Vehicle Census of Massachusetts
Documentation and Data Dictionary – v.1
March 10, 2014

Contributors:
Tim Reardon, MAPC Assistant Director of Data Services
Susan Brunton, MAPC Data Services
Meghna Dutta, MAPC Data Services
Holly St. Clair, MAPC Director of Data Services
Paul Reim, Central Transportation Planning Staff
Ken Gillingham, PhD, Yale University School of Forestry and Environmental Studies
Eric Minikel, IBI Group
Benedict Holland, Clark University

The Vehicle Census of Massachusetts is a catalog of information about vehicle registered in the Commonwealth from 2008 to 2011. It is a valuable new resource for everyone seeking to understand the key factors that influence auto ownership patterns and miles driven, and will help public agencies and communities in their efforts to build a more efficient and sustainable transportation system in the Commonwealth. The Vehicle Census combines information from vehicle registrations, inspection records, mileage ratings, and other sources to document the ownership and mileage history of each vehicle, which is presented in two different data formats designed to protect the privacy of individual owners. Massachusetts will be the first state in the nation to publish a dataset with this level of detail, so this release has significance for researchers and policy makers across the country. The Vehicle Census of Massachusetts was created by the Metropolitan Area Planning Council, in partnership with the Massachusetts Registry of Motor Vehicles, with assistance from Central Transportation Planning Staff and Yale University, and with the generous support of the Barr Foundation.

The Vehicle Census is presented in two different formats:

- A *vehicle-level dataset* with a history for each vehicle, including inspection and transaction dates, zip code, mileage estimates, fuel economy ratings, estimated fuel consumption, make, model, year, and MSRP, and over the course of the study period (2008- 2011.) To protect driver confidentiality, anonymous registration and vehicle identification numbers (VINs) have been substituted for the actual plate number and VIN. No address information or other personal information is included in the data.
  (*16 million records*)
- A *grid cell dataset* that estimates the number of registered vehicles, number of households, average mileage per vehicle, mileage per household, and estimated fuel usage in 250-meter grid cells statewide, at a given point in time. This snapshot, taken every three months, allows for temporal analysis of registrations and driving patterns. No vehicle-level data will be included in this dataset, only summary statistics of all the vehicles geocoded to the grid cell.
  (*5.6 million records*)

Michelle Ciccolo, *President* • Lynn Duncan, *Vice President* • Marilyn Contreas, *Secretary* • Taber Keally, *Treasurer* • Marc Draisen, *Executive Director*

Core Data Sources

The Vehicle Census is based on two administrative datasets maintained by the Massachusetts Registry of Motor Vehicles, a division of the Massachusetts Department of Transportation. The Automated License and Registration System (ALARS) contains the owner, license plate, and address information for each registered vehicle and is updated each time there is a 'transaction,' including transfer of title (purchase or sale), cancellation (junked vehicle), registration renewal, or change of address. A separate database contains records from safety and emissions inspections conducted as part of the state's Vehicle Check program. Vehicles are required to be inspected annually and within seven days of sale. The date of inspection, VIN, and odometer reading are all recorded on the inspection record. Comparison of odometer readings recorded during successive inspections can be used to estimate average daily mileage driven by the vehicle during the intervening period.

Data Processing

MAPC and its partners conducted extensive processing and modification of the raw RMV data to create the vehicle census in a form that could be publicly shared. The sections below describe this process.

Creating the "Registrations" Table

Registration transaction records for 2008 through 2011 were processed and geocoded by the Central Transportation Planning Staff to create a registration history for each vehicle, which contains a distinct record for each combination of VIN, registration ID (license plate), and address. Each transaction record includes the date and nature of the transaction, the beginning date of the 2-year registration period (same as previous if renewal or new registration), and the end date of the 2 year registration period, as well as the effective address. Records created as a result of registration renewal, with no change in address or plate ID, were merged to create a continuous record that spans the registration date. The earliest available transaction records were received from MassDOT in 2008, and for vehicles with a new or renewed registration in that year there may be no information about the previous registration (if any). *As a result, the 2008 records should be considered an incomplete inventory of registered vehicles*. By 2009, the volume of transactions is sufficient to allow construction of a complete dataset.

Registration addresses were geocoded to X-Y locations using a combination of NAVTEQ and TIGER address information. Of approximately 10 million unique addresses in the registration data, 90% were geocoded. Approximately 2.6% of the records contain no street address (P.O. Box only) and an additional 4.2% are registered to out-of-state addresses. Each geocoded record was assigned to a 250 meter grid cell. Neither addresses, X-Y coordinates, nor grid cell ID is included in the public version of the dataset. Non-geocoded vehicles were inspected to determine whether they contained a valid garaging zip code or municipal ID that could be used for spatial analysis.

Creating Mileage Estimates

Odometer readings from vehicle inspection records (from 2005 to the beginning of 2012) and the dates of the inspections were compared to calculate the mileage driven in the intervening period and the average daily mileage. In many cases, input mistakes or other errors result in lower odometer readings being recorded in later inspections. To increase the number of valid records, MAPC used a 'triplets' method to extract the best estimate from three successive inspection records, with a bias toward discarding the odometer readings that would overestimate mileage (e.g., if reading at Time A was 5,000 miles, Time B was 15,000 miles, and Time C was 10,000 miles, the method would use readings A and C.) The result was a series of mileage estimates for each VIN with an inspection date at the beginning and end, the odometer reading at the beginning inspection, and estimated daily mileage during the intervening period.

