# Driving and Land Use: An Explanatory Model

## An analysis prepared for the 37 billion mile challenge, April 19, 2014

**Paul Schimek and Zia Sobhani with Kim Ducharme**

What are the relative contributions of land use and other geographic factors and demographic factors in explaining variation in driving in Massachusetts? A new database released by the Metropolitan Area Planning Council in cooperation with state agencies provides an unprecedented opportunity to examine this question. The Vehicle Census of Massachusetts includes the location, number and type of motor vehicles registered in Massachusetts, and estimates of miles traveled per day in these vehicles based on matched pairs of annual vehicle inspection records.[1] In order to better focus on small-area land use characteristics, we used the "grid cell dataset," which consists of summary data geocoded to square grid cells measuring 250m on each side (approximately 15 acres per cell). These uniform, statewide grid cells do not correspond to any other geographic boundaries. However, MAPC geocoded other characteristics at the grid cell level including data about population, employment, street network, buildings, and land use types in a supplementary database. The data used for this analysis can be viewed and explored on a complementary website: http://zsobhani.github.io/37bill/vizSubmission/

The grid cell dataset consists of snapshots in time by quarter. Inspection of the data, combined with the documentation of how it was produced, shows that the earliest years of the data are incomplete (because of few matching inspection records), and also suggests the most recent (2011) data may be incomplete. We decided to look at a single snapshot, the second quarter of 2010, because this is the point in time that most closely corresponds with the April 1, 2010 date of the 2010 U.S. Census, which was used as the source of population data.

The first question is how best to measure car use so that it can be compared to geographic and demographic factors. We are only interested in household travel, so we exclude commercial vehicles and miles traveled in commercial vehicles. (This may exclude some personal travel done in commercial vehicles, as well as in taxis and carshare vehicles, but these are a small enough share of household vehicle use that they almost certainly do not affect the results.) The Vehicle Census includes data on the number of vehicles and driving per vehicle. Miles traveled per vehicle is not the best variable to consider because there is a large variation in the number of vehicles per household. A household with two adults sharing a single vehicle may show higher than average miles per vehicle, whereas a similar household owning two (or three) vehicles may have much less driving per vehicle but much more total driving per household. Therefore we instead look at miles of driving per day per *person*. Specifically, we calculated Miles Per Day Per Person:

$$MPDPP = mipdaybest * pass\_veh / pop10$$

where

[1] Vehicle Census of Massachusetts, Metropolitan Area Planning Council 2014, documented here: http://www.mapc.org/sites/default/files/VehicleCensusofMA_Documentation_v1.pdf. The Massachusetts Vehicle Census dataset is licensed by MAPC under a Creative Commons Attribution 4.0 International Public License, http://creativecommons.org/licenses/by/4.0/legalcode.

mipdaybest =Average daily mileage for "best" estimate vehicles. Estimates greater than 200 miles per day were excluded.

pass_veh = Count of total passenger vehicles assigned to the cell, either via geocoding or assigning a proportional amount of vehicles where only the municipality or only the zipcode is known.

pop10 = Estimated population per grid cell. MAPC allocated U.S. Census 2010 block-level population counts to grid cells based on point-level household records from third-party data vendor (InforGroup) and the distribution of residential land uses (MassGIS,Land Use 2005).

Only 144,453 of the 355,728 grid cells in the state have population according to the allocated MAPC data (the rest therefore drop out of the analysis). In addition, in 43,242 cases populated grid cells had no data for miles traveled or vehicles registered (mipdaypass, which is less restrictive in selecting paired inspection records, has almost as many cells with no data). In other cases, the calculated MPDPP suggests hundreds or even thousands of miles of driving per person per day. There are clearly anomalies in the data which could be the result of:

- Vehicles registered to addresses that do not represent the places where the vehicle owner lives.
- Errors in entering registration address or assigning them to grid cells.
- Errors in estimating population per grid cell.

To compensate for these problems, we restricted the analysis to grid cells with fewer than 200 miles of driving per person per day. There were 1,406 such cells excluded from the analysis. In total, 100,004 cells had valid data for MPDPP.

## FACTORS RELATED TO AUTO USE

Personal travel is related to the desire to access destinations to acquire goods and services, earn a living, attend school, and conduct all the other functions of life. In most places in America, including most of Massachusetts, it is difficult to conduct all of these activities without access to an automobile, at least some of the time. It is clear from prior research in this area that both demographics and the built environment affect the amount of driving.

