

## **Freight Demand Statistical Modeling: A Classification and Review**

**Boile, M.P.**

Assistant Professor, Dept. of Civil and Environmental Engineering, Rutgers University, NJ  
Phone: (732) 445 7979, Fax: (732) 445 0577, E-mail: boile@rci.rutgers.edu

**Golias, M.**

Research Assistant, Dept. of Civil and Environmental Engineering, Rutgers University, NJ  
Phone: (732) 445 3162, Fax: (732) 445 0577, E-mail: golias@eden.rutgers.com

---

### **ABSTRACT**

Linear Regression is used as a prediction tool in transportation planning, traffic data analysis and safety. Many researchers have attempted to address several transportation issues using these types of models. The accuracy and stability of these models is mainly dependent on the size of the available data. Unlike other science fields, where sophisticated algorithms are used to deal with problems of small datasets, the majority of freight demand modeling relies on simple statistical techniques. Limitations of these modeling methodologies, caused by their high dependence on data availability, and several assumptions that need to be made can result in erroneous models. In cases in which limited data are available, more advanced algorithms that can be legitimately used on small datasets should be applied. In this paper a description and classification of algorithms and processes used for creating these types of models under limited data is presented. To demonstrate the applicability of these algorithms, along with implementation problems, limitations, and the different performance measures, a case study is used. Different models are created and results are presented and discussed.

---

### **INTRODUCTION**

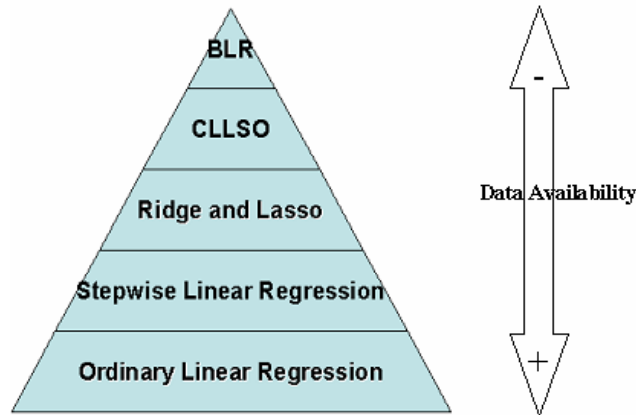
Linear regression is one of the most commonly used predictive tools in transportation planning and modeling applications. It is a simple statistical technique that enables relationships between an output (predicted) and several input (predictive) variables to be constructed. Transportation planners and engineers have extensively used different types of regression techniques to model various transportation problems that include traffic demand and supply, roadway accident prediction, pavement analysis, origin-destination matrix estimation, and in transportation economic analysis on problems such as the estimation of highway maintenance costs. These models range from simplistic to sophisticated depending on the type of problem. The accuracy and stability of these models is mainly dependent on the size of the available data.

Freight demand modeling and truck trip activity estimation is one of the transportation areas where linear regression algorithms have been extremely used. Insufficient data and accuracy is a critical limitation in freight demand modeling (Ortuzar and Willumsen 2001, Allaman et al. 1982) affecting the accuracy and validation methods of the produced models. Unlike other science fields where sophisticated algorithms are used to deal with problems of small datasets, the majority of freight demand modeling relies on simple statistical techniques such as ordinary and stepwise linear regression (NCHRP 298, 2001). Limitations of these modeling methodologies, caused by their high dependence on data availability, and several assumptions that need to be made can result in erroneous models. Transportation planners are forced to adopt other demand modeling techniques that encounter their own type of challenges (Holguin-Veras and Thorson, 2000) and do not necessarily perform better. In cases in which limited data are available, more advanced algorithms that can be legitimately used on small datasets should be applied. In this paper a description and classification of such algorithms and processes used for creating linear regression models under limited data is presented. A systematic approach for selecting the appropriate approach depending on the conditions of the problem is also obtainable within this paper.

### **BACKGROUND**

There are two main reasons why the least squares estimates, from linear regression models, may not be satisfactory: a) Prediction Accuracy, and b) Interpretation. Subsequently, computational problems from multi-collinearity between the predictors, instability of the estimated parameters and ill-conditioned problems occur under limited training data (Bjorkstrom 2001). As the available dataset decreases more advanced algorithms should be used that provide flexibility over the modeling process, allowing the introduction of assumptions that will make the model more representative of the problems' conditions and hypothesis. Furthermore algorithms that allow cross-validation to be performed without affecting the

accuracy of the produced model should be used so that validation measures are not merely based on the statistical properties of the model. Fig. 1 presents a suggestive approach selection hierarchy based on the size of the available training dataset of the algorithms presented in this paper.



