SPATIAL PREDICTION OF AADT IN UNMEASURED LOCATIONS BY UNIVERSAL KRIGING

Brent Selby
Graduate Student Researcher
The University of Texas at Austin.
6.9 E. Cockrell Jr. Hall
Austin, TX 78712-1076
bfselby@mail.utexas.edu

Kara M. Kockelman,
(Corresponding author)
Professor and William J. Murray Jr. Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
6.9 E. Cockrell Jr. Hall
Austin, TX 78712-1076
kkockelm@mail.utexas.edu
Phone: 512-471-0210

Presented at the 90th Annual Meeting of the Transportation Research Board (2010).
Journal of Transport Geography 29: 24-32 (2013)

Key Words: Annual average daily traffic (AADT) prediction, universal kriging, traffic counts

ABSTRACT

This work explores the application of kriging methods for prediction of average daily traffic counts across the Texas network. Results based on Euclidean distances are compared to those using network distances, and both allow for strategic spatial interpolation of count values while controlling for each roadway's functional classification, lane count, speed limit, and other site attributes. Universal kriging is found to reduce errors (in practically and statistically significant ways) over non-spatial regression techniques, though errors remain quite high at some sites, particularly those with low counts and/or in less measurement-dense areas. Interestingly, the estimation of kriging parameters by network distances showed no enhanced performance over Euclidean distances, which require less data and are much more easily computed.

INTRODUCTION

Traffic flow volumes represent key information for proper transportation engineering and planning decisions. Sampling, tracking, interpolating, and extrapolating Annual Average Daily Traffic (AADT) counts is fundamental to road construction and maintenance scheduling, as well as to demand modeling and validating estimates of network activity. However, assembly of accurate and robust traffic counts is not straightforward, due to difficulties in measurement and

calculation. To obtain counts at a sample of specific sites across extensive road networks, departments of transportation (DOTs) tend to use a set of permanently located automatic traffic recorders (ATRs) in league with portable traffic counters (PTCs, for short-term count samples). (FHWA 2005) While a U.S. state DOT may have 100 ATRs across its network, it is likely to sample at tens of thousands of short-period traffic count (SPTC) sites, for two or three days each, typically. Overall, spacing between count sites can easily average 5 miles or more, due to limited resources and competing interests (see, e.g., Wang and Kockelman 2009).

The U.S. Federal Highway Administration's (FHWA's) standards state that interstates and other high-volume roads must be measured on a maximum three-year cycle, while other highways' count can be sampled up to every six years. Day-of-week averages are calculated for ATR sites each month. These are averaged over the months for each day's AADT and then all averaged to get a single AADT for that site. In the case of SPTC sites, data collection over at least 48 consecutive hours in a measurement cycle is recommended. Seasonal, day-of-week and month-of-year adjustments (to adjust SPTC values to AADT values) are calculated as ratios of ATR-site counts in the relevant time period to the year's count, based on average adjustments from groups of ATRs at similar/matched locations. (FHWA 2005)

The FHWA (2005) Highway Performance Monitoring System suggests that AADT estimates should lie within ± 10 percent of actual AADT values with a 90% confidence interval on urban arterials and 80% interval on all other roadway types. Using Minnesota and Florida ATR counts, Gadda et al. (2007) estimated that average errors in AADT estimation using one- and two-days' counts produced estimates that, on average, fell within 10 to 20 percent of actual counts, when using own-site adjustment factors (i.e., best-case scenario). The 95% confidence interval notion would suggest an even wider range, generally[1], particularly as one relies on other sites for the adjustment factors. Of course, modelers must also anticipate counts at locations wholly unobserved by ATRs and PTCs, where average error rates can rise quickly to 100% or above. (Gadda et al. 2007)

As discussed below, standard regression techniques, geographic weighted regression (GWR), and geostatistical methods have been used to estimate traffic counts at unmeasured locations. This paper endeavors to harness known local conditions that influence count and road network spatial information about measured locations using a geostatistical technique known as universal kriging. Kriging essentially involves spatial interpolation, and universal kriging makes use of local information (such as lane count and population density) while also drawing on residuals in prediction from nearby sites. Such methods cannot replace counting entirely, but can reduce the need, if spatial interpolation errors are low. Furthermore, this methodology is also useful for real-time count and speed and other predictions (to anticipate and avoid congestion via ITS techniques, for example), as well as demographic prediction (e.g., population densities, annual household travel distances, and/or vehicle ownership levels throughout a region – based on a sample of sites or households).

[1] If the average is 15 percent error in prediction, then there are probably many errors falling well beyond the 15-percent average – far more than 5 percent of the estimates will tend to fall out there. Perhaps the 70 percent confidence interval is on the order of 10 to 20 percent error.

Kriging's origins lie in the prediction of mineral contents by mining engineer D.G. Krige in the early 1950s. Mathematician George Matheron outlined kriging for geostatistics, using a "semivariogram" variance function (for latent effects or prediction residuals) that depends on the distance between data points. In general, there are three types of kriging: simple kriging, ordinary kriging, and universal kriging. In simple kriging, the value of interest at a location is predicted directly from nearby values, based on the semivariogram and a known global mean value. Ordinary kriging is slightly more complicated, requiring the process to estimate an unknown mean as well as the semivariogram. Universal kriging is used when a global-meal assumption cannot be used, and combines the distance-based variance with a trend, such as a linear, parametric function, as pursued here (following Box-Cox transformation of the traffic count [AADT] response variable).

