Truck Flow Prediction in Tennessee Using Various Regression Models

Semester project report for Data Mining (COMP 8118)

Prepared by Dimitrios Giampouranis (U00706450)

ABSTRACT

In this project a methodology for preparing truck GPS data and freight business data to predict truck flows is being presented. Various regression models, including trees and linear robust regression, are being tested for their efficiency to predict truck flows using freight business indices. Computational experiments show that some performed better than others with a general conclusion that freight business indices, such as freight tonnage, value, units, and milage can be used as indicators of how truck flow in the US behaves.

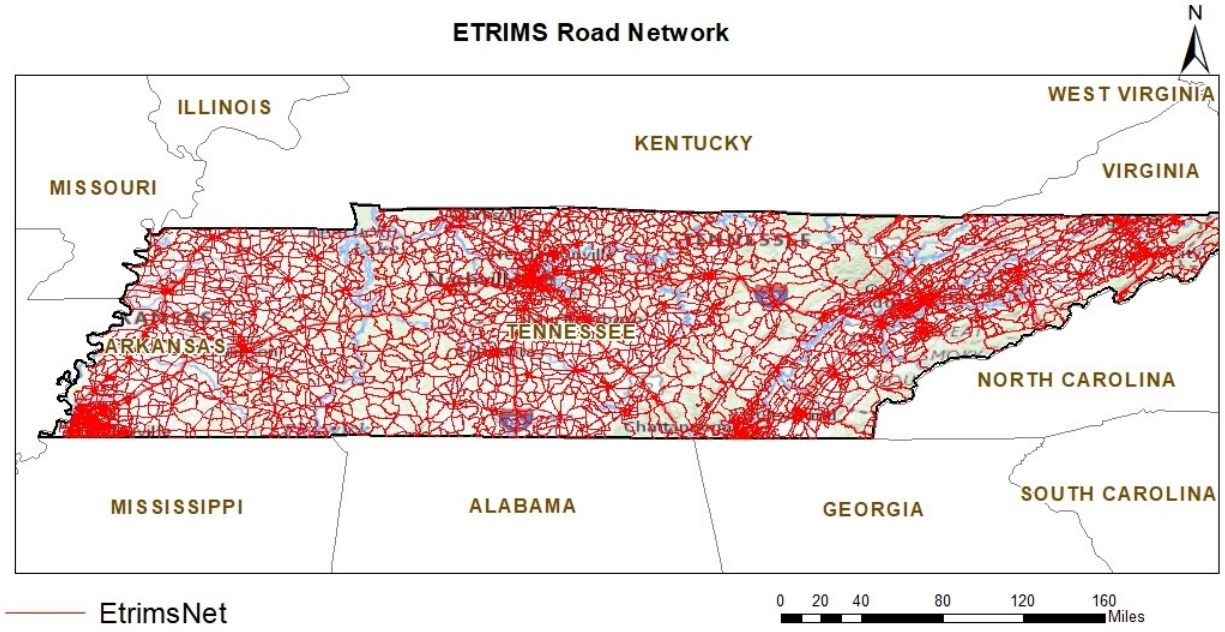
INTRODUCTION

Problem statement

The freight transportation system in the United States (U.S.) has one of the most valuable contributions to the nation's economy and growth. Long-term economic growth, as well as our nation's dramatic shift to e-commerce, is expected to result in even greater demand for truck transportation. Whether needed for planning, infrastructure, or policy making, being able to predict truck flows is a critical component of freight operations. Finding suitable predictors for truck flows is an important part of that. In this study the correlation between truck flows and freight tonnage, value, units, and mileage will be tested.

Datasets Description

ETRIMS Tennessee Road Network: This dataset was acquired from the ETRIMS (Enhanced Tennessee Roadway Information Management System) website. It contains the Tennessee road network shapefile as shown below.



