Predicting a Vehicle's Distance Traveled from Short-duration Data

Ruohan Li[*], Kara M Kockelman[†]

Transport Findings

This article uses one year's worth of daily travel distance data for 252 Seattle households' vehicles to ascertain that one day's distance (plus day of week and month of year information) accounts for 10.7% of the variability in that vehicle's annual (total) distance traveled, while two and seven consecutive days' distance values predict 16.7% and 33.6%, respectively. In analyzing Gini coefficients (which average $0.546 +/- 0.117$ across these instrumented vehicles), one finds that full-time employed females have the most stable day-to-day driving patterns, allowing for shorter-duration surveys of such households.

RESEARCH QUESTION AND HYPOTHESIS

Vehicle-miles traveled (VMT) or vehicle-kilometers (km) travelled (VKT) is a key measure of household travel behaviors and, therefore, regional and interregional travel demand (Cervero and Hansen 2002). Single-day and two-day travel surveys are the norm, with households completing detailed trip diaries and providing vehicle odometer values for 24- to 48-hour durations. Individuals and their households travel patterns, however, can vary considerably over time (Pendyala and Pas 2000). This day-to-day and month-to-month variability differs across households, drivers, and their vehicles. To reduce uncertainty in vehicle travel estimates by being more strategic in determining sample sizes and survey durations across demographic and other strata, this article explores day-to-day variability in vehicles tracked for an entire year of driving. The work investigates how valuable one-day versus two-day versus seven-day data are for inferring a year's worth of driving behaviors and which persons and household types tend to offer the most stable versus variable driving patterns, across days of the year.

METHODS AND DATA

The data used here comes from the Puget Sound Regional Council's Traffic Choices Study (NREL 2017), for the Seattle region of Washington in the United States. The data was collected between November 2004 and August 2006 by placing GPS tolling meters on the vehicles of volunteer households. The final data set contains 329 unique households and 484 vehicles. The duration of GPS device placement on vehicles varied across households and

* The University of Texas at Austin; Department of Civil Architectural and Environmental Engineering

† The University of Texas at Austin; Department of Civil Architectural and Environmental Engineering
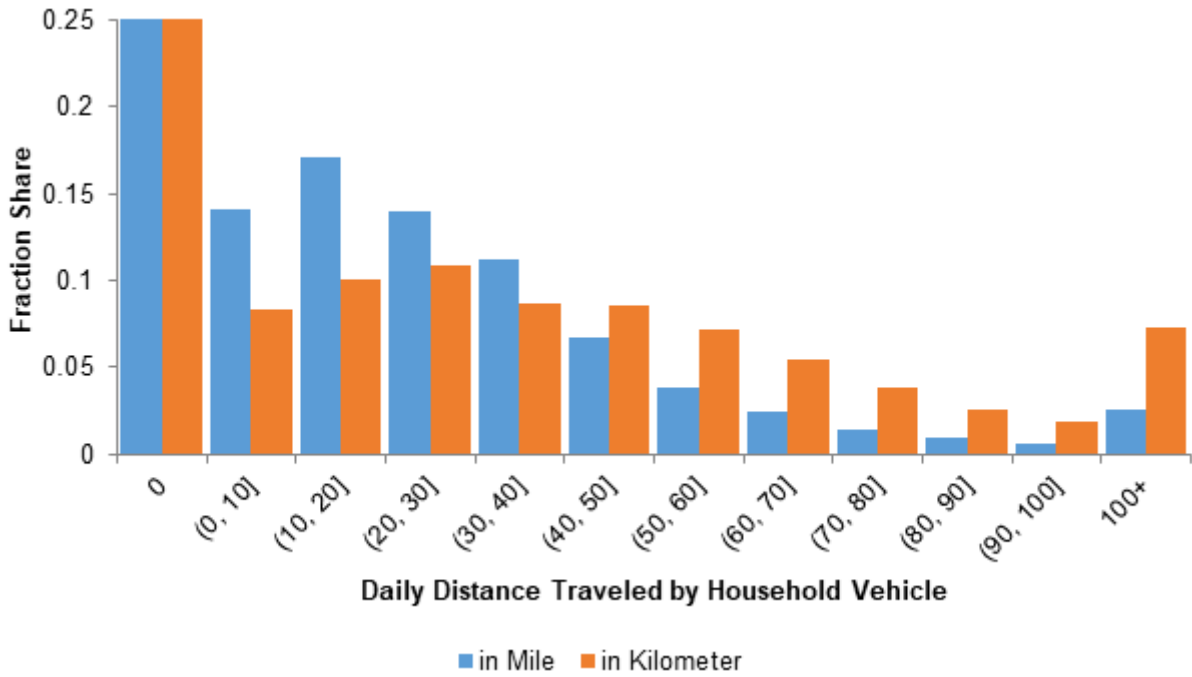 ORCID iD: 0000-0003-4763-1304 Link: http://www.caee.utexas.edu/prof/kockelman/