EXISTING WORK

A variety of techniques have been implemented to estimate traffic counts. Each method takes known counts and uses additional information (e.g., local land use data, time-steps, road attributes, and nearby sites' residuals in count prediction) to make a prediction. These can be divided into future-year (or future-period) prediction and same-year prediction methods. Future year prediction uses current and past traffic data to estimate counts at the same locations at future dates. This is important for many applications, including planning maintenance or capacity increases on roadways, as well as real-time transportation systems management decisions (like signal timing, ramp metering, and variable tolling). In contrast, current-year prediction methods estimate counts at locations whose traffic flow have not been measured, using data from nearby locations during the same time period. This paper's applications center on current-year prediction only, but there is insight to be gained from both streams of work.

Future Year Prediction

Tang et al. (2003) tested and compared the Box-Jenkins, neural network (NN), nonparametric regression (NPR), and Gaussian maximum likelihood (GML) methods for short-term (less than one year into the future) prediction with data from densely urban Hong Kong. Their Box-Jenkins used autoregressive integrated moving average (ARIMA) specifications, which require an evenly-spaced time series data set. Their NNs iteratively adjusted a network of weighted sigmoidal equations using past-year traffic counts. Their NPR approach predicted counts by calculating similarity indices between the current state and prior states with known counts. Finally, their GML method used both flows and flow increments. They found the Box-Jenkins and NN methods to require considerably more calibration work, while producing higher errors. The GML and NPR techniques were easier to implement and performed better for their data sets.

Recently, Castro-Neto et al. (2009) implemented a "support vector regression with data-dependent parameters" for Tennessee's highway counts. This approach has similarities to standard least-squares regression techniques as well as NN methods. Its objective is to keep all residuals below a certain value, rather than minimizing the global sum of squared errors. This is useful when a modeler desires a certain level of accuracy for all points rather than maximum overall accuracy. They compared it to "Holt exponential smoothing" and found superior for longer prediction time steps and seasonal data.

Jiang et al. (2006) used a growth factor in conjunction with satellite images to enhance future year predictions. Satellite photos were reviewed for visible vehicles and adjusted by factors for time and season. The image-based estimates were then averaged with estimates from growth factor methods (using weights based on estimated variances of the two methods). Results suggested a great improvement in accuracy.

Current Year Prediction

Zhao and Chung (2001) used local employment and population attributes along with roadway details for current-year predictions across Broward County, Florida by ordinary least squares (OLS) regression. They compared several models and found that number of lanes, functional classification, regional access to employment, employment in an adjacent buffer zone (ranging from 0.25 to 3 miles on either side of the highway, based on road type), and direct access to expressways (via an indicator variable) worked best in predicting their data set's AADT values (with $n = 816$). 66 to 83 percent of the variability was explained by these variables, particularly the number of lanes and functional class. On their top performing model, they saw a mean squared error (MSE) of 50,000 vehicles per day and a bias of +0.25%.

Zhao and Park (2004) pursued a similar study, using geographically weighted regression (GWR) in place of OLS. This regression calculates local parameters using a distance-based weighting function (with a separate regression around each data point, essentially). The expectation is that variables have effects that may differ by location. Table 1 shows the variables included in their (and another's) model. The GWR specification was clearly better in terms of MSE, maximum error in prediction (136%), and error distributions, over the OLS method for the same data.

Zhao and Park (2004)	Eom et al. (2006)
Lanes	Lanes
Direct access to expressway (binary)	Suburban (indicator)
Employment in buffer zone	Urban (indicator)
Population in buffer zone	Median income (Census bloc level)
Job accessibility index (by travel time)	Functional class (indicators)

Table 1: Explanatory variables used in two previous studies

Wang and Kockelman (2006) used ordinary kriging functions built into ESRI's ArcGIS to estimate AADT counts, thus offering the advantage of being easily replicated by anyone with this popular software package. However, ordinary kriging does not allow the analyst to control for point-specific characteristics. Their findings suggested that given limited information, ordinary kriging can provide estimates of counts of unmeasured sites throughout a network, though errors can be significant. Their median (non-absolute) errors were 33%, meaning half of all predictions were more than 33% over the actual value. They also found ArcGIS to be very limiting.

Eom et al. (2006) used universal kriging to predict Box-Cox transformed AADT counts (as discussed below) on non-freeway facilities in Wake County, North Carolina. They tested three semivariogram models (Gaussian, exponential, and spherical) and four estimation methods (OLS, weighted least squares, maximum likelihood, and restricted maximum likelihood (REML)). Their results suggest that universal kriging improved prediction overall, particularly in urban locations. REML and WLS performed well in terms of errors, with REML slightly ahead. Improvements over non-spatial methods were more pronounced in the urban areas, where denser placement of measurement locations provides more nearby data points.

This study expands on the work of Wang and Kockelman (2006), Eom et al. (2006), and Zhao and Park (2004) by modeling AADT counts in Texas via a universal kriging model to distances measured on a network.

METHODOLOGY

In this work universal kriging was used with a Box-Cox transformation of all traffic counts. Box-Cox is a likelihood-maximizing power transform that gives skewed data a more normal distribution, thereby stabilize variation. It is performed by maximizing the likelihood function over a power variable, λ. The transformation equation is:

$$Y(\lambda) = \begin{cases} \left(Y^{\lambda} - 1\right)/\lambda & \lambda \neq 0 \\ \ln Y & \lambda = 0 \end{cases} \quad (1)$$

λ's estimation was performed during the data set's pre-processing, using an in-built STATA software command.

The following kriging theory and implementation were derived from content in Schabenberger and Gotway (2005) and Cressie (1993). WLS was chosen over REML techniques for relative ease of implementation, as well as comparable performance seen in Eom et el.'s (2006) work. Moreover, WLS does not require the assumption of the error term's distribution. The general equation for universal kriging is as follows:

$$z_i = \mu(x_i) + \varepsilon_i \quad (2)$$

where $\mu(x_i)$ typically is a linear function of explanatory variables at location i, ε_i is a spatially dependent error term, z_i is the dependent variable, and $i=\{1,2,\ldots,N\}$.