ATRI truck GPS data: Raw truck GPS data for the month of October. For each GPS point, the set contains the truck ID, a timestamp (without time zone), the longitude and latitude, and instantaneous vehicle speed. A summary of the data is shown in the table below (Note: This dataset is confidential and thus cannot be provided as part of the deliverables)

# of GPS pings (in millions)	# of unique trucks	% of unique trucks in dataset
~171.4	~140K	~0.08%

IHS Global Insight Transearch Database: Freight tonnage, units, mileage, and value for freight movements in the USA by origin, destination, commodity, and transportation mode. Approximately 16 million records (movements) are included in this dataset. below (Note: This dataset is confidential and thus cannot be provided as part of the deliverables)

METHODOLOGY

Data preprocessing

A substantial part of this project was to prepare the datasets for use as input to create the various regression models (presented in the next section). This section shows the steps taken to prepare the datasets.:

Step 1: The Tennessee road network shapefile was loaded into PostgreSQL and a buffer of 100 feet was creates on every link of the network. Then, the .csv file containing all the GPS points was loaded into PostgreSQL as well. The timestamp of every GPS ping was adjusted to local Tennessee time zones based on their coordinates. After that the two layers (i.e., polygon and point layer) were intersected to produce hourly truck flows on the network for every day that data were available. The average flow for every hour was calculated in order to acquire the final product, which was daily truck flows for the State of Tennessee. The daily truck flows for every link of the network will act as the depended variable in the regression models described in the next section.

Step 2: The IHS Global Insight Database was loaded into PostgreSQL using MATLAB coding interface. At first, the dataset was filtered using SQL queries to produce two different tables. One with all the freight movements originated from Tennessee (i.e., productions table) and one with all the freight movements that had Tennessee as destination (i.e., attractions table). After that, the subtotals for every origin or destination and commodity type were calculated. Note that origins and destinations in this dataset are on the county level. In the next step, they will be transformed to the link level in order to match the GPS truck flow dataset from step 1.

Step 3: To transform the IHS dataset from county level to link level a weight was created for every link of the network. The flow of every link of each county were added to calculate a table that contains the truck flow of every county in Tennessee. Then, link flows were divided by county flows to create a weight that when multiplied with county freight indices (tonnage, units, mileage, and value) of the two tables (i.e., productions and attractions) will give us the freight indices of every link. This resulted to a table that has as many rows as the links in our network and one column for every freight index (tonnage, units, mileage, and value), for every commodity type, for both attractions and productions. This produced 300 different columns that will be used as the independent variables (i.e., predictors) in the models described in the next section.

Step 4: The products from step 1 and step 3 were combined to create the final table "R", which has as many rows as the links in our network and 301 columns. The first column corresponds to the depended variable and the rest 300 to the independent variables.

Regression Models

For this study the following regression models were considered, tested, and validated. All four models were validated with 5-fold cross validation to avoid overfitting.

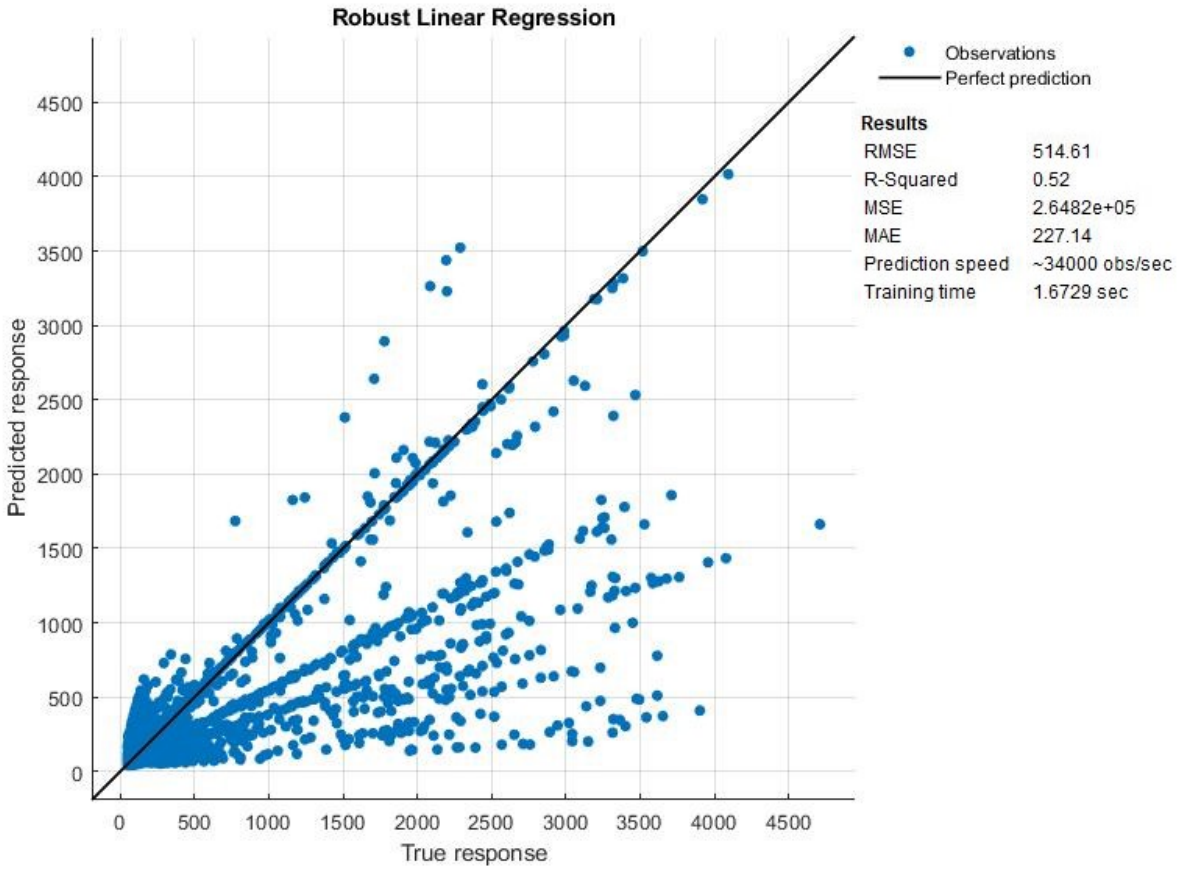
Robust linear regression works by assigning a weight to each data point. Weighting is done automatically and iteratively using a process called iteratively reweighted least squares. In the first iteration, each point is assigned equal weight and model coefficients are estimated using ordinary least squares. At subsequent iterations, weights are recomputed so that points farther from model predictions in the previous iteration are given lower weight. Model coefficients are then recomputed using weighted least squares. The process continues until the values of the coefficient estimates converge within a specified tolerance *(1)*.

