Jaclyn Scholl
Maria Prieto

November 29, 2007
Version 5

# Geocoding report

*The Tax99 dataset contains sale prices and assessment values, as well as location and owner information, for most properties in the town of New Haven, CT. Exact latitude and longitude information for each property was desired to represent the data visually.. We had two competitive resources at our disposal to aid with this task: Google and Tiger (a database containing coordinate information for New Haven street segments). We explored the performance of both sources, and after close examination found Tiger to be more accurate. Following this conclusion, we used Tiger data in combination with a linear interpolation formula to compute Tax99 coordinate information.*

## I. Introduction—our data and Tiger

Our goal for this project was to geocode, or assign locations to all the properties in Tax99, a dataset which contains 1998 tax assessment information for 27,323 properties in New Haven, CT. Tax99 lists each property by a unique identifying code, the MBP,[1] and provides various sorts of land and building valuations for each property, as well as owner information, latest sale date, latest sale price, and property type among others.

To obtain geographical information for each property we used data from the 2006 Second Edition TIGER/Line® Files, which can be obtained online from http://www.census.gov/geo/www/tiger/tiger2006se/tgr2006se.html. The TIGER/Line® Files are extracts from the Census TIGER® (Topologically Integrated Geographic Encoding and Referencing) database of the U.S. Census Bureau,[2] and contain information for all counties (and statistical equivalents) in the United States. They consist of 19 record types that present different types of geographic information.[3] We extracted a record type 1 subset for the New Haven county,[4] and from there we extracted the records roughly corresponding to New Haven town. The latter extraction required comparing the fips (federal information processing standards) codes on the left and right of each segment. We kept the segments for which either the fips code on the left or on the right corresponded to the one for the town of New Haven.[5]

The record type 1 Tiger/Line® file for the New Haven town (Tiger, for short) contains data describing line segments for roads, railroads, shorelines, rivers, and non-visible features such as

---

[1] MBP stands for "map", "block", "parcel". It is a 12 digit number in the form "mmm bbbb ppppp".

[2] The US Census Bureau's Census TIGER® System contains a digital geographic data base that includes coverage of all of the US, and automates the mapping required to support census and sample survey programs of the US Census Bureau. The US Census bureau releases periodic extracts of the database for public use. These include the Tiger/Line® files.

[3] For full technical documentation for the Tiger files is available in pdf format at http://www.census.gov/geo/www/tiger/tiger2006se/TGR06SE.pdf .

[4] The fips code for the New Haven county is "09 009", where 09 is the code for CT and 009 the code for the New Haven county.

[5] New Haven town fips code: 52070

jurisdictional boundaries[6]. It contains unique records for each segment, which are identified by a permanent 10-digit number (TLID). Tiger provides beginning and ending street numbers for the right and left sides of the road for over 70% of the road segments corresponding to New Haven town. Exceptions include parkways, connectors, and other roads without street numbers. Tiger also provides beginning and ending latitude and longitude information for both the left and right side of the segment.

We extracted the following subset of variables:

| Field | Description | Notes |
|---|---|---|
| TLID | TIGER/Line® ID, Unique segment identifier | 10 Digit Identifier |
| FEDIRP | Feature Direction, Prefix | Levels: N, S, E, W |
| FENAME | Feature Name | Road Name |
| FETYPE | Feature Type | E.g. Ave, Pky, St, Aly |
| FEDIRS | Feature Direction, Suffix | Levels: N, S |
| FRADD (R/L)* | Start Address, Left | |
| TOADD (R/L)* | End Address, Left | |
| BLOCKL | Census Block Number, 2000 Left | |
| BLOCKR | Census Block Number, 2000 Right | |
| FRLONG (R/L)* | Beginning of Segment Longitude | |
| FRLAT (R/L)* | Beginning of Segment Latitude | |
| TOLONG (R/L)* | End of Segment Longitude | |
| TOLAT (R/L)* | End of Segment Latitude | |

*(R/L) fields with this notation correspond to two variables; one corresponds to the right and the other to the left of the segment.

## II. Tiger Geocoding

In order to map information from Tiger onto the Tax99 data, we first had to identify a feature common to both datasets. Block information proved to be of no utility in this regard, because the census block numbers in Tiger do not correspond to the block component of the MBP in tax99. Instead we decided to match by street name and street number. For our first matching attempt we focused solely on Hillhouse Avenue properties. We extracted the following information from Tiger for each property: TLID, FRLAT, TOLAT, FRLONG, TOLONG, FRADDR, TOADDR. Then we used linear interpolation to estimate individual longitudes and latitudes.