WLS can be applied to this with the matrix notation:

$$Z = X\beta + \varepsilon \quad (3)$$

where z is the vector of response outcomes (e.g., AADT or Box-Cox-transformed AADT values), and X is an N by $(K+1)$ data matrix, interacted with the linear parameters (β).

The variances of the N error terms (ε) are assumed to follow a semivariogram relation, $\gamma(h_{ij})$, as a function of distances (h_{ij}) between the locations of data points i and j. Here, such distances were

calculated both using Euclidean distances (the standard approach) and network distances, to see whether the latter enhances prediction. The semivariogram's parameters can be estimated with mean or trend removed using WLS – or simultaneously with mean parameters when using REML. Three types of theoretical semivariogram functions, each with parameter set $\theta = \{c_0, c_e, a_s\}$, were tested, to ascertain the best performance (Cressie 1993):

$$\begin{aligned}
&\text{Gaussian} && \gamma(h_{ij}; c_0, c_e, a_e) = c_0 + c_e\left(1 - e^{-h_{ij}/{a_s}^2}\right) \\
&\text{Spherical} && \gamma(h_{ij}; c_0, c_e, a_s) = c_0 + c_e\left(1.5\, h_{ij}/a_s - \left(.5\, h_{ij}/a_s\right)^3\right) \\
&\text{Exponential} && \gamma(h_{ij}; c_0, c_e, a_e) = c_0 + c_e\left(1 - e^{-h_{ij}/a_s}\right)
\end{aligned} \tag{4}$$

Feasible generalized least squares (FGLS) regression was used to recognize heteroskedasticity in error terms (but neglecting spatial autocorrelation across pairs of points), and enhanced estimates of AADT residuals. After performing the two-step FGLS estimation process, the squares of differences in all residuals were used in the Cressie-Hawkins robust estimator. This estimator divides the distances between points into a series of bins, from 0 miles to some maximum distance, and creates an empirical semivariogram using the following equation (Schabenberger and Gotway's [2005] Eq. 4.26):

$$\tilde{\gamma}(H) = {}^1\!/_2\left({}^1\!/_{|N(H)|}\textstyle\sum_{N(H)}|e_i - e_j|\right)^4 \Big/ \left(0.457 + {}^{0.494}\!/_{|N(H)|}\right) \tag{5}$$

where H is the distance bin, $N(H)$ is the number of ij pairs in that bin and e_i is the FGLS residual from point i.

An iterative least-squares approach converges on the values for c_0, c_e, and a_s that minimize the sum of squared residuals (between observed/empirical and smoothed/theoretical semivariogram values). In equation form, the objective was $\min\left(\tilde{\gamma}(H) - \gamma(H; c_0, c_e, a_s)\right)^2$ with respect to c_0, c_e, and a_s.

The covariance matrix for kriging, C_{dd}, is then estimated from the theoretical semivariogram and the FGLS error-term variance, σ^2. Additionally, a vector of covariances, c_{d0}, for error terms across all known-response locations and all target (predicted) locations can be estimated. Each value in these two matrices is given by the following equation:

$$C_{ij} = C(h_{ij}) = C(0) - \gamma(h_{ij}; c_0, c_e, a_s) = \sigma^2_{FGLS} - \gamma(h_{ij}; c_0, c_e, a_s) \tag{6}$$

With the inverse of the covariance matrix, C_{dd}, as the weight matrix, the β values can be re-estimated using a full-matrix-weighted least-squares regression, and response predictions (of Z) can be derived at all "new" locations x_0 as follows (Schabenberger and Gotway 2005):

$$\begin{aligned}
\hat{\beta} &= \left(X^T {C_{dd}}^{-1} X\right)^{-1} X^T {C_{dd}}^{-1} Z \\
\hat{Z}_0 &= \left(X_0 - c_{d0}^T {C_{dd}}^{-1} X\right)\hat{\beta} + c_{d0}^T {C_{dd}}^{-1} Z
\end{aligned} \tag{7}$$

where X_0 is the data matrix for the predicted locations and $\hat{Z}_0$ is their predicted Box-Cox transformed value.

All these equations were coded into MATLAB software, and the run time was under 20 seconds for the largest data subset described here (i.e., 4,979 known-count locations, and 1281 new/unknown-count locations [with AADT values available for all sites, and used for predictive-model validation purposes]).

Comparison of Results

The model parameters were estimated using a randomly selected collection of the data points from each regional sample analyzed. The remaining 10 to 20 percent of count sites, from each subset, were used for model validation. The prediction errors were measured using MSE and averages of absolute percentage errors. Though the model uses Box-Cox transformed AADT values, the reverse transformation was used to work directly with AADT estimates, $\hat{Z}_i$, before generating the MSE and percentage errors shared here.

$$MSE = {}^{1}\!/_{n_{unknown}} \sum_{i=1}^{n_{unknown}} \left(\hat{Z}_i - Z_i\right)^2 \quad (8)$$

$$Percentage\ Error = 100 \cdot {}^{\left(\hat{Z}_i - Z_i\right)}\!/_{Z_i} \quad (9)$$

In both equations there is a known value for each traffic count, $Z(s_i)$, from the data and an estimated count from the model, $\hat{Z}_i(s_i)$. Results are given as median percentage error and average absolute percentage error.

As noted earlier, both shortest-path network distances and Eucliean distances were used for *h*. Their prediction errors were compared to determine the value, if any, of using network distance. In each case all three semivariogram functions – linear, spherical, and exponential – were used. As a point of comparison, an aspatial FGLS approach (reflecting heteroskedasticity in count volume residuals) was used with the same.