In explaining differences in auto use per *person*, we have to consider demographic factors such as the number of children, the number of seniors, the number of households (household size), and household income. All else equal, we would expect driving to decrease as the number of children and seniors increases, and increase as household income and the number of households increases. (Increasing households while holding population constant effectively means that household size decreases; with smaller households there is less ability to share driving.) These and other variables are shown in Table 1. In addition to area household income (HHIncBG), we have two other variables that are correlated with income: percent of housing units that are owner-occupied (OwnPct) and total assessed value of all property in the grid cell (pttlasval).

*Table 1: Description and Source of Variables Used in the Analysis*

| Name | Description | Source |
|---|---|---|
| MPDPP | Mean miles per day per person, based on best matching inspection records. | Mass Vehicle Census, 2010, second quarter and 2010 U.S. Census data as coded to grid by MAPC. |
| ChildPct | Share of children (age 5-17) | 2010 U.S. Census data as coded to grid by MAPC. |
| SeniorPct | Share of seniors (65+) | 2010 U.S. Census data as coded to grid by MAPC. |
| OwnPct | Share of owned homes | 2010 U.S. Census data as coded to grid by MAPC. |
| pop10 | Population based on 2010 U.S. Census | 2010 U.S. Census data as coded to grid by MAPC. |
| hh10 | Households based on 2010 U.S. Census | 2010 U.S. Census data as coded to grid by MAPC. |
| Total_emp | Total employment. | InfoGroup, 2011 |
| pbld_sqm | Rooftop area (square meters) of all buildings in this grid cell, in square meters. | MassGIS Building Rooftops, 2011 |
| prow_sqm | Total area of road and rail rights-of-way within this grid cell. | MassGIS Level 3 Parcel Data, MAPC analysis |
| pttlasval | Total assessed value of all land, buildings, and other improvements in this grid cell, in dollars. Based on local assessing records from 2009 - 2014. | MassGIS Level 3 Parcel Data, MAPC analysis. |
| ppaved_sqm | Estimated paved area, in square meters, excluding roads within public rights-of-way. | MassGIS Impervious Surface Data 2005, MassGIS Building Structures 2010, MassGIS Level 3 Parcels, MAPC analysis. |
| far_agg | Building floor area divided by lot area of fee parcels (non right-of-way). A common measure of density. | MassGIS Level 3 Parcel Data, MAPC analysis. |
| Intsctnden | Count of roadway intersections in this grid cell. | MassDOT Road Layer, MAPC Analysis |
| Sidewlksqm | Linear length of sidewalks in this grid cell. Sidewalks on opposite sides of a roadway are counted separately. | MassDOT Road Layer, MAPC Analysis |
| Schwlkindx | Index of school walkability developed by MAPC, based on number of public, private, or charter school grades within one mile walking distance of grid centroid | MAPC |
| exit_dist | On-road distance to nearest highway interchange | MassDOT Road Layer, MAPC Analysis |
| HHIncBG | Median household income (2008-2012 average) of Census block group in which the grid cell is located. | ACS 2008-2012 block group summary file for Massachusetts. |
| SLD_D4c | Aggregate frequency of transit service within 0.25 miles of Census block group boundary (in which the grid CELL is located) per hour during evening peak period | EPA Smart Location Database, based on GTFS data for MBTA, Massport, and 9 of the 15 RTAs in Massachusetts. |

The built environment affects driving in several different ways:

