A Copula-Based Closed-Form Binary Logit Choice Model for Accommodating Spatial Correlation Across Observational Units

Chandra R. Bhat*
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
1 University Station C1761, Austin, TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
Email: bhat@mail.utexas.edu

and

Ipek N. Sener
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
1 University Station, C1761, Austin, TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
Email: ipek@mail.utexas.edu

*corresponding author

ABSTRACT

This study focuses on accommodating spatial dependency in data indexed by geographic location. In particular, the emphasis is on accommodating spatial error correlation across observational units in binary discrete choice models. We propose a copula-based approach to spatial dependence modeling based on a spatial logit structure rather than a spatial probit structure. In this approach, the dependence between the logistic error terms of different observational units is directly accommodated using a multivariate logistic distribution based on the Farlie-Gumbel-Morgenstein (FGM) copula. The approach represents a simple and powerful technique that results in a closed-form analytic expression for the joint probability of choice across observational units, and is straightforward to apply using a standard and direct maximum likelihood inference procedure. There is no simulation machinery involved, leading to substantial computation gains relative to current methods to address spatial correlation. The approach is applied to teenagers' physical activity participation levels, a subject of considerable interest in the public health, transportation, sociology, and adolescence development fields. The results indicate that failing to accommodate heteroscedasticity and spatial correlation can lead to inconsistent and inefficient parameter estimates, as well as incorrect conclusions regarding the elasticity effects of exogenous variables.

Keywords: Spatial analysis, copula, maximum likelihood estimation, teenager physical activity, public health

1. INTRODUCTION

Spatial effects are quite ubiquitous in urban and economic data, whether the data is in aggregate form (such as crime rates, cancer rates, land cover change, and employment rates in each of several defined spatial units) or in disaggregate form (such as shopping activity location choice/commute mode choices for each of several sampled individuals, and pavement surface deterioration levels for each of several sampled roadway sections). The pervasive nature of spatial effects has spawned a vast literature on accommodating spatial effects in different fields such as earth sciences, epidemiology, transportation, land use analysis, geography, and ecology (see Páez and Scott, 2004 for a relatively recent review). The studies in these fields have focused on one or more of three spatial analytic issues in analyzing the dependent variable of interest: spatial dependency, spatial heterogeneity, and spatial heteroscedasticity (see Bhat, 2000).

Spatial dependency (also referred to as spatial autocorrelation) refers to the tendency of the data points to be similar when closer in space. This may occur because of diffusion effects, social interaction effects, or unobserved location-related effects influencing the level of the dependent variable (see Jones and Bullen, 1994; Miller, 1999). In general, ignoring spatial dependency can result in mis-estimated standard errors in linear models (Anselin and Griffith, 1988) and (in addition) inconsistent parameter estimation in non-linear models (Case, 1992). Spatial heterogeneity refers to differences in the data-generating urban process over space due to location-specific effects, as demonstrated by Fotheringham *et al.* (1996, 1997). Fotheringham and Brunsdon (1999) and Griffith and Layne (1999) discuss the reasons for these variations in detail, identifying two equally plausible but indistinguishable sources in analysis. One source is intrinsic behavioral differences in the process across spatial units. The other source is the lack of information (on the part of the analyst) regarding some process-related or spatial-unit related attributes. In either case, the result is spatial non-stationarity. In particular, a single global relationship in a study region may not reflect the urban process appropriately in any local part of the study region. Further, this potential mismatch in the global relationship and local relationships can lead to inconsistent estimates of the effect of variables at the global-level if the process at work has a non-linear form. Spatial heteroscedasticity refers to heterogeneity in the variance of the unobserved process across spatial units. Ignoring spatial heteroscedasticity when it is present leads to inconsistent parameter estimates in non-linear models (see McMillen, 1992; 1995).

As discussed in Bhat (2000), and more recently in Páez (2007), much of the spatial analysis literature to date has focused on accommodating spatial effects in models with continuous dependent variables or proportions. In contrast, and as indicated by Páez (2007), "…the explicit incorporation of spatial effects in discrete choice models is still in its infancy". In this study, the focus is on accommodating spatial dependency in data indexed by geographic location. In particular, the emphasis is on accommodating spatial error autocorrelation in binary discrete choice models, while also controlling for heteroscedasticity across observational units. In the next section, we review the literature on discrete choice models with spatial autocorrelation. Then, Section 1.2 positions the current study within the existing literature and motivates the current research.

1.1 Discrete Choice Models with Spatial Error Autocorrelation

Spatial error autocorrelation (or simply spatial correlation in the rest of this paper) may arise because of error correlation across alternatives or across units of observation. Spatial correlation across alternatives arises naturally when the alternatives correspond to spatial units. This type of spatial correlation has been examined primarily in the transportation and geography literatures (see Hunt *et al*., 2004). The common model structures to accommodate such inter-alternative spatial correlation include the mixed logit model (see Bolduc *et al.*, 1996 and Miyamoto *et al.*, 2004), the multinomial probit model (see Garrido and Mahmassani, 2000 and Bolduc *et al.*, 1997), and a GEV-based spatially correlated logit (SCL) model (see Bhat and Guo, 2004). While spatial correlation across alternatives is an important component of modeling choice among multinomial spatial units, there are several choice occasions where the alternatives themselves are not spatial units. However, the choice among the aspatial alternatives may be moderated by space in a way that generates spatial correlation across the choice decisions of observational units. It is this aspect of spatial correlation that is of interest in the current study, as discussed next.

Spatial correlation across observational units has been the focus of attention in the regional science and political/social science literature, though much of this literature is oriented toward non-discrete dependent variables. However, there has been increasing interest recently in accommodating spatial correlation across observational units in models with discrete dependent variables. A brief overview of the most commonly used estimation techniques is provided below.

Case (1992) proposes a spatial probit-based maximum likelihood method that allows spatial dependence using a structure that generates correlation among observations within a region, but

assumes that there is no correlation among observations in different regions. McMillen (1992) proposed an Expectation Maximization (EM) algorithm to account for autocorrelation across observational units using a more general spatial autocorrelation structure in a probit model. The idea is to replace the latent dependent variable of the probit structure with an *expectation* based on the observed binary choice, and then to estimate the resulting model using standard *maximum likelihood* techniques for the case of a continuous dependent variable. The estimation results from the second *"maximization"* step provide new estimates of parameters that are used to construct updated *expectations* of the latent variable, and the procedure is iterated to convergence in the parameters.

LeSage (2000) uses an approach similar to the EM algorithm of McMillen for a binary probit model, but adopts a Bayesian estimation approach for the *maximization* step using a Monte Carlo Markov Chain (MCMC) procedure (*i.e.*, Gibbs sampling with a Metropolis-Hastings algorithm). This procedure entails specifying a complete conditional distribution for the model parameters and iteratively sampling from these conditional distributions (with the conditioning variables for each set of model parameters being the most recent draws of other model parameters). The sequence of the resulting parameter draws converge to the joint posterior distribution of the parameters after a sufficient number of draws (see Casella and George, 1992 or Train, 2003 for clear expositions of the MCMC approach). To construct values of the latent dependent variable (as in the *expectation* step of the EM technique), LeSage adds an additional conditional distribution for the posterior of this latent variable conditional on all other parameters, which takes the form of a truncated normal distribution. Both McMillen and LeSage apply their models to neighborhood crime, categorized as low crime or high crime based on the number of burglaries and vehicle thefts per thousand households in each neighborhood. The number of observational units (or neighborhoods) in their application is 49.

Pinkse and Slade (1998) propose a two-step Generalized Method of Moments (GMM) estimation technique for probit models that considers the induced heteroscedasticity from spatial correlation effects, but does not accommodate the dependency across observations due to the spatial correlation effects.[1] A related GMM method that is equivalent to a weighted non-linear version of the familiar feasible generalized least squares estimator is described in Fleming (2004), and is based off the work of Kelejian and Prucha (1999) for the continuous-dependent variable case.

[1] Klier and McMillen (2007) propose a linearized logit variant of Pinkse and Slade's estimator, but this linearization technique does not work in the purely spatial error model since the gradient with respect to the spatial correlation term is zero for all observations at the starting linearization point that corresponds to the correlation term being equal to zero.

Beron *et al.* (2003) and Beron and Vijverberg (2004) adopt, in their binary probit model with spatial error correlation, a recursive importance sampling (RIS) technique to directly evaluate the likelihood function that involves a multidimensional integral of the order of the number of observational units. This constitutes a maximum simulated likelihood (MSL) method using a GHK simulator, different from the EM or the Bayesian approaches of McMillen and LeSage.

A problem with the approaches just discussed is that they are not feasible for moderate-to-large samples since they require the inversion and determinant computation of a square matrix of the order of the number of observational units (for McMillen's EM method, LeSage's MCMC method, and Pinkse and Slade's heteroscedastic approach), or treat spatial dependence as a nuisance with no provision of an estimate of the standard error of the spatial error parameter (for the GMM method described in Fleming, 2004), or require the simulation of a multidimensional integral of the order of the number of observational units (for the RIS-based method).[2] Another possible approach is to maintain a relatively restrictive spatial correlation structure that allows a constant correlation within observational units in pre-specified spatial regions, but no correlation in observational units in different spatial regions. Bhat (2000) and Dugundji and Walker (2005) address unordered multinomial discrete choices in this manner. Specifically, Bhat (2000) examines work travel mode choice, allowing for error correlation across decision-makers based on residential location as well as work location. Dugundji and Walker (2005) also examine work mode choice, but allow for spatial error correlation only among decision-makers in the same residential location. The result of such restrictive error correlation specifications is a considerable reduction in the dimensionality of integrals in the likelihood function, allowing relatively easy estimation within a mixed logit framework. However, these studies are likely to be more affected by the modifiable areal unit problem (MAUP) than general autocorrelation structures that are not as dependent on the definition of spatial regions (see Páez and Scott, 2004).[3]

[2] See Franzese and Hays, 2007 for a detailed discussion of the drawbacks of the various methods in general, and the MCMC method in particular.

[3] In Anselin's (2003) taxonomy, the work of Bhat (2000), Dugundji and Walker (2005), and Case (1992) (described earlier in the section) corresponds to "local" spatial effects, while more general correlation structures allow "global" spatial effects.

1.2 The Current Paper in Context and Paper Structure

The current paper focuses on general spatial correlation structures across observational units for the case of a discrete dependent variable (more specifically, a binary discrete variable). While earlier studies discussed in Section 1.1 have addressed this problem, these methods become infeasible with a high number of observational units. The methods of Bhat (2000) and Dugundji and Walker (2005) are readily applicable to the case of binary outcomes, but they rely on restrictive local specifications of the spatial correlation structure across observations.

