Road network structure and speeding using GPS data

Toshihiro Yokoo
Graduate Student
University of Minnesota, Twin Cities.
Department of Civil, Environmental, and Geo- Engineering
500 Pillsbury Drive SE
Minneapolis, MN 55455 USA
Email: yokoo002@umn.edu

David Levinson
RP Braun-CTS Chair of Transportation Engineering
Director of Network, Economics, and Urban Systems Research Group
University of Minnesota
Department of Civil, Environmental, and Geo- Engineering
500 Pillsbury Drive SE
Minneapolis, MN 55455 USA
Tel: 612-625-6354; Fax: 612-626-7750; Email: dlevinson@umn.edu
http://nexus.umn.edu

Word count: 3,175 words text + (12 figures + 4 tables) x 250 words (each) = 7,175 words
July 25, 2015

ABSTRACT

This paper analyzes the relationship between road network structure and the percentage of speeding using GPS data collected from 152 individuals over a 7 day period. To investigate the relationship, we develop an algorithm and process to match the GPS data and GIS data accurately. Comparing actual travel speed from GPS data with posted speed limits we measure where and when speeding occurs, by whom. We posit that road network structure shapes the decision to speed. Our result shows that the percentage of speeding, which is calculated by travel distance, is large in high speed limit zones (e.g. 60 mph) and low speed limit zone (less than 25 mph); in contrast, the percentage of speeding is much lower in the 30 - 50 mph zone. The results suggest driving pattern depends on the road type. We also find that if there are many intersections in the road, average link speed (and speeding) drops. Long links are conducive to speeding.

INTRODUCTION

Driving speed and speed variance are usually found to increase crash risk (1). Traffic crashes are due to driver "manipulation error" (e.g., speeding and illegal overtaking) and "perception error" (e.g., wrong assessment of speed and distance). As the speed increases, the percentage of both errors increases (2), and Stanton and Salmon (*3*) illustrate that "Road infrastructure", "Vehicle", "Road user", "Other road users", and "Environmental conditions" are factors that affect to driver's error, and therefore, 45 % to 75 % of all roadway crashes are related to driver's error (4). In contrast Moore (*5*) found a decline in highway death rates after the speed limits were raised.

It seems that speeding is generally related to human factors; however, we posit that road network structure shapes the decision to speed. For example, Lai and Carsten (*6*) explain that the percentage of speeding depends on the speed limit zone. We test that hypothesis in this paper.

MATERIAL

This paper uses GPS data from the 2010 Twin Cities (Minneapolis - St. Paul area) Travel Behavior Inventory (TBI2010) administered by the Metropolitan Council between 2010 and 2012 (most data was collected in 2011). Each subject in the survey carried a GPS pendant for 7 days. The raw GPS data contains the following trip information: Speed (km/h), Longitude, Latitude, Altitude (meters), Date (year/month/day), Time (hour/minute/second), Distance (meters), Course (degree), Number of satellites, HDOP. Daily movement of each person is recorded in the data and the information of the trajectories was recorded every second.

However, this GPS data by itself does not contain personal information (e.g. gender, age). Therefore, we use the associated records from the TBI2010 Household Interview Survey (7) as complementary data (e.g., Household data, Personal information, Trip data). These data can be matched on TripID. Among 274 GPS subjects (drawn from 250 households), 152 travelers have trip ID matching the survey and no other data problems, allowing us to analyze 152 participants. We append the GPS data with personal information describing each subject's gender, age, and education.

Each traveler made many trips, and not all trips were by automobile, therefore, we need to extract driving trip data from trip data by the algorithm.

With the purpose of understanding actual speed limit data, we used a GIS map maintained by the Metropolitan Council and The Lawrence Group (TLG) that covers majority of routes in the Twin Cities seven county metropolitan area. This GIS street map has the most accuracy at the present time. The map contains 290,231 links, and each link has several attributes (e.g. speed limit, length of link, street name, one-way).

To conduct the analysis, we use QGIS (8), an open source geographic information system.

METHODOLOGY

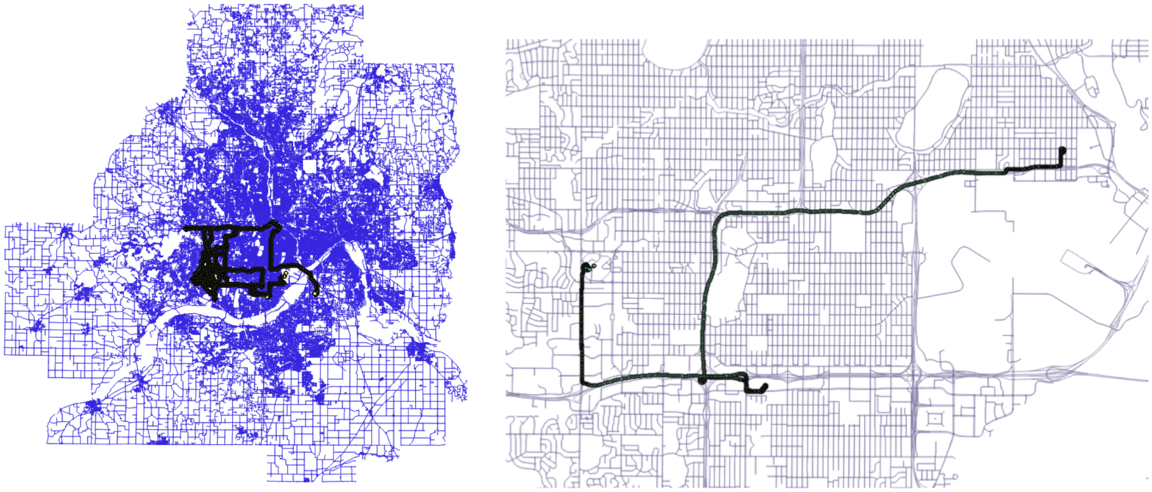
Trip Definitions

Trip generation

As Figure 1(a) shows, GPS travel records initially contains data from multiple trips.