When we attempted to extend the method to all properties in Tax99, we discovered the following:

a) There is only one street in New Haven called Hillhouse. This is not true for others such as Church where there is a Church Street and a Church Street South.
b) Matching streets is a matter of matching not only street names in Tax99 and Tiger, but also prefixes, suffixes, and cardinal directions. The spelling of some of the prefixes and suffixes is different in Tiger and Tax99. (Ex. Tiger: Ave, Pky vs. Tax99: Av, Pkwy)
c) Tiger does not have segment information for all the street names in the Tax99 data. In fact, after various manipulations described below, we only managed to establish correspondence between 652

[6] Tiger provides segment type codes for the various types of line segments in the CFCC (Census Feature Class Code) field. A's are roads, B's are railroad tracks, Fs are non-visual boundaries, and H's are shorelines, rivers and creeks.

out of the 699 street names in the Tax99 data and the Tiger data.[7] That left us with 47 street names with no matching potential. Those 47 streets contain information for 386 properties, and include the Foxon Hill Rd and the Spring Street with more than 90 properties in each. Considering that the Tax99 data has information for 27,323 properties, 386 does not seem like too large of a number, but it is still of some concern.

d) Not every entry in Tax99 has a street number (1437 entries do not have one at all, and others have characters in the number—lot numbers, As, Bs, etc). Without this street number, our interpolation mechanism fails.

e) Our loop maps a street number to a street segment for the corresponding street by checking if the street number is smaller than the ending street number (E) and greater than the beginning street number (B). However, E is not strictly larger than B for all segments.

Here are some examples from the Tiger data. Rows 1-4 exemplify "a" above, and rows 5-6 exemplify "e").

| TLID | FEDIRP | FENAME | FETYPE | FEDIRS | FRADDL | TOADDL | FRADDR | TOADDR |
|---|---|---|---|---|---|---|---|---|
| 3697307 | W | Prospect | St | | 70 | 112 | 61 | 107 |
| 3701157 | | Prospect | St | | 116 | 190 | 125 | 155 |
| 3701160 | | Prospect | Pl | | 12 | 24 | 11 | 25 |
| 3704124 | | Prospect | Ave | | 276 | 354 | 279 | 343 |
| 3754598 | | Hoover | St | | 298 | 2 | 299 | 1 |
| 3704027 | | East Shore | Pky | | 598 | 2 | 599 | 1 |

In order to circumvent items b and d above and in order to make it easier to map a property to the proper side of the segment, we reorganized the Tiger data by implementing the following changes:

1. We extracted segments corresponding only to road information based on the CFCC field (Census Feature Class Code) in Tiger. All A's correspond to different types of roads.
2. We changed the spelling of various suffixes, prefixes, cardinal directions, numbers, and street names in Tiger to match those of the Tax99 data.
3. We checked for parity matches between the beginning and ending number of the right and left sides of each segment. We confirmed this is true for all segments.
4. We split each line segment into left and right segments, doubling the number of rows in our data structure. We assigned a number to establish the parity of the segment (0 if even and 1 if odd), and attached it as an additional entry in that row (essentially creating a new "parity" column).
5. We rearranged the *to* and *from* street addresses so address numbers increased along each segment.
6. We created a column with the full street name information for each segment, containing a string of prefix, street name, suffix, street number, and direction. We wrote this string in upper case to fully match the format of the Tax99 data.

The following are some examples of the data transformations. We can see that now every TLID has been broken into two different entries.

---

[7] We combined the FEDIRP, FENAME, FETYPE, FEDIRS for each feature, to create a full street name variable. Eg: "W Prospect St". We used the street name variable to do the matching with Tax99.

| TLID | FEDIRP | FENAME | FETYPE | FEDIRS | from | to |
|---|---|---|---|---|---|---|
| 2549 3754598 | | Hoover | St | | 1 | 299 |
| 2550 3754598 | | Hoover | St | | 2 | 298 |
| 3927 3704124 | | Prospect | Ave | | 276 | 354 |
| 3928 3704124 | | Prospect | Ave | | 279 | 343 |

| TLID | FRLONG | FRLAT | TOLONG | TOLAT | street.name | parity |
|---|---|---|---|---|---|---|
| 2549 3754598 | -72897545 | 41291697 | -72896085 | 41291556 | HOOVER ST | 1 |
| 2550 3754598 | -72897545 | 41291697 | -72896085 | 41291556 | HOOVER ST | 0 |
| 3927 3704124 | -72894245 | 41287097 | -72893745 | 41288597 | PROSPECT AV | 0 |
| 3928 3704124 | -72894245 | 41287097 | -72893745 | 41288597 | PROSPECT AV | 1 |

We used this rearranged subset of the Tiger data to produce our final results. The specific methodology we used will be described in section IV.

## III. Google Geocoding

As explained in part II, we used street names as the matching element to geocode Tax99 from Tiger data. However, given that Tiger did not have information for all the streets in Tax99, we looked for an alternate source of coordinate information. We utilized a Geocoding function, which enabled us to extract Google coordinates from inputted street addresses. We called it the Google Geocoder. We also wrote a function we called Mapmaker, which allowed us to plot the obtained coordinates onto fully functional Google maps.