In the current paper, we propose a new approach to accommodate spatial correlation across observational units. The resulting spatial logit model retains a simple closed-form expression, obviating the need for any kind of simulation machinery and methods. In particular, the model can be estimated using a standard and direct maximum likelihood technique and is computationally tractable even for a high number of observational units. The methodology proposed here highlights the power of closed-form techniques, and serves as yet another reminder that researchers would do well to formulate closed-form models rather than get carried away by the simulation advancements of the day.

The rest of this paper is structured as follows. The next section discusses the concept of a copula, which is a multivariate functional form for the joint distribution of random variables derived purely from the marginal distribution of each random variable. Section 3 employs a particular type of copula to generate spatial dependence among observations in any binary discrete choice structure. Section 4 describes the data source and sample formation procedures for an empirical application of the proposed spatial logit model to teenagers' physical activity participation. Section 5 presents the corresponding empirical results. The final section summarizes the important findings from the study.

2. THE COPULA APPROACH

The incorporation of dependency effects can be greatly facilitated by using a copula approach for modeling joint distributions, so that the resulting model can be in closed-form and can be estimated using direct maximum likelihood techniques (the reader is referred to Trivedi and Zimmer, 2007 or Nelsen, 2006 for extensive reviews of copula approaches and their benefits). A copula approach basically involves the generation of a multivariate joint distribution, given the marginal distributions of the correlated variables. The word copula itself was coined by Sklar, 1959 and is derived from the Latin word "copulare", which means to tie, bond, or connect (see Schmidt, 2007). Thus, a copula is a

device or function that generates a stochastic dependence relationship among random variables with pre-specified marginal distributions. In essence, the copula approach separates the marginal distributions from the dependence structure, so that the dependence structure is entirely unaffected by the marginal distributions assumed. This provides substantial flexibility in correlating random variables, which may not even have the same marginal distributions. The effectiveness of a copula approach has been recognized in the statistics field for several decades now (see Schweizer and Sklar, 1983, Chapter 6), but it is only recently that Copula-based methods have been explicitly recognized and employed in the financial risk analysis and econometrics fields (see, for example, Smith, 2005, Zimmer and Trivedi, 2006, Cameron *et al.*, 2004, Embrechts *et al.*, 2003, Cherubini *et al.*, 2004, Junker and May, 2005, Quinn, 2007, and Bhat and Eluru, 2008).

The precise definition of a copula is that it is a multivariate distribution function defined over the unit cube linking uniformly distributed marginals. Let C be a Q-dimensional copula of uniformly distributed random variables $U_1, U_2, U_3, \ldots, U_Q$ with support contained in $[0,1]^Q$. Then,

$$C_\theta(u_1, u_2, \ldots, u_Q) = \Pr(U_1 < u_1, U_2 < u_2, \ldots, U_Q < u_Q), \quad (1)$$

where θ is a parameter vector of the copula commonly referred to as the dependence parameter vector. A copula, once developed, allows the generation of joint multivariate distribution functions with given marginals. Consider Q random variables $V_1, V_2, V_3, \ldots, V_Q$, each with standard univariate continuous marginal distribution functions $F_q(v_q) = \Pr(V_q < v_q)$, $q = 1, 2, 3, \ldots, Q$. Then, by the integral transform result, and using the notation $F_q^{-1}(.)$ for the inverse standard univariate cumulative distribution function, we can write the following expression for each q ($q = 1, 2, 3, \ldots, Q$):

$$F_q(v_q) = \Pr(V_q < v_q) = \Pr(F_q^{-1}(U_q) < v_q) = \Pr(U_q < F_q(v_q)). \quad (2)$$

Finally, by Sklar's (1973) theorem, a joint Q-dimensional distribution function of the random variables with the continuous marginal distribution functions $F_q(v_q)$ can be generated as follows:

$$\begin{aligned} F(v_1, v_2, \ldots, v_Q) &= \Pr(V_1 < v_1, V_2 < v_2, \ldots, V_Q < v_Q) = \Pr(U_1 < F_1(v_1), U_2 < F_2(v_2), \ldots, U_Q < F_Q(v_Q)) \\ &= C_\theta(u_1 = F_1(v_1), u_2 = F_2(v_2), \ldots, u_Q = F_Q(v_Q)). \end{aligned} \quad (3)$$

Conversely, by Sklar's theorem, for any multivariate distribution function with continuous marginal distribution functions, a unique copula can be defined that satisfies the condition in Equation (3).

Copulas themselves can be generated in several different ways, including the method of inversion, geometric methods, and algebraic methods (see Nelsen, 2006; Chapter 3). A rich set of copula types have been generated using these methods, including the Farlie-Gumbel-Morgenstern (FGM) family, the Archimedean group (including the Clayton, Gumbel, Joe, and Frank family of copulas), the Gaussian copula, and the Student's *t*-copula. In this paper, we consider the simplest of these copulas, the FGM family, which is particularly well-suited to incorporate spatial correlation. As we will discuss later, the FGM family also imposes certain restrictions that are not maintained by other copulas, but these other copulas are practically infeasible for the case of accommodating spatial correlation across observational units. Based on the FGM copula family, we will develop a particular multivariate variant of Gumbel's (1961) Type III bivariate logistic distribution for use in binary choice models.

2.1 The FGM Family-Based Multivariate Logistic Distributions

The FGM family of copulas was first proposed by Morgenstern (1956), and has been well known for some time in statistics (see Conway, 1983, Kotz *et al.*, 2000; Section 44.13). However, until Prieger's (2002) application for sample selection, it does not seem to have been used in econometrics. In the bivariate case, the FGM copula takes the following form:

$$C(U_1 < u_1, U_2 < u_2) = u_1 u_2[1 + \theta(1-u_1)(1-u_2)]. \tag{4}$$

For the copula above to be 2-increasing (that is, for any rectangle with vertices in the domain of [0,1] to have a positive volume based on the function), theta must be in [-1,1] (see Nelsen, 2006; pg. 77). The presence of the theta term allows the possibility of correlation between the uniform marginals U_1 and U_2. Specifically, the density function for the FGM copula is:

$$c_{U_1U_2}(u_1, u_2) = 1 + \theta(1-2u_1)(1-2u_2). \tag{5}$$

From above, it is clear that, when θ is positive, the density is higher if u_1 and u_2 are both high (both close to 1) or both low (both close to zero). On the other hand, when θ is negative, the density is higher if u_1 is high and u_2 is low, or if u_2 is high and u_1 is low. Thus, when θ is zero, it corresponds to independence. Otherwise, depending on whether θ is positive or negative, a positive or negative correlation, respectively, is generated between the continuous variables u_1 and u_2. Thus,

the FGM copula has a simple analytic form and allows for either negative or positive dependence. However, the correlation between U_1 and U_2 is restricted in the range of [-1/3, 1/3].

The standard bivariate logistic distribution corresponding to the copula in Equation (4) is obtained using Equation (3) as follows:

$$\Lambda(V_1 < v_1, V_2 < v_2) = \Lambda_1(v_1)\Lambda_2(v_2)\left[1 + \theta(1-\Lambda_1(v_1))(1-\Lambda_2(v_2))\right], \tag{6}$$

where $\Lambda_q(v_q) = \dfrac{1}{1+e^{-v_q}}$ for q = 1, 2.

The above distribution is Gumbel's (1961) bivariate logistic distribution, with the Spearman's correlation coefficient $\rho(V_1, V_2)$ being $\dfrac{3\theta}{\pi^2}$. The restriction that θ should be in the [-1, 1] range implies that the maximal correlation between V_1 and V_2 is restricted by $|\rho(V_1, V_2)| \leq 0.304$. Thus, the logistic distribution of Equation (6) allows only moderate dependence. However, in a modeling context, the correlation refers to the association between unobserved elements after controlling for observed factors. In fact, high correlations between unobserved factors suggest a model that is not well-specified in its exogenous variables. Further, in the typical context of spatial structure-based dependence, the correlation between observational units drops off sharply with geographic distance (see Anselin, 2003). Thus, the correlation range of the FGM logistic distribution may not be too limiting.[4] At the same time, the advantages of using the FGM distribution for spatial correlation analysis are several. First, the method offers a simple linear copula structure that can easily be extended to multivariate correlation structures to accommodate spatial correlation across several observational units. Second, the resulting structure provides a closed-form solution for the choice outcomes in the binary choice context that has been examined in the literature on spatial models. Thus, model estimation can proceed using standard simulation-free direct maximum likelihood methods. Finally, the model easily accommodates flexible patterns of spatial correlation across observational units (rather than the strict space-based ordering effects that generally are imposed on spatial correlation in extant models).

[4] There are other copulas that are not as limiting as the FGM copula. However, these copulas are extremely difficult to apply in a spatial modeling context, as discussed later. Also, as indicated by Prieger (2002), "allowed correlation is only one dimension along which to judge a model". In fact, Prieger goes on to show that the multivariate FGM distribution with limited correlation substantially outperforms the multivariate normal distribution with a full range of correlation in his empirical application.

The bivariate logistic distribution can be readily extended using a multivariate version of the FGM copula (see Nelsen, 2006; pg. 108). A particularly appealing approach to constructing a multivariate logistic distribution for spatial correlation analysis is to allow pairwise correlation across observational units (see Karunaratne and Elston, 1998 for such a pairwise correlation structure):

$$\Lambda(V_1 < v_1, V_2 < v_2, ..., V_q < v_q, ..., V_Q < v_Q) = \left[\prod_{q=1}^{Q} \Lambda_q(v_q)\right] \times \left[1 + \sum_{q=1}^{Q-1} \sum_{k=q+1}^{Q} \theta_{qk} \cdot \left(1 - \Lambda_q(v_q)\right)\left(1 - \Lambda_k(v_k)\right)\right], \quad (7)$$

where θ_{qk} is the dependence parameter between V_q and V_k ($-1 \le \theta_{qk} \le 1$), $\theta_{qk} = \theta_{kq}$ for all q and k, and $\Lambda_q(v_q) = \dfrac{1}{1+e^{-v_q}}$.

