TRAVEL PATTERN AND VARIABILITY PREDICTION

Ruohan Li
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
lruohan_322@yahoo.com

Kara M. Kockelman
(Corresponding Author)
E.P. Schoch Professor of Engineering
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
kkockelm@mail.utexas.edu
Phone: 512-471-0210 & FAX: 512-475-8744

Under review for publication in the *Journal of the Transportation Research Forum*
January 2017

ABSTRACT

It is often important to know the travel demand of an area, which can be represented by the annual vehicle-miles travelled (VMT) of the population in vehicles belonging to this area. However, VMT can be very costly and difficult to track. Surveys are one of the most effective tools ingathering travel demand data, by interviewing survey respondents about their travel mileage across one or two consecutive days which can ultimately be used to model annual VMT. In this study, using the data from Puget Sound Regional Council (PSRC), a regression of annual VMT over both daily and 2-day VMTs ran for 20 times with randomly selected dates, resulting in an average R-squared value of 0.1928 for annual vs. daily and 0.2233 for annual vs. 2-day. Demographic variables include household income, age of the head of the household, number of children, number of drivers per vehicle, as well as the month and week day of the selected date. In order to keep track of the variation in travel among individual days across the year, the Gini coefficient of each vehicle's travel pattern is also determined. The 214 vehicles have a mean Gini coefficient of 0.2465. However, the adjusted R-square value for this regression turns out to be 0.1242, indicating that it's not an easy task to predict the Gini coefficient of a vehicle from variables such as annual VMT, household income, age of head of household, number of children, and number of drivers per vehicle.

INTRODUCTION AND MOTIVATION

Vehicle-miles traveled (VMT) is a metric that can be used to represent travel demand (Cervero et al., 2002). In order to find the VMT for a population of vehicles, single day surveys are performed in which each household participating completes a trip diary capturing all trips undertaken during the 24-hour survey period. Surveys, are still the most prevalent among the possible methods used to calculate VMT, due to the high drop-off rate and respondent fatigue in multi-day surveys (Stopher et al., 2008). People's travel patterns, however, can vary considerably over time (Pendyala and Pas, 2000). There can be days of extremely heavy travel as well as days on which no travel takes place at all. Compared to obtaining only one day of trip data, surveying for two days has the advantage of better capturing this variation. As a result, there is an increasing use of multi-day surveys of travel behavior usually two to three days in length, aiming at capturing more variance than one day can provide (Axhausen et al., 2000). Thus, when evaluating how descriptive of the annual VMT the short term travel records are in this paper, the primary focus are on single day and 2-day values.

The variability, then, can be expressed by another term called Gini coefficient, initially derived from economics, typically for income distribution. It is the area between the line of equality and the Lorenz curve over the area of the triangle formed by the x-axis and the line of equality. The Lorenz curve is a curve of the cumulative percentage of total travel plotted over the domain of 0 to 1. The line of equality is the Lorenz curve when the distribution is completely even across the population. The value of the Gini coefficient varies between 0 and 1, and the smaller its value, the less variation, and thus the more stable and predictable the travel pattern. A regression of the Gini coefficient over annual VMT and a series of demographic variables is then run, to see how they are correlated.

DATA SET

The data came from the Puget Sound Regional Council (PSRC) when it conducted the Traffic Choices Study by placing GPS tolling meters in the vehicles of volunteer households. The final data set contains 329 unique households and 484 vehicles. In order to get rid of the correlation of travel among different vehicles in the same household, one vehicle per household was used, and after removing certain households due to a low tracking period or missing demographic information, the study was carried out with 214 vehicles, each from a different household.