With the Google Geocoder we were able to obtain coordinate information for virtually every address we entered. For reasons that still remain unclear, the function returned zeros every now and then. Simply running it again returned non-zero coordinate values. Since the Geocoding function looks up individual entries online, we think this might have to do with failed matches and the second attempt is equivalent to refreshing a window that didn't load on the first place. To avert this minor problem we repeated the geocoding until the results were non-zero values.

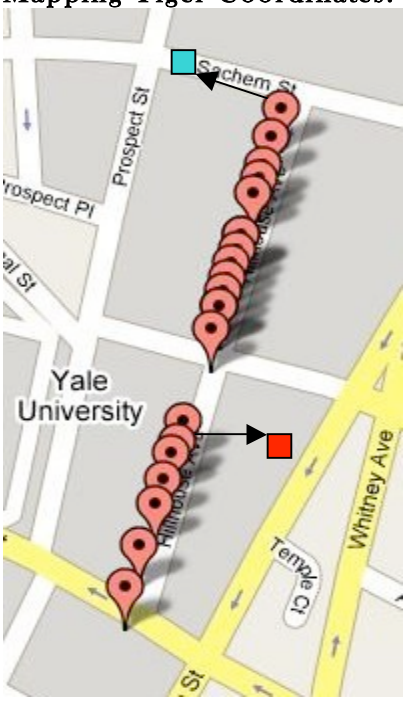
The initial advantages of using this method include the fact that Google accepts the property location in multiple forms. It is not necessary, for example, to input an address in the standard street number, street name, city, state format; we were actually able to enter the information in any order, and Google's geocoding could handle this. Likewise, google interprets various forms of spellings for suffixes and street types. For instance it understands that both "Av" and "Ave" stand for "Avenue". Another advantage is getting exact coordinate information for each property without having to interpolate it from other data. One major disadvantage, however, is that it takes much longer to extract information from Google than from Tiger, as the Google Geocoder looks up each entry online. We also noticed that Google was able to return results for a wider range of street addresses than what we were able to geocode with Tiger data. However, some of these additional matches correspond to addresses without a street number. We are not sure that this necessarily good. For long streets, the Google estimations for un-numbered properties have the potential to be far away from their actual locations. Perhaps it is better to leave such properties with no assigned coordinates.

## IV. Performance Comparison

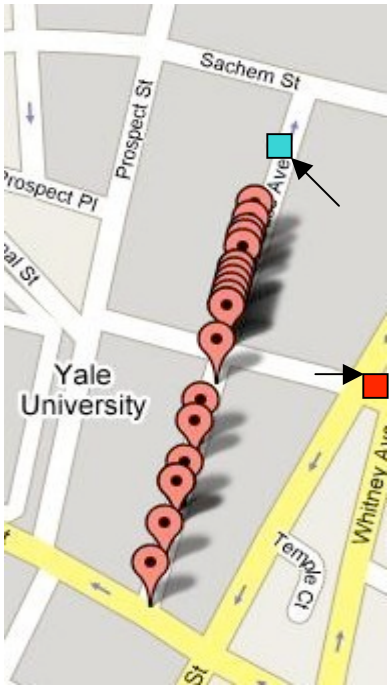
With these two resources at our disposal, a performance evaluation was necessary to compare the coordinates obtained from Tiger with Google's estimations. In an effort to gain a small-scale understanding, we first considered the data for Hillhouse Avenue. We utilized the Google Mapmaker to plot both the

Google and Tiger coordinates for each of the Hillhouse properties on Google-generated maps. We noticed that the Tiger coordinates produced a consistent distribution of points, while Google clumped many of them too close together. (The red and blue squares were added to identify the same property on both maps. The red corresponds to 24 Hillhouse Av, and the blue to 55 Hillhouse Ave.)

