# VARIATIONS IN VEHICLE-MILES TRAVELLED: PREDICTING A VEHICLE'S ANNUAL VMT FROM SHORT-DURATION DATA

Ruohan Li
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin – 6.9 E. Cockrell Jr. Hall
Austin, TX 78712-1076
lruohan_322@yahoo.com

Kara M. Kockelman
(Corresponding Author)
Dewitt Greer Centennial Professor of Transportation 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

Published in *Transportation Findings* (2019).

## ABSTRACT

A region's daily and annual vehicle-miles travelled (VMT) are important for moderating congestion, evaluating transportation policy and investment decisions. VMT is difficult to track and surveys of households offer low sample size and only a day or two of odometer readings. This paper uses a year's worth of daily VMT data for 215 Seattle pax. vehicles to see how useful short-duration VMT data really are and how variable each vehicle's VMT really is. A day's worth of VMT plus month of the year and day of the week reflects just 27% of the demographic variables' annual totals while 2 days of data predicts $R_{adj}^2$ = 33%, can recover 47% of the annual VMT's variation. The average Gini coefficient across these 215 vehicles is 0.51, and average coefficient of variation (standard deviation over mean) is greater than 1.0, suggesting substantial variation day to day and month to month. Vehicles owned by households of lower annual income, with middle-aged, full-time workers have most stable daily VMT values, allowing researchers to place greatest value on short-term VMT data from households of this type.

## INTRODUCTION AND MOTIVATION

Vehicle-miles traveled (VMT) is a key measure of household and regional travel demand (Cervero et al., 2002). Single day surveys are the norm with households completing detailed trip dairies and providing for all 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). There can be days of extremely heavy travel, as well as days on which no travel takes place. Compared to one day of trip data, two-day surveys better capture such variation. While 2 and 3-day surveys have become more common (Axhausen et al., 2000), respondent fatigue limits anything longer.

## DATA SET

The data came from the Puget Sound Regional Council (PSRC) when it conducted the Traffic Choices Study from 2005 to 2006 by placing GPS tolling meters on vehicles of volunteer

households. The final data set contains 329 unique households and 484 vehicles. To remove correlation in travel among different vehicles in the same household, one vehicle per household was used. Moreover, households with a low tracking period or missing demographic information were also removed, resulting in a dataset of 215 vehicles from 215 households.

