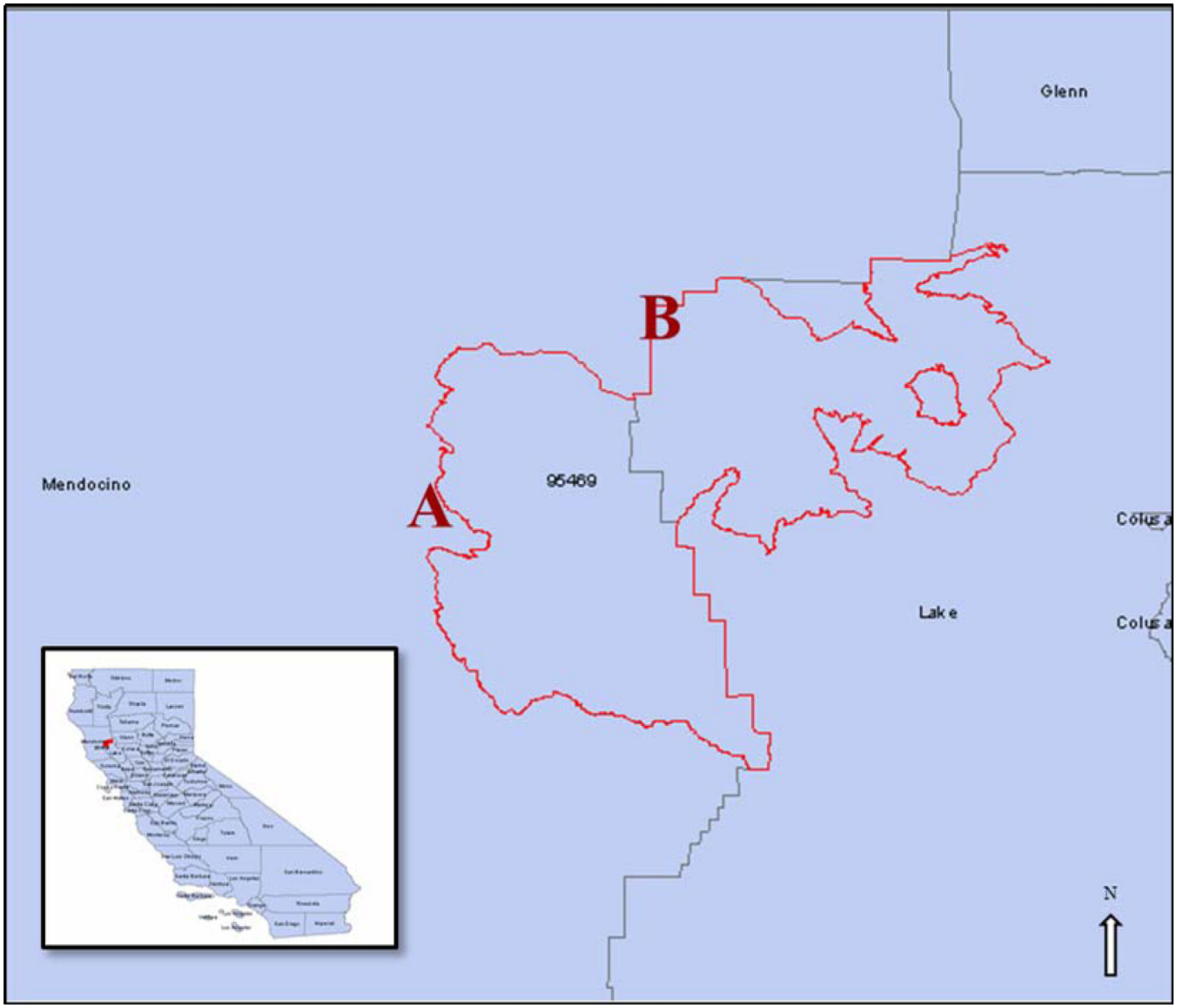

Figure 3.
Example of ZIP code 95469 which crosses the county boundaries of Mendocino and Lake Counties

