Relaxing the Multivariate Normality Assumption in the Simulation of Transportation System Dependencies

ManWo Ng (corresponding author)
Ph.D. Candidate
The University of Texas at Austin
Department of Civil, Architectural, and Environmental Engineering
1 University Station C1761
Austin, TX 78712-1076
Email: MWNg@mail.utexas.edu
Ph: +1 512 906 7960
Fax: +1 512 475 8744

Kara M. Kockelman
Professor and William J. Murray Jr. Fellow
The University of Texas at Austin
Department of Civil, Architectural, and Environmental Engineering
6.9 E. Cockrell Jr. Hall
Austin, TX 78712-1076
Email: kkockelm@mail.utexas.edu
Ph: +1 512 471 0210
Fax: +1 512 475 8744

S. Travis Waller
Associate Professor
The University of Texas at Austin
Department of Civil, Architectural, and Environmental Engineering
1 University Station C1761
Austin, TX 78712-1076
Email: stw@mail.utexas.edu
Ph: +1 512 471 4539
Fax: + 512 475 8744

Committee: ABJ80, Statistical Methodology and Statistical Computer Software in Transportation Research

The following paper is a pre-print and the final publication can be found in *Transportation Letters: The International Journal of Transportation Research*, Vol. 2 (2):63-74, April 2010

ABSTRACT

By far the most popular method to account for dependencies in the transportation network analysis literature is the use of the multivariate normal (MVN) distribution. While in certain cases there is some theoretical underpinning for the MVN assumption, in others there is none. This can lead to misleading results: results do not only depend on *whether* dependence is modeled, but also *how* dependence is modeled. When assuming the MVN distribution, one is limiting oneself to a specific set of dependency structures, which can substantially limit validity of results. In this paper a more flexible, correlation-based approach (where just marginal distributions and their correlations are specified) is proposed, and it is demonstrated that, in simulation studies, such an approach is a generalization of the MVN assumption. The need for such generalization is particularly critical in the transportation network modeling literature, where oftentimes there exists no or insufficient data to estimate probability distributions, so that sensitivity analyses assuming different dependence structures could be extremely valuable. However, the proposed method has its own drawbacks. For example, it is again not able to exhaust all possible dependence forms and it relies on some not-so-known properties of the correlation coefficient.

1. INTRODUCTION

The prevailing assumption in transportation network modeling has been that parameters are deterministically known (e.g.. Abdulaal and LeBlanc, 1979; Suwansirikul et al., 1987). More recent research recognizes that uncertainty is a critical consideration (e.g., Peeta and Ziliaskopoulos, 2001; Waller et al., 2001; Ng and Waller, 2009a, b). However, while this new stream of publications relaxed the assumption of determinism, it imposed a new assumption, namely, that of statistical independence (Siu and Lo, 2008; Ng and Waller, 2009c). In certain cases, the independence assumption can be justified (e.g., Lo and Tung, 2003; Ng and Waller, 2009c); in other cases such an assumption is questionable, if not, clearly unreasonable.

The ideal method for modeling dependencies is to define a joint probability distribution to characterize the joint behavior of the random elements under consideration. However, specification and estimation of such a joint distribution can be a formidable task, especially as the number of random elements increases. Moreover, specialized algorithms are needed to sample from these case-specific multivariate distributions in simulation studies (see, e.g., Ghosh and Henderson, 2002). By assuming that the joint probability distribution comes from a particular parametric family of multivariate distributions, the above difficulties can be overcome.

In the transportation literature, the assumption of multivariate normality[1] is clearly the most popular. For example, Zhao and Kockelman (2002) investigated the propagation of uncertainty in the classical four-step travel demand model using a multivariate normal (MVN) distribution to describe demographic inputs. Pradhan and Kockelman (2002) and Krishnamurthy and Kockelman (2003) performed similar analyses, integrating the added uncertainty emerging from the application of land-use models. In Clark and Watling (2005) the MVN distribution was used to model the joint behavior of link flows in a transportation network with uncertain demand in the context of travel time reliability assessment (Hazelton, 2000). Subsequent papers on traffic assignment and network design adopted the same assumption (e.g., Sumalee et al., 2006; Lam et al., 2008). The MVN distribution has also proven popular for trip table estimation problems (e.g., Maher, 1983; Hazelton, 2000; Lo and Chan, 2003). More recently, Duthie et al. (2009) used the MVN distribution to model correlated travel demand between origin-destination pairs in a network. They found that neglecting correlation in demand can lead to misleading predictions of system performance and, hence, suboptimal network improvement decisions. Siu and Lo (2008) also intimated the use of a MVN distribution to model travel demand in their reliability-based network equilibrium models. In Watling (2006) the joint behavior of the link travel times was assumed to follow a MVN distribution. Finally, the MVN assumption is fundamental to the multinomial probit model's specification (Daganzo, 1979).

While in certain cases there is some theoretical motivation for the MVN assumption, in many cases there is none. The MVN distribution is then simply assumed for reasons such as mathematical tractability or the availability of simple and efficient sampling algorithms. However, this can lead to misleading results, since results do not only depend on whether dependence is modeled, but also how dependence is modeled (Livny et al., 1993). When assuming a MVN distribution, one is limiting oneself to a relatively narrow set of dependence structures (see Section 2), which may limit the validity of the results.

The modeling of dependence via the specification of marginal distributions (or "marginals", for short) and their correlations is very rare in the transportation literature (Chen et al., 2002, 2007). In contrast, such an approach has been standard in fields like finance (e.g., see Embrechts et al., 2002 and the references therein) and risk analysis (e.g. Ferson and Burgman, 1995 and the references therein). This paper proposes such a "correlation-based" method as an alternative to the MVN assumption. The paper also demonstrates how the correlation-based approach is a natural generalization of the MVN assumption. Consequently, the method is able to represent a much wider range of dependence structures than the MVN distribution. This is particularly useful in settings like transportation network modeling, where often there is no or insufficient data to ascertain the most appropriate probability distributions (Ng et al., 2009d). It seems that only Chen et al. (2002, 2007) have adopted the correlation-based approach to model dependencies in the transportation literature. However, they presented no unified framework for further applications, and no mention was made of limitations and potential complexities that can arise. To address this gap in the literature, this paper presents a unified framework and discusses limitations and complexities of the correlation-based approach in detail.

[1] For a detailed discussion of the MVN and its properties, please see Rencher (2002).

The remainder of this paper is organized as follows. Section 2 briefly reviews some properties of the correlation coefficient and the MVN distribution. Section 3 introduces the correlation-based approach and illustrates its flexibility in representing multifarious dependence structures. As the correlation-based approach is not without its own limitations, Section 4 describes potential complexities. Concluding remarks are provided in Section 5.

