A Flexible Spatially Dependent Discrete Choice Model: Formulation and Application to Teenagers' Weekday Recreational Activity Participation

Chandra R. Bhat*

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

Ipek N. Sener

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

Naveen Eluru
The University of Texas at Austin
Dept of Civil, Architectural & Environmental Engineering
1 University Station C1761, Austin TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
E-mail: naveeneluru@mail.utexas.edu

*corresponding author

ABSTRACT

This study proposes a simple and practical Composite Marginal Likelihood (CML) inference approach to estimate ordered-response discrete choice models with flexible copula-based spatial dependence structures across observational units. The approach is applicable to data sets of any size, provides standard error estimates for all parameters, and does not require any simulation machinery. The combined copula-CML approach proposed here should be appealing for general multivariate modeling contexts because it is simple and flexible, and is easy to implement

The ability of the CML approach to recover the parameters of a spatially ordered process is evaluated using a simulation study, which clearly points to the effectiveness of the approach. In addition, the combined copula-CML approach is applied to study the daily episode frequency of teenagers' physically active and physically inactive recreational activity participation, a subject of considerable interest in the transportation, sociology, and adolescence development fields. The data for the analysis are drawn from the 2000 San Francisco Bay Area Survey. The results highlight the value of the copula approach that separates the univariate marginal distribution form from the multivariate dependence structure, as well as underscore the need to consider spatial effects in recreational activity participation. The variable effects indicate that parents' physical activity participation constitutes the most important factor influencing teenagers' physical activity participation levels. Thus, an effective way to increase active recreation among teenagers may be to direct physical activity benefit-related information and education campaigns toward parents, perhaps at special physical education sessions at the schools of teenagers.

Keywords: Spatial econometrics, copula, composite marginal likelihood (CML) inference approach, children's activity, public health, physical activity.

1. INTRODUCTION

Spatial dependence in data may occur for several reasons, including diffusion effects, social interaction effects, or unobserved location-related effects influencing the level of the dependent variable (see Jones and Bullen, 1994; Miller, 1999). Accommodating such spatial dependence has been an active area of research in spatial statistics and spatial econometrics, and has spawned a vast literature in different application fields such as earth sciences, epidemiology, transportation, land use analysis, geography, social science and ecology (see Páez and Scott, 2004; Franzese and Hays, 2008). However, while this literature abounds in techniques to address spatial dependence in continuous dependent variable models, there has been much less research on techniques to accommodate spatial dependence in discrete choice models, as already indicated by several researchers (see Bhat, 2000; Páez, 2007). This, of course, is not because there is a dearth of application contexts of spatial dependence in discrete choice settings, but because of the estimation complications introduced by spatial interdependence in non-continuous dependent variable models.

In the past decade, several alternative approaches have been introduced that attempt to address the estimation complications of spatial dependence across observational units in discrete choice models (see Fleming, 2004). Almost all of these efforts are focused on the binary spatial probit model, which is predicated on a multivariate normality assumption to characterize the spatial dependence structure. However, an approach referred to as the "Copula" approach has recently revived interest in a whole set of alternative couplings that can allow non-linear and asymmetric dependencies. A copula is essentially a multivariate dependence structure for the joint distribution of random variables that is separated from the marginal distributions of individual random variables, and derived purely from pre-specified parametric marginal distributions of each random variable. Under the copula approach, the multivariate normal distribution adopted in almost all spatial binary choice models in the past is but one of a suite of different types of error term couplings that can be tested. In particular, since it is difficult to know *a priori* what the best structure is to characterize the distribution of the univariate observation-specific error terms, as well as the dependence between the error terms across observations, it behooves the analyst to empirically test different univariate error distributions and multivariate dependence functions rather than pre-imposing particular error distributions. The copula approach enables such testing by allowing different specifications for the univariate marginal distributions and the dependence structure (see Bhat and Eluru, 2009; Trivedi and Zimmer, 2007).