- *Proximity*: People living next to more potential opportunities (jobs, shopping, parks, etc) can reach these places without driving as far as others who are more isolated, or can reach them by other means (by walking, biking, or using public transit). However, just because people can go to a nearby job or grocery store does not mean they will necessarily choose that option over a preferable one further away. For proximity, we tested measures of population (pop10), employment (total_emp), buildings (pbld_sqm), and building size (far_agg) as well as an index related to the number of nearby schools (schwlkindx). Since the grid cells are all the same size, these figures can be taken as rough measures of the density of population, employment, and buildings.[2]
- *Cost of driving*: the major variable costs of driving that vary over space within Massachusetts (as opposed to over time) are tolls, parking fees, and driver time. The first two only affect a small portion of trips, since roads and parking are generally free. However, traffic congestion and parking scarcity increases the time cost of driving greatly, especially in the older portions of the state that were built with narrower roads and limited off-street parking. Congestion may also be related to the size of the urban area, since in larger metro areas there is more traffic. For the cost of driving, we tested the extent of the road network (prow_sqm), the amount of parking lots and driveways (ppaved_sqm), and the on-road distance to the nearest highway interchange. However, the proximity measures may also be taken as related to the amount of roadway congestions – and thus are also potentially indicators of the cost of driving.
- *Transit accessibility*: Transit requires both a concentration of residences (for economical walk access) and a concentration of employment (or other attractions at the destination end of the trip, since driving egress is generally not feasible). In the U.S. and in Massachusetts, transit does best in locations that had already developed these patterns prior to the widespread suburbanization of people and jobs beginning after the Second World War. For transit accessibility, we tested the aggregate frequency of transit service within 0.25 miles of the Census block group (in which the grid cell is located) boundary per hour during evening peak period (SLD_d4c). This latter measure is imperfect because the missing data includes the bus routes of the Worcester Regional Transit Authority and 5 more of the 15 RTAs.
- *Walk and bike access*: Walking and biking are favored not only by proximity (as discussed above), but also by grid networks that minimize travel distance and provide the possibility of walking or biking along low-volume, low-speed streets. A built-out street front also makes walking safer and more pleasant (compared to walking along buildings fronted by parking lots, for example.) The presence of sidewalks and bike paths may also increase walking and biking; alternatively, these facilities may be installed primarily where there already is the demand for walking and biking. For walk and bike access we tested the number of intersections (intsctnden), which may differentiate older grid-based street networks from rural areas and suburban areas with cul-de-sac networks. We also tested a measure of the number of sidewalks within the grid (sidewlksqm).

## THE MODEL

It is clear that many factors are related to the observed variability in automobile miles traveled over space. Comparing each of these factors one at a time to miles per day per person can be misleading,

[2] The cells do not necessarily have the same amount of land area, however, since they were not drawn with respect to water boundaries. However, total land area per cell was not available.

because many of them vary together. For example, frequent transit service is generally only available in higher density areas, but high density also is expected to affect driving via congestion, parking, proximity, and biking and walking favorability. Thus any observed relationship between transit availability and driving may actually be the result of one of these other related variables. One way to address this problem is to estimate a multivariate model that uses all of these variables simultaneously. Multivariate regression models estimate the independent effect of each variable, controlling for all the others in the model.

The variable we are modeling – miles per day per person – is skewed. Most grid cells are around the median value of 17 miles, but some average much more, up to the cap of 200 miles (values above that threshold were dropped as obviously unreasonable, since this figure represents a daily average, for household vehicles, typically over many months of driving). Because of the long tail of the data, we transformed the dependent variable by taking its natural log (ln), which better captures the non-linear form of the data. Summary statistics for MPDPP and all the independent (explanatory) variables are shown in Table 2.

*Table 2: Summary Statistics of Variables Used in Model*

| Name | Description | Mean | median | min | Max | Std. Dev. |
|---|---|---|---|---|---|---|
| MPDPP | Miles of Driving per Day per Person | 23.7 | 17.3 | 0 | 199.9 | 23.5 |
| ChildPct | Population 5 to 17 as share of total | 2% | 0% | 0 | 66% | 3% |
| SeniorPct | Population 65+ as share of total | 11% | 9% | 0 | 100% | 12% |
| OwnPct | Owned housing units as a percent of all housing units. | 71% | 80% | 0 | 100% | 31% |
| pop10 | Total population, 2010 | 46 | 18 | 1 | 3,867 | 95.14 |
| hh10 | Households, 2010 | 18 | 6.5 | 0.01 | 1,779 | 40.4 |
| total_emp | Employment | 18 | 0 | 0 | 18,422 | 166 |
| pbld_sqm | Building Footprint | 2,717 | 1,597 | 0 | 266,480 | 3,296 |
| prow_sqm | Right of Way Area | 7,104 | 5,162 | -0.07 | 62,490 | 7,445 |
| pttlasval | Total Assessed Value of All Parcels | 6,238,300 | 2,921,200 | 0 | 1,404,500,000 | 18,385,000 |
| ppaved_sqm | Parking Lots and Driveways | 4,392 | 2,860 | 0 | 739,200 | 5,538 |
| far_agg | Floor Area Ratio | 0 | 0.030 | 0 | 14.38 | 0.188 |
| intsctnden | Intersections | 60 | 34 | 0 | 695 | 74 |
| sidewlksqm | Sidewalk Density | 320 | 0 | 0 | 6,556 | 683 |
| schwlkindx | Schools Within a Mile | 0 | 0 | 0 | 9.78 | 0.77 |
| exit_dist | Highway Exit Distance | 5,677 | 7,897 | 3.99 | 57,313 | 7,437 |
| HHIncBG | Median household income of block group | 86,174 | 80,197 | 2,499 | 250,000 | 33,853 |
| SLD_D4c | Peak Transit Frequency per Hour within 1/4 Mile | 26 | 0 | 0 | 3,535 | 124 |

