

Description & Data Dictionary for PAYD Analytic Dataset

This document provides the Data Dictionary for the 'PAYD Analytic Dataset' of Massachusetts automobile mileage, fuel economy, and accident costs that were used in the PAYD study, "Pay-As-You-Drive Auto Insurance in Massachusetts: A Risk Assessment and Report on Consumer, Industry, and Environmental Benefits," by MIT Prof. Joseph Ferreira, Jr. and Eric Minikel for the Conservation Law Foundation and Environmental Insurance Agency, Nov. 2010.¹ The dataset was developed from the DVD of raw data released in March, 2010, by the Massachusetts Executive Office of Energy and Environmental Affairs (EOEEA). That DVD contained 1.3 GB of compressed data from the state's Commonwealth Automobile Reinsurer (CAR) and the Registry of Motor Vehicles (RMV). The CAR data included several years of auto insurance policy and claims transaction records for all insured private passenger vehicles in the state. The RMV data included odometer readings from all state-mandated annual safety inspections of all private passenger vehicles. Public notice of the availability of these data is posted on the EOEEA website at:

http://www.mass.gov/?pageID=eoeaterminal&L=3&L0=Home&L1=Grants+%26+Technical+Assistance&L2=Data+Resources&sid=Eoea&b=terminalcontent&f=eea_data-resources_2005-08-auto-insur-data&csid=Eoea

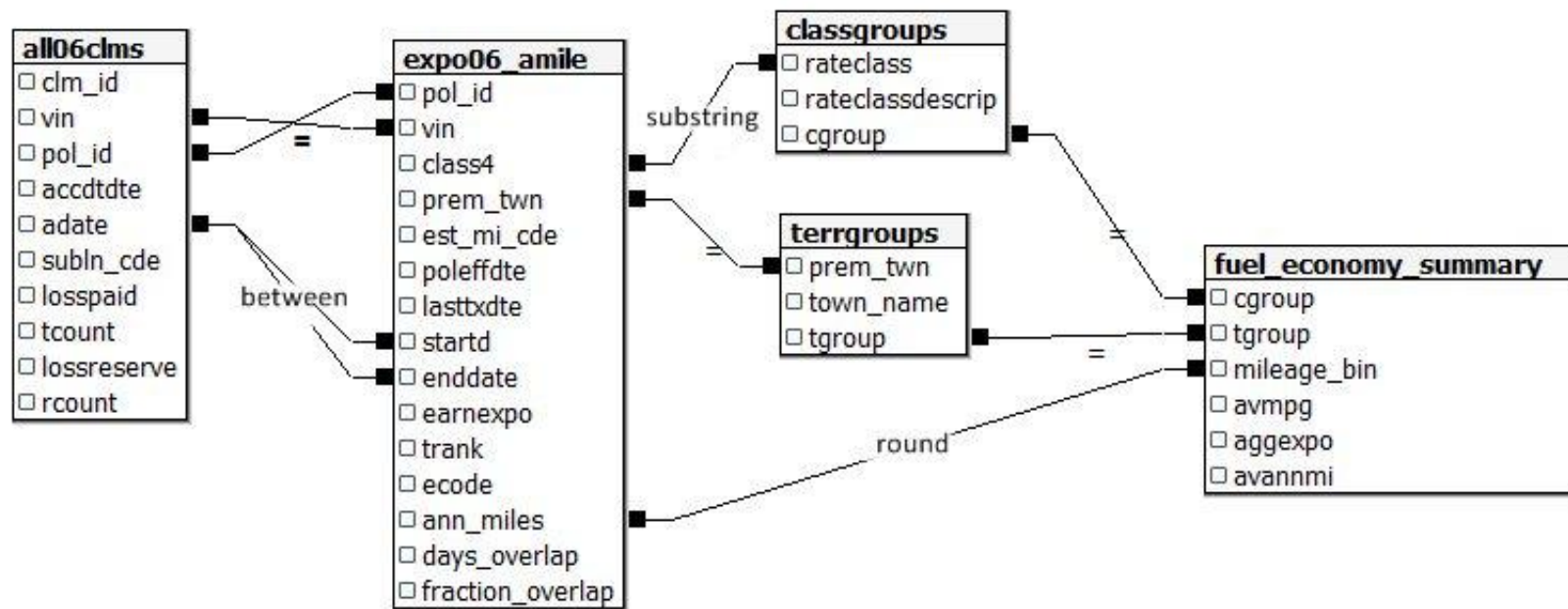
Considerable processing was required to convert the raw policy transaction data from CAR into earned exposure records that matched each vehicle to an appropriate rating class, territory and mileage estimate during each insured month. Additional processing was also required to convert the raw claims transaction data into net paid losses and reserves (as of December 2008) and then match each claim (for bodily injury and property damage liability or personal injury projection) to the appropriate vehicle and earned exposure month. This processing of the raw CAR data was done for all policy and claim transactions involving vehicles insured during policy year 2006. The processing was done by Prof. Joseph Ferreira at the Massachusetts Institute of Technology (MIT) through the partial support of University Transportation Center Region One research grant MITR22-5.

