Improving Geocoding Accuracy in Network Analysis by Calculating the Altitude of a Triangle Generated by the Parcel Centroid and Road Segment Endpoints

By Scott Harding

There are many network analysis applications that require addresses of parcels to be located along a road segment. Using a GIS (e.g. ESRI's ArcGIS) it is possible to place nodes that represent parcels along a line segment using the address range to estimate the percentage along the line segment (a.k.a linear referencing) it should be placed. However, the address range can be a poor indicator of where an address should be located along a road segment.

A common reason why addresses are misplaced along a line segment using a parcel address is that the range of addresses in an urban setting for the corresponding address is much greater than the range of addresses used by the parcels along the entire length of the road segment (Figure 1.).

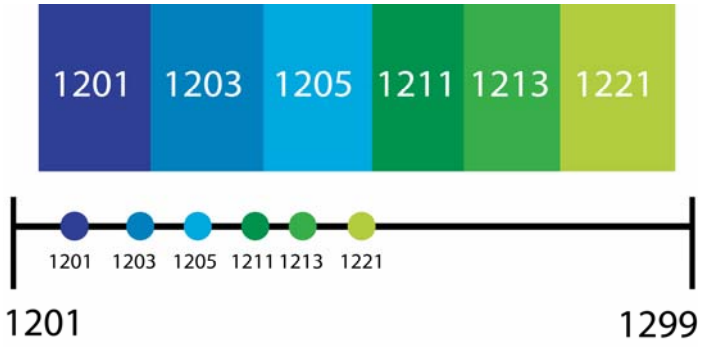

Figure 1. Addresses are geocoded in the wrong location when using address ranges.

One solution to overcome the difference in parcel addresses location and road range of addresses is to match the road segment range to the parcel range (this new address range is only used to determine the linear referenced location). The problem with this method is that parcels with irregular shapes and skipped address can significantly distort the linear referencing process. As Figure 2. shows, parcel 1207 is much wider than the rest of the parcels along the same road segment and 1207 jumps to address 1221. This causes address 1207 to be located farther to the left than it should be.

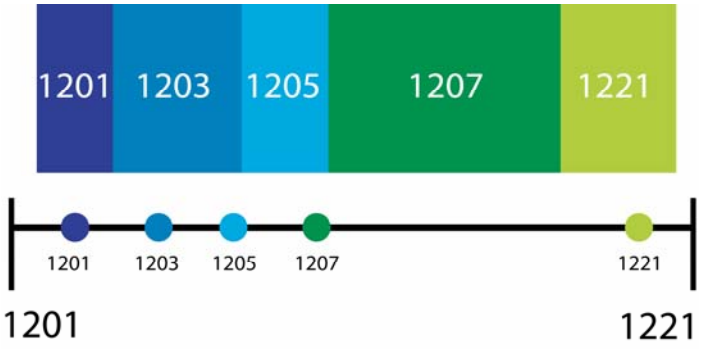

Figure 2.

A better solution to geocoding addresses is to use the location of the parcel, if this data is available, to determine where the geocoded address should be placed along the road segment.

The first step in using the parcel location along the road segment in geocoding addresses is to determine the centroid of the parcel. The centroid coordinates provide the best method to represent the location of the parcel in most circumstances. In some cases, where the shape of the parcel is highly irregular, the centroid coordinates may not fall within the parcel. However, this does not necessarily mean that the centroid is a poor representation of the parcel location along the road segment. In most cases, the importance of geocoding the location of a parcel on to a road segment is to represent where the people, products, and etc. interface between the parcel and the road. Knowing this interface location accurately is critical when running network analyses. One way to determine the position of the interface node more accurately it to observe where there are entrances from the road or sidewalk to the parcel, in which some cases may be more than one location, on more than one road (e.g. a mall with multiple entrances). However, when hundreds or even thousands of addresses are being geocoded, it becomes impractical to locate and represent theses entrances along the road network by hand and a method of automation is paramount.