It is important to note that the multivariate distribution above is legitimate only when the corresponding multivariate density is nonnegative, which implies the following restriction on the θ_{qk} parameters:

$$\left[1 + \sum_{q=1}^{Q-1} \sum_{k=q+1}^{Q} \theta_{qk} \cdot \left(1 - 2\Lambda_q(v_q)\right)\left(1 - 2\Lambda_k(v_k)\right)\right] \ge 0. \quad (8)$$

Theoretically speaking, and since each $\Lambda_q(v_q)$ can take any value between 0 and 1, the requirement for a nonnegative density globally at any (and all) points of the entire Q-dimensional space of possible combination values of $\Lambda_q(v_q)$ across observational units is equivalent to the condition that the density is nonnegative at each of the 2^Q vertices of the Q-dimensional unit cube $[0,1]^Q$. This requires the very strong and limiting condition on the θ_{qk} values (see Cambanis, 1977 and Armstrong and Galli, 2002):

$$\left[1 + \sum_{q=1}^{Q-1} \sum_{k=q+1}^{Q} \theta_{qk} \Delta_q \Delta_k\right] \ge 0, \text{ for all } \Delta_1, \Delta_2, ... \Delta_Q \in \{-1, +1\}. \quad (9)$$

However, from a practical standpoint, the range of space spanned by the $\Lambda_q(v_q)$ terms across observational units in discrete choice models is quite far away from the vertex points of the Q-dimensional unit cube. This is so even for prediction purposes on a new set of data and observations.

In the spatial correlation setting in which the FGM copula-based multivariate logistic distribution is being applied, the θ_{qk} terms are also quite small for pairs of observations that are not geographically proximate. The net effect of all this is that the restriction in Equation (9) will seldom be violated in the context of spatial correlation across observations in discrete choice models, either in estimation or in application, as long as each θ_{qk} term is bounded between 0 and 1. Thus, in general, there will be no practical need to expressly impose the very strong restriction implied by Equation (9). We found this to be the case in our own estimations and application of the model developed in the current paper.

3. THE BINARY CHOICE MODEL WITH SPATIAL CORRELATION

Consider that the data (z_q, x_q) for q = 1, 2, …, Q are generated by the following latent variable framework:

$$z_q^* = \beta' x_q + \varepsilon_q$$

$$z_q = \begin{cases} 0 \text{ if } z_q^* < 0 \\ 1 \text{ if } z_q^* \geq 0 \end{cases} \tag{10}$$

where z_q^* is an unobserved propensity variable, β is a vector of coefficients to be estimated, and ε_q is a logistically distributed idiosyncratic error term with a scale parameter of σ_q (this allows spatial heteroscedasticity).[5] Define $V_q = \varepsilon_q / \sigma_q$, where V_q is standard logistic distributed. Let the V_q terms (q = 1, 2, …, Q) follow the standard multivariate logistic distribution in Equation (7). Also, let d_q be the actual observed value of z_q in the sample. Then, the probability of the observed vector of choices $(d_1, d_2, d_3, ..., d_Q)$ can be written, after some algebraic manipulations, as:

[5] As indicated by Franzese and Hays (2007), almost all earlier methodological and applied research on spatial discrete choice models have considered a normally distributed error term, leading to a spatial probit model. However, as we show below in the current paper, a univariate logistical error distribution for each individual ε_q, combined with the FGM copula to generate dependence among the ε_q terms, leads to a simple spatial logit model structure with a closed-form solution for the joint choice probabilities.

$$P(z_1 = d_1, z_2 = d_2, ..., z_Q = d_Q) = \left[\prod_{q=1}^{Q} \frac{e^{\left(\frac{\beta' x_q}{\sigma_q}\right) \cdot d_q}}{1 + e^{\left(\frac{\beta' x_q}{\sigma_q}\right)}} \right]$$
$$\left[1 + \sum_{q=1}^{Q-1} \sum_{k=q+1}^{Q} (-1)^{d_q + d_k} \cdot \theta_{qk} \left\{ 1 - \frac{e^{\left(\frac{\beta' x_q}{\sigma_q}\right) \cdot d_q}}{1 + e^{\left(\frac{\beta' x_q}{\sigma_q}\right)}} \right\} \left\{ 1 - \frac{e^{\left(\frac{\beta' x_k}{\sigma_k}\right) \cdot d_k}}{1 + e^{\left(\frac{\beta' x_k}{\sigma_k}\right)}} \right\} \right] \quad (11)$$

The above probability function considers the spatial correlation across observation units through the θ_{qk} terms (see previous section). When all the θ_{qk} terms are zero, the expression above collapses to a heteroscedastic binary choice model with no spatial correlation.[6]

In the probability expression of Equation (11), it is not possible to estimate a separate θ_{qk} term for each pair of observational units from the data. So, we propose that these terms be parameterized by writing:

$$\theta_{qk} = \pm \left[\frac{(e^{\delta})' s_{qk}}{1 + (e^{\delta})' s_{qk}} \right], \quad (12)$$

where s_{qk} is a vector of variables that influence the level of spatial correlation between observational units *q* and *k*, and δ is a parameter vector whose elements are associated with the elements of s_{qk}. By functional form, the expression in parenthesis in the above equation is bounded by 0 and 1, ensuring that the θ_{qk} terms are between –1 and 1. The form also ensures the symmetry of the θ_{qk} terms (*i.e.*, $\theta_{qk} = \theta_{kq}$). In the current empirical context, we expect observational units in close proximity to have similar preferences. For this reason, we impose the '+' sign in front of the expression in Equation (12), which generates positive correlation between pairs of observational units.

[6] The simple structure for the spatially correlated binary logit model in Equation (11) is the reason for using the FGM-based multivariate logistic distribution. While more flexible multivariate distributions that do not restrict the magnitude of correlation to be 0.31 or less can be developed using the Archimedean family of copulas (such as Frank's copula or Gumbel's copulas) or other copulas, the problem with these is that the equivalent probability expression to Equation (11) can have up to 2^Q terms on the right side. Even for medium-sized *Q* values (sample sizes), computation of the probability expression becomes prohibitive. In the case of the FGM copula, the expressions simplify, so we get the simple and elegant expression in Equation (11).

The functional form of Equation (12) can accommodate multiple variables in the s_{qk} vector that may influence spatial correlation across observational units, including whether or not two observational units are in the same spatial unit, whether or not two observational units are contiguous, boundary length of shared border between spatial units, and inverse of time or distance. The parameterization in Equation (12) can also be used to test for "thresholds" of distance or time or cost beyond which there is effectively no spatial correlation between observation units (through the appropriate specification of independent variables in the s_{qk} vector). Finally, the specification in Equation (12) nests the typical spatial correlation patterns used in the literature as special restrictive cases. For instance, the case of observational units correlated only if they are in the same spatial unit corresponds to only one variable in the s_{qk} vector, which takes a value of 1 if q and k are in the same spatial unit and 0 otherwise. If there is only one observation per spatial unit (as in McMillen, 1992 and LeSage 2000), then first-order adjacency-based correlation can be obtained by restricting the specification so that there is one element in s_{qk} which takes a value of 1 if q and k are in adjacent spatial units and zero otherwise. The case of a distance-based decay effect corresponds to a restricted version of Equation (12) with an inverse function of distance as the only element in the s_{qk} vector.

The parameter σ_q in Equation (11) is next parameterized as:

$$\sigma_q = g(\lambda' \varpi_q) = \exp(\lambda' \varpi_q), \tag{13}$$

where ϖ_q includes variables specific to pre-defined "neighborhoods" (or other groupings) of observational units and individual-related factors (see Páez, 2006 and Bhat and Zhao, 2002).

Finally, the likelihood function of Equation (11) can be directly maximized to estimate the β, δ, and λ vectors. The starting parameters may be obtained by constraining $\theta_{qk} = 0$ for all q and k, and estimating β and λ (this corresponds to a heteroscedastic binary logit model with independence across observation units). The resulting values of β and λ can be used to start the iterations for the likelihood maximization. In the current study, the GAUSS matrix programming language was employed to undertake the maximum likelihood estimation.

A unique feature of our spatial model is that we directly capture heteroscedasticity and correlation in the error terms ε_q across observational units, rather than using a pre-specified spatially autoregressive structure to generate dependence - $\varepsilon = \rho W\varepsilon + u$, where ε is a Q-dimensional vector of all the ε_q terms stacked vertically, ρ is the spatial autoregressive parameter, W is a row-standardized spatial weights matrix that represents an average of values from neighboring spatial units, and u is a Q-vector of independently distributed error terms. This autoregressive structure also indirectly generates heteroscedasticity, though this heteroscedasticity is artificial and not directly associated with inherent variations across spatial units (see McMillen, 1992). In our copula approach, the spatial correlation effect is completely delineated from heteroscedasticity effects, allowing a clear capture and testing of each of the heteroscedasticity and correlation effects distinctly. At the same time, our approach enables the accommodation of flexible patterns of spatial correlation, is simple to implement using traditional maximum likelihood methods, and does not require any simulation machinery whatsoever.

4. EMPIRICAL CONTEXT AND DATA

In this paper, the copula-based binary choice model with spatial correlation is applied to examine the factors that influence whether or not a teenager participates in physical activity during the course of a day. Physical activity is an inherent part of a healthy lifestyle with the potential to increase the quality and years of life in the U.S. (see U.S. Department of Health and Human Services, USDHHS, 2000). Epidemiological research studies have emphasized a strong association between the lack of physical activity and the increasing rates of morbidity and mortality due to obesity, coronary heart disease, stroke, diabetes, high blood pressure, colon cancer, depression, and anxiety (see, for instance, Feldman *et al.*, 2003, Dong *et al.*, 2004, Nelson and Gordon-Larsen, 2006, and Ornelas *et al.*, 2007). Other studies have established that regular physical activity helps increase cardiovascular fitness, enhances agility and strength, reduces the need for medical help, and improves mental health (USDHHS, 1996; Center for Disease Control, CDC, 2006).

While the physical, emotional, and social benefits of physical activity are well established, physical inactivity is quite prevalent in today's developed world, particularly among adolescents (Dong *et al.*; 2004, Warburton *et al.*, 2006). According to the results of the National Health and Life Style Surveys (2003), only about two-thirds of teenage boys and one-third of teenage girls reported

participating in some form of vigorous physical activity during the day.[7] A report by the Center for Disease Control (CDC, 2002) also indicates that about a third of teenagers do not engage in adequate physical activity for health, and that the high school physical education class participation rate has been steadily declining over the past decade. In addition, studies have shown that physical activity participation decreases with age during the adolescence period (USDHHS, 1996; Mhuircheartaigh, 1999).

The low physical activity participation, as well as the decrease in physical activity rates with age among adolescents, constitutes a national health concern, since inactive lifestyles may be transferred from adolescence to adulthood. Aaron *et al.* (2002) indicate that "[t]he adolescent years are thought to be the period during which adult health behavior, such as dietary and physical activity patterns, begin to develop". Therefore, public health professionals strongly emphasize the importance of developing effective strategies to motivate and increase physical activity participation as an individual moves through the adolescence stage of life.

