Transportation Research Forum

Truck Volume Estimation via Linear Regression Under Limited Data
Author(s): Maria Boilé and Michail Golias
Source: *Journal of the Transportation Research Forum*, Vol. 45, No. 1 (Spring 2006), pp. 41-58
Published by: Transportation Research Forum
Stable URL: http://www.trforum.org/journal

Truck Volume Estimation via Linear Regression Under Limited Data

This paper employs linear regression algorithms in order to train models under the presence of limited training data. Usually in transportation applications, these models are built via Ordinary Least Squares and Stepwise Regression, which perform poorly under limited data. The algorithms presented in this paper have been extensively used in other scientific fields for problems with similar conditions and seem to partially or fully remedy this problem and its consequences. Four different algorithms are presented and several models are built. The models are used for truck volume prediction on highway sections in New Jersey, and results are compared to Stepwise Linear regression models.

by Maria Boilé and Michail Golias

INTRODUCTION

Trucks negatively impact the roadway network, primarily because of their massive weight, poor operating characteristics, and large dimensions. These impacts intensify the need for better truck traffic estimation techniques. Such estimates are used by state DOTs (Departments of Transportation) and MPOs (Metropolitan Planning Organizations) in pavement and bridge design and management, reconditioning and reconstruction of highway pavement, planning for freight movements, environmental impact analyses, and investment policies. Transportation planners and researchers have attempted to address the issue of predicting freight movements at the regional, state, and local level. The successful implementation of this type of analysis is limited, compared with similar analyses in passenger transportation due to the lack of appropriate freight transportation modeling methodologies, (models and data), and the complexity of the freight transportation system. The Quick Response Freight Manual (FHWA 1996) provides simple techniques and transferable parameters that can be used to develop commercial vehicle trip tables. The Trip Generation Handbook (2003), which provides guidelines for the preparation and application of trip generation data for a wide range of land-use categories, albeit not explicitly for freight trips, are widely used in practice even though the sources they are derived from are limited and outdated (Transportation Research Board 2001).

The current state of practice in truck trip activity estimation and freight modeling in general falls short of today's needs. According to the Federal Highway Administration (Transportation Research Board 2001) there are three major widely reported approaches to estimate truck trip data: a) Estimation of simple rates, b) Linear regression models, and c) Commodity flow models. Linear regression is one of the main methods used in vehicle-based modeling and can be considered rather simple and straightforward. Critical limitations of vehicle-based freight modeling include (Ortuzar and Willumsen 2001; Allaman et al. 1982): a) insufficient data and accuracy of the measured truck counts, b) different vehicle classification methods that further limit the available data, c) limited traffic data, d) time and place dependence of the models, and e) choice of the independent variable set.

Typical approaches to develop truck trip estimates using land-use and socio-economic data include, acreage of land used, square footage of building floor area, and employment or activity indicators (e.g., number of container lifts and import/export container moves). The selection of land-use categories is a critical question and one for which little guidance is available. The general approach in truck demand modeling applications is to employ land-use categories that correspond closely to

industry/employment categories. In addition, state DOTs obtain truck activity information on state highways through their traffic monitoring systems. These systems typically include information on traffic counts taken at various locations throughout the state. Although these counts usually provide a good geographic and temporal coverage for the overall traffic, there are a limited number of classification counts providing information explicitly on truck volumes. To fill the gap of limited availability of observed truck traffic data, various models may be used as predictive tools.

This paper describes the implementation of different linear regression techniques, which may be used to obtain more accurate estimation of commercial traffic. The goal is to create linear relationships between point estimates of truck volumes and surrounding land use activities and economic development in the form of: *Observed Truck Volume = F(Socioeconomic Variables),* where F is a linear function of the independent variables that represent land use and socioeconomic activity measures. Mittal et al. (2004) employed Stepwise Linear Regression (SLR) to build linear models that in some instances experienced: a) negative or extremely high predictions for the validation dataset, b) a negative sign on variables that have a positive effect on truck volumes, and c) over-fitting or null-model results (no predictors enter the model and the mean is used as the prediction for each observation). These problems are typical in linear regression modeling with limited data. In this paper, more advanced regression algorithms, which are shown in the literature (Hastie et al. 2001; Helland and Almoy 1994; Hubert and Vanden Branden 2003; Ngo et al. 2003) to have the capability to deal with issues such as the ones mentioned above, have been implemented and tested. Different approaches are used to create relationships between truck traffic volumes on roadways and their adjacent land use and economic activity. The resulting models are used to estimate truck volumes on roadway sections where such information does not exist through classification counts. Model results are compared with actual observations to determine their accuracy. The paper concludes with a discussion of the implementation of the proposed modeling framework within a Geographical Information System framework for easy access to data available to state DOTs, quick update of the models whenever new data becomes available, improved user friendliness, and visualization of model results.

BACKGROUND

Using Ordinary Least Squares (OLS) regression analysis for the estimation of truck trips as a function of a set of variables has been widely used. (See the Transportation Research Board, 2001, for a detailed list of truck demand modeling studies using linear regression.) The following is a brief description of the OLS method and its limitations with a presentation of alternative modeling approaches that may overcome these limitations.

OLS Regression

Suppose we have a training dataset[1] $(X_{ij}, y_i), \ldots\ldots, (X_{ij}, y_i)$, where i=1,..,n is the number of observations, j=1,..,m is the number of independent variables/predictors, X_{ij} are column vectors (observed values of the independent variables), and y_i is the vector of the observed values of the dependent variable (in the case of truck trip estimation y_i are observed truck trips). The comparison class consists of the linear function **Y=X*b.** Throughout this paper $\mathbf{X}=\{X_{1j}, X_{2j}, \ldots\ldots X_{ij}\}$ is referred to as the independent variable dataset and $\mathbf{Y}=\{y_1, y_2, \ldots\ldots y_i\}$ as the dependent variable dataset. The least squares linear regression method recommends computing the column vector $\hat{b}$ (regression coefficient vector) that minimizes the squared difference of the observed values from the models' predictions:

$$\hat{b} = \arg\min_{b} \sum_{i=1}^{n} (y_i - bo + \sum_{j=1}^{m} X_{ij} \hat{b}_j)^2$$

where: *bo* is the intercept. A basic criterion for the goodness of fit of the model is the R^2 value (i.e., the fraction of the variance in the data that is explained by the regression model) while the significance of the model and the variables are expressed through other statistical measures (such as the F or p values). Prior to creating the model, it is assumed that a number of independent variables (***X***) have a causal

effect on the dependent variable (***Y***), but it is rarely known with certainty which independent variables should be included in the final model.

Often in OLS, exact multicollinearity, caused by limited data,[2] will cause instability of the estimated parameters[3] (regression coefficients) unrealistic models that overfit the data, and conceptually incorrect models (independent variables that appear in the model with an "*incorrect*" coefficient sign; e.g., negative coefficient for an independent variable that has been known to have a positive effect on the dependent variable). Several different approaches have been presented in the literature to deal with this problem and can be classified as: a) variable elimination, b) variable combination, and c) variable shrinkage techniques. The main idea behind these techniques is to try to reduce either the number (variable elimination or variable combination) or the influence (variable shrinkage) of the independent variables.