MODEL INPUTS

All traffic counts and highway data used here come from the year 2005 in Texas, the U.S.'s second largest state (in population and area). Texas contains a number of major metropolitan areas, including Houston and Dallas-Fort Worth, as well as large swaths of sparsely populated land. AADT values vary tremendously across the state DOT's geocoded 79,000+ centerline-mile network. The sampled counts come from all types of roads, from local roads to interstates freeways, in both highly urban and very rural settings. Figure 1 shows where SPTCs are concentrated, with an average of 111.3 count sites per county (or one count site every 10 square miles, or every 3 miles of highway centerline, on average) in this particular data set (and a standard deviation of 70.8 sites per county).

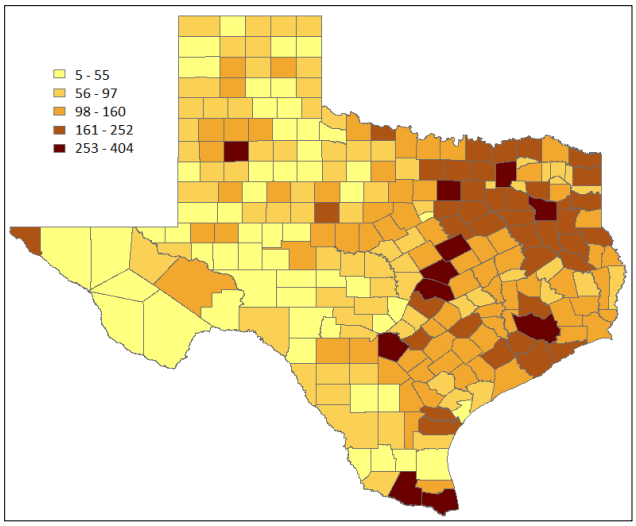

Figure 1: Number of annual count locations in each county

Data Sets

Traffic counts were derived by the Texas Department of Transportation (TxDOT) using ATR and short counts. Their annual count data set includes over 28,000 geocoded locations in 2005. Road information, (including number of lanes, functional class, and speed limits) was given in a GIS file of roads, represented as lines, with associated data from the TxDOT RHINO database. Population files contained block-group level data from the US Census. Employment data from the Department of Labor Statistics was collected and then geocoded into a map of county level polygons.

All data were spatially merged using ESRI's ArcGIS. Points with very or erroneously low traffic counts (less than 200 vehicles per day) were removed This resulted in 25,183 data points representing locations where counts were taken. These were divided into smaller, regional data sets for more local kriging to reduce computational demands on a standard PC. It is unnecessary and time consuming to, for example, include points from El Paso when predicting values in Houston, 700 miles away. A subset of 3,145 points from 24 counties around the southeastern, Gulf coast region of Texas, including the Houston Region[2] were evaluated in one set of models, followed by another set of 667 points for the 5-county Austin region. Both sets were selected by hand, with boundaries guided by areas of sparse point coverage, and included all functional classes of roads. Additionally, regressions were performed on point subsets with interstates only (n=1053), urban-classification only[3] (n=6,256), minor roads only (n=3,532), and the Houston Region only with interstates removed (n=3,017).

The resulting count concentrations (0.16 counts per square mile for the Houston region and 0.21 for the Austin region) are noticeably lower than those enjoyed by Eom et al. (2006) and Zhao and Park (2004) (at 1.35 and 0.65 counts per square mile, respectively). The more closely

[2] The Houston Region subset has 0.162 counts per square mile, and the average distance to each count site's three nearest neighbors is 3.06 miles, which is rather sparse.

[3] The Urban subset of sites is spread across the state (such that a density measure [in count per square mile] is not very meaningful here), but the average distance to each site's three nearest neighbors is relatively low, at 2.32 miles.

located data points are, the lower the resulting errors are likely to be, following spatial interpolation, ceteris paribus. Nevertheless, intelligent application of kriging remains a key tool of interest, particularly as data become costly and, typically, more sparse.

Variables Used

The speed limit, number of lanes and functional class were taken from the road segment associated with the count location using the overlay function in ArcGIS. The high values for median and average speed limit (55 and 56 mph, respectively, as shown in Table 2) hint at the fact that relatively few count sites lie in the city and town centers. The population and employment densities were derived for the county in which each count was taken. This is a very coarse measure of local density, of course, but does help reflect some of the longer distance travel that many regularly take (e.g., the NHTS 2005 data suggest that the average one-way commute trip in the U.S. is 12 miles long, while the average "radius" of a Texas county is 18 miles [if one were to form circles with the area of land present in Texas' 254 counties]). All variables' summary statistics, for only the data points included in the subsets, are given in Table 2.

	Mean	Std. Dev.	Min	Max
AADT 2005 (vehs/day)	17,843	33,601	210	341,940
Speed Limit (mph)	53.6	10.4	20	80
Lanes (number)	3.18	1.48	1	12
Persons / Acre	0.251	0.503	2.73E-4	2.55
Jobs / Sq Mile	0.576	0.962	1.04E-3	4.22
Rural Interstate (indicator)	0.049	-	0	1
Rural Major Road	0.188	-	0	1
Urban Interstate	0.047	-	0	1
Urban Principal Arterial	0.058	-	0	1
Local & Collector Roads*	0.658	-	0	1
Number of data points = 10,978 * Used as base case in regression				

Table 2: Summary statistics of model variables of data in all subsets

Fourteen functional classes of highway exist in Texas (as designated by TxDOT), with seven being rural in designation and seven urban. As shown later (and noted in Table 3), these were combined into six categories, based on regression results that indicated a lack of statistical distinction on coefficients for certain classes.

