Transportation Planning and Technology, 2013
Vol. 36, No. 2, 201–217, http://dx.doi.org/10.1080/03081060.2013.770944

A Bayesian analysis of the effect of estimating annual average daily traffic for heavy-duty trucks using training and validation data-sets

Ioannis Tsapakis[a]*, William H. Schneider IV[b] and Andrew P. Nichols[c]

[a]Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, UK; [b]Department of Civil Engineering, The University of Akron, Akron OH 44325-3905, USA; [c]Weisberg Division of Engineering and Computer Science, Marshall University, Huntington, WV 25755, USA

(*Received 15 January 2012; accepted 12 October 2012*)

The precise estimation of annual average daily traffic (AADT) is of significant importance worldwide for transportation agencies. This paper uses three modeling frameworks to predict the AADT for heavy-duty trucks. In total, 12 models are developed based on regression and Bayesian analysis using a training data-set. A separate validation data-set is used to compare the results from the 12 models, spanning the years 2005 through 2007 and taken from 67 continuous data recorders. Parameters of significance include roadway functional class, population density, and spatial location; five regional areas – northeast, northwest, central, southeast, and southwest – of the state of Ohio in the USA; and average daily truck traffic. The results show that a full Bayesian negative binomial model with a coefficient offset is the most efficient model framework for all four seasons of the year. This model is able to account for between 87% and 92% of the variability within the data-set.

Keywords: Bayesian analysis; AADT; average daily truck traffic; regression analysis; traffic monitoring program

Introduction

The accurate estimation of annual average daily traffic (AADT) plays a vital role for day-to-day operations within a government department of transportation. As a result, there have been many research studies over recent decades focused on developing more efficient methods to estimate AADT. Additionally, most recent emphasis is being placed on predicting specific vehicle classes (especially heavy-duty vehicles – known in the USA as classes 4–13) instead of total AADT. One common or traditional approach is to develop individual monthly adjustment factors from continuous count locations, then group these factors together, and finally assign short-term counts to each group. The main concern associated with this traditional approach includes the development of errors and uncertainties throughout each step of the process.

As a result of this shortcoming, some research has focused on developing new methods based on local conditions in order to estimate AADT directly. Some of the

*Corresponding author. Email: i.tsapakis@ucl.ac.uk more common approaches include artificial neural networks and ordinary least squares regression (Faghri and Hua 1995; Lam and Xu 2000; Lingras et al. 2000; Sharma et al. 2000, 2001). The data requirements from these methods vary from simple regression models based on roadway functional classification, to increased data needs such as socioeconomic and land use parameters. For example, Neveu (1983) developed the regression models to predict AADT for roads of each functional class; Mohamad et al. (1998) developed a multiple regression model using the roadway type, the accessibility of the road, the county population, and the total arterial mileage of a county; and Fricker and Sinha (1987) used population, vehicle registration, and employment as predictors in their models. Xia et al. (1999) developed multiple regression models for Florida based on roadway characteristics, such as number of lanes and functional classification for non-state roads. Zhao and Chung (2001) used roadway data, socioeconomic characteristics, expressway accessibility, and accessibility to regional employment centers to develop four multiple regression models for expressway roads in a Florida county and Zhao and Park (2004) used weighted regression models to estimate AADT, conducting multiple linear regression analyses separately for selected rural and urban areas to identify explanatory variables for interpreting seasonal traffic patterns.

In addition to more traditional regression models, researchers have added innovative statistical modeling to improve the overall performance of their predictions. Lingras, Sharma, and Zhong (2002) used genetically designed regression models for individual hours, while Tang, Lam, and Ng (2003) built a nonparametric regression model to forecast short-term traffic volumes for one year. Zhong, Sharma, and Lingras (2004) developed a locally weighted regression model, a form of memory-based algorithm for learning continuous mapping from real-valued input vectors to real-valued output vectors. In most cases, the more advanced models improve the findings by a few percentage points.

Two areas for potential development include the creation of models that directly predict AADT for heavy-duty trucks (specifically vehicle classes 4–13), and the altering of the model framework to include negative binomial models into a full Bayesian framework. Since AADT is a nonnegative count variable, it is reasonable to use a negative binomial model instead of a more traditional ordinary least squares regression model. One negative of the regression model is the generation of negative AADT values. The potential benefits of the Bayesian framework include the implementation of prior knowledge into the prediction model, as well as developing a posterior distribution of the beta coefficients.

Of the three objectives of this study on which this paper is based, the first is to develop eight data-sets – a training set and validation set for each of the four seasons in a year. The second objective is to develop three individual modeling frameworks using the four seasonal training data-sets. Model One is an ordinary least squares regression model, while Models Two and Three are full Bayesian negative binomial models – the difference being that Model Two includes a coefficient offset and Model Three does not. The third objective is to compare the three model frameworks using the validation data-sets across all seasonal durations. The end result of this research study shows the effectiveness of the three model frameworks for directly predicting seasonal AADT for heavy-duty vehicles.

Study data

There are 67 continuous count stations across the state of Ohio from the years 2002 through 2007 that collected volume counts for heavy-duty vehicles in classes 4–13 (Federal Highway Administration 2001). Based on these continuous count stations, the final development of the training and validation data-sets used in this study is based on two criteria. The first criterion requires a minimum number of collected hourly heavy-duty vehicle volumes per day per continuous count station. The second requires a minimum number of complete, no missing hourly volumes, for days of the year per continuous count station. Additional land use, socioeconomic, and population density data are provided at the county level using US census data (US Census Bureau 2008).

