Capturing the Built Environment - Travel Connection for Strategic Planning: Development of A Multi-Modal Travel Module for VisionEval

Research brief submitted for presentation at ITM 2018

Liming Wang$^{1a}$, Huajie Yang$^{1}$, Brian Gregor$^{2}$, and Tara Weidner$^{3}$

$^{1}$Portland State University $^{2}$Oregon Systems Analytics LLC $^{3}$Oregon Department of Transportation $^{a}$ lmwang@pdx.edu

# Abstract

There are now a wide spectrum of integrated land use and transportation models spanning from simple sketch planning models to complex microsimulation models. The Regional Strategic Planning Model (RSPM) aims to balance the flexibility and realism of microsimulation and the speed and interactivity of simple models by taking a microsimulation approach while making other simplifications in order to model first-order effects quickly. While recent efforts on creating a unified modeling framework, VisionEval, for RSPM and a suite of other GreenSTEP family of models are documented elsewhere, this paper focsuses on the innovations of a multi-modal travel demand module developed for VisionEval - VETravelDemandMM. In the development of this new module, we follow the best practices for model development recommended in the literature. In particular, we address the uncertainty and validity in our models by going through rigorous cross validation and model selection (in addition to variable selection). We also use a unique high resolution nationwide dataset that combines 2009 National Household Travel Survey, EPA's Smart Location Database, and regional roadway and transit services information. Due to space limit, this paper presents the results of model estimation, validation, model selection, and sensitivity tests for the VMT model, before concludes with an assessment of the innovations of VETravelDemandMM over alternative approaches and a discussion of the implications for travel model development and applications.

Word count: 2,243 + 1 table x 250 = 2,493

##### `

# Objectives and Motivations

There are now a wide spectrum of integrated land use and transportation models spanning from simple sketch planning models to complex microsimulation models. The Regional Strategic Planning Model (RSPM) aims to balance the flexibility and realism of microsimulation and the speed and interactivity of simple models by taking a microsimulation approach while making other simplifications in order to model first-order effects quickly.

While the development of VisionEval is documented in other papers (Gregor, 2018; Wang, Gregor, Yang, Weidner, & Knudson, 2018, Flyte & Raw (2018)), this paper focuses on describing the innovations in the multi-modal travel demand module developed for VisionEval - VETravelDemandMM. The purpose of the VETravelDemandMM module is to capture the effects of household characteristics, built environment, and regional transportation system on travel for strategic planning. It works along with other modules in RSPM, including Household Synthesis, Land Use Characteristics, Transportation Supply, Household Vehicle Ownership, and Household Vehicle Travel Budgets, to enable the assessment of a range of policy and technoligy scenarios in the domain of land use and transportation.

In the development of this new module, we follow the best practices recommended in the literature to address many of the challenges in the development and applications of integrated models (Waddell, 2011). Compared with similar models, the VETravelDemandMM incorporates the following innovations, which we will elaborate further in later sections:

* Utilizing a novel dataset of highly detailed information of households and built environment with a US nationwide coverage;
* Modeling annual average daily Vehicle Miles Traveled (AADVMT), instead of the VMT during the day of the survey commonly used in similar models;
* Use of cross validation in model development to avoid overfitting;
* Conducting model selection, in addition to variable selection, in the model development process to select the best model structure among a handful of options;
* Implemented in the popular statistical programming language R as a self-contained package for VisionEval, which contains documents and functions for model (re-)estimation and prediction;
* The model development process, from data processing, model estimation, validation and testing, to report and document compiling, follows the practice of reproducible research; results and products derived from our work, including this paper, is fully reproducible by others (sans confidential data).