**Mapping Tiger Coordinates:**




**Mapping Google Coordinates:**



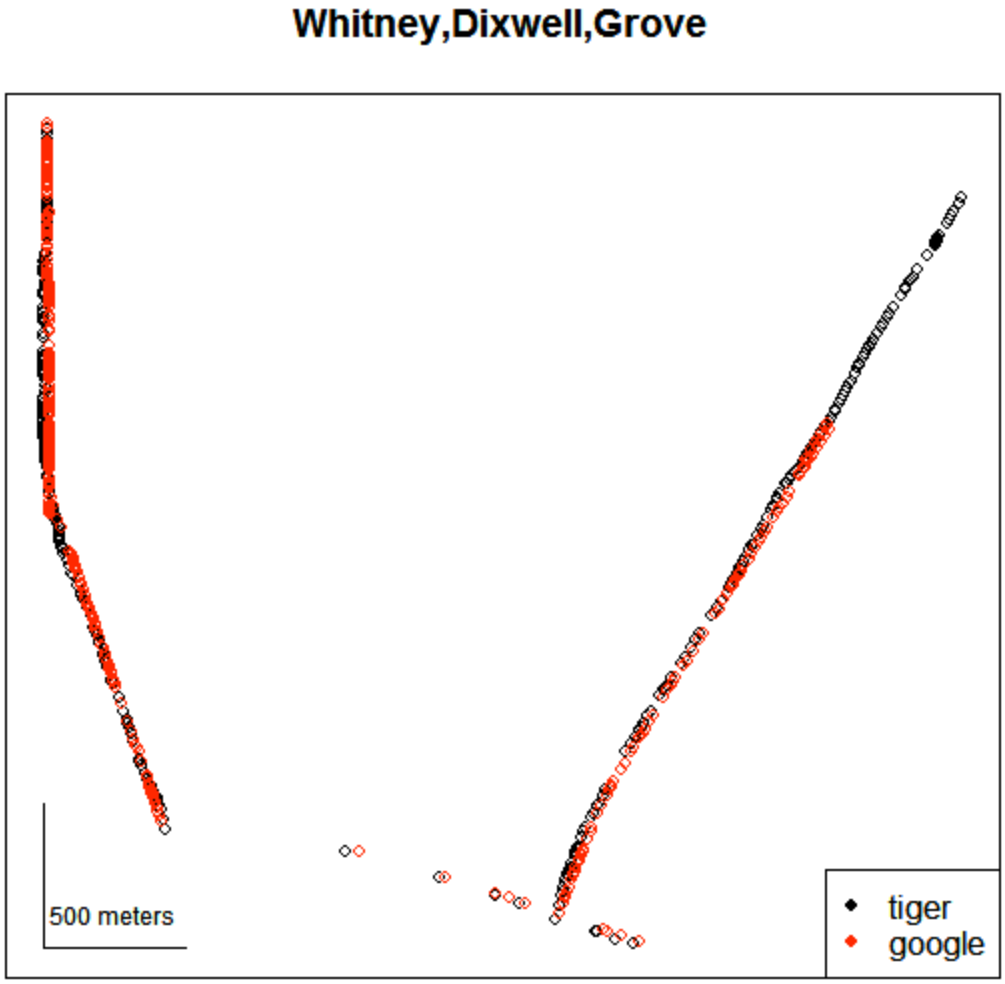

After establishing that there was a difference between the Google and Tiger coordinates, we explored whether this difference was a constant number that would allow us to convert between the coordinates produced by both methods. We computed the differences between the latitudes and longitudes for each Hillhouse property. As seen from the results below, we detected no distinct constant or pattern. We realized we would not be able to easily manipulate one set of coordinates to convert to the other. We needed to make a decision: which source was best?

**Differences in Tiger and Google Coordinates for selected Hillhouse Avenue Properties (in meters)**

| Street Number | Latitude | Longitude |
|---|---|---|
| 5 | -13 | -5.8 |
| 9 | -15.1 | -0.9 |
| 15 | -4.9 | -1.8 |
| 17 | 4.7 | -1.8 |
| 24 | 11 | -0.8 |
| 27 | 17.9 | 0.3 |
| 30 | 23.7 | 1.3 |