In terms of estimation of discrete choice models with a general spatial correlation structure, the analyst confronts, in the familiar probit model, a multi-dimensional integral over a multivariate normal distribution, which is of the order of the number of observational units in the data. While a number of approaches have been proposed to tackle this situation (see McMillen, 1995; LeSage, 2000; Pinske and Slade, 1998; Fleming, 2004; Beron *et al.*, 2003; Beron and Vijverberg, 2004), none of these remain practically feasible for moderate-to-large samples.[1] These methods are also quite cumbersome and involved. An approach to deal with these estimation complications in the spatial probit or other non-normal copula-based spatial models is the technique of composite marginal likelihood (CML), an emerging inference approach in the statistics field. The CML estimation approach is a simple approach that can be used when the full likelihood function is near impossible or plain infeasible to evaluate due to the underlying complex dependencies, as is the case with spatial discrete choice models. The CML approach also represents a conceptually, pedagogically, and implementationally simpler procedure relative to simulation techniques, and also has the advantage of reproducibility of results.

In the current paper, we combine a copula-based formulation with a CML estimation technique to propose a simple and practical approach to estimate ordered-response discrete choice models with spatial dependence across observational units. Our approach subsumes the familiar and extensively studied spatial binary probit model as a special case. The approach is applicable to data sets of any size, provides standard error estimates for all parameters, and does not require any simulation machinery, which is in contrast to extant spatial approaches for binary choice models. In essence, the current paper brings together two emerging areas in the statistics field – the copula approach to construct general multivariate distributions and the CML approach to estimate models with an intractable likelihood function – to develop and estimate spatial discrete choice models.

The rest of this paper is structured as follows. The next section provides an overview of copula concepts and the composite marginal likelihood estimation method. Section 3 presents the structure of the copula-based spatial ordered response model and discusses the estimation/inference approach utilized in the current paper. Section 4 focuses on a simulation study to evaluate the

[1] McMillen's EM method, LeSage's MCMC method, and Pinkse and Slade's heteroscedastic approach require the inversion and determinant computation of a square matrix of the order of the number of observational units. Most practical algorithms require $O(Q^3)$ steps to evaluate the determinant and inverse of a $Q \times Q$ matrix, which becomes prohibitive for moderate-to-large Q (see Caragea and Smith, 2006). Beron and Vijvberg's method requires the simulation of a multidimensional integral of the order of the number of observational units, which again becomes prohibitive for large Q.

performance of the CML approach. Section 5 describes the data source and sample formation procedures for an empirical application of the proposed spatial model to teenagers' recreational activity participation. Section 6 presents the corresponding empirical results. The final section summarizes the important findings from the study and concludes the paper.

2. OVERVIEW OF COPULA CONCEPTS AND THE CML METHOD

2.1. Copula Concepts

A copula is a device or function that generates a stochastic dependence relationship (*i.e.*, a multivariate distribution) 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 developing dependence among random variables (see Bhat and Eluru, 2009; Trivedi and Zimmer, 2007).

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

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

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 continuously distributed marginals. Consider K random variables $Y_1, Y_2, Y_3, \ldots, Y_K$, each with univariate continuous marginal distribution functions $F_k(y_k) = \Pr(Y_k < y_k)$, $k = 1, 2, 3, \ldots, K$. Then, by Sklar's (1973) theorem, a joint K-dimensional distribution function of the random variables with the continuous marginal distribution functions $F_k(y_k)$ can be generated as follows:

$$\begin{aligned} F(y_1, y_2, \ldots, y_K) &= \Pr(Y_1 < y_1, Y_2 < y_2, \ldots, Y_K < y_K) = \Pr(U_1 < F_1(y_1), U_2 < F_2(y_2), \ldots, U_K < F_K(y_K)) \\ &= C_\theta (F_1(y_1), F_2(y_2), \ldots, F_K(y_K)). \end{aligned} \quad (2)$$

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 (2).

Thus, given a known multivariate distribution $F(y_1, y_2, \ldots, y_K)$ with continuous and strictly increasing margins $F_k(y_k)$, the inversion method may be used to obtain a unique copula using Equation (2) (see Nelsen, 2006):[2]

$$\begin{aligned} C_\theta(u_1, u_2, \ldots, u_K) &= \Pr(U_1 < u_1, U_2 < u_2, \ldots, U_K < u_K) \\ &= \Pr(Y_1 < F^{-1}{}_1(u_1), Y_2 < F^{-1}{}_2(u_2), \ldots, Y_k < F^{-1}{}_k(u_k)) \\ &= F(F^{-1}{}_1(u_1), F^{-1}{}_2(u_2), \ldots, F^{-1}{}_k(u_k)). \end{aligned} \tag{3}$$

Once the copula is developed, one can revert to Equation (2) to develop new multivariate distributions with arbitrary univariate margins.