The effects of social-economic characteristics, built environment, and transportation supply on multi-model travel are a well researched topic (Ewing & Cervero, 2010). VMT, Person Miles Traveled (PMT), trip frequency and length are the most common travel outcomes modelled (Ewing & Cervero, 2010). There are more than a dozen model structures (types) used in the literature, although a handful of them are most most common (Ewing & Cervero, 2010). Authors often adopt one model structure based on idiosyncrasy or theroetical rationale; rarely more than a couple of model structures are applied and benchmarked against each other on the same dataset. For our module, we model VMT for driving, and PMT, trip frequency and length for transit, bike, and walk travel (due to space limit, only the VMT model is discussed in detail in this paper). We test multiple common model structures and select the best with cross validation. Another innovation is that we model long term VMT in the form of AADVMT, instead of VMT during the day of the survey commonly used in similar models. For the purpose of strategic planning, which focuses on assessing long term effects for various inputs, long term VMT works better.

# Methodology and Data

In this section, we discuss the data, model structure options, and our approach to cross validation and model selection.

## Data

We create a unique high resolution nationwide dataset for model estimation, validation and testing.

### NHTS

The 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation, Federal Highway Administration, 2009) collected trips taken by surveyed household members, as well as their socio-demographic characteristics. In addition, the 2009 NHTS also includes vehicle odometer readings. Based on odometer readings and other information, Oak Ridge National Laboratory (ORNL) imputed and validated annual miles driven for each vehicle (Oak Ridge National Laboratory, 2011).

In addition, we get access the confidential residence Census Block Group in the NHTS. This information allows us to join household characteristics and travel with the Smart Location Database to create a unique nationwide travel dataset with rich built environment measures.

### Smart Location Database

The Smart Location Database (SLD) is a nationwide database with more than 90 built environment measures organized by 5D - Density, Diversity, Design, Destination, Distance to transit. Most attributes are available for every Census block group in the United States (Ramsey & Bell, 2014).

### Regional Transportation Supply

To supplement the block group level transportation supply measures in SLD, we add urbanized area (UZA) level transportation supply information, including freeway lane-miles and annual transit vehicle revenue miles, from Highway Performance Monitoring System (HPMS) and National Transit Database (NTD).

Since transit service related measures from SLD and NTD are only available for UZAs, we segment our dataset similar to Gregor (2015): a UZA segment with complete information and a non-UZA segment for which transit service related measures are missing.

After joining these 4 datasets, we have a household-level dataset with about 200 variables. A summary table for the variables is available in our project report (Wang, 2017).

## Structure of the AADVMT Model

Like other models in RSPM, all models in the VETravelDemandMM module take the microsimulation approach and model multi-modal travel for individual households. VMT/PMT for each of the 4 modes: driving, transit, bike, and walk, are modeled separately. The household travel budget module of RSPM captures the cross-elasticites among these modes. Below we present the AADVMT model as an example to demonstrate our model development process. The process for developing other models for non-driving modes are identical and the results are available in our project report (Wang, 2017).

For the AADVMT model, we first calculate AADVMT for each household in the NHTS from the annual miles driven of each vehicle in the household. Computed Household AADVMT is then regressed on independent variables, including household socio-economic status (SES), built environment, and regional transportation supply:

$$\begin{matrix}AADVMT\_{h}=f(&household SES, built environment,\\&regional transportation supply).\end{matrix}  (1)$$

We consider 5 model structure options ($f(.)$ in Equation (1)) for VMT models common in the literature (Ewing & Cervero, 2010): a linear regression model, a non-linear regression model, a hurdle model, a zero-inflated negative binomial model, and a 2-step model of binomial logit and linear/non-linear regression model. Models for each of the structures are fit to our data through a variable selection process, taking into consideration behavioral validity, model goodness-of-fit and statistical significance. In the specification for each of the options, we control for household SES before add built environment and transportation supply variables. We aim to include at least one variables from each of the 5D categories and avoid variables with high correlation (measured by variance inflation factor). After we identify the best model specification for each model structure, the best model structure is selected from the 5 options through cross validation.