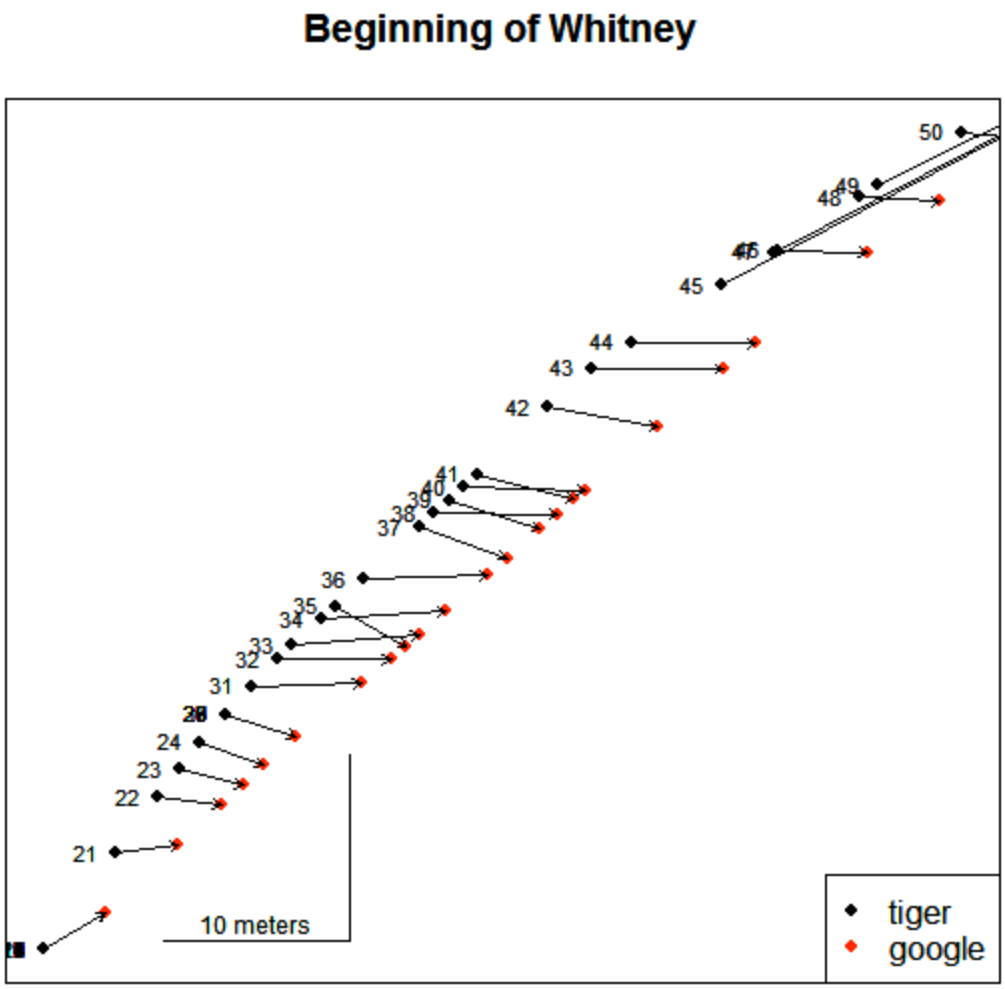
| Street Number | Latitude | Longitude |
|---|---|---|
| 34 | 30.9 | 2.3 |
| 37 | 37.4 | 3.4 |
| 43 | 57.1 | 6.3 |
| 46 | 61.8 | 7 |
| 47 | 70.3 | 8.1 |
| 51 | 83.6 | 9.9 |
| 55 | 96.7 | 11.6 |

We extended our Google/Tiger comparison to include the mapping of additional streets. With more examples of each method's geocoding capabilities, we felt we would be able to draw a stronger conclusion about overall performance. We chose to concentrate on Whitney Avenue, Dixwell Avenue, and Grove Street as their close proximities would allow for all their points to be seen on one map. As anticipated from the Hillhouse maps, we found significant differences between Tiger's and Google's estimated coordinates. Of the three streets, Dixwell stood out as the one with the greatest Tiger/Google similarities, as its black and red points were clustered closest together. Noticing greater spreads on Whitney and Grove, we chose to further examine those two streets. (Note that for the following four plots the scale applies in the horizontal and vertical axis.)