Figure 1: Histogram for Daily VMT and VKT (N= 91,980 survey days, across 252 household vehicles)

across vehicles within each household. In this study, vehicles with less than a full year (365 days) of data were excluded. To avoid correlations in travel between pairs or triplets of vehicles owned by the same household, only one vehicle per household has been randomly selected for analysis.

Figure 1 provides histograms of daily VMT and VKT values for all 252 vehicles. Common values are zero-kilometer days (25% doing no driving on that GPS-survey day), and between 16.1 and 32.2 km (10 and 20 miles) per day (17.1%). Average daily travel is 38.9 ± 55.8 km (24.2 ± 34.7 miles) per vehicle, which is reasonably consistent with the average daily vehicle travel of 46.62 km (28.97 miles) per day per driver, as found in the 2009 National Household Travel Survey (NHTS) (Santos et al. 2011).

A total of 60 runs of regression were made, with 20 different sets of starting dates for all vehicles randomly selected and the short period capturing one, two, or seven days of travel data, respectively. Table 1 shows adjusted R^2 values of all 60 ordinary least squares (OLS) regressions, which increase as the sampled length is extended. An entire week's data outperforms single- and two-day distance data in predicting annual VMT. As technology advances, GPS devices have been used more and more often in longer travel behavior studies, as in Stopher et al.'s (2007) Sydney, Australia survey, and the Seattle, Washington region's Traffic Choices study (Khan and Kockelman 2012).

Additionally, when multiple days are sampled, the adjusted R^2 value is higher when each sampled day vehicle travel is individually used as a regressor than when they are summed up to be one variable.

Table 1: Summary Statistics for Adjusted R^2 Values over 20 OLS Regressions for Y = Single Vehicle's Annual Travel Distance

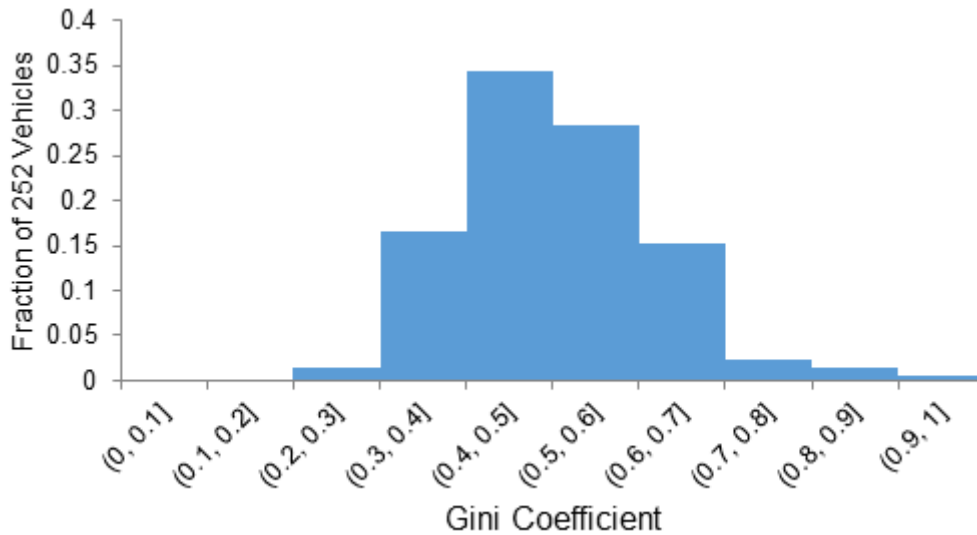| Adjusted R^2 Values	X's = Single Day Distance with Day of Week & Month of Year	X's = Single Day Distance without Day of Week & Month of Year	X's = Two Individual Days' Data	X's = Sum of Two Days' Distances	X's = Seven Individual Days' Distance Data	X = Sum of Entire Week's Travel Distance
Mean	0.1023	0.1046	0.1855	0.1667	0.4423	0.3361
Standard Deviation	0.0730	0.0652	0.0695	0.0724	0.0539	0.0805
Median	0.0864	0.1094	0.1801	0.1595	0.4415	0.3343