Variable Elimination

Eliminating variables from a model is a special case of model selection procedure and includes Stepwise and all-subsets regression. In stepwise regression (SR) the basic procedures involve: a) identifying an initial model, b) iteratively altering the model from the previous step by adding or removing an independent variable in accordance to a certain criterion (usually the F or p values of the independent variable under consideration), and c) terminating the search when improvement of the model is no longer possible given a certain criterion, or when a specified maximum number of steps has been reached. A limitation of the SR approach is that it assumes there is a single "best" subset of the independent variables and seeks to identify it. Furthermore, if, during stepwise variable selection, a predictor is ultimately excluded from a model due to its low significance (F or p value), the coefficients of the other variables will change (Neter et al. 1996). Thus the use of SR may exclude explanatory variables that are actually highly correlated with the dependent variable.

All-possible-subset regression (APSR) can be used as an alternative to stepwise regression. Using this approach, one first decides on the range of subset sizes that could be considered useful. For example, one might expect that inclusion of at least three independent variables in the model is necessary to adequately explain the dependent variable, and also might expect there is no advantage to considering models with more than six independent variables. Only the "best" of all possible subsets of three, four, five, and six independent variables would then be considered. The problem with APSR is that the number of possible models increases very rapidly as the number of independent variables in the whole model increases. For example, for the all-possible subsets regression with up to 12 independent variables to be performed, about 2.7 million different models need to be estimated.

Both SR and APSR are very sensitive to the size of the dataset and overfitting under limited data is a major problem with both approaches (Hastie et al. 2001).

Variable Combination

In some situations, it is not feasible to use variable selection to reduce the number of independent variables or it is not desirable to do so because the experience of the modeler with the problem suggests that all of the considered variables should be present in the final model. In these situations the general method used, based on Principle Components Analysis (PCA), is the Principal Component Regression (PCR). The idea of PCR is to combine all the independent variables into a new group of variables (principal components), and then regress the dependent variable on the newly created group. Major limitations of this approach include choosing the number of the new variables, interpretation of the principal components, and complexity of applying the method.

Another method that belongs to this category is Partial Least Squares Regression (PLSR). It is a recent technique that generalizes and combines features from PCA and multiple regression. It is particularly useful when a set of dependent variables needs to be predicted from a very large set of independent variables (Abdi 2003).

Variable Shrinkage

A less complicated approach compared with variable combination is the use of shrinkage estimators.[4] Ridge Regression (RR), and Lasso Regression (LR) are two of the most widely used shrinkage techniques that perform well under multicollinearity (LR can be also considered as a variable elimination technique). Ridge regression is probably the strongest competitor for PLSR in terms of flexibility and robustness of the predictive models. Both methods, which can be considered as constrained versions of OLS, require the setting of arbitrary "constant/ tuning parameters" (explained in more detail in the Model Description section), which is used to *shrink* the regression coefficients from their original OLS value. This can be considered as a major limitation of the methods because selecting the constant could become a very cumbersome and time-consuming procedure. Both methods though, have been known to produce more robust results when compared to OLS or SR (Hastie et al. 2001).

MODEL FORMULATION

Approach Selection

In this section a brief description of the different model formulations, with links to related literature, are presented. PLSR, RR, and LR are probably the least restrictive of the various multivariate extensions of the multiple linear regression models. This flexibility allows them to be used in situations where the use of traditional multivariate methods (OLS, SR) is severely limited, such as when there are fewer observations than predictor variables. An extensive simulation study comparing variable selection regression methods is presented in Frank et al. (1993), and although results are conditional on the simulation design, they indicated that PCR, RR, and PLSR are, in the case of limited data or multicollinearity problems, highly preferable. Wentzell and Montoto (2003) present theoretical and empirical comparisons of PCR and PLSR,[5] concluding that the two methods produce similar optimal predictions and perform in a similar way, while Helland and Almoy (1994) and Helland (2001) proved that neither PCR nor PLSR dominate one another. On the other hand, unlike LR, none of these methods perform variable selection/elimination.

PLSR, RR, and LR have been used in various disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science where predictive linear modeling, especially with a large number of predictors is necessary. These techniques will also be examined in this paper. Motivated by the use of constraints in all three methods, a classical constrained optimization approach is also presented. The main reason for introducing this approach, described later in detail, is that in contrast to the other techniques, it is quite suitable to include certain decision maker preferences if these need to be reflected by the final model and its predictions. The authors would like to note that least absolute deviation and least median of squares linear regression, variations of OLS that can be used, were not considered in this paper as they can very frequently exhibit instability (Ellis 1998).

Ridge and Lasso Regression

Ridge and Lasso regression are shrinkage methods that constrain large values of the coefficients ($\boldsymbol{b_j}$) of the linear model. The difference between these two algorithms is that, while Ridge regression does not omit any of the independent variables, Lasso, due to the type of the constraint used, can zero-out some of the coefficients. The formulas for both methods are given below in equations 1 and 2. In these equations *bo* is the intercept, b_j are the regression coefficients, and $\boldsymbol{X_{ij}}$ is the value of the independent variable j at $y_{i.}$ Adjusting for the tuning parameters $\boldsymbol{s}$ and $\boldsymbol{t}$ (parameters that constrain/penalize the regression coefficients) in equations 1 and 2 produces different model estimates. Notice that when $\boldsymbol{s}$ and $\boldsymbol{t}$ are equal to 0, the least squares estimate is obtained. However, as $\boldsymbol{s}$ and $\boldsymbol{t}$ get bigger, over fitting gets more expensive as larger values of $\boldsymbol{b_j}$ penalize the criterion more.

Ridge Regression Formulation

$$(1)\ \bar{b} = \arg\min_{b} \sum_{i=1}^{N} (y_i - \bar{bo} - \sum_{j=1}^{p} X_{ij}\bar{b}_j)^2,$$

Subject to: $\sum_{i=1}^{p} \bar{b}_j{}^2 \le s$

Lasso Regression Formulation

$$(2)\ \bar{b} = \arg\min_{b} \sum_{i=1}^{N} (y_i - \bar{bo} - \sum_{j=1}^{p} X_{ij}\bar{b}_j)^2,$$

Subject to.: $\sum_{i=1}^{p} \left|\bar{b}_j\right| \le t$

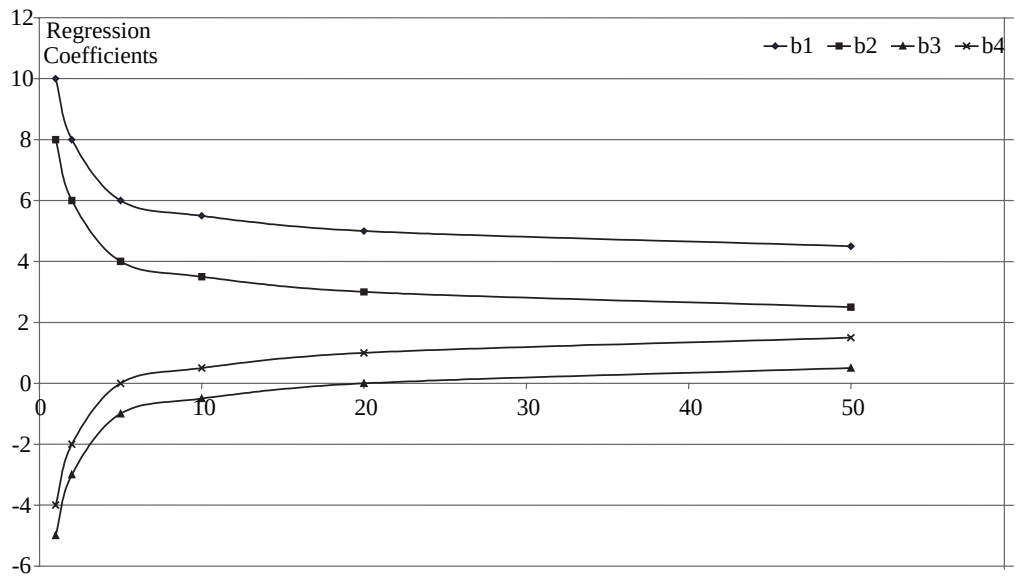
Ngo et al. (2003) presented three practical cases where the application of ridge regression is studied and illustrated through mathematical derivation and computer simulation. In all three cases, the improvement over the OLS method was tremendous in both the relevant mean square error function[6]and the ridge trace. (A plot of the regression coefficients as a function of the ridge parameter is shown in Figure 1 for a four-dimensional hypothetical regression scenario.)