**Fig. 1.** Statistical Approach Ranking with Data Availability

At the bottom of the pyramid (fig. 1) lies the classical ordinary least squares linear regression. If a substantial amount of training data is available (20-50 training cases per variable) then usually this method performs adequately and cross-validation can be used in order to test the model. As the available dataset decreases variable selection techniques come into play to reduce collinearity problems. Stepwise linear regression (SLR) enters the most significant variables into the model and removes the non-significant variables based on statistical criteria. The basic criterion for a variable to enter or exit the model is the F-statistic. Freedman, 1983, pointed out though that when many predictors are used and there is no relationship between the predictors and the response, classical variable selection techniques lead to models with high statistical goodness-of-fit measures ( $R^2$  and F values). Furthermore under limited training datasets the coefficients become correlated and are affected by outliers entering the model with incorrect size and sign (Pazzani and Bay 1999). Ridge and Lasso regression (Hastie et al. 2003) are shrinkage/variable selection methods that penalize the coefficients of the linear model. They both produce more stable results than simple regression and can partially remedy/reduce multi-collinearity effects. Based on the idea of penalization of the regression coefficients ordinary linear least squares regression can be examined as an optimization problem. The Constrained Linear Least Squares Optimization (CLLSO) algorithm uses an objective function that minimizes the sum of squares and at the same time adds constraints, not only to the values of the coefficients, but also to the values of the predicted variables, producing more stable and logical results. Linear regression can also be approached from its Bayesian perspective where both observable quantities and model parameters are considered to be random. Using a Bayesian framework allows the training of models that depart from the classical linear regression assumptions, i.e normality of the error terms, and independent observations with equal variances, allowing for a variety of parametric models with unequal variances and different error structures to be implemented and evaluated (Gelman et al., 2003).

In this paper these approaches are implemented on a vehicle-based freight-modeling problem with limited training data. Different models are trained and linear relationships between truck traffic volumes on roadways and their adjacent land use and economic activity are created. Through this case study the applicability of these algorithms, implementation problems, limitations, and the different performance measures are demonstrated. Furthermore different software packages that can implement the above approaches are presented and their performance is briefly discussed.

#### **STATISTICAL ALGORITHM DESCRIPTION**

In this section the general formulation of the statistical techniques previously mentioned is described.

##### **Stepwise Linear Regression (OLR)**

Suppose we have a training set  $(X_{ij}, y_1), \dots, (X_{ij}, y_i)$ , where  $j=1, \dots, m$  number of predictors and  $i=1, \dots, n$  number of training cases,  $X_{ij}$  are column vectors in  $R$  and  $y_i \in R$ ,  $i=1, 2, 3, \dots, n$ . The comparison class consists of the linear function  $Y=X*b$  (throughout this paper  $X=\{X_{1j}, X_{2j}, \dots, X_{ij}\}$  is referred to as the

independent variable dataset and  $Y=\{y_1, y_2, \dots, y_i\}$  as the dependent variable dataset). The least squares linear regression method recommends computing the column vector  $\hat{b}$  (coefficient vector) that minimizes:

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

where: bo is the intercept

The 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. Often, it is not known which independent variables should be included in the model. In order to select among a set of candidate models, different approaches have attracted considerable attention that include forward, backward and stepwise regression, model choice criteria (Akaike Information Criterion (Akaike, 1973) and Bayes Information Criterion (Schwarz, 1978) which are based on the maximum likelihood estimates of the models' parameters) and different Bayesian techniques. SLR is one of the best-known approaches for variable selection. Drawbacks of this method include the choice for the appropriate value of the F-statistic and the difficulty in performing cross-validation under limited training data. On the other hand SLR can be easily performed using standard software statistical packages (SAS, SPSS, Minitab) with minimum computational effort.

### Ridge Regression and Lasso Regression (RR and LR)

RR and LR (Hastie et al. 2003) regression are linear regression methods that penalize the coefficients of the linear model. The formulas for both methods are given below in equations (2) and (3). In these equations bo is the intercept,  $b_j$  are the regression coefficients, and  $X_{ij}$  is the value of the independent variable j at  $y_i$ . The difference between these two algorithms is that while RR does not omit any of the independent variables, LR, due to the type of the constraint used, can zero-out some of the coefficients. Adjusting for the tuning parameters s and t produces different model estimates. Validation of RR and LR models relies on the calculation of the  $R^2$  value and both methods respond similarly to cross-validation under limited training data, as SLR.

$$\text{RR: } \hat{b} = \arg \min_b \sum_{i=1}^N (y_i - bo - \sum_{j=1}^p X_{ij} \hat{b}_j)^2, \text{ s.t.: } \sum_{j=1}^p \hat{b}_j^2 \leq s \quad (2)$$

$$\text{LR: } \hat{b} = \arg \min_b \sum_{i=1}^N (y_i - bo - \sum_{j=1}^p X_{ij} \hat{b}_j)^2, \text{ s.t.: } \sum_{j=1}^p |\hat{b}_j| \leq t \quad (3)$$

It can be shown (Hastie et al. 2003) that RR has a closed form solution as shown in equation 4 and is easy to implement. The independent set of observations is first standardized and the intercept is calculated as the mean value of the  $y_i$ 's as shown in equation 5.

$$\bar{b} = (X^T X + \lambda I)^{-1} X^T Y \quad (4)$$

$$bo = \frac{1}{N} \sum_{i=1}^N y_i \quad (5)$$

where:  $\lambda$  is a complexity parameter and  $I$  is the identity matrix

Grandvalet, 1998 proved that Least Absolute Shrinkage is equivalent to Quadratic Penalization and derived an EM (expectation maximization) algorithm. This allows for the computation of the LR solution and has been used in this paper. One drawback of these two methods is the difficulty in deciding on the values of the tuning parameters. Usually, cross-validation is used but this requires a significant amount of training data to be available. Under limited training data instead of cross-classification multiple values for the parameters can be used. As shown in figure 1 the values for the tuning parameters are first initialized. RR

and LR are performed and if all the predictions ( $\hat{y}_i = \sum_j X_{ij} * \hat{b}_j$ ) are positive the process stops. If not the

parameters are decreased by 5% and the algorithms are re-performed. Out of all the different values used, the ones that produce the highest  $R^2$  value and the most positive predictions should be chosen. RR and LR are algorithms conceptually easy to apply. Unfortunately, only RR is part of the most common statistical packages (SAS, SPSS, MatLab, R, SPLUS), while LR requires programming effort.