2. PRELIMINARIES

The parameter that plays a fundamental role here is the correlation coefficient (also known as the Pearson correlation coefficient, product moment correlation or linear correlation parameter). Let X_i and X_j be two random variables with finite expected values (i.e., $E(X_i) < \infty, E(X_j) < \infty$) and finite variances (i.e., $Var(X_i) \equiv \sigma_i^2 < \infty$, $Var(X_j) \equiv \sigma_j^2 < \infty$). Then their correlation $\rho(X_i, X_j)$ can be defined as:

$$\rho(X_i, X_j) = \frac{E(X_i X_j) - E(X_i)E(X_j)}{\sigma_i \sigma_j}.$$

Key properties include the fact that ρ must lie between -1 and 1, and, if $|\rho(X_i, X_j)|=1$, then with probability one, $X_j = aX_i + b$ for some real numbers *a* and *b* (with $a > 0$ if $\rho(X_i, X_j) = 1$ and $a < 0$ if $\rho(X_i, X_j) = -1$). As evident, the correlation coefficient can be interpreted as a measure of *linear* dependence between two random variables. Other well-known properties include that the correlation matrix Σ_X, where $(\Sigma_X)_{ij} = \rho(X_i, X_j)$, is symmetric positive semidefinite, with unit diagonals (here $(A)_{ij}$ is used to denote the value in the *i*-th row and *j*-th column of matrix *A*). For instance, travel time on link *i* cannot be positively correlated with those on links *j* and *k*, while the travel times on links *j* and *k* are negatively correlated. Furthermore, zero correlation does not imply independence (while the converse is true), and the correlation coefficient is not invariant under nonlinear monotonic transformations.

This paper proposes a method to generalize the MVN assumption *in simulation studies* (our emphasis on simulation studies will become clear in Section 3). Figure 1 shows typical ellipsoidal scatter plots resulting from a bivariate normal distribution for correlation coefficients ranging from $\rho(X_i, X_j) = -0.8$ (top left) to $\rho(X_i, X_j) = 0.8$ (bottom right) in increments of 0.2.

Figure 1's plots rely on the most popular normal random vector generation algorithm (as described in Scheuer and Stoller, 1962): To simulate a MVN random vector $Z \sim N(\mu, \Sigma_Z)$ with mean μ and correlation matrix Σ_Z, one simply evaluates $Z = CY$, where C is a lower triangular matrix such that $\Sigma_Z = CC^T$ (where C^T denotes the transpose of matrix *C*) and *Y* a vector of independent and identically distributed standard normal random values.

More interesting dependence structures, such as those shown in Figure 2, cannot be captured by the MVN distribution. In the upper figure, the random variable on the *y*-axis is bounded (whereas a normal random variable will be unbounded). In the lower figure, the dependence in the left-tail is much stronger than in the right-tail. MVN distributions cannot capture such asymmetry.

Before proceeding to the next section, it is important to note that one may argue (Duthie et al., 2009) that one can vary the correlation coefficient to investigate the "*entire*" range of possible dependence structures. This type of sensitivity analysis is particularly popular when there is insufficient empirical data to gain insights into the true dependence structure. However, from Figures 1 and 2 it is clear that varying the correlation coefficient in a MVN distribution will not exhaust all conceivable dependence structures. Such limitations can lead to suboptimal decisions (Livny et al., 1993).

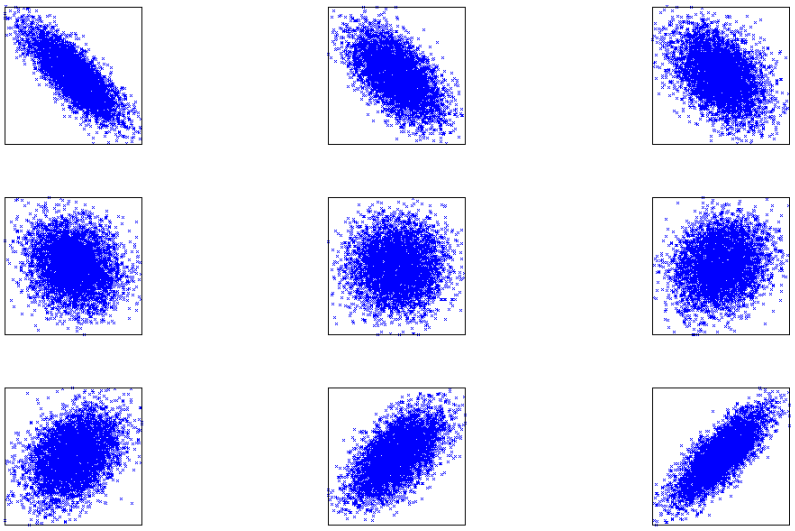

Figure 1: Typical bivariate ellipsoidal scatter plots of the MVN distribution for correlation coefficients ranging from -0.8 (top left figure) to 0.8 (bottom right figure), in increments of 0.2.

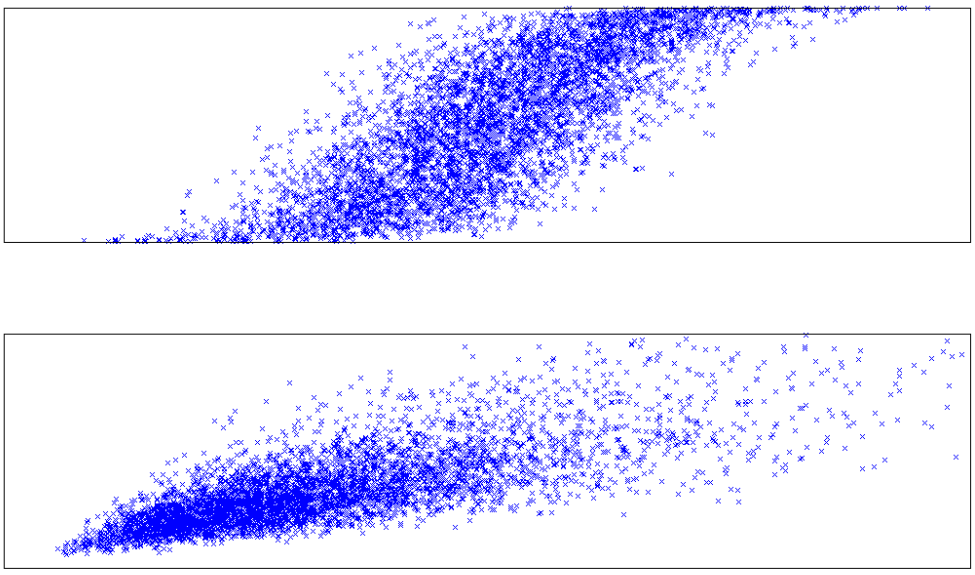

Figure 2: Examples of dependence structures that MVN distributions cannot capture.