Not considered here is the measurement-type (ATR or PTC) for the counts. The dataset provided has no such distinction, so it was not an option in these analyses. In other data contexts, weighting by measurement type could be used to give more consideration to counts from permanent counters (ATR), thanks to their added reliability as known traffic count values.

Rural		*Urban*	
Functional Type	*Frequency*	*Functional Type*	*Frequency*
Interstate	533	Interstate	520
Principal Arterial[1]	315	Principal Arterial (Freeway/Expressway)[2]	630
Minor Arterial[1]	491	Principal Arterial (Other)[3]	2889
Major Collector[1]	1250	Minor Arterial[3]	1768
Minor Collector[3]	2108	Collector[3]	438
Local[3]	25	Local[3]	11
[1] Rural Major Road, [2] Urban Principal Arterial, [3] Minor Road			

Table 3: Frequency of traffic counts by functional class for data used in all subsets

Distance Measures

In previous applications of kriging for AADT count estimation, only Euclidean distances have been used (to estimate covariances via the semivariogram). Many experts would expect actual travel distance or impedance (time plus cost) to be a better indicator of count relationships; however, computing the hundreds of thousands of inter-point distances is challenging (if not impossible) for software like ArcGIS. Here, TransCAD travel demand modeling software was used to obtain shortest-path distances. (This activity required 7 hours to produce almost 800 million distance calculations across the Texas network.) Euclidean distances were calculated using the Vincenty formula for great circle distance. (Thomas and Featherstone 2005) All estimates and model performances are described below.

RESULTS

Tests on the various subsets of data show a marginal preference for the exponential semivariogram. Table 4 shows semivariogram parameter estimates under a variety of specifications for the Houston region subset and the minor roads subset (which traverses the entire state of Texas). Though there is some variability in the range parameter, it has different meanings for each equation, and there was no clear winner (in terms of error) for function choice.

		Semivariogram function	Parameters			Performance of model using these specifications	
			Nugget, c_0	Sill, c_e	Range, a_s	MSE	Avg abs. % error
Houston Region	Network Distance	Spherical	4.03	5.44	12.81	4.70E+08	63.9%
		Exponential	3.12	6.65	5.19	3.97E+08	62.5%
		Gaussian	4.91	4.59	6.41	5.07E+08	64.7%
	Euclidean Distance	Spherical	4.41	5.31	12.56	4.70E+08	63.0%
		Exponential	3.42	6.57	4.97	4.12E+08	62.6%
		Gaussian	5.26	4.49	6.27	5.28E+08	63.7%
Minor Roads	Network Distance	Spherical	4.93	6.28	17.41	6.77E+07	62.8%
		Exponential	3.73	7.89	6.81	5.97E+07	60.4%
		Gaussian	5.74	5.44	8.24	7.06E+07	63.9%
	Euclidean Distance	Spherical	5.23	5.89	16.76	7.35E+07	60.5%
		Exponential	4.22	7.34	6.83	6.56E+07	59.0%
		Gaussian	6.01	5.10	8.01	7.82E+07	61.6%

Table 4: Semivariogram parameter estimates

Figure 2 shows how the parameterized semivariogram function estimates compare to each other as a function of distance. The functions flatten at different points relative to their ranges, a_s, such that they are very close (despite a factor of two difference in their range estimates). Given that the effect tapers off at such a far distance (about 12 miles), it seems that this method, with this data, captures local effects, but not traffic flow effects from nearby highways

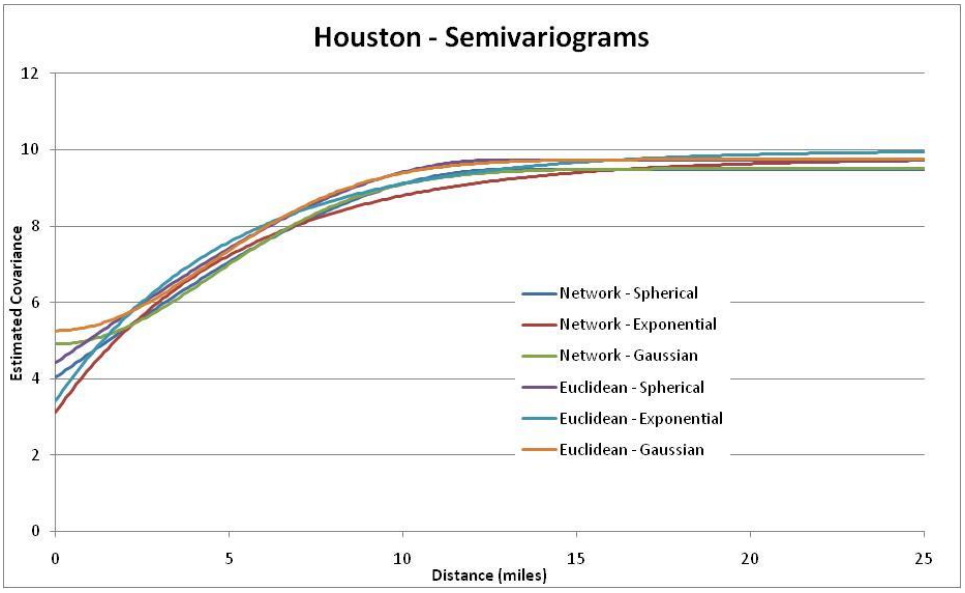

Figure 2: Estimated semivariogram functions

The other regional data sets similarly showed no strongly favored model. Each one has a slightly lower error for the exponential model but neither distance measure was consistently favored.