Registrations and Estimates Table

The mileage estimates were intersected with the registration data to create a new "Registrations and Estimates" table that includes a distinct record for each combination of VIN, registration ID, address, and mileage estimate. Registration records were split where a mileage estimate begins or ends. Registration periods without a corresponding mileage estimate are retained and assigned a "false" value for the insp_match field. The temporal overlap between the mileage estimate and the registration record is compared to the length of the mileage estimate period as a measure of data reliability. High values for this "Percent Overlap" field mean that the vehicle had the same owner and was garaged in the same location for a large portion of the mileage estimate period; a low value means that a substantial portion of the estimated mileage may have been driven while the vehicle was owned by another person or garaged in a different location. To support temporal analysis, MAPC created a set of sixteen Boolean fields that indicate whether a given record was valid at a specified point in time; in this case, we used the median day of each calendar quarter from January 2008 to December 2011. MAPC commonly uses the second quarter of 2010 as a reference point for summary statistics, since the registration and inspection datasets are most complete at this point and vehicle information can be related to contemporaneous census population and household counts.

Vehicle Characteristics

With assistance from researchers at Yale University, MAPC assigned vehicle characteristics using information from a commercially-available vehicle database augmented with additional values researched by Yale staff. Available vehicle attributes include make, model, model year, MSRP, curb weight, vehicle type (a composite field developed by Yale researchers), fuel type, flags for hybrid vehicles and motorcycles, and estimated fuel efficiency (miles per gallon, or MPG) based on U.S. Environmental Protection Agency economy ratings issued in 2008. Because fuel efficiency declines with vehicle age and mileage, MAPC calculated an adjusted MPG rating based on the odometer reading at the start of the mileage estimate, using efficiency decay factors from scientific and engineering literature. Adjusted MPG and estimated daily mileage were combined to generate estimated daily fuel consumption and associated greenhouse gas emissions (CO2 equivalents) based on the GHG density of the associated fuel type.

Vehicle History Table (Public Version)

After assigning vehicle characteristics, MAPC created a publicly-accessible version of the dataset, called the Vehicle History table. The registration ID, X-Y coordinate, grid cell ID, and zip +4 fields were all removed from the table, leaving the zip code and municipality as the most specific level of geographic detail. It is important to note that some zip codes have no geographic "footprint"—they are associated with post office boxes and not street addresses.

MAPC took additional steps to prevent individual vehicles from being identified on the basis of their make, model, and year. At any given point in time, approximately 15% of registered passenger vehicles have a make, model, and model year that is unique within the zip code where they are garaged (e.g., only one 2008 Toyota Camry in zip code 02043.) To prevent such vehicles from being identified only on the basis of their make, model, and year, MAPC suppressed the zip code and municipal ID of these passenger vehicles for the period when they would be considered unique. The vehicle information is retained, but there is no location information for the suppressed vehicle record. MAPC also found 109 one-of-a-kind vehicles that remained unique, even after suppressing the zip code; the make, model, and year of these vehicles was suppressed to prevent individual vehicles from being identified in the dataset.

As a result of the "no-footprint" zip codes and data suppression, the Vehicle History records within a given zip code or municipality cannot be considered to be a complete inventory of all registered vehicles. This table can be used to assess vehicle characteristics and mileage patterns within a municipality or zip code, but analysis that requires a complete count of vehicles should use the Grid Cell allocation, described below.

Allocation to Grid Cells

Information from the Registrations and Estimates table was summarized into a statewide layer of 250 meter grid cells. This grid cell framework, first developed by MassGIS, can be populated with a variety of statistics developed by MAPC, including estimated 2010 population and households (based on Census 2010 block-level counts), business establishments, and other data. Each grid cell is represented sixteen times, once for each of the sixteen calendar quarters described above. The Registrations and Estimates table was sampled on the median day of each quarter and the results were used to populate the associated record in the grid data.

For passenger vehicles, MAPC summarized the total number of geocoded vehicles in each grid cell, the number of geocoded vehicles with valid mileage estimates (between zero and 200 miles per day), and the number of geocoded vehicles with the "best" mileage estimates (more than 95% of the mileage estimate is associated with the same owner and address.) We also calculated the average daily mileage for the latter two classes of vehicles; the "best" vehicles used a straight average; for all valid estimates we weighted each record by the percent inspection days so that the least reliable estimates have relatively less impact on the resulting average. The distribution of mileage is also represented in five fields that count the number of "best" estimates within specified ranges representing the quartiles and top decile of mileage for all vehicles.

Since not all vehicles could be geocoded, MAPC allocated vehicles with only zip code or garage town information to each grid cell *pro rata* based on that cell's share of the total households in the zip code or municipality. The number of geocoded passenger vehicles as well as those assigned based on zip code or municipality were summed to estimate the total number of passenger vehicles garaged in that grid cell. The number of commercial vehicles assigned to each grid cell was also summed up, and un-geocoded commercial vehicles were allocated to grid cells *pro rata* based on each cell's share of the total geocoded commercial vehicles in the zip code or municipality. Commercial VMT is summarized in a single field which is the average of all valid mileage estimates for geocoded commercial vehicles, weighted by the percent inspection days. An estimate of total registered vehicles includes both geocoded and assigned passenger and commercial vehicles.

MAPC estimated an "effective MPG" for the grid cell based on the aggregate mileage and fuel consumption of all passenger vehicles with valid mileage estimates, and used that figure to estimate daily fuel consumption and GHG emissions associated with vehicles in the grid cell. To prevent users from extracting information about specific vehicles in grid cells with a very small number of households and vehicles, MAPC suppressed the passenger mileage estimates and fuel consumption estimates for all grid cells with fewer than two households and fewer than five geocoded passenger vehicles.