3. A CORRELATION-BASED APPROACH

This section introduces a dependence modeling method in which the user only specifies the marginal distributions $F_i(x_i)$ and their correlations $\rho(X_i, X_j)$. This approach has been used in other fields (e.g., Li and Hammond, 1975; Whitt, 1976; Ferson and Burgman, 1995; Wang and Dhaene, 1998; Embrechts et al., 2002) – fields where there is considerably more experience with modeling stochasticity, risk, and dependence than in the transportation arena.

Clearly, such an approach is less restrictive that the MVN assumption: to specify a MVN distribution, one has to assume normal marginals, a set of correlation coefficients (the only two ingredients in a correlation-based approach) *and* the assumption that the joint distribution is MVN. That is, while the converse is true, *normal marginal distributions do not necessarily imply that the joint distribution is MVN*. In addition, the estimation of the individual marginal distributions is generally an easier task than the estimation of the entire multivariate probability density function. Indeed, it is easy to see that if one is able to estimate a MVN distribution, then one is able to estimate the essential elements in a correlation-based approach. Of course, as shown in the next section, the correlation-based is not perfect either. However, it is demonstrated next that in simulation studies it represents a true improvement of the MVN assumption, in the sense that the MVN assumption is a special case of the proposed method.

The correlation-based approach is best introduced by examining the perhaps most popular algorithm to generate random vectors with prespecified marginals and correlation structure. This algorithm is known as the NORTA (i.e., NOrmal-To-Anything) algorithm developed by Cario and Nelson (1997). The NORTA algorithm can be summarized as follows.

Algorithm NORTA

Input: Desired marginal cumulative distribution functions $F_i(x_i)$ and their correlations $\rho(X_i, X_j)$

Output: Random vector X with marginals $F_i(x_i)$ and correlation matrix Σ_X

Step 1 Generate a normal random vector $Z \equiv [Z_1, Z_2, ..., Z_N]^T \sim N(0, \Sigma_Z)$ with $(\Sigma_Z)_{ii} = 1$ such that

$$\rho\left(F_i^{-1}(\phi(Z_i)), F_j^{-1}(\phi(Z_j))\right) = \rho(X_i, X_j)$$

Step 2 Evaluate

$$X = \begin{bmatrix} F_1^{-1}(\phi(Z_1)) \\ F_2^{-1}(\phi(Z_2)) \\ \vdots \\ F_N^{-1}(\phi(Z_N)) \end{bmatrix}$$

where $\phi(x)$ and $F_i^{-1}(u) \equiv \inf\{x : F_i(x) \geq u\}$ denote the cumulative distribution function of a standard normal random variable and the generalized inverse of $F_i(x_i)$, respectively.

The reason for the name NORTA is clear: the algorithm starts with a MVN random vector and transforms it to a random vector with "any" (see Section 4) desired marginal distributions and correlation structure. To see this, recall that $\phi(Z_i)$ has a uniform distribution on [0,1], so that $F_i^{-1}(\phi(Z_i))$ has the desired marginal distribution $F_i(x_i)$, e.g., see Casella and Berger (2001). The crux of the algorithm lies in Step 1, where a suitably chosen correlation matrix Σ_Z ensures that *X* has the desired correlation structure. Note that if the correlation coefficient *were* invariant under nonlinear transformations, Step 1 would be trivial. In particular, to find Σ_Z the following equation needs to be solved:

$$\sigma_i \sigma_j \rho(X_i, X_j) + E(X_i)E(X_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_i^{-1}(\phi(z_i))F_j^{-1}(\phi(z_j))\varphi_{(\Sigma_Z)_{ij}}(z_i, z_j)dz_i dz_j \qquad (1)$$

where $\varphi_{(\Sigma_Z)_{ij}}$ denotes the bivariate normal density function. Note that this equation is an equation in one unknown, namely $(\Sigma_Z)_{ij}$. By definition, $(\Sigma_Z)_{ii} = 1$ and since Σ_Z is symmetric, Step 1 amounts to the solution of $N(N-1)/2$ single-variable equations. Fortunately, it turns out that (1) behaves nicely: the right-hand side of (1) is continuous and non-decreasing as a function of $(\Sigma_Z)_{ij}$ under very mild conditions (for details, see Cario and Nelson, 1997). Consequently, efficient solution algorithms exist (Press et al., 2007). Finally, note that in Step 1 of the algorithm the vector *Z* is typically constructed via Scheuer and Stoller's algorithm.

As indicated at the beginning of this section, normal marginals do not necessarily imply a MVN distribution. That is, it is not trivial that adopting normal marginals will result in a MVN distribution as NORTA's output. Proposition 1

below demonstrates that normal marginals *in conjunction with* NORTA *does* generate MVN data, in the sense that the output could have been obtained from Scheuer and Stoller's algorithm.

Proposition 1: When the desired marginal distributions $F_i(x_i)$ in NORTA are normal, then NORTA is equivalent to Scheuer and Stoller's algorithm.

Proof: Recall that in NORTA

$$X = \begin{bmatrix} F_1^{-1}(\phi(Z_1)) \\ F_2^{-1}(\phi(Z_2)) \\ \vdots \\ F_N^{-1}(\phi(Z_N)) \end{bmatrix}$$

where $Z = CY$ and C is a lower triangular matrix such that $\Sigma_Z = CC^T$ and Y is a vector of independent and identically distributed standard normal random variables. Next it is shown that one could have obtained vector X using Scheuer and Stoller's algorithm with a particular choice of the lower triangular matrix C. To see this, note that $F_i^{-1}(u) = \mu_i + \sigma_i \phi^{-1}(u)$ in case of a normal distribution with mean μ_i and standard deviation σ_i. Hence, one can rewrite X as follows:

$$X = \begin{bmatrix} \mu_1 + \sigma_1 Z_1 \\ \mu_2 + \sigma_2 Z_2 \\ \vdots \\ \mu_N + \sigma_N Z_N \end{bmatrix} = \begin{bmatrix} \mu_1 + \sigma_1 C_{1.} Y \\ \mu_2 + \sigma_2 C_{2.} Y \\ \vdots \\ \mu_N + \sigma_N C_{N.} Y \end{bmatrix}$$

where $C_{i.}$ denotes the i-th row of the matrix C. From this last equation it is clear that X could have been obtained directly from Scheuer and Stoller's algorithm using a lower triangular matrix with rows equal to $\sigma_i C_{i.}$. That is, let $\overline{C}$ denote the lower triangular matrix with rows $\sigma_i C_{i.}$, then $X = \mu + \overline{C}Y$ and $\Sigma_X = \overline{C}\overline{C}^T$. Q.E.D.