Site-specific requirements

The first criterion in the empirical setting is site-specific. In this case, each site used in the development of the final data-sets requires a continuous 24-hour data collection period to calculate average daily traffic (ADT). Depending on the time of day, and location of the continuous count, the 24-hour ADT may include hourly volumes with zero recorded heavy-duty vehicle counts. The selection of 280 complete days provides an adequate amount of data when using the American Association of State Highway and Transportation Officials (AASHTO) recommended formula for estimating AADT (AASHTO 1992; Spiegelhalter et al. 2004). Using a lower number of complete days hinders the overall performance of the AASHTO recommended formula.

Temporal aggregation of the data-sets

The temporal aggregation of the data is based both annually as well as seasonally. The initial data-set is comprised of 48,893 daily traffic volumes for the years 2002 through 2007 in the state of Ohio. The data from 2002 to 2004 are only used as prior knowledge for the full Bayesian models, while 75% of the data collected from 2005 to 2007 is used for both training and modeling, and the remaining 25% is used to populate the validation data-set for the three model frameworks. No data provided within this study are used for both the training and validation data-sets. The second temporal aggregation is based on the seasonality of the data. The winter months are December through February, spring months are March through May, summer is June through August, and fall is September through November. Table 1 presents a summary of the parameters populated in the final data-sets for 2005 through 2007. It represents urban and rural counties, spatial distribution, and multiple roadway function classes creating a representative statewide empirical setting.

Statistical methodology

Three model methodologies are developed in this study. Model One is an ordinary least squares regression model and Models Two and Three are full Bayesian negative binomial models. The overall model structure for each model includes ADT volumes for heavy-duty vehicles; roadway functional classes, which include interstate, freeway,

Table 1. Summary statistics for the final data-set.

Entire Data-set[a]			
Variable	Average	Minimum	Maximum
Functional class[b]	N/A	1	12
Lanes	4.3	2	8
ADT classes 4–13	3608.2	0	24,234
AADT classes 4–13	3609.7	15	14,050
Population density[c] (population/mi^2)	993.5	33.1	3035.7
Percent interstate[d]	43.9 (21,464)[e]		
Percent freeway[d]	17.5 (8556)[e]		
Percent principal arterial[d]	32.6 (15,939)[e]		
Percent minor arterials[d]	6.0 (2934)[e]		

Notes: [a]The entire data-set comprises 48,893 ADT observations.
[b]The roadway functional classification is based on the guidance provided by the HPMS.
[c]Population density is defined as the total number of people per county divided by the number of square miles per county.
[d]Percent refers to the total number of observations for each roadway functional class for the entire data-set.
[e]Values within the parentheses are the number of observations per variable.

and principal arterial highways (based on the US Highway Performance Monitoring System [HPMS] classification); spatial locations, which include northeast, northwest, central, southeast, and southwest locations within Ohio; and temporal variables such as Mondays, midweek, and Fridays.

Model one: ordinary least squares regression

Several studies have used regression models effectively to predict AADT (Xia et al. 1999; Zhao and Chung 2001; Lingras, Sharma, and Zhong 2002). There is, however, one potential limitation associated when using an ordinary least squares method when predicting AADT. AADT is considered to be count data, and should not be modeled as a continuous variable. As a result of modeling AADT as a continuous variable, there is some potential for predicting negative values, which is impossible for a positive parameter such as AADT.

Models two and three: negative binomial

As a result of the limitation with the basic regression model, there are two common ways to model count data. These two methods are the Poisson and negative binomial models. One criterion for the correct use of the Poisson model is that the mean and variance of the prediction should be equivalent. If this criterion is not satisfied, the negative binomial model should be used. As a result of the nonequivalent mean and variance, the negative binomial model is selected over the Poisson model. In this study, there are two negative binomial model frameworks, as shown in Equations (1) and (2):

$$\lambda_i = \beta_1 x_1 e^{(\beta_0+\beta_2 x_2+\ldots+\beta_n x_n)} \tag{1}$$

$$\lambda_i = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)} \tag{2}$$

where, $\beta_0 =$ constant term; $\beta_1, \ldots, \beta_n =$ estimated parameters in vector form; and $x_1, \ldots, x_n =$ explanatory variables developed for the individual models.

Equation (1) includes a coefficient offset while Equation (2) does not include a coefficient offset. The use of an offset is common when there is a wide range of values for an individual parameter such as ADT. In this study, the models require all variables to have a p-value less than 0.05 which corresponds to a 95% confidence level. Once the key parameters are identified, the next phase is to develop our eight predictive models, one model for each season using both Equations (1) and (2) implementing a full Bayesian methodology.

Full Bayesian methodology

Once the initial models are developed, a Bayesian methodology with Gibbs sampling is adopted in order to obtain the predictive posterior simulation of AADT. In a fully developed Bayesian framework for modeling and inference of estimated parameters, the Bayesian model specification requires a likelihood function, and prior distribution to obtain the posterior density of the estimated parameters from the given data. The Bayesian methodology is developed from Equation (3):

$$p(\Theta \Big| \chi) = \frac{p(\Theta)p(\chi|\Theta)}{\int p(\Theta)p(\chi|\Theta)d\Theta} \propto p(\Theta)p(\chi|\Theta) \tag{3}$$

where the prior distribution $p(\theta)$ expresses the uncertainty before and the posterior distribution $p(\theta|\mathrm{x})$ describes the uncertainty after seeing the data. In order to obtain the predictive posterior simulation associated with AADT, three Monte Carlo Markov Chains are developed with Gibbs sampling. In the Monte Carlo simulation, informative priors with some precision are assigned to the parameter coefficients. These informative priors are based on the model results using data provided from 2002 through 2004. In addition to the prior knowledge, a three chain approach is used in the simulation. In each of the chains, different initial values are selected for each chain and, after sufficient simulation iterations, the three chains are evaluated for convergence. In this study, the convergence of the three chains is based on Gelman–Rubin statistics, Kernel density, autocorrelation, trace plots, and times series plots (Gelman et al. 2003; Robichaud & Gordon 2003). Once the model converges, the posterior distributions are summarized (shown later in Tables 3–6).