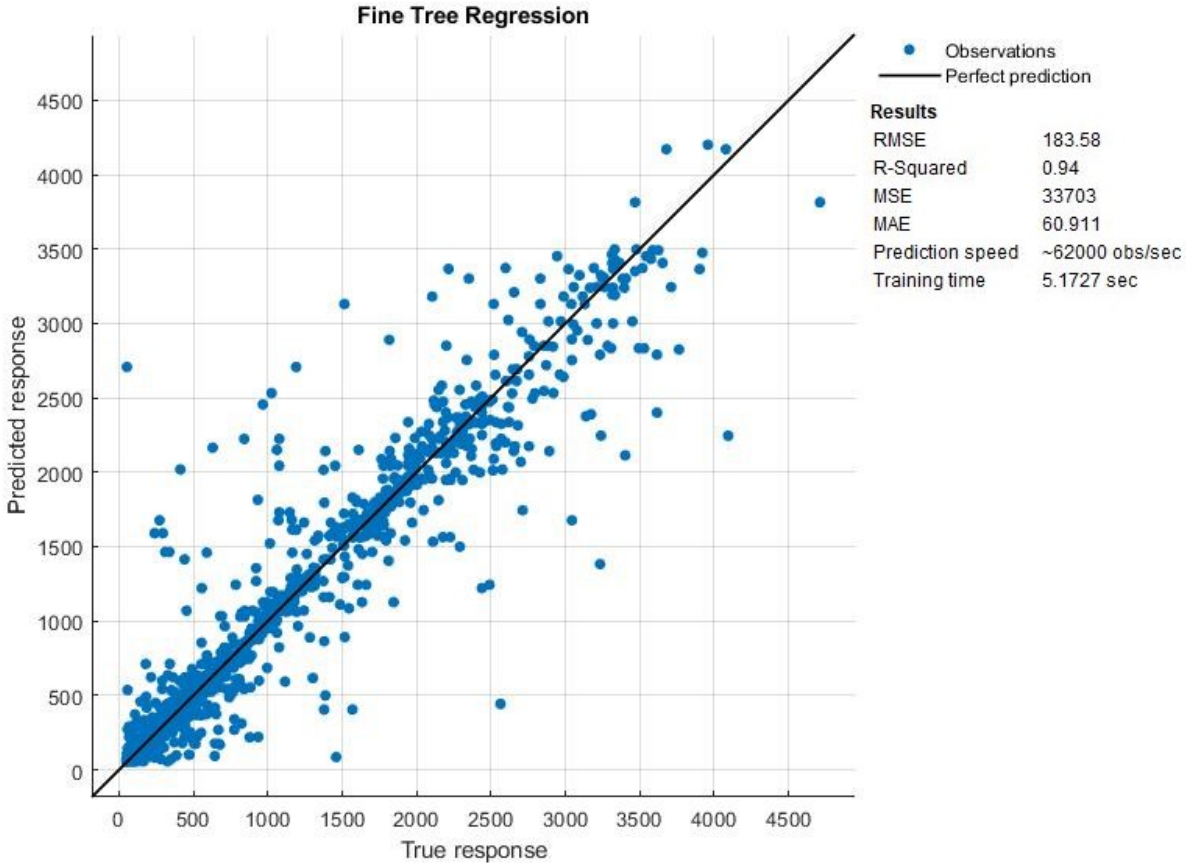
Decision Tree Regression are used for regression or classification models as tree structures. A dataset is broken down into subsets while the associated decision tree is developed in parts. The product is a tree with decision and leaf nodes. Decision nodes have two or more branches, each corresponding to an attribute that was tested. Leaf node represents a decision on the numerical target. The best predictor, which is called root node, is at the top of the tree. Decision trees can be useful for both categorical and numerical data *(2)*. In this study three different decision trees were used. Fine tree, Boosted tree, and Bayesian optimized tree regression.