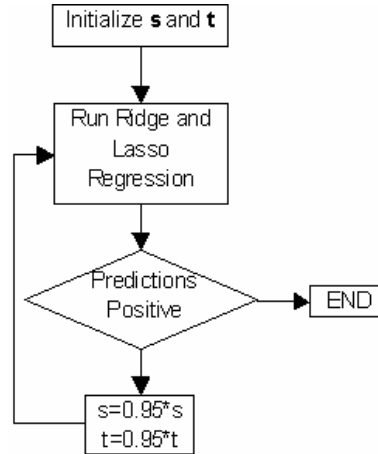


Fig. 1 RR and LR Tuning Parameter Value Selection Process

### Constrained Linear Least Squares Optimization (CLLSO)

Linear regression modeling with inequality constraints arises very commonly in the literature (Liew 1976, Judge and Takayama 1966). Based on the idea of establishing constraints on the regression coefficients the use of the CLLSO algorithm can be introduced as means of creating more logical and stable models. The CLLSO algorithm uses an objective function, as shown in equation 6, that minimizes the sum of squares and at the same time adds constraints, not only to the values of the coefficients, but also to the values of the predicted variables.

#### CLLSO Formulation

$$\min_b \left[ \frac{1}{2} (X_{ij} \bar{b}_j - y_i)^2 \right] \quad (6)$$

$$\text{s.t.: } X_{ij} \bar{b}_j \leq d_i \quad (6a)$$

$$X_{ij} \bar{b}_j = d^{eq}_i \quad (6b)$$

$$lb \leq b_j \leq ub \quad (6c)$$

where:  $lb$  and  $ub$  are the lower and upper bound vectors for the beta values and  $d_i$  and  $d^{eq}_i$  are the upper and equality bound for prediction

The above constraints provide a better control over the logic of the model formulation. From the engineering point of view the first constraint (6a) captures the range of the expectation for the predicted variable. This way the model takes into account uncertainty of the accuracy on the measurement of each station. The second constraint (6b) can be considered as a weighting factor for the predicted variable used for training. At some cases it is known that the observed measurement is accurate and at some cases it may not be very accurate. Setting up equality constraints for some or all of the accurate measurements forces the model to give more weight to them minimizing the transferring of error that exists in the observed measurement. The third constraint (6c) can be considered as 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 we can constrain that variable's beta coefficient to positive values and vice versa. This is especially important since outliers can seriously bias the results by "pulling" or "pushing" the regression line in a particular direction, thereby leading to biased regression coefficients. Similar to RR and LR an iterative process (fig. 2) can be used to select the two bounds so that feasibility is obtained. The values of the lower and upper bounds are first initialized (lower bound = 75%, upper bound = 100%) and then the CLLSO 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 CLLSO is re-performed. This process continues until both bounds provide a feasible solution.

The structure of the algorithm relieves the variable selection process from statistical criteria decisions usually found in OLR, RR, and LR i.e. F-value range, tuning parameter values etc. We should note that the models' goodness of fit is still based on the  $R^2$  value and, that if no constraints are used, the prediction corresponds to the least squares regression solution. The flexibility and power of this algorithm allows for the initial consideration of any number of prediction variables to be included in the model without any computational burden, alleviating the planner from previous limitations of similar models, where only a few and specific type of variables were used. The CLLSO algorithm eliminates variables that have no predictive power without adding any complexity to the training or use of the model(s). Cross-validation is feasible and efficient only by relaxation of the constraint bounds that can resolve in a decrease of the models'  $R^2$  value. This algorithm can be easily implemented in any software that implements optimization procedures, such as Matlab.

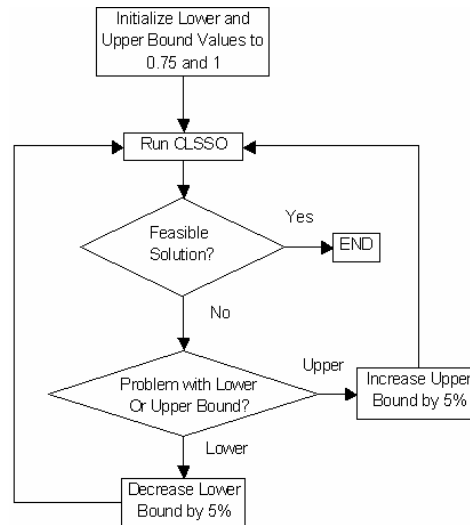


Fig. 2. Iterative Process for Upper and Lower Bound Determination

### Bayesian Linear Regression (BLR)

Looking at the linear regression problem from its Bayesian perspective both observable quantities ( $Y$ ) and parameters ( $\beta$ =regression coefficients) are considered to be random. The components of a Bayesian inference problem can be identified as: a) the prior distribution of the parameters involved ( $P(\beta)$ , and  $P(Y)$ ) that expresses the uncertainty or the information that is available at the start of the study about the unknown variables by means of a probability distribution, b) the likelihood of the data given the unknown parameters that relates all the variables into a full probability model that summarizes the current knowledge of the phenomenon, and c) the posterior distribution for the unknown parameters ( $P(\hat{\beta}|X, Y)$ , and  $P(\hat{Y}|\hat{\beta}, X)$ ), that expresses our uncertainty about the parameters after seeing the data. The task of each Bayesian analysis is to build a model for the relationship between parameters and observable, and then calculate the probability distribution of the parameters conditional on the data. In addition, the Bayesian analysis may calculate the predicted distribution of unobserved data  $P(Y'|\hat{\beta}, X')$ , where  $X'$  is the new input data and  $Y'$  are the new predictions). This of course is not a free-trouble method. Advantages and disadvantages of the Bayesian approach are summarized in table 1.