It can be shown (Hastie et al. 2001) that Ridge regression has a closed form solution. Grandvalet (1998) derived an EM (expectation maximization) algorithm that allows for the computation of the Lasso solution. The algorithm is used in this paper. The main drawback of these two methods is the difficulty in deciding on the values of the tuning (***s*** and ***t***) parameters. Usually, cross-validation[7] is used but this requires a significant amount of training data. In this paper we use an iterative process to set the values for these parameters. Both RR and LR algorithms are conceptually easy to apply and are part of many statistical packages (SAS, SPSS, MatLab, R, SPLUS), which facilitates their implementation.

Partial Least Squares Regression

PLSR is a linear regression technique developed to deal with high-dimensional regressors (large number of independent variables) and one or several dependent variables. In PLRS we assume that ***X*** and ***Y*** are related through a bilinear model. The main idea of the bilinear structure is to first construct k_p variables as a linear combination of the X variables

Figure 1: Hypothetical Ridge Trace Plot



Note: As the shrinkage parameter increases the regression coefficients shrink

($k_p=F_p(X)$, $p<||X||$) and then regress the response (Y) onto these k_p variables ($Y=G(k_p)$). It is well known that popular algorithms for PLSR (Wakeling and Macfie 1992; De Jong 1993) are very sensitive to outliers in the data set. Hubert and Vanden Branden (2003) present robust algorithms that can handle high dimensional spaces.[8] Their algorithms are extremely suitable for high-dimensional data (when the number of the independent variables are larger than half the number of observations). Their approach combines the goodness-of-fit and the predictive power of the model when selecting the best sub-model. The selection of the number of variables to be used is based on the $RCS_k = \sqrt{\gamma A_k^2 + (1-\gamma) A_k^2}$ statistic, where γ is a tuning parameter between 0 and 1, and A is a robust statistic. For further details see Hubert and Verboven (2003), and Hubert and Vanden Branden (2003). The γ parameter is used to decide whether the model needs to be strong in prediction (γ > 0.5) or whether the goodness-of-fit is a primary interest (γ < 0.5). For this paper a combination of three values (0, 0.5, and 1) for the γ parameter was used in order to select the number of components as suggested in Engelen and Hubert (2005). For more details the reader is referred to Engelen and Hubert (2005). A Matlab[9] toolbox enabling the use of these algorithms is available and is used in this paper (Verboven and Hubert 2005).

Constrained Linear Least Squares Optimization (CR)

Constraint regression has been extensively treated in the literature (Mukerjee and Tu 1995; Geweke 1996; Knautz 1998; Koenker and Ng 2004; Klugkist 2004; Li 2005; Zhu et al. 2005; Edlund and Ekblom 2005). Similar to this concept and using an objective function (equation 3) that minimizes the sum of squares, constraints are added to the values of the coefficients as well as to the values of the predicted variables.

Constrained Regression Formulation

$$(3) \quad \min_b \left[\sum_i \{ (\sum_j X_{ij} \bar{b}_j) - y_i \}^2 \right]$$

subject to:

$$(3a) \quad \sum_j X_{kj} \bar{b}_j \le d_k \quad \forall k \subseteq i$$

$$(3b) \quad \sum_j X_{lj} \bar{b}_j = d_l^{eq} \quad \forall l \subseteq i$$

$$(3c) \quad lb \le b_j \le ub \quad \forall j$$

$$(3d) \quad k + l = n$$

Where: *lb* and *ub* are the lower and upper bound vectors for the beta values, d_k and d_l^{eq} are the upper and equality bounds, and *n* is the number of observations.

From the engineering point of view the first constraint (3a) captures the range of the expectation for the observed truck volumes, taking into account the uncertainty of the accuracy on the measurement of each station. The second constraint (3b) can be considered a weighting factor for the observed truck volumes. In some cases it is known that the observed truck measurement is accurate (i.e. weigh-in-motion [WIM] station counts) and in some cases it may not be very accurate (i.e., 48-hour count stations). Setting up equality constraints for some or all of the accurate measurements forces the model to give more weight to these measurements, minimizing transferring of error that may exist in vehicle counts. The third constraint (3c) can be considered a weighting factor of the decision variables. The upper and lower bounds of the constraints are based on the training data and possibly the engineers' experience with the study area. If a priori knowledge for a variable's positive effect exists, that variables' beta coefficient can be constrained to positive values and vice versa.

CASE STUDY

The statistical methods described above were tested with data consisting of classification

traffic counts as the dependent variable and socioeconomic data as the independent variables. The dependent variable dataset was obtained from various locations throughout New Jersey. It consists of 270 long and short duration truck traffic counts (FHWA 2001) taken at different locations in the state. Long duration counts were obtained by permanent WIM locations. Initially traffic counts of vehicle classes 5 through 13 in the FHWA vehicle classification system were to be considered trucks. Following NJDOT officials' suggestion, vehicle class 5 was removed from the analysis because of the arguable way that class 5 vehicles are classified, resulting in cars, small pick-up trucks, and vans to often be classified as class 5 trucks.

Data for the independent variable dataset included population, the number of employees, sales volume, and number of establishments for each Standard Industrial Classification (SIC) code. A total of 34 independent variables, including population, were considered in the final model training process (Table 1). Both the dependent and the independent variables and the estimates are based on 2001 data. Data have been extracted from the ESRI BIS (Environmental Systems Research Institute Business Information Solutions), database,[10] a comprehensive list of businesses licensed from InfoUSA.

Uniform highway sections were defined around each classification count location. In addition, for model validation and testing purposes, uniform sections were defined on 14 major highways. The sections were defined based on a set of criteria such as major interchanges, changes in roadway functionality and in roadway geometry, and traffic count availability. Socioeconomic data associated with these sections were extracted and used as input in the model training and testing process. ArcView, a GIS software package, was used to buffer and aggregate the independent variable dataset for nine different bandwidths of influence[11] (0.25, 0.50, 0.75, 1.0, 1.25, 1.5, 2, 3, and 5 miles area around each section).

Creating models based on different buffer zone sizes permits the determination of the sensitivity of a model with the increasing size of the area of influence of the independent variables (as the buffer area size increases the model accuracy fluctuates). This procedure will identify the most appropriate buffer zone size and model for a particular type of roadway. In order to reduce the prediction error and maximize the correlation between the prediction variables and the predicted truck volumes, the dataset was clustered into six subsets (Table 2) according to the functional class (FC) of the roadway (Weinblatt 1996).