Figure 1. Histogram for Daily VMT

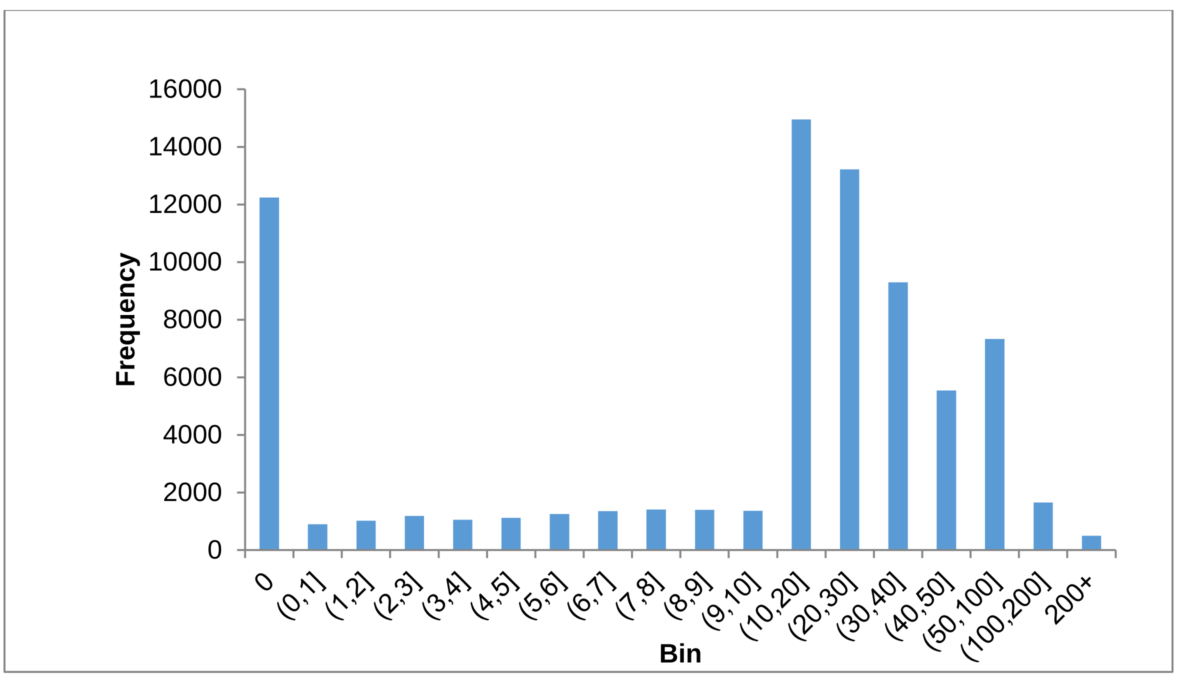


Figure 1 shows the histogram for daily VMTs of all the vehicles. It can be seen from the graph that 0 takes up a heavy proportion. Another peak in the histogram takes place between 10 miles and 20 miles per day. This indicates that although no travel happening on a day at all is a very common phenomenon, if a car does travel, a very probable amount it travels on one day falls between 10 miles and 20 miles. However, this falls short when compared to the average annual VMT of 19,850 miles found in the 2009 National Household Travel Survey (NHTS), and the daily VMT of 28.97 (Santos, 2009). One explanation can be that the sample being surveyed in the data set used in this paper might not be the most representing group of the overall travel pattern nationally, and might on average travel less.

A regression is run of annual VMT (used in the study is the sum of 360 days of travel) over the travel mileage of one day or two consecutive days of each vehicle along with demographic information including household income, number of children, age of the head of the household, and number of drivers per vehicle, as well as date and day of the week the travel happened. After running the regression with randomly picked dates, the t-statistic and P-value of each variable is examined.

DATA ANALYSIS

Table 1. OLS of Annual VMT over Daily and 2-Day VMT (n_{obs} = 214 vehicles)

	Coefficients	*t Stat*	*P-value*		*Coefficients*	*t Stat*	*P-value*
Intercept	7128.3	16.54	0.000	Intercept	6757.4	15.81	0.000
Daily VMT	56.5	7.44	0.000	Two-Day VMT	34.4	8.60	0.000
Number of Children	485.4	1.98	0.049	Number of Children	351.3	1.48	0.139
Thursday	-1343.8	-1.97	0.050	Thursday	-1455.7	-2.21	0.028
Saturday	-1508.0	-2.20	0.029	Saturday	-833.9	-1.28	0.201
February	1430.5	1.73	0.085	February	1573.9	1.97	0.050
March	2354.8	2.49	0.013	March	2601.0	2.86	0.005
April	2468.4	3.18	0.002	April	2316.6	3.09	0.002
June	1706.1	1.61	0.108	June	1993.0	1.96	0.051
July	891.5	1.06	0.291	July	813.0	1.00	0.319
August	1697.7	2.31	0.022	August	1061.2	1.48	0.140

Adjusted R^2 = 0.273 Adjusted R^2 = 0.322

Table 1 above is the regression result for one set of random dates. Because of the instability of the correlation between annual VMT and the selected parameters, the regression is repeated 20 times. Table 2 contains all 20 runs' R^2 values, suggesting that Table 1's example is a relatively high fit result..