The results of the model are presented in Table 3. As mentioned earlier, only about 100,000 cells had valid data for MPDPP. There were missing values for at least one of the independent variables leaving almost 95,000 grid cells available for the analysis. In a model with so many observations, it is easy to find relationships that pass conventional measures of statistical significance. By the same token, it will be more likely to find the independent significance of variables that are closely related. As long as there are some cases where the miles traveled varies differently with one of these variables compared to another closely related one, it should be possible to estimate separate coefficients for each. In fact, we find that all of the variables included in the model were statistically significant at a high level of confidence except two.

One of these statistically insignificant variables was ChildPct, the share of children in the grid cell. By contrast, SeniorPct was significant and negative. In other words, a higher percentage of seniors is associated with less driving than average (that is, compared to the average for adults 18 to 64), but a higher percentage of children is not. Perhaps children generate enough demands for chauffeuring to compensate for the fact that none but the 16-17 year olds among them are drivers.

Population and household were included in the model after transforming them by taking their natural logarithms. As with the dependent variable, this was necessary because they are highly non-linear: although grid cells are of uniform size, they have a very wide variation in population, with many cells having very few people and a few having very many. The population and household variables work in opposite ways: greater population results in less driving per person, but more households results in less. In other words, as average household size increases, there is less driving per person as driving tasks are shared among household members. The strength of the relationship with population may be the result of the use of the same variable (pop10) to construct the dependent variable. Population may also be measuring proximity effects of higher density areas (such as traffic congestion and inconvenient parking) that are not completely described by the other variables in the model.

The household income (HHIncBG), homeownership (OwnPct), and assessed value (pttlasval) variables, which were also included in log form, were all positively related to driving per person. All of these variables are related to individual income. This finding agrees with previous work which finds that the demand for driving is a “normal” good that increases with increasing income. Household income was measured at the block group level; block groups are larger than grid cells, especially outside of urban areas. Block group median income is only a good indicator of individual income to the extent that populations are sorted by income. Homeownership is related to income in that most low-income people are renters. However, homeownership may also be related to land use in that owned housing is disproportionately single-family and rented housing is disproportionately multi-family and therefore higher-density. The assessed value includes both residential and non-residential property. Thus it may be related not only to household income but the value (and quantity) of non-residential property. People who live next to office towers, for example, would be located in the high end of the range of total assessed value.

Among the hypothesized proximity measures, population (pop10), employment (total_emp), and the number of nearby schools (schwlkindx) were all negatively related to driving per person. However both the measure of the number of buildings (pbld_sqm) and building size (far_agg) were positively related.