By using the coordinates of the parcel centroid and the coordinates of the beginning and end of the corresponding road segment it is possible to greatly improve the accuracy of the placement of addresses along a road segment. If we label the X and Y coordinates of the centroid as CX and CY respectively, the start X and Y coordinates of the road segment SX and SY respectively, and the end coordinates of the road segment EX and EY respectively, the distance between the points can be calculated using the distance formula (See Formula 1. and Figure 3.).

Formula 1.

$$\overline{SC} = \sqrt{(CX - SX)^2 + (CY - SY)^2}$$
$$\overline{EC} = \sqrt{(CX - EX)^2 + (CY - EY)^2}$$
$$\overline{SE} = \sqrt{(EX - SX)^2 + (EY - SY)^2}$$

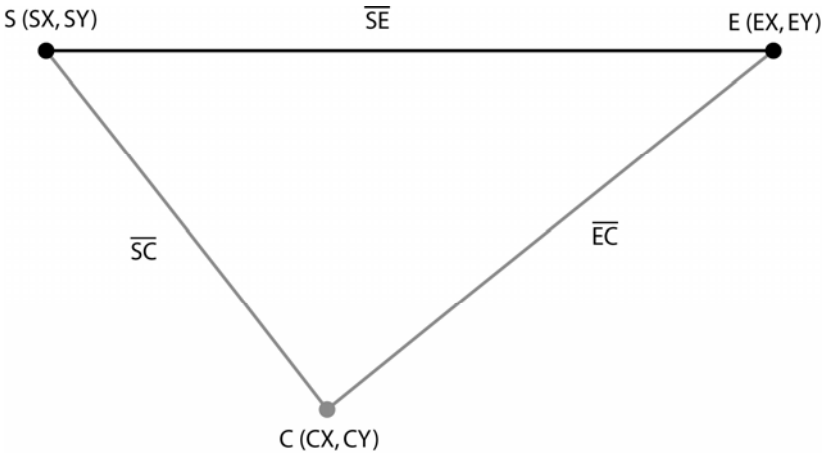

Figure 3.

Using the law of cosines $\angle S$, $\angle E$, and $\angle C$ can be calculated using the distance between the points (Formula 2. and Figure 4.).

Formula 2.

$$\angle S = \cos^{-1}\left[\frac{-\overline{EC}^2 + \overline{SC}^2 + \overline{SE}^2}{2 * \overline{SC} * \overline{SE}}\right]$$

$$\angle E = \cos^{-1}\left[\frac{\overline{EC}^2 - \overline{SC}^2 + \overline{SE}^2}{2 * \overline{EC} * \overline{SE}}\right]$$

$$\angle C = \cos^{-1}\left[\frac{\overline{EC}^2 + \overline{SC}^2 - \overline{SE}^2}{2 * \overline{EC} * \overline{SC}}\right]$$

Note: Knowing $\angle C$ is not required for the calculations.

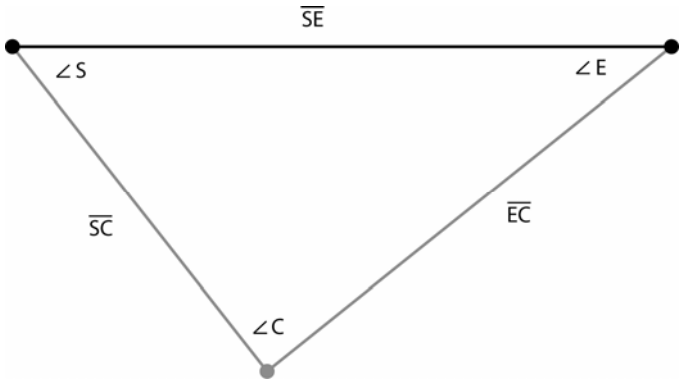

Figure 4.

To project the centroid onto the road segment at the best possible location, the altitude of the triangle is used. The advantage of the altitude is that it is the shortest distance between the road segment and the parcel centroid (assuming the road segment is a straight line) and therefore would most often be the entrance point from the parcel to the road or vise versa (Figure 5.). In addition, the distance from the centroid to the road can be easily calculated and used in a network analysis assuming that the starting point is at the center of the parcel.