Figure 2: Histogram for Gini Coefficients on 252 Vehicles' Daily Travel

Table 2: Variability Measures of Household and Vehicle Travel

| | Gini Coefficient | |
	Household	Vehicle
Mean	0.490	0.546
Standard Deviation	0.106	0.117
Minimum	0.265	0.277
Maximum	0.875	0.937
25th Percentile	0.415	0.460
Median	0.470	0.528
75th Percentile	0.553	0.616

GINI COEFFICIENT

The Gini coefficient is a common measure of income inequality (Dorfman 1979), and can be used for other variables as well. In economics, the Lorenz curve illustrates the cumulative distribution of income, with percentage of individuals or households arranged in an ascending order along the x-axis (Kakwani 1977). Gini (1921) called the area between the Lorenz curve and the $y = x$ line the "area of concentration," and the Gini coefficient is the ratio between that area and the total area under the $y = x$ line (or 0.5). This study calculates a Gini coefficient for each of the 252 vehicles using the 365 daily travel distance values.

Table 2 provides summary statistics for day-to-day variability in vehicle travel per household and per vehicle. Household vehicle travel values are less variable than the per vehicle values.
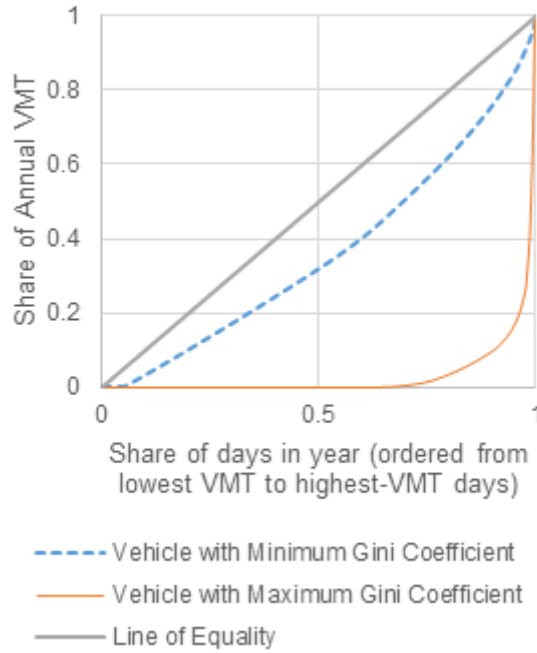
Figure 3: Lorenz Curves for One Year's Worth of Daily Travel Values for Vehicles with Highest and Lowest Gini Coefficients

Table 3: Gini Coefficient Regression Summary (OLS Method)

Model Covariates	Coefficient	T-Statistic
Intercept	0.5299	27.55
Annual vehicle travel under 8,047 km (5,000 miles)	0.1495	6.081
Annual vehicle travel over 16,093 km (10,000 miles)	−3.784E-02	−2.860
Annual household income for 2004	2.239E-07	1.673
Age of assigned driver: 20–29 years	2.539E-02	1.063
30–39 years	1.910E-02	1.242
70–79 years	6.859E-02	1.681
Whether assigned driver works full-time	−5.581E-02	−3.199
Whether assigned driver is a homemaker	−5.655E-02	−1.734

The Gini coefficients of the vehicles studied have an approximately bell-shaped distribution, with 162 vehicles' values concentrated between 0.4 and 0.6, only 2 vehicles having a Gini coefficient less than 0.35, and 8 vehicles' coefficients being greater than 0.8. Figure 3 presents the Lorenz curves of the two extreme Gini coefficient cases in the dataset.

A regression was then run for 252 vehicles, with the Gini coefficient as the dependent variable; results are summarized in Table 3. Apart from the demographic variables, indicators for whether a vehicle's annual travel is under 8,047 km (5,000 miles) or above 16,093 km (10,000 miles) are also used as predictors. With a total of eight independent variables, the adjusted R^2 equals 0.2807.