The PAYD analytic dataset also includes annual mileage estimates for each policy year 2006 vehicle and fuel economy estimates for each class, territory, and mileage category. The DVD from EOEEA also contained annual mileage estimates developed by MassGIS from the safety inspection records. However, we redid the mileage estimates (using the same safety inspection records) in order to base each vehicle's estimate on the pair of inspections that had the largest overlap with that vehicle's 2006 policy year. The fuel economy estimates for each class, territory, and mileage category are averages of the adjusted EPA estimates (reported by VINquery.com) for each individual vehicle (out of the 3 million in our study) that fell into each class, territory, and mileage category. The mileage (re)estimation and fuel economy processing was done by the authors as part of this PAYD study.

A downloadable zip file with the **PAYD Analytic Dataset** is available from this MIT website: <http://mit.edu/jf/www/payd>. The tables included in the PAYD Analytic Dataset are explained in the following entity-relationship diagram and table definitions:

¹ This document is the same as Appendix III of the PAYD report. The full report is available from the Conservation Law Foundation website: <http://www.clf.org/PAYD>.

Entity-Relationship Diagram for PAYD Analytic Dataset



all06clms

Summarizes claims data, including total paid losses and outstanding reserves, for all claims that matched an earned exposure period during the 2006 policy year.

681,423 records.

Primary key: vin||pol_id||clm_id||adate||subln_cde

clm_id	integer	Claim ID, anonymized by MassGIS
vin	varchar(17)	Vehicle Identification Number (VIN)
pol_id	varchar(14)	Insurance policy ID, anonymized by MassGIS
accdtdte	varchar(8)	Accident date in string format (YYYYMMDD)
adate	date	Accident date in date format
subln_cde	character(1)	Subline code: 1 = Liability (property damage and bodily injury) 5 = No fault (personal injury protection)
losspaid	bigint	Net total losses (and loss adjustment expenses) paid through 2008 Q4 for each unique claim ID and subline code. 2433 out of 625632 paid claim records are capped at \$25000. 71,829 records in this table have net losspaid+lossreserve <= \$50 (e.g., after subrogation) and so were excluded from frequency and pure premium calculations.
tcount	integer	Number of payment transactions related to this claim
lossreserve	bigint	Outstanding reserves as of 2008 Q4 for each unique claim ID and subline code. 2568 of 36635 records with outstanding loss reserves are capped at 25000. 71,829 records in this table have net losspaid+lossreserve <= \$50 (e.g., after subrogation) and so were excluded from frequency and pure premium calculations.
rcount	integer	Number of transactions updating the reserve amount for this claim

expo06_amile

Earned exposure months and annual mileage estimates for each period of consistent policy endorsement conditions during policy year 2006 for 3.25M policy-vehicle combinations.

3,991,012 records.