One notable finding is that the intercepts are very high, and the coefficient for the daily travel amount is significantly less than 360. This might be due to the zeros in the sample. Since there are days that the vehicles sit still without any travel, and these zero days can be sampled into the daily VMT. When being plotted, they occupy the positive y axis, and thus bringing the intercept up and the slope down.

Table 2. R-Squared Value for 20 Additional Regressions (n = 214 vehicles)

Single day	2-day

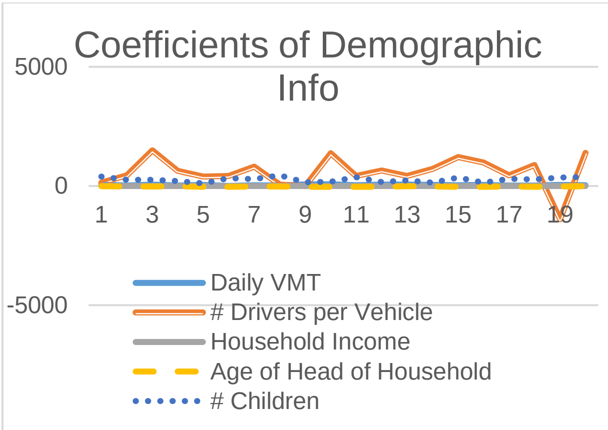
0.319	0.362
0.100	0.070
0.224	0.308
0.116	0.215
0.221	0.103
0.086	0.100
0.134	0.254
0.185	0.225
0.287	0.356
0.282	0.334
0.144	0.197
0.247	0.284
0.215	0.255
0.167	0.237
0.170	0.205
0.287	0.312
0.121	0.089
0.178	0.258
0.278	0.090
0.210	0.213

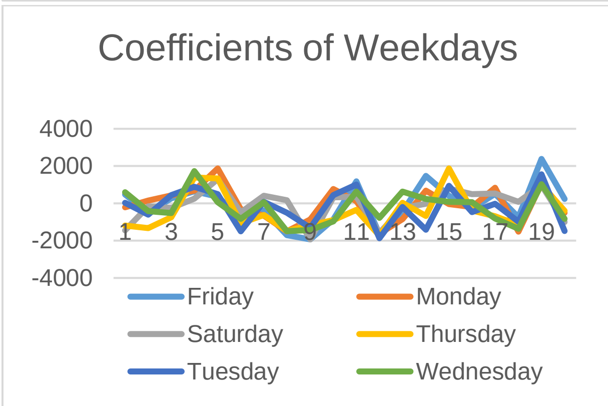
In the 20 pairs of regression models, the R-squared values vary significantly, but, as expected, most (80 percent) of the pairs delivered a higher R-squared value for the 2-day sample. This indicates that the prediction result is unstable, and also varies with the date being selected. However, within the same sample, 2 consecutive days of travel tends to be more descriptive of how much the vehicle travels throughout the year than only one day.

But the R-squared value for 2-day regressions can drop very low, one of the reasons being that when two days are added together, variability can fall, but can also increase when both days sampled are unusual (e.g., a summertime driving vacation or two consecutive days of zero driving). There 12,595 zero-VMT days in the data set (16.3% of the total), and half (6,263) of these zero-VMT days are directly followed by a zero-VMT day.

Plotting the coefficients can make evident which parameters are of more significance. The following plots are divided into three categories: demographic information, months, and weekdays.