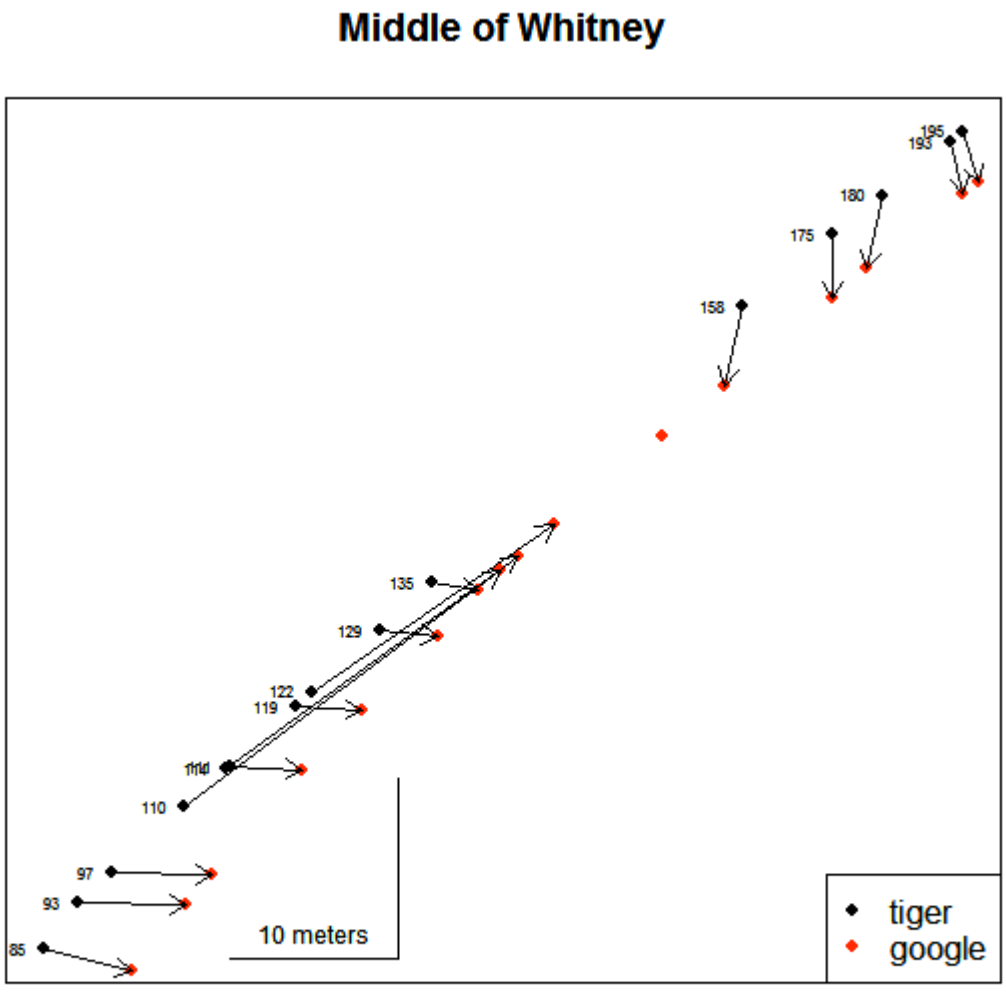


The plot of Grove Street further supported our convictions; however, while the Google and Tiger coordinates clearly differed, there was no excitingly strange or unusual pattern. All of the Google estimates, for example, were consistently positioned to the right of Tiger's. In addition, the Google latitudes were, for the most part, always a little larger. It should also be noted that the centrally located Google point lacking a corresponding Tiger estimate corresponds to the Grove properties that did not have a listed street number. Because street numbers were an essential part of our interpolation formula, we were unable to assign specific coordinates to the properties where this information was missing. Google, however, estimated these unknown street-numbered properties by assigning them the coordinates of a street segment midpoint (the middle of the streets for short streets like Grove). In addition to this, it should also be noted that the differences between estimations on this plot are only a couple meters. This might be due to each source's differences in handling street width or deciding the exact start of property lines.

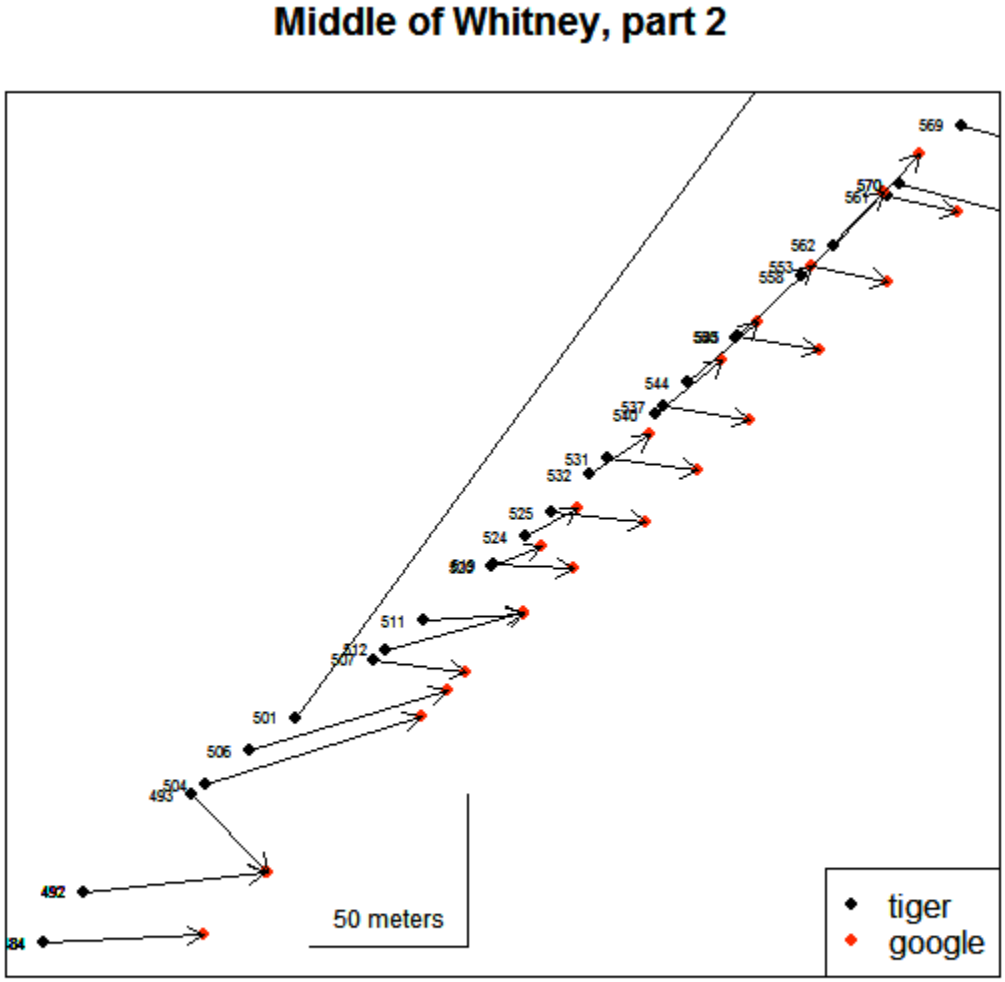
To examine Whitney Avenue, we had to break it into smaller segments due to the large number of properties on it. The plots below show some interesting findings. (Before describing them, note that the places where the street numbers look extra dark represent properties with multiple units; the points and labels were mapped to the one location multiple times.) A unique observation can be gained from each of the three maps. The first shows that the Google coordinates do not map street addresses in increasing or decreasing order. This would not matter if it only had to do with a mismatch in increasing odd and even sequences (such as 2,7,4,9,6,11), because odd and even numbers represent different sides of the street. This plot, however, shows that Google sometimes plot points with the same parity out of order. For instance, we can see that Google puts 35 Whitney Av. before 33 Whitney Av (seq.: 31,32,35,33,34,36). This threw up a red flag. Google puts 45, 47, and 49 Whitney Av. right outside the area of this plot, on a point that corresponds to a segment midpoint.