First, the algorithm divides the travel time into the trip data when there is a time difference of GPS data is more than 300 seconds.

Figure 1(b) illustrates the GPS data after it is decomposed into individual trips.



(a) Travel data (all trips) (b) Trip data

FIGURE 1 Example of GPS data

Study area

In our research, the TLG map covers the seven county metropolitan area in Minnesota; however, some GPS data are located this map. We use the following boundary to define the study area:

$$-94.0123 <= Longitude <= -92.7397$$

$$44.4714 <= Latitude <= 45.4139$$

Mode selection

Moreover, GPS data contains not only driving data but also walking, biking, and transit data (typically lower speed data). The purpose of our research is to focus on driving speeding behavior, so the algorithm excludes the walking data. The condition to remove the data is as follows;

1. Speed data is less than 5 km/h or
2. Trip average speed is less than 10 km/h

Algorithm

This section describes the map matching algorithm used to connect the GPS data to the GIS map, and issues and errors that need to be considered.

Convert coordinate system

In order to match the GPS data and GIS data, the coordinate system of these two data should be the same; however, GPS data is geographic latitude and longitude; while the TLGMap uses the Universal Transverse Mercator (UTM). To convert UTM data into longitude and latitude data, we use the formula of Karney (*9*) and Kawase (*10*).

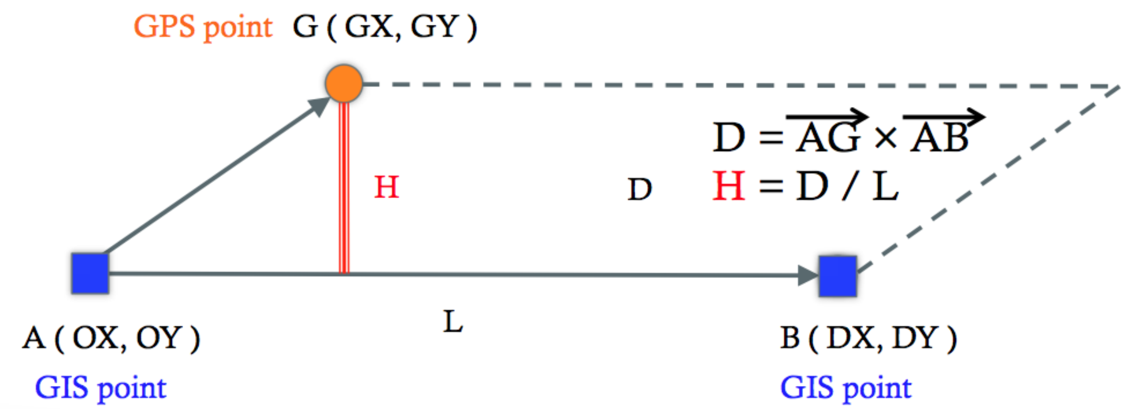


FIGURE 2 Method of calculating nearest GIS link

Match GPS and GIS data

Because of measurement error and map resolution, GPS location is not the exact location of the GIS link. Therefore, several map-matching methods have been developed to identify the correct GIS road link (11) (12). Firstly we assume that closest length between GPS and GIS data indicates the accurate map matching. To begin, we match the GPS data and GIS data when the length between GPS and GIS link is smallest. In order to estimate the length, the algorithm calculates the area of parallelogram (D) by vector outer product operations, then calculates the height of parallelogram (H) by dividing the area (D) by the length of the base (L), as shown in Figure 2.

When the height is smallest, the algorithm extracts the road information (e.g., link length, street name, roadtype, and speed limit) from GIS data, and combines this data into GPS data.

Curve section

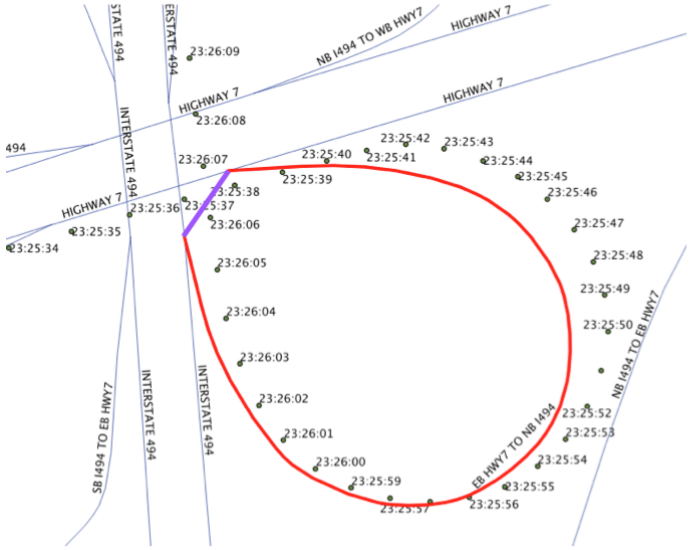
The GIS link between two points is not necessarily a straight line like Figure 2; some GPS link are curved as Figure 3(a). When the algorithm above calculates the shortest length between GPS data and GIS link in curve section, GPS data is likely to choose different GIS link. This failure significantly occurs in ramp section of interstate or freeway.