A rich set of bivariate copula types have been generated using inversion and other methods, including the Gaussian copula, the Farlie-Gumbel-Morgenstern (FGM) copula, and the Archimedean class of copulas (including the Clayton, Gumbel, Frank, and Joe copulas). Of these, the Gaussian and FGM copulas can be extended to more than two dimensions in a straightforward manner, allowing for differential dependence patterns among pairs of variables.[3] In fact, the multivariate normal distribution used in the spatial probit model corresponds to the Gaussian copula with univariate normal distributions. Recently, Bhat and Sener (2009) proposed the use of the FGM copula with univariate logistic distributions for spatial modeling in a binary choice context, but point out that the maximal correlation allowable between pairs of variables is 0.303.

In the current paper, we use the Gaussian and FGM copulas to formulate spatial ordered response models. This allows us to test different distributions for the individual observation-specific error terms as well as the multivariate dependence structure. For reference, the multivariate Gaussian copula takes the following form:

$$C_\theta(u_1, u_2, \ldots, u_Q) = \Phi_Q(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_Q), \theta), \tag{4}$$

where Φ_Q is the Q-dimensional standard normal cumulative distribution function (CDF) with zero mean and correlation matrix (obtained by scaling an arbitrary covariance matrix so that each component has a variance of one) whose off-diagonal elements are captured in the vector θ, and

[2] The strictly increasing nature of the margins ensures the existence of the inverse of the margins.

[3] More generally, it is possible to specify specific copula forms in two dimensions, and then examine the compatibility issues in developing multivariate copula forms with the predetermined copula forms in two dimensions as the bivariate marginals (see Aas *et al.*, 2009, Aas and Berg, 2009).

$\Phi^{-1}(.)$ is the inverse (or quantile function) of the univariate standard normal CDF. This copula collapses to the independence copula when all elements of θ take the value of zero. In the bivariate case, the Gaussian copula takes the form given below:

$$C_\theta(u_1,u_2) = \Pr(U_1 < u_1, U_2 < u_2) = \Phi_2(\Phi^{-1}(u_1),\Phi^{-1}(u_2),\theta) = \int_{s_1=-\infty}^{\Phi^{-1}(u_1)}\int_{s_2=-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{s_1^2 - 2\theta s_1 s_2 + s_2^2}{2(1-\theta^2)}\right) ds_1 ds_2, \quad (5)$$

where θ now includes a single parameter that is the correlation coefficient of the standard bivariate normal distribution, and also represents the direction and magnitude of dependence between the standard uniform variates U_1 and U_2.

The multivariate FGM copula that allows for pairwise dependence for spatial analysis takes the form shown below:

$$C_\theta(u_1,u_2,\ldots u_Q) = \prod_{q=1}^{Q} u_q \left[1 + \sum_{q=1}^{Q-1}\sum_{k=q+1}^{Q} \theta_{qk}(1-u_q)(1-u_k)\right], \quad (6)$$

where θ_{qk} is the dependence parameter between U_q and U_k $(-1 \le \theta_{qk} \le 1)$, $\theta_{qk} = \theta_{kq}$ for all q and k.

2.2. Composite Marginal Likelihood Approach

The composite marginal likelihood (CML) approach is an estimation technique that is gaining substantial attention in the statistics field, though there has been little to no coverage of this method in econometrics and other fields.[4] While the method has been suggested in the past under various pseudonyms such as quasi-likelihood (Hjort and Omre, 1994; Hjort and Varin, 2008), split likelihood (Vandekerkhove, 2005), and pseudolikelihood or marginal pseudo-likelihood (Molenberghs and Verbeke, 2005), Varin (2008) discusses reasons why the term composite marginal likelihood is less subject to literary confusion.[5]

[4] The first conference dedicated to the composite likelihood method was held at the University of Warwick in the United Kingdom last year. For a recent review of the method, see Varin (2008).

[5] For instance, the term "quasi-likelihood" is already reserved for a well-established statistical estimating function method that is applicable to cases where the analyst is unable to posit (or would rather not posit) a statistical model for a given set of data, but is willing to identify a link function that relates the mean of the dependent variable vector to a set of covariates, and a variance function that relates the covariance of the dependent variables to the mean vector of the variable (see Wedderburn, 1974; Heyde, 1997, provides an extensive treatment).