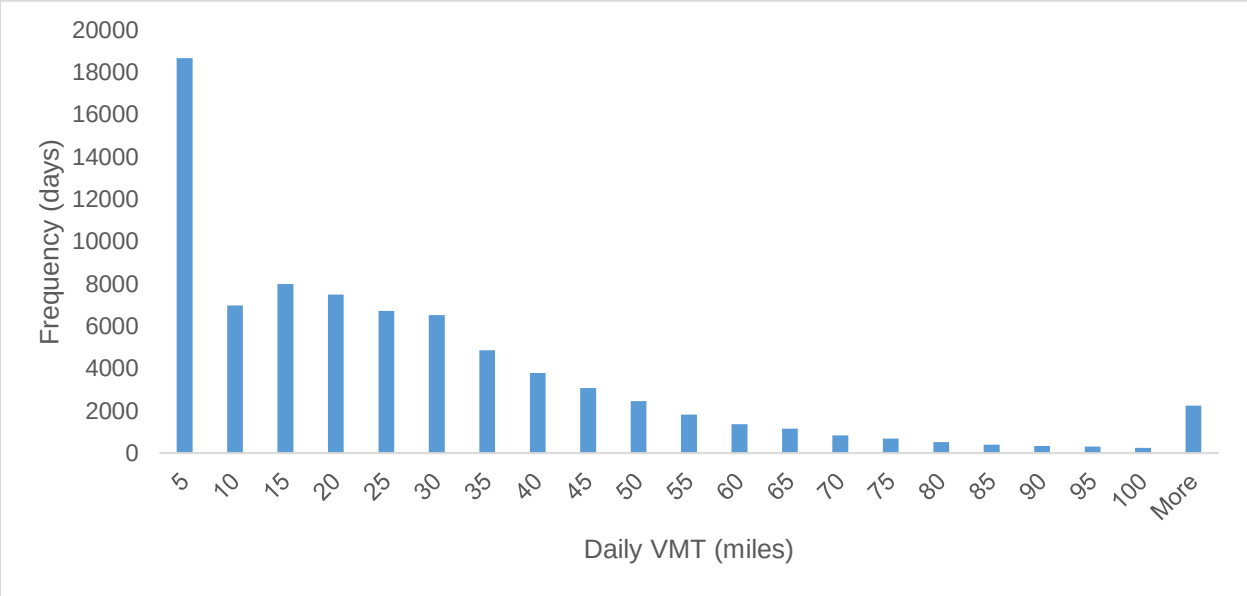


**FIGURE 1 Histogram for daily VMT**

Figure 1 provides a histogram of daily VMTs of all 215 vehicles. Common values are zero (for no driving). 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 covers on one day falls between 10 miles and 30 miles. Among the total 81618 surveyed vehicle-days, 29331 of them experience travel distances between 10 miles and 30 miles. The average daily VMT is 26.37 ± 35.53 miles, which is reasonably consistent with the average daily VMT of 28.97 miles per day per driver, found in the 2009 National Household Travel Survey (NHTS) (Santos, 2009). But the median found to be 18.64 miles per day turns out to be significantly lower than this value.

## DATA ANALYSIS

To get a sense of how well one can predict a household vehicles' annual VMT from such short-duration data, regressions were run of annual VMT (from April 3, 2005 to April 2, 2006) as a function of 1-day, 2-day, or 1-week distances, along with demographic information and month of year and day of week the travel happened. Random days are selected at least 1 week before 4/2/06, and along with them the following one day or six days for the 2-day or 1-week data. The independent variables, of which the annual VMT is considered a function, include the short-duration VMT, household income, age of the driver, number of children within the household, number of drivers per vehicle, driver's years of education, and month of year and day of week of the selected day or the day with which the sampled dates start.

Using OLS regression, the coefficients on the daily, two-day, and weekly VMT are 18.63 (rather than 52 weeks/year), 36.93 (rather than 182 weeks/year), and 47.37 (rather than 364 days/year). and much flatter slopes than. This is largely due to the abundance of zero-VMT days in any household vehicle travel sample. When a no-travel day is sampled, the independent variable is 0, which is not so helpful in predicting annual VMT.

The adjusted R-squared values of all 60 OLS regressions are shown in Table 1. As sampled duration rises, the adjusted R-squared increases, and thus, the accuracy in predicting annual VMT improves. With weekly VMTs known, the prediction is not bad, with an average adjusted R-squared of 0.4711.

**TABLE 2 Adjusted R-Squared Value for 20 OLS Regressions**

| | Weekly | Two-Day | Daily |
|---|---|---|---|
| Run 1 | 0.6339 | 0.4486 | 0.3383 |
| Run 2 | 0.3475 | 0.2936 | 0.2457 |
| Run 3 | 0.5217 | 0.4691 | 0.3830 |
| Run 4 | 0.4739 | 0.2680 | 0.2689 |
| Run 5 | 0.4679 | 0.2908 | 0.2420 |
| Run 6 | 0.4661 | 0.2717 | 0.2558 |
| Run 7 | 0.4102 | 0.3565 | 0.3042 |
| Run 8 | 0.5753 | 0.4384 | 0.3242 |
| Run 9 | 0.4167 | 0.2492 | 0.1941 |
| Run 10 | 0.4356 | 0.2913 | 0.2569 |
| Run 11 | 0.4011 | 0.2454 | 0.2028 |
| Run 12 | 0.4000 | 0.2919 | 0.2536 |
| Run 13 | 0.5458 | 0.3558 | 0.3043 |
| Run 14 | 0.4590 | 0.2951 | 0.2430 |
| Run 15 | 0.5249 | 0.3367 | 0.2375 |
| Run 16 | 0.5053 | 0.3571 | 0.3129 |
| Run 17 | 0.4199 | 0.3754 | 0.3019 |
| Run 18 | 0.5959 | 0.3984 | 0.2926 |
| Run 19 | 0.3923 | 0.3329 | 0.2889 |
| Run 20 | 0.4302 | 0.3224 | 0.2506 |
| Average | 0.4711 | 0.3344 | 0.2751 |