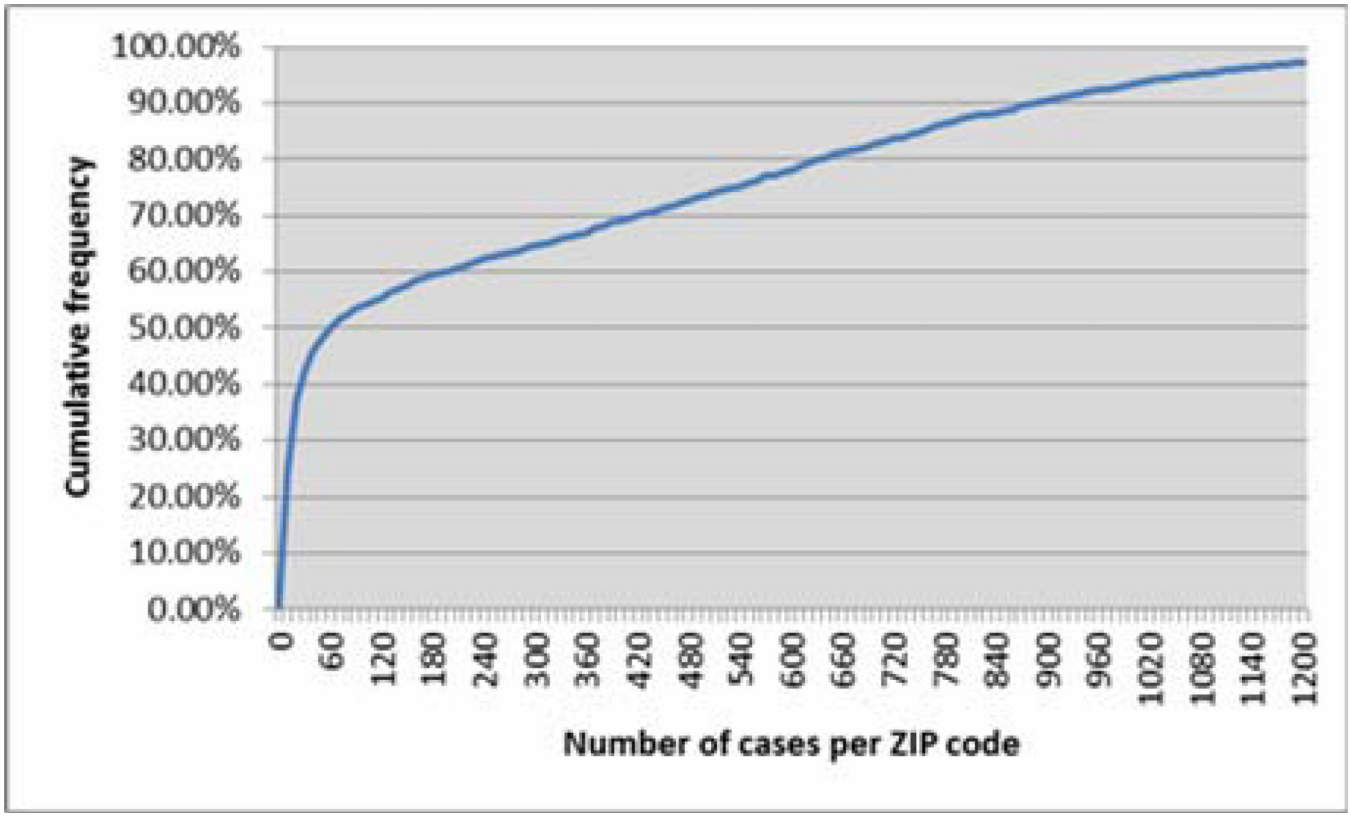

Figure 4.
Cumulative frequency distribution of cases per ZIP code; 2006–2010