Results

The AADT predictions developed in this study are based on the initial results of the three models for each season of the year. The initial results are based solely on the model training data-set, which randomly samples 75% of the 2005–2007 data. The second set of results is based on the estimated model performance. The model performance is evaluated with the 2005–2007 validation data-set. As described previously, it is important to note that no data from the validation data-set is used in the initial model development. The model performance is based on the model prediction values versus the actual validation data-set AADTs.

Initial results: ordinary least squares regression models

The results from the regression models for the four seasons are shown in Table 2. The variables with a statistical significance at the 95% confidence level include heavy-duty truck ADTs as well as interstate, freeway, and principal arterial roadway functional classifications. Other variables of significance include population density as well as spatial location. The final sets of variables are temporal based on the day of week: Monday, midweek, and Friday. Additionally, other variables including number of lanes, socioeconomic and additional land use categories are also tried. Unfortunately, these parameters are not considered statistically significant, or, in the case of the number of lanes, create multicollinearity problems with other variables.

The results show the AADT predictions are higher with an increase in heavy-duty ADT. The roadway classification for interstates produces higher predicted AADTs than do freeway and principal arterials. Other findings of interest show the northwest and central areas predict higher AADTs than southern Ohio. This result is expected because northwest and central Ohio includes more urban areas such as Toledo, in the northwest, and the state capital Columbus, in the central area, when compared with the southeast. The final overall results are developed for the day of the week. In general, the ADTs are lower for Monday followed by Friday with the highest during the midweek, Tuesday through Thursday. As a result of the higher prediction values associated with the ADT and the other variables remaining constant, the net effect of the day of the week is similar to a daily adjustment factor which results in a larger subtraction of values with the midweek, followed by Friday, and lastly Monday. The

Table 2. Regression model coefficients.

	Model coefficients			
Variable name	Spring model	Summer model	Fall model	Winter model
Constant	700.036	563.920	977.529	982.359
Truck ADT classes 4–13	0.638	0.634	0.597	0.573
Interstate	3199.530	2941.930	2665.136	3756.312
Freeway	1105.856	1121.471	998.724	1295.978
Principal arterial	335.872	254.929	258.363	392.389
Population density (population/mi^2)	−1.63E-01	−1.74E-01	−1.97E-01	−3.71E-01
Northwest Ohio	845.400	933.432	462.919	466.904
Central Ohio	842.494	702.819	476.775	375.207
Southwest Ohio	226.096	565.635	N/A	N/A
Southeast Ohio	464.130	565.500	N/A	N/A
Monday	−1566.896	−1495.514	−1313.483	−1136.265
Midweek (Tuesday through Thursday)	−2044.769	−1889.574	−1664.600	−1593.948
Friday	−1604.455	−1619.532	−1306.301	−1308.538
Number of observations	3546	3363	4315	3119
χ^2	7915.67	7799.64	9039.98	6249.31
Adjusted R^2	0.89	0.90	0.88	0.86

Notes: All variables are statistically significant at the 0.05 level.
N/A = the variable is not statistically significant and was not included in the final model.

Table 3. Full Bayesian framework with coefficient offsets for spring and summer.

Variable[a,b]	Mean	Standard deviation	MC Error[c]	2.5%[d]	Median[d]	97.5%[d]
Spring full Bayesian model with coefficient offsets						
Constant	1.937	0.035	0.002	1.870	1.936	2.010
Truck ADT classes 4–13	0.754	0.006	0.000	0.741	0.754	0.765
Interstate	1.005	0.038	0.002	0.931	1.005	1.079
Freeway	0.687	0.033	0.002	0.623	0.687	0.753
Principal arterial	0.414	0.028	0.001	0.359	0.414	0.467
Population density (population/mi^2)	N/A	N/A	N/A	N/A	N/A	N/A
Northwest Ohio	0.068	0.015	0.000	0.040	0.068	0.098
Central Ohio						
Southwest Ohio	−0.101	0.022	0.001	−0.143	−0.102	−0.058
Southeast Ohio	−0.133	0.030	0.001	−0.192	−0.133	−0.073
Monday	−0.805	0.017	0.000	−0.839	−0.805	−0.771
Midweek (Tuesday through Thursday)	−0.967	0.015	0.001	−0.996	−0.967	−0.938
Friday	−0.880	0.018	0.001	−0.916	−0.880	−0.845
Inverse dispersion	11.810	0.287	0.002	11.250	11.810	12.380
Summer full Bayesian model with coefficient offsets						
Constant	2.161	0.042	0.002	2.073	2.162	2.240
Truck ADT classes 4–13	0.570	0.006	0.000	0.558	0.571	0.582
Interstate	2.035	0.047	0.002	1.947	2.035	2.128
Freeway	1.583	0.041	0.002	1.505	1.583	1.663
Principal arterial	0.914	0.034	0.002	0.849	0.914	0.982
Population density (population/mi^2)	1.02E-05	8.73E-06	2.94E-07	−6.67E-06	1.02E-05	2.74E-05
Northwest Ohio	0.174	0.020	0.001	0.135	0.174	0.214
Central Ohio						
Southwest Ohio	−0.053	0.027	0.001	−0.106	−0.053	−0.002
Southeast Ohio						
Monday	−0.541	0.022	0.001	−0.584	−0.541	−0.497
Midweek (Tuesday through Thursday)	−0.603	0.017	0.001	−0.637	−0.602	−0.570
Friday	−0.568	0.023	0.001	−0.613	−0.567	−0.523
Inverse dispersion	6.786	0.168	0.001	6.464	6.784	7.122

[a] All variables are statistically significant at the 95% confidence level.