Intersection

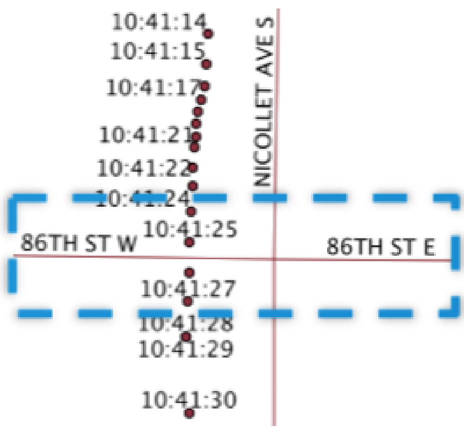
When the GPS point is near the intersection area, the matching failure is likely to occur. As Figure 3(b) shows, the vehicle moves on the vertical road (NICOLLET AVE S); however, some GPS points choose horizontal road (86TH ST W) as driving route in the algorithm.

Missing link

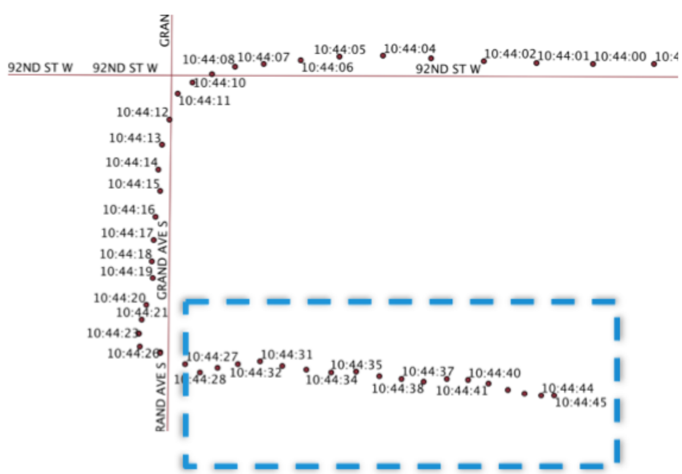
Although the TLG network data is the most accurate digital map in the Metropolitan area, some road information is missing. When vehicles drive on these missing roads, it is difficult to match the data correctly (Figure 3(c)).



(a) Curve



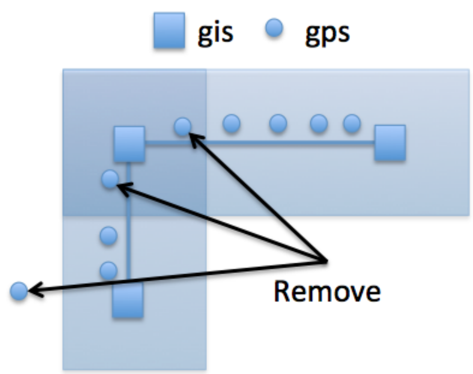
(b) Intersection



(c) Missing link

(a) Red line: Real GIS link, Purple line: GIS link of algorithm, Green dot: GPS data

FIGURE 3 Example of error




```
if (ox_gis - 10<= x_gps <= dx_gis + 10):
    if (oy_gis - 10) <= y_gps <= (dy_gis + 10):
        count += 1
    elif (dy_gis - 10) <= y_gps <= (oy_gis + 10):
        count += 1
elif (dx_gis - 10 <= x_gps <= ox_gis + 10):
    if (oy_gis - 10) <= y_gps <= (dy_gis + 10):
        count += 1
    elif (dy_gis - 10) <= y_gps <= (oy_gis + 10):
        count += 1

if count == 1:
    print >> outputFile, ','.join(row)
```

FIGURE 4 Image and explanation of extraction algorithm

Process of improving accuracy

Based on the result of the mapmatching algorithm, simply matching to the closest point between GPS and GIS data is far from perfect, and many errors are found at Intersections and due to Missing link. To control for that, when GPS data is near an intersection area or far from a mapped link, this GPS point is removed. This process is useful for removing the error, but wastes data.

