1	
2	SPATIAL PREDICTION OF AADT IN UNMEASURED LOCATIONS
3	BY UNIVERSAL KRIGING
4	
5	Brent Selby
6	Graduate Student Researcher
7	The University of Texas at Austin.
8	69E Cockrell Ir Hall
9	Austin TX 78712-1076
10	hfselby@mail.utexas.edu
11	orseno y e man.uczas.edu
12	Kara M. Kockelman
12	(Corresponding author)
13	Professor and William I. Murray Ir. Follow
14	Department of Civil Architectural and Environmental Engineering
15	The University of Texas at Austin
10 17	6 0 E. Cookroll In Holl
1/	0.9 E. COCKIEII JI. Hall Anotin TX 79712 1076
10	Austin, $1X / 8 / 12 - 10 / 0$
19	KKOCKEIM@mail.utexas.edu
20	Phone: 512-4/1-0210
21	
22	Presented at the 90th Annual Meeting of the Transportation Research Board in
23	Washington DC and under review for publication to the
24	Journal of Transport Geography, January 2011
25	
26	<b>Key Words:</b> Annual average daily traffic (AADT) prediction, universal kriging, traffic counts
27	
28	ABSTRACT
29	
30	This work explores the application of kriging methods for prediction of average daily traffic
31	counts across the Texas network. Results based on Euclidean distances are compared to those
32	using network distances, and both allow for strategic spatial interpolation of count values while
33	controlling for each roadway's functional classification, lane count, speed limit, and other site
34	attributes. Universal kriging is found to reduce errors (in practically and statistically significant
35	ways) over non-spatial regression techniques, though errors remain quite high at some sites,
36	particularly those with low counts and/or in less measurement-dense areas. Interestingly, the
37	estimation of kriging parameters by network distances showed no enhanced performance over
38	Euclidean distances, which require less data and are much more easily computed.
39	
40	INTRODUCTION
41	
42	Traffic flow volumes represent key information for proper transportation engineering and
43	planning decisions. Sampling, tracking, interpolating, and extrapolating Annual Average Daily
44	Traffic (AADT) counts is fundamental to road construction and maintenance scheduling, as well
45	as to demand modeling and validating estimates of network activity. However, assembly of
46	accurate and robust traffic counts is not straightforward, due to difficulties in measurement and

- 47 calculation. To obtain counts at a sample of specific sites across extensive road networks,
- 48 departments of transportation (DOTs) tend to use a set of permanently located automatic traffic
- 49 recorders (ATRs) in league with portable traffic counters (PTCs, for short-term count samples).
- 50 (FHWA 2005) While a U.S. state DOT may have 100 ATRs across its network, it is likely to
- 51 sample at tens of thousands of short-period traffic count (SPTC) sites, for two or three days each,
- 52 typically. Overall, spacing between count sites can easily average 5 miles or more, due to limited
- 53 resources and competing interests (see, e.g., Wang and Kockelman 2009).
- 54
- 55 The U.S. Federal Highway Administration's (FHWA's) standards state that interstates and other
- 56 high-volume roads must be measured on a maximum three-year cycle, while other highways'
- 57 count can be sampled up to every six years. Day-of-week averages are calculated for ATR sites
- 58 each month. These are averaged over the months for each day's AADT and then all averaged to
- 59 get a single AADT for that site. In the case of SPTC sites, data collection over at least 48
- 60 consecutive hours in a measurement cycle is recommended. Seasonal, day-of-week and month-
- 61 of-year adjustments (to adjust SPTC values to AADT values) are calculated as ratios of ATR-site
- 62 counts in the relevant time period to the year's count, based on average adjustments from groups
- 63 of ATRs at similar/matched locations. (FHWA 2005)
- 64
- 65 The FHWA (2005) Highway Performance Monitoring System suggests that AADT estimates
- should lie within  $\pm$  10 percent of actual AADT values with a 90% confidence interval on urban 66
- 67 arterials and 80% interval on all other roadway types. Using Minnesota and Florida ATR counts,
- 68 Gadda et al. (2007) estimated that average errors in AADT estimation using one- and two-days'
- 69 counts produced estimates that, on average, fell within 10 to 20 percent of actual counts, when 70
- using own-site adjustment factors (i.e., best-case scenario). The 95% confidence interval notion would suggest an even wider range, generally<sup>1</sup>, particularly as one relies on other sites for the 71
- 72 adjustment factors. Of course, modelers must also anticipate counts at locations wholly
- 73
- unobserved by ATRs and PTCs, where average error rates can rise quickly to 100% or above. 74 (Gadda et al. 2007)
- 75
- 76 As discussed below, standard regression techniques, geographic weighted regression (GWR),
- 77 and geostatistical methods have been used to estimate traffic counts at unmeasured locations.
- 78 This paper endeavors to harness known local conditions that influence count and road network
- 79 spatial information about measured locations using a geostatistical technique known as universal
- 80 kriging. Kriging essentially involves spatial interpolation, and universal kriging makes use of
- 81 local information (such as lane count and population density) while also drawing on residuals in
- 82 prediction from nearby sites. Such methods cannot replace counting entirely, but can reduce the
- 83 need, if spatial interpolation errors are low. Furthermore, this methodology is also useful for
- 84 real-time count and speed and other predictions (to anticipate and avoid congestion via ITS
- 85 techniques, for example), as well as demographic prediction (e.g., population densities, annual
- 86 household travel distances, and/or vehicle ownership levels throughout a region - based on a
- 87 sample of sites or households).
- 88