At the same time that public health professionals are focusing on ways to promote physical activity participation among adolescents, urban transportation planners are becoming more interested in expanding their focus from studying only adults' activity-travel patterns to also explicitly examining children's activity-travel patterns, including children's participation in physical activity pursuits. This is because young teenagers depend, to a large extent, on household adults or other adults to drive them to physically active and non-physically active events, which in turn can influence adults' activity-travel patterns in important ways (Reisner, 2003). For instance, the need to drop a teenager off at soccer practice at a certain time and location would influence the temporal and spatial dimensions of a parent's activity-travel pattern. In addition to serve-passenger activities, children's desires and needs can also influence adults' activity-travel patterns through joint activity participation in such activities as going to the park, walking together, or playing tennis. Of course, the consideration of children's activity-travel patterns is also important in its own right since these patterns contribute directly to travel demand. Due to the above reasons, the transportation field has witnessed an increasing interest in children's activity-travel patterns, including participation in

[7] Vigorous physical activity is one that requires an energy expenditure that is more than 6 times the energy expended when sitting quietly, or equivalently, an energy expenditure of more than 7 kilo-calories per minute. Sample vigorous physical activities include jogging or running, mountain climbing, and bicycling more than 10 mph or bicycling uphill (see USDHHS, 2000).

physical activity (see Transportation Research Board and Institute of Medicine, 2005; Goulias and Kim, 2005; Copperman and Bhat 2007a, 2007b; and Sener *et al*., 2008).

In the current application, we model teenagers' participation in physical activity using a comprehensive set of socio-demographic and physical environment variables, while also accommodating spatial effects based on residence patterns. In the physical activity literature, the factors affecting the physical activity behavior of an individual have been broadly classified into three categories (see GAO, 2006): (1) Demographic factors (including individual and household demographics), (2) physical environment factors (including community design/built environment factors such as land-use and transportation system attributes and urban form factors, and other environment factors such as perceived safety, weather, and season of the year), and (3) Cognitive and behavioral factors (including individual-level attitudes, beliefs, and perceptions, and family and social influences). While there is an extensive literature focusing on the effects of the first and third category of factors on children's/teenagers' physical activity (see Feldman *et al.*, 2003, Kelly *et al.*, 2006, and Salmon, 2007 for some recent examples), there has been limited research on the physical environment determinants of physical activity. Further, the few research studies that have accommodated physical environment factors in the analysis of children's/teenagers' physical activity have predominantly considered variables relating to highway network measures and the presence of, or access to, recreational facilities and programs (see, for example, Sallis *et al.*, 2000, and Gordon-Larsen *et al.*, 2005). In this paper, we consider a more extensive set of physical environment variables associated with teenagers' residential neighborhoods, including (a) land-use attributes, (b) size and density measures, (c) accessibility to shopping, employment, and recreational activities, (d) ethnic composition, (e) demographics and housing cost attributes, (f) the intensity and density of activity opportunities, and (g) transportation network measures. But the current study also has its limitations. Specifically, we do not accommodate cognitive and behavioral determinants of physical activity participation. Further, as in earlier studies, the current research study is based on cross-sectional outcome data from which one can, strictly speaking, only infer correlations and not causal effects. This caveat should be kept in mind as we interpret the effects of variables on physical activity. The reader will note that cross-sectional data still remains the main basis for modeling and interpreting physical activity behavior, since collecting panel data or process data has its own set of problems and limitations.

4.1 Data Sources

The primary source of data is the 2000 San Francisco Bay Area Travel Survey (BATS), which was designed and administered by MORPACE International, Inc. for the Bay Area Metropolitan Transportation Commission. The survey collected detailed information on individual and household socio-demographic and employment-related characteristics from over 15,000 households in the Bay Area. The survey also included data on all activity-travel episodes for a two-day period of time (see MORPACE International Inc., 2002 for details on survey, sampling, and administration procedures).

In addition to the 2000 BATS survey data set, several other secondary data sets were used to obtain spatial variables that may influence spatial correlation across observational units as well as physical environment variables that may characterize the choice behavior of individuals. These secondary data sets include: 1) Land-use and demographic coverage data, 2) Zone-to-zone travel level-of-service (LOS) data, 3) The 2000 Census of population and household summary files (SF1), 4) GIS layers of highways (including interstate, toll, national, state and county highways), 5) GIS layers of local roadways (including local, neighborhood, and rural roads), 6) GIS layers of bicycle facilities, and 7) GIS layers of businesses (shopping and grocery stores, medical facilities and personal services, automotive businesses, food stores, sports and fitness centers, parks and gardens, restaurants, recreational businesses, and schools).

Among the secondary data sets indicated above, the land-use/demographic coverage data, LOS data, and the GIS layer of bicycle facilities were obtained from the Metropolitan Transportation Commission (MTC). The GIS layers of highways and local roadways were obtained from the 2000 Census Tiger Files. The GIS layers of businesses were obtained from the InfoUSA business directory. The details of the variables extracted from these data sets are discussed in the next section.

4.2. Sample Formation

4.2.1 Sample Extraction

The final sample used for the analysis is confined to teenagers residing in nine Counties (Alameda, Contra Costa, San Francisco, San Mateo, Santa Clara, Solano, Napa, Sonoma and Marin) of the San Francisco Bay Area. Several steps were pursued in extracting the final sample for the analysis, which includes 722 teenagers. First, since the focus is on the physical activity participation of teenagers, only individuals aged between 13 and 19 years in the BATS survey data were considered

in the analysis. Second, each episode was classified as being physically active or physically passive.[8] Third, a binary flag variable was created for each teenager-day combination taking the value of '1' if the individual participated in one or more physically active episodes during the course of the day, and '0' otherwise. This binary flag serves as the dependent variable in the current empirical application. Fourth, only weekday data were selected, since the focus of the current analysis is on weekdays. Fifth, only one randomly chosen day was selected among the two-day activity diary data for each teenager, and only one randomly chosen teenager was picked from each household, to keep the sample size manageable and prevent problems that may occur due to repeated data measurement from the same individual and/or household.

Of the 722 teenagers, 186 (26%) participate in physical activity during the course of their sampled weekday and 536 (74%) do not participate in physical activity. These results are similar to those found by the Center for Disease Control, CDC, 2006.

4.2.2 Physical Environment Measures

The six secondary data sources identified in Section 2 provide a rich set of physical environment variables for each TAZ.[9] These physical environment variables include:

1) *Zonal land-use structure variables*, including housing type measures (fractions of single family, multiple family, duplex and other dwelling units), land-use composition measures (fractions of zonal area in residential, commercial, and other land-uses), and a land-use mix diversity index computed as a fraction based on the land-use composition measures with values between 0 and 1 (zones with a value closer to one have a richer land-use mix than zones with a value closer to zero; see Bhat and Guo, 2007 for detailed explanation on the formulation of this index).
2) *Zonal size and density measures*, including total population, number of housing units, population density, household density, and employment density by several employment

[8] A physically active episode requires regular bodily movement during the episode, while a physically passive episode involves maintaining a sedentary and stable position for the duration of the episode. For example, playing basketball or walking around the neighborhood would be a physically active episode, while watching television constitutes a physically passive episode. The designation of an episode as physically active or physically passive was based on the nature of the episode and the location type at which it is pursued, as reported in the survey. Thus, an episode designated as "recreation" by a respondent and pursued at a health club is labeled as physically active. Due to space constraints, we are unable to provide a detailed description of the activity episode classification procedure. Interested readers may obtain the procedure from the authors.

[9] The use of a TAZ as a spatial unit of resolution for computing physical environment variables is admittedly rather coarse. Future studies should consider more micro-scale measures to represent physical environment variable effects.

categories, as well as dummy variables indicating whether the area corresponds to a central business district (CBD), urban area, suburban area, or rural area.

3) *Regional accessibility measures*, which include Hansen-type employment, shopping, and recreational accessibility indices that are computed separately for the drive and transit modes.
4) *Zonal ethnic composition measures,* constructed as fractions of Caucasian, African-American, Hispanic, Asian and other ethnic populations for each zone.
5) *Zonal demographics and housing cost variables*, including average household size, median household income, and median housing cost in each zone.
6) *Zonal activity opportunity variables*, characterizing the composition of zones in terms of the intensity or the density of various types of activity centers. The typology used for activity centers includes five categories: (a) maintenance centers, such as grocery stores, gas stations, food stores, car wash, automotive businesses, banks, medical facilities, (b) physically active recreation centers, such as fitness centers, sports centers, dance and yoga studios, (c) physically passive recreational centers, such as theatres, amusement centers, and arcades, (d) natural recreational centers such as parks, gardens, and e) restaurants and eat-out places.
7) *Zonal transportation network measures*, including highway density (miles of highway facilities per square mile), local roadway density (miles of roadway density per square mile), bikeway density (miles of bikeway facilities per square mile), street block density (number of blocks per square mile), non-motorized distance between zones (*i.e.*, the distance in miles along walk and bicycle paths between zones), and transit availability. The non-motorized distance between zones was used in the empirical analysis to develop an accessibility measure by non-motorized modes, computed as the number of zones (a proxy for activity opportunities) within "x" non-motorized mode miles of the teenager's residence zone. Several variables with different thresholds for "x" were formulated and tested.

4.2.3 Spatial Correlation Variables

Among the secondary data sets, the land use coverage data were used to obtain variables to characterize spatial correlation patterns across observational units (these are the elements of the s_{qk} vector in Section 3). Specifically, Geographic Information System (GIS) procedures were implemented to compute measures that may contribute to spatial correlation in physical activity choices between each pair of observational units (teenagers in the current application). These

included (1) whether or not two teenagers reside in the same TAZ, (2) whether or not two teenagers reside in contiguous TAZs, (3) the boundary length of the shared border between the residence zones of two teenagers, and 4) several functional forms of the distance between the residence TAZ activity centroids of the two teenagers, such as inverse of distance, square of inverse of distance, and distance "cliff " measures (the latter form essentially allows the spatial correlation between two teenagers to go to zero beyond a certain distance threshold).[10]

5. MODEL RESULTS

5.1 Variable Specification

Several variable specifications and functional forms were considered in the model. These included (1) teenager demographics (age, sex, race, driver's license holding, physical disability status, *etc.*), (2) household demographics (number of adults, number of children, household composition and family structure, household income, dwelling type, whether the house is owned or rented, *etc.*), (3) activity-day variables (season of the year, day of week, rain-fall, *etc.*), (4) physical environment measures (see Section 4.2.2), and (5) spatial correlation variables (see Section 4.2.3). In addition, several interaction effects of the variables were considered.