[b] N/A suggests that the variables are not statistically significant in the final model.

[c] MC error shows the Markov chain error. This measure helps identify model convergence.

[d] 2.5% and 97.5% show the middle 95% of the data while being the midpoint of the parameter distribution.

Table 4. Full Bayesian framework with coefficient offsets for fall and winter.

Variable[a,b]	Mean	Standard deviation	MC Error[c]	2.5%[d]	Median[d]	97.5%[d]
Fall full Bayesian model with coefficient offsets						
Constant	3.311	0.039	0.002	3.234	3.312	3.384
Truck ADT classes 4–13	0.389	0.005	0.000	0.380	0.389	0.399
Interstate	2.512	0.038	0.002	2.437	2.513	2.585
Freeway	1.763	0.039	0.002	1.687	1.764	1.840
Principal arterial	1.000	0.035	0.002	0.929	1.000	1.069
Population density (population/mi^2)	−8.38E-05	9.13E-06	2.17E-07	−1.02E-04	−8.36E-05	−6.63E-05
Northwest Ohio	0.211	0.022	0.001	0.167	0.211	0.254
Central Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Southwest Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Southeast Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Monday	−0.436	0.025	0.001	−0.486	−0.436	−0.386
Midweek (Tuesday through Thursday)	−0.451	0.020	0.001	−0.489	−0.451	−0.412
Friday	−0.378	0.025	0.001	−0.427	−0.378	−0.329
Inverse dispersion	4.036	0.085	0.001	3.871	4.035	4.205
Winter full Bayesian model with coefficient offsets						
Constant	3.141	0.041	0.002	3.062	3.142	3.218
Truck ADT classes 4–13	0.593	0.007	0.000	0.580	0.593	0.606
Interstate	1.294	0.047	0.002	1.200	1.294	1.382
Freeway	0.839	0.042	0.002	0.754	0.840	0.917
Principal arterial	0.403	0.035	0.002	0.332	0.404	0.473
Population density (population/mi^2)	−6.97E-05	8.86E-06	2.40E-07	−8.72E-05	−6.94E-05	−5.28E-05
Northwest Ohio	0.069	0.023	0.001	0.022	0.069	0.115
Central Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Southwest Ohio	−0.320	0.029	0.001	−0.379	−0.319	−0.263
Southeast Ohio	−0.287	0.046	0.001	−0.377	−0.286	−0.198
Monday	−0.525	0.024	0.001	−0.572	−0.525	−0.479
Midweek (Tuesday through Thursday)	−0.874	0.020	0.001	−0.914	−0.873	−0.835
Friday	−0.806	0.025	0.001	−0.854	−0.806	−0.757
Inverse dispersion	6.005	0.152	0.001	5.713	6.002	6.307

[a] All variables are statistically significant at the 95% confidence level.
[b] N/A suggests that the variables are not statistically significant in the final model.
[c] MC error shows the Markov chain error. This measure helps identify model convergence.
[d] 2.5% and 97.5% show the middle 95% of the data while being the midpoint of the parameter distribution.

Table 5. Full Bayesian framework with no coefficient offsets for spring and summer.

Variable[a,b]	Mean	Standard deviation	MC Error[c]	2.5%[d]	Median[d]	97.5%[d]
Spring full Bayesian model with no coefficient offsets						
Constant	4.927	0.056	0.003	4.821	4.923	5.040
Truck ADT classes 4–13	1.44E-04	4.24E-06	1.57E-07	1.36E-04	1.44E-04	1.53E-04
Interstate	3.250	0.063	0.003	3.121	3.253	3.374
Freeway	2.647	0.056	0.003	2.535	2.648	2.754
Principal arterial	1.913	0.048	0.002	1.816	1.914	2.005
Population density (population/mi^2)	−4.84E-05	1.18E-05	3.47E-07	−7.12E-05	−4.86E-05	−2.51E-05
Northwest Ohio	0.196	0.030	0.001	0.136	0.196	0.254
Central Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Southwest Ohio	−0.398	0.039	0.001	−0.473	−0.398	−0.321
Southeast Ohio	−0.643	0.057	0.001	−0.753	−0.643	−0.531
Monday	−0.384	0.031	0.001	−0.446	−0.385	−0.322
Midweek (Tuesday through Thursday)	−0.498	0.026	0.001	−0.549	−0.498	−0.446
Friday	−0.423	0.031	0.001	−0.484	−0.423	−0.362
Inverse dispersion	3.454	0.079	0.000	3.300	3.453	3.610
Summer full Bayesian model with no coefficient offsets						
Constant	4.721	0.049	0.002	4.633	4.729	4.829
Truck ADT classes 4–13	1.35E-04	4.07E-06	1.37E-07	1.27E-04	1.35E-04	1.43E-04
Interstate	3.456	0.057	0.003	3.341	3.450	3.563
Freeway	2.794	0.050	0.002	2.695	2.794	2.893
Principal arterial	2.002	0.042	0.002	1.919	2.003	2.082
Population density (population/mi^2)	−8.98E-05	1.30E-05	3.67E-07	−1.15E-04	−8.96E-05	−6.46E-05
Northwest Ohio	0.186	0.034	0.001	0.120	0.186	0.250
Central Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Southwest Ohio	−0.358	0.039	0.001	−0.435	−0.357	−0.280
Southeast Ohio	−0.585	0.062	0.002	−0.706	−0.585	−0.461
Monday	−0.348	0.033	0.001	−0.412	−0.348	−0.283
Midweek (Tuesday through Thursday)	−0.427	0.027	0.001	−0.478	−0.428	−0.375
Friday	−0.372	0.035	0.001	−0.438	−0.372	−0.304
Inverse dispersion	3.061	0.071	0.000	2.922	3.060	3.203

[a] All variables are statistically significant at the 95% confidence level.