This was unexpected, because the network distances contain more information relating to the effective separation between points. Contrary to what may be expected, taking the interstate highways out of the data set produced a slight increase in average errors. (This was not expected since interstates have dramatically higher traffic counts [and thus higher variance values, whereas kriging assumes homoskedastic error terms] and may carry a lot of through traffic [offering less correlated information for sharing with neighboring sites].)

Table 5 shows the range of results which can be seen in the different data sets. The median percentage errors suggest an inconsistent bias hovering on either side of zero. The average absolute errors are very high for many of the subsets. The Houston and urban-road regressions exhibit the highest absolute average errors, though the latter offers the densest point structure. However, urban regions and facilities can be quite a bit more complex in their flow variation over space (thanks to land use diversity and network complexity). Lower percentage errors were associated with sites of higher count and higher count density, thanks, presumably, to a higher correlation between traffic flow and lane numbers (capacity) on larger facilities and the benefits of more information from nearby points.

		Austin Region	Houston Region	Houston Region, No IS	Minor Roads	Interstates	Urban Roads
Network	MSE	2.62E+08	3.97E+08	2.66E+08	5.97E+07	2.84E+08	3.68E+08
	Avg abs err	53.9%	62.5%	63.7%	60.4%	19.4%	62.4%
	Median err	-2.8%	5.6%	5.2%	1.5%	-4.6%	-3.1%
	Best Model	Spherical	Exponential	Exponential	Exponential	Exponential	Spherical
Euclidean	MSE	2.72E+08	4.12E+08	2.71E+08	6.56E+07	3.24E+08	4.11E+08
	Avg abs err	54.4%	62.6%	63.6%	59.0%	20.3%	62.2%
	Median err	-3.6%	5.1%	4.5%	1.3%	-2.7%	-2.8%
	Best Model	Spherical	Exponential	Exponential	Exponential	Exponential	Exponential
FGLS	MSE	4.20E+08	7.98E+08	6.79E+08	1.48E+08	1.14E+09	5.37E+08
	Avg abs err	115.3%	103.0%	103.6%	114.0%	38.4%	80.6%
	Median err	-8.6%	9.1%	8.5%	6.8%	-10.9%	-3.4%

Table 5: Errors in prediction for all six data subsets

Interstates performed relatively well, as a modeled subset, perhaps as a result of the nearby count locations often being on the same route. With the lowest number of nearby count locations and a steep semivariogram function, the interstate count estimates were influenced by counts of only the nearest 7 (on average) count sites (compared to roughly 30 or more nearby sites for other data sets analyzed here).

While each model's percentage errors in prediction on the hold-out samples are significant, they are a dramatic improvement over non-spatial FGLS techniques, averaging between 19 and 62 percentage points lower. The greatest improvement from kriging application was seen in the Austin data set, as well as those with lower traffic counts in general (i.e., minor roads). Figure 3 illustrates error comparisons for the Houston data sets. Unfortunately, all three models have a number of severe outliers. However, kriging's improvement is still quite apparent.

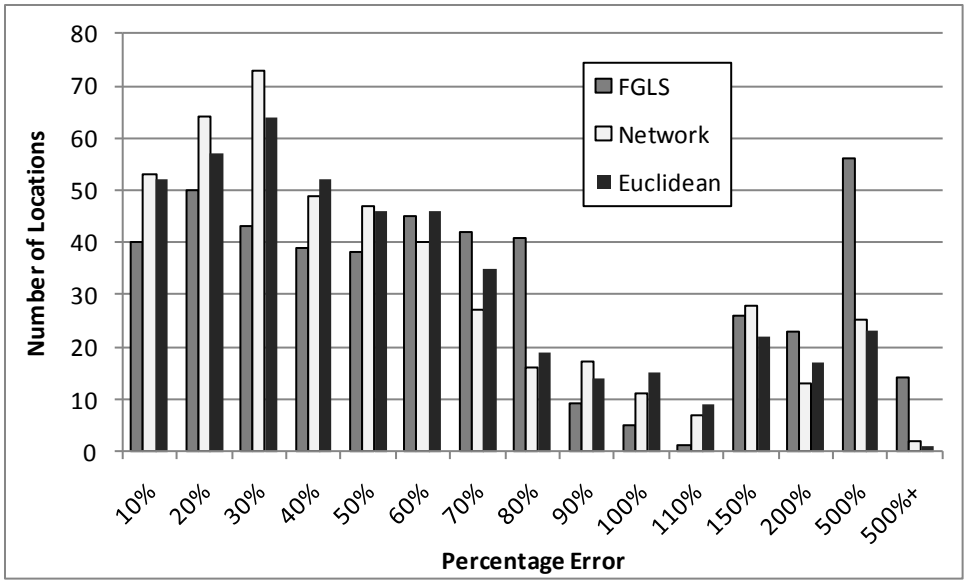

Figure 3: Comparison of percentage prediction errors for aspatial (FGLS) versus kriging techniques, using the Houston data set

Figure 4 illustrates the errors at Houston's 473 hold-out (prediction) locations. The vast majority of extreme outliers (i.e., those with the highest error percentages lie among low-count sites. These are generally rural non-interstate roads and some small urban roads with two lanes. Count prediction along Houston's interstates and urban arterials performed far better. The map shows that estimates at points closer to Houston's downtown tend to be lower errors, but there are few strong, regional geographic trends seen.