<sup>&</sup>lt;sup>1</sup> If the average is 15 percent error in prediction, then there are probably many errors falling well beyond the 15percent 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.

89 Kriging's origins lie in the prediction of mineral contents by mining engineer D.G. Krige in the

90 early 1950s. Mathematician George Matheron outlined kriging for geostatistics, using a

- 91 "semivariogram" variance function (for latent effects or prediction residuals) that depends on the
- 92 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
- 94 predicted directly from nearby values, based on the semivariogram and a known global mean 95 value. Ordinary kriging is slightly more complicated, requiring the process to estimate an
- 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
- 97 assumption cannot be used, and combines the distance-based variance with a trend, such as a
- 98 linear, parametric function, as pursued here (following Box-Cox transformation of the traffic
- 99 count [AADT] response variable).
- 100

## 101 EXISTING WORK

102

103 A variety of techniques have been implemented to estimate traffic counts. Each method takes

104 known counts and uses additional information (e.g., local land use data, time-steps, road

105 attributes, and nearby sites' residuals in count prediction) to make a prediction. These can be

106 divided into future-year (or future-period) prediction and same-year prediction methods. Future

107 year prediction uses current and past traffic data to estimate counts at the same locations at future

108 dates. This is important for many applications, including planning maintenance or capacity

109 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

113 prediction only, but there is insight to be gained from both streams of work.

114

## 115 Future Year Prediction

116

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

110 regression (NPR), and Gaussian maximum fikelihood (GML) methods for short-term (less maximum fikel

120 Jenkins used autoregressive integrated moving average (ARIMA) specifications, which require

- 121 an evenly-spaced time series data set. Their NNs iteratively adjusted a network of weighted
- 122 sigmoidal equations using past-year traffic counts. Their NPR approach predicted counts by
- 123 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.
- 126 The GML and NPR techniques were easier to implement and performed better for their data sets.
- 127
- 128 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

130 standard least-squares regression techniques as well as NN methods. Its objective is to keep all

- 131 residuals below a certain value, rather than minimizing the global sum of squared errors. This is
- 132 useful when a modeler desires a certain level of accuracy for all points rather than maximum
- 133 overall accuracy. They compared it to "Holt exponential smoothing" and found superior for
- 134 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.