Figure 2(a). Coefficients of each Variable in 3 Categories for Annual VMT vs. Daily VMT


Coefficients of Demographic Info
5000
0
-5000
1
3
5
7
9
11
13
15
17
19
Daily VMT
Drivers per Vehicle
Household Income
Age of Head of Household
Children


Coefficients of Weekdays
4000
2000
0
-2000
-4000
1
3
5
7
9
11
13
15
17
19
Friday
Monday
Saturday
Thursday
Tuesday
Wednesday

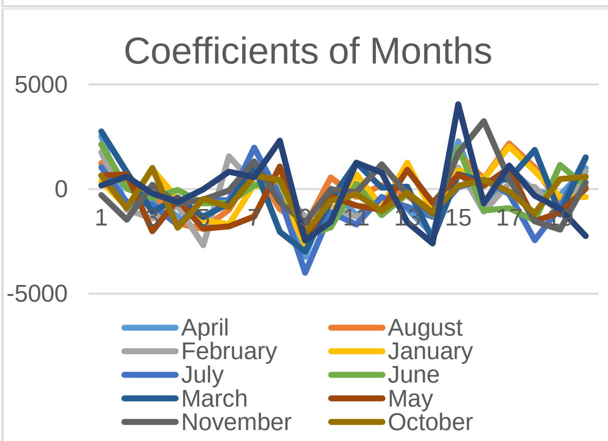
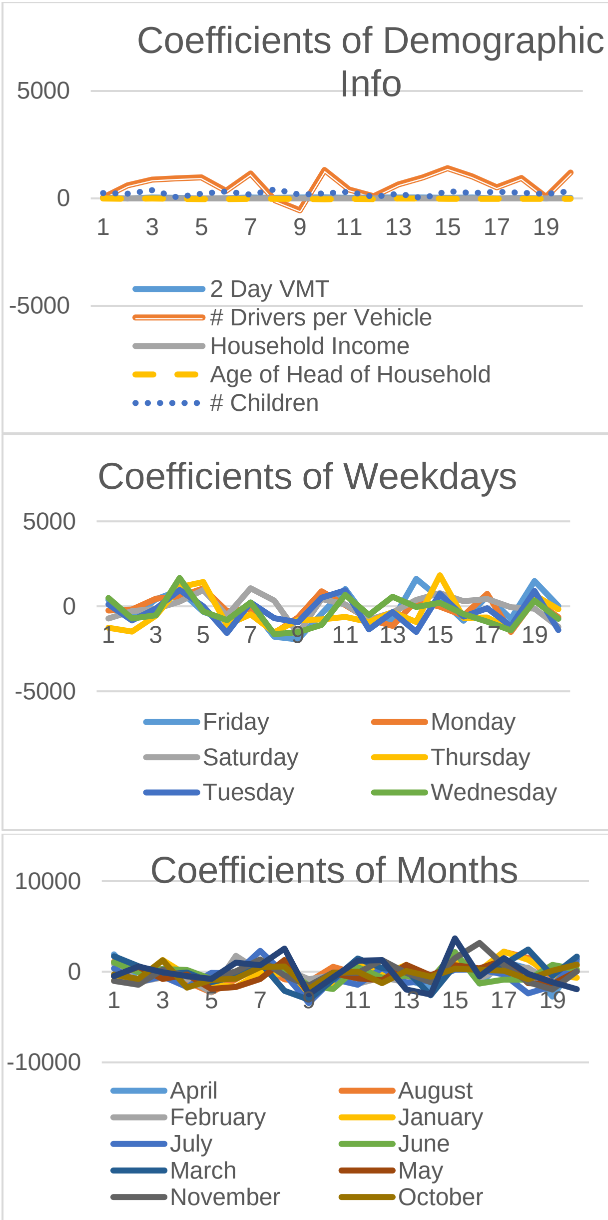

Coefficients of Months
5000
0
-5000
1
15
April
August
February
January
July
June
March
May
November
October

Figure 2(b). Coefficients of each Variable in 3 Categories for Annual VMT vs. 2-Day VMT



When looking at the results of the regression, not many consistent patterns can actually be drawn, other than the fact that the age coefficient of the head of the household tends to be negative and the number of children coefficient is usually positive. This is due to the variability of people’s travel patterns. Even ithe miles a vehicle travels on a day and all household information are known, it is difficult to determine whether that day is a day of heavy or light travel for the household. This is analogous to the fact that randomly selecting one sample from the population is often not relevant enough to predict the entire normal distribution. Surveying for more than one day can eliminate this type of error to some extent, but the extended survey period reduces reporting accuracy as well. It is possible, however, with the technology of GPS devices, to keep track of the surveyed vehicles’ travel pattern over a longer period of time, like an entire week, without sacrificing accuracy. As shown in Table 3 and as expected, R2 values for a week's worth of VMT data out-perform a day or two of such data, in predicting the household's annual VMT with that vehicle. It should be noted though, that the month to which a week is considered to belong is determine by which month the first day of the week falls in (so the month may change, and there is a greater chance of overlap in some of the sampled days across the 6 regression results, due to a week-long sampling period).

