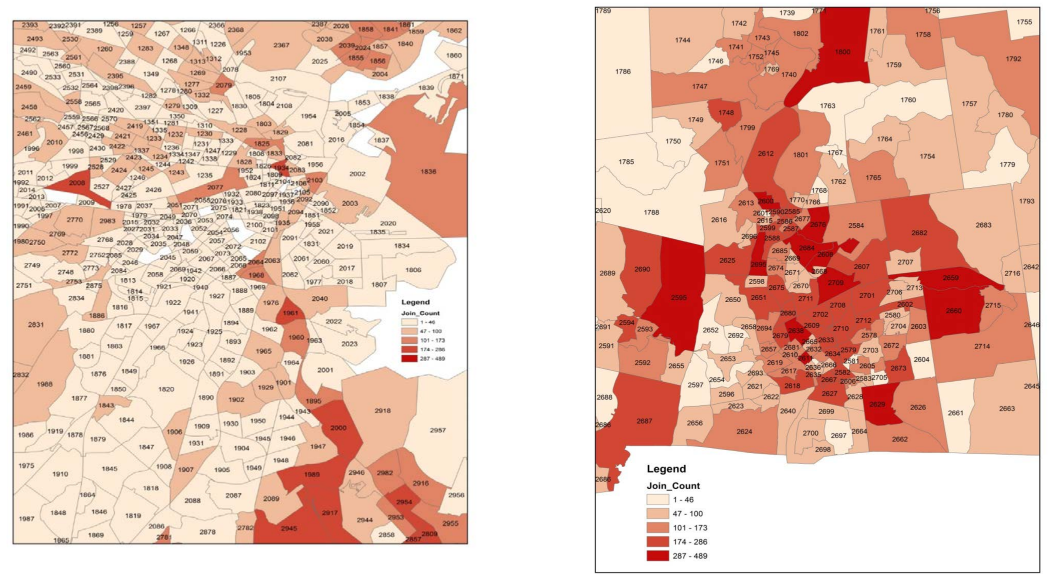
APPLICATIONS OF MACHINE LEARNING METHODS IN MACROSCOPIC CRASH ANALYSIS FOR TRANSPORTATION SAFETY MANAGEMENT

Somaye Dovirani, and Michael Knodler, PhD

Department of Civil and Environmental Engineering, University of Massachusetts, Amherst

INTRODUCTION

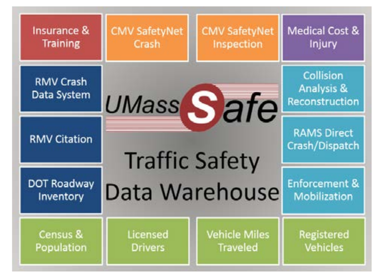
Traffic safety, which continues to remain a critical issue worldwide, has led to a myriad of modeling techniques to improve analytical capabilities with respect to crash modeling and prediction. The State and Metropolitan transportation planning processes must be consistent with Strategic Highway Safety Plans. This research aims to identify machine learning models and methods to improve the ability to capture variables that have the most significant impact on traffic safety through crash prediction with modeling parametric and nonparametric approaches tin machine learning models to use different models for prediction and inference with the aim of minimizing the reducible error.

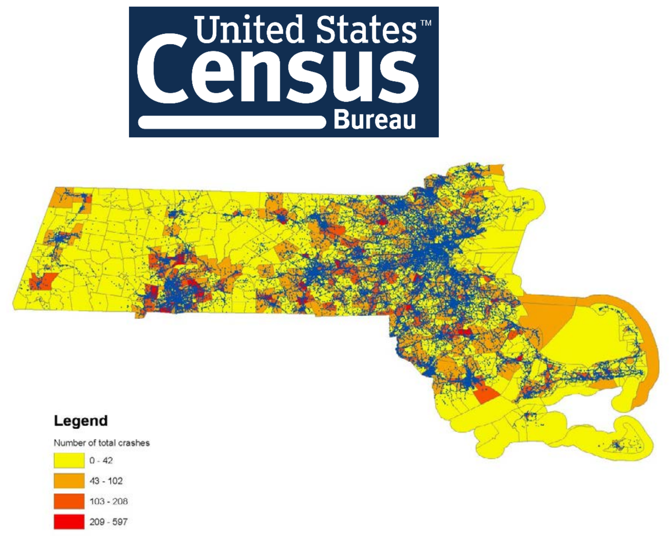


Total number of crashes in each TAZ in Springfield and Boston, MA.

DATA PREPARATION

Crash data has been obtained from UMass Safe Data Warehouse and has been geocoded and provided with several attributes. This GIS layer represented the actual locations of crashes. Information including Roadway Characteristics Inventory, traffic characteristics, demographic, and socioeconomic data have been collected from U.S. Census Bureau and the Office of Geographic Information (Mass GIS). The linked dataset allows for the creation of a model that directly relates level of service and level of safety. The final product creates a mechanism by which hot spots of crashes are readily identified and in which specific geometric elements can be manipulated to evaluate safety effects.