141

## 142 Current Year Prediction

143

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

146 (OLS) regression. They compared several models and found that number of lanes, functional

147 classification, regional access to employment, employment in an adjacent buffer zone (ranging

148 from 0.25 to 3 miles on either side of the highway, based on road type), and direct access to

149 expressways (via an indicator variable) worked best in predicting their data set's AADT values

150 (with n = 816). 66 to 83 percent of the variability was explained by these variables, particularly

151 the number of lanes and functional class. On their top performing model, they saw a mean

152 squared error (MSE) of 50,000 vehicles per day and a bias of +0.25%.

153

154 Zhao and Park (2004) pursued a similar study, using geographically weighted regression (GWR)

155 in place of OLS. This regression calculates local parameters using a distance-based weighting

156 function (with a separate regression around each data point, essentially). The expectation is that

157 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.

160

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

161

Table 1: Explanatory variables used in two previous studies

162

163 Wang and Kockelman (2006) used ordinary kriging functions built into ESRI's ArcGIS to

164 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

166 for point-specific characteristics. Their findings suggested that given limited information,

167 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

170 limiting.

- 172 Eom et al. (2006) used universal kriging to predict Box-Cox transformed AADT counts (as
- 173 discussed below) on non-freeway facilities in Wake County, North Carolina. They tested three
- semivariogram models (Gaussian, exponential, and spherical) and four estimation methods
- 175 (OLS, weighted least squares, maximum likelihood, and restricted maximum likelihood
- 176 (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
- 178 ahead. Improvements over non-spatial methods were more pronounced in the urban areas, where
- denser placement of measurement locations provides more nearby data points.
- 180

181 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 distancesmeasured on a network.
- 183 184

## 185 **METHODOLOGY**

186

187 In this work universal kriging was used with a Box-Cox transformation of all traffic counts.

188 Box-Cox is a likelihood-maximizing power transform that gives skewed data a more normal

189 distribution, thereby stabilize variation. It is performed by maximizing the likelihood function 190 over a power variable,  $\lambda$ . The transformation equation is:

191

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

(1)

193

194  $\lambda$ 's estimation was performed during the data set's pre-processing, using an in-built STATA 195 software command.

196

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:

202 203

 $z_i = \mu(x_i) + \varepsilon_i \tag{2}$ 

204

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

WLS can be applied to this with the matrix notation:

 $Z = X\beta + \varepsilon \tag{3}$ 

210 211

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 ( $\beta$ ).

The variances of the *N* error terms ( $\varepsilon$ ) 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):

222 223

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

225 226

224

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):

234

235

236

$$\widetilde{\gamma}(H) = \frac{1}{2} \left( \frac{1}{|N(H)|} \sum_{N(H)} |e_i - e_j| \right)^4 / \left( 0.457 + \frac{0.494}{|N(H)|} \right)$$
(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  $\min(\tilde{\gamma}(H) - \gamma(H;c_0,c_e,a_s))^2$  with respect to  $c_0$ ,  $c_e$ , and  $a_s$ .

244

The covariance matrix for kriging,  $C_{dd}$ , is then estimated from the theoretical semivariogram and the FGLS error-term variance,  $\sigma^2$ . 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:

251

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

With the inverse of the covariance matrix,  $C_{dd}$ , as the weight matrix, the  $\beta$  values can be reestimated 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):

256  $\hat{\beta} = (X^T C_{dd}^{-1} X)^{-1} X^T C_{dd}^{-1} Z$ 

258

 $\hat{Z}_{0} = (X_{0} - c_{d0}^{T} C_{dd}^{-1} X) \hat{\beta} + c_{d0}^{T} C_{dd}^{-1} Z$ (7)

- where  $X_0$  is the data matrix for the predicted locations and  $\hat{Z}_0$  is their predicted Box-Cox 259
- 260 transformed value.
- 261

262 All these equations were coded into MATLAB software, and the run time was under 20 seconds

263 for the largest data subset described here (i.e., 4,979 known-count locations, and 1281

264 new/unknown-count locations [with AADT values available for all sites, and used for predictive-265 model validation purposes]).

266

#### 267 **Comparison of Results**

268 269

The model parameters were estimated using a randomly selected collection of the data points 270 from each regional sample analyzed. The remaining 10 to 20 percent of count sites, from each 271 subset, were used for model validation. The prediction errors were measured using MSE and 272 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 273 274 generating the MSE and percentage errors shared here.