Building models by considering roadway classes is significant, as different roadways attract different truck volumes that are dependent on different variables. Roadways are classified

Table 1: SIC Titles and Corresponding Independent Variables*

SIC Title and Corresponding Independent Variables					
Mining	**Agriculture**	**Manufacturing**	**Construction**	**Transportation**	**Utilities**
Number of Employees	Number of Employees	Number of Employees	Number of Employees	Number of Employees	Number of Employees
Sales Volume	Sales Volume	Sales Volume	Sales Volume	Sales Volume	Sales Volume
Number 0f Establishments	Number of Establishments	Number of Establishments	Number of Establishments	Number of Establishments	Number of Establishments
Retail Trade	**Wholesale Trade**	**Real Estate**	**Finance/ Insurance**	**Services**	
Number of Employees	Number of Employees	Number of Employees	Number of Employees	Number of Employees	**Population**
Sales Volume	Sales Volume	Sales Volume	Sales Volume	Sales Volume	
Number of Establishments	Number of Establishments	Number of Establishments	Number of Establishments	Number of Establishments	

*Data provided by NJDOT is in SIC, not NAICS

under different FCs based on the type of the roadway, lane width, traffic, and functionality. Roadway information was obtained through the NJDOT Statewide Truck Model (STM) and the 2002 New Jersey Straight Line Diagrams (NJSLD).

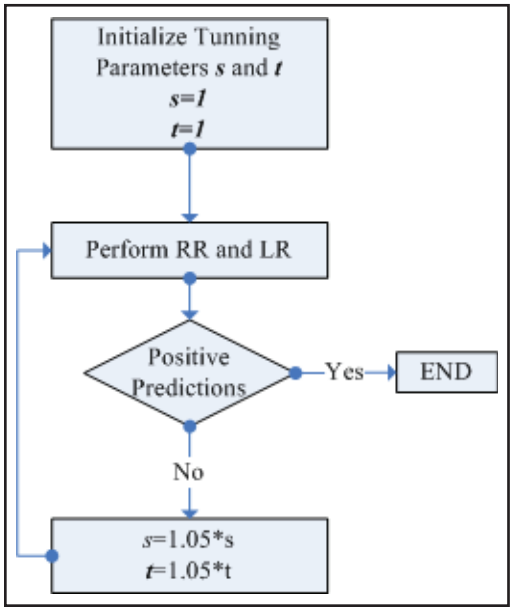
Table 2: Clustered Dataset by Highway Functional Class and Count Availability

Functional Class	*Counts*
FC = 1, 2 (rural interstate and major arterials)	31
FC = 6, 7, 8, 9 (rural minor arterials, collectors, and local)	51
FC = 11 (urban interstate)	29
FC = 12 (urban expressways and parkways)	20
FC = 14 (urban major arterials)	59
FC = 16, 17, 19 (urban minor arterials, collectors, and local)	80

MODEL APPLICATION

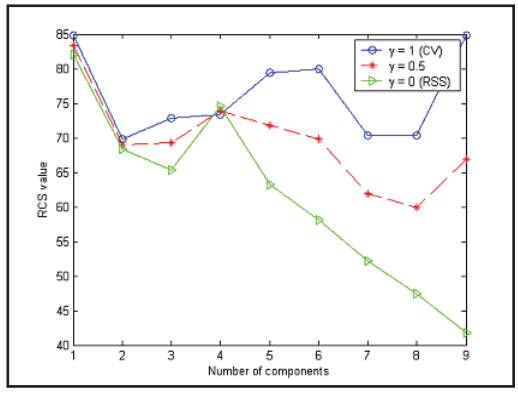
The main issue with RR and LR was the choice of the values for the tuning parameters ***s*** and ***t***. The limited training data did not allow cross-validation to be performed. Instead, multiple values for the parameters were used. As shown in Figure 2, the values for the tuning parameters are first initialized (s=t=1). Ridge and Lasso regression are performed, and if all the predicted truck volumes ($\hat{y}_i$) are positive the process stops. If not, the parameters are increased by 5% and the algorithms are re-performed. Out of all the different values of ***s*** and ***t*** that were used in the iterative process shown in Figure 2, two are chosen for each approach (RR and LR): a) the values of ***s*** and ***t*** that produce the models with the highest R^2 value, and b) the values of ***s*** and ***t*** that produce a model with all the predictions positive.

Figure 2: Ridge and Lasso Regression Tuning Parameter Value Selection Process



The main issue with PLSR as discussed previously is the choice of the number of components (latent variables) used. For that purpose an extra parameter (γ) was used to select the optimal number of components. The values of γ range between zero and one. High γ values would improve goodness of fit at the expense of predictive power and vice versa. For the example presented herein, Figure 3 indicates that the number of components that achieves a balance between minimizing the error in prediction while maximizing the goodness of fit would range between two and five.

Figure 3: RCS Value for Three Different γ Values



The constraint optimization approach was the last to be implemented. The formulation implemented for this study is given in equations 4a-4c.

CR Final Formulation

$$\min_{b}\left[\sum_{i}\{(\sum_{j} X_{ij}\bar{b}_j) - y_i\}^2\right] \quad (4a)$$

subject to

(4b) $0.25 * y_i \le \sum_j X_{ij} \bar{b}_j \le 1.25 * y_i, \ \forall i$

(4c) $0 \le b_j, \ \forall j$

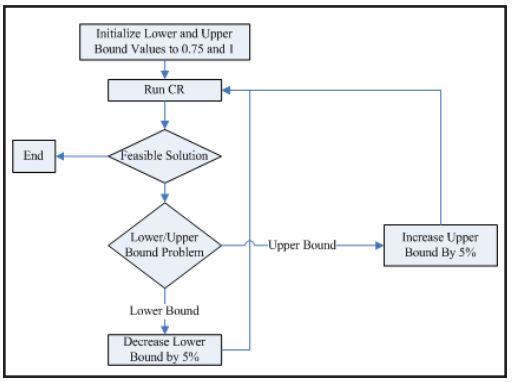
Constraints on the minimum and maximum value, for both the coefficients and the predicted variables, may vary so that the models account for space variations corresponding to the functional class of the highway and the geographical location of the count. The first constraint (4b) requires that the values of the estimated truck volumes fall within 25% to 125% of the observed value. This range of the predicted truck volumes is not necessarily the same for all the stations. It may vary based on the functional class of the roadway, the type of the observed count, and the count location. These bounds were determined using the iterative process shown in Figure 4. The values of the lower and upper bounds are first initialized (lower bound = 75%, upper bound = 100%) and then the CR is performed. If a feasible solution is obtained, the algorithm stops. If a feasible solution is not obtained, the bound that causes the feasibility problem is identified and the value is increased (upper bound)/decreased (lower bound) by 5% and the CR is re-performed. This process continues until both bounds provide a feasible solution. The limitation of using constraint 4b is that for relatively small training datasets and strict lower and upper bounds, the solution may be infeasible (as was the case here where the lower bound dropped to 25% of the observed value in order to obtain a feasible solution). A pseudo-increase[12] of the data, similar to the bootstrap method[13], was performed for all the subsets and the results showed that both interval bounds are inversely correlated to the amount of the training data. In other words: *the larger the dataset the stricter the bounds that can be enforced*.

The second constraint (4c) indicates that the predictive variables should have a positive effect on truck volume production. This constraint was used because, due to the small amount of data, one or two outliers were enough to enter a variable into the model with an incorrect sign (which was the case with SLR). To verify the assumption of the positive effect for all the independent variables, Mean Coefficient Regression (Pazzani and Bay 1999) was performed for each dataset and the results showed positive correlation between predictors and predicted variables in isolation.