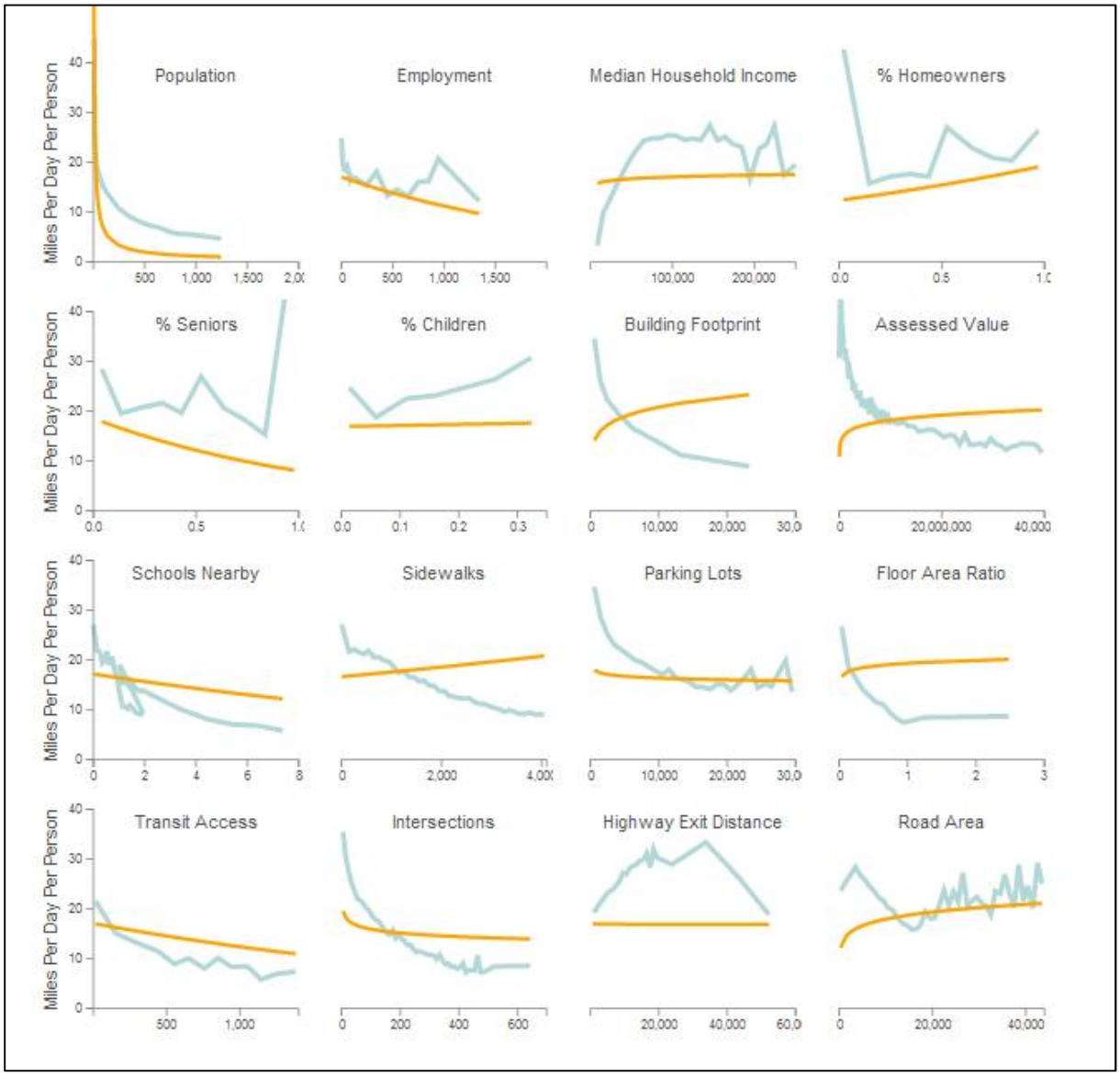
With regard to the cost of driving variables, the extent of the road network (prow_sqm) was positively related to driving, as we would expect. However, the extent of parking lots and driveways (ppaved_sqm) was negatively related. There is no clear explanation for this result. The on-road distance to the nearest

highway interchange was not statistically significant; this may be because being near an interstate makes driving only a little more convenient and because urban areas, which have less driving than rural areas, also have more interstates (and more exits)—essentially counteracting the first effect.
The transit accessibility measure was significant and negative: areas with more frequent transit have less driving, even controlling for population density and other land use factors. The number of intersections was significant and negative; this may be because walking and biking is more convenient in grid-based neighborhoods, and it may also be because average travel speeds are slower in these older neighborhoods. The number of sidewalks (sidewlksqm) was *positively* related to driving, suggesting that more sidewalks are associated with more driving (after controlling for all the other factors included in the model). It is not clear why this is the case.