Table 3. R-Squared Value for 6 Annual VMT vs. Weekly VMT Regressions

	Adj R^2
Run 1	0.389
Run 2	0.454
Run 3	0.385
Run 4	0.445
Run 5	0.389
Run 6	0.499
Average	0.427

The next task, then, would be to sort out the demographic patterns of households that might have more regularly traveling vehicles, and an indication for this would be the Gini coefficient.

GINI COEFFICIENT

The Gini coefficient is a term originated from economics and has been used as a conventional measure of income inequality (Dorfman, 1979). However, it can be used in other areas as well. In this study, for example, it is possible to look at the Gini coefficient of the daily VMT of a certain vehicle across a year. It is defined as the area enclosed by the line of equality and the Lorenz curve (Turrell and Mathers, 2001). In economics, the Lorenz curve represents the cumulative distribution of income while income units such as individuals or households are arranged in an ascending order from left to right along the x-axis (Kakwani, 1977). The line of equality is a line in the first quadrant passing through the origin that forms a 45 degree angle with the positive x-axis, and it is the Lorenz curve when income is evenly distributed among all individuals. The income inequality Gini index of the United States was 0.480 in 2014, increasing by 5.9 percent since 1993 (DeNavas-Walt and Proctor, 2015). In this study, the line of equality shows the cumulative distribution of a car's daily VMT over a year's period if it travels the same amount every day throughout the year, while the Lorenz curve is the actual cumulative distribution across days of the week arranged from left to right in the order of ascending daily VMT. The area between the 2 curves then, is the Gini coefficient.

In this study, since it is picked 360 days as a year, there are a sufficiently large number of points on the x-axis to use the rectangular approximation method. The rectangular approximation method is used to approximate the area under the curve as the sum of the areas of n rectangles each having a width of $\frac{1}{n}$, while n is the number of days being looked at. In this case, the distance between the points on the line of equality and the Lorenz curve on each day is multiplied by the width of $\frac{1}{360}$ to get the area of 360 rectangles, and summing up the areas of them would be the approximated Gini coefficient.

Figure 3. Histogram for Gini Coefficient of Uniformity in Daily Travel Distances

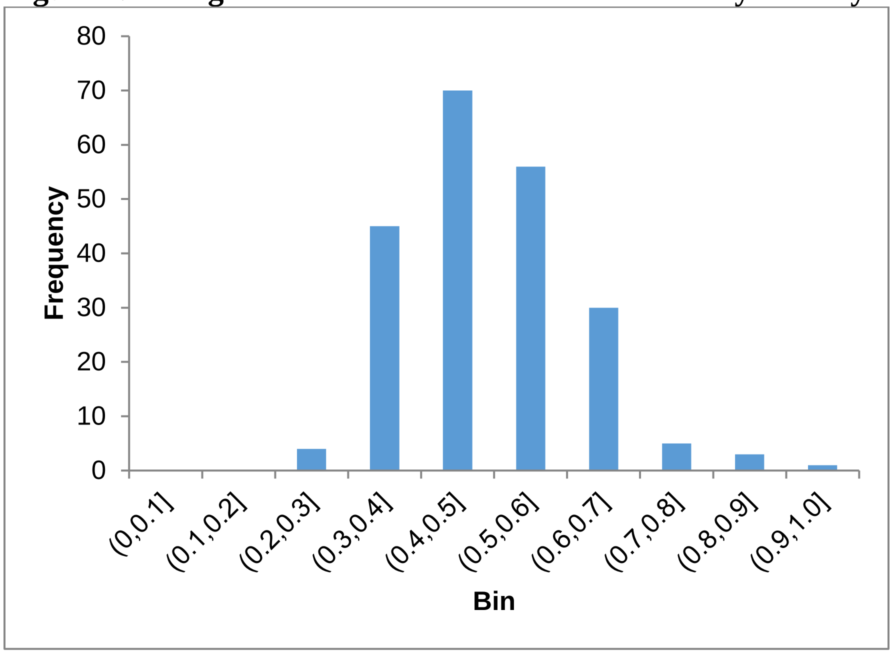


Figure 3 shows the histogram of the vehicles' Gini coefficient for daily VMT. Among the 214 Gini coefficients calculated, more than a third are between 0.4 and 0.5, indicating that a large percentage of Lorenz curves form an area with the line of equality approximately half the size of the triangle formed by the line of equality with the x and y axes. It is very rare that this value drops below 0.3 or goes beyond 0.8.