[b] N/A suggests that the variables are not statistically significant in the final model.

[c] MC error shows the Markov chain error. This measure helps identify model convergence.

[d] 2.5% and 97.5% show the middle 95% of the data while being the midpoint of the parameter distribution.

Table 6. Full Bayesian framework with no coefficient offsets for fall and winter.

Variable[a,b]	Mean	Standard deviation	MC Error[c]	2.5%[d]	Median[d]	97.5%[d]
Fall full Bayesian model with no coefficient offsets						
Constant	5.439	0.040	0.002	5.361	5.440	5.518
Truck ADT classes 4–13	1.36E-04	3.70E-06	1.20E-07	1.29E-04	1.36E-04	1.43E-04
Interstate	2.744	0.046	0.002	2.657	2.743	2.836
Freeway	2.235	0.045	0.002	2.150	2.235	2.322
Principal arterial	1.494	0.040	0.002	1.416	1.494	1.572
Population density (population/mi^2)	−8.07E-05	1.09E-05	2.53E-07	−1.02E-04	−8.07E-05	−5.99E-05
Northwest Ohio	N/A	N/A	N/A	N/A	N/A	N/A
Central Ohio	0.091	0.031	0.001	0.031	0.091	0.150
Southwest Ohio	−0.431	0.031	0.000	−0.493	−0.431	−0.370
Southeast Ohio	−0.439	0.048	0.001	−0.533	−0.439	−0.346
Monday	−0.339	0.029	0.001	−0.396	−0.339	−0.283
Midweek (Tuesday through Thursday)	−0.422	0.023	0.001	−0.466	−0.422	−0.377
Friday	−0.341	0.029	0.001	−0.399	−0.341	−0.284
Inverse dispersion	3.055	0.063	0.000	2.932	3.055	3.179
Winter full Bayesian model with no coefficient offsets						
Constant	5.396	0.064	0.003	5.270	5.399	5.517
Truck ADT classes 4–13	1.36E-04	5.09E-06	1.68E-07	1.26E-04	1.36E-04	1.46E-04
Interstate	2.965	0.070	0.004	2.830	2.964	3.106
Freeway	2.292	0.064	0.003	2.166	2.291	2.417
Principal arterial	1.510	0.059	0.003	1.395	1.512	1.623
Population density (population/mi^2)	−1.33E-04	1.36E-05	3.79E-07	−1.59E-04	−1.33E-04	−1.06E-04
Northwest Ohio	0.136	0.037	0.001	0.062	0.136	0.207
Central Ohio	0.179	0.042	0.001	0.098	0.178	0.261
Southwest Ohio	−0.558	0.043	0.001	−0.644	−0.557	−0.473
Southeast Ohio	−0.579	0.069	0.002	−0.713	−0.581	−0.442
Monday	−0.271	0.037	0.001	−0.343	−0.271	−0.199
Midweek (Tuesday through Thursday)	−0.412	0.030	0.001	−0.470	−0.412	−0.353
Friday	−0.332	0.037	0.001	−0.405	−0.333	−0.260
Inverse dispersion	2.612	0.063	0.000	2.490	2.612	2.738

[a] All variables are statistically significant at the 95% confidence level.
[b] N/A suggests that the variables are not statistically significant in the final model.
[c] MC error shows the Markov chain error. This measure helps identify model convergence.
[d] 2.5% and 97.5% show the middle 95% of the data while being the midpoint of the parameter distribution.

temporal factor would in turn create a relatively similar AADT estimate for one particular section of roadway for all weekday samples.

Initial results: full Bayesian negative binomial model

The second set of results is developed for Models Two and Three using a full Bayesian model framework. The results produced by Model Two are shown in Tables 3 and 4, and those of Model Three are shown in Tables 5 and 6.

The results for Model Two with the offset indicate that, as the heavy-duty ADT increases, the predicted AADT also increases. In terms of the HPMS roadway classification, the interstate has the greatest influence on AADT, followed by freeway and principal arterials. Field data estimates developed in the northwest produce higher AADT estimates than in the southwest and southeast. The central geographic location is no longer a significant variable. Similar to the regression findings, the highest ADT volumes on average occur during the midweek followed by Friday and Monday. In order to predict relatively similar AADT for an individual segment, independent of the day of the week, the temporal results show the requirement to lower AADT predictions for the midweek, followed by Friday and Monday. The highest inverse dispersion results in the lowest overall dispersion, a measurement of model performance (the variance divided by the mean), indicates that spring is the most efficient model followed by summer, winter, and finally the fall.

The results for Model Three with no offset generally have similar trends in terms of sign and magnitude when directly compared with Model Two. As the ADT values increase, so do the predicted AADT predictions. The interstate roadway class still has the greatest influence on AADT followed by freeway and principal arterials. ADT estimates for the northwest increase the AADT prediction while southwest and southeast lower AADT estimates. The central geographic location is not significant in the spring and summer while the northwest is not significant in the fall.

The results remain consistent with the other models that have the highest ADT volumes occurring during the midweek, followed by Friday and Monday. This in turn requires temporal adjustment as seen with the midweek followed by Friday and then Monday to remain as relatively similar AADT final predictions for individual segments. The lowest overall dispersion remains the spring followed by summer, fall, and winter. In comparison to Model Two, the inverse dispersion values are smaller; therefore, these models are not as efficient as Model Two.

Comparison of results

A comparison of the performance of each of the three model frameworks is shown in Figures 1–4. In each case, the models developed with the training data-sets are compared with the AADTs provided by the validation data-set. As described previously, the training and validation data-sets are both developed randomly for each season, and no data are used in both data-sets. In each of the four figures, the horizontal axis is the AADT validation data-set for heavy-duty trucks (classes 4–13) and the vertical axis is the predicted AADTs from each of the three model frameworks for each section.