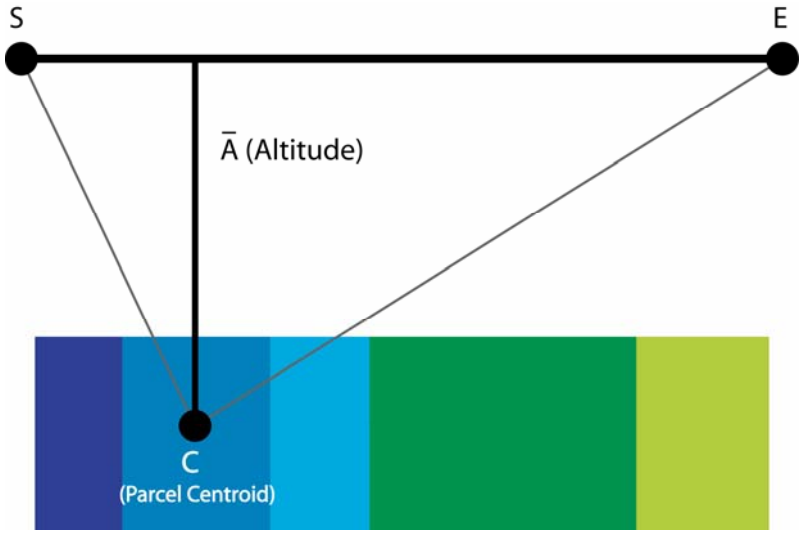

Figure 5.

Determining where the altitude intersects $\overline{SE}$ is calculated by first finding the angle, θ, made by the perpendicular bisector, $\overline{A}$, and line $\overline{SC}$. The $\angle\theta$ is $180^0 - 90 - \angle S$, which can be simplified to $90 - \angle S$ because line $\overline{A}$ will always intersect $\overline{SE}$ at a right angle (See Formula 3. and Figure 7.).

Formula 3.

$$\angle\theta = 90 - \angle S$$

For these properties to hold true, $\angle S$ and $\angle E$ must be less than 90˚. When geocoding, a limited number of parcel centroids will be located in a location where $\angle S$ and $\angle E$ are greater than or equal to 90˚. If $\angle S$ is greater than or equal to 90˚ then the projected point should be placed at or close to point S (Figure 6.). If $\angle E$ is greater than or equal to 90˚ then the projected point should be placed at or close to point E.

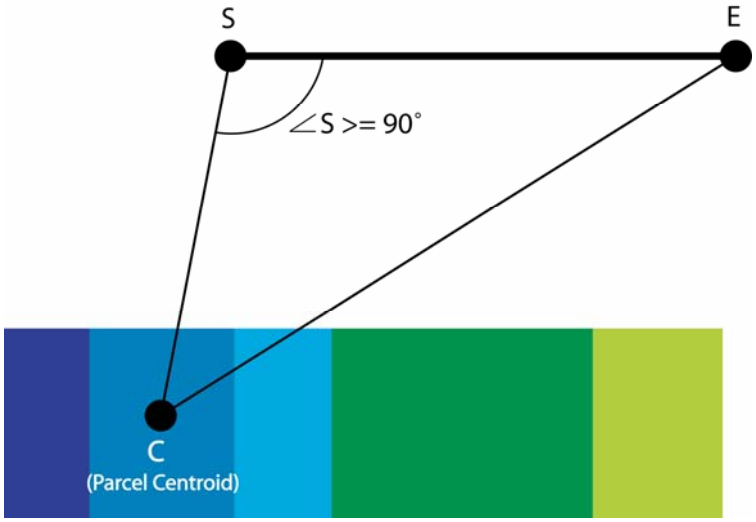

Figure 6.