In the case of OLR the observations are assumed to be independent and have equal variation:  $Y|\beta, \sigma^2, X \sim N(X*\beta, \sigma^2 * I)$ , where  $I$  is the identity  $n*n$  matrix,  $n$  is the number of cases (observations), and  $\sigma^2$  is the variance. This case of regression makes several assumptions: a) normality of the error terms, and b) independent observations with equal variances. As mentioned before BLR allows the training of models that depart from these assumptions and a variety of parametric models for unequal variances and different error structures have been successfully used (Gelman et al., 2003).

Many methodologies have been proposed in the context of Bayesian regression model/variable selection. Some of the papers proposing related procedures include: a) the Stochastic Search Variable Selection (SSVS) of George and McCulloch (1993), b) the model selection approach of Carlin and Chib

(1995), c) model averaging and accounting for a models uncertainty using ‘Occam’s Window’ by Madigan and Raftery (1993) d) simultaneous variable selection and outlier identification based on the computation of posterior model probabilities by Hoeting et. al. (1996), and e) the Gibbs Variable Selection (GVS) by Dellaportas et. al. (2000, 2002). In this paper we present a Bayesian hierarchical setup, analogous to the ones that exist in the literature used to perform Gibbs variable selection.

**Table 1.** Advantages and Disadvantages of Bayesian Inference Methods

<b>Advantages of the Bayesian Approach</b>	<b>Disadvantages of the Bayesian Approach</b>
Basis of Inference is Probability Theory	Inferences Need to Be Justified
Less Computational Burden for Small/Medium Problems	Computational Burden for Very Complex Models
No Need for Significance Tests, P-values etc	Reasonable Prior Distribution Selection
Complex models to meet reality demands	Model Adequacy for the Data

Let us assume a linear model of the form:  $Y_i = \sum_{j=1}^n \gamma_j X_{ij} \beta_j$ , where  $X_{ij}$  is the design covariate matrix and  $\beta_j$  the parameter vector (regression coefficients) of the  $j^{\text{th}}$  term. The indicator  $\gamma_j$  identifies if covariate  $j$  will enter the final model ( $\gamma_j=1$ ) or not ( $\gamma_j=0$ ). The model likelihood takes the form of:  $f(Y | \beta, \gamma)$ , while the model prior the form of:  $f(\beta, \gamma) = f(\beta | \gamma) * f(\gamma)$ . We denote  $f(\beta | \gamma)$  as the coefficient prior and by  $\beta_\gamma$  and  $\beta_{\setminus \gamma}$  the coefficients included and excluded from the final model. The covariates included and excluded in each model are then sampled by:  $f(\beta_\gamma | \beta_{\setminus \gamma}, \gamma, y) \propto f(y | \beta, \gamma) * f(\beta_\gamma) * f(\beta_{\setminus \gamma} | \beta_\gamma, \gamma)$  and  $f(\beta_{\setminus \gamma} | \beta_\gamma, \gamma, y) \propto f(\beta_{\setminus \gamma} | \beta_\gamma, \gamma)$  respectively, an assumption introduced by Carlin and Chib, 1995. The variable indicator  $\gamma_i$  is sampled from a Bernoulli distribution with success probability  $a_j$ . Typically, Bernoulli priors with probability 0.5 can be assigned to the probability of each selection index  $\gamma_j$  being 1 (Congdon, 2003). In order to quantify uncertainty of the success probability, a hierarchical framework is introduced where the success probability follows a beta distribution:  $a_j \sim B(\alpha_1, \alpha_2)$ . In most cases introducing a third level of hierarchy by entering a probability distribution for the parameters  $(\alpha_1, \alpha_2)$  increases the level of detail, but not necessarily the level of accuracy (Gelman et. al., 2003). An indicative prior distribution for both  $\alpha_1$ , and  $\alpha_2$  is the uniform between 1 and 50.

An advantage of the Bayesian framework is that it allows the modeler to introduce certain constraints to the maximum and minimum values of the predicted variable and the regression coefficients, similarly to the CLLSO approach, by truncation of the prior distributions (right and left truncation respectively). The major advantage of BLR though lies in the fact that multi-case cross-validation can be performed under limited training datasets without affecting the accuracy of the final models. This means that testing results obtained from a BLR model are what should be expected from the models’ performance in a real world application. Two main drawbacks of this approach are: a) the effort required to build the models and b) the computational time that increases with the available data. Furthermore, justification of the proposed prior distributions can be difficult to establish. The question that should be asked though is: “If a model performs adequately in real life problems is any other justification needed?”.