The results for the three models developed for spring 2005–2007 are shown in Figure 1. The results show Model One underpredicts AADT by 4%, while Model

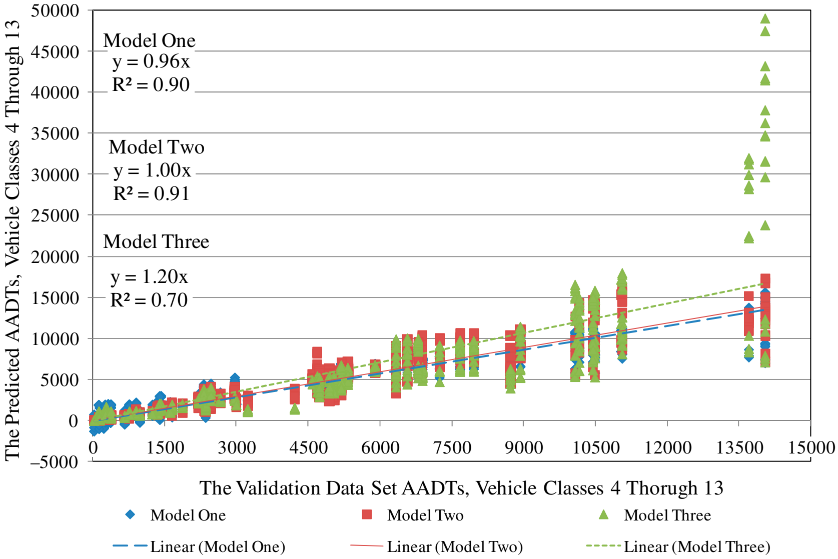

Figure 1. Comparison of the spring model results.

Three overpredicts by 20%. The overall explanation of the variability within the validation data-set by the prediction models ranges from 70 to 91%. When comparing the negative binomial models, Model Two with the ADT offset performs better than Model Three. Model Three does not have an ADT offset, and the predicted AADTs are significantly overestimated when the AADTs are greater than 10,000 veh/day. In terms of model selection, the impact of the ADT offset is diminished as the predicted AADTs are reduced.

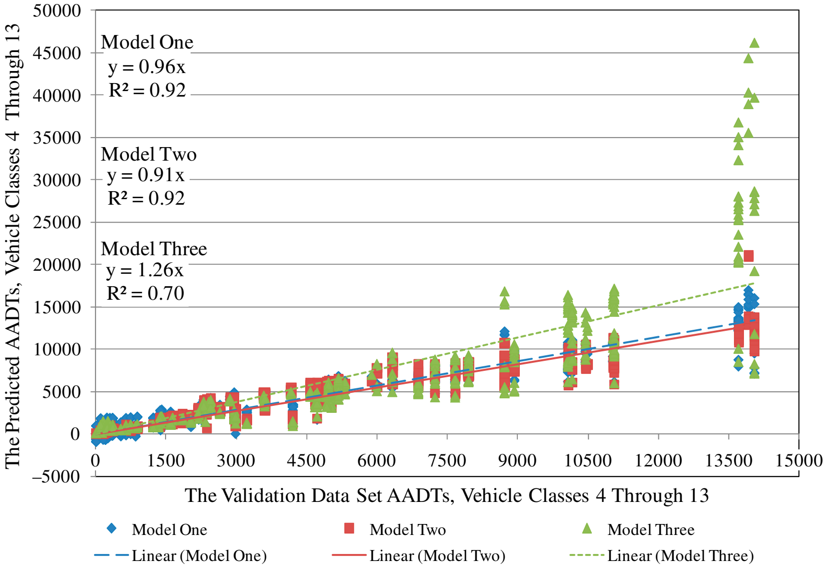

Figure 2. Comparison of the summer model results.

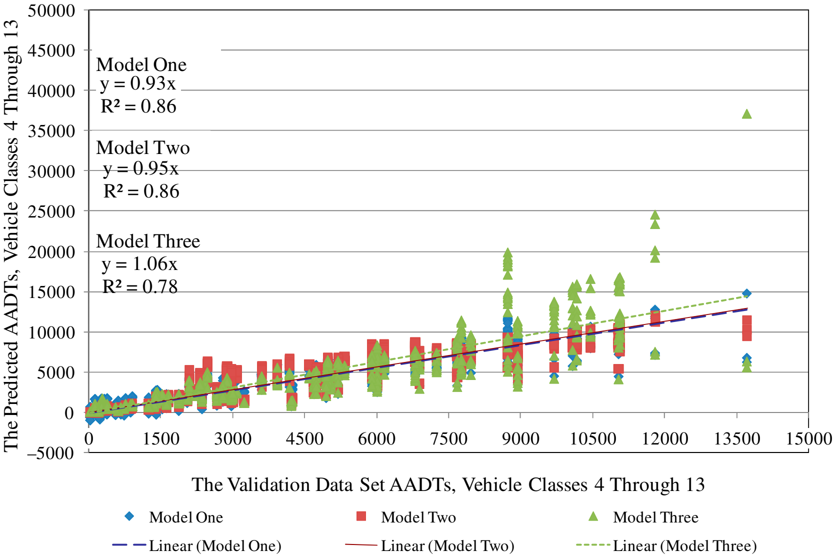

Figure 3. Comparison of the fall model results.

The results for the summer season are shown in Figure 2. Models One and Two slightly underpredict by 4% and 9%, respectively, while Model Three, on average, overpredicts by 26%. This overprediction is influenced by the ADT values. The models' ability to describe the variability within the validation data-set ranges from 70% for Model Three to 92% for Models One and Two.