The final model specification was based on intuitive considerations, insights from previous literature, parsimony in specification, and statistical fit/significance considerations. The final specification includes some variables that are not highly statistically significant, because of their intuitive effects and potential to guide future research efforts in the field.

5.2 Estimation Results

Table 1 presents the estimation results for teenagers' weekday physical activity participation choice. The coefficients provide the effects of variables on the latent propensity to participate in physical activity. The second main column provides the aspatial binary logit (ABL) model results, while the third main column presents the copula-based spatially correlated heteroscedastic binary logit model (SCHBL) results. Overall, the parameter estimates for the two models have the same sign (except the presence of bicycle variable), though the asymptotic t-statistic values of several determinants of

[10] Due to privacy considerations, we do not have the point coordinates of each teenager's residence. We only have the TAZ of residence of each teenager. Thus, for two teenagers in the same zone, we assigned a distance that was one-half of the distance between that zone and its closest neighboring zone.

physical activity are much improved in the SCHBL relative to the ABL model. This may be attributed to the increased efficiency from accounting for spatial correlation among teenagers' physical activity choices in the SCHBL model.

In the next few sections, we discuss the estimation results by variable category, discussing differences in the results among the ABL and SCBHL models.

5.2.1 Individual Demographics

The effect of individual demographics indicates that, among teenagers, males are more likely to participate in physical activity than females, a result also found in other public health studies (see, for example, Mhuircheartaigh, 1999). This variable is statistically significant at beyond the 5% level in the SCHBL model, but not in the ABL model. The race variables suggest that Caucasian and Hispanic teenagers have a higher propensity to partake in physical activity relative to African Americans, Asians, and other ethnic groups (see Sallis *et al.*, 2000 and Sener *et al.*, 2008, for similar results). Both the race variables are statistically significant at the 10% level in the SCBHL model, though this is true only for the Caucasian variable in the ABL model.

The final variable in the category of individual demographics indicates that teenagers with a driver's license are more likely (than those without a driver's license) to participate in physical activity, presumably due to less dependency on others to access physical activity opportunity locations. Perhaps, an appropriate policy strategy to increase the physical activity participation among non-driving teenagers would be to improve accessibility to activity opportunity centers and/or natural parks, as well as improving the neighborhood in terms of walking/cycling facilities (see Hoefer *et al.*, 2001 and Sjloie and Thuen, 2002). Interestingly, we also considered age variables in the specification to examine if physical activity participation among teenagers drops off with age as suggested by some earlier studies, but did not find any such statistically significant effects.

5.2.2 Household Demographics

The household demographic variable effects reflect the higher prevalence of physical activity participation among teenagers living in large families, possibly due to increased opportunities for joint physical activities with siblings and/or parents. We also examined if the type of household members (*i.e.*, number of children, active adults, and senior adults) had an impact, but the results suggested no differential impacts among these categories of household members after controlling for the household structure effects, as discussed next. Teenagers living in single parent households are more likely to be physically active compared to those in other household structure types (nuclear families, roommate families, and joint families with several adults), a result that needs to be further explored. Finally, the coefficient on the "Presence of bicycle" variable has opposite signs in the ABL and SCHBL models, with the sign being intuitive in the SCHBL model though significant only at the 20% level. The corresponding result from the ABL model is counter-intuitive, though also insignificantly different from zero. Interestingly, we did not find any statistically significant effects of household income and household car ownership on the physical activity levels of teenagers.

5.2.3 Household Location and Season Variables

The next set of variables in Table 1 indicates the impact of household location and season variables. The results show a higher tendency to pursue physical activity among teenagers residing in San Francisco county compared to the rest of the counties in the region (*i.e*. San Mateo, Santa Clara, Alameda, Contra Costa, Solano, Napa, Sonoma, and Marin counties), though this effect is significant only at the 20% level in the SCHBL model. The seasonal variables imply that the summer and fall seasons correspond to higher levels of physical activity participation among teenagers in the San Francisco Bay area (see Sener and Bhat, 2007 for similar seasonal variations). This may be a result of more time availability in the summer, and temperate weather conditions in the summer and fall, for physically active outdoor pursuits.

5.2.4 Zonal Structure, Density, and Race Composition Variables

In addition to the demographic, location, and season-related variables, the built environment and race composition measures identified in Section 4.2.2 were also considered in the model specification. Many of these variables did not turn out to be statistically significant at the 15% level

or lower in the SCHBL model, and hence do not appear in Table 1. The insignificance of several variables in this category can be attributed to the high correlation among these variables.[11]

The effect of the zonal structure and density variables in Table 1 show that there is higher propensity to participate in physical activity among teenagers residing in zones with a high share of multi-family units, though this effect is tempered if the teenager is in a high household density neighborhood. The first result may be a reflection of more opportunities for joint physical activity participation with peers and other individuals in neighborhoods with a high share of multi-family units, while the second result is perhaps related to the lack of open space in areas of high household density that may discourage physical activity (see Copperman and Bhat, 2007a).

Among the zonal ethnic composition measures, the negative sign on the "fraction of African-American population" reveals that teenagers living in an area with a high percentage of African-American population are less likely to participate in physical activity relative to teenagers in other areas. Gordon-Larsen *et al.* (2005; 2006) also find similar results and suggest that this is due to poor neighborhood quality, and lack of good recreational facilities in areas with a high fraction of African-American population. Our results also indicate higher levels of physical activity participation among teenagers residing in highly populated Asian areas, though this result is not very statistically significant even in the SCHBL model. Interestingly, the race composition variables are not at all significant in the ABL model.

5.2.5 Zonal Activity Opportunity, Housing Cost, and Transportation Network Variables

As expected, the number of physically active recreation centers such as fitness centers, sports centers, dance, and yoga studios in a zone has a positive influence on the physical activity levels of teenagers residing in that zone, indicating that, as suggested by Trost *et al.* (1997), an effective policy to increase physical activity among teenagers would be to facilitate more opportunities for community-based physical activity outlets including sport/fitness programs, summer parks, and recreation programs.[12] The zonal housing cost variable has a positive impact on the physical activity

[11] The following discussion of the effects of built environment measures should be viewed with some caution, since we have not considered potential residential self-selection effects. That is, it is possible that highly physically active families self-select themselves into zones with built environment measures that support their active lifestyles (see Bhat and Guo, 2007 and Bhat and Eluru, 2008 for methodologies to accommodate such self-selection effects).

[12] In addition to the number of activity opportunities, we also considered the presence of activity opportunities in the zone as well as accessibility to shopping, recreation, and employment opportunities. But these variables did not turn out to be statistically significant after including the number of opportunities.

levels of teenagers, perhaps because of better recreational opportunities in such zones (Lacar *et al.*, 2000 and Gordon-Larsen *et al.*, 2005 also find a similar result). Finally, the transportation network measure effects emphasize the positive influence of good bicycle facilities and walk/bicycle accessibility to activity opportunities on physical activity levels. This suggests that urban and transportation planners should consider the provision of well designed bicycle paths, and the design of dense, mixed land-use, neighborhoods as a means to increase physical activity levels of teenagers (Krizek *et al.*, 2004 also discuss the importance of community design on the physical activity participation of the youth). The reader will note that the bicycle density variable is not significant in the ABL model, while it is statistically significant at about the 5% level of significance in the SCBHL model.

5.3. Heteroscedasticity, Spatial Dependency, and Data Fit

This section presents the parameter estimates characterizing heteroscedasticity and spatial correlation in the SCHBL model, and discusses data fit measures for the ABL and SCHBL models.

5.3.1 Heteroscedasticity

The SCBHL model accommodates heteroscedasticity in the variance of the error term across individuals due to both individual attributes as well as spatial residence zone attributes. In the current application, only three variables turned out to be statistically significant in influencing heteroscedasticity, including two household attributes and one built environment measure. Note that the estimates reported in the table correspond to the λ vector in Equation (13). The results indicate a much tighter variation (*i.e.*, less spread) in the propensity to be physically active among teenagers who (1) live in single parent households, (2) belong to households owning a bicycle, and (3) live in areas with a high share of multi-family units. Specifically, the scale parameter is normalized to 1 for teenagers in non-single parent family households with no bicycles in the household and living in single-family dwelling units, but is statistically significantly smaller for other teenagers. For instance, if we consider the pool of teenagers in a single parent family with no bicycles in the household and living in a zone with no multi-family dwelling units, the scale parameter estimate is exp(–2.177) = 0.11 (with a standard error estimate of 0.06), while the corresponding value for teenagers in a non-single parent family with bicycles in the household and living in a zone with no multi-family dwelling units is exp(–0.305) = 0.73 (with a standard error estimate of 0.18).

In combination with the direct positive influence of the single parent, presence of bicycle, and fraction of multi-family dwelling unit variables on the physical activity propensity, as discussed in earlier sections, the overall implication is that teenagers in single parent households with a bicycle, and living in areas with a high share of multi-family units, are uniformly more likely to participate in physical activity. The estimates of the ABL model, which ignores such heteroscedasticity, are therefore inconsistent.

5.3.2 Spatial Correlation Effects

In addition to heteroscedasticity, the SCBHL model also incorporates spatial correlation across observational units. In this regard, several spatial variables and functional forms of these variables were considered to accommodate spatial correlation across teenagers' propensity to participate in physical activity (see Section 3). The best specification included a single "inverse of distance" variable in the s_{qk} vector of Equation (12). The corresponding δ coefficient on this variable is reported in Table 1, and has a value of 3.862 (with a standard error estimate of 2.13). Given the range of the distance between teenagers' residences in the sample, this results in θ_{qk} values (see Equation (12)) that range from 0.252 (for two teenagers located 141 miles apart) to 0.997 (for two teenagers located 0.14 miles apart), with a mean value of 0.590. The corresponding range of Spearman's correlation between the physical activity propensities of teenagers is from 0.075 to 0.303, with a mean correlation coefficient of 0.180. Table 2 provides the estimated θ_{qk} and correlation coefficient values for pairs of teenagers located (in terms of their residences) at different distances from each other. The table also provides the share of teenager pairs in the sample for each distance value. The asymptotic t-statistics for θ_{qk} and the correlation coefficient for each distance value are computed for the null hypothesis of no dependence (*i.e.*, $\theta_{qk} = 0$ and the correlation is zero).[13] The results show, as expected, the correlation decay with distance. As importantly, the correlation values are statistically significant at about the 10% level (for a one-tailed t-test) up to a distance of about 35 miles. The share of teenage pairings within 35 miles is 50.6%, indicating that spatial correlation is present and statistically significant for a high fraction of teenage pairings. The

[13] The standard errors for θ_{qk} and the correlation coefficient are computed based on the standard error of the estimated δ coefficient using the familiar "delta" method for the asymptotic distribution of a nonlinear function (see Greene, 2000,

aspatial binary logit (ABL) model ignores these spatial correlations, and so it is not surprising that the spatially correlated heteroscedastic binary logit (SCHBL) model provides more efficient parameter estimates (as discussed in Sections 5.2.1 through 5.2.5).