In all three approaches, two criteria to be met are set as shown in equation (5). The first criterion requires that R^2 values fall between two extreme values. It is assumed that if the R^2 is greater than 0.9, then the model is over-fitting the learning dataset, while for R^2 less than 0.5 the predictive power of the model is not adequate. The second criterion requires that the predicted truck volume at *i*, $\hat{y}_i$, is greater than or equal to a lower bound value (y_0)

(5) $0.5 \le R^2 \le 0.9$, and $\hat{y}_i \ge y_o$

Figure 4: Iterative Process for Upper and Lower Bound Determination



MODEL EVALUATION

For the first part of the evaluation, the adjusted-R^2 values[14] of the models obtained are compared to the ones obtained from the SLR approach. Table 3 presents the best R^2 value for each type of roadway and the corresponding band buffer used to extract the socioeconomic data. For all statistical approaches, the results show that the best model for a roadway depends on the type and the function that the roadway serves, but is also dependent on the size of the buffer zone of influence. Table 3 shows that higher-level roadways (expressways [FC=12] and urban interstates [FC=11]) have a larger optimal band size compared with lower level roadways. This result satisfies the underlying assumption that trucks will use local roads only to access local

Table 3: R^2 Values and Band Buffer for the Best Model for Each Functional Class

Roadway Functional Class	CR (No R^2 Problem, No Prediction Negativity Problems)		SLR (Negativity In Some Predictions and R^2 Problem)		RR (Small R^2 Values for some models)		LR (Small R^2 Values for some models)		PLSR (Negativity In Predictions)	
	R^2	Band	R^2	Band	R^2	Band	R^2	Band	R^2	Band
1-2	0.82	0.25	0.97	0.25	0.54	0.25	0.65	0.25	0.69	0.25
6-9	0.79	0.25	0.84	0.5	0.62	0.25	0.76	0.25	0.65	0.25
11	0.77	0.5	0.92	0.75	0.34	0.5	0.41	0.5	0.65	0.75
12	0.87	0.75	0.99	1.0	0.28	1.0	0.44	1.0	0.80	0.75
14	0.87	1.0	0.13	0.25	0.1	1.0	0.1	1.0	0.38	1.5
16-19	0.82	1.25	0.59	0.25	0.1	1.25	0.18	1.25	0.33	1.00

facilities and they will travel over higher level roadways for the rest of their trip.

It can be seen that SLR produces some models that are unrealistic (R^2 values close to 1) and most probably over-fit the learning dataset (negative predictions). On the other hand, the other approaches seem to produce models with more reasonable R^2 values. Only the CR models managed to meet both of the criteria set in equation (5), and in that sense produced better results than SLR. On the other hand, RR and LR models did not always meet both criteria simultaneously. For large values of the parameters, the correlation coefficient was more than satisfactory (R^2>0.65, p<0.05) but some of the predicted values on the learning dataset were negative. Increasing the values of the tuning parameters (***s*** and ***t***), the constraints in equation (2) and (3) become less restrictive and the solution approaches the least squares solution. As the values of the tuning parameters were decreased, the predictions were positive but the R^2 value was below satisfactory levels (as set in equation 5). The concept behind the effect of changing the value of these parameters is discussed in more detail in Hastie et al. (2003).

PLSR exhibited similar behavior (results depend on the chosen γ value and the number of components used) and produced negative values for some of the observations. Results shown in Table 3 are for γ=0.5 (a value for which the error between prediction and goodness-of-fit models is averaged). Compared to the SLR approach, however, RR, LR, and PLSR models produced better results. They reduced the number of negative predicted truck volumes by 80% to 100% compared to the same number in the SLR models and produced models for all the band buffers of the six different clusters of roadways. Models 1-6 present the final models built using CR, which are the only ones meeting both criteria (as shown in equation 5). All the variables that entered the model had a p-value<0.05. However, it should be noted that under such limited data p-values lose most of their explanatory power.

Model 1: Trucks on Rural Interstate (R^2=0.82, p<0.05)

Daily Truck Volume = 48+8.5442*EMP_TRANSP+ 1.2641*EMP_FINANC+ 2.8996*EMP_REAL+ 0.1758*SALES_TRAN+ 0.0114*SALES_UTIL

Where: EMP_TRANSP is the employment in the transportation industry, EMP_FINANC is the employment in the finance industry, EMP_REAL is the employment in real estate, SALES_TRANS is the number of sales in the transportation industry, and SALES_UTIL are the sales in the utilities industry.

Model 2: Trucks on Rural Minor (R^2=0.79, p<0.05)

Daily Truck Volume = 3+0.245*EMP_AGRICU+ 1.02*EMP_CONSTR+ 0.013*EMP_UTILIT+ 0.001*SALES_AGRI+ 0.001*SALES_MANU+ 15.574*COUNT_

TRAN+ 3.142*COUNT_WHOL

Where: EMP_AGRICU is the employment in the agricultural industry, EMP_CONSTR is the employment in the construction industry, EMP_UTILIT is the employment in the utilities industry, SALES_AGRI is the number of sales in the agricultural industry, SALES_MANU are the sales in the manufacturing industry, COUNT_TRAN is the number of establishments in the transportation industry, and COUNT_WHOL is the number of establishments in the wholesale industry.

Model 3: Trucks on Urban Interstates (R^2=0.77, p<0.05)

Daily Truck Volume = 267+10.258*EMP_AGRICU+ 7.71*EMP_MINING+ 3.556*EMP_CONSTR+ 0.157*EMP_MANUFA+ 0.073*SALES_MINI

Where: EMP_AGRICU is the employment in the agricultural industry, EMP_CONSTR is the employment in the construction industry, EMP_MINING is the employment in the mining industry, EMP_MANUFA is the employment in the manufacturing industry, and SALES_MINI is the sales in the mining industry.

Model 4: Trucks on Expressways (R^2=0.87, p<0.05)

Daily Truck Volume = 110+0.348*EMP_WHOLES+ 0.428*EMP_RETAIL+ 0.008*SALES_CONS+ 268.57*COUNT_MINI+ 29.976*COUNT_TRAN

Where: EMP_WHOLES is the employment in the wholesale industry, EMP_RETAIL is the employment in the retail industry, SALES_CONS is the sales in the construction industry, COUNT_MINI is the number of establishments in the mining industry, and COUNT_TRAN is the number of establishments in the transportation industry.

Model 5: Trucks on Urban Major (R^2=0.87, p<0.05)

Daily Truck Volume = 26+0.673*EMP_CONSTR+ 0.129*EMP_MANUFA+ 0.076*EMP_WHOLES+ 0.007*SALES_TRAN+ 0.001*SALES_UTIL+ 13.213*COUNT_AGRI+ 257.39*COUNT_MINI

Where: EMP_CONSTR is the employment in the construction industry, EMP_MANUFA is the employment in the manufacturing industry, EMP_WHOLES is the employment in the wholesale industry, SALES_TRANS is the number of sales in the transportation industry, SALES_UTIL is the number of sales in the utility industry, COUNT_MINI is the number of establishments in the mining industry, and COUNT_AGRI is the number of establishments in the agricultural industry.

Model 6: Trucks on Urban Minor (R^2=0.82, p<0.05)

Daily Truck Volume = 4+0.004*SALES_MINI+ 0.002*SALES_TRAN+ 2.98*COUNT_AGRI+ 24.995*COUNT_MINI

Where: SALES_MINI is the sales in the mining industry, SALES_TRANS is the number of sales in the transportation industry, COUNT_AGRI is the number of establishments in the agricultural industry, and COUNT_MINI is the number of establishments in the mining industry.

The second part of the evaluation compared the predictive power of the models on 14 selected New Jersey highways. The RR, LR, and PSLR models had a better predictive power (less negative predictions) than SLR. The CR models produced the best results among the three methods and are used in the following discussion. Results for highways I-80 and US 206 are presented in Figures 5 and 6, and in Tables 4 and 5. The reason for choosing these highways is the high number of observed truck traffic counts. These figures show the predicted truck volumes for each section of the highway. Observed counts are also shown for sections of the highway, for which such information exists. As can be seen in Figure 5 and Figure 6, the negativity problem in the predictions has been resolved. It is also obvious that the CR approach tends to reduce, but not eliminate, the over-estimation problem. This pattern is followed in all 14 highways (205 sections) selected to test the models.