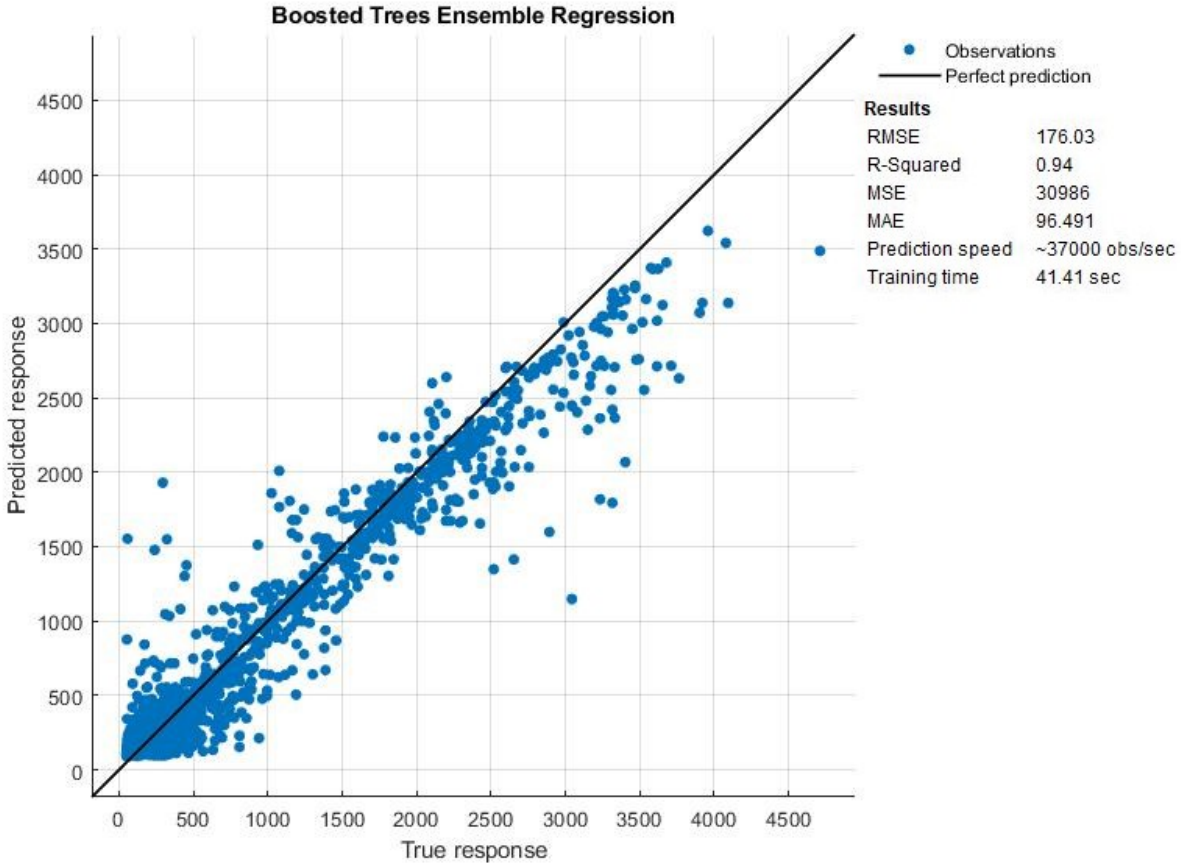
EXPERIMENTS

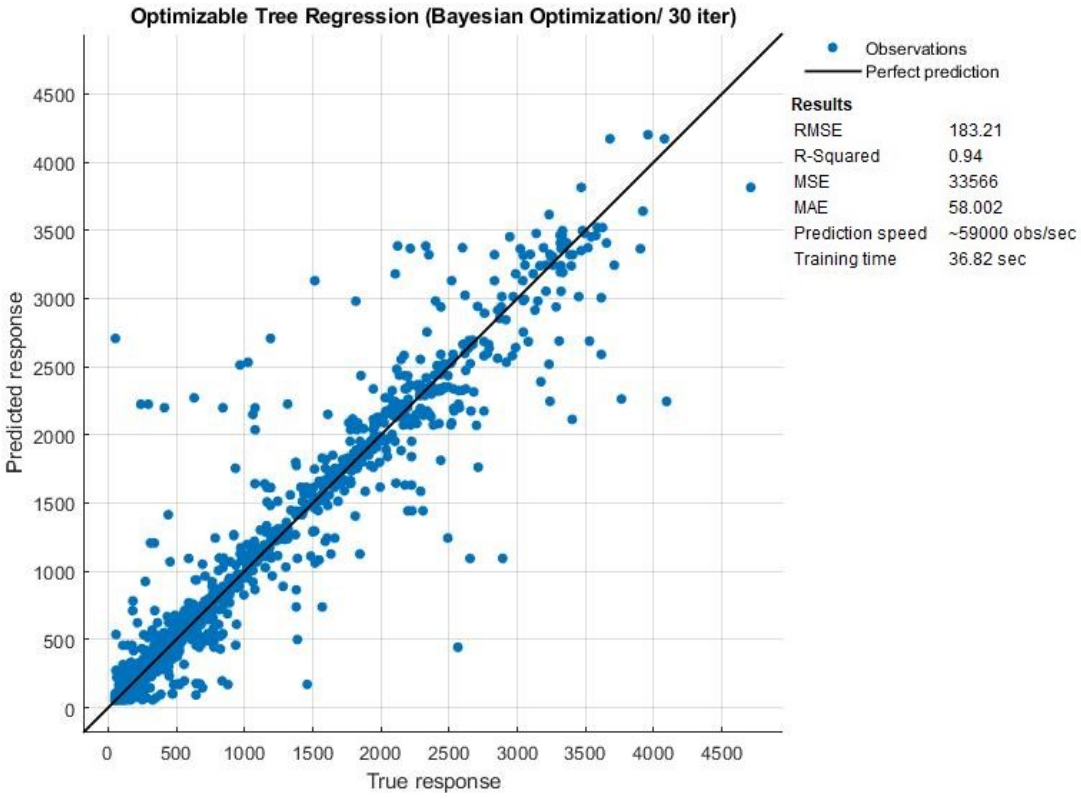
For the numerical experiments, after the datasets were prepared, the Regression Learner App included with MATLAB was used to train, test, and validate the proposed models. The Regression Learner App interactively trains, validates, and tunes regression models to make predictions using supervised machine learning. All models were validated using 5-fold cross validation to avoid overfitting. The next four figures show how each of the proposed models performed. Indices such as RMSE (i.e., Root Mean Square Error), R-squared (indicates the percentage of the variance in the dependent variable that the independent variables explain collectively), MSE (i.e., Mean Squared Error), MAE (i.e., Mean Absolute Error) and others are included in order for the reader to be able to compare the models. For the Bayesian Optimized Tree, the observed and estimated MSEs for every iteration of optimizing are also included. For the