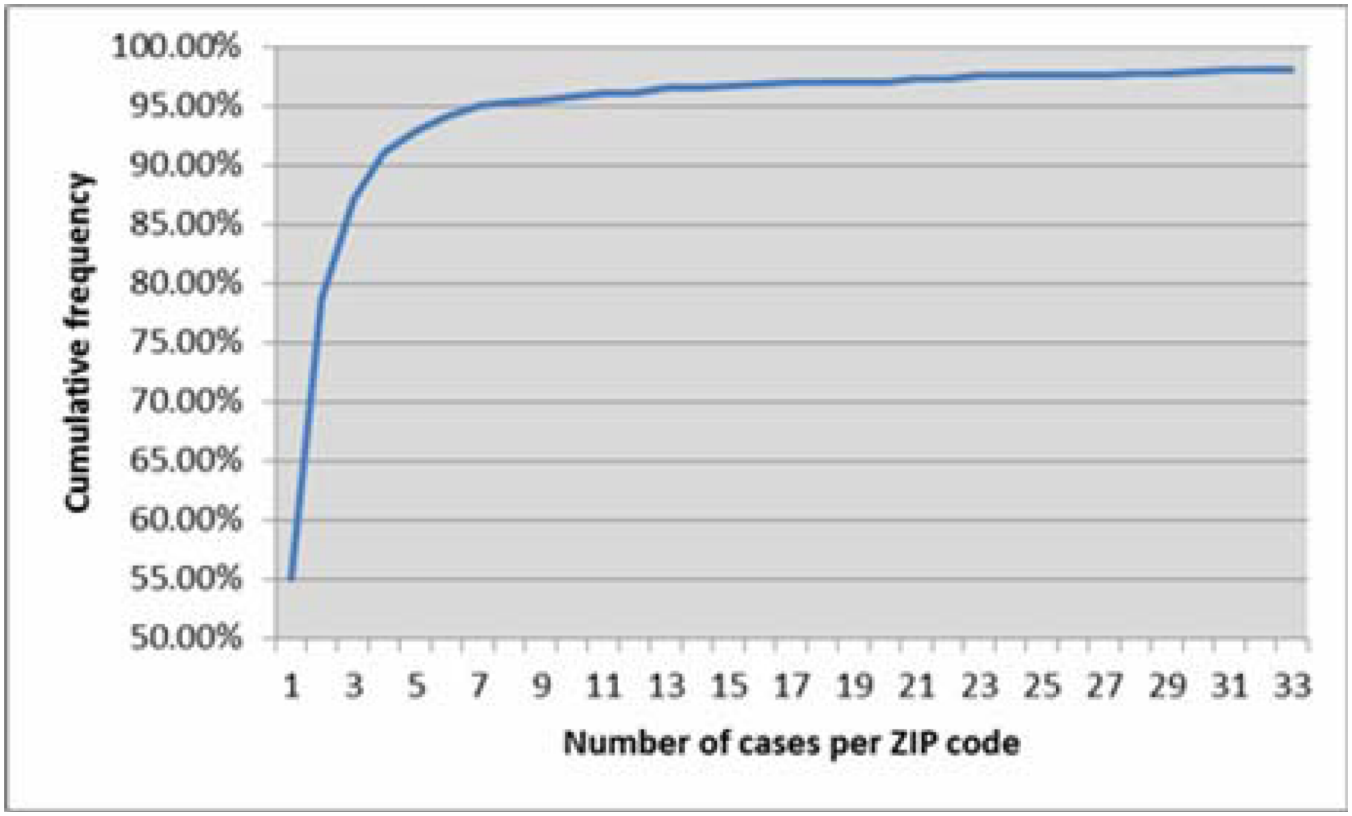

Figure 5.
Cumulative frequency distribution of cases per ZIP code for unmatchable ZIP codes; 2006–2010