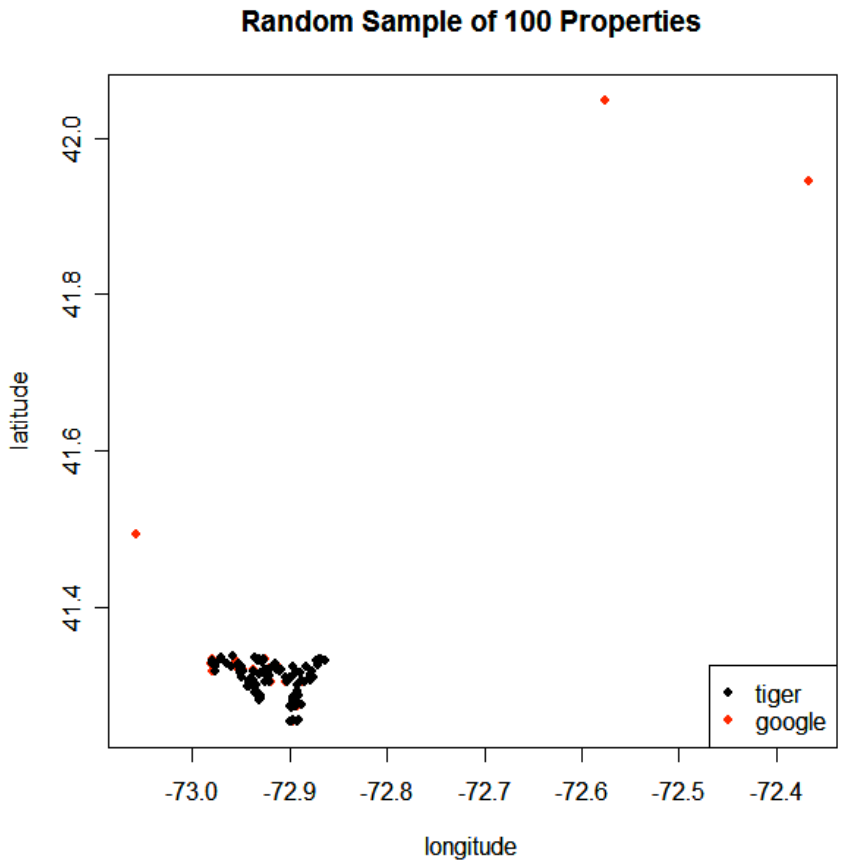
The next plot examines a mid-section of Whitney Avenue. It shows larger differences between Google and Tiger coordinates, with particularly striking disagreements for 110, 114, and 122 Whitney Av. It is interesting to note the red point in the middle with no corresponding Tiger coordinate. This point represents 155 Whitney Av. Our interpolation method with Tiger failed to provide an estimate because the segment ranges available for Whitney Av. have a gap between 151 and 157. Google plots it on a segment midpoint different from the one mentioned in the previous paragraph. This highlights one of the disadvantages associated with using the Tiger data.



Many things are occurring in this final Whitney Ave plot. It is first interesting to notice that Google maps different street addresses to the same location. For example, 492 and 493 Whitney, both of which are given their own Tiger coordinates, have the exact same Google location. In addition to this, it is interesting to see that starting at 519 Whitney, Google seems to be mapping properties to their appropriate street side; all the even-numbered properties are significantly positioned to the left of the odds. Finally, we can see a striking difference in the mapping of 501 Whitney. The Google estimation lies just 10 meters North of the range of the plot, another segment midpoint.