The R-squared values vary a fair across each of 20 regressions, and annual VMT vs. 2-day VMTs turns out to be higher than that of annual VMT vs. daily VMTs. The R-squared value for 2-day regressions dropped just 0.24 in Run 11, presumably due to day to day correlation in VMT values as when one travels out of the region leaving a household vehicle without a driver, or gets

sick and doesn't leave home. In fact, 49.7% of the zero-VMT days are followed by another zero-VMT day.

A balance then, needs to be reached, so that the surveyed period is long enough to capture the variation, while short enough not to cause respondent fatigue.

When looking at vehicles individually, each one behaves differently, as shown in Figure 2, some travel more regularly and are easier to predict an annual VMT for, even with only a one-day survey, while others may really need an extended survey period. It is helpful to know which vehicles need longer or shorter survey durations.

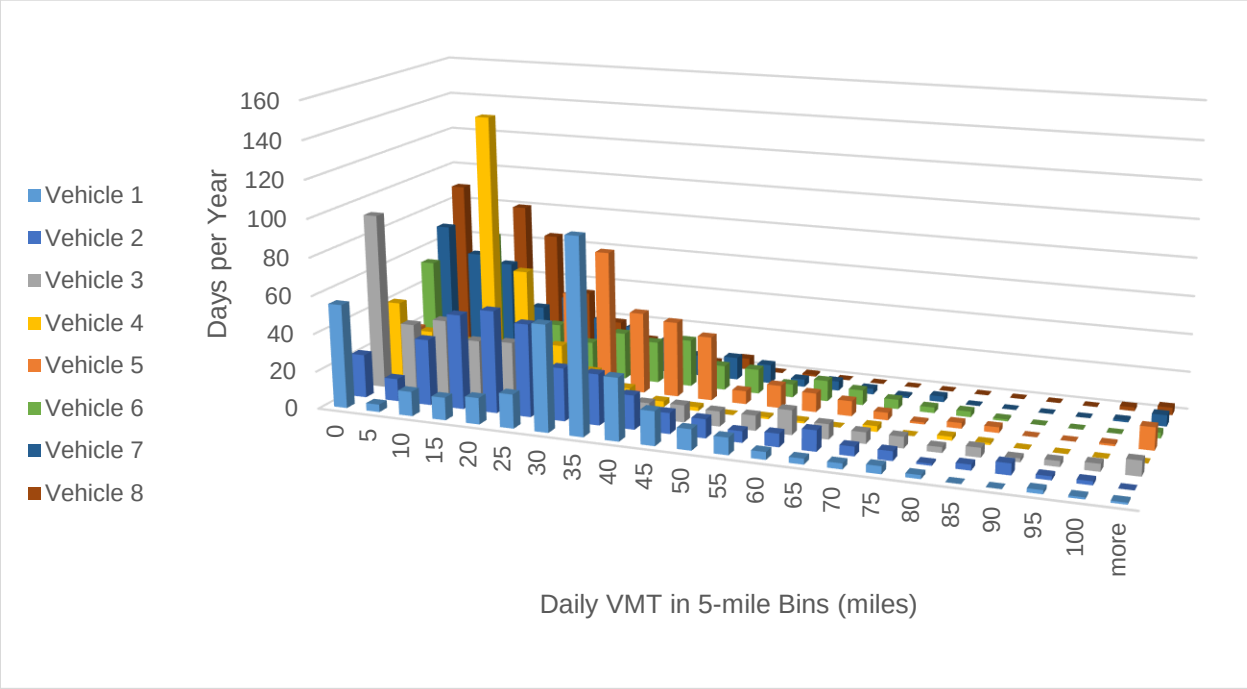


**FIGURE 2 Daily VMT distribution for 8 randomly selected vehicles**

Two distinctive measures of variability are the coefficient of variation and Gini's coefficient.