Another problem of the SLR method was related to the values of the intercept. The intercept in these models can be interpreted as: "*How many trucks should we expect at a specific section if there is no influence from adjacent land use and economic activity.*" The

Figure 5: Observed and Predicted Truck Volumes from CR and SLR Models for Highway I-80.

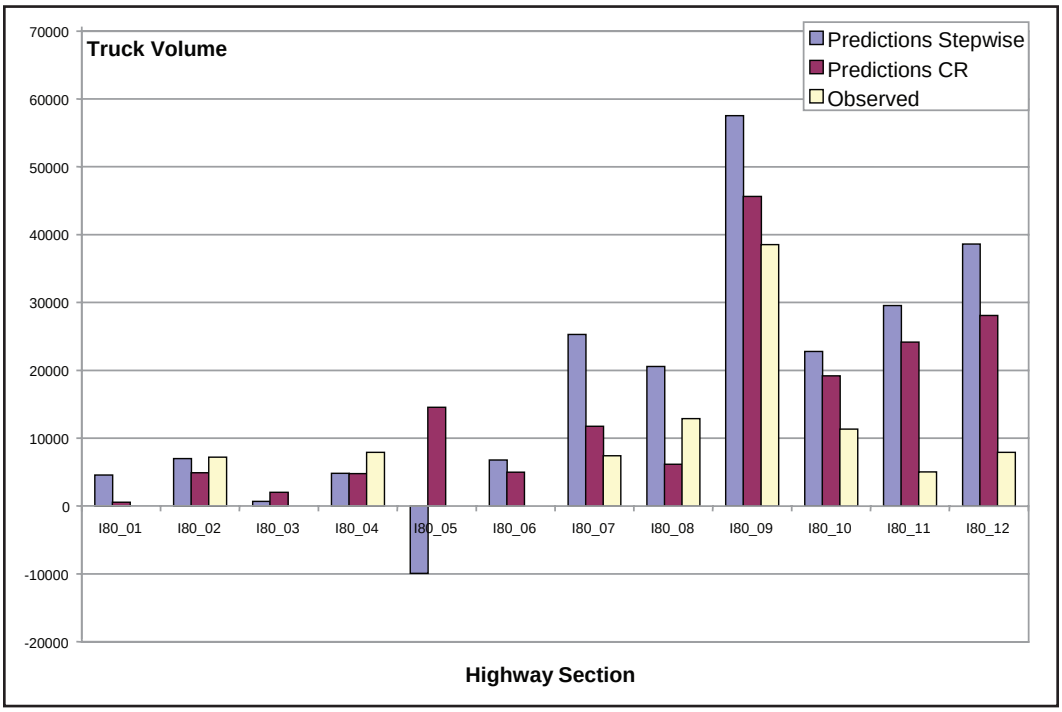


Figure 6: Observed and Predicted Truck Volumes from CR and SLR Models for Highway US 206

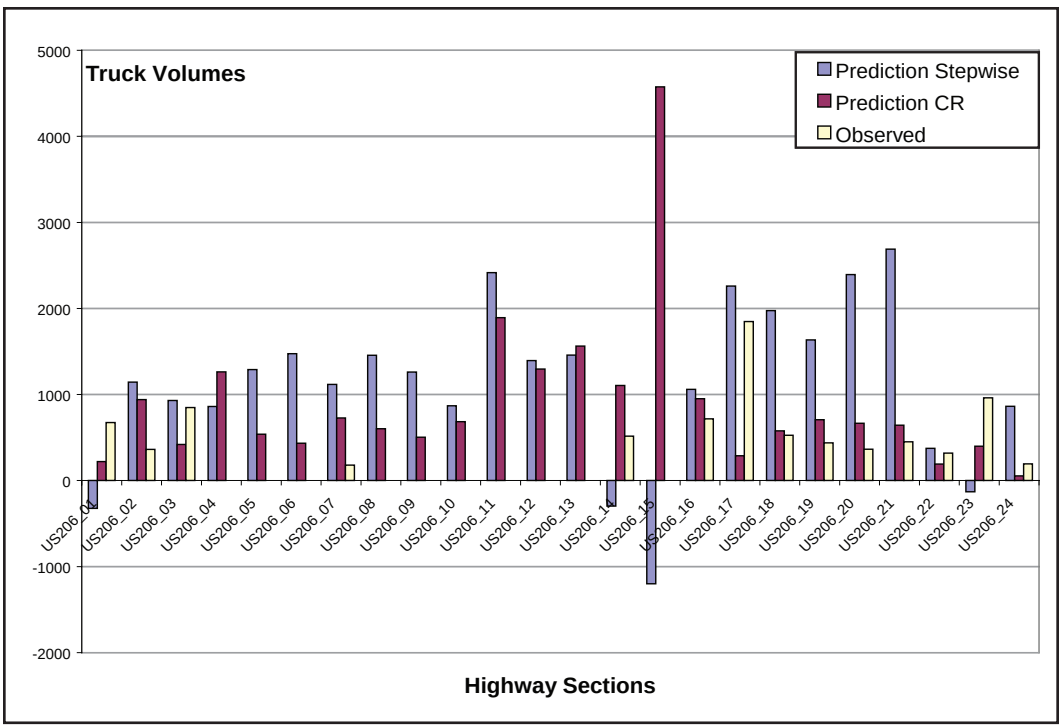


Table 4: Observed, Predicted and Relative Errors of Truck Volumes for Highway I-80 (Stepwise and Constrained Regression)

Highway Section	SLR Predictions	CR Predictions	Observed	CR Relative Error	SR Relative Error
I80_01	4576	551	N.O	N.A	N.A
I80_02	7015	4885	7178	32%	2%
I80_03	694	2038	N.O	N.A	N.A
I80_04	4831	4773	7928	40%	39%
I80_05	-9889	14564	N.O	N.A	N.A
I80_06	6778	4994	N.O	N.A	N.A
I80_07	25284	11773	7426	37%	71%
I80_08	20583	6182	12913	52%	37%
I80_09	57541	45608	38518	16%	33%
I80_10	22793	19192	11353	41%	50%
I80_11	29546	24168	5014	79%	83%
I80_12	38607	28090	7906	72%	80%

N.O: No Observations, N.A.: Not Applicable

Table 5: Observed, Predicted and Relative Errors of Truck Volumes for Highway US 206 (Stepwise and Constrained Regression)

Highway Section	SLR Predictions	CR Predictions	Observed	CR Relative Error	SR Relative Error
US206_01	-324	221	675	67%	148.0%
US206_02	1144	941	363	61%	68.3%
US206_03	931	419	848	51%	8.9%
US206_04	860	1263	N.O	N.A	N.A
US206_05	1289	539	N.O	N.A	N.A
US206_06	1474	433	N.O	N.A	N.A
US206_07	1117	728	180	75%	83.9%
US206_08	1455	603	N.O	N.A	N.A
US206_09	1261	503	N.O	N.A	N.A
US206_10	869	685	N.O	N.A	N.A
US206_11	2416	1894	N.O	N.A	N.A
US206_12	1395	1296	N.O	N.A	N.A
US206_13	1458	1562	N.O	N.A	N.A
US206_14	-296	1105	515	53%	157.5%
US206_15	-1200	4576	N.O	N.A	N.A
US206_16	1059	951	718	25%	32.2%
US206_17	2261	288	1848	84%	18.3%
US206_18	1976	577	526	9%	73.4%
US206_19	1635	706	437	38%	73.3%
US206_20	2393	665	364	45%	84.8%
US206_21	2689	644	450	30%	83.3%
US206_22	375	191	319	40%	14.9%
US206_23	-130	400	961	58%	113.5%
US206_24	863	54	194	72%	77.5%

N.O: No Observations, N.A.: Not Applicable

intercept should be independent of the band buffer area used to train the models and thus changing the band size should not significantly affect this value. Table 6 presents the intercept of each model for each type of roadway.