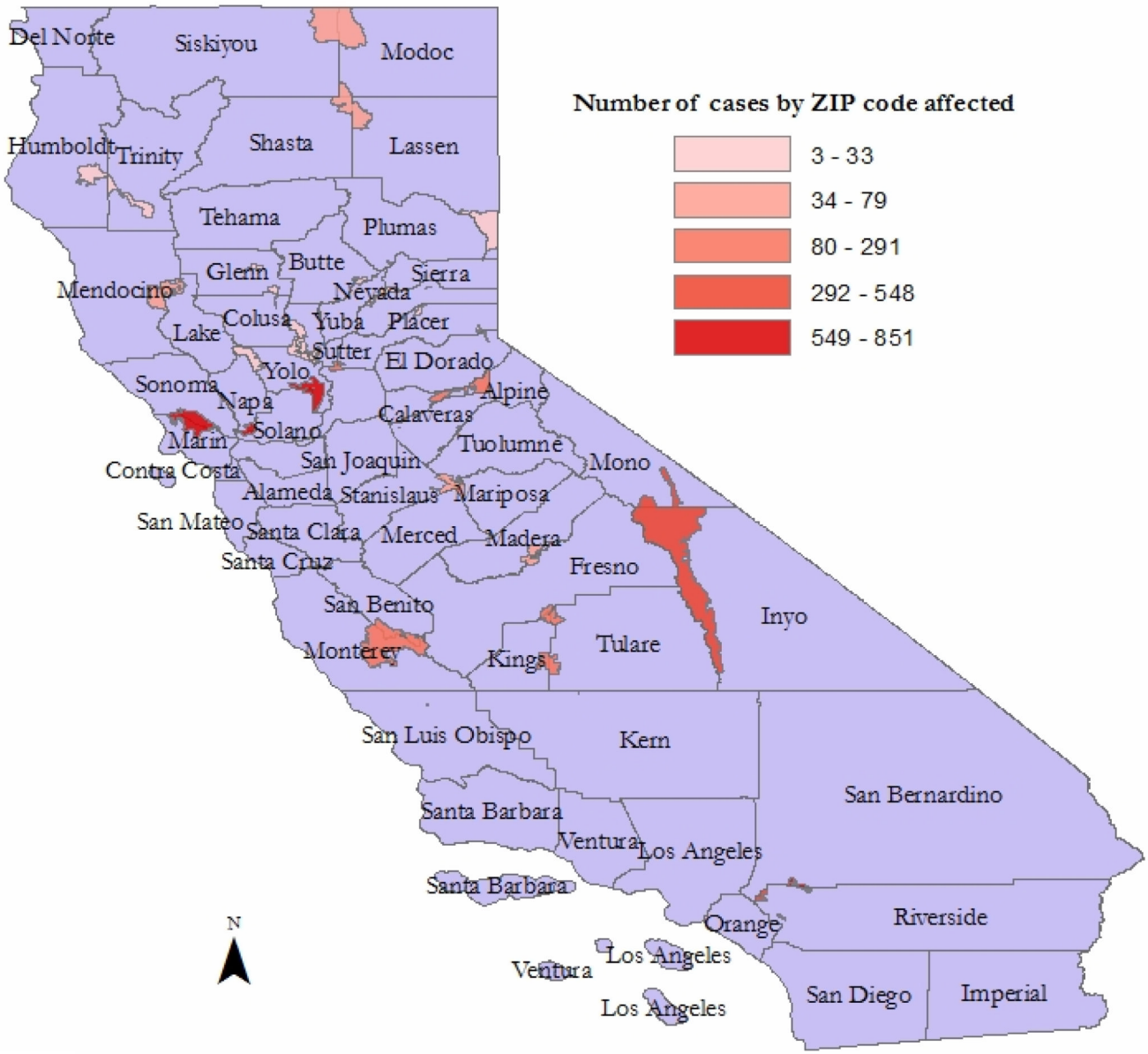

Figure 6.
Geographic distribution of ZIP codes affected by different areal unit computation; 2006–2010