To find distance along $\overline{SE}$ from the starting point to where $\overline{A}$ intersects $\overline{SE}$, the opposite side, L1, of $\angle\theta$ is computed (See Formula 4. and Figure 7.).

Formula 4.

$$L1 = \sin(\theta) * \overline{SC}$$
$$L2 = \sin(\theta) * \overline{EC}$$

Formula 5.

$$P = \frac{L1}{L1 + L2} or \frac{L1}{\overline{SE}}$$

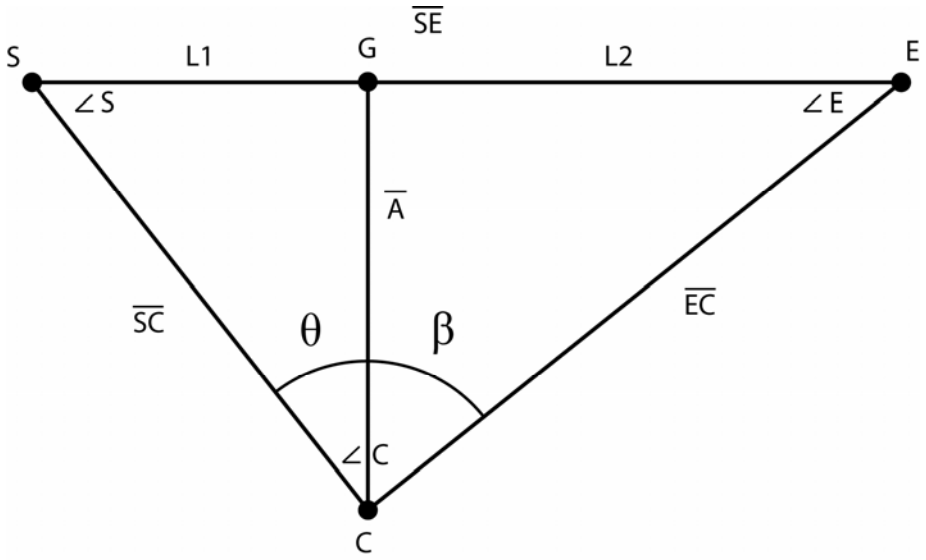

Figure 7.

The final step is to find the percentage, *P*, that G is along the road segment $\overline{SE}$ (See Formula 5. and Figure 7.). The percentage, *P*, can be used to properly geocode an address along a road segment either by using the value directly if the addresses locator allows or by creating a pseudo-address.

Test Case

The accuracy of this method for geocoding addresses was tested with approximately 40,000 parcel addresses and 10,000 roads of an urban city located in the Midwest. ESRI ArcView 9.1 was used to geocode the addresses and Microsoft Excel and Access was used in addition to ArcView to generate the pseudo addresses.

The first step in the analysis was to geocode the addresses with the traditional method of using the standardized US Alphanumeric Ranges. This method located 95% of the addresses along the road network, although it was evident that this method did a poor job in placing the addresses in the correct location in relation to the parcels. As Figure 8. shows, the addresses tend to get clumped towards the start point of the road segment because the range of the addresses of the road network are typically much larger than that of the range of parcel addresses.