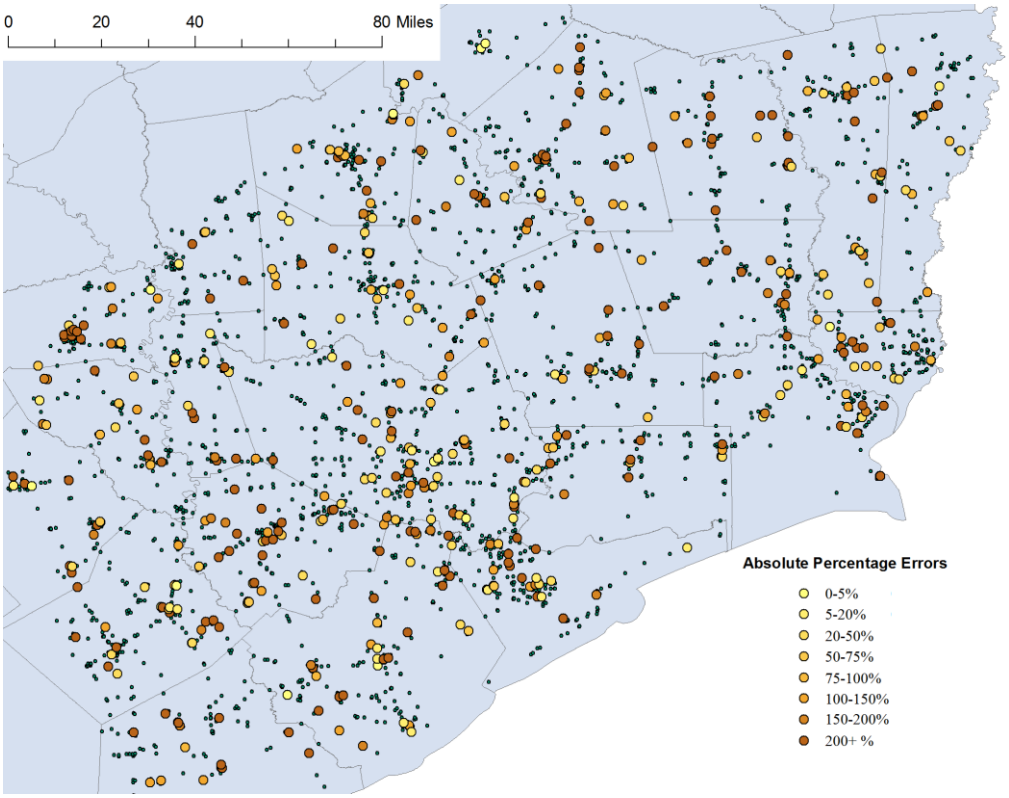

Figure 4: Errors in Houston Area using universal kriging with network distances
(Note: Small points indicate "known" locations.)

Table 6 provides the estimated parameter values for each model type using the Houston area data set, along with adjusted R^2 values. As shown in Table 6, and earlier, in Table 5, the network- and Euclidean-based kriging models performed almost identically. Some coefficient values vary a fair bit between the spatial and aspatial models (e.g., those on the number of lanes and the indicator for Rural Major Road), but very little between the two kriging models (with Euclidean and network distances). The kriging methods show a decisive improvement in the amount of variance explained by the variables (as exemplified by the adjusted R2 values).

	FGLS		Network		Euclidean	
	Beta	t-stat	Beta	t-stat	Beta	t-stat
Constant	10.66	23.63	7.72	17.11	7.85	17.40
Speed	0.0004	0.05	0.0276	4.04	0.0339	4.96
Lanes	2.01	32.99	1.49	24.47	1.55	25.32
Employment / Acre	-9.47	-10.14	-8.71	-9.33	-7.48	-8.01
Population / Acre	5.57	12.24	5.55	12.18	4.62	10.16
Rural Interstate	5.16	12.70	7.39	18.18	7.25	17.84
Rural Major Road	0.52	3.28	2.08	13.12	2.09	13.21
Urban Interstate	3.79	7.76	4.79	9.81	4.68	9.58
Urban Arterial	3.57	9.81	3.71	10.20	3.65	10.04
	Adj R^2 = .674		Adj R^2 = 0.971		Adj R^2 = 0.970	

Table 6: Coefficient estimates and results from FGLS and kriging with the exponential model for Houston subset

Note: Values apply to Box-Cox transformed AADT values.

One final issue deserving attention relates to the covariance matrices used here. When non-Euclidean distances are used in kriging, the covariance matrix may not be positive semi-definite (PSD), a condition necessary for mode validity (Curriero, 2006). In the cases when this occurred here, the estimates' errors could be consistent with those of the other models or wildly high. Interestingly, the data subsets that had this problem exhibited it for just one or two – but not all three – semivariogram models. Since the Euclidean-distance methods used here perform nearly as well as their network-distance counterparts (and are far easier to estimate in practice), it seems wise to simply use Euclidean distances.

CONCLUSIONS

This study has shown that universal kriging can provide more accurate traffic count estimates than aspatial regression, across a variety of road types in Texas. Moreover, Euclidean distance-based kriging fared just about as well as network-based metrics, suggesting that the latter's complexity is not warranted in such applications. Universal kriging reduced error by controlling for local attributes and recognizing distance-based correlation structures that exploit information found in nearby residuals. Recognition of such covariates resulted in average absolute error reductions between 16% and 79% here, depending on the data set and model specification used. Both the spatial and aspatial methods examined here offered their lower error percentages in the Austin application, and, in particular, on the interstate system's data point subset. Errors tended to be lower at locations with higher counts and more nearby count locations, though the urban set, which was above average for both, offered substantial count variation and thus was amongst the highest in overall errors.

It was interesting to find that network distances offered little improvement to models' predictive performance than Euclidean distances. This was the case for every subset of data tested. It is possible that with a more dense set of count locations the model would benefit more from the additional information provided in network distances, but with TxDOT's current data, this is the

limit. Given the number of links and sites of interest in large networks, like the ones used here, calculation of shortest path distances is probably unwarranted (especially since it can be very computationally intensive and requires additional information on network structures).