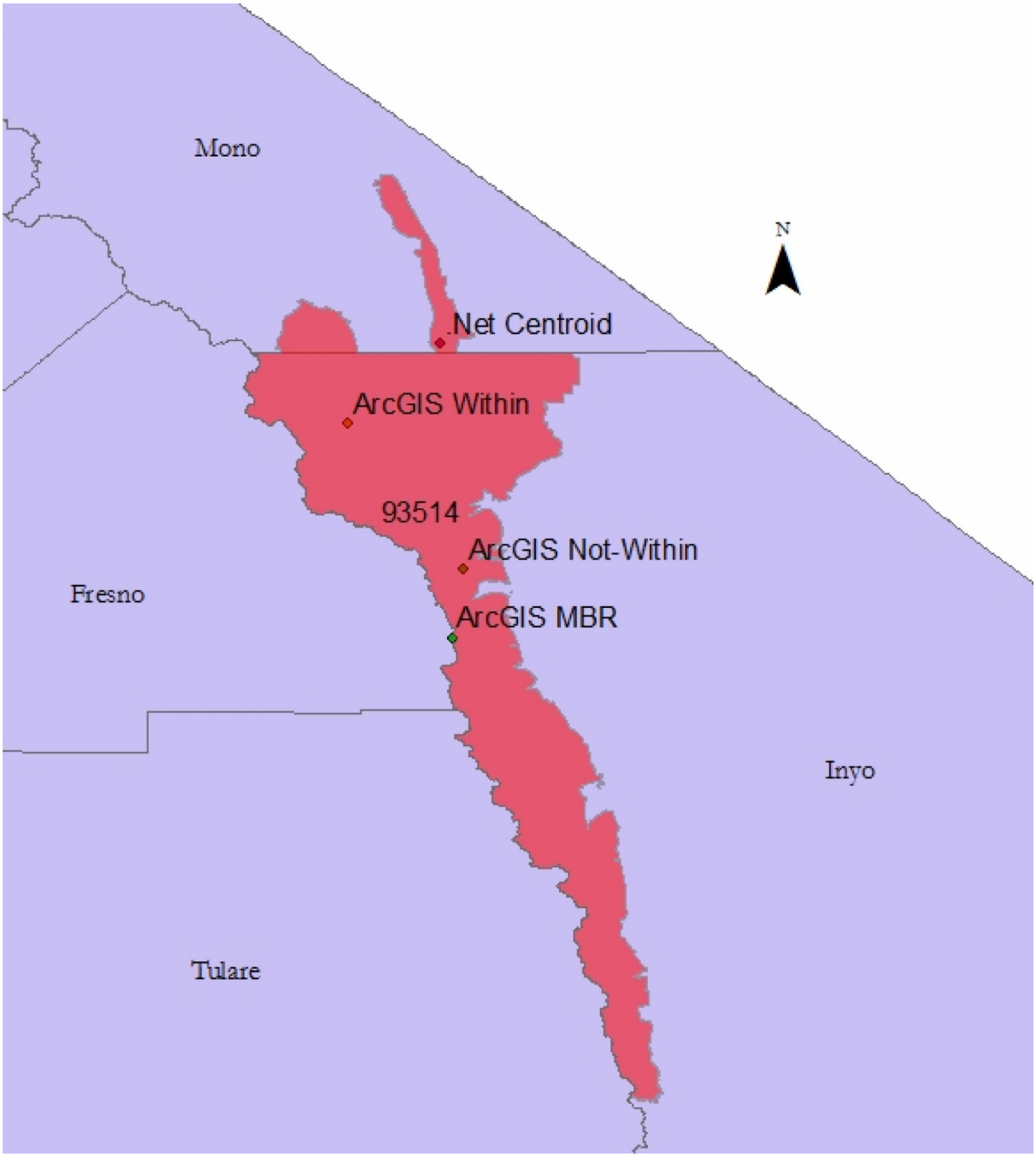

Figure 7.
Example ZIP code 93514 resulting in three different county assignments: Mono, Fresno, and Inyo Counties