The composite marginal likelihood (CML) estimation approach is a relatively simple approach that can be used when the full likelihood function is near impossible or plain infeasible to evaluate due to the underlying complex dependencies, as is oftentimes the case with spatial and time-series models. For instance, in discrete choice models with spatial dependence based on a multivariate normal form, the full likelihood function entails a multidimensional integral of the order of the number of observational units. While there have been recent advances in simulation techniques within a classical or Bayesian framework that assist with such complex model estimation situations (see Bhat, 2003; Beron and Vijverberg, 2004; LeSage, 2000), these techniques are impractical and/or infeasible in situations with a moderate to high number of observations. These simulation-based methods are also not straightforward to implement. In contrast, the CML method, which belongs to the more general class of composite likelihood function approaches (see Lindsay, 1988), is based on forming a surrogate likelihood function that compounds much easier-to-compute, lower-dimensional, marginal likelihoods.[6] The simplest CML, formed by assuming independence across observations, entails the product of univariate densities (for continuous data) or probability mass functions (for discrete data). However, this approach does not provide estimates of dependence that are of central interest in spatial application situations. Another approach is the pairwise likelihood function formed by the product of power-weighted likelihood contributions of all or a selected subset of couplets (*i.e.*, pairs of observations).[7] Almost all earlier research efforts employing the CML technique have used the pairwise approach, including Apanasovich *et al.* (2008), Bellio and Varin (2005), de Leon (2005), Varin and Vidoni (2006, 2009), and Varin *et al.* (2005).

Attention to the CML estimation approach in spatial analysis has been confined to the spatial statistics field thus far, primarily in the context of characterizing spatial dependence for spatial random fields or spatial points or spatial lattices (see, for example, Caragea and Smith, 2006; Guan, 2006; Oman *et al.*, 2007; and Apanasovich *et al.*, 2008). There is little to no mention of the CML approach in the spatial econometrics field, even in recent reviews of, and dedicated paper collections

[6] The general class of composite likelihood function approaches includes composite marginal likelihood approaches and composite conditional likelihood approaches (see, for example, Besag, 1974; Hanfelt, 2004; Mardia *et al.*, 2007). The research of Hjort and Varin (2008) suggests that, in general, composite marginal likelihood approaches are much more efficient than composite conditional likelihood approaches. However, more research is needed to test this result for different types of model structures.

[7] The power weights for each couplet's likelihood contribution may be optimally chosen based on estimating equation theory (Heyde, 1997; Chaganty and Joe, 2004; Kuk, 2007; Guan, 2006), though this is still an open area of debate and research.

in, the field (see Fleming, 2004; Paelinck, 2005; Beck *et al.*, 2006; Páez, 2007; Franzese and Hays, 2008).

3. MODEL FORMUATION

3.1. Copula-based Spatial Ordered Response Model Structure

In the usual framework of an ordered-response model based on a censoring mechanism involving the partitioning of an underlying latent continuous random variable into non-overlapping intervals, let the data (z_q, x_q) be generated as follows:

$$z_q^* = \beta' x_q + \varepsilon_q \tag{7}$$

$$z_q = \begin{cases} 0 \text{ if } -\infty < z_q^* \le \psi_0 \\ 1 \text{ if } \psi_0 < z_q^* \le \psi_1 \\ 2 \text{ if } \psi_1 < z_q^* \le \psi_2 \\ \vdots \\ M \text{ if } \psi_{M-1} < z_q^* < \infty \end{cases}$$

where $\{\psi_0 < \psi_1 < \psi_2 ... < \psi_{M-1}\}$ is a set of thresholds to be estimated, x_q is a vector of exogenous variables whose elements are not linearly dependent (x_q does not include a constant), β is a vector of parameters to be estimated, and ε_q is a random error term. Note that since the underlying scale is unobserved, we normalize the scale without any loss of generality in a translational sense by not including a constant in the x_q vector. The univariate distribution of ε_q can be any parametric distribution in our copula approach, though we will confine ourselves to a logistic or normal distribution in the current study. The mean of ε_q is set to zero. Let σ_q be a scale parameter such that $\eta_q = \varepsilon_q / \sigma_q$ is standard logistic or standard normal. Of course, it is not possible to estimate a separate σ_q parameter for each q. Thus, we parameterize σ_q as $\sigma_q = g(\lambda' w_q) = \exp(\lambda' w_q)$ where w_q includes variables specific to pre-defined "neighborhoods" or other groupings of observational units and individual related factors, and λ is a corresponding coefficient vector to be estimated. For identification purposes caused by scale invariance in the ordered-response model, w_q cannot include a constant. Additionally, consider that the η_q terms are spatially dependent based on the multivariate