*Table 3: Model Results*

| Heteroskedasticity-corrected least squares, using observations 1-138,740 (n = 94,758)<br>Missing or incomplete observations dropped: 43,982 | | | | |
|---|---|---|---|---|
| Dependent variable: l_MPDPP | | | | |
| **Variable** | **coefficient** | **std. error** | **t-ratio** | **p-value** |
| Constant | 1.231 | 0.094 | 13.031 | <.001 |
| ChildPct | 0.127 | 0.093 | 1.371 | 0.170 |
| SeniorPct | -0.845 | 0.024 | -35.840 | <.001 |
| OwnPct | 0.448 | 0.011 | 41.120 | <.001 |
| l_pop10 | -0.758 | 0.009 | -87.629 | <.001 |
| l_hh10 | 0.430 | 0.009 | 48.452 | <.001 |
| total_emp | -0.00042 | 0.00002 | -20.979 | <.001 |
| l_pbld_sqm | 0.136 | 0.007 | 18.545 | <.001 |
| l_prow_sqm | 0.110 | 0.003 | 34.179 | <.001 |
| l_pttlasval | 0.078 | 0.005 | 15.495 | <.001 |
| l_ppaved_sqm | -0.032 | 0.004 | -7.944 | <.001 |
| l_far_agg | 0.044 | 0.004 | 10.522 | <.001 |
| l_intsctnden | -0.072 | 0.003 | -21.009 | <.001 |
| Sidewlksqm | 0.000 | 0.000 | 19.457 | <.001 |
| Schwlkindx | -0.046 | 0.002 | -19.027 | <.001 |
| l_exit_dist | -0.001 | 0.003 | -0.477 | 0.633 |
| l_HHIncBG | 0.032 | 0.006 | 5.361 | <.001 |
| SLD_D4c | 0.000 | 0.000 | -22.899 | <.001 |
| Statistics based on the weighted data | | | | |
| Sum squared resid | 414427.4 | S.E. of regression | 2.091498 | |
| R-squared | 0.272185 | Adjusted R-squared | 0.272054 | |
| F(17, 94740) | 2084.14 | P-value(F) | 0 | |
| Log-likelihood | -204367 | Akaike criterion | 408769.8 | |
| Schwarz criterion | 408940 | Hannan-Quinn | 408821.6 | |

The following figure shows a comparison of the independent variables, one at a time, with the dependent variable. The blue line in each plot represents the actual data, averaged over specified intervals. The orange line represents the best estimate from the regression model, *holding all other factors constant.* In the figures on the first line, the predicted values are roughly similar to the actual ones. In some of the other cases, however, there is a wide divergence. In the case of building footprint, assessed value, and sidewalks, the slope of the line changes directions. The explanation is that the apparent relationship seen in the raw data does not hold up after controlling for all the other factors.



One way of comparing the relative influence of the different variables in this regression model is to calculate the *elasticity* of each one with respect to the dependent variable, Miles per Day per Person. Elasiticity is defined as the percent change in Miles per Day per Person that results from a 1% change in the independent variable. It is a ratio and has no units. An elasticity that is larger in absolute value than another one means that the factor of interest (Miles per Day per Person) is more sensitive to changes in that variable. The elasticities derived from the model estimated above are shown in Table 4. Three of them stand out as significantly higher than the rest: owner-occupied housing share, population, and households. It should be remembered that population and households are inevitably linked. Since they have opposite signs, they tend to cancel each other out. However, since the population elasticity is greater in absolute value, all else equal, increasing average population density will still have a large impact on reducing driving. Increasing the share of owner-occupied housing would tend to increase driving. As mentioned previously, this variable has both an income effect (since homeowners tend to have higher than average income) and a spatial effect (since owned housing is most likely to be single-family housing).

All of the remaining variables have elasticities that are less than 0.15 in absolute value. After controlling for everything else, they appear to have a measurable, but small, impact on driving. The most notable one is intersections (which represents intersection density, given that the grid cells are the same size). This association could be because areas with a grid of streets are more conducive to walking, biking, and transit, and it could also be because such areas have narrow streets and limited off-street parking, making driving slower and more expensive.

*Table 4: Elasticities of Explanatory Variables with Respect to Miles per Day per Person*

| Name | Description | Elasticity* |
|---|---|---|
| ChildPct | Population 5 to 17 as share of total | 0 |
| SeniorPct | Population 65+ as share of total | -0.09 |
| OwnPct | Owned housing units as a percent of all housing units | 0.32 |
| ln_pop10 | Total population, 2010 | -0.76 |
| ln_hh10 | Households, 2010 | 0.43 |
| total_emp | Employment | -0.01 |
| ln_pbld_sqm | Building Footprint | 0.14 |
| ln_prow_sqm | Right of Way Area | 0.11 |
| ln_pttlasval | Total Assessed Value of All Parcels | 0.08 |
| ln_ppaved_sqm | Parking Lots and Driveways | -0.03 |
| ln_far_agg | Floor Area Ratio | 0.04 |
| ln_intsctnden | Intersections | -0.07 |
| sidewlksqm | Sidewalk Density | 0.02 |
| schwlkindx | Schools Within a Mile | -0.01 |
| ln_exit_dist | Highway Exit Distance | 0 |
| ln_HHIncBG | Median household income of block group | 0.03 |
| SLD_D4c | Transit Frequency within 1/4 Mile | -0.01 |
| *Calculated at the sample mean for independent variables not in log form. | | |