275 276

$$MSE = \frac{1}{n_{unknown}} \sum_{i=1}^{n_{unknown}} \left(\hat{Z}_i - Z_i\right)^2 \tag{8}$$

Percentage Error = 
$$100 \cdot \frac{\hat{Z}_i - Z_i}{Z_i}$$
 (9)

278

279 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 280 281 average absolute percentage error.

282

283 As noted earlier, both shortest-path network distances and Eucliean distances were used for h. 284 Their prediction errors were compared to determine the value, if any, of using network distance. 285 In each case all three semivariogram functions – linear, spherical, and exponential – were used. 286 As a point of comparison, an aspatial FGLS approach (reflecting heteroskedasticity in count 287 volume residuals) was used with the same. 288

#### 289 **MODEL INPUTS**

290

291 All traffic counts and highway data used here come from the year 2005 in Texas, the U.S.'s 292 second largest state (in population and area). Texas contains a number of major metropolitan 293 areas, including Houston and Dallas-Fort Worth, as well as large swaths of sparsely populated 294 land. AADT values vary tremendously across the state DOT's geocoded 79,000+ centerline-mile 295 network. The sampled counts come from all types of roads, from local roads to interstates 296 freeways, in both highly urban and very rural settings. Figure 1 shows where SPTCs are 297 concentrated, with an average of 111.3 count sites per county (or one count site every 10 square 298 miles, or every 3 miles of highway centerline, on average) in this particular data set (and a 299 standard deviation of 70.8 sites per county).





Figure 1: Number of annual count locations in each county

#### 304 **Data Sets**

305 306 Traffic counts were derived by the Texas Department of Transportation (TxDOT) using ATR 307 and short counts. Their annual count data set includes over 28,000 geocoded locations in 2005. 308 Road information, (including number of lanes, functional class, and speed limits) was given in a 309 GIS file of roads, represented as lines, with associated data from the TxDOT RHINO database. 310 Population files contained block-group level data from the US Census. Employment data from 311 the Department of Labor Statistics was collected and then geocoded into a map of county level polygons.

312 313

314 All data were spatially merged using ESRI's ArcGIS. Points with very or erroneously low traffic 315 counts (less than 200 vehicles per day) were removed This resulted in 25,183 data points 316 representing locations where counts were taken. These were divided into smaller, regional data 317 sets for more local kriging to reduce computational demands on a standard PC. It is unnecessary 318 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, 319 Gulf coast region of Texas, including the Houston Region<sup>2</sup> were evaluated in one set of models, 320 followed by another set of 667 points for the 5-county Austin region. Both sets were selected by 321 322 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 323

- 324 (n=1053), urban-classification only<sup>3</sup> (n=6,256), minor roads only (n=3,532), and the Houston
- 325 Region only with interstates removed (n=3,017).
- 326
- 327 The resulting count concentrations (0.16 counts per square mile for the Houston region and 0.21 328 for the Austin region) are noticeably lower than those enjoyed by Eom et al. (2006) and Zhao 329
  - and Park (2004) (at 1.35 and 0.65 counts per square mile, respectively). The more closely

 $<sup>^{2}</sup>$  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.

 $<sup>^{3}</sup>$  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.

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

#### 333 334 Variables Used

335

336 The speed limit, number of lanes and functional class were taken from the road segment 337 associated with the count location using the overlay function in ArcGIS. The high values for 338 median and average speed limit (55 and 56 mph, respectively, as shown in Table 2) hint at the 339 fact that relatively few count sites lie in the city and town centers. The population and 340 employment densities were derived for the county in which each count was taken. This is a very 341 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 342 343 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 344 345 variables' summary statistics, for only the data points included in the subsets, are given in Table 346 2.

347

	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							

348

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

349

350 Fourteen functional classes of highway exist in Texas (as designated by TxDOT), with seven 351 being rural in designation and seven urban. As shown later (and noted in Table 3), these were 352 combined into six categories, based on regression results that indicated a lack of statistical

353 distinction on coefficients for certain classes.