When SLR was used, the limited training data forced the intercept to become correlated to the independent variables and vary with the band buffer size. In RR and LR, the intercept is calculated as the mean value of the predicted variable and is constant for all the models. This creates the problem, that under a limited training dataset, the calculated intercept may be over-estimated. Using PLSR and CR, the intercept is part of the decision variables and is calculated along with the rest of the beta coefficients of each model. It should be emphasized that PLSR and CR produced approximately the same intercept for all nine different bandwidths (0.25-5 miles) for each FC. The values of the intercept also indicate that the models account to a certain extent for the through (non-local) traffic. High intercept values indicate that through trucks use major facilities, such as interstates and expressways. In lower level highways, the small intercept value indicates that the truck traffic depends primarily on the local socioeconomic activity, which generates local traffic.

GIS MODEL IMPLEMENTATION

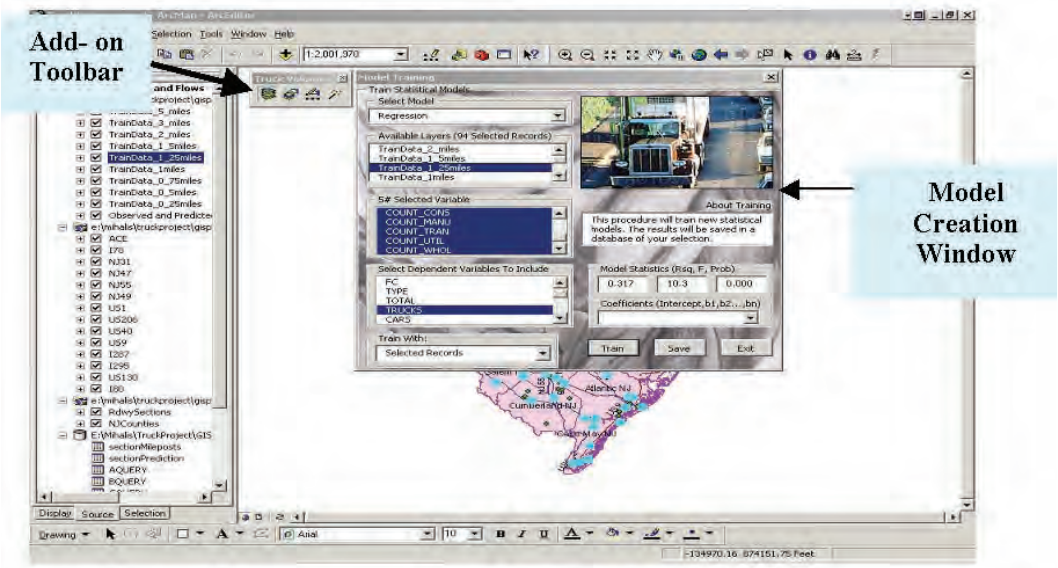
One major disadvantage of the proposed methods, making them unattractive to the transportation practitioner, is that their use necessitates some degree of sophisticated computation, and their application depends on the availability of software. In order to make the algorithms more widely and easily applicable to the practitioner, an add-on toolbar that incorporates these methods within a GIS environment (Boile and Golias 2004) has been developed. This tool has been created as an easy-to-use, in-house application (Figure 7) for state DOTs and MPO transportation planners, giving them the ability to develop regression models and use them in order to obtain an estimate of truck activity throughout the state.

Table 6: Variation of Intercept with Band Buffer Size

Statistical Technique Used	Band (miles)	Intercept per FC (number of vehicles)					
		FC 1-2	FC 6-9	FC 11	FC 12	FC 14	FC 16-19
CR	0.25-5	48	3	267	110	26	4
RR & LR	0.25-5	900	68	5000	2200	500	70
PLSR	0.25-5	1096	267	5764	2319	611	99
SLR	0.25	412[a]	75	6057	2766	1501[a]	119[a]
	0.5	677	64[a]	9999[a]	NM[b]	NM[b]	123
	0.75	1008	89	6374	4890	NM[b]	119
	1	1116	101	5537	3505[a]	NM[b]	125
	1.25	609	68	5396	3977	NM[b]	127
	1.5	1210	63	1453	3782	NM[b]	149
	2	354	37	4870	4454	2200	130
	3	659	-6	5059	7224	NM[b]	171
	5	1333	-56	11348	2845	2429	168

[a]SLR Best Model, [b]N.M: No Model

Figure 7: GIS Add-on Toolbar



The tool allows for the creation and update of regression models using the statistical methods presented in the paper. Furthermore, it allows for the creation and display of thematic maps of truck volumes (both observed and estimated) on state highways, taking advantage of the GIS capabilities of the software. The desktop platform for this application is ArcInfo 9.0 by ESRI.[15] The add-on toolbar is available at http://www.cait.rutgers.edu/miemp/ as a beta version.

CONCLUSIONS

This paper described the formulation and implementation of different regression techniques in vehicle-based freight modeling under limited training data, following the work of Mittal et al. (2004). The objective was to use "*simple and easy-to-use"* techniques that are powerful enough to provide more robust models than SR, and provide a comparative analysis of their performance using a real world example. Five different algorithms, including SR, were used and linear relationships between truck traffic volumes on roadways and their adjacent land use and economic activity were created. Due to the limited dataset, cross validation was not feasible and thus generalizations on the models' performance should not be made. The implementation of the proposed techniques, however, showed that these algorithms have the computational ability to overcome many of the problems that OLS and SLR face. In practice, when dealing with problems under limited data, all the techniques should be implemented and the one with the best results should be selected. The criterion for selecting the best models should be a combination of the models R^2 value, the significance of the model and its parameters (p-value or t-statistic of the model and the regression coefficients), and if data availability permits, on cross validation.

The major advantage of these algorithms is that they are conceptually easy to apply and are part of many statistical packages, which facilitates their implementation. To further facilitate their use, the procedure has been automated within a GIS environment. This tool provides the framework for analyzing limited transportation data in an efficient manner.

Endnotes

1. Training dataset: Dataset used to create the model.
2. The number of observations is less then the number of predictors.
3. Omitting one or two observations or omitting one variable from the same dataset may produce significant changes in the regression coefficients.
4. Shrinkage refers to the decrease of the regression coefficient ($\boldsymbol{b}_j$) values compared to the OLS estimate.
5. An excellent literature review is provided to support the conclusions.
6. $\text{Mean Square Error} = \sum_{j=1}^{m} Var(\bar{b}_j) + \sum_{j=1}^{m} [bias(\bar{b}_j)]^2$
7. Cross validation is a model evaluation method that is better than residuals. For further information the interested reader is referred to: http://www.cs.cmu.edu/~schneide/tut5/node42.html.
8. Large number of independent variables.
9. A data-manipulation software package that allows data to be analyzed and visualized using existing functions and user-designed programs.
10. www.esribis.com
11. Area around a highway section that could produce/attract truck traffic.
12. Randomly select and clone a number of existing observations.
13. Bootstrapping is a method for estimating the sampling distribution of an estimator by re-sampling with replacement from the original sample (http://www.icp.ucl.ac.be/~opperd/private/bootstrap.html).
14. Calculated on the estimation dataset.
15. http://www.esri.com/software/arcgis/arcinfo/index html

References

Abdi H. "Partial Least Squares (PLS) Regression." M. Lewis-Beck, A. Bryman, and T. Futing eds. *Encyclopedia of Social Sciences Research Methods*. Thousand Oaks CA: Sage (2003).