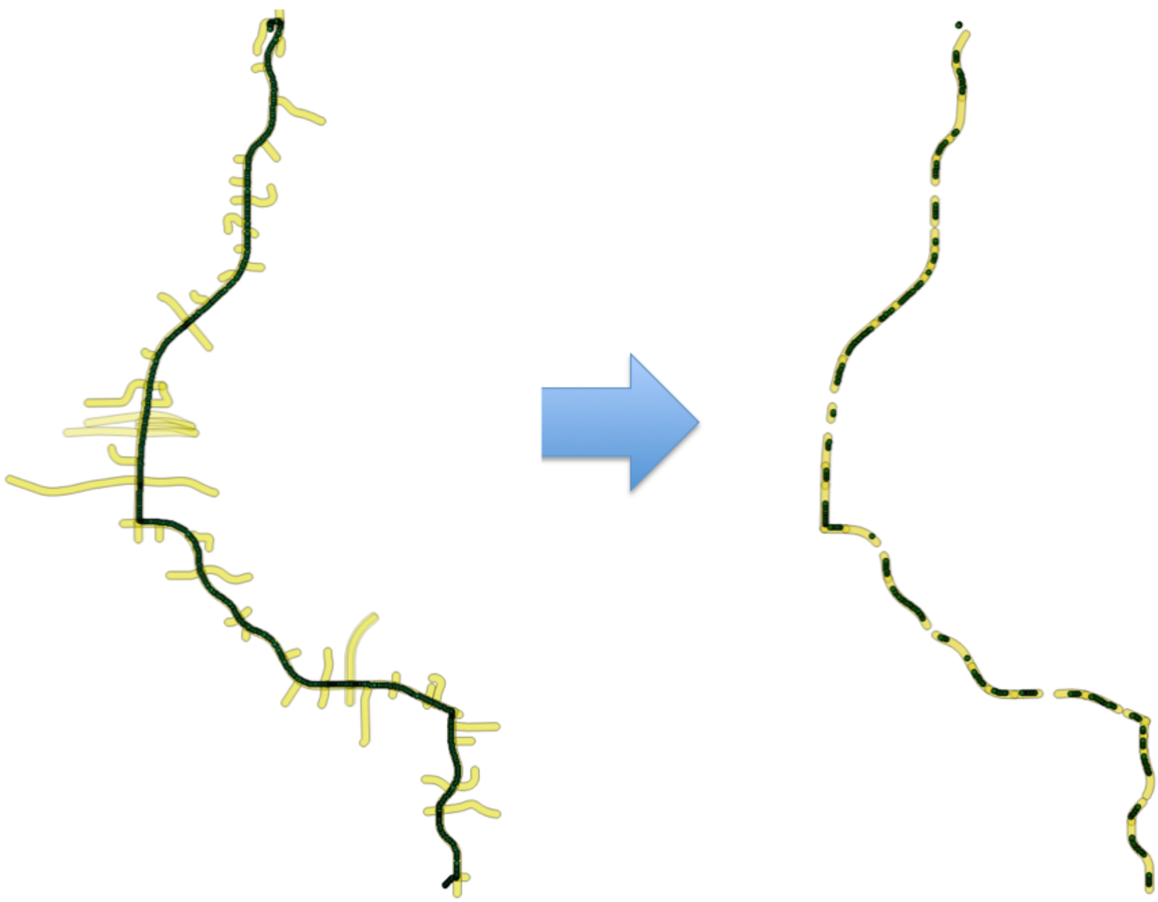
Extract GIS link

Firstly, we create a buffer of GIS link by using QGIS. We used 10m buffer distance because most of the GPS data are located within 10m of the centerline of road (GIS link). After matching the coordinate system of GPS and GIS data, we extract GIS links that intersect 10m buffer area and GPS data. As Figure 5 (left) shows, this process can extract the link. But this is not perfect because extracted link contains intersected link, so we need to remove these links.

Remove intersection area

In order to remove unnecessary buffer area, we need to remove GPS data which is near the intersection area or far from GIS link. Figure 4 is the image and explanation of the algorithm. Each GIS link creates the layer of range area (10m), so intersected GIS link area has more than two layers. The algorithm counts the number when GPS data is inside the range area. When the GPS point is near the intersection area, the number of count becomes more than two. Also, when GPS data is outside of the range of GIS data, the number of count is zero. The algorithm only saves the GPS data that count of number is one (Figure 4). After completing this process, we extract the GIS link again based on the same process of *Extract GIS link* by using selected GPS data. Figure 5 (right) is the result of this process, and the figure shows this method can remove intersected link and save actual driving route link.

The problem of this process is that most of GPS data is removed when several road links are close to each other. For example, there is a parallel road near the interstate, and therefore, GPS data that drive on interstate is likely to be removed.



Left: Before, Right: After (Yellow area: GIS buffer, Dot: GPS data)

FIGURE 5 Result of algorithm and process

Result of accuracy

After finishing the process, we matched the extracted GPS data and GIS link. Also, we analyzed the accuracy from the result of map-matching one at a time by checking the data visually on QGIS. Visually inspection shows that 97.4% of matched points on 20 trips are accurate (Table 1), and therefore this result is quite satisfactory. The problem is that about 51.0% of the GPS data are removed and some driving information during the trip is missing due to this process. Therefore, we cannot analyze continuous trip data. In order to ensure few false positives, there are many false negatives.

After this algorithm, 152 GPS travel data create 2,891 trip data (Total GPS point: 1,223,753).

TABLE 1 Accuracy of matching data

Trip	# of accuracy	# of total data	Accuracy (%)		# of initial trip data	% of removal
Trip1	7,103	7,283	97.5%		12,144	40.0%
Trip2	3,502	3,544	98.8%		5,786	38.7%
Trip3	2,945	3,048	96.6%		16,152	81.1%
Trip4	2,980	3,063	97.3%		6,056	49.4%
Trip5	2,645	2,655	99.6%		3,857	31.2%
Trip6	2,502	2,540	98.5%		5,814	56.3%
Trip7	2,496	2,536	98.4%		4,688	45.9%
Trip8	2,442	2,473	98.7%		4,666	47.0%
Trip9	2,337	2,439	95.8%		5,152	52.7%
Trip10	2,300	2,351	97.8%		4,783	50.8%
Trip11	2,280	2,390	95.4%		3,739	36.1%
Trip12	2,325	2,344	99.2%		5,637	58.4%
Trip13	2,176	2,311	94.2%		4,189	44.8%
Trip14	2,003	2,236	89.6%		3,722	39.9%
Trip15	2,138	2,171	98.5%		3,885	44.1%
Trip16	2,103	2,160	97.4%		3,586	39.8%
Trip17	2,063	2,117	97.4%		4,032	47.5%
Trip18	2,096	2,106	99.5%		5,468	61.5%
Trip19	2,044	2,100	97.3%		3,051	31.2%
Trip20	2,115	2,136	99.0%		3,730	42.7%
Total	52,595	54,003	97.4%		110,137	51.0%

HYPOTHESES

We hypothesize that speeding is affected by road type (hierarchy of the network) and road characteristics. Therefore, we analyze the relationship between two road network variables and the degree of speeding by matching GPS data and GIS map:

1. Hierarchy (Road type) – *Hypothesis: High hierarchy is correlated with speeding*

   We analyze the effect of hierarchy in the network. Compared with local road, the driver is less affected by external factors on freeways. GIS data contains the information of speed limit, street name and road type, and therefore we can calculate the degree of speeding depending on each road type.

2. Link length – *Hypothesis: Long link length is correlated with speeding*

   We investigate the effect of the link length on the speeding. When the length is small, there are many intersections in the network. GIS network has link length data, so it is possible to calculate the relationship between link length and speeding behavior.