354

355 Not considered here is the measurement-type (ATR or PTC) for the counts. The dataset

356 provided has no such distinction, so it was not an option in these analyses. In other data

357 contexts, weighting by measurement type could be used to give more consideration to

358 counts from permanent counters (ATR), thanks to their added reliability as known traffic

359 count values.

Rural		Urban				
Functional Type	Frequency	Functional Type	Frequency			
Interstate	533	Interstate	520			
		Principal Arterial				
Principal Arterial <sup>1</sup>	315	(Freeway/Expressway) <sup>2</sup>	630			
Minor Arterial <sup>1</sup>	491	Principal Arterial (Other) <sup>3</sup>	2889			
Major Collector <sup>1</sup>	1250	Minor Arterial <sup>3</sup>	1768			
Minor Collector <sup>3</sup>	2108	Collector <sup>3</sup>	438			
Local <sup>3</sup>	25	Local <sup>3</sup>	11			
$1 p_{1} + 1 M_{1}^{2} + p_{2} + 2 M_{1}^{2} + p_{1}^{2} + 1 M_{2}^{2} + \frac{1}{3} M_{1}^{2} + p_{1}^{2} + \frac{1}{3} M_{1}^{2} + \frac{1}$						

<sup>1</sup> Rural Major Road, <sup>2</sup> Urban Principal Arterial, <sup>3</sup> Minor Road

360

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

361

362 **Distance Measures** 

363

364 In previous applications of kriging for AADT count estimation, only Euclidean distances have 365 been used (to estimate covariances via the semivariogram). Many experts would expect actual 366 travel distance or impedance (time plus cost) to be a better indicator of count relationships; 367 however, computing the hundreds of thousands of inter-point distances is challenging (if not 368 impossible) for software like ArcGIS. Here, TransCAD travel demand modeling software was 369 used to obtain shortest-path distances. (This activity required 7 hours to produce almost 800 370 million distance calculations across the Texas network.) Euclidean distances were calculated 371 using the Vincenty formula for great circle distance. (Thomas and Featherstone 2005) All 372 estimates and model performances are described below.

## 374 **RESULTS**

375

373

Tests on the various subsets of data show a marginal preference for the exponential

377 semivariogram. Table 4 shows semivariogram parameter estimates under a variety of

378 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 differentmeanings for each equation, and there was no clear winner (in terms of error) for function

- 381 choice.
- 382
- 383

				Parameters	Performance of model		
		Semivariogram function	N	0.11	D	using these specifications	
			Nugget, c <sub>0</sub>	SIII, C <sub>e</sub>	Kange, a <sub>s</sub>	MSE	Avg abs. % error
ſ	Network Distance	Spherical	4.03	5.44	12.81	4.70E+08	63.9%
gior		Exponential	3.12	6.65	5.19	3.97E+08	62.5%
Houston Re		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%

385

 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

391



392 393 394

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.

- 397 This was unexpected, because the network distances contain more information relating to the
- 398 effective separation between points. Contrary to what may be expected, taking the interstate
- 399 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,
- 401 whereas kriging assumes homoskedastic error terms] and may carry a lot of through traffic
- 402 [offering less correlated information for sharing with neighboring sites].)
- 403

Table 5 shows the range of results which can be seen in the different data sets. The median

- 405 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 regressionsexhibit the highest absolute average errors, though the latter offers the densest point structure.
- 407 exhibit the highest absolute average errors, though the latter offers the densest point structure.
  408 However, urban regions and facilities can be quite a bit more complex in their flow variation
- 409 over space (thanks to land use diversity and network complexity). Lower percentage errors were
- 410 associated with sites of higher count and higher count density, thanks, presumably, to a higher
- 411 correlation between traffic flow and lane numbers (capacity) on larger facilities and the benefits
- 412 of more information from nearby points.
- 413