5.3.3 Overall Likelihood-Based Measures of Fit

The data fit of the ABL and SCHBL models may be compared formally using likelihood ratio tests. The log-likelihood value at convergence for the ABL model is –318.3, while that for the SCHBL model is –308.3. The likelihood ratio test for testing between these two models is 20.1, which is larger than the critical χ^2 value with 4 degrees of freedom at any reasonable level of significance, confirming the importance of accommodating heteroscedasticity and spatial correlation when modeling physical activity participation choice of teenagers. Further, the log-likelihood value for the model with only heteroscedasticity is –310.8. The likelihood ratio between this model and the ABL model yields a value of 15, which is again larger than the critical χ^2 value with 3 degrees of freedom at any reasonable level of significance, rejecting the null hypothesis of absence of heteroscedasticity. Finally, a comparison of the SCHBL model and the model with only heteroscedasticity yields a likelihood ratio test value of 5, indicating that the null hypothesis of the absence of spatial correlation can be rejected at the 2.5% level of significance. Overall, the likelihood ratio tests show that heteroscedasticity and spatial correlation are both individually and jointly statistically significant.

5.4 Aggregate-Level Elasticity Effects

The parameters on the exogenous variables in Table 1 do not directly provide the magnitude of the effects of the variables on the probability of teenagers' physical activity participation. Further, the parameters in the ABL and SCHBL models cannot be directly compared because the SCHBL model allows heteroscedasticity across individuals. To address these issues, we compute the aggregate-level "elasticity effects" of each variable. In particular, to compute the aggregate-level elasticity of a dummy exogenous variable (such as the "male" variable), we change the value of the variable to one for the subsample of observations for which the variable takes a value of zero and to zero for the subsample of observations for which the variable takes a value of one. We then sum the shifts in

---

page 120). The asymptotic t-statistics for θ_{qk} and the correlation coefficient are identical because the correlation coefficient is a simple linear function of θ_{qk}.

expected aggregate shares in the two subsamples after reversing the sign of the shifts in the second subsample, and compute an effective percentage change in the expected aggregate share of teenagers participating in physical activity due to a change in the dummy variable from 0 to 1. On the other hand, to compute the aggregate level elasticity effect of an ordinal variable (such as household size), we increase the value of the variable by 1 and compute a percentage change in the expected aggregate share of teenagers participating in physical activity. Finally, the aggregate-level "arc" elasticity effect of a continuous exogenous variable (such as fraction of multi-family dwelling unit) is obtained by increasing the value of the corresponding variable by 10% for each individual in the sample, and computing a percentage change in the expected aggregate share of teenagers participating in physical activity due to the increase in the continuous variable.

The elasticity effects by variable category and for both the ABL and SCHBL models are presented in Table 3. The numbers in the table may be interpreted as the percentage change in the share of teenagers participating in physical activity. For instance, the first number "12.93" corresponding to the "male" variable in the ABL model indicates that the share of male teenagers participating in physical activity is about 13% higher than the share of female teenagers participating in physical activity. Similarly, the number "34.07" corresponding to the "household size" variable in the ABL model reflects that an increase in household size by 1 leads to a 34% increase in the level of teenager participation in physical activity, while the number "1.33%" for the effect of the "zonal fraction of multi-family dwelling units" implies that teenager physical activity participation levels increase by 1.33% due to a 10% increase in the zonal fraction of multi-family dwelling units.

The elasticity results provide several insights. First, among the demographics and locational/seasonal variables, seasonality (whether the season is fall or not) constitutes the most important factor influencing teenagers' physical activity participation levels, followed by family structure (whether the teenager is part of a single parent family or not). This suggests that public health policies aimed at encouraging year-round teenager physical activity participation should focus on providing more indoor recreational activity opportunities at affordable cost during the non-fall seasons in general, and the winter and spring seasons in particular. Also, the result related to family structure needs further exploration. Perhaps teenagers in single parent families are more independent and have a less-structured activity schedule because of the multiple responsibilities shouldered by single parents. Conversely, it may be that teenagers in two-parent (and other non-single parent) households have less independence and feel more "monitored", contributing to less opportunity for

physically active free play. Overall, there is a suggestion that physical activity participation may be related to independence and empowerment within the household. The behavioral dynamics of interpersonal interactions and how they manifest themselves in physical activity participation is a research topic that should benefit from joint collaborative research efforts by sociologists, public health researchers, and transportation professionals.

Second, among the built environment measures, the two major factors determining teenagers' physical activity participation levels are zonal household density and the number of zones accessible within four non-motorized mode miles, suggesting that policies to provide planned open spaces in high density neighborhoods and increased accessibility by walk/bicycle modes have the potential to increase teenager physical activity levels. However, teenager physical activity levels appear to be quite inelastic to built environment changes.

Third, there are differences in the elasticity effects between the ABL and SCHBL models. This, combined with the better data fit of the SCHBL model, points to the inconsistent elasticity effects from the ABL model. For instance, the ABL model underestimates gender differences and family structure differences in physical activity participation levels among teenagers, and overestimates the impact of residing in San Francisco County. Further, the ABL model predicts that teenager physical activity participation levels in the pool of households owning bicycles is 15% less than that in the pool of households not owning bicycles. This result is rather unintuitive. The SCHBL model, on the other hand, shows a marginally higher level of physical activity participation among teenagers in households with bicycles. Thus, ignoring spatial effects, when present, can lead to inconsistent estimation of variable effects that, in turn, can lead to misinformed policy actions and recommendations.

6. CONCLUSIONS

Spatial dependence is rather ubiquitous in many choice decisions in geography, transportation, political science, economics, and other social sciences. The current methods to address such dependence range from (1) ignoring the dependence entirely to (2) sampling from the data systematically to reduce the dependence among observations to (3) accommodating the full spatial dependency using simulation techniques. The first approach is known to produce inconsistent and inefficient estimates in the presence of spatial heteroscedasticity and correlation. The second

approach is inefficient because it reduces the size of the available data for estimation. The third approach considers the full spatial specification explicitly, but the techniques are too computationally intensive and not feasible for sample sizes of the type frequently encountered in practice.

In the current paper, we propose a new copula-based approach that adopts the full spatial specification approach. In contrast to current full spatial specification methods, our approach is based on a spatial logit structure rather than a spatial probit structure. The dependence between the logistic error terms of different observational units is directly accommodated using a multivariate logistic distribution based on the Farlie-Gumbel-Morgenstein (FGM) copula. The approach represents a simple, powerful, technique that results in a closed-form analytic expression for the joint probability of choice across observational units, and is straightforward to apply using a standard and direct maximum likelihood inference procedure. There is no simulation machinery involved, leading to substantial computation gains relative to current methods. The method is computationally tractable even for a high number of observational units. In addition to computational efficiency gains, there is another more basic reason to prefer the closed-form copula-based spatial logit model over the extant spatial probit model. This is related to the fact that closed-form analytic structures should be used whenever feasible, because they are always more accurate than the simulation evaluation of analytically intractable structures (see Train, 2003; pg. 191). Of course, one issue with our spatial logit approach is that the correlation between observations is limited to moderate levels. However, in the typical context of spatial structure-based dependence, where the correlation between observational units drops off sharply with geographic distance, the correlation range of the FGM logistic distribution is not likely to be too limiting.

In the current paper, we demonstrate an application of the model to teenagers' physical activity participation levels, a subject that is of considerable interest in the public health, transportation, sociology, and adolescence development fields. The data for the analysis is drawn from the 2000 San Francisco Bay Area Survey, supplemented with several other sources to obtain measures of built environment determinants. The empirical results indicate the important effects of individual demographics (gender, race, driver license holding), household demographics (family structure, household size and presence of bicycle), household location (whether teenager residence is in San Francisco County or not), and season of year. Physical environment variables are also

statistically significant determinants of teenagers' physical activity levels, though these variable effects are inelastic.

A comparison of the aspatial binary logit (ABL) model and the spatially correlated heteroscedastic binary logit (SCHBL) model proposed in the paper indicates the significant presence of heteroscedasticity across observations and spatial correlation between teenager pairs. The ABL model, which ignores these effects, provides inconsistent and inefficient parameter estimates. The SCHBL model also leads to a statistically superior data fit. In addition, the results indicate that failing to accommodate heteroscedasticity and spatial correlation can lead to incorrect conclusions regarding the elasticity effects of exogenous variables.

To summarize, this paper introduces a copula-based approach to addressing spatial dependency and heteroscedasticity issues in binary choice models. The study, to our knowledge, represents the first formulation and application of such an approach for spatial analysis, and highlights the power of closed-form techniques for accommodating spatial effects. The authors are currently focusing on extending the approach to consider spatial effects in unordered multinomial choice models and exploring the use of other more flexible copula structures for incorporating spatial effects.

ACKNOWLEDGEMENTS

This research was funded in part by Environmental Protection Agency Grant R831837. The authors are grateful to Lisa Macias for her help in formatting this document.

REFERENCES

Aaron, D.J., K.L. Storti, R.J. Robertson, A.M. Kriska, and R.E. LaPorte (2002) Longitudinal Study of the Number and Choice of Leisure Time Physical Activities from Mid to Late Adolescence. *Archives of Pediatric and Adolescent Medicine*, 156, 1075-1080.

Anselin, L., and D.A. Griffith (1988) Do Spatial Effects Really Matter in Regression Analysis? *Papers of the Regional Science Association*, 65, 11-34.

Anselin, L. (2003) Spatial Externalities, Spatial Multipliers and Spatial Econometrics. *International Regional Science Review*, 26, 153 – 166.

Armstrong, M., and A. Galli (2002) Copulas. Presented at the *SPE ATW Risk Analysis Applied to Field Development Under Uncertainty*, Rio de Janeiro, Brazil, 29-30 August 2002.

Beron, K.J., J.C. Murdoch, and W.P.M. Vijverberg (2003) Why Cooperate?Public Goods, Economic Power, and the Montreal Protocol. *Review of Economics and Statistics,* 85(2), 86-97.