From Proposition 1 one can make a subtle but important observation. As noted earlier, the MVN assumption imposes one more restriction than the proposed correlation-based approach: In addition to the marginal distributions (that are assumed to be normal) and a correlation matrix, the MVN assumption implies that the joint behavior of the marginals is MVN. In light of Proposition 1, i.e., if using NORTA, it is clear that the correlation-based approach also implicitly assumes some joint distribution for the marginals. Indeed, as will be seen in the next section, this is necessary since the specification of marginal distributions and correlation coefficients *alone* does not uniquely determine a joint distribution. Hence in some sense the correlation-based approach requires as many assumptions as the MVN distribution. The only *big* difference is that the correlation-based approach allows one to relax the assumption of normal marginal distributions. That is, *in simulation studies*, there is no reason at all why the correlation-based approach is not preferred over the MVN assumption! However, as can be expected, no single method is perfect. Clearly, the correlation-based approach is computationally more intensive than Scheuer and Stoller's algorithm (in particular in Step 1 of NORTA). Other potential difficulties and limitations exist and they will be discussed in the next section.

To illustrate the versatility of the correlation-based approach, consider Figure 3 that shows four more (Figure 2 has also been generated using NORTA) dependence structures that the MVN distribution is not able to capture. The top left figure depicts some nonlinear increasing trend (the correlation coefficient in this case equals 0.63) where the marginal distributions are distributed according to an extreme value distribution with location parameter 5 and scale parameter 0.1 and an exponential distribution with mean 1. The top right figure shows a dependence relationship where dependence is stronger in the right tail than in the left. The marginals underlying this plot were Weibull (with location parameter 3 and scale parameter 5) and extreme value (with location parameter 10 and scale parameter 1) with a correlation coefficient of 0.78. In the lower left figure another non-standard dependence structure is depicted with negative correlation (-0.65) and normal (with mean 10 and variance 1) and beta (with parameters 8 and 1) marginal distributions. Thus far it might seem that the marginal distributions in NORTA need to be continuous. In fact, marginals can be discrete in which case one has to interpret (1) slightly differently (for more details, one may

refer to Cario and Nelson, 1997). The lower right figure in Figure 3 depicts a dependence relation with correlation -0.5, where one of the marginals is exponential (with mean 10) and the other uniform *discrete* (with mean 5).

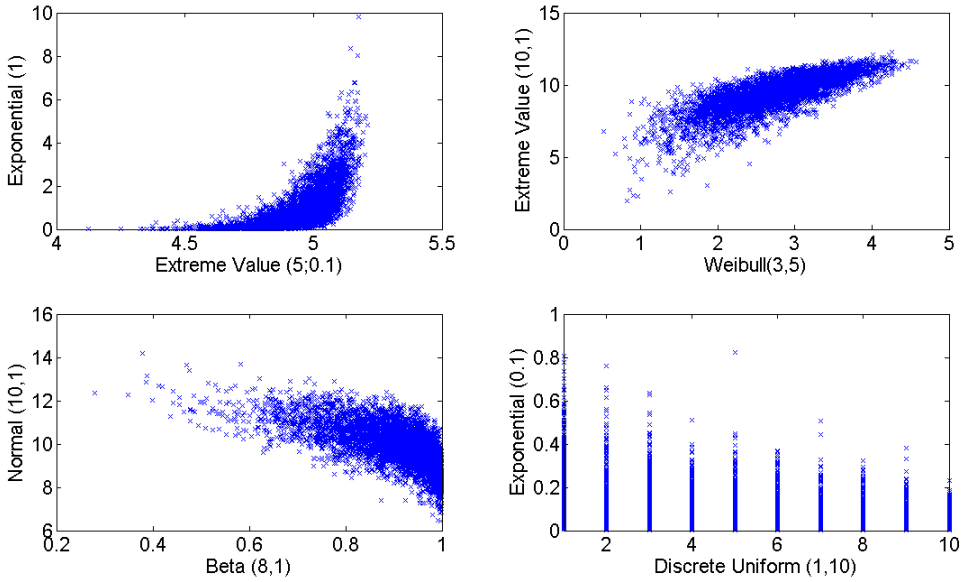

Figure 3: Examples of dependence structures NORTA can capture.

4. SOME LIMITATIONS OF THE CORRELATION-BASED APPROACH

In the previous section it has been demonstrated that the correlation-based approach is able to represent more forms of dependence than the MVN can. However, it turns out that it is also not exhaustive. As an example, suppose that there are two origins generating travel demand. Furthermore, assume that each demand has a normal distribution with mean 10 and standard deviation 1. Figure 4 depicts their dependence structure (using 5000 samples) with "weak" dependence in the right tail and "strong" dependence in the left tail[2]. Since for normal marginals, the correlation-based approach is equivalent to the MVN assumption (see Proposition 1), it is clear that NORTA is unable to represent Figure 4's dependence structure.

The above discussion implies that marginal distributions together with their correlations do not uniquely specify the dependence relationship. Indeed, it can be verified that the sample correlation implicit in Figure 4's bivariate scatter plot is approximately +0.95. Given normal marginal distributions with means of 10 and standard deviations of 1 and a desired correlation of +0.95, NORTA would generate the familiar ellipsoidal scatter plot shown in the upper part in Figure 5. The lower part of Figure 5 repeats Figure 4's scatter plot for ease of comparison. Clearly, while the marginal distributions and correlation coefficient are the same, the dependence structures differ substantially. Therefore, prior to using any pre-coded software that requires as input only marginal distributions and their correlations, it can be critical that users consult the associated documentation in order to ensure that the desired dependence structure has been generated. For example, does the software reproduce Figure 5's lower or upper dependence structure?

[2] Figure 4 was generated using the theory of copulas (e.g. see Srinivas et al., 2006; Bhat and Eluru, 2009; Ng and Waller, 2009e), which is outside the scope of this paper. Here, the purpose is to emphasize the fact that the depicted marginal distributions are normal random variables.