In this paper WinBugs has been used to build and implement the BLR models. WinBugs is an open source software package that enables a flexible approach to Bayesian modeling, in which the specification of the full conditional densities is not necessary and so small changes in program code can achieve a wide variation in modeling options. This enables sensitivity analysis to likelihood and prior assumptions to be performed with ease. Additional software packages that can perform Bayesian analysis, with more intensive programming effort, include R, MatLab and JAGS.

## **CASE STUDY: VEHICLE-BASED TRUCK VOLUME ESTIMATION**

### **DATA DESCRIPTION**

The statistical methods were implemented with data consisting of classification traffic counts as the dependent variable and socioeconomic data as the independent variables. The dependent dataset was obtained from various locations throughout New Jersey. It consisted of 270 long and short duration truck traffic counts (vehicle classes 5 through 13). Long duration counts were obtained by permanent Weight-In-

Motion (WIM) locations. All short duration vehicle classification counts were adjusted for axle correction and pattern factors. The data for the independent variable dataset included, population, number of employees, sales volume, and number of establishments for each SIC code. In total 34 independent variables were included in the final training process. Both the dependent and the independent variables and the estimates are based on year 2001 data.

Econometric data associated with these sections was extracted and used as input in the model training and testing process. ArcView, a GIS software package, was used in order to buffer and aggregate the independent variable dataset for 9 different bandwidths of influence (0.25, 0.50, 0.75, 1.0, 1.25, 1.5, 2, 3 and 5 miles). 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 models accuracy fluctuates). Employing 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 6 subsets (Table 1) according to the functional class (FC) of the roadway. Building models by considering roadway classes is significant as different roadways attract different truck volumes that are dependent on different variables. Roadways are classified under different FC 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 FC 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 FORMULATION

Several models were created using: a) SLR, b) RR, c) LR, d) CLLSO, and e) BLR. The SLR approach is similar to OLR and is very easy to implement. For that reason the formulation of the SLR models are omitted from this paper. An analytical description of SLR modeling can be found at Kleinbaum et al., 1988.

### RR and LR Models

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-classification to be performed. Instead multiple values for the parameters were used. The values of the tuning parameters were calculated following the procedure described in section 3. Out of all the different values used two were chosen for both approaches: a) the one that produces the model with the highest  $R^2$  value, and b) the one that produces a model with all the predictions positive. Computational time for each functional class varied from a few seconds to a maximum of 1 minute.

### CLLSO Model

The CLLSO model implemented for this study is given in equations 7a-7c.

#### CLLSO Final Formulation

$$\min_b \left[ \frac{1}{2} (X_{ij} \bar{b}_j - y_i)^2 \right] \quad (7a)$$

$$\text{st.}: 0.25 * y_i \leq X_{ij} \bar{b}_j \leq 1.25 * y_i \quad (7b)$$

$$0 \leq b_j \quad (7c)$$

CLLSO offers the advantage of controlling the values of the predictions in terms of size and sign. For both the coefficients and the predicted variables, corresponding to the functional class of the highway and the geographical location of the count, the constraints on the minimum and maximum value of the expected traffic volumes may vary so that the models account for space variations.

The second constraint (7b) requires that the value of the estimated truck volume 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. The limitation of using constraint 7b is that for relatively small training datasets and strict lower and upper bounds the solution may be infeasible. A pseudo-increase of the data was performed for all the subsets and the results showed that both interval bounds are positively correlated to the amount of the training data. In this study lower and upper bounds (0.25 and 1.25 respectively in equation 7b) were determined using the iterative process described in section 3. Computational time for each functional class varied from a few seconds to a maximum of 3 minutes.

The third constraint (7c) indicates that the predictive variables should have a positive effect on truck volume production. This constrain 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.

**BLR Model**

For the Bayesian approach a linear regression form on the expectation of Y (predicted truck volumes), with a variety of different error structures was assumed. Specifically:

$$mu_i = \beta_0 + \beta_{i1} + \beta_{i2} + ..... + \beta_{ij} \quad (8)$$

$$Y_i \sim N(mu_i, 1/\sigma^2) \quad (9)$$

where j=1:34, i=1:n, n=number of training cases, m<sub>i</sub> is the mean, σ<sup>2</sup> is the standard deviation

Priors for the regression coefficients were set to a neutral value so that all the terms have priory a zero mean value (Dellaportas and Forster, 1999). Further prior information for the beta values did not exist. Results from the SLR models were used as prior information for the intercept that was removed from models with small values (FC=6-9, FC=14, FC=16-19). The assumption of zero intercept for models used on local access roads is valid since truck traffic on these types of roadways should not be expected if the traffic generating variables are all zero. Both distributions (beta coefficients: β and intercept: inter) were truncated at zero (10 and 11). A gamma distribution (12) was used instead of a vague prior for the coefficient precision (betaTau<sub>j</sub>).

$$\beta_j \sim N(0, betaTau_j) \quad (10)$$

$$inter \sim N(0, 1.0E - 6) \quad (11)$$

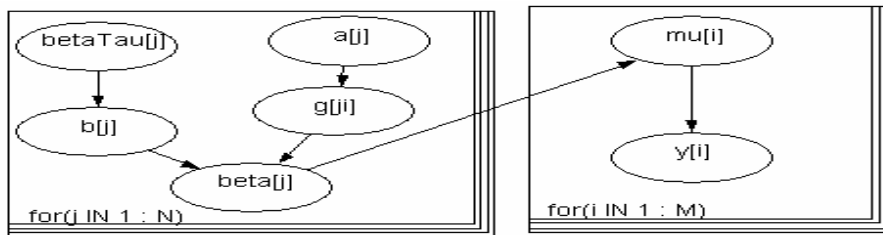
$$betaTau_j \sim Gamma(1.0E - 2, 1.0E - 2) \quad (12)$$