Beron, K.J., and W.P.M. Vijverberg (2004) Probit in a Spatial Context: A Monte Carlo Analysis. In *Advances in Spatial Econometrics: Methodology, Tools and Applications*, L. Anselin, R.J.G.M. Florax, and S.J. Rey (eds.), Springer-Verlag, Berlin.

Bhat, C.R. (2000) A Multi-Level Cross-Classified Model for Discrete Response Variables. *Transportation Research Part B*, 34(7), 567-582.

Bhat, C.R., and H. Zhao (2002) The Spatial Analysis of Activity Stop Generation. *Transportation Research Part B*, 36(6), 557-575.

Bhat, C.R., and J.Y. Guo (2004) A Mixed Spatially Correlated Logit Model: Formulation and Application to Residential Choice Modeling. *Transportation Research Part B*, 38(2), 147-168.

Bhat, C. R., and J. Y. Guo (2007) A Comprehensive Analysis of Built Environment Characteristics on Household Residential Choice and Auto Ownership Levels. *Transportation Research Part B*, 41(5), 506-526.

Bhat, C. R., and N. Eluru (2008) A Copula-Based Approach to Accommodate Residential Self-Selection in Travel Behavior Modeling. Technical Paper, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin. Available at: http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/sample_selection_16June08.pdf

Bolduc, D., B. Fortin, and M. Fournier (1996) The Effect of Incentive Policies on the Practice Location of Doctors: A Multinomial Probit Analysis. *Journal of Labor Economics*, 14(4), 703-732.

Bolduc, D., B. Fortin, and S. Gordon (1997) Multinomial Probit Estimation of Spatially Interdependent Choices: An Empirical Comparison of Two New Techniques. *International Regional Science Review*, 20(1 & 2), 77-101.

Cambanis, S. (1977) Some Properties and Generalizations of Multivariate Eyraud-Farlie-Gumbel-Morgenstern Distributions. *Journal of Multivariate Analysis,* 7, 551-559.

Cameron, A.C., T. Li, P. Trivedi, and D. Zimmer (2004) Modelling the Differences in Counted Outcomes Using Bivariate Copula Models with Application to Mismeasured Counts. *The Econometrics Journal*, 7(2), 566–584.

Case, A. (1992) Neighborhood Influence and Technological Change. *Economics*, 22, 491-508.

Casella, G. and E. I. George (1992) Explaining the Gibbs Sampler. *The American Statistician*, 46(3), 167-174.

Center for Disease Control (CDC) (2002) Youth Risk Behavior Surveillance – United States, 2001. Morbidity and Mortality Weekly Report Surveillance Summaries, 51(SS-4).

Center for Disease Control (CDC) (2006) Youth Risk Behavior Surveillance – United States, 2005. Morbidity and Mortality Weekly Report, 55(SS-5).

Cherubini, U., E. Luciano and W. Vecchiato (2004) *Copula Methods in Finance*. John Wiley & Sons, Hoboken, NJ.

Conway, D.A. (1983) Farlie-Gumbel-Morgenstern Distributions. *Encyclopedia of Statistical Sciences*, 7, 28-31.

Copperman, R.B., and C. R. Bhat (2007a) An Analysis of the Determinants of Children's Weekend Physical Activity Participation. *Transportation,* 34(1), 67-87.

Copperman, R.B., and C.R. Bhat (2007b) An Exploratory Analysis of Children's Daily Time-Use and Activity Patterns Using the Child Development Supplement (CDS) to the US Panel Study of Income Dynamics (PSID). *Transportation Research Record,* 2021, 36-44.

Dong, L., G. Block, and S. Mandel (2004) Activities Contributing to Total Energy Expenditure in the United States: Results from the NHAPS Study. *International Journal of Behavioral Nutrition and Physical Activity*, 1(1): 4.

Dugundji, E.R. and J.L. Walker (2005) Discrete Choice with Social and Spatial Network Interdependencies. *Transportation Research Record*, 1921, 70-78.

Embrechts, P., F. Lindskog and A. McNeil (2003) Modelling Dependence with Copulas and Applications to Risk Management. In *Handbook of Heavy Tailed Distributions in Finance*, S. Rachev (ed.), Elsevier.

Feldman, D.E., T. Barnett, I. Shrier, M. Rossigni, and L. Abenhaim (2003) Is Physical Activity Differentially Associated with Different Types of Sedentary Pursuits? *Archives of Pediatric and Adolescent Medicine*, 157, 797-802.

Fleming, M. (2004) Techniques for Estimating Spatially Dependent Discrete Choice Models. In *Advances in Spatial Econometrics*, R. Florax and L. Anselin (eds.), Springer, Berlin.

Fotheringham, A. S. (1983) Some Theoretical Aspects of Destination Choice and Their Relevance to Production-Constrained Gravity Models. *Environment and Planning*, 15(8), 1121-1132.

Fotheringham, A. S., M. E. Charlton, and C. Brunsdon (1996) The Geography of Parameter Space: An Investigation into Spatial Nonstationarity. *International Journal of GIS*, 10, 605-627.

Fotheringham A. S., M. E. Charlton and C. Brunsdon (1997) Two Techniques for Exploring Nonstationarity in Geographical Data. *Geographical Systems*, 4, 59-82.

Fotheringham, A. S., and C. Brunsdon (1999) Local Forms of Spatial Analysis. *Geographical Analysis*, 31(4), 340-358.

Franzese, R.J. and J.C. Hays (2007) Empirical Models of Spatial Interdependence. In *Oxford Handbook of Political Methodology*, J. Box-Steffensmeier, H. Brady, and D. Collier (eds.), forthcoming.

Garrido, R.A., and H.S. Mahmassani (2000) Forecasting Freight Transportation Demand with the Space-time Multinomial Probit Model. *Transportation Research Part B*, 34(5), 403-418.

Gordon-Larsen, P., R.G. McMurray, and B.M. Popkin (2005) Determinants of Adolescent Physical Activity and Inactivity Patterns. *Pediatrics*, 105(6), E83.

Gordon-Larsen, P., M. Nelson, P. Page, and B.M. Popkin (2006) Inequality in the Built Environment Underlies Key Health Disparities in Physical Activity and Obesity. *Pediatrics*, 117(2), 417-424.

Goulias, K.G., and T. Kim (2005). An Analysis of Activity Type Classification and Issues Related to the With Whom and For Whom Questions of an Activity Diary. Presented at the *84th Annual Meeting of the Transportation Research Board*, Washington, D.C., January.

Government Accountability Office (GAO) (2006) Childhood Obesity: Factors Affecting Physical Activity. GAO-07-260R Childhood Obesity and Physical Activity. Washington, D.C.

Griffith, D., and L. Layne (1999) *A Casebook for Spatial Statistical Data Analysis*, Oxford University Press, New York.

Greene, W.H. (2000) Econometric Analysis, Fourth Edition. Prentice-Hall Inc., New Jersey.

Gumbel, E., (1961) Bivariate Logistic Distributions. *American Statistical Association Journal*, 56(294), 335-349.

Handy, S.L. (2004) Community Design and Physical Activity: What do We Know? - and What DON'T We Know? Presented at the National Institute of Environmental Health Sciences conference on "Obesity and the Built Environment: Improving Public Health through Community Design," Washington, D.C., May 2004.

Hoefer, W.R., R.D. Thomas, L. McKenzie, J.F. Sallis, S.J. Marshall, and T.L. Conway (2001) Parental Provision of Transportation for Adolescent Physical Activity. *American Journal of Preventive Medicine*, 21(1), 48-51.

Hunt, L.M., B. Boots, and P.S. Kanaroglou (2004) Spatial Choice Modelling: New Opportunities to Incorporate Space into Substitution Patterns. *Progress in Human Geography*, 28(6), 746-766.

Jones, K., and N. Bullen (1994) Contextual Models of Urban Home Prices: A Comparison of Fixed and Random Coefficient Models Developed by Expansion. *Economic Geography*, 70, 252-272.

Junker, M., and A. May (2005) Measurement of Aggregate Risk with Copulas. *The Econometrics Journal*, 8(3), 428–454.

Karunaratne, P. M., and R. C. Elston (1998) A Multivariate Logistic Model (MLM) for Analyzing Binary Family Data. *American Journal of Medical Genetics Part A*, 76(5), 428-437.

Kelejian, H., and I. R. Prucha (1999) A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International Economic Review*, 40(2), 509-533.

Kelly, L.A., J.J. Reilly, A. Fisher, C. Montgomery, A. Williamson, J.H. McColl, J.Y. Paton., and S. Grant (2006). Effect of Socioeconomic Status on Objectively Measured Physical Activity. *Arch. Dis. Child*. 91: 35–38.

Klier, T., and D. P. McMillen (2007) Clustering of Auto Supplier Plants in the U.S.: GMM Spatial Logit for Large Samples. Available at: http://tigger.uic.edu/~mcmillen/papers/Clustering%20of%20Auto%20Supplier%20Plants%20in%20the%20US,%20revision.pdf

Kotz, S., N. Balakrishnan, and N.L. Johnson (2000) *Continuous Multivariate Distributions*, *Vol. 1, Models and Applications,* 2nd edition. John Wiley & Sons, New York.

Krizek K., A. Birnbaum, and D. Levinson (2004) A Schematic for Focusing on Youth in Investigation of Community Design and Physical Activity. *American Journal of Health Promotion*, 19(1), 33-38.

Lacar, E.S., X. Soto, and W.J. Riley (2000) Adolescent Obesity in a Low-income Mexican-American District in South Texas. *Archives of Pediatrics & Adolescent Medicine*, 154(8), 837-840.

LeSage, J.P. (2000) Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models. *Geographical Analysis*, 32(1), 19-35.

McMillen, D.P. (1992) Probit with Spatial Autocorrelation. *Journal of Regional Science*, 32, 335-348.

McMillen, D.P. (1995) Spatial Effects in Probit Models: A Monte Carlo Investigation. In *New Directions in Spatial Econometrics*, pp. 189-228, L. Anselin and R. Florax (eds.) Springer-Verlag, Heidelberg.

Mhuircheartaigh, J.N. (1999) Participation in Sport and Physical Activities among Secondary School Students. Department of Public Health, Western Heath Board.

Miller, H.J. (1999) Potential Contributions of Spatial Analysis to Geographic Information Systems for Transportation (GIS-T). *Geographical Analysis*, 31(4), 373-399.

Miyamoto, K., V. Vichiensan, N. Shimomura, and A. Páez (2004) Discrete Choice Model with Structuralized Spatial Effects for Location Analysis. Presented at the *83rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January.