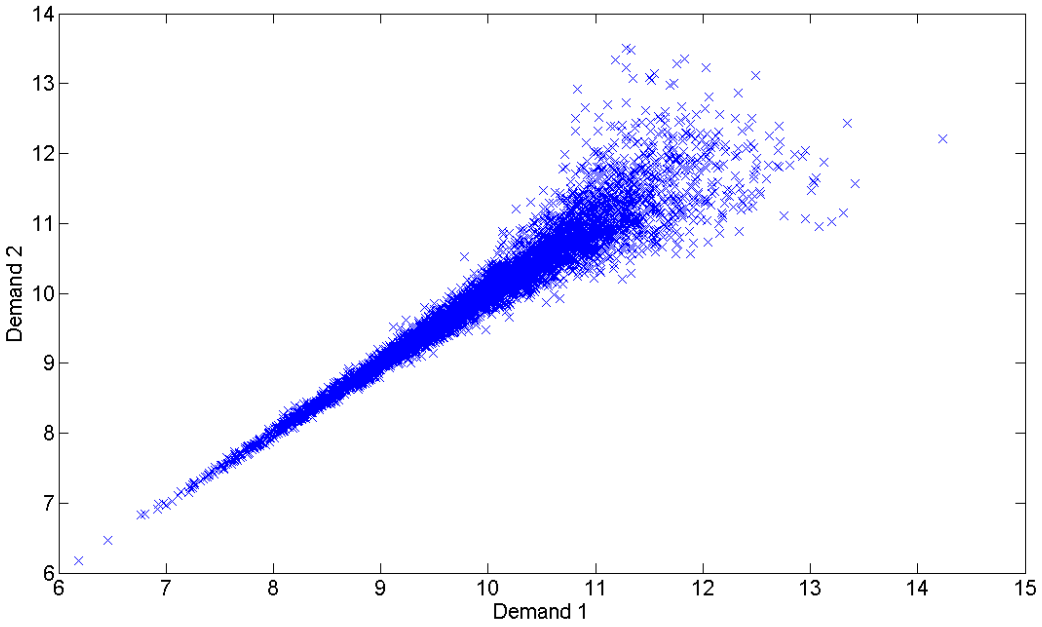

Figure 4: A dependence structure with normal marginals.

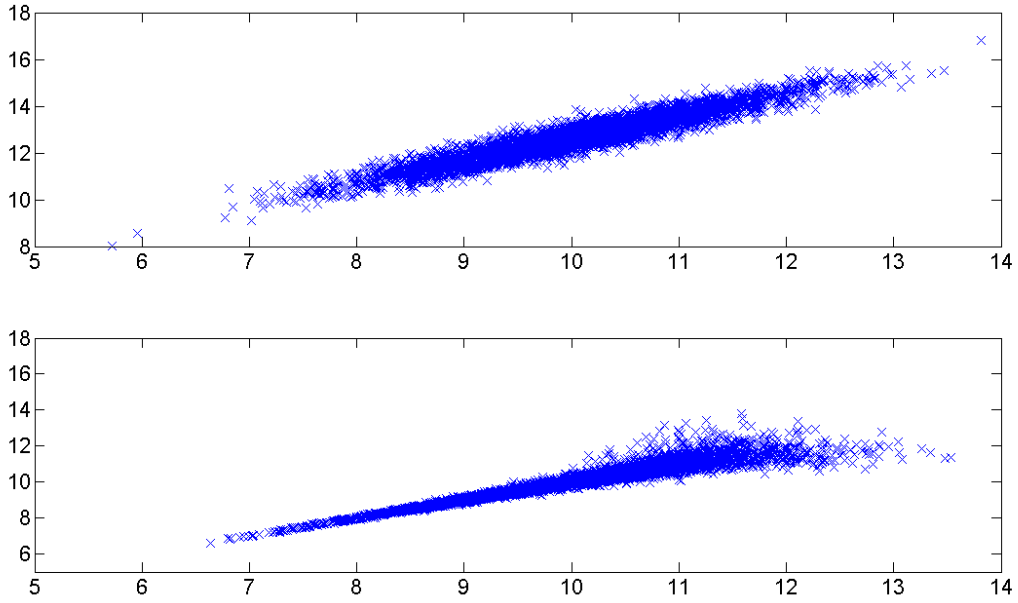

Figure 5: Two bivariate distributions exhibiting the *same* marginal distributions and sample correlation.

Now, consider the following reasoning. Suppose that the travel times T_1 and T_2 on two links are known to be correlated. Moreover, assume that $\log T_1 \sim N(0,1)$ and $\log T_2 \sim N(0,\sigma^2)$; thus, T_1 and T_2 are lognormal random variables. The variance of the sum of these two random variables is given by the following:

$$Var(T_1 + T_2) = Var(T_1) + Var(T_2) + 2\rho(T_1, T_2)\sqrt{Var(T_1)Var(T_2)}$$

where $Var(T_1) = (e-1)e$ and $Var(T_2) = (e^{\sigma^2}-1)e^{\sigma^2}$ (Johnson et al., 1995). Suppose that one is interested in the maximum and minimum variability of the sum of these travel times. Since for a given value of σ^2, the individual variances $Var(T_1)$ and $Var(T_2)$ are fixed, one might argue that the maximum and minimum variances of the sum are obtained when $\rho(T_1, T_2) = 1$ and $\rho(T_1, T_2) = -1$, respectively. In other words:

$$Var(T_1 + T_2)\big|_{\max} = e(e-1) + (e^{\sigma^2}-1)e^{\sigma^2} + 2\sqrt{e(e-1)(e^{\sigma^2}-1)e^{\sigma^2}} \quad (2)$$

$$Var(T_1 + T_2)\big|_{\min} = e(e-1) + (e^{\sigma^2}-1)e^{\sigma^2} - 2\sqrt{e(e-1)(e^{\sigma^2}-1)e^{\sigma^2}} \quad (3)$$

This argument is seemingly valid. However, there is a little known and rather surprising property of the correlation coefficient that explains why the above argument is wrong and even misleading.

Hoeffding (1940) proved the following property: Depending on the marginal distributions of X_i and X_j, it is possible that $\rho_{\min} < \rho(X_i, X_j) < \rho_{\max}$, where $-1 < \rho_{\min} < 0 < \rho_{\max} < 1$. Furthermore, the set of all possible correlations forms a closed interval $[\rho_{\min}, \rho_{\max}]$. That is, the extremal correlations of -1 and +1 may not be achievable. A failure to recognize this fact can give rise to misleading conclusions. Fortunately, this issue is not relevant for all distributions. For example, consider the case of two normal random variables: $X_i \sim N(0,1)$ and $X_j \sim N(\mu, \sigma^2)$. One can write $X_j = \sigma V + \mu$ where V is some standard normal random variable. Clearly, the largest possible correlation $\rho_{\max}(X_i, X_j)$ is obtained when $V = X_i$, which directly yields that $\rho_{\max}(X, Y) = 1$. Likewise, setting $V = -X_i$ gives the smallest possible correlation $\rho_{\min}(X, Y) = -1$ since in this case $X_j = -\sigma X_i + \mu$ with probability one. That is, in case of normal marginals, there is no need to worry about the range of achievable correlations (Duthie et al., 2009). However, for other marginal distributions, care must be exercised, as demonstrated by the following example due to Embrechts et al. (2002).