A Bernoulli distribution, with success probability a<sub>j</sub> (13), was used as the means for the variable selection. In order to quantify the uncertainty of the success probability, a hierarchical framework was introduced (14 and 15) where the success probability follows a beta distribution. Using this distribution for the success probability we assume that priory all of the covariates have the same probability (50%) of entering the model. The full model is graphically presented in fig.3

$$\gamma_j \sim Bern(a_j) \quad (13)$$

$$beta_j = \beta_j * \gamma_j \quad (14)$$

$$a_j \sim Beta(2,2) \quad (15)$$



Note: g=γ, b=β

Fig. 3. WinBugs Graphical Presentation of the Model

**MODEL EVALUATION**

This part of the paper evaluates and compares the performance of each approach. Furthermore, at the end of the evaluation and based on the results of the case study, a scoring table is presented.

For the first part of the evaluation the  $R^2$  values of the best model from each approach for each roadway functional class are compared. For all approaches 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 buffer zone size of influence of the independent variable considered. Table 3 and 4 present the best  $R^2$  value for each type of roadway and the corresponding band buffer used to extract the socioeconomic data. Table 3 and 4 show that higher-level roadways (Expressways (FC=12) and Urban interstates (FC>12)) have a larger optimal band size compared to lower level roadways. This result satisfies the underlying assumption that trucks will use local roads only to access local 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 RR, LR, and CLLSO produce models with more reasonable  $R^2$  values. CLLSO models managed to meet both of the criteria, set in equation 8, and produce better results than SLR. 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. As the values of  $s$  and  $t$  are increased constrains in (2) and (3) become less restrictive and the solution approaches the least squares. When the value of the tuning parameters was decreased, the predictions were positive but the  $R^2$  value was below satisfactory levels (as set in equation 8). Compared to the SLR approach however, RR and LR 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. The two different models that are presented in table 3 for the RR and LR approach as described in the Model Formulation section.

Table 3.  $R^2$  values and Band Buffer for the best model for each FC (CCLSO and SLR)

FC	CLLSO (No $R^2$ Problem, No Prediction Negativity Problems)		SLR (Prediction Negativity Problem and $R^2$ Problem)		RR (Negative Predictions)		RR ( $R^2$ Problem, No Prediction Negativity Problem)		LR (Negative Predictions)		LR ( $R^2$ Problem, No Prediction Negativity Problem)	
	$R^2$	Band	$R^2$	Band	$R^2$	Band	$R^2$	Band	$R^2$	Band	$R^2$	Band
1-2	0.82	0.25	0.97	0.25	0.9	0.25	0.54	0.25	0.65*	0.25	0.65	0.25
6-9	0.79	0.25	0.84	0.5	0.85	0.25	0.62	0.25	0.76*	0.25	0.76	0.25
11	0.77	0.5	0.92	0.75	0.55	0.5	0.34	0.5	0.75	0.5	0.41	0.5
12	0.87	0.75	0.99	1.0	0.77	1.0	0.28	1.0	0.65	1.0	0.44	1.0
14	0.87	1.0	0.13	0.25	0.49**	1.0	0.1	1.0	0.29**	1.0	0.1	1.0
16-19	0.82	1.25	0.59	0.25	0.44**	1.25	0.1	1.25	0.38**	1.25	0.18	1.25

\*No Negative Prediction Problem, \*\*  $R^2$  Problem

Table 4 summarizes the  $R^2$  values of the BLR models BLR produces a distribution for the predicted truck volumes and not a point estimate. Thus for the Bayesian approach 3 different  $R^2$  values are presented each corresponding to the 2.5%, median and 97.5% interval of the predicted truck volumes. It can be seen that over-fitting has been remedied (Median  $R^2$  values). Looking at the  $R^2$  values the RR, LR, and BLR models for FC=12 do not perform adequately, something that was expected since only 20 observations were available, with an  $R^2$  range from 0.08 to 0.15. On the other hand the SLR model assigns 99% accuracy to a model, showing a complete over-fit of the training data, and should be rejected. Following the same pattern CLLSO overestimates the accuracy of the model for FC=12.

Table 4.  $R^2$  Values from Bayesian Regression and SLR

FC	Band Used	$R^2$		
		2.50%	Median	97.50%
1-2	0.25	0.44	0.54	0.84
6-9	0.50	0.37	0.50	0.56
11	0.50	0.54	0.63	0.99
12	0.75	0.08	0.08	0.15
14	1.00	0.14	0.12	0.15
16-19	1.25	0.48	0.36	0.04

The second part of the evaluation compared the predictive power of the models on 14 selected highways. Although RR and LR models had a better predictive power (less negative predictions) than SLR, results are presented only for CLLSO since it was the method with the strongest predictive power (zero negativity in the predictions). Results for highway US9 and US206 are presented in Fig. 3 and Fig. 4. These figures show predicted truck volumes for each section of the highway (light blue for SLR, red for CLLSO). Observed counts (yellow) are also shown for sections of the highway, for which such information exists. As can be seen in figures 4 and 5, the negativity problem in the predictions has been answered. It is also obvious that the CLLSO approach tends to reduce, but not eliminate, the over-estimation problem. This pattern is followed in all the 14 highways (205 sections) that were selected to test the models.