The results from the fall season, Figure 3, show that Models One and Two both underpredict the AADTs by 7 and 5%, respectively. In both cases, the R^2 values are

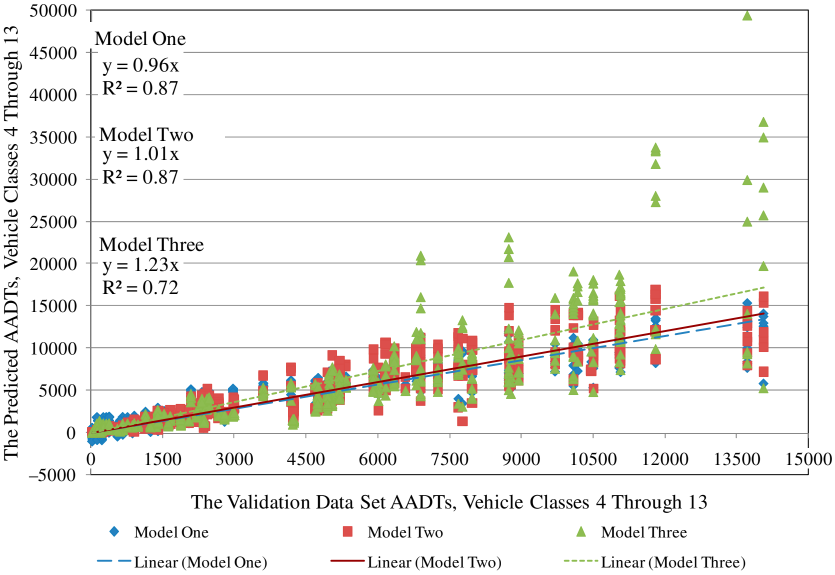

Figure 4. Comparison of the winter model results.

0.93 and 0.95. The results for Model Three explain 70% of the variability and on average are closer in prediction, overpredicting by 6% with the validation data-set AADTs. The overall results show less variability between the three models as well as the three other seasons.

The final seasonal comparison of the data is shown in Figure 4 for the winter season. The overall findings remain consistent with the spring and summer seasons. Models One and Two predict similar values with Model One underestimating AADT by 4% and Model Two overpredicting by 1%. In both cases, the average prediction values are complementary to the validation data-set. Model Three still overpredicts on average by 23%, and the ability to explain 72% of the variability within the validation data-set.

Summary of model performance

A summary of the three model performance evaluations is provided in Table 7. There are two sets of results from this table. The first set of results shows the percentage of predicted heavy-duty AADTs that are within 10% of the validation data-set AADTs per season. The overall result shows that Model Two has the highest percentage of values within 10% of the actual value for all seasons. The results when comparing Models One and Three are mixed. Model Three performs slightly better for the spring and summer months while Model One is more efficient in prediction for the fall and winter months. Generally, there are fewer variables required in Model Three than Model One. All three models have the overall highest accuracy rating for the spring season, while the lowest predicting accuracy varies per model across the other three seasons. Other results stem from the inclusion of the offset as shown in Model Two. Model Two with the offset has less variation, with an overall improvement of 5–10% over the non-offset model, Model Three. The inverse dispersion term is higher, indicating a better performance for all negative binomial models with coefficient offsets (Model Two) than Model Three with no coefficient offsets. The higher inverse dispersion parameters are for spring followed by summer in both Models Two and Three while the fall and winter provide lower inverse dispersion values.

Additional findings illustrate the influence of the model framework between the ordinary least squares regression model and the two negative binomial models. In the case of Model One, for each season the individual models predict negative heavy-duty

Table 7. Summary of model performance.

	Spring	Summer	Fall	Winter
Observations within 10% of the validation data-set AADT				
Model One	20.1%	19.7%	18.6%	19.1%
Model Two	33.9%	20.7%	21.1%	22.7%
Model Three	22.4%	21.7%	16.7%	15.0%
Number of predicted AADTs less than Zero				
Model One	9.8%	12.9%	8.3%	8.7%
Model Two	0	0	0	0
Model Three	0	0	0	0

Note: The predicted values may be ±10%.

truck AADTs for approximately 8–13% of the total number of observations, while the negative binomial models on the other hand predict no AADTs less than zero. The negative predictions show the potential limitations when using ordinary least squares regression.

Conclusions

The overall scope of the study on which this paper is based has been to develop seasonal regression and negative binomial models to predict heavy-duty truck AADT directly. One of the strengths of this approach is the reduction in prediction errors and, unlike the traditional method, does not require grouping automated traffic recorders or the assignment of short-term counts to groups (Robichaud and Gordon 2003).

The objectives of this study included the development of seasonal training and validation data-sets, the initial assessment of model coefficients across the seasons, and the comparison of individual model performance. The ultimate goal was the development of an accurate modeling approach for predicting heavy-duty AADT for a segment of road. In order to validate the models, the initial data-sets were separated seasonally with 75% of the observations into a training data-set and the remaining 25% of the observations into a validation data-set. The random separation of the training and validation data-sets for each season of the year allowed for a non-biased assessment of the prediction capabilities developed per season for each of the models.

The flow of commodities across the state of Ohio may vary for each season of the year and, therefore, it may not be accurate to constrain model coefficients across the various seasons. Initial models using all the training data were developed with seasonal indicator variables within the model. Based on the level of significance, it is reasonable to split the model seasonally. For example, the model coefficient results shown in Table 2, including the influence of the midweek, Tuesday through Thursday, changed seasonally from −2044, −1889, −1664, −1593 (spring, summer, fall, and winter) throughout the year. A second example is shown in Tables 3 and 4, illustrating the potential negative for constraining the model coefficients. These coefficients varied from 0.754, 0.570, 0.389, to 0.593 for the spring, summer, fall, and winter seasons. Generally, all three model frameworks showed that an increase in ADT increased the predicted AADTs. The roadway classifications demonstrated that the interstate indicator variable had the greatest influence on AADTs followed by freeway and principal arterial highways. Northern samples over southern areas increased AADT estimates. Temporally, the weekday volumes showed higher ADT values for midweek over Friday and Monday. This in turn required a greater adjustment for the midweek followed by Friday and Monday.