As in the travel time example above, suppose that the travel times on two links of a transportation network have lognormal distributions, i.e., $\log T_1 \sim N(0,1)$ and $\log T_2 \sim N(0, \sigma^2)$. One can write $T_1 = e^V$ where V is some standard normal random variable. Likewise, it is clear that $T_2 = e^W$ where $W \sim N(0, \sigma^2)$ and $W/\sigma \sim N(0,1)$. Figure 6 depicts two instances of the lognormal densities ($\mu = 0, \sigma = 1$ and $\mu = 0, \sigma = 5$). Nothing seems to suggest that the correlation between these lognormal travel times cannot be any arbitrary value. Next it is shown that the only theoretically consistent correlation is zero!

Clearly, the maximum correlation that can be induced between T_1 and T_2 is when $W/\sigma = V$. Therefore, using the fact that $E(U) = e^{\mu + \sigma^2/2}$ and $Var(U) = (e^{\sigma^2}-1)e^{2\mu+\sigma^2}$ for a lognormal random variable U with parameters μ and σ^2, one can write

$$\rho_{\max}(T_1, T_2) = \rho(e^V, e^{\sigma V}) = \frac{Ee^{(1+\sigma)V} - Ee^V Ee^{\sigma V}}{\sqrt{Var(e^V)Var(e^{\sigma V})}} = \frac{e^{(1+\sigma)^2/2} - e^{1/2}e^{\sigma^2/2}}{\sqrt{e(e-1)e^{\sigma^2}(e^{\sigma^2}-1)}} = \frac{e^{\sigma} - 1}{\sqrt{(e-1)(e^{\sigma^2}-1)}} \quad (4)$$

On the other hand, the minimum correlation that can be induced between T_1 and T_2 occurs when $W/\sigma = -V$. Hence,

$$\rho_{\min}(T_1, T_2) = \rho(e^V, e^{-\sigma V}) = \frac{Ee^{(1-\sigma)V} - Ee^V Ee^{-\sigma V}}{\sqrt{Var(e^V)Var(e^{-\sigma V})}} = \frac{e^{(1-\sigma)^2/2} - e^{1/2}e^{\sigma^2/2}}{\sqrt{e(e-1)e^{\sigma^2}(e^{\sigma^2}-1)}} = \frac{e^{-\sigma} - 1}{\sqrt{(e-1)(e^{\sigma^2}-1)}} \quad (5)$$

Figure 7 shows the extremal correlations (4) and (5) as a function of σ. It is readily seen that the range of possible correlations diminishes very rapidly as σ grows. If $\sigma = 5$, then $\rho_{\min} \approx \rho_{\max} \approx 0$! In the travel time example above, it was assumed that the correlation coefficient could achieve the values -1 and 1. From Figure 7 it is clear that perfect negative correlation can *never* be achieved, whereas perfect positive correlation is only (approximately) achieved for a single value of σ. In other words, the assumption that perfect positive and negative correlation can be

achieved could lead to overestimates of the maximum variance (i.e., one unnecessarily believes that the variance is large) and, more importantly, to underestimates (up to 45%) of the minimum variance (i.e., one will falsely believe that the variance is small). Figure 8 depicts the theoretical lower and upper bounds (4) and (5), together with the "incorrect" bounds (2) and (3). Note that, strictly speaking, the "incorrect" bounds are valid bounds since they enclose the entire theoretically feasible region. However, as explained, they provide misleading information.

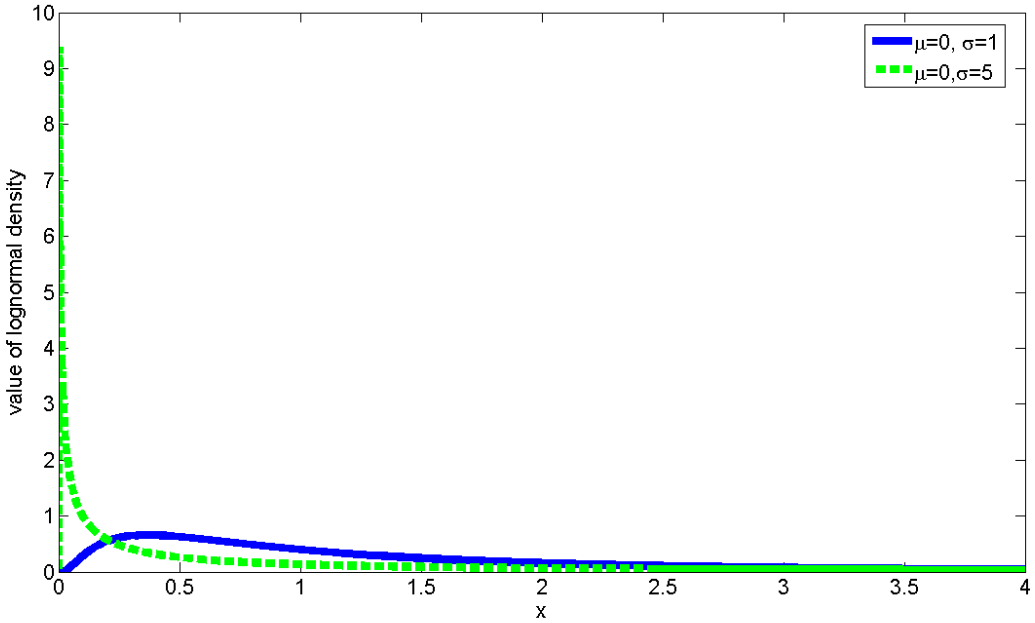

Figure 6: Lognormal densities with parameters $\mu = 0, \sigma = 1$ and $\mu = 0, \sigma = 5$.