		Austin	Houston	Houston	Minor		Urban
		Region	Region	Region, No IS	Roads	Interstates	Roads
k	MSE	2.62E+08	3.97E+08	2.66E+08	5.97E+07	2.84E+08	3.68E+08
vor	Avg abs err	53.9%	62.5%	63.7%	60.4%	19.4%	62.4%
etv	Median err	-2.8%	5.6%	5.2%	1.5%	-4.6%	-3.1%
N	Best Model	Spherical	Exponential	Exponential	Exponential	Exponential	Spherical
ıclidean	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%
Ει	Best Model	Spherical	Exponential	Exponential	Exponential	Exponential	Exponential
S	MSE	4.20E+08	7.98E+08	6.79E+08	1.48E+08	1.14E+09	5.37E+08
FGL	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%

<sup>414</sup> 415

416

417

418

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

- 419 the nearest 7 (on average420 data sets analyzed here).
- 421

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

- 426 illustrates error comparisons for the Houston data sets. Unfortunately, all three models have a427 number of severe outliers. However, kriging's improvement is still quite apparent.
- 428

Table 5: Errors in prediction for all six data subsets





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

442 443

(Note: Small points indicate "known" locations.)

444 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-445 446 and Euclidean-based kriging models performed almost identically. Some coefficient values vary 447 a fair bit between the spatial and aspatial models (e.g., those on the number of lanes and the 448 indicator for Rural Major Road), but very little between the two kriging models (with Euclidean 449 and network distances). The kriging methods show a decisive improvement in the amount of 450 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 <sup>2</sup>	= .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

- 454
- 455

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

456 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.

460 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

462 as well as their network-distance counterparts (and are far easier to estimate in practice), it seems

- 463 wise to simply use Euclidean distances.
- 464

### 465 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
- 469 complexity is not warranted in such applications. Universal kriging reduced error by controlling
- 470 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
- 474 Austin application, and, in particular, on the interstate system's data point subset. Errors tended
- 475 to be lower at locations with higher counts and more nearby count locations, though the urban 476 set, which was above average for both, offered substantial count variation and thus was amongst
- set, which was above average for both, offered substantial count variation and thus was amongstthe highest in overall errors.
- 478

479 It was interesting to find that network distances offered little improvement to models' predictive

- 480 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
- 482 additional information provided in network distances, but with TxDOT's current data, this is the

- 483 limit. Given the number of links and sites of interest in large networks, like the ones used here,
- 484 calculation of shortest path distances is probably unwarranted (especially since it can be very485 computationally intensive and requires additional information on network structures).
- 486
- 486 487 The issue with non-PSD covariance matrices makes network distances even less compelling. To
- 488 combat or work around this problem, some solutions have been suggested, including spatial
- 489 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.
- 491 492
- 492493 As a way of exploiting spatial information (while capitalizing on local attributes), universal
- 494 kriging is worthy of application in a variety of transportation and other contexts. Opportunities495 to improve upon universal kriging, to better reflect heteroskedasticity in response variability,
- 496 would be useful. Though the implementation here included the Cressie-Hawkins method and
- 497 FGLS (as opposed to only OLS), estimation of the covariance matrix,  $C_{dd}$ , still relies a constant-
- 498 variance assumption. An interesting alternative to be considered, for future work, involves
- 499 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
- 502 correlations, with variance estimates coming from the FGLS procedure's estimates (as used in
- 503 this work). A full (non-diagonal) variance-covariance matrix that does not presume
- 504 homoskedasticity and does allow for interesting spatial (and other) effects could then be applied,
- 505 providing more information in the spatial interpolation process for AADT estimation and for 506 myriad other spatial application contexts that analysts across a variety of disciplines may be
- 506 myrad other spatial application contexts that analysis across a variety of disciplines may be 507 studying (e.g., pavement quality, population densities, home values, vegetative cover, soil 508 quality and water chemistry). Spatial data surround us, and more spacifications (and data
- quality, and water chemistry). Spatial data surround us, and more specifications (and datacontexts) should be evaluated.
- 510