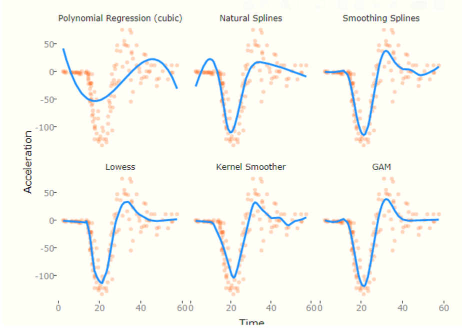


MODELS

Moving Beyond Linearity (GAM)

There are several methods that offer a lot of flexibility, without losing the ease and interpretability of linear models:

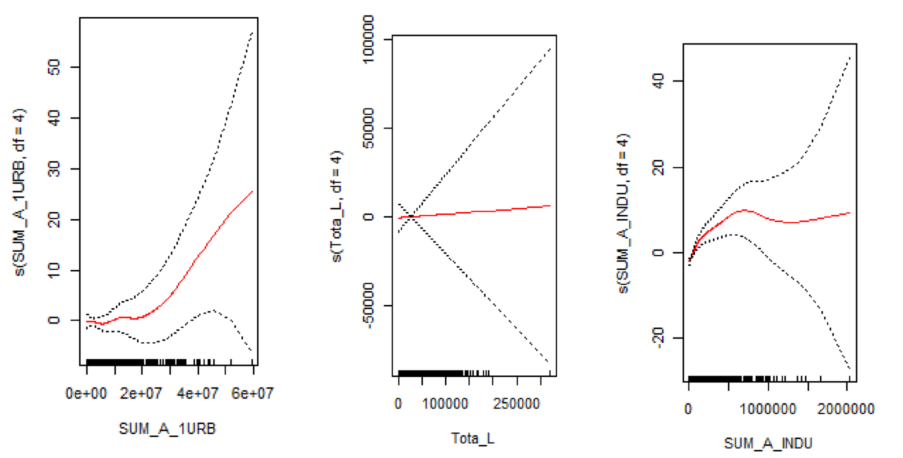
- Polynomial regression
- Step functions
- Regression and smoothing splines
- Local regression
- **Generalized additive**
- **models (GAMs) which use**
- **above models**



In particular, we replace each linear component of the multiple linear regression model with a (smooth) non-linear function.

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned}$$

In this model, we define separate fj for each Xj in each TAZ, and then add together all of their contributions.



GAM are a very nice and effective way lead to interpretable Models. easily mix terms in GAMs, some linear and some Non Linear terms

MODELS

Support Vector Machine(SVM)

The n-SVM model produces a learning model based on the training set and subsequently makes predictions on the testing set. The n-SVM model learns the relations between the TAZs-level crash frequency and explanatory variables based on the training dataset.

$\{(\mathbf{x}_i, y_i)\}_{i=1..n}$, and $\mathbf{x}_i \in \mathbf{R}$ which represents the **full set of zone-level contributing factors of each TAZ**

$y_i \in \mathbf{R}$ **represents the crash frequency that occurred in the TAZ.**

To linearize the nonlinear relation between x(i) and y(i). The estimation function of y(i) :

$$\hat{y} = f(x) = w^T \phi(x) + b$$

Where w ∊ R and b ∊ R are coefficients

$$MinZ(W, \varepsilon, \xi_i, \xi_i *) = C\{\upsilon\varepsilon + \frac{1}{N}\sum_{n=1}^{N}(\xi_n + \xi_n^*) + \frac{1}{2}W^T W$$

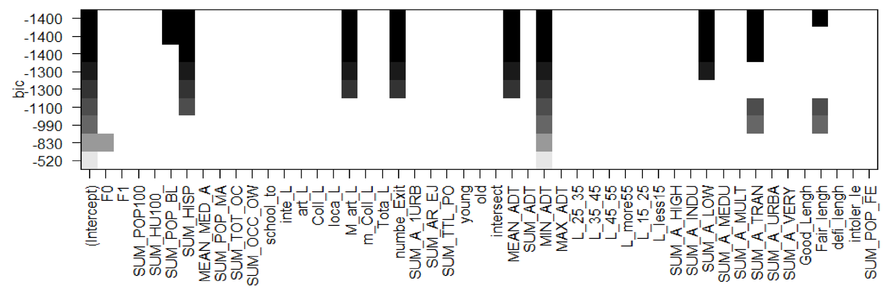
subject to: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all i

RESULTS

- ***Splitting dataset into training and validation sets***

Model Type	MSPE
Train Data 70% test 30%	657.45
Train Data 80% test 20%	659.10
Train Data 90% test 10%	649.47
Cross Validation 10-fold CV	610.78

Model Types	MSPE
Total Number of Crashes in Each TAZ	
NB	634.92
Best Subset Selection	623.17
Lasso	646.95
SVM	599.24



Model Types	MSPE
Total number of crashes in each TAZ	
NB	659.7116
Lasso	646.0906
SVM	643.1633

Comparison of the significant variable in total number of crashes and total number of severe crashes

Variable	Total number of crashes	Total number of Severe Crashes
SUM_POP_BL	-	
SUM_HISP	+	+
M_art_L	+	+
numbe_Exit	+	+
SUM_A_1URB	+	+
MEAN_AADT	+	+
MIN_AADT	+	+
SUM_A_HIGH	+	+
SUM_A_LOW		-
SUM_A_TRAN	+	
Fair_lengh		-

Few variables were associated with both total and severe crashes (red color). Seven variables in particular were present in the top of the variable rankings in terms of increasing for both total and severe crashes. The effect of these variables should be considered while developing a strategy for improving the safety of a zone. For example, a TAZ with higher number of AADTs can be prioritized for allocating funds for safety treatment, if necessary. TAZs with higher lengths of roadways for Minor Arterial or Rural Major Collector roadways may be scrutinized carefully by transportation officials to reduce severe crashes as well as total crashes. Alternatives such as installing speed-calming devices or lowering the speed limit may be additionally taken into account to improve safety.

DISCUSSION

Output of this study can improve modeling of transportation safety planning at the macro level. Using machine learning models to improve prediction of traffic crashes and improve safety with considering the boundary analysis can make these predictions more accurate. Development of new models to analyze drivers' contribution with developing policies regarding these analyses can decrease the crashes and using variables that have not traditionally been used in previous studies can improve the TSP.