ANALYSIS

In this analysis, we computed 152 travel data, 2,891 trip data, and analyzed the relationship between road network variable and speeding. The total amount of data are 1,223,753 GPS points and the date and route is different depends on the trip. As Table 2 illustrates, most of the data are concentrated on 30, 35, and 40 mph zone. Moreover, relatively few points are in low speed limit zone (less than 20 mph) and high speed limit zone (70 mph).

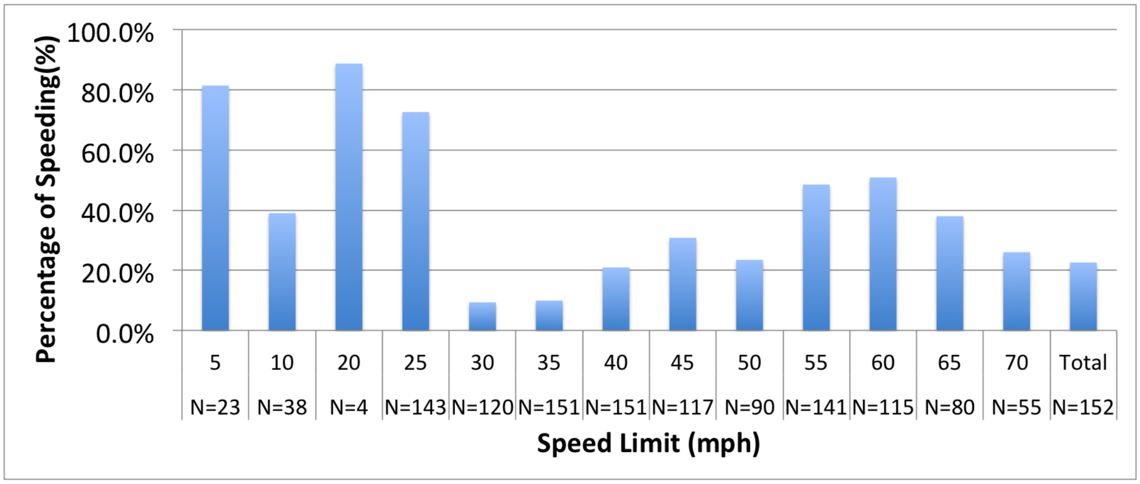
TABLE 2 Total data in each speed limit zone

	GPS data		GIS link	
Speed limit (mph)	**# of data**	**percentage**	**# of link**	**percentage**
5	792	0.1%	181	0.1%
10	2,612	0.2%	354	0.1%
20	80	0.0%	227	0.1%
25	21,758	1.8%	4,245	1.5%
30	152,642	12.5%	24,204	8.3%
35	394,777	32.3%	223,198	76.9%
40	356,521	29.1%	29,209	10.1%
45	23,224	1.9%	1,921	0.7%
50	20,030	1.6%	939	0.3%
55	122,146	10.0%	3,523	1.2%
60	91,181	7.5%	836	0.3%
65	28,497	2.3%	768	0.3%
70	9,493	0.8%	214	0.1%
Total	1,223,753	100.0%	290,231	100.0%

Road network structure

Hierarchy (Road type)

The bar chart shows the percentage of speeding on each speed limit zone (Figure 6). Overall, 22.6 percent of GPS driving data exceeded the speed limit. It can also be seen that speeding behavior was significant in low speed limit zone (e.g. from 5 to 25 mph zone) and high speed limit zone (e.g. from 55 to 65 mph zone). It suggests that most participants encountered speed limits between 25 and 65 mph while driving.



(N = *) indicates number of individuals persons encountering that particular speed limit

FIGURE 6 Percentage of speeding across speed limit zone

Link length

Figure 7 illustrates the relationship between link length and speeding behavior. Y axis is *Speeding* defined in Equation 1.

$$Speeding = \frac{DrivingSpeed}{SpeedLimit} \tag{1}$$

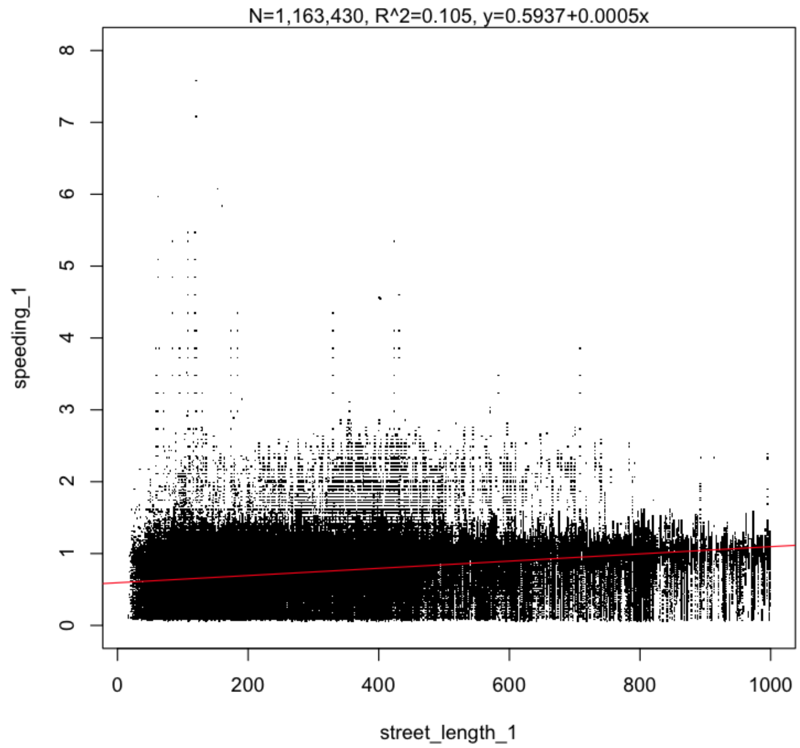
When the value is 1, driving speed is equals the speed limit, and when the value is more than 1, the data shows speeding. To reduce data issues, we only analyzed the data for link length <= 1km.