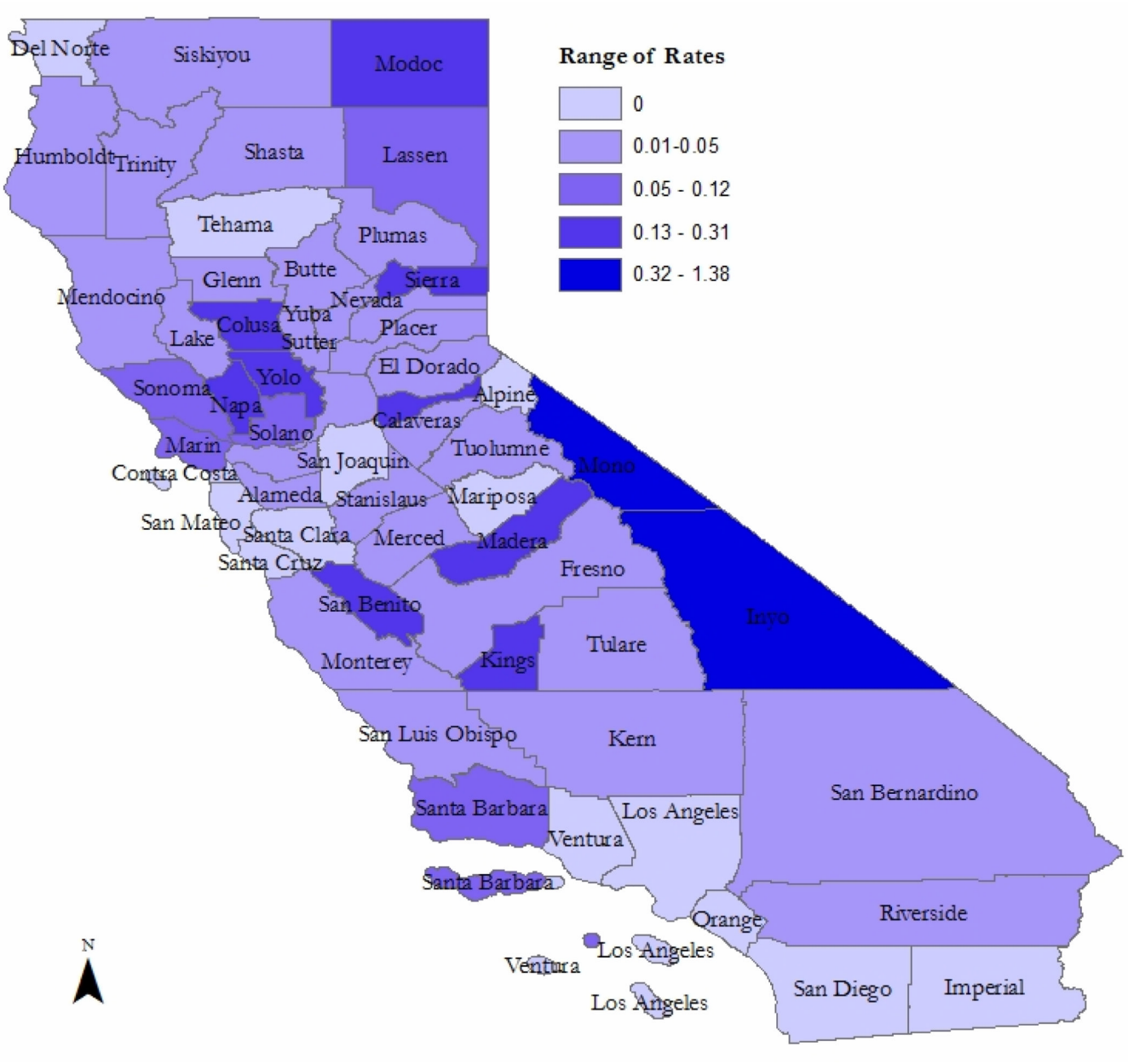

Figure 8.
Levels of percent change in cumulative five year incidence rates across methods; 2006–2010

Table 1

Descriptive statistics of the counts of cases per ZIP code; 2006–2010

Minimum	**1**
Maximum	2082
Mean	290
5% trimmed mean	252
Median	62
Standard deviation	377

Table 2

ZIP Codes resulting in different county assignments by method and number of cases affected; 2006–2010

ZIP code	Method 1	Method 2	Method 3	Method 4	Cases
94952	Marin	Sonoma	Sonoma	Sonoma	851
95616	Solano	Yolo	Yolo	Yolo	831
92324	San Bernardino	Riverside	San Bernardino	San Bernardino	678
94589	Napa	Solano	Napa	Solano	638
92880	San Bernardino	Riverside	Riverside	Riverside	548
94707	Alameda	Alameda	Alameda	Contra Costa	454
94708	Alameda	Alameda	Alameda	Contra Costa	417
93514	Fresno	Inyo	Inyo	Mono	360
93631	Tulare	Kings	Tulare	Tulare	291
93212	Kings	Tulare	Tulare	Kings	279
95666	El Dorado	Amador	Amador	Amador	196
93930	Monterey	San Benito	Monterey	Monterey	160
95626	Placer	Sacramento	Placer	Sacramento	145
93447	San Luis Obispo	San Luis Obispo	San Luis Obispo	Madera	79
95502	Humboldt	Humboldt	Humboldt	Madera	72
95329	Stanislaus	Stanislaus	Tuolumne	Tuolumne	67
93626	Madera	Fresno	Madera	Madera	60
96134	Siskiyou	Siskiyou	Siskiyou	Modoc	56
96056	Shasta	Lassen	Shasta	Shasta	52
95736	Placer	Placer	Placer	Madera	47
95469	Lake	Mendocino	Lake	Lake	46
95937	Colusa	Colusa	Colusa	Yolo	42
93216	Kern	Kern	Kern	Madera	38
95645	Sutter	Yolo	Sutter	Sutter	33
95225	Calaveras	Calaveras	Calaveras	Madera	32
95960	Nevada	Sierra	Nevada	Nevada	31
96109	Plumas	Lassen	Lassen	Lassen	24