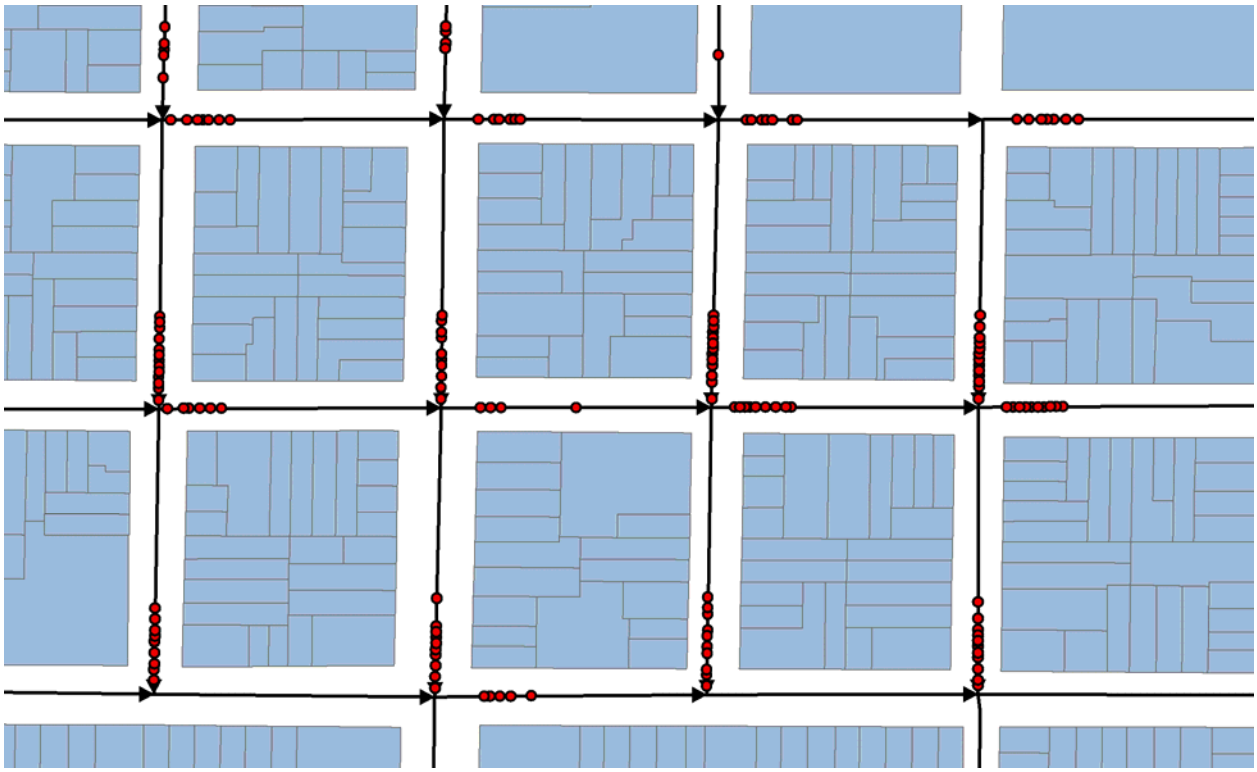

Figure 8.

The next step involved using the geocoded addresses is to observe which road segments the parcels are located. With the appropriate road segments correlated, the coordinates of the centroid for each parcel and the coordinates of the start and end point of the associated road segment were added to the parcel address database file. Using excel the distance of the three sides of the triangle, the angles made by the segments, and the intersection of the altitude were calculated (Figure 3 & 4.). The proportion of L1 of the sum of L1 and L2 produces a percentage that was used to create a pseudo-address by multiplying it by the address range and rounded to the nearest integer (Figure 7.).

With the pseudo-address, the parcels were geocoded again. This time the addresses along the road segments are much closer to the associated parcel.

