Technical Document:
Procedures for cleaning, geocoding, and aggregating crime incident data

John R. Hipp, Charis E. Kubrin, James Wo, Young-an Kim, Christopher Contreras, Nicholas Branic, Michelle Mioduszewski, Christopher Bates
Irvine Lab for the Study of Space and Crime
University of California, Irvine
November 1, 2016

Introduction

This technical document describes the geocoding and cleaning procedures used when geocoding crime data in Southern California cities for the project, *Crime in Metropolitan America: Patterns and Trends across the Southern California Landscape*, a project funded by the National Institute of Justice (NIJ) and led by principal investigators John Hipp and Charis Kubrin (Professors in the Department of Criminology, Law and Society at the University of California, Irvine). A key component of this project entailed the collection of incident crime data for cities located in the greater Southern California region. Cooperating police agencies reported incident crime data with geographic information for as many years of the study period as possible (2000-2013). Crime events were geocoded through a procedure using a geographic information system (ArcGIS 10), as well as two other procedures to attempt to geocode cases not geocoded by ArcGIS. All cases were aggregated to their corresponding census block (in both 2000 and 2010 boundaries). The following Part 1 crimes are represented by the data: homicide, robbery, aggravated assault, larceny, burglary, and motor vehicle theft.

Cleaning

Police departments provided data files in csv, excel, or dbf formats. Addresses associated with each crime incident were then "cleaned" using Stata 12/13 syntax in order to make the

addresses amenable for geocoding. For example, addresses that contained cross street connectors such as "/" were substituted by "&". In another example, apartment numbers were removed from all addresses. Accordingly, we corrected for any systematic spelling or abbreviation errors contained by addresses.

Some of the agencies only provided crime data at the 100 block level: that is, addresses were rounded down to the nearest 100 block due to privacy concerns. For example, a crime event that occurred at "2107 Culver Avenue" would be rounded down and reported by the police as, "2100 Culver Avenue." A crime event that occurred at "2188 Culver Avenue" would also be rounded down and reported by the police as, "2100 Culver Avenue." Given crime events that occurred on the same street (by intervals of 100) likely could have occurred on *either* side of that street, a concern was that geocoding of 100 block addresses could cause systematic error if this randomization process was not accounted for. Specifically, we would not want all crime events that occurred between 2100-2199 Culver Avenue to be allocated to the same side of the street and thus be placed in the same census block (note that different sides of the same street typically belong to different census blocks), with the accompanying assumption that none of them occurred on the other side of the street and therefore a different census block. Therefore, we needed an appropriate strategy for placing 100 block crime events to both sides of a street.

Our approach therefore randomly allocated a number from 0 to 99 for every crime event. We then add the street number (that was rounded and reported by the police) to this random number. For example, if the police reported that a burglary occurred at "2300 Culver Avenue" and the random number allocated was "45", then, the final address that would be used for geocoding would be, "2345 Culver Avenue".

Geocoding

Stage 1: Geocoding in ArcGIS

We geocoded crime events using a three-step procedure. First, all crimes (for all years) of a city were geocoded using Esri ArcGIS 10. Although many researchers use the national address locator ("Street_Addresses_US") along with default geocoding specifications (minimum match score: 85; minimum candidate score 10; spelling sensitivity: 80) provided by Esri, we performed tests to assess the appropriateness of these default settings. We discuss the results of these tests next.

Settings for spelling sensitivity and match rate in ArcGIS

There is little guidance in the literature regarding the settings to use in a program such as ArcGIS when geocoding events. A challenge is twofold: 1) we do not want to geocode an event to an incorrect location; but 2) we do not want to discard observations that we can actually geocode. Indeed, this verges quickly into a missing data problem, although little attention has been given to it. Whereas some researchers adopt a strict strategy of only accepting geocodes when they are very certain of the accuracy, this contains an often unacknowledged assumption that events that were not geocoded in fact did not occur. This is clearly wrong, and thus discarding cases due to an inability to geocode them is quite dangerous. Existing guidance in the literature that a match rate of 70% is acceptable has virtually no statistical support (Ratcliffe 2004). An important consideration is how much accuracy the study needs, and a decision about the possible tradeoffs between type 1 and type 2 error regarding the accuracy of the geocoding. If the research is aggregating the observations to street segments, or to blocks, then a geocoded observation that is a few houses off is effectively a perfect match. However, those that are geocoded to one or two blocks away will necessarily be geocoded to the wrong block. The

question then is whether it is better to geocode the observation to this incorrect block, or to discard the observation entirely. The problem with dropping such observations is that this relies on an assumption of complete randomness to these now "missing" observations. But in fact we know that crime events tend to be clustered. Furthermore, if these events are only one or two blocks away from the correct location, then there is a systematic spatial pattern to these excluded observations, which violates the assumption of randomness. Thus, dropping these observations is almost certainly not justified statistically in most practical research settings for criminologists.

We explored this issue with our data in the following way.[1] First, we geocoded crime events in three example cities in which we systematically altered the settings for the spelling sensitivity and the match rate. Second, we geocoded these same observations using the Google geocoder. We then could compare the results in several fashions: 1) we computed the distance between the ArcGIS geocoded result and the Google geocoded result; 2) we compared the initial address and the one that ArcGIS geocoded the observation to.

We briefly describe some of the results. When setting the spelling sensitivity to 70, Table 1 shows the median distance between the Google geocoded location and the ArcGIS one for various match scores for the city of Fontana (we performed similar analyses for the city of Colton, and the results were similar). As can be seen, the distances are relatively short even for relatively low scores. For example, scores above 40 typically had a median distance less than ¼ mile, and often considerably less. Even match scores between 25-30 are less than 1/10 mile off.