## Cross Validation and Model Selection

We make model selection decision using k-fold cross-validation. k-fold cross-validation first randomly partitions a sample evenly into $k$ subsamples. Then in $k$ iterations, each one of the $k$ subsamples is reserved for cross-validation (testing) in turn, while the remaining $k-1$ subsamples are combined and used for estimation (training). k-fold validation is an efficient approach for cross-validation with low variance (Hastie, Tibshirani, & Friedman, 2016). We select the model structure with the best prediction performance in cross-validation.

# Results

We compute the prediction accuracy measured by root mean squared error (rmse) for each of 5-fold cross-validation for each model structure and then average them over the 5 folds to get the average accuracy for each model structure. The power-transformed model and the 2-step models have the best accuracy (rmse = 31.0 and 31.1, respectively while other model structures have rmse ranging from 32.2 to 34.3) in cross-validation (The k-fold cross validation results are available in Wang (2017).). We choose the power-transformed model based on the pricinple of simplicity, easy of use, and computational performance. The power parameter is determined via a Box-Cox transformation (Box & Cox, 1964): $λ=0.38$.

Table 1 presents the final estimation results for the power-transformed model. All the coefficients have expected sign. We try our best to capture the non-linearity and interaction effects among independent variables. For example, for the VehPerDriver variable, we use a cubic spline on VehPerDriver to capture the non-linear effect of households' vehicle ownership on driving. The k-fold cross-validation helps ensure that specifications with such non-linear transformations do not overfit the sample.

|  |
| --- |
| **Table 1 Estimation results for the power-transformed non-linear model** |
|  |
|  | AADVMT0.38 |
|  |  |
|  | UZA | Non-UZA |
|  |
| #Drivers | 0.719\*\*\* (0.011) | 0.755\*\*\* (0.011) |
| Household Size | 0.004 (0.008) | 0.018\*\* (0.009) |
| #Workers | 0.178\*\*\* (0.008) | 0.167\*\*\* (0.007) |
| LogIncome | 0.250\*\*\* (0.007) | 0.299\*\*\* (0.006) |
| #Persons age 0-14 | 0.095\*\*\* (0.011) | 0.097\*\*\* (0.011) |
| #Persons age 65+ | -0.066\*\*\* (0.008) | -0.073\*\*\* (0.007) |
| Vehicles per drivers natural spline: linear | 2.770\*\*\* (0.047) | 2.730\*\*\* (0.043) |
|  quadratic | 5.870\*\*\* (0.192) | 5.600\*\*\* (0.171) |
|  cubic | 2.950\*\*\* (0.208) | 3.530\*\*\* (0.173) |
| Life Cycle (ref=Couple w/o children):  Empty Nester | -0.227\*\*\* (0.016) | -0.188\*\*\* (0.015) |
|  Parents w/ children | 0.034\* (0.017) | 0.019 (0.017) |
|  Single | -0.186\*\*\* (0.020) | -0.176\*\*\* (0.020) |
| Census Tract Population Density | -0.028\*\*\* (0.010) | -0.039\*\*\* (0.013) |
| Total Employment within 5-mile buffer of residence | -0.057\*\*\* (0.005) | -0.037\*\*\* (0.003) |
| Census District: East South Central | 0.079 (0.054) | 0.087\*\*\* (0.030) |
|  Middle Atlantic | -0.108\*\*\* (0.027) | -0.168\*\*\* (0.022) |
|  Mountain | -0.084\*\*\* (0.026) | -0.111\*\*\* (0.029) |
|  New England | -0.130\*\*\* (0.045) | -0.025 (0.030) |
|  Pacific | -0.077\*\*\* (0.023) | -0.202\*\*\* (0.024) |
|  South Atlantic | 0.018 (0.023) | 0.030 (0.019) |
|  West North Central | -0.030 (0.058) | -0.057\*\* (0.025) |
|  West South Central | 0.073\*\*\* (0.024) | 0.084\*\*\* (0.021) |
| Freeway lane miles per 1000 person in region | 0.064\*\*\* (0.024) |  |
| Annual Transit Revenue Miles per 1000 person in region (TranRevMiP1k) | -0.001 (0.001) |  |
| Block Group population density (D1B) | -0.001\*\*\* (0.0004) | 0.010\*\*\* (0.004) |
| Block Group Household Workers per Job (D2A\_WRKEMP) | -0.0003\*\* (0.0001) |  |
| #Pedestrian-oriented intersections per square mile (D3bpo4) | -0.0004\*\* (0.0002) |  |
| Employment accessibility (D5cr) | -12.000\*\*\* (2.930) |  |
| TranRevMiP1k interacting with aggregate frequency of block group level transit service (D4c) | -0.00001\*\*\* (0.00000) |  |
| Employment and household entropy (D2A\_EPHHM) |  | 0.044\* (0.026) |
| D1B interacting with D2A\_EPHHM |  | -0.026\*\*\* (0.007) |
| Intercept | -1.900\*\*\* (0.112) | -2.220\*\*\* (0.094) |
|  |
| Observations | 47,288 | 55,103 |
| R2 | 0.456 | 0.464 |
| Adjusted R2 | 0.456 | 0.464 |
| Residual Std. Error | 0.979 (df = 47258) | 1.000 (df = 55076) |
| F Statistic | 1,366.000\*\*\* (df = 29; 47258) | 1,834.000\*\*\* (df = 26; 55076) |
|  |
| Note: | \*p< 0.10 \*\*p<0.05 \*\*\*p<0.01 |
|  | Std error in parentheses. |