## COEFFICIENT OF VARIATION

Coefficient of variation is simply the standard deviation in a set of values divided by the mean or average value. It is easily understood but does not have an upper bound and can be overly sensitive to outliers (Kvålseth, 2017).

Among all 215 vehicles' 365 daily VMT values, the average coefficient of variation is 1.24, with a standard deviation of 0.569. The 25% percentile is 0.87, while the 75% percentile is 1.39, indicating that most vehicle's standard deviation in daily VMT exceeds their mean, and thus high day to day variability in driving distances.. The vehicle with the highest coefficient of variation of 5.5912, is driven by a student who lives alone with 1 vehicle, while the vehicle with the lowest coefficient of variation of 0.53, belongs to a household with two cars, no kids, and two drivers, both full-time workers.

## GINI COEFFICIENT

Gini coefficient is an economic measure of income inequality (Dorfman, 1979) and it can be used for other areas as well. This study calculates a Gini coefficient for each of 215 vehicles using the 365 daily VMT values. The Gini coefficient is the area enclosed by the y=x line of equality and the Lorenz curve, also known as the cumulative distribution function (Turrell and Mathers, 2001). In economics, the Lorenz curve illustrates the cumulative distribution of income, percentage of individuals or households arranged in an ascending order along the x-axis (Kakwani, 1977). The line of equality coincides with the Lorenz curve when income is evenly distributed among all individuals. The U.S.'s income inequality Gini index was 0.480 in 2014, or 5.9 percent higher than it was in 1993 (DeNavas-Walt and Proctor, 2015). In this study, the line of equality shows the cumulative distribution of a household vehicle'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 mileage traveled, with days of the year arranged from left to right in the order of ascending daily VMT. The area between the two curves is the vehicle's Gini coefficient.

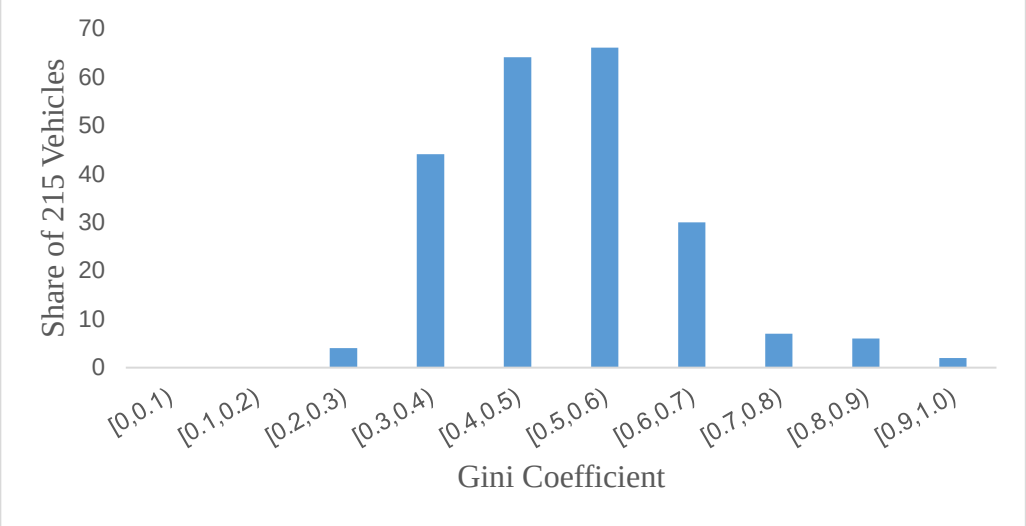


**FIGURE 3 Histogram for gini coefficient**

The studied vehicles have an average Gini coefficient of 0.5065, and they are distributed in an approximately bell-shaped curve, with 130 concentrated in the range between 0.4 and 0.6, only 4 less than 0.3, and 5 greater than 0.8. Thus, there is some degree of variability in most vehicles' travel pattern, while only a few that travel rather constant and a few that travel extremely unpredictably throughout the year. To better look at the variations graphically, Figure 4 is produced by plotting all 215 Lorenz curves in the same coordinate.