## 511 ACKNOWLEDGEMENTS

512

513 The authors appreciate the technical assistance provided by Dr. Jianming Ma of the Texas

- 514 Department of Transportation and the help and support of Annette Perrone at the University of 515 Texas, Austin. The National Science Foundation Award SES-0818066 provided the support for 516 this research.
- 517

# 518 **REFERENCES**

- 519
- 520 AASHTO Guidelines for Traffic Data Programs. 1992 AASHTO, Washington, DC.
- 521 Collins, S. 1991. *Prediction Techniques for Box-Cox Regression Models*. Washington, DC:
   522 Board of Governors of the Federal Reserve System.
- 523 Cressie, N. 1993. *Statistics for Spatial Data*. New York: Wiley Interscience.
- 524 Cressie, N. and Johannesson, G. 2008. Fixed-rank kriging for very large spatial data sets. *Journal* 525 *of the Royal Statistical Society B*, 70(1): 209-226.
- 526 Curriero, F. 2006. On the Use of Non-Euclidean Distance Measures in Geostatistics1.
- 527 *Mathematical Geology*, 28(8): 907-926.

- Draper, N. R., and Cox, D. R. 1969. On Distributions and Their Transformation to Normality.
   *Journal of the Royal Statistical Society B*, 31:472–476.
- Eom, J.K., Park, M.S., Heo, T.Y. and Huntsinger, L.F. 2006. Improving the Prediction of Annual
   Average Daily Traffic for Nonfreeway Facilities by Applying a Spatial Statistical Method.
   *Artificial Intelligence and Advanced Computing Applications* 1968: 20-29.
- FHWA. 2005. Highway Performance Management System Field Manual Federal Highway
  Administration, Washington, DC.

535 <u>http://www.fhwa.dot.gov/policy/ohpi/hpms/fieldmanual/</u>.

- Gadda, S., Kockelman, K., and Maggon, A. 2007. Estimates of AADT: Quantifying the
  Uncertainty. Presented at the June 2007 *World Conference on Transportation Research*, *Berkeley, California*. Accessed May 20, 2010.
  (www.ce.utexas.edu/prof/kockelman/public html/TRB07AADTUncertainty.pdf.)
- Hu, P and Reuscher, T, 2004. Summary of Travel Trends: 2001 National Household Travel
  Survey. Washington, DC: Federal Highway Administration. Accessed November 2010.
  (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.
- Tang, Y.F. Lam, W.H.K. and Ng, P.L.P. 2003. Comparison of Four Modeling Techniques for
  Short-Term AADT Forecasting in Hong Kong. *Journal of Transportation Engineering* 129
  (3): 271-277.
- Thomas, C. and Featherstone, W. E. 2005. Validation of Vincenty's Formulas for the Geodesic
  Using a New Fourth-Order Extension of Kivioja's Formula. *Journal of Surveying Engineering* 131 (1): 20-26.
- Ver Hoef, J., Peterson, E, and Theobald. D., 2006. Spatial Statistical Models that Use Flow and
   Stream Distance. *Environmental and Ecological Statistics*. 13(4): 449-464.
- Wang, X and Kockelman, K. 2009. Forecasting Network Data: Spatial Interpolation of Traffic
   Counts Using Texas Data. *Transportation Research Record*, 2105: 100-108.
- Wang, H and Ranalli, M.G. 2007. *Low-Rank Smoothing Splines on Complicated Domains*.
  Biometrics 63:209–217.
- Webster, R. and Oliver, M. 2001. *Geostatistics for Environmental Scientists*. New York: John
  Wiley and Sons.
- Zhao, F. and Chung, S. 2001. Contributing Factors of Annual Average Daily Traffic in A Florida
   County Exploration with Geographic Information System and Regression Models.
   *Pavement Management, Monitoring, and Accelerated Testing* 1769: 113-122.
- 562 Zhao, F. and Park, N. 2004. Using Geographically Weighted Regression Models to Estimate
- Annual Average Daily Traffic. *Transportation Research Record: Journal of the Transportation Research Board* 1879: 99-107.