Figure 7 and Table 3 shows that when the link length is longer, the driver is likely to exceed the speed limit. The polynomial model (Equation 3) fit the data better.

$$E(speeding|streetlength) = \beta_0 + \beta_1 streetlength \tag{2}$$

$$E(speeding|streetlength) = \beta_0 + \beta_1 streetlength + \beta_2 streetlength^2 \tag{3}$$

The result of Figure 7 is from all road network , but it is considered that driving condition is significantly different depending on road type, therefore, we analyze this relationship according to each speed limit zone. Figure 8 illustrates that the relationship between link length and speeding is different depending on the speed limit zone. Interesting finding is that the graph shows there is a little correlativity between link length and speeding at 30mph zone. On the other hand, there is strong positive correlation at 25 mph and 40mph zone.



(link length is less than 1,000m)

FIGURE 7 Relationship between link length and speeding

TABLE 3 Regression Results

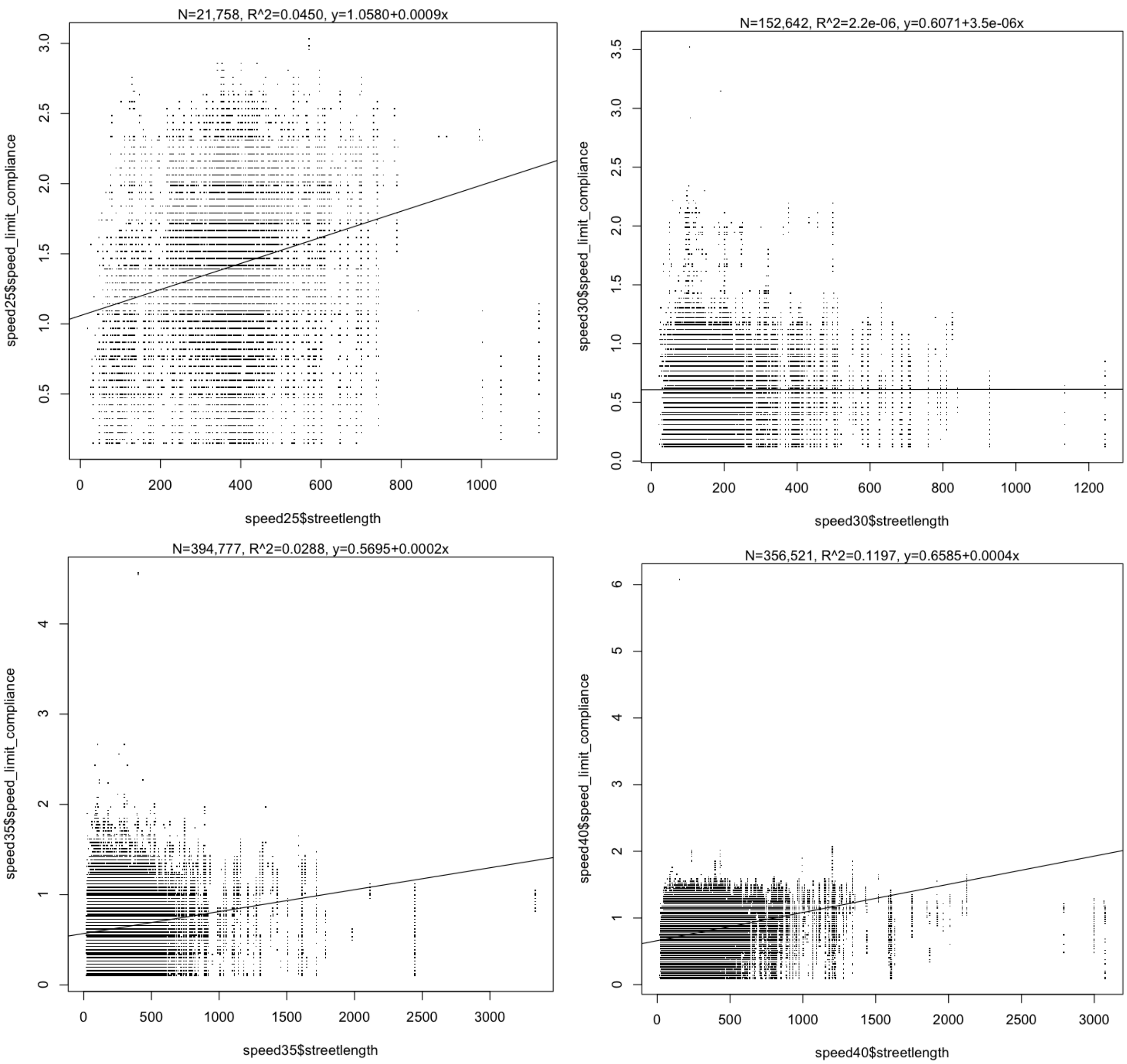
Dependent variable : Speeding (link length is less than 1,000m)

	Equation 2			**Equation 3**		
	Estimate	**t-value**		**Estimate**	**t-value**	
Intercept	5.937e-01	1183.9	***	4.995e-01	585.5	***
street_length	5.020e-04	369.4	***	1.164e-03	230.3	***
street_length^2	na	na		-7.619e-07	-135.9	***

—

signif. Codes: 0 '***' 0.001,'**',0.01 '*' 0.05 '.' 0.1 ' ' 1

	Equation 2	**Equation 3**
Multiple R-squared	0.105	0.119
F-statistic	1.365e+05	7.856e+04
p-value	<2.2e-16	<2.2e-16



(Top left: 25mph, Top right: 30mph, Bottom left: 35mph, Bottom right: 40mph)