## Implementation

The VETravelDemandMM module is implementated as a self-contained R package and works as a module for the VisionEval framework. The VisionEval framework, itself an R package, provides functionalities for data loading and saving, validation, tests, and coordination with other modules. These functions are invoked through configuration and API calls to the framework, which facilitates the development of the VETravelDemandMM module. Besides the functions, the framework establishes a style and development guideline that enforces a consistent coding style across VisionEval code base, including modules like VETravelDemandMM (Flyte & Raw, 2018; Gregor, 2018).

Besides following the VisionEval guideline, the implementation takes advantage of the tidyverse suite of R packages for efficiency, concision and code readability (Wickham & Grolemund, 2017). The source code and documentation of the VETravelDemandMM package is available at <https://github.com/cities-lab/VETravelDemandMM> and subject to automated continuous tests via TravisCI. To ensure the implementation follows the best practices for developing and documenting planning support tools, FHWA and Oregon DOT recently gathered a team of domain experts and stakeholders and orchestrated a contribution review of the module (Details of the review process is documented by Stabler, Weidner, & Hull (2018)).

Given the nonlinear nature of our model, we compute the elasticities for variables of interest (number of vehicles, income, built environment) and compare them against the literature (Ewing & Cervero, 2010; Stevens, 2017) and the current model in RSPM. The elasticities are in line with both sources. We then apply the VETravelDemandMM module to the Rogue Valley MPO area for sensitivity testing and assessing its computational performance in application. The new models have better sensitivity to all built environment inputs than the current model in RSPM with reasonable running time (the whole module takes 1-1.5 minute for a population of about 170,000). The results of these comparisons and tests are available in our project report (Wang, 2017).

# Conclusion and Discussion

In this paper, we introduce the VETravelDemandMM module for the VisionEval framework. We document the process and techniques we utilize to do cross validation and model selection with the household AADVMT model. We aim for the simplest model with the best predicting power and behavior validity and select the power-transformed structure among a handful of options with a k-fold cross-validation process, a practice that has not been done systematically in research or development of other models to our knowledge. The cross-validation process also ensure we don't overfit the sample data as we add more independent variable while consider transformations of independent variables to capture non-linear and interaction effects.

We utilize a unique dataset by joining four nationwide datasets: the NHTS, the SLD, and transportation supply information from HPMS and NTD. To the best of our knowledge, the household AADVMT model presented in this paper is the first model utilizing such a nationwide dataset with high resolution built environment variables and household characteristics. Finally, we describe the implmentation of the VETravelDemandMM module as an open source R package following the best practices outlined in the VisionEval development guide (Gregor, 2018), R package development(Wickham, 2015), and open and reproducible research(Gandrud, 2015). To further ensure we follow the best practices, the module, including its method, implementation, and documentation, recently went through a review process by a panel of domain experts and stakeholders (Stabler et al., 2018).