To supplement our conclusions from the street maps just presented, we took a random sample of 100 Tax99 properties and mapped them using both the Tiger and Google coordinates. The plot below shows a very striking finding. The shape of the cluster of points provides a good representation of the actual shape of New Haven. We would expect such a thorough covering of the city due to the nature of a random sample. The three Google estimates outside this cluster, however, are what is so significant about this plot. Google seems to be mapping points that we know are located in New Haven far outside the city bounds! Most strikingly perhaps, Google maps (www.maps.google.com) plots these points inside the boundaries of the town of New Haven.




The final factor in our determination of the most accurate geocoding resource was a website conversion calculator.[8] After prompting for the input of an address, the site retrieves coordinate information for that property from multiple sources (Google, Geocoder, Yahoo, Terraserver). To test out what we were seeing in the previous plots, we decided to look up 110 Whitney Avenue, one of the points where Tiger and Google differed the most. According to our Tiger interpolation formula, we estimated this property to be located at latitude 41.3135 and longitude -72.92178. As can be seen from the website's

---

[8] source(http://stevemorse.org/jcal/latlon.php)

display, our results are more consistent with Geocoder, Yahoo, and Terraserver; Google's estimations are most off!

| from google | latitude | longitude |
|---|---|---|
| decimal | 41.314479 | -72.921284 |
| deg-min-sec | 41° 18' 52.1244" | -72° 55' 16.6224" |

| from geocoder | latitude | longitude |
|---|---|---|
| decimal | 41.313497 | -72.921779 |
| deg-min-sec | 41° 18' 48.5892" | -72° 55' 18.4044" |

110 Whitney Ave , New Haven CT 06511

| from yahoo | latitude | longitude |
|---|---|---|
| decimal | 41.314061 | -72.921741 |
| deg-min-sec | 41° 18' 50.6196" | -72° 55' 18.2676" |

110 Whitney Ave, New Haven, CT 06510

| from terraserver | latitude | longitude |
|---|---|---|
| decimal | 41.31361200 | -72.92188100 |
| deg-min-sec | 41° 18' 49.0032" | -72° 55' 18.7716" |

110 Whitney Ave, New Haven, CT 06510-1235

*Differences between our Tiger estimate (41.3135, -72.92178) and the various sources (in meters):*

**Google**:

Latitude: 108.7 meters
Longitude: -18.9 meters

**Geocoder**:

Longitude: -0.3 meters
Latitude: .04 meters

**Yahoo**:

Longitude: 62.3 meters
Latitude: 1.5 meters

**Terraserver**:

Longitude: 12.4 meters
Latitude: -3.9 meters

## V. Methodology to Produce Final Results and Assessment

Combining the conclusions we drew from our findings, we decided to use Tiger to map all of the Tax99 properties. We constructed a loop to run through each of the entries in the Tax99 dataset; it operated in the following manner: First, the street name, suffix, and number were temporarily held aside. Next, it searched the Tiger dataset and isolated all entries containing information for that street name. It then tested the street number to be even or odd and further isolated the Tiger set to include only the entries with both the same street name and parity. Finally, the appropriate Tiger information was assigned to the Tax99 property when the street number fell within the range of the Tiger segment. Once the coordinates and address ranges for each segment were matched to each property, our interpolation formula was utilized.

After running through Tax99 in its entirety, we were unable to geocode 2,268 properties (out of 27,323) with Tiger for the following reasons:

1. 1437 properties lack a street number. This often occurred with vacant land.
2. Tiger does not have segment information for all the streets in Tax99. One example is Andress Street, a location listed only in Tax99. A reason this is occurring might be because the street no longer exists.
3. The Tax99 street number did not always fit within a Tiger address range.
    - There are gaps within the different Tiger intervals. For example, on one side of Trumbull Street, the segments go from 38 to 48 and then from 60 to 70. This means that an even-numbered property between 48 and 60 would not be matched.
    - The segment ranges fall short of a particular street number. Holmes Street, for example, has one Tiger segment listed, ranging from 59 to 65. Tax99, however, lists three properties on Holmes all with street numbers less than 59.
    - The Tiger segments do not include the correct street side. For example, 11 Newport Street

was unable to be assigned coordinates since the only available Tiger information ranges from 2 to 46.

After utilizing Tiger's resources, we decided to use Google to geocode the places where our Tiger methodology failed. After sending the remaining 2,268 addresses through the Google Geocoder function, we were able to obtain coordinates for all but forty-nine properties. Google was unable to locate properties on:

| | |
|---|---|
| Roosevelt St Ex | (29 properties) |
| Hughes St Ex | (2 properties) |
| Hillis Street | (3 properties) |
| Mill River Front | (2 properties) |

In addition, Google also failed to estimate coordinates for land lots:

| | |
|---|---|
| Division Street | (6 lots) |
| Newhall Street | (3 lots) |
| Winchester Ave | (3 lots) |
| Whalley Ave | (1 lot) |

Combining both methods we were able to assign coordinates to 99.8% of the properties.