ZIP code	Method 1	Method 2	Method 3	Method 4	Cases
95957	Colusa	Sutter	Sutter	Sutter	21
95481	Mendocino	Mendocino	Mendocino	Madera	20
95970	Colusa	Glenn	Glenn	Glenn	20
95526	Trinity	Humboldt	Trinity	Trinity	19
95233	Calaveras	Calaveras	Calaveras	Madera	17
95913	Glenn	Glenn	Glenn	Madera	14
94972	Sonoma	Sonoma	Sonoma	Madera	13
96142	El Dorado	Placer	Placer	El Dorado	12
95941	Butte	Butte	Butte	Yuba	11
92278	San Bernardino	San Bernardino	San Bernardino	Madera	11
95312	Merced	Merced	Merced	Madera	8
91719	Riverside	Riverside	Riverside	Madera	6
95606	Napa	Yolo	Yolo	Yolo	6
95715	Nevada	Placer	Placer	Placer	4
95981	Plumas	Yuba	Plumas	Plumas	3

Table 3

Five year incidence counts and cumulative five year incidence rates of cases by county per method (1–4) for counties with rates that do vary; 2006–2010

County	Population	Count (1)	Count (2)	Count (3)	Count 4)	Rate (%) (1)	Rate (%) (2)	Rate (%) (3)	Rate (%) (4)	Rate Mean (%)	Rate Range(%)
Mono	12,853	170	170	170	530	1.32	1.32	1.32	4.12	2.02	2.80
Inyo	17,945	170	530	530	170	0.95	2.95	2.95	0.95	1.95	2.01
Sierra	3,555	92	123	92	92	2.59	3.46	2.59	2.59	2.81	0.87
Yolo	168,660	2,711	3,581	3,548	3,590	1.61	2.12	2.10	2.13	1.99	0.52
Modoc	9,449	221	221	221	277	2.34	2.34	2.34	2.93	2.49	0.59
Colusa	18,804	447	406	406	364	2.38	2.16	2.16	1.94	2.16	0.44
Napa	124,279	4,244	3,600	4,238	3,600	3.41	2.90	3.41	2.90	3.15	0.52
Amador	35,100	1,094	1,290	1,290	1,290	3.12	3.68	3.68	3.68	3.54	0.56
Madera	123,109	2,591	2,531	2,591	2,948	2.10	2.06	2.10	2.39	2.16	0.34
San Benito	53,234	983	1,143	983	983	1.85	2.15	1.85	1.85	1.92	0.30
Kings	129,461	1,954	1,966	1,675	1,954	1.51	1.52	1.29	1.51	1.46	0.22
Lassen	33,828	602	678	626	626	1.78	2.00	1.85	1.85	1.87	0.22
Marin	247,289	8,799	7,948	7,948	7,948	3.56	3.21	3.21	3.21	3.30	0.34
Solano	394,542	8,756	8,563	7,925	8,563	2.22	2.17	2.01	2.17	2.14	0.21
Santa Barbara	399,347	9,266	9,266	8,463	8,463	2.32	2.32	2.12	2.12	2.22	0.20
Sonoma	458,614	10,950	11,801	11,801	11,788	2.39	2.57	2.57	2.57	2.53	0.19
Plumas	20,824	564	537	540	540	2.71	2.58	2.59	2.59	2.62	0.13
El Dorado	156,299	4,858	4,650	4,650	4,662	3.11	2.98	2.98	2.98	3.01	0.13
Tulare	368,021	6,657	6,645	6,936	6,657	1.81	1.81	1.88	1.81	1.83	0.08
San Bernardino	1,709,434	29,736	28,510	29,188	29,177	1.74	1.67	1.71	1.71	1.71	0.07
Siskiyou	44,301	1,404	1,404	1,404	1,348	3.17	3.17	3.17	3.04	3.14	0.13
Trinity	13,022	482	463	482	482	3.70	3.56	3.70	3.70	3.66	0.15
Contra Costa	948,816	22,367	22,367	22,367	23,238	2.36	2.36	2.36	2.45	2.38	0.09
Calaveras	40,554	1,421	1,421	1,421	1,372	3.50	3.50	3.50	3.38	3.47	0.12
Tuolumne	54,501	1,992	1,992	2,059	2,059	3.65	3.65	3.78	3.78	3.72	0.12