Figure 4. Lorenz Curves for n = 194 Vehicles over 360 Days versus Line of Equality

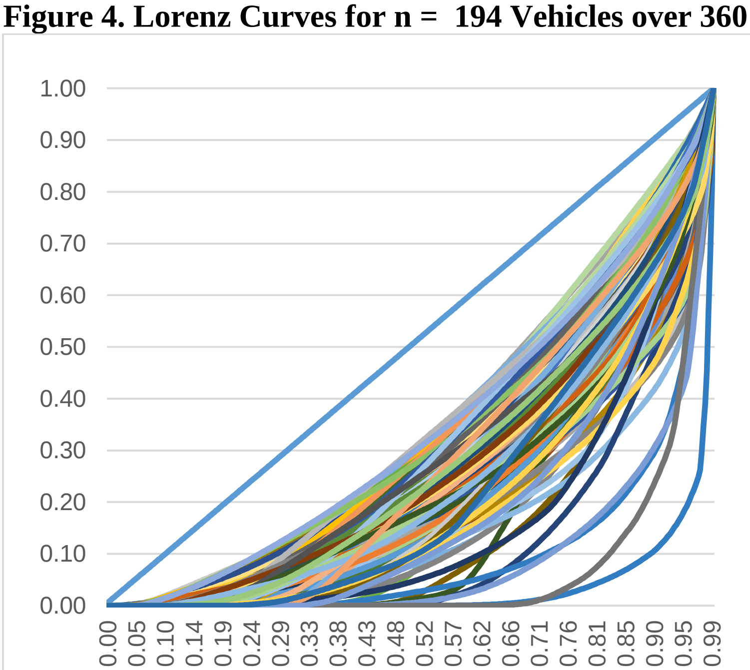


Figure 4 displays the plot for the Lorenz curve of 194 different vehicles, and the area between each Lorenz curve and the line of equality is the Gini coefficient of that specific vehicle's daily VMT throughout the year. This area reaches its maximum when all travel occurs on one day. As the total number of days increases, the maximum Gini coefficient approaches 1. When the travel pattern is homogeneous every day throughout the year, the Gini coefficient reaches its minimum of 0, and the Lorenz curve coincides with the equality line. Since there are vehicles whose travel mileage are not kept track of for a full 360-day period, they are removed from the list when graphing, and that's why the number of vehicles on the graph are less than that included in the regressions.

When surveying households, the lower the Gini coefficient of travel, the easier it is to predict the annual VMT from the daily value, since it indicates a more homogeneous travel. If the Gini coefficient is high because of how far the vehicle travels per day varies widely throughout the year a one day survey doesn't provide enough information on where this day falls on the spectrum of travel distance for this vehicle, and might be an indicator that suggests coming back to the household for another time, or asking them to provide travel details over a longer period of time.
What demographic characteristics may deliver more day-to-day VMT variability and a higher Gini coefficient for the vehicle? Table 2 provides the results of such a regression, for y = Gini coefficient versus household demographic information, the following result is shown in Table 2.

Table 2. OLS of Gini Coefficient over Annual VMT and Demographic Variables (n = 214 vehicles)

	Coefficients	t Stat	P-value
Intercept	0.2679	10.45	0.000
Annual VMT	0.0000	-5.09	0.000
Annual Household Income	0.0000	1.07	0.287
Number of Children	-0.0019	-0.43	0.666
Age of Household Head	0.0003	0.87	0.384
#Drivers per Vehicle	0.0058	0.30	0.761

From the regression, however, it can be concluded that the Gini coefficient heavily depends on annual VMT. This discovery makes sense, since generally the regular travelers that go to certain places (for example, school or work) every single day accumulate a higher annual VMT. Due to the low R-squared value of annual VMT over both single day and 2-day VMTs, using a predicted annual VMT to find out the Gini coefficient may cause even more significant error. But it can be possible to read from the odometer of the vehicle to extract the value of the annual VMT, and use this value to estimate the Gini coefficient.