Primary key: vin||pol_id||poleffdte||startd

pol_id	varchar(14)	Insurance policy ID, anonymized by MassGIS
vin	varchar(17)	Vehicle Identification Number (VIN)
class4	character(4)	Insurance rating class as defined in Code Descriptions.pdf
prem_twn	character(3)	3-digit code indicating the garage 'town' for insurance purposes. Includes 360 towns and neighborhoods in Massachusetts plus codes for out-of-state policies. Described in PremAccdtTownTables.pdf.
est_mi_cde	character(3)	Code for driver-reported annual mileage category described in Code Descriptions.pdf. This data comes from Commonwealth Automobile Reinsurers, not from our own mileage estimates.
poleffdte	varchar(8)	Year and month (YYYYMM) of policy effective date.
lasttxdte	varchar(8)	Year and month (YYYYMM) of last transaction modifying policy.
startd	date	Estimated start date for the period during which the policy endorsement conditions reported in this record are earned. (This is different from the policy effective date if an endorsement occurred).
enddate	date	Estimated end date of the period during which the policy endorsement conditions reported in this record are earned. (This is different from the policy end date if an endorsement occurred).
earnexpo	numeric(3,0)	The number of months of 'earned exposure' for the policy conditions reported in this data record.

trank	integer	Number of transactions modifying policy during the policy year.
ecode	integer	Indicates what type of adjustment to start and end dates was necessary in order to reconcile earned exposure months with transaction dates and policy effective date. 0: No endorsements, no adjustment needed 10: Endorsements made but no adjustment needed 22: End date moved earlier by 17 days 24: End date moved later by 14 days
ann_miles	numeric	Estimated annual miles traveled by the vehicle. New estimates developed from odometer readings in RMV safety inspection data from all_rmv_insp_06.txt, etc. 0 to 100,000 -1 = no mileage estimate available
days_overlap	integer	Number of days of overlap of the policy effective period and the two odometer inspections used to create the annual mileage estimate. A measure of the quality of the estimate. 0 to 366 -1: no mileage estimate available
fraction_overlap	numeric	The fraction of the policy effective period that is overlapped by the period between two odometer inspections used to create the annual mileage estimate. A measure of the quality of the estimate. 0.00 = No overlap ... 1.00 = Full overlap -1 = no mileage estimate available

terrgroups

Lookup table matching insurance 'towns' to six territory groupings.

360 records. Primary key: prem_twn

prem_twn	character(3)	3-digit code indicating the garage 'town' for insurance purposes. Includes 360 towns and neighborhoods in Massachusetts. Described in PremAccdtTownTables.pdf.
town_name	varchar(30)	Name of town or neighborhood.
Tgroup	character(1)	Six territory groupings: '1' = Least risky '2' '3' '4' '5' '6' = Most risky

classgroups

Lookup table matching the last character of the four-digit rating class code (class4) to the five class groupings.

9 records. Primary key: rateclass

Rateclass	character(1)	Indicates the rating class of the insurance policy. Corresponds to the fourth digit of class4 in expo06_amile. '1' through '9'. Detailed descriptions stored in rateclassdescrip.
rateclassdescrip	varchar(42)	Description of the rate class
Cgroup	character(1)	Five class groupings: 'A': Adults 'B': Business 'I': <3 yrs experience 'M': 3-6 yrs experience 'S': Senior citizens

fuel_economy_summary

Summarizes vehicle fuel economy data, aggregating vehicles into groups by class, territory and annual mileage bin.

5,170 records.

Primary key: cgroup||tgroup||mileage_bin

cgroup	character(1)	Five class groupings: 'A': Adults 'B': Business 'I': <3 yrs experience 'M': 3-6 yrs experience 'S': Senior citizens
tgroup	character(1)	Six territory groupings: '1' = Least risky '2' '3' '4' '5' '6' = Most risky
mileage_bin	integer	Vehicle's annual mileage by 500-mile bin
avmpg	numeric	Average fuel economy of vehicles in this group. Each vehicle's fuel economy is considered to be 88% of the simple average of its city and highway fuel economy. Average fuel economy for all vehicles in the group is then weighted by aggregate miles traveled (i.e. annual mileage times months of exposure)
aggexpo	numeric	Aggregate car years of exposure for vehicles in this group.
avannmi	numeric	Average annual miles traveled for vehicles in this group.