Table 1.

ArcGIS Match Score	Median distance to Google geocoded point (miles)
67	.02
50-67	.47

[1] These files are in this folder: Z:\natl\neighs\crime_SoCal\Geocoding_Sens.

43	.01
42	.13
41	.11
40	.23
30-39	.55
25-30	.07
20-25	.82
< 20	1.56

In looking closer at these cases, we found that the cases with a match score of 43 typically were off by just a parcel or two (i.e., 9900 Briarwood vs. 9902 Briarwood; 7600 Lemon Street vs. 7598 Lemon Street; etc.). Given our aggregation to blocks or street blocks, many of these are placed in the same unit in either case. For those incorrectly placed, it is typically just one block away from the correct location. The spatial specificity of this error would have problematic consequences if these cases were simply omitted. Those with a match score of 41 or 42 were often off by about a 100 block; and those with a match score of 40 were typically off by about 300 to 500 block (i.e., 4100 Main St. vs. 4400 Main St.). Those with match scores in the 30's were often off by a larger amount, which is why their distance from the google geocoding is getting relatively high. However, the ones between 25 and 30 often were missing a part of the address name (i.e., "Street" or "Drive") and were perhaps a 100 block off, and thus were frequently quite close to the Google geocoded points even though they have relatively lower match scores.

We found that pushing the spelling sensitivity below 70 produced unsatisfactory results. Typically, such cases were matched incorrectly (i.e., Bark St. as Burke St.; Brant St. as Brandon St., etc).

Given these results, we implemented the address locator with modified geocoding specifications (minimum match score: 10; minimum candidate score 10; spelling sensitivity: 70).

Based on our own sensitivity checks, we treated addresses that received a score of 25 to 100 as matched.

Using a custom address locator in ArcGIS

One challenge we faced when geocoding was that on *rare* occasions the address locator provided by Esri, "Street_Addresses_US," produced low overall match results. These instances almost always occurred when we were geocoding data for a County Sheriff that served several cities in a county, and almost never happened when the agency served a single city. That is, for a given city/county, less than 70% of crime events were matched. In these instances, we created custom address locators in ArcGIS based on street files provided by the U.S. Census Bureau (see, https://www.census.gov/geo/maps-data/data/tiger-line.html). Geocoding using these custom address locators for initially low match rate cities/counties, resulted in improved match rates greater than 90%.

Stage 2: using another geocoding platform for unmatched observations

In the second stage, cases that were not matched in ArcGIS (match score < 25) were then geocoded using the MapQuest Geocoding API. Unlike ArcGIS that provides a score from 0 to 100, the geocoding accuracy of MapQuest is indicated by one of the following geographic aggregations: state, county, city, zip, street, address, and point. That is, the addresses produced by MapQuest are designated the most spatially precise aggregation. We deemed all cases that received "address" or "point" to be matched. One wrinkle is that for a given crime event MapQuest can provide more than one address as a possible "match," although we found this to be rare. As a result, we later account for this feature when aggregating (discussed later).[2]

Stage 3: using Google Earth Pro for remaining unmatched observations

[2] For a few cities, we used the Google geocoding API, rather than the Mapquest API.

The third stage of the geocoding used Google Earth Pro to geocode cases that remained unmatched from the first two stages. Google Earth Pro was cumbersome to use, as the output data was in a format that was difficult to work with. Stata code was written to clean the data for Google Earth Pro, and then to read the cases back in to Stata after geocoding and merging with the other observations.

This three stage geocoding procedure yields results broken down by ArcGIS matched (%), MapQuest matched (%), Google Earth Pro matched (%), and unmatched crimes (%). Therefore, the total match rate is the sum of the ArcGIS and MapQuest and Google Earth Pro percentages. All but a few cities/counties produced total match rates greater than 90%. Using ArcGIS and MapQuest, we attach the corresponding census block FIPS code to every matched crime address (for 2000 and 2010 census boundaries).

Aggregating

Census block FIPS codes pertaining to crime events were used to aggregate to three spatial units (blocks, block groups, and tracts). Given that both block group and tract FIPS codes are contained within the fifteen digit FIPS codes of blocks, aggregating to geographically larger spatial units than the latter was straightforward.

MapQuest geocoding produced multiple "matches" for a small percentage of crime events, thereby necessitating a weight for such events. Specifically, we calculated a weight in these situations as 1 divided by the number of "matched" addresses. For example, if MapQuest geocoding produced 4 "matched" addresses, then, each of these addresses would be considered ¼ of a crime.

For most of the crime data, it was straight forward for us to discern the type of crime that occurred, as reporting agencies explicitly delineated the type of Part 1 crime (or provided the corresponding penal code). However, for a small number of cities, such agencies only provided a description of the crime. Thus, we had to code these crimes ourselves. As a result, we felt it necessary to ascertain the accuracy of our crime data. Specifically, we empirically compared our summed Part 1 crime counts for a city to those counts published by the FBI's Uniform Crime Reports (UCR). In instances in which there was a significant discrepancy between the UCR reported values and our summed values (defined as a difference of +/- 10 percent between the totals), we revisited our code to assess whether other categories of the crime code needed to be included in the larger classification. For most of the cities, we were able to obtain results that were relatively similar to the UCR values (there will be differences, given the reclassification of crime events over time by police agencies due to new information about the event, etc.).

References

Ratcliffe, Jerry H. 2004. "Geocoding crime and a first estimate of a minimum acceptable hit rate." *International Journal of Geographical Information Science* 18:61-72.