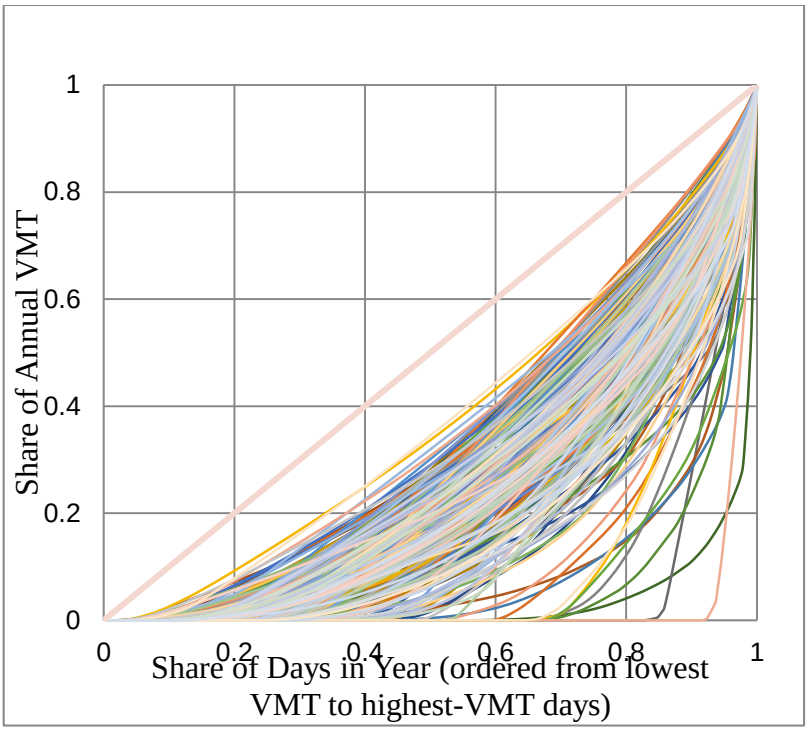


**FIGURE 4 Lorenz curves for 1 year's worth of daily VMT values of 215 vehicles**

The straight line that forms a 45-degree angle with the axes is the line of equality, which can be regarded as the Lorenz curve of a hypothetical vehicle whose daily VMT remains constant throughout the year. The farther away from this line is vehicle's Lorenz curve located, the more variability there is in this vehicle's annual mileage distribution. The lines that rise smoothly and gradually over the entire length correspond to those with a low Gini coefficient while those that remain flat over much of the x-axis and slopes up all of a sudden correspond to those with a high Gini coefficient.

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. With a high Gini coefficient, how far the vehicle travels per day varies widely throughout the year, and data from a single-day survey might not provide sufficient information to make a close estimation on how much the vehicle travels over an entire year. It would be a good idea to increase the study period to two days or even a week for such vehicles.

The next question would be, how to distinguish these vehicles? What demographic characteristics can suggest a higher Gini coefficient, or, in other words, a less equally distributed travel pattern throughout the year? A regression is then run, for 215 vehicles, with Gini coefficient as the dependent variable, and the demographic traits as independent variables.

According to the regression result, variables that contribute most positively to the Gini coefficient is the household income. The higher a household's income, the less likely drivers within this household would travel evenly throughout a year. Commuting or driving carpool

often, having an age between 30 and 49, and working as a full-time employee all are factors contributing negatively to the Gini coefficient. Among all the driver age groups, being between the ages of 20 and 29 has the largest regression coefficient.