According to Table 3, the variable that has the most positive impact on the Gini coefficient is whether the annual mileage travelled by the studied vehicle falls under 8,047 km (5,000 miles). Meanwhile, the three major factors related to a lower Gini coefficient are whether the primary driver of the vehicle works full time, whether it travels more than 16,093 km (10,000 miles) a year, and whether the driver is a homemaker. Although how much the vehicle will travel is unknown, a possible approach to estimate is to ask its primary driver to select a range for vehicle travel over the past year, under the assumption that the vehicle's travel pattern, though varying on a daily basis, tends to be constant from year to year (Oak Ridge National Laboratory 2011).

Among the drivers in the lowest quartile in Gini coefficient, more than 90% are employed full-time, significantly higher than the 75% sample average. Thirty percent are male, lower than the sample average of 46%. Their average annual vehicle travel is ~800 km (500 miles) short of the average annual travel across all the sampled drivers. When regression of annual vehicle travel is run with the same variables limited to only this quartile of drivers, an average adjusted R^2 of 0.374 is reached, which is higher than if run with the entire group using a full week's data. In other words, the vehicles within the top 25% in travel homogeneity have daily vehicle travel that provides predictive power as strong as a whole week's data from an average vehicle in the dataset for annual vehicle travel.

FINDINGS

Individuals' travel patterns vary from day to day, so dramatic mistakes are regularly made when extrapolating one day's vehicle travel to a full year's. Full-week trip entries facilitated by technologies such as GPS devices improve the accuracy in annual vehicle travel prediction significantly compared to single-day and two-day trip diaries. When a daily vehicle travel value comes from a vehicle with a low (first-quartile) Gini coefficient, that vehicle's annual vehicle travel can be better predicted than when using a full week's data from a vehicle with an average Gini coefficient. Such consistently used vehicles are more likely to be driven by full-time employed female drivers who travel slightly less than average. However, overall higher mileage throughout the year tends to correspond to lower day-to-day variability.

ACKNOWLEDGMENTS

REFERENCES

Cervero, R., and M. Hansen. 2002. "Induced Travel Demand and Induced Road Investment: A Simultaneous Equation Analysis." *Journal of Transport Economics and Policy* 36 (3): 469–90.

Dorfman, Robert. 1979. "A Formula for the Gini Coefficient." *The Review of Economics and Statistics* 61 (1): 146. https://doi.org/10.2307/1924845.

Gini, Corrado. 1921. "Measurement of Inequality of Incomes." *The Economic Journal* 31 (121): 124–26. https://doi.org/10.2307/2223319.

Kakwani, N. 1977. "Applications of Lorenz Curves in Economic Analysis." *Econometrica* 45 (3): 719–27. https://doi.org/10.2307/1911684.

Khan, M., and K. Kockelman. 2012. "Predicting the Market Potential of Plug-in Electric Vehicles Using Multiday GPS Data." *Energy Policy* 46 (July): 225–33. https://doi.org/10.1016/j.enpol.2012.03.055.

NREL. 2017. "2004-2006 Puget Sound Traffic Choices Study." Transportation Secure Data Center, at National Renewable Energy Laboratory. https://www.nrel.gov/transportation/secure-transportation-data/tsdc-puget-sound-traffic-study.html.

Oak Ridge National Laboratory. 2011. "Developing a Best Estimate of Annual Vehicle Mileage for 2009 NHTS Vehicles."

Pendyala, R.M., and E.I. Pas. 2000. "Multi-Day and Multi-Period Data for Travel Demand Analysis and Modeling." No. E-C008.

Santos, A., N. McGuckin, H.Y. Nakamoto, D. Gray, and S. Liss. 2011. "Summary of Travel Trends: 2009 National Household Travel Survey." (No. FHWA-PL-11-022). Washington, D.C. https://nhts.ornl.gov/2009/pub/stt.pdf.

Stopher, P. R., K. Kockelman, S. P. Greaves, and E. Clifford. 2008. "Reducing Burden and Sample Sizes in Multiday Household Travel Surveys." *Transportation Research Record: Journal of the Transportation Research Board* 2064 (1): 12–18. https://doi.org/10.3141/2064-03.