FIGURE 8 Speeding by Link Length by Speed Limit

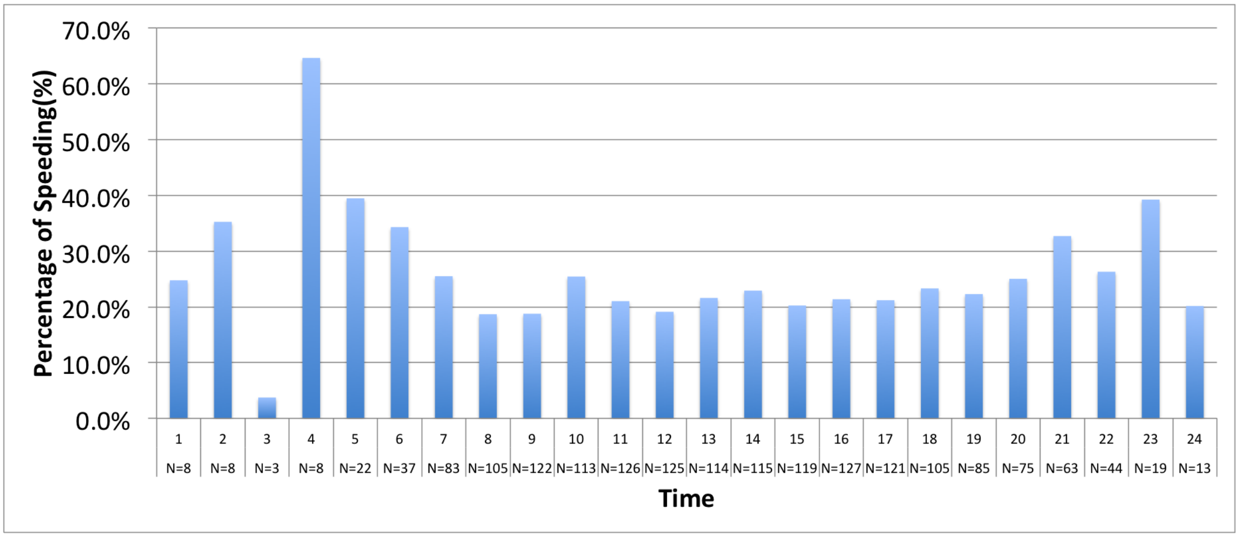


FIGURE 9 Percentage of speeding by time of day

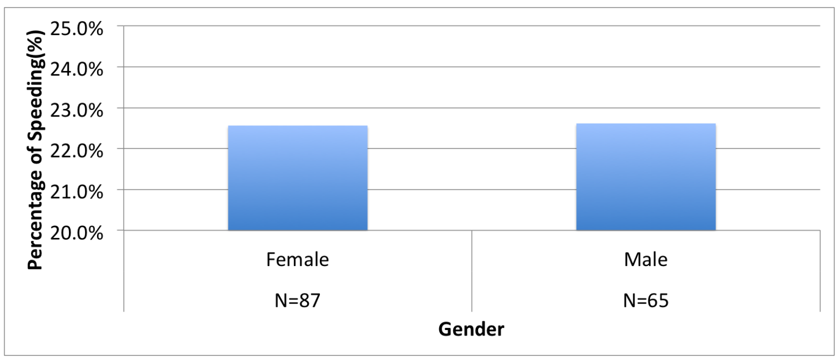


FIGURE 10 Speeding by gender

Time of the day

The bar chart illustrates the proportion of exceeding the speed limit across time of the day (Figure 9). Overall, speeding behavior varies by time of day and the percentage of speeding during daytime driving is lower than nighttime driving, with a peak at 4 am.

Personal information

Several researchers mention that speeding behavior depends on the driver (2) (6). Unlike the data of loop detector, GPS data is linked to personal information. Therefore, we choose gender, age, and education as personal data to analyze the data of 152 participants.

Figure 10 shows speeding is undifferentiated by gender . While the number of participants was large (male driver: 65, female driver: 87), the percentage of speeding for male driver is same as female driver (Male, Female: 22.6%).

Figure 11 presents the result of the percentage of speeding across age. The age group who are from 25 to 34 illustrated the highest speeding behavior at 32.7%.

Figure 12 illustrates the result of education, and education level is divided into 2 groups; low education (below high school graduate) [low_educ] and high education (above college) [high_educ].

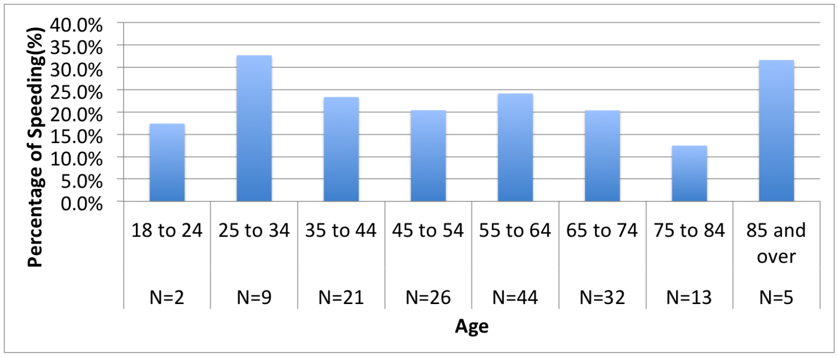


FIGURE 11 Speeding by age

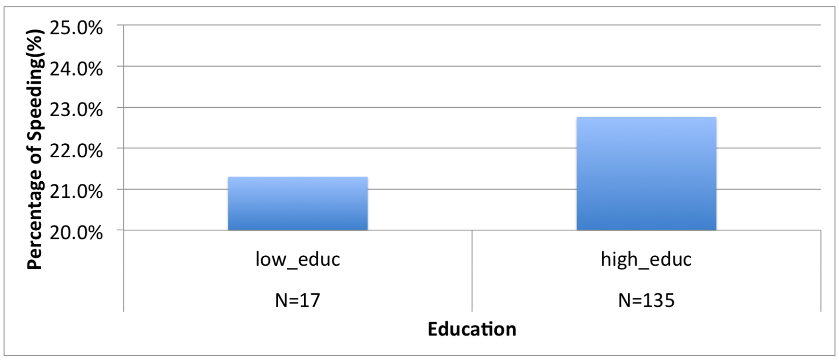


FIGURE 12 Speeding by education level

The percentage of high educated driver (22.8%) was slightly larger than low educated driver (21.3%), therefore high educated driver tends to exceed the speed limit.