Morgenstern, D. (1956), Einfache Beispiele Zweidimensionaler Verteilungen. *Mitteilingsblatt fur Mathematische Statistik*, 8, 234-235.

MORPACE International, Inc. (2002) Bay Area Travel Survey Final Report.

Nelsen, R. B. (2006) *An Introduction to Copulas* (2nd ed.), Springer-Verlag, New York.

Nelson, M. C., and P. Gordon-Larsen (2006) Physical Activity and Sedentary Behavior Patterns are Associated with Selected Adolescent Heath Risk Behaviors. *Pediatrics*, 117(4), 1281-1290.

Ornelas, I.J., K.M. Perreira, and G.X. Ayala (2007) Parental Influences on Adolescent Activity: A Longitudinal Study. *The International Journal of Behavioral Nutrition and Physical Activity*, 4:3.

Páez, A., and D. Scott (2004) Spatial Statistics for Urban Analysis: A Review of Techniques with Examples. *GeoJournal*, 61(1), 53-67.

Páez, A. (2006) Exploring Contextual Variations in Land Use and Transport Analysis Using a Probit Model with Geographical Weights, *Journal of Transport Geography*, 14, 167-176.

Páez, A. (2007) Spatial Perspectives on Urban Systems: Developments and Directions. *Journal of Geographic Systems*, 9, 1-6.

Pinkse, J., and M.E. Slade (1998) Contracting in Space: An Application of Spatial Statistics to Discrete-Choice Models. *Journal of Econometrics*, 85(1), 125-154.

Prieger, J.E. (2002) A Flexible Parametric Selection Model for Non-Normal Data with Application to Health Care Usage. *Journal of Applied Econometrics*, 17(4), 367-392.

Reisner, E. (2003) Understanding Family Travel Demands as a Critical Component in Work-family Research, Transportation and Land-use. Presented at *From 9 to 5 to 24/7: How Workplace Changes Impact Families, Work and Communities, Academic Work and Family Research Conference*, March.

Quinn, C. (2007), The Health-economic Applications of Copulas: Methods in Applied Econometric Research. HEDG Working Papers 07/22, University of York.

Sallis, J. F., J.J. Prochaska, and W.C. Taylor (2000) A Review of Correlates of Physical Activity of Children and Adolescents. *Medicine & Science in Sports & Exercise*, 32(5), 963-75.

Salmon, J., M.L. Booth, P. Phongsavan, N. Murphy, and A. Timperio (2007). Promoting Physicla Activity Particiaption among Children and Adolescents. *Epidemiologic Reviews*, 29: 144-159.

Schmidt, T. (2007) Coping with Copulas. In *Copulas - From Theory to Application in Finance*, J. Rank (ed.), 3-34, Risk Books, London.

Schweizer, B., and A. Sklar (1983) *Probabilistic Metric Spaces*. North-Holland, New York.

Sener, I.N., and C.R. Bhat (2007) An Analysis of the Social Context of Children's Weekend Discretionary Activity Participation. *Transportation*, 34(6), 697-721.

Sener, I.N., R.B. Copperman, R.M. Pendyala, and C.R. Bhat (2008) An Analysis of Children's Leisure Activity Engagement: Examining the Day of Week, Location, Physical Activity Level, and Fixity Dimensions. Forthcoming, *Transportation*.

Sjloie, A.N., and F. Thuen (2002) School Journeys and Leisure Activities in Rural and Urban Adolescents in Norway. *Health Promotion International*, 17(1), 21-30.

Sklar, A. (1959) Fonctions de Répartition à *n* Dimensions et Leurs Marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8, 229-231.

Sklar, A. (1973) Random Variables, Joint Distribution Functions, and Copulas. *Kybernetika*, 9, 449-460.

Smith, M. (2005) Using Copulas to Model Switching Regimes with an Application to Child Labour. *Economic Record*, 81, S47-S57.

The National Health & Lifestyle Surveys (2003) Regional Results of the National Health & Lifestyle Surveys SLÁN (Survey of Lifestyle, Attitudes & Nutrition) & HBSC (Health Behavior in School Aged Children). Health Promotion Unit.

Train, K. (2003) *Discrete Choice Methods with Simulation*, Cambridge University Press.

Transportation Research Board and Institute of Medicine (2005) Does the Built Environment Influence Physical Activity? Examining the Evidence. TRB Special Report 282, National Research Council, Washington, D.C.

Trivedi, P. K. and D. M. Zimmer (2007) Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1), Now Publishers.

Trost, S. G., R. R. Pate, R. Saunders, D. Ward, M. Dowda, and G. Felton (1997) A Prospective Study of the Determinants of Physical Activity in rural Fifth-Grade Children. *Preventive Medicine*, 26, 257-263.

United States Department of Health and Human Services (USDHHS) (1996) Physical Activity and Health: A Report of the Surgeon General. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Heath Promotion, Atlanta, GA.

United States Department of Health and Human Services (USDHHS) (2000) Healthy People 2010: Understanding and Improving Health, 2nd ed. U.S. Government Printing Office, Washington, D.C., November.

Warburton, D.E.R., C.W. Nicol, and S.S.D. Bredin (2006) Health Benefits of Physical Activity: the Evidence. CMAJ-*Canada's Leading Medical Journal*, 174(6), 801-9.

Zimmer, D. M., and P. K. Trivedi (2006) Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand. *Journal of Business and Economic Statistics*, 24, 63-76.

LIST OF TABLES

Table 1. Estimation Results for Teenagers' Weekday Physical Activity Participation Choice

Variables	(Aspatial) Binary Logit Model		Copula-based Spatially Correlated and Heteroscedastic Model	
	Parameter	**t-stat**	**Parameter**	**t-stat**
Constant	-5.534	-7.45	-3.211	-3.56
Individual demographics				
Male	0.238	1.18	0.259	2.22
Caucasian	0.722	2.42	0.320	1.82
Hispanic	0.457	0.95	0.336	1.72
Driver's license	0.661	3.02	0.309	2.23
Household demographics				
Household size	0.562	5.40	0.275	2.73
Single parent family	1.264	2.95	1.070	3.35
Presence of bicycle	-0.266	-0.93	0.168	1.35
Household location and season variables				
San Francisco County	1.309	1.84	0.341	1.36
Summer	0.816	3.94	0.450	3.28
Fall	4.265	8.47	2.459	3.37
Physical environment measures				
Zonal structure, density, and race composition variables				
Fraction of multi-family dwelling units	1.100	1.57	0.883	2.66
Household density	-0.308	-3.54	-0.155	-3.32
Fraction of African-American population	-1.299	-0.59	-2.379	-1.65
Fraction of Asian population	0.152	0.17	0.459	1.49

Table 1 (Continued). Estimation Results for Teenagers' Weekday Physical Activity Participation Choice

Variables	(Aspatial) Binary Logit Model		Copula-based Spatially Correlated and Heteroscedastic Model	
	Parameter	**t-stat**	**Parameter**	**t-stat**
Zonal activity opportunity, housing cost, and transportation network variables				
Number of physically active recreation centers such as fitness centers, sports centers, dance and yoga studios	0.031	1.07	0.021	1.68
Average of median housing value	0.132	2.09	0.052	1.67
Bicycling facility density (miles of bike lanes per square mile)	0.033	0.66	0.035	1.89
Number of zones within 4 non-motorized mode miles[14]	0.032	2.16	0.019	2.54
(Spatial) heteroscedasticity variables				
Single parent family	---	---	-2.177	-3.95
Presence of bicycle	---	---	-0.305	-1.23
Fraction of multi-family dwelling units	---	---	-0.982	-2.02
Spatial correlation variables (δ) in the θ parameter				
Inverse of distance between zonal centroids	---	---	3.862	1.81
Number of Observations	722		722	
Log-likelihood at convergence	-318.323		-308.273	

[14] The non-motorized distance between zones, which is computed based on the actual bike/walk way, was used in the empirical analysis to develop an accessibility measure by non-motorized modes, computed as the number of zones (a proxy for activity opportunities) within "x" non-motorized mode miles of the teenager's residence zone. Several variables with different thresholds for "x" were formulated and tested.

Table 2. Spatial Correlation Patterns in Physical Activity Propensity

Distance between teenagers (in miles)	Cumulative share of teenagers within specified distance	θ_{qk}	Spearman's correlation coefficient	
		Estimate	Estimate	t-stat*
0.1	0.0	0.998	0.303	223.3
1.0	0.4	0.979	0.298	22.8
5.0	2.9	0.905	0.275	4.9
10.0	7.8	0.827	0.251	2.7
15.0	15.1	0.760	0.231	2.0
20.0	23.1	0.704	0.214	1.6
25.0	32.1	0.655	0.199	1.4
35.0	50.6	0.576	0.175	1.1
50.0	72.4	0.488	0.148	0.9
100.0	98.8	0.322	0.098	0.7
150.0	100.0	0.241	0.073	0.6

*The t-statistic values apply to both the θ_{qk} and the correlation coefficient estimates.

Table 3. Aggregate-level Elasticity Effects

	Formulation of the Change on the Variable	(Aspatial) Binary Logit Model	Copula-based Spatially Correlated and Heteroscedastic Model
Individual socio-demographics			
Male	changed from 0 to 1	12.93	25.90
Caucasian	changed from 0 to 1	36.33	29.63
Hispanic	changed from 0 to 1	26.86	36.72
Driver's license	changed from 0 to 1	37.85	31.61
Household socio-demographics			
Household size	increased by 1	34.07	30.72
Single parent family	changed from 0 to 1	82.22	137.63
Presence of bicycle	changed from 0 to 1	-15.02	1.10
Household location variables			
San Francisco	changed from 0 to 1	87.46	38.62
Seasonal variables			
Summer	changed from 0 to 1	45.53	47.94
Fall	changed from 0 to 1	274.37	243.60
Zonal structure, density, and race composition variables			
Fraction of multi-family dwelling units	increased by 10%	1.33	1.48
Household density	increased by 10%	-3.99	-4.69
Fraction of African-American population	increased by 10%	-0.29	-0.95
Fraction of Asian population	increased by 10%	0.13	0.90
Zonal activity opportunity, housing cost, and transportation network variables			
Number of physically active recreation centers such as fitness centers, sports centers, dance and yoga studios	increased by 10%	0.50	0.62
Average of median housing value	increased by 10%	2.99	2.33
Bicycling facility density (miles of bike lanes per square miles)	increased by 10%	0.42	1.01
Number of zones within 4 non-motorized mode miles	increased by 10%	2.82	3.63