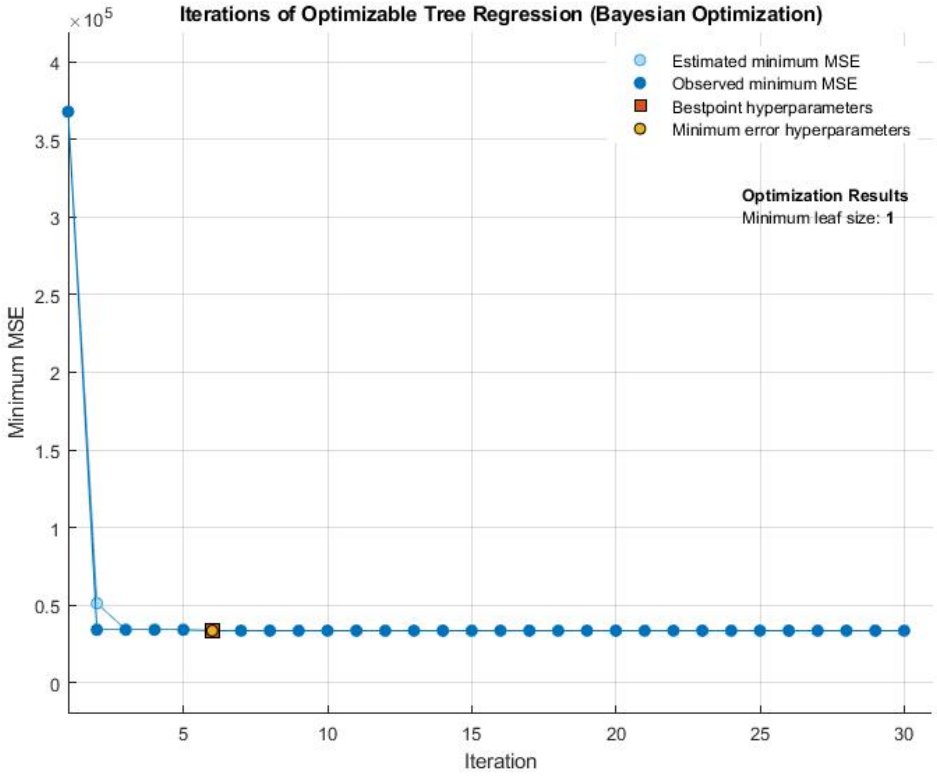
Robust Linear Regression, the PCA (principal component analysis) option was enabled to reduce the independed variables based on their significance.




Fine Tree Regression
Predicted response
True response
0
500
1000
1500
2000
2500
3000
3500
4000
4500
Observations
Perfect prediction
Results
RMSE 183.58
R-Squared 0.94
MSE 33703
MAE 60.911
Prediction speed ~62000 obs/sec
Training time 5.1727 sec


Boosted Trees Ensemble Regression
Observations
Perfect prediction
Results
RMSE 176.03
R-Squared 0.94
MSE 30986
MAE 96.491
Prediction speed ~37000 obs/sec
Training time 41.41 sec
Predicted response
True response
0
500
1000
1500
2000
2500
3000
3500
4000
4500


Optimizable Tree Regression (Bayesian Optimization/ 30 iter)
Observations
Perfect prediction
Results
RMSE 183.21
R-Squared 0.94
MSE 33566
MAE 58.002
Prediction speed ~59000 obs/sec
Training time 36.82 sec
Predicted response
True response
0
500
1000
1500
2000
2500
3000
3500
4000
4500


Iterations of Optimizable Tree Regression (Bayesian Optimization)
×10^5
Estimated minimum MSE
Observed minimum MSE
Bestpoint hyperparameters
Minimum error hyperparameters
Optimization Results
Minimum leaf size: 1
Minimum MSE
Iteration
0
0.5
1
1.5
2
2.5
3
3.5
4
5
10
15
20
25
30

DISCUSSION

In this project a methodology for preparing truck GPS data and freight business data to predict truck flows is being presented. Various regression models, including fine tree, boosted tree, Bayesian optimized tree, and linear robust regression, were tested for their efficiency to predict truck flows using freight business indices

Computational experiments showed that all prediction tree models performed really well with R-squared = 0.94. The robust linear regression predicted truck flows with percentage of the variance in the dependent variable that the independent variables explained collectively at 52 percent.

The most important conclusion of this project is that freight business indices, such as freight tonnage, value, units, and mileage can be used as indicators for truck flows. Of course further research is needed with regards to testing more models with the same and different datasets.

REFERENCES

1. DuMouchel, W. H., and F. L. O'Brien. “Integrating a Robust Option into a Multiple Regression Computing Environment.” *Computer Science and Statistics*: *Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.
2. Shalev-Shwartz, S.; Shai, BD. "18. Decision Trees". Understanding Machine Learning. Cambridge University Press. 2014.