Allaman P.M., T.J. Tardiff, and F.C. Dunbar. *New Approaches to Understanding Travel Behavior*. National Cooperative Research Program Report 250, Transportation Research Board, Washington, D.C., 1982.

Boile, M. and M. Golias. "A Dynamic GIS-based Tool for Truck Volume Estimation." *Institute of Transportation Engineers Journal*, December (2004): 42-46.

De Jong S. "SIMPLS: An Alternative Approach to Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems* 18, (1993): 251–263.

Edlund, O. and H. Ekblom. "Computing the Constrained M-estimates for Regression." *Computational Statistic & Data Analysis* 49(1), (2005): 19-32.

Ellis, P.S. "Instability of Least Squares, Least Absolute Deviation and Least Median of Squares Linear Regression with a Comment by Stephen Portnoy and Ivan Mizera and a Rejoinder by the Author." *Statistical Science* 13 (4), (1998): 337–350.

Engelen, S. and M. Hubert. "Fast Model Selection for Robust Calibration Methods." *Analytica Chemica Acta* 544, (2005): 219-228.

Federal Highway Administration, Office of Planning and Environment, Technical Support Services for Planning Research. *Quick Response Freight Manual*. Report prepared by Cambridge Systematics, COMSIS, and the University of Wisconsin-Milwaukee, 1996.

Federal Highway Administration, Office of Highway Policy Information. *Traffic Monitoring Guide*. FHWA-PL-01-021 online document 2001.

Frank, I.E., J.H. Friedman, S. Wold, T. Hastie, and C. Mallows. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics* 35(2), (1993): 109-148.

Geweke, J. "Bayesian Inference for Linear Models Subject to Linear Inequality Constraints." In A. Zellner and J.S. Lee. eds. *Modeling and Prediction: Honoring Seymour Geisser.* New York, Springer, (1996).

Grandvalet, Y. "Least Absolute Shrinkage is Equivalent to Quadratic Penalization." *Perspectives in Neural Computing*, ICANN, (1998):201-206.

Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer Series in Statistics, New York, 2001.

Helland, I.S. "Some Theoretical Aspects of Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems* 58, (2001): 97-107.

Helland, I.S. and T. Almoy. "Comparison of Prediction Methods when Only a Few Components are Relevant." *Journal of American Statistical Association* 89, (1994): 583-591.

Hubert M. and K. Vanden Branden. "Robust Methods for Partial Least Squares Regression." *Journal of Chemometrics* 17, (2003): 537-549.

Hubert, M. and S. Verboven. "A Robust PCR Method for High-dimensional Regressors." *Journal of Chemometrics* 17, (2003): 438-452.

Institute of Transportation Engineers. *Trip Generation Handbook. 2nd ed.: An ITE Recommended Practice*. Institute of Transportation Engineers, 2003.

Klugkist, I. "Inequality Constrained Normal Linear Models." Dissertation (PhD). Utrech University, 2004.

Knautz, H. "Inference in Linear Models with Inequality Constrained Parameters." (www.rrz.uni-hamburg.de/IfStOek/zrabstr.htm).

Koenker, R., and P. Ng. "Inequality Constrained Quantile Regression." 2004. (www.econ.uiuc.edu/~roger/research/sparse/alg5.pdf).

Li, M. L. "An Algorithm for Computing Exact Least-Trimmed Squares Estimate of Simple Linear Regression with Constraints." *Computational Statistics & Data Analysis* 48(4), (2005): 717-734.

Mittal N., M. Golias, M. Boile', L. Spasovic, and K. Ozbay. "Estimating Truck Volumes on State Highways – A Statistical Approach." Presented at the 84th Annual Transportation Research Board, Washington D.C., 2004.

Mukerjee, H. and R. Tu. "Order-restricted Inferences in Linear Regression." *Journal of the American Statistical Association* 90, (1995): 717-728.

Transportation Research Board, National Academy Press, Truck Trip Generation Data. "A Synthesis of Highway Practice." Washington, D.C., 2001.

Neter, J, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. Irwin, Chicago, Illinois, 1996.

Ngo, S. H., S. Kemény and A. Deák. "Performance of the Ridge Regression Method as Applied to Complex Linear and Nonlinear Models." *Chemometrics and Intelligent Laboratory Systems* 67(1) , (2003): 69-78.

Ortuzar, J. D. and G.L. Willumsen. *Modeling Transport.* Third Edition, John Wiley and Sons, LTD, New York, 2001.

Pazzani, M. J. and S. D. Bay. "The Independent Sign Bias: Gaining Insight from Multiple Linear Regression." *In Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society*, 1999 (http://www.ics.uci.edu/~pazzani/Publications/).

Verboven, S. and M. Hubert. " LIBRA: A MATLAB Library for Robust Analysis." *Chemometrics and Intelligent Laboratory Systems* 75, (2005): 127-136.

Wakeling, IN and H.J.H. Macfie. "A Robust PLS Procedure." *Journal of Chemometrics* 6, (1992): 189–198.

Weinblatt, H. "Using Seasonal and Day-of-Week Factoring to Improve Estimates of Truck Vehicle Miles Traveled." *Transportation Research Record* 1522, (1996): 1-8.

Wentzell, P.D. and L.V. Montoto. "Comparison of Principal Components Regression and Partial Least Squares Regression Through Generic Simulations of Complex Mixtures." *Chemometrics and Intelligent Laboratory Systems* 65, (2003): 257-279.

Zhu, J., R.Santerre, and W.X. Chang. "A Bayesian Method for Linear Inequality-Constrained Adjustment and its Application to GPS Positioning." *Journal of Geodesy* 78, (2005): 528–534.

Acknowledgements

The work presented in this paper has been supported by a NJDOT grant and by the Center for Advanced Infrastructure and Transportation. The authors would also like to thank Dr. D. Madigan for comments on the choice and implementation of the statistical models. This support is gratefully acknowledged but implies no endorsement to the modeling approach and implementation or the findings. The authors would finally like to thank the four anonymous referees for their valuable input and corrections.

Maria Boilé *is assistant professor of transportation in the Department of Civil and Environmental Engineering and director of research and education of the Maritime Infrastructure Engineering and Management Program (MIEMP) of the Center for Advanced Infrastructure and Transportation (CAIT) at Rutgers University. Her areas of research and interest are intermodal network modeling, freight and maritime systems analysis, port and marine terminal operations and logistics systems. Recent research work includes empty intermodal container management, modeling of container terminal operations; modeling of shipper and carrier behavior and interactions within an intermodal network environment; estimating the economic impact of shipper and carrier decisions; commodity flow modeling; IT assisted Intermodal freight network modeling for transportation contingency planning; and statistical analysis of freight data. Boilé holds a M.S. degree in civil and environmental engineering from Rutgers University and a Ph.D. in transportation engineering from New Jersey Institute of Technology.*

Michail Golias *is a graduate research assistant with the Department of Civil Engineering at Rutgers University. He holds an MS in civil engineering from Rutgers University and is currently pursuing a Ph.D. in transportation engineering. His research interests and experience include intermodal and maritime transportation, port operations and management, freight network modeling, statistical analysis in transportation, and geographical information systems.*