Household income is another important variable in predicting the Gini coefficient. There is a positive coefficient linked to this variable, indicating that the Gini coefficient tends to be higher when the surveyed household has a high income. This pattern makes sense as well, since usually wealthier people have more control over their time, including what to do and where to go. They are more likely to take vacations and go on long journeys. Long, driven journeys would appear in the data as very large daily VMTs, while non-driven (or rented-car) journeys can result in a series of zeros, since the owners are away and their personal vehicles are left unattended. Either situation delivers a higher Gini coefficient.

The Gini coefficient tends to increase with age of the household head, which may be due to retirees not having a daily work or school commitment, and thus less regular travel needs.

The number of children in a household and the drivers-to-vehicles ratio seems are not statistically significant predictors of the Gini coefficient, but a larger data set will tend to result in smaller p-values on any coefficient estimate. Other variables that may be quite helpful to have include location of the household within its region (e.g., local population and jobs densities, land use balance metrics, and distance to worker workplaces), distances to nearby major cities, household income, number of students in the household, and ages of all household members.

CONCLUSION

People's travel patterns vary from day to day, and after knowing how much one travels on a specific day, it is may be difficult to predict how much a person travels due to the discrepancy in how one day of surveying might compare to other days in the year. Longer or more repetitive surveys, then, can be carried out in order to make a closer estimation, but the respondent's own accuracy in reporting trips decrease as the length of survey period increases. There are certain households that have vehicles that follow a more regular travel pattern, of which we can take the advantage to survey for only one day to reduce the cost of longer surveys and the loss of accuracy due to respondent fatigue. Higher annual VMT, lower household income, and lower age of the head of the household are all such indicators. If not all variables appear to indicate a low Gini coefficient, the researchers can consider surveying another day. But if it occurs that all variables suggest a high Gini coefficient, it might be a better idea to turn towards another household since the extended survey period to make up for the variability might result in even lower accuracy due to the survey length. One major difference between the daily VMT used in the study and that obtained in traffic surveys is that in the study, the daily VMT is collected by the GPS device, while in actual traffic surveys, this value is often self-reported. However, it was found that self-reported car trips compare well with the actual distance, especially when the trip is short (Salon, 2013). Also, because of the use of 360 days instead of 365 days when selecting data, the generated value can be scaled up by $\frac{365}{360}$ in order to get the travel mileage of a full 365-day year.

REFERENCES

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., & Haupt, T. (2002). Observing the Rhythms of Daily Life: A Six-Week Travel Diary. *Transportation*, *29*(2), 95-124.

Cervero, R., & Hansen, M. (2002). Induced Travel Demand and Induced Road Investment: A Simultaneous Equation Analysis. *Journal of Transport Economics and Policy (JTEP), 36*(3), 469-490.

DeNavas-Walt, C., and B. D. Proctor. "US Census Bureau. Income and poverty in the United States: 2014." *Current Population Reports* (2015): 60-252.

Dorfman, R. (1979). A Formula for the Gini Coefficient. *The Review of Economics and Statistics*, 61(1), 146-149.

Kakwani, N. (1977). Applications of Lorenz Curves in Economic Analysis. *Econometrica*, 45(3), 719-727.

Pendyala, R. M., & Pas, E. I. (2000). *Multi-Day And Multi-Period Data for Travel Demand Analysis and Modeling* (No. E-C008,).

Salon, D. (2014). Comparison of Self-Reported to Network-Calculated Trip Distances for the California Add-on to the 2009 National Household Travel Survey. In Transportation Research Board 93rd Annual Meeting (No. 14-5389).

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., & Liss, S. (2011). *Summary of Travel Trends: 2009 National Household Travel Survey* (No. FHWA-PL-ll-022).

Stopher, P., Kockelman, K., Greaves, S., & Clifford, E. (2008). Reducing Burden and Sample Sizes in Multiday Household Travel Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, (2064), 12-18.

Turrell, Gavin, and Colin Mathers. "Socioeconomic Inequalities in All-Cause and Specific-Cause Mortality in Australia: 1985–1987 and 1995–1997." *International Journal of Epidemiology* 30, no. 2 (2001): 231-239.