Since the adjusted R-squared is calculated to be 0.1130, showing not a very well-fit linear regression. Compared to the fitted model, it might be more helpful to look at the traits of certain individuals rather than the overall trend of the entire sample.

The vehicles with Gini coefficients lower than 0.3 correspond to drivers with ID labeled 10042, 10041, 210, 308, and 58. Two traits that all five of them share in common is that their ages all fall in the range between 50 and 59, and that they are all full-time employees. It is also found that none of them have children within their households, however, when the drivers whose vehicles have the highest Gini coefficients for daily VMT are analyzed, it is interesting to see that all six drivers with vehicle Gini coefficients over 0.8 do not have children within their households either. Thus, number of children within the driver's household cannot be used as a deterministic factor for estimating whether or not the vehicle would experience a stable travel pattern distributed evenly across the year. No obvious pattern in age or employment has been discovered within the high Gini coefficient group, but five out of six of them are female, with education years between 17 and 19.

## CONCLUSION

People's travel patterns vary from day to day. So knowing how much one travels on a specific day does not make it easy to predict a year's travel distance. Longer and more burdensome surveys can be carried out, but without GPS, accuracy will suffer. Gini coefficients are used to evaluate the heterogeneity of each vehicle's travel pattern across the year. To maximize the efficiency and accuracy for annual VMT prediction, different vehicles can be assigned different survey period lengths due to their potential Gini coefficients. Full-time employed drivers between the age of 50 and 59 tend to be the most stable drivers, for most of whom a single-day survey might be sufficient. Female drivers with education years between 17 and 19 tend to have the most variable travel pattern, and a week of survey period might be needed to eliminate the effect of the instability. For drivers with both or neither traits, a two-to-three-day survey might be considered, depending on the situation.

However, some idealizations and approximations made in this paper might contribute to some extent of inaccuracy or error. For example, each vehicle is assumed to be linked to one and only one driver, but in reality, some vehicles are shared by multiple drivers, while some drivers have access to more than one vehicle. Sometimes a vehicle is linked to the demographic information of a certain driver, using all of it in the regressions, while there is actually another driver using it whose information is not taken into account. Another one is that in order to avoid correlation within the same household, one vehicle is selected per household to be analyzed. However, the Gini coefficient in travel pattern of the vehicles whose data have been discarded due to this reason are not kept track of. It can be hypothesized intuitively that being the primary or secondary vehicle of the household can be a significant factor impacting the Gini coefficient as

well. But the dataset doesn't show which vehicles are the primarily used ones by the household drivers. Another way to solve this problem is to include all the vehicles instead of keeping only one per household, and use weighted least squares rather than ordinary least squares.

## AUTHOR CONTRIBUTION STATEMENT

The authors confirm the contribution to the paper as follows: study conception and design: Li, R. and Kockelman, K.; Data analysis and interpretation of results: Li. R. and Kockelman, K., Draft manuscript preparation: Li, R., and Kockelman, K. All authors reviewed the results and approved the final version of the manuscript.

## ACKNOWLEDGMENTS

This work owes much credit to the NSF Sustainable Research Networks project that funded it, and Scott Schauer-West who provided great help in editing.

## 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 (2015). US Census Bureau. Income and poverty in the United States: 2014. *Current Population Reports*: 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.

Klakto, T., Saeed, T. U., Volovski, M., Labi, S., Fricker, J. D., & Sinha, K. (2017). Addressing the local road vmt estimation problem using spatial interpolation techniques. *Journal of Transportation Engineering Part A Systems*, *143*(8).

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 Proceedings of the 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.

Tarald O. Kvålseth (2017) Coefficient of variation: the second-order alternative, *Journal of Applied Statistics*, 44:3, 402-415.

NREL, "Transportation Secure Data Center" (2015). National Renewable Energy Laboratory. |Date TSDC data was accessed|. www.nrel.gov/tsdc.

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.

Yitzhaki, Shlomo and Edna Schechtman. *The Gini Methodology: A Primer on a Statistical Methodology*. Heidelberg;New York;: Springer, 2013.