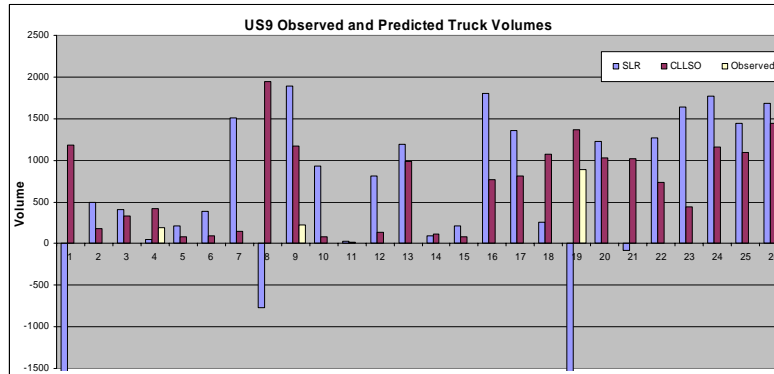


Fig. 4. Observed and Predicted Truck Volumes from CLLSO and SLR models for Highway US9

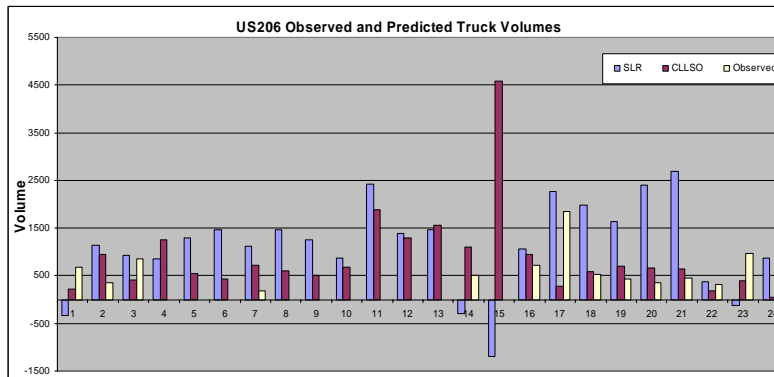


Fig. 5. Observed and Predicted Truck Volumes from CLLSO and SLR models for Highway US206

Table 5 presents estimations from SLR, CLLSO, and BLR models for the 6 highway functional classes that were defined. Predictions from the BLR models are made using selective multi-case cross-validation. The cases removed and used for validation purposes were selected to cover the full range of the observed truck volumes of each training dataset. Thus we were able to test if the posterior distributions cover the full range of the expected truck volumes. FC=12, where a single-case omission is used, corresponds to the exception, due to the extremely small number of observed cases (20 cases with 34 independent variables). Predictions from the SLR and CLLSO models are made from data already used for training the models for reasons explained in section 2. This means that though the results obtained from the Bayesian approach are what should be expected from the models' performance in a real world application, claiming the same for the other models is rather questionable. As can be seen in table 5 for the BLR models, the majority of the observed truck volumes lie between the predicted range for the Bayesian models and the negativity problem has been solved.

Some of the locations on the 14 highways experienced extreme values in the predictions. One major advantage of using CLLSO and BLR is the power to add constrains to the maximum and minimum values of the predicted variable via truncation. Table 6 presents results from 4 models (SLR, CLLSO, BLR, and BLR Right Truncated models) for a specific highway location (Highway US 1, FC=1-2) where an extreme prediction was produced.

**Table 5.** Observed and Predicted (SLR, RR, LR, BLR) Truck Volumes

FC	Observed	SLR	CLLSO	BLR			
				Mean	2.50%	Median	97.50%
1-2	3506	2427	4165	2845	705	2793	5149
	7178	4455	4885	3129	865	3100	5571
	1038	-571	108	1048	8	928	2730
6-9	165	280	176	311	15	273	818
	1266	1579	1296	812	96	778	1711
11	1514	-25284	7738	3115	2079	3084	4297
	3914	-304	1101	1775	784	1735	2982
	7426	3989	2060	6175	4386	6135	8157
12	2258	-4094	6187	5345	4475	5322	6348
14	1738	1443	1732	1159	64	1050	2924
	167	887	983	1156	58	1054	2801
	8497	909	1628	1366	87	1281	3171
	926	1547	402	1009	52	889	2619
16-19	515	-292	1105	304	140	286	574
	1618	530	1640	1947	1459	1942	2465
	178	1205	499	318	136	307	569
	885	1940	1469	839	551	828	1200
	310	497	442	707	35	574	2140
	48	149	168	303	2	268	795

**Table 6.** Extreme Prediction Value Problem and Solution

	Location: Highway US1, Section 13			
	Mean	2.50%	Median	97.50%
<b>Observed Truck Volumes</b>	7124	Does not apply		
<b>SLR</b>	273500	Does not apply		
<b>CLLSO</b>	27171	Does not apply		
<b>Bayesian Model</b>	86860	22580	65290	166400
<b>Truncated Bayesian Model</b>	15830	7686	16550	19840
<b>Difference Between Bayesian Models for US1, Section 13</b>	82%	66%	75%	88%

Initially the Bayesian model was used with no truncation. The result was an extremely high prediction following the pattern of the SLR model. We introduced a constraint, right truncating the posterior density of the predicted truck volumes so that the values should not exceed an upper limit of 20,000 (150% of the maximum observed truck volume for highways of FC=1-2). The truncated model was applied to make predictions for highway sections with FC=1-2. The change in the prediction for the specific location was more than satisfactory (86%) and had an insignificant effect to the remaining locations (change in the mean value of the predictions varied from 0% to 23% between the two Bayesian models, with 53 observations varying from 0 to 7%, 5 observations varying from 9% to 11%, and 6 observations varying from 15% to 23%).