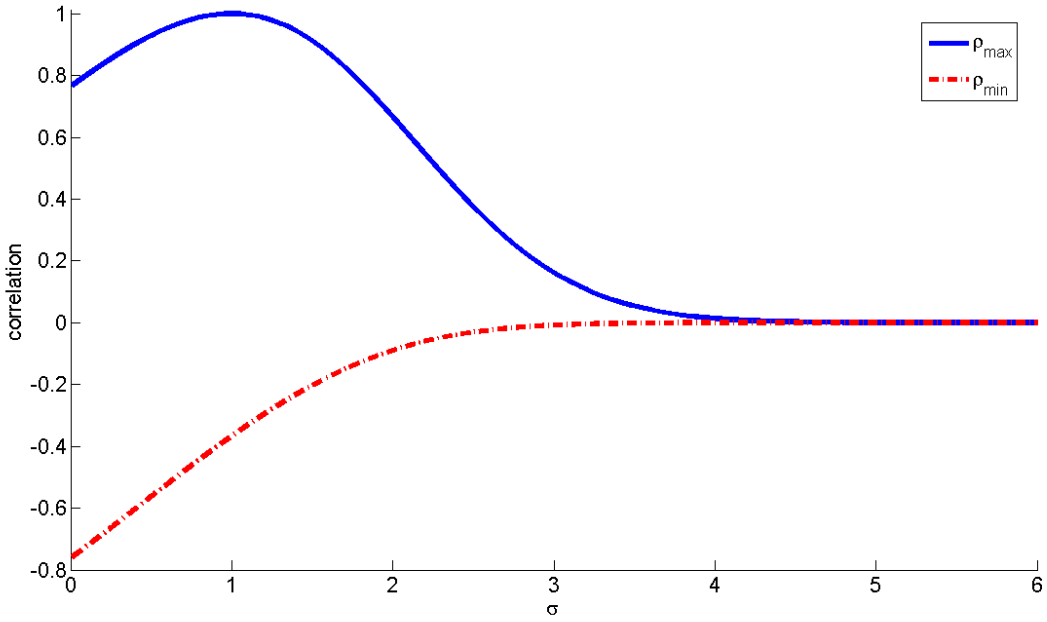

Figure 7: Upper and lower bounds on the correlation coefficient between two lognormally distributed random variables.

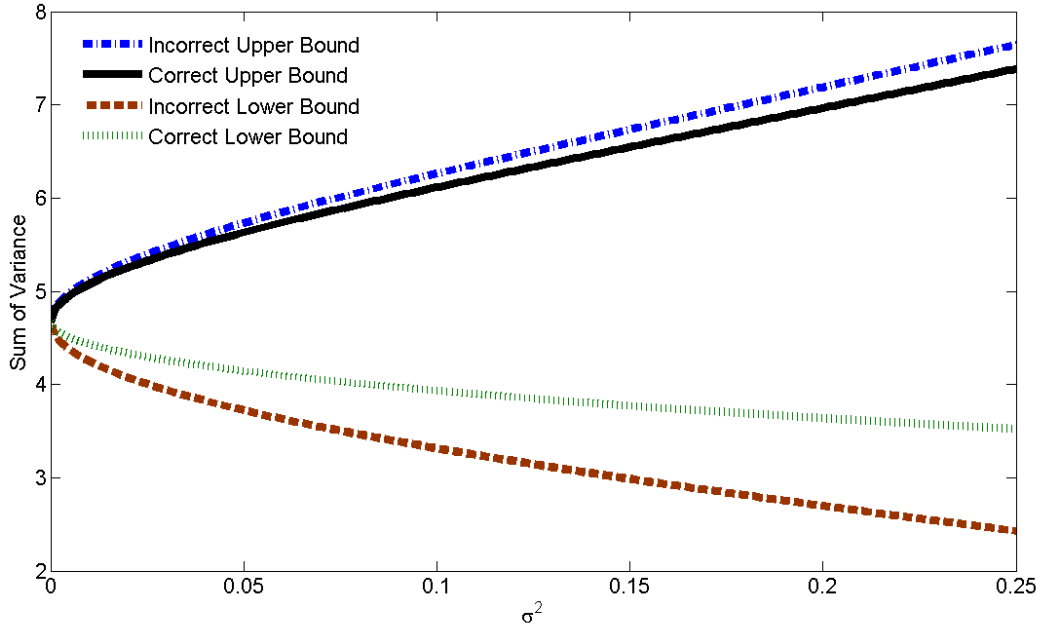

Figure 8: Theoretical and "incorrect" bounds for the travel time example.

It is clear that in order to avoid misleading conclusions, it is necessary to ensure that all (theoretically) assigned correlations lie within the bounds imposed by the marginal distributions under consideration. In other words, they must be feasible. This issue is particularly relevant when real-life data are not available to support the selection of feasible correlation coefficients (Duthie et al., 2009). However, a complicating factor is that it might not be tractable to theoretically derive this feasible range for arbitrary marginal distributions, although complex computational procedures have been developed to determine the feasibility of correlation matrices (*once* the correlation matrices are specified), as discussed in Ghosh and Henderson (2002). To further complicate matters, Ghosh and Henderson (2002) formally demonstrated that NORTA fails to generate random vectors with certain feasible correlation matrices. In these cases, the algorithm can only generate a random vector with approximately the given correlation structure.

5. CONCLUSIONS

The modeling of dependencies is becoming recognized as fundamental to addressing important questions in transportation network analysis, including traffic forecasting, reliability assessment, investment decision-making, and transportation planning. Specification of MVN distributions is the primary method for accounting for parameter and input dependencies, though in many cases there may be no empirical or theoretical motivation for such assumptions.

Using the NORTA algorithm, this work demonstrates how a correlation-based approach (with marginal distributions also specified) offers a generalization of the MVN assumption regularly used in simulation studies. It has been shown that the correlation-based approach is able to represent more dependence structures than the MVN can. Such flexibility may be critical in the transportation and urban systems modeling arenas, where interactions are regularly complex, heteroskedasticity, non-linearities and a variety of other behaviors can emerge – yet there is little data to empirically illuminate the multivariate nature of various relationships. As in any complex science, sensitivity analyses assuming different dependence structures could be extremely useful, and result in more robust network design, policy making, and operations management decisions.

Unfortunately, the correlation-based approach is not perfect either. In particular, examples were provided to demonstrate that:

- The correlation-based approach is not able to capture all possible forms of dependence. That is, one effective discards a whole set of possible dependence structures in any type of sensitivity analysis where the analyst varies the dependence structure. Nevertheless, it represents a true improvement over the MVN assumption.

- Marginal distributions and their correlation coefficients do not uniquely determine the dependence structure. This issue is particularly relevant when using pre-coded random vector generation algorithms in which case it is critical to consult the associated documentation before its use. In certain cases, a visual inspection of the resulting scatter plots might also help to ensure that the desired correlation structure has been generated.

- The range of values a correlation coefficient can assume *depends* on the marginal distributions involved. This characteristic of the correlation coefficient is perhaps the least known of all of its properties, even in fields where the correlation coefficient has been extensively used in modeling dependence. This property is particularly relevant to the transportation network modeling community where oftentimes real data are not available to support the selection of appropriate correlations and distributions. An example with lognormal random variables was provided in which the only theoretically feasible correlation was a singleton. That is, any other (assumed) correlation would provide misleading results. To complicate matters further, the verification of the feasibility of the (theoretically assigned) correlation coefficients can be challenging. Finally, NORTA is not able to generate all *theoretically feasible* correlated random vectors. In such cases, it only generates random vectors with approximately the given correlation structure.

Despite these undesirable properties, the correlation-based approach is a true generalization of the MVN assumption. It seems there is little reason why the correlation-based approach should not become a more widespread simulation tool in the transportation community.

References

Abdulaal, M., L.J. LeBlanc. 1979. Continuous Equilibrium Network Design Models. *Transportation Research B*, Vol. 13, pp. 19–32.