The issue with non-PSD covariance matrices makes network distances even less compelling. To combat or work around this problem, some solutions have been suggested, including spatial moving averages (Ver Hoef et al., 2006) and low-rank thin plate splines (Wang and Ranalli, 2007). Cressie and Johannesson's (2008) "fixed rank kriging" scheme uses scales of spatial dependence to create the covariance matrix, which they show is always positive semi-definite.

As a way of exploiting spatial information (while capitalizing on local attributes), universal kriging is worthy of application in a variety of transportation and other contexts. Opportunities to improve upon universal kriging, to better reflect heteroskedasticity in response variability, would be useful. Though the implementation here included the Cressie-Hawkins method and FGLS (as opposed to only OLS), estimation of the covariance matrix, C_{dd}, still relies a constant-variance assumption. An interesting alternative to be considered, for future work, involves calculation of covariances as a function of inter-site distances and other attributes using site-pair interaction variables (such as indicators for similarity in road type and lane counts). To do this, an expanded logistic function (to ensure estimates within the [-1,+1] range) may be used for correlations, with variance estimates coming from the FGLS procedure's estimates (as used in this work). A full (non-diagonal) variance-covariance matrix that does not presume homoskedasticity and does allow for interesting spatial (and other) effects could then be applied, providing more information in the spatial interpolation process for AADT estimation – and for myriad other spatial application contexts that analysts across a variety of disciplines may be studying (e.g., pavement quality, population densities, home values, vegetative cover, soil quality, and water chemistry). Spatial data surround us, and more specifications (and data contexts) should be evaluated.

ACKNOWLEDGEMENTS

The authors appreciate the technical assistance provided by Dr. Jianming Ma of the Texas Department of Transportation and the help and support of Annette Perrone at the University of Texas, Austin. The National Science Foundation Award SES-0818066 provided the support for this research.

REFERENCES

AASHTO Guidelines for Traffic Data Programs. 1992 AASHTO, Washington, DC.

Collins, S. 1991. *Prediction Techniques for Box-Cox Regression Models*. Washington, DC: Board of Governors of the Federal Reserve System.

Cressie, N. 1993. *Statistics for Spatial Data.* New York: Wiley Interscience.

Cressie, N. and Johannesson, G. 2008. Fixed-rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society B,* 70(1): 209-226.

Curriero, F. 2006. On the Use of Non-Euclidean Distance Measures in Geostatistics1. *Mathematical Geology,* 28(8): 907-926.

528 Draper, N. R., and Cox, D. R. 1969. On Distributions and Their Transformation to Normality.
529 *Journal of the Royal Statistical Society B*, 31:472–476.

530 Eom, J.K., Park, M.S., Heo, T.Y. and Huntsinger, L.F. 2006. Improving the Prediction of Annual
531 Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method.
532 *Artificial Intelligence and Advanced Computing Applications* 1968: 20-29.

533 FHWA. 2005. Highway Performance Management System Field Manual Federal Highway
534 Administration, Washington, DC.
535 http://www.fhwa.dot.gov/policy/ohpi/hpms/fieldmanual/.

536 Gadda, S., Kockelman, K., and Maggon, A. 2007. Estimates of AADT: Quantifying the
537 Uncertainty. Presented at the June 2007 *World Conference on Transportation Research,*
538 *Berkeley, California*. Accessed May 20, 2010.
539 (www.ce.utexas.edu/prof/kockelman/public_html/TRB07AADTUncertainty.pdf.)

540 Hu, P and Reuscher, T, 2004. Summary of Travel Trends: 2001 National Household Travel
541 Survey. Washington, DC: Federal Highway Administration. Accessed November 2010.
542 (http://nhts.ornl.gov/2001/pub/STT.pdf)

543 Schabenberger, O. and Gotway, C. 2005. *Statistical Methods for Spatial Data Analysis*. New
544 York: Chapman & Hall/CRC.

545 Tang, Y.F. Lam, W.H.K. and Ng, P.L.P. 2003. Comparison of Four Modeling Techniques for
546 Short-Term AADT Forecasting in Hong Kong. *Journal of Transportation Engineering* 129
547 (3): 271-277.

548 Thomas, C. and Featherstone, W. E. 2005. Validation of Vincenty's Formulas for the Geodesic
549 Using a New Fourth-Order Extension of Kivioja's Formula. *Journal of Surveying*
550 *Engineering* 131 (1): 20-26.

551 Ver Hoef, J., Peterson, E, and Theobald. D., 2006. Spatial Statistical Models that Use Flow and
552 Stream Distance. *Environmental and Ecological Statistics*. 13(4): 449-464.

553 Wang, X and Kockelman, K. 2009. Forecasting Network Data: Spatial Interpolation of Traffic
554 Counts Using Texas Data. *Transportation Research Record*, 2105: 100-108.

555 Wang, H and Ranalli, M.G. 2007. *Low-Rank Smoothing Splines on Complicated Domains*.
556 Biometrics 63:209–217.

557 Webster, R. and Oliver, M. 2001. *Geostatistics for Environmental Scientists*. New York: John
558 Wiley and Sons.

559 Zhao, F. and Chung, S. 2001. Contributing Factors of Annual Average Daily Traffic in A Florida
560 County - Exploration with Geographic Information System and Regression Models.
561 *Pavement Management, Monitoring, and Accelerated Testing* 1769: 113-122.

562 Zhao, F. and Park, N. 2004. Using Geographically Weighted Regression Models to Estimate
563 Annual Average Daily Traffic. *Transportation Research Record: Journal of the*
564 *Transportation Research Board* 1879: 99-107.