Based on the case study results a performance matrix has been created (table 7) in which each row describes a different approach and each column describes the performance or the level of difficulty of each method against a criterion. The individual performance assessment is alphabetical with each letter corresponding to a specific numerical range in order to assist the model selection process. The expected consequences of each option are assigned a numerical score, between the suggested range, based on their strength for each option for each criterion and the modelers' opinion.

**Table 7. Suggestive Scoring of Different Approaches**

Method	Computational Effort	Cross Validation Accuracy		Models' Expected Accuracy		Software Availability/ Implementation Difficulty	Flexibility to Incorporate Extra Assumptions/ Hypothesis	Modeler Mathematical and Statistical Background
		Data (+)	Data (-)	Data (+)	Data (-)			
<b>OLS</b>	A	A	D	A	D	A	D	A
<b>SLR</b>	A	A	D	A	D	A	D	A
<b>RR</b>	B	A	C	A	C	B	D	B
<b>LR</b>	C	A	C	A	B-C	D	D	B
<b>CLLSO</b>	B-C	A	B	A	B-C	B	B	B
<b>BLR</b>	C-D	A	A	A	A	D	A	D

A=100-75, B=74-50, C=49-25, D=24-0

## CONCLUSIONS

In this paper a description and classification of algorithms and processes used for creating linear predictive models was presented. A case study was used and the applicability of these algorithms, implementation problems, limitations, and the different performance measures, was presented and discussed. A scoring table that can assist transportation planners and engineers in choosing the most appropriate approach based on several features of the problem, that are known at the beginning of a study, was also presented.

## ACKNOWLEDGEMENTS

Part of 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.

## REFERENCES

- Akaike, H. (1973). "Information theory and an extension of tmaximum likelihood principle." 2nd International Symposium Information Theory, Tsahkadsor 1971, 267-281.
- Allaman P.M., T.J. Tardiff, and F.C. Dunbar (1982) New Approaches to Understanding Travel Behavior. National Cooperative Research Program Report 250, Transportation Research Board, Washington, D.C.
- Bjorkstrom, A. (2001). "Ridge regression and inverse problems." Research Report in Mathematical Statistics, Stockholm University.
- Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society B, 157, 473-484.
- Congdon, P. (2003). Applied Bayesian Modelling. Willey Series in Probability and Statistics, John Willey and Sons Ltd, West Sussex, England.
- Dellaportas, P. and Forster, J.J. (1999). Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-linear Models. Biometrika, 86,615-633.
- Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2000). Bayesian Variable Selection Using the Gibbs Sampler. Generalized Linear Models: A Bayesian Perspective. D.K. Dey, S. Ghosh, and B Mallick, eds. New York: Marcel Dekker, 271-286, 2000.
- Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2002). On Bayesian Model and Variable Selection Using MCMC. Statistics and Computing. Vol. 12, 27-36.
- Freedman, D. (1983). "A Note on Screening Regression Equations." The American Statistician, Vol. 37, 152-155.
- Gelman, A., Carlin, B. J., Stern, S. H. and Rubin B. D. (2003) Bayesian Data Analysis, Second Edition, Chapman & Hall/CRC, Florida, US.
- George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. Journal of the American Statistical Association. 88, 881-889.
- Grandvalet, Y. (1998) "Least absolute shrinkage is equivalent to quadratic penalization." Perspectives in Neural Computing, ICANN'98,201-206.

- Hastie, T., Tibshirani, R., and Friedman J. (2003) *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer Series in Statistics, New York.
- Hoeting, J., Raftery, E. A., and Madigan D. (1996). A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression. *Computational Statistics and Data Analysis*. Vol. 22, 251-270.
- Holguín-Veras, J. and E. Thorson. (2000). "An Investigation of the Relationships Between the Trip Length Distributions in Commodity-based and Trip-based Freight Demand Modeling." *Transportation Research Record*, 1707, 37-48.
- Judge, G. C. and Takayama, T. (1966). Inequality Restrictions in Regression Analysis. *Journal of the American Statistical Association*. Vol. 61, 166-181.
- Liew, C. K. (1976). Inequality Constrained Least-Squares Estimation. *Journal of the American Statistical Association*. Vol. 71, 746-751.
- Madigan, D.M. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association*, 89, 1335-1346.
- NCHRP Synthesis 298, Truck Trip Generation Data. (2001). "A Synthesis of Highway Practice." Transportation Research Board, National Academy Press, Washington, D.C.
- Ortuzar, J. D., and Willumsen, G.L. (2001) *Modeling Transport*. Third Edition, John Wiley and Sons, LTD, New York.
- Pazzani, M. J. and Bay, S. D. (1999). The Independent Sign Bias: Gaining Insight from Multiple Linear Regression. In *Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society*.
- Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, Vol. 6, 461-464.
- Kleinbaum, G. D., Kupper, L. L., and Muller, E. K. (1988). *Applied regression analysis and other multivariable methods*. 2nd Edition, WS-Kent Publication Company, Boston, Massachusetts.