Bhat, C.R., Eluru, N., 2009. A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling. *Transporation Research Part B*, 43(7), pp. 749-765.

Biller, B., Ghosh, S., 2006. Multivariate Input Processes, *Handbooks in Operations Research and Management Science*, Vol.13, pp. 123-154.

Cario, M. C., Nelson, B. L., 1997. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Casella, G., R. Berger. 2001. Statistical Inference, 2nd ed. Pacific Grove, CA: Wadsworth.

Clark, S. D., Watling, D. P., 2005. Modeling network travel time reliability under stochastic demand. *Transportation Research B* 39(2), 119-140.

Chen, A., Yang, H., Lo, H.K, Tang, W.H., 2002. Capacity reliability of a road network: an assessment methodology and numerical results. *Transportation Research B*, 36. 225–252.

Chen, A., Kim, J., Zhou, Z., Chootinan, P., 2007. Alpha Reliable Network Design Problem, In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2029, TRB, National Research Council, Washington, D.C., pp. 49–57.

Daganzo, C., 1979. Multinomial Probit: The Theory and its Application to Demand Forecasting, Academic Press, New York.

Duthie, J.C., Unnikrishnan, A., Waller, S.T., 2009. Robust Traffic Network Analysis: Efficient Techniques for Sampling Multivariate Demand, *Computer-Aided Civil and Infrastructure Engineering*, accepted.

Embrechts, P., McNeil, A., Straumann, D., 2002. Correlation and dependence in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond*. ed. M.A.H. Dempster, pp. 176–223.

Ferson, S., Burgman, M. A., 1995. Correlations, dependency bounds, and extinction risks. Biological Conservation 73: 101–105.

Ghosh, S., Henderson, S.G., 2002. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research* 50, 820–834.

Hazelton, M., 2000. Estimation of origin–destination matrices from link flows on uncongested networks. *Transportation Research Part B* 34 (7), pp. 549–566.

Hoeffding, W. 1940. Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik des Universitat Berlin* 5, 179–233.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate distributions, Vol.1, New York, Wiley.

Krishnamurthy, S., Kockelman, K., 2003. Propagation of Uncertainty in Transportation Land Use Models: Investigation of DRAM-EMPAL and UTTP Predictions in Austin, Texas. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1831, TRB, National Research Council, Washington, D.C., pp. 219–229.

Lam, W.H.K., Shao, H., Sumalee, A., 2008. Modeling impacts of adverse weather conditions on a road network with uncertainties in demand and supply, *Transportation Research Part B*, 42(10), pp. 890-810.

Li, S.T,, Hammond, J.J., 1975. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man and Cybernetics* 4:557-561.

Livny, M., Melamed, B., Tsiolis, A.K.., 1993. The impact of autocorrelation on queueing systems. *Management Science* 39, pp. 322–339.

Lo, H.K., Tung, Y.K., 2003. Network with degradable links: capacity analysis and design. *Transportation Research Part B* 37 (4), 345–363.

Lo, H.-P., Chan, C.-P., 2003. Simultaneous estimation of an origin-destination matrix and link choice proportions using traffic counts, *Transportation Research. Part A*, 37(9), 771–88.

Maher, M.J., 1983. Inferences on trip matrices from observations on link volumes: a Bayesian statistical approach, *Transportation Research Part B* 17 (6), pp. 435–447.

Ng, M.W., Waller, S.T., 2009a. A Dynamic Route Choice Model in Face of Uncertain Capacities (submitted to *Networks and Spatial Economics*).

Ng, M.W., Waller, S.T., 2009b. Reliable System Optimal Network Design: A Convex Mean-Variance Type Model with Implicit Chance Constraints, *Transportation Research Record: Journal of the Transportation Research Board*, accepted.

Ng, M.W., Waller, S.T., 2009c. A Computationally Efficient Methodology to Characterize Travel Time Reliability using the Fast Fourier Transform (submitted to *Transportation Research Part B*).

Ng, M.W., Szeto, W.Y., Waller, S.T., 2009d. Distribution-free Travel Time Reliability Assessment with Probability Inequalities (submitted to *Transportation Research Part B*).

Peeta, S., Ziliaskopoulos, A.K., 2001. Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Networks and Spatial Economics*, Vol.1 (3/4), pp. 233–266.

Pradhan, A., Kockelman, K., 2002. Uncertainty Propagation in an Integrated Land Use-Transportation Modeling Framework: Output Variation via UrbanSim, *Transportation Research Record: Journal of the Transportation Research Board*, 1805, pp. 128-135.

Press, W. H., Tuekolsky, S. A., Vetterling, W. T., Flannery, B. P., 2007. Numerical Recipes: The Art of Scientific Computing, Cambridge University Press.

Rencher, A.C., 2002. Methods of Multivariate Analysis, New York, USA, John Wiley & Sons.

Scheuer,E.M., Stoller, D.S., 1962. On the generation of normal random vectors. *Technometrics* 4, 278-281.

Siu, B.W.Y., Lo, H.K, 2008. Doubly uncertain transportation network: Degradable capacity and stochastic demand. *European Journal of Operational Research* 191, pp. 166–181.

Srinivas, S., Menon, D., Prasad, A.M., 2006, Multivariate Simulation and Multimodal Dependence Modeling of Vehicle Axle Weights with Copulas, *Journal of Transportation Engineering* 132 (12), 945–955.

Sumalee, A., Watling, D.P., Nakayama, S., 2006. Reliable Network Design Problem: Case with Uncertain Demand and Total Travel Time Reliability. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1964, Transportation Research Council of the National Academies, Washington, D.C., pp. 81–90.

Suwansirikul, C., Friesz, T. L., Tobin, R. L., 1987, Equilibrium Decomposed Optimization: A Heuristic for the Continuous Equilibrium Network Design Problem. *Transportation Science*, Vol. 21, No. 4, pp. 254–263.

Waller, S. T., J. L. Schofer, and A. K. Ziliaskopoulos, 2001. Evaluation with Traffic Assignment under Demand Uncertainty. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1771, TRB, National Research Council, Washington, D.C., pp. 69–74.

Wang, S., Dhaene, J., 1998. Comonoticity, correlation order and premium principles. *Insurance: Mathematics and Economics*, 22, 235–242.

Watling, D., 2006. User equilibrium traffic network assignment with stochastic travel times and late arrival penalty, *European Journal of Operational Research* 175 (3), pp. 1539–1556.

Whitt, W., 1976. Bivariate distributions with given marginals, *The Annals of Statistics*, 4, 1280-1289.

Zhao, Y., Kockelman, K., 2002. The Propagation of Uncertainty Through Travel Demand Models: An Exploratory Analysis, *Annals of Regional Science* 36, pp. 145–63.