County	Population	Count (1)	Count (2)	Count (3)	Count 4)	Rate (%) (1)	Rate (%) (2)	Rate (%) (3)	Rate (%) (4)	Rate Mean (%)	Rate Range(%)
Riverside	1,545,387	38,672	39,898	39,220	39,214	2.50	2.58	2.54	2.54	2.54	0.08
Glenn	26,453	630	650	650	636	2.38	2.46	2.46	2.40	2.43	0.08
Alameda	1,443,741	29,906	29,906	29,906	29,035	2.07	2.07	2.07	2.01	2.06	0.06
Humboldt	126,518	3,184	3,203	3,184	3,112	2.52	2.53	2.52	2.46	2.51	0.07
Mendocino	86,265	2,352	2,398	2,352	2,332	2.73	2.78	2.73	2.70	2.73	0.08
Fresno	799,407	14,270	13,970	13,910	13,910	1.79	1.75	1.74	1.74	1.75	0.05
Lake	58,309	1,867	1,821	1,867	1,867	3.20	3.12	3.20	3.20	3.18	0.08
Monterey	401,762	7,091	6,931	7,091	7,091	1.76	1.73	1.76	1.76	1.76	0.04
Placer	248,399	9,433	9,304	9,449	9,245	3.80	3.75	3.80	3.72	3.77	0.08
Sutter	78,930	1,779	1,767	1,800	1,800	2.25	2.24	2.28	2.28	2.26	0.04
Nevada	92,033	2,835	2,800	2,831	2,831	3.08	3.04	3.08	3.08	3.07	0.04
San Luis Obispo	246,681	6,789	6,789	6,789	6,710	2.75	2.75	2.75	2.72	2.74	0.03
Shasta	163,256	5,067	5,015	5,067	5,067	3.10	3.07	3.10	3.10	3.10	0.03
Yuba	60,219	1,402	1,405	1,402	1,413	2.33	2.33	2.33	2.35	2.33	0.02
Stanislaus	446,997	9,273	9,273	9,206	9,206	2.07	2.07	2.06	2.06	2.07	0.01
Sacramento	1,223,499	28,402	28,547	28,402	28,547	2.32	2.33	2.32	2.33	2.33	0.01
Kern	661,645	11,755	11,755	11,755	11,717	1.78	1.78	1.78	1.77	1.78	0.01
Butte	203,171	4,983	4,983	4,983	4,972	2.45	2.45	2.45	2.45	2.45	0.01
Merced	210,554	3,616	3,616	3,616	3,608	1.72	1.72	1.72	1.71	1.72	0.00

Table 4

Counts and cumulative five year incidence rates of cases by county for counties that do not vary between methods (1–4); 2006–2010

County	Population	Count (1–4)	Rate (%) (1–4)
Alpine	1,208	20	1.66
Del Norte	27,507	680	2.47
Imperial	142,361	2,405	1.69
Los Angeles	9,519,338	171,726	1.80
Mariposa	17,130	508	2.97
Orange	2,846,289	56,420	1.98
San Diego	2,813,833	58,746	2.09
San Francisco	776,733	19,429	2.50
San Joaquin	563,598	11,809	2.10
San Mateo	707,161	17,553	2.48
Santa Clara	1,682,585	34,708	2.06
Santa Cruz	255,602	5,544	2.17
Tehama	56,039	2,549	4.55
Ventura	753,197	18,054	2.40