Statistical result

Next, we analyzed the relationship between several variables (road network structure and personal information) and speeding statistically.

During the statistical analysis, we removed some data because of error record, such as 'driver's age is less than 16' or 'no answer' in the data.

Table 4 showing regression results is calculated using the statistical package *R*. While t-value of [morning] is small, t-value of other variables are large, therefore they are statistically significant. Overall F-test shows that F value is 1.664e+04 and p-value is less than 2.2e-16. Therefore, this model shows high significance. The findings from earlier are corroborated in this multivariate analysis.

Coefficient value indicates that when the street length is long, speeding is likely to occur. Also, speed limit zone at 30 and 35 mph zone are negatively associated with speeding. Persons who are 35 to 44, 75 to 84, male, drive during the daytime, and/or are educated speed less. Overall the R^2 is 0.23

TABLE 4 Regression Results
Dependent variable : Speeding

Coefficients:

	Estimate	**t-value**	
Intercept	7.093e-01	159.540	***
street length	1.662e-04	174.138	***
speed limit 25	6.291e-01	170.941	***
speed limit 30	-1.184e-01	-35.550	***
speed limit 35	-1.124e-01	-34.510	***
speed limit 40	2.930e-02	9.022	***
speed limit 45	5.625e-02	15.142	***
speed limit 50	5.104e-02	13.509	***
speed limit 55	1.161e-01	35.913	***
speed limit 60	1.158e-01	35.839	***
speed limit 65	7.707e-02	21.896	***
age 25 - 34	4.081e-02	17.169	***
age 35 - 44	-5.793e-03	-2.601	**
age 45 - 54	2.017e-02	9.097	***
age 55 - 64	2.870e-02	13.059	***
age 65 - 74	2.934e-02	13.245	***
age 75 - 84	-2.210e-02	-9.389	***
age 85+	7.506e-02	26.541	***
male	-1.546e-02	-27.639	***
educated	-1.961e-02	-21.955	***
morning	-3.865e-04	-0.172	
afternoon	-1.461e-02	-6.530	***
night	1.660e-02	7.237	***

—
signif. Codes: 0 '***' 0.001,'**',0.01 '*' 0.05 '.' 0.1 ' ' 1

Multiple R^2	0.2303
F-statistic	1.664e+04
p-value	<2.2e-16

DISCUSSION AND CONCLUSION

This paper investigates GPS-based speed data from 152 participants in Minneapolis to examine the relationship between road network structure and speeding. The most pertinent findings from the results are that speeding behavior was significant in low and high speed limit zones, and long link length is correlated with speeding. Moreover, our algorithm and process shows sufficient accuracy of map-matching although half of GPS data are removed due to the process. Unlike the speed data of loop detector, GPS-based analysis enables to investigate the speed information over a broad area. Multi-variate regression analysis finds persons who are 35 to 44, 75 to 84, male, drive during the daytime, and/or are educated speed less than other age groups, female drivers, nighttime drivers, and/or less educated drivers. Future research will aim to reduce the amount of excluded data (false negatives).

REFERENCES

[1] Aarts, L. and I. Van Schagen, Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, Vol. 38, No. 2, 2006, pp. 215–224.

[2] Kanellaidis, G., J. Golias, and K. Zarifopoulos, A survey of drivers' attitudes toward speed limit violations. *Journal of safety Research*, Vol. 26, No. 1, 1995, pp. 31–40.

[3] Stanton, N. A. and P. M. Salmon, Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, Vol. 47, No. 2, 2009, pp. 227–237.

[4] Medina, A. L., S. E. Lee, W. W. Wierwille, and R. J. Hanowski, Relationship between infrastructure, driver error, and critical incidents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications, 2004, Vol. 48, pp. 2075–2079.

[5] Moore, S., *Speed doesn't kill: The repeal of the 55-mph speed limit*. Cato Institute, 1999.

[6] Lai, F. and O. Carsten, What benefit does Intelligent Speed Adaptation deliver: A close examination of its effect on vehicle speeds. *Accident Analysis & Prevention*, Vol. 48, 2012, pp. 4–9.

[7] Council, M., *TBI 2010 Household Interview Survey Data*. http://datafinder.org/metadata/TravelBehaviorInventory2010HomeInterviewSurvey.html, Accessed May. 10, 2015.

[8] Sutton, T. and O. Dassau, *QGIS*. http://www.qgis.org/en/site/index.html, Accessed March. 30, 2015.

[9] Karney, C. F., Transverse Mercator with an accuracy of a few nanometers. *Journal of Geodesy*, Vol. 85, No. 8, 2011, pp. 475–485.

[10] Kawase, K., Concise derivation of extensive coordinate conversion formulae in the Gauss-Krüger projection. *Bulletin of the Geospatial Information Authority of Japan*, Vol. 60, 2012, pp. 1–6.

[11] Ochieng, W. Y., M. Quddus, and R. B. Noland, Map-matching in complex urban road networks. *Revista Brasileira de Cartografia*, Vol. 2, No. 55, 2003.

[12] Quddus, M. A., W. Y. Ochieng, and R. B. Noland, Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, Vol. 15, No. 5, 2007, pp. 312–328.