The final results from this research were based on the overall predictive results of the models. Three models were developed within this study for each season. The first model was an ordinary least squares regression model, while the second and third were full Bayesian negative binomial models, the first with an offset and the second without an offset. The results demonstrated that Model Two, the negative binomial with an offset, performed the best while the other two models had mixed results. The main rationale of the offset was a direct result of the wide variation in the ADT values for heavy-duty vehicles. The offset limited the variation as seen in Model

Three with the higher ADTs and, therefore, produced more accurate results. The final conclusions indicate the limitations of Model One with respect to the prediction of negative AADTs. In this study, Model One predicted negative values approximately 10% of the time for each season, while the negative binomial by its very nature did not predict negative values. The final comparison showed that Model Two with the offset was the most efficient model form for predicting seasonal AADTs.

Acknowledgements

The authors thank Dave Gardner, Tony Manch, and Lindsey Pflum from the Ohio Department of Transportation (ODOT) for their valuable advice and help. The contents of this paper reflect the views of the authors and do not necessarily reflect the official views or polices of ODOT or FHWA.

References

AASHTO, 1992. *Guidelines for Traffic Data Programs.* Joint Task Force on Traffic Monitoring of the AASHTO Highway Subcommittee on Traffic Engineering. Washington, DC: American Association of State Highway and Transportation Officials.

Faghri, A., and J. Hua. 1995. "Roadway Seasonal Classification Using Neural Network." *Journal of Computing in Civil Engineering* 9 (1): 209–215. doi:10.1061/(ASCE)0887-3801(1995)9:3(209).

Federal Highway Administration. 2001. *Traffic Monitoring Guide.* Washington, DC: Office of Highway Policy Information.

Fricker, J., and K. Sinha. 1987. *Traffic Volume Forecasting Methods for Rural State Highways.* FHWA/IN/JHRP-86/20. West Lafayette, IN: School of Engineering, Purdue University.

Gelman, A., J. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian Data Analysis.* 2nd ed. New York: Chapman & Hall/CRC.

Lam, W., and J. Xu. 2000. "Estimation of AADT from Short Period Counts in Hong-Kong – A Comparison Between Neural Network Method and Regression Analysis." *Journal of Advanced Transportation* 34 (2): 249–268. doi:10.1002/atr.5670340205.

Lingras, P., S. C. Sharma, P. Osborne, and I. Kalyar. 2000. "Traffic Volume Time-Series Analysis According to the Type of Road Use." *Computer-Aided Civil and Infrastructure Engineering* 15 (5): 365–373. doi:10.1111/0885-9507.00200.

Lingras, P., S. C. Sharma, and M. Zhong. 2002. "Prediction of Recreational Travel Using Genetically Regression and Time-Delay Neural Network Models." *Transportation Research Record* 1805: 16–24. doi:10.3141/1805-03.

Mohamad, D., K. C. Sinha, T. Kuczek, and C. F. Scholer. 1998. "Annual Average Daily Traffic Prediction Model for County Roads." *Transportation Research Record* 1617: 69–77. doi:10.3141/1617-10.

Neveu, A. J. 1983. "Quick-Response Procedures to Forecast Rural Traffic." *Transportation Research Record* 944: 47–53.

Robichaud, K., and M. Gordon. 2003. "An Assessment of Data Collection Techniques for Highway Agencies." *Transportation Research Record* 1855: 129–135. doi:10.3141/1855-16.

Sharma, S. C., P. Lingras, G. X. Liu, and F. Xu. 2000. "Estimation of Annual Average Daily Traffic on Low-Volume Roads; Factor Approach Versus Neural Networks." *Transportation Research Record* 1719: 103–111. doi:10.3141/1719-13.

Sharma, S. C., P. Lingras, F. Xu, and P. Killburn. 2001. "Application of Neural Networks to Estimate AADT on Low-Volume Roads." *Journal of Transportation Engineering* 127 (5): 426–432. doi:10.1061/(ASCE)0733-947X(2001)127:5(426).

Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. 2004. "WinBUGS User Manual." Version 1.4, MRC Biostatistics Unit. Cambridge, UK. Accessed January 2013. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf.

Tang, Y. F., W. H. Lam, and L. G. Ng. 2003. "Comparison of Four Modeling Techniques for Short-Term AADT Forecasting in Hong-Kong." *Journal of Transportation Engineering* 129 (3): 271–277. doi:10.1061/(ASCE)0733-947X(2003)129:3(271).

US Census Bureau. 2008. "US Census Datasets." Accessed January 2013. http://www2.census.gov/census_2000/datasets/.

Xia, Q., F. Zhao, Z. Chen, D. Shen, and D. Ospina. 1999. "Estimation of Annual Average Daily Traffic for Nonstate Roads in Florida County." *Transportation Research Record* 1660: 32–40. doi:10.3141/1660-05.

Zhao, F., and N. Park. 2004. "Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic." *Transportation Research Record* 1879: 99–107. doi:10.3141/1879-12.

Zhao, F., and Z. Chung. 2001. "Contributing Factors of Annual Average Daily Traffic in a Florida County; Exploration with Geographic Information System and Regression Models." *Transportation Research Record* 1769: 113–122. doi:10.3141/1769-14.

Zhong, M., S. C. Sharma, and P. Lingras. 2004. "Genetically Designed Models for Accurate Imputations of Missing Traffic Counts." *Transportation Research Record* 1879: 71–79. doi:10.3141/1879-09.