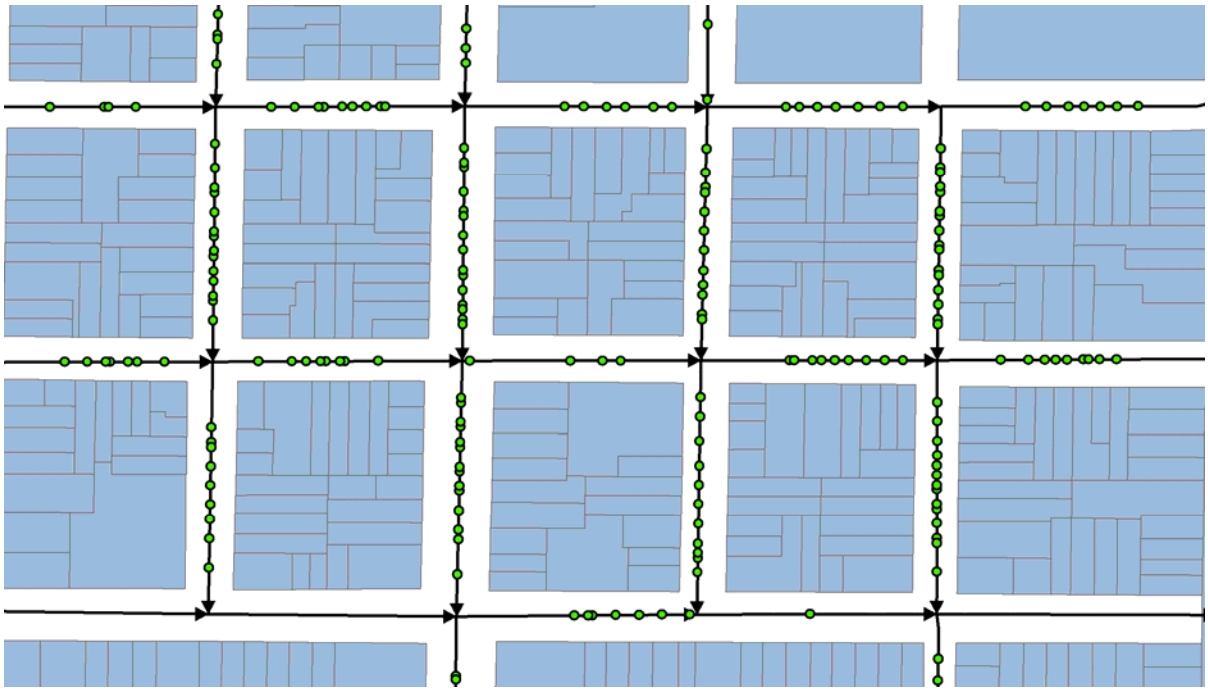

Figure 7.

To determine the degree of improvement, the distance from the parcel centroid to the geocoded addresses between the two methods was compared. Of the 39,965 geocoded parcels, 92.4% of the addresses using the new method were closer to the parcel centroid with a standard deviation of 362.27 ft. 5.1% of the addresses were farther from the parcel centroid using the new method with a standard deviation of 1191.11 ft and 2.5% addresses showed no difference (Table 1.). Using the parcel centroids to calculate a pseudo-address value significantly increases the accuracy of the geocoding process.

Table 1. The difference of distance between traditional and new method.

Parcel to Road Node Distance	Number of Addresses	Percent of Total	Mean	St. Dev.
Decreased Distance	36923	92.4	162.99	362.27
Increase Distance	2044	5.1	-116.89	1191.11
No Difference	998	2.5	N/A	N/A

The two main reasons why some addresses are farther away from the parcel centroid when using the new method is that some of the data have the wrong and reversed address ranges and the road line segments are always represented as straight lines. When calculating the distance between the start and end point of the line segment, the distance is calculated as a straight line. When the road significantly deviates from a straight line, the length of the road may be much longer in length and therefore the pseudo-address will place the addresses either too far along the road segment or

not far enough. The traditional method of geocoding addresses produces the same type of error. In addition, the occurrences where the traditional method places the geocoded address closer to the parcel occur merely by chance. One way to adjust for the irregular shape of road segments is to break them up into shorter pieces. However, doing so can be a complicated and a time consuming process. Another option is to calculate where the perpendicular bisector intersects the line segment and uses the actual length of the road segment in the calculations. To do this, the equation of line has to be created to represent the bisector and the intersection solved mathematically. This process is complicated and may produce undesired results when the road passes in front of a parcel at an odd angle.

Since both the calculation of the distance from the parcel centroid and geocoded addresses using the traditional and new method is easily computed, both methods can be combined to create better results. Additionally, the calculated distance can be added to the network to represent the cost of travel from the centroid to the road network. It is important to note, however, that the shorter distance form a parcel centroid to a geocoded address on a road segment does not always equate to a more accurately located address.

The method presented is an efficient and easily implemented solution to geocoding addresses when parcel information is available. It can be combined with other methods to create an accurate address road network.