Overall, we believe the development of the VETravelDemandMM described in this paper demonstrates a more rigorous and open process to travel model development, which we hope would inspire other model developers and researchers to adopt and improve upon. We hope that the modular structure, openly available source code, and documentation will facilitate applications and adoptation of the module by interested users and enable modifications and extensions by model developers and researchers.

# Acknowledgements

The paper results from a research project with financial support from Oregon Department of Transportation (SPR 788) and National Institute for Transportation and Communities (NITC-881).

# References

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Ewing, R., & Cervero, R. (2010). Travel and the Built Environment - A Meta-Analysis. *Journal of the American Planning Association*, *76*(3), 265–294.

Flyte, D., & Raw, J. (2018). Modeling In The Open: Best Practices And Lessons Learned For Development Of VisionEval, An Open Source Strategic Planning Project. In. Presented at the 7th International Conference on Innovations in Travel Modeling (ITM), Atlanta, GA.

Gandrud, C. (2015). *Reproducible Research with R and R Studio, Second Edition* (2 edition). Boca Raton: Chapman and Hall/CRC.

Gregor, B. (2015, December). GreenSTEP & RSPM Model Version 3.5 Technical Documentation. Retrieved March 26, 2017, from <https://github.com/gregorbj/RSPM/tree/master/Version_3.5/Documentation>

Gregor, B. (2018). Creating An Open-Source Model System To Facilitate Agile Model Development, Collaboration, And Connecting Transportation Research With Practice. In (Vol. Submitted). Presented at the 7th International Conference on Innovations in Travel Modeling (ITM), Atlanta, GA.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd edition). New York, NY: Springer.

Oak Ridge National Laboratory. (2011). *Developing a Best Estimate of Annual Vehicle Mileage for 2009 NHTS Vehicles*. Washington, D.C.: Federal Highway Administration. Retrieved from <http://nhts.ornl.gov/2009/pub/DerivedAddedVariables2009.pdf>

Ramsey, K., & Bell, A. (2014). *Smart Location Database: Version 2.0 User Guide*. Washington, DC 20460: US Environment Protection Agency. Retrieved from <https://www.epa.gov/smartgrowth/smart-location-mapping#SLD>

Stabler, B., Weidner, T., & Hull, K. (2018). Piloting A Contribution Review Process For The VisionEval Strategic Planning Modeling Framework. In. Presented at the 7th International Conference on Innovations in Travel Modeling (ITM), Atlanta, GA.

Stevens, M. R. (2017). Does Compact Development Make People Drive Less? *Journal of the American Planning Association*, *83*(1), 7–18. <https://doi.org/10.1080/01944363.2016.1240044>

U.S. Department of Transportation, Federal Highway Administration. (2009). *2009 National Household Travel Survey*. Retrieved from <http://nhts.ornl.gov>

Waddell, P. (2011). Integrated Land Use and Transportation Planning and Modelling: Addressing Challenges in Research and Practice. *Transport Reviews*, *31*(2), 209–229. <https://doi.org/10.1080/01441647.2010.525671>

Wang, L. (2017). *Performance-Based Planning and Decision Making* (No. SPR-788). Salem, OR: Oregon DOT.

Wang, L., Gregor, B., Yang, H., Weidner, T., & Knudson, T. (2018). Development of a Multi-modal Travel Demand Module for the Regional Strategic Planning Model. In *97th Annual Meeting of Transportation Research Board*. Washington, D.C.

Wickham, H. (2015). *R Packages: Organize, Test, Document, and Share Your Code* (1 edition). Sebastopol, CA: O’Reilly Media.

Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1 edition). Sebastopol, CA: O’Reilly Media.