copula $C_\theta(.)$. The vector θ includes pairwise dependence terms between an observational unit and other observational units (if only a selected subsample of observational units *k* within a threshold distance of observational unit *q* is considered in the CML estimation approach, then the vector θ includes only the θ_{qk} terms for the selected observational units *k*; alternatively, θ_{qk} =0 for all observational units *k* beyond the threshold distance from observational unit *q*). Since it is not possible to estimate a separate dependence term for each pair of observational units *q* and *k*, and assuming that the spatial process is isotropic, we parameterize θ_{qk} for the Gaussian and FGM copulas as[8]:

$$\theta_{qk} = \pm \frac{(e^{\varsigma})' s_{qk}}{1 + (e^{\varsigma})' s_{qk}}, \tag{8}$$

where s_{qk} is a vector of variables (taking on non-negative values) corresponding to the {*q*,*k*} pair and that influence the level of spatial dependence between observational units *q* and *k*, and ς is a corresponding set of parameters to be estimated.[9] By functional form, $-1 \le \theta_{qk} \le 1$, as required in the FGM and Gaussian copulas (see Bhat and Eluru, 2009). Further, in a spatial context, we expect observational units in close proximity to have similar preferences, because of which we impose the '+' sign in front of the expression in Equation (8). The functional form of Equation (8) can accommodate various (and multiple) forms of spatial dependence through the appropriate consideration of variables in the vector s_{qk}. In particular, the dependence form nests the typical

[8] Note that if we have a separate dependency term for each pair of observational units, there will be $Q(Q-1)/2$ parameters to estimate, which would be more than the number of observations available for parameter estimation for $Q > 3$.

[9] A few issues about the functional form. A continuous transformation mapping to the -1 to +1 range, such as $\theta_{qk} = (e^{\varsigma' s_{qk}} - 1)/(e^{\varsigma' s_{qk}} + 1)$, would pose problems in interpretation if there are negative coefficients in the ς vector. For example, consider that there are two variables in the vector s_{qk}, one being whether *q* and *k* are in the same spatial neighborhood or not and another being the distance between *q* and *k*. Let the coefficient on the first variable be positive and that on the second negative. Then, the implication would be that for *q* and *k* in the same spatial neighborhood, the dependence is positive and decreasing up to a certain distance threshold, but then abruptly changes to a negative dependence beyond the threshold. Also, for *q* and *k* not in the same spatial neighborhood, the spatial dependence is always negative and lower in magnitude for *q* and *k* closer to one another than farther from one another. These dependency forms are difficult to explain. On the other hand, the functional form used in Equation (8) allows a dampening of the magnitude of dependence effects without changing the sign of dependency in response to variables in s_{qk}. Finally, note that the functional form $\theta_{qk} = e^{\varsigma' s_{qk}}/(e^{\varsigma' s_{qk}} + 1)$ is also not appropriate. For example, consider the case when s_{qk} includes a single dummy variable indicating whether or not *q* and *k* are in the same spatial neighborhood. Then, if one uses the above functional form, a spatial dependency is implied even for observational units in different spatial neighborhoods. On the other hand, the functional form in Equation (8) ensures no spatial dependence.

spatial dependence patterns used in the extant literature as special cases, including dependence based on (1) whether observational units are in the same "neighborhood" or in contiguous "neighborhoods", (2) shared border length of the "neighborhood" of two observational units, and (3) time or distance between observational units.[10]

Let $F_q(.)$ be the cumulative distribution of z_q^* and let $f_q(.)$ be the corresponding density function. Also, let d_q be the actual observed categorical response for 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 as:

$$P(z_1 = d_1, z_2 = d_2, ..., z_Q = d_Q) = \int_{D_z^*} c_\theta\left(F_1(z_1^*), F_2(z_2^*), ..., F_Q(z_Q^*)\right) \cdot \left(\prod_{q=1}^{Q} f_q(z_q^*)\right) dz_1^* dz_2^* ... dz_Q^*, \quad (9)$$

where $D_z^* = \{z_1^*, z_2^*, ..., z_Q^* : \psi_{(d_q - 1)} < z_q^* < \psi_{d_q}$ for all $q = 1,2,...,Q\}$ and c_θ is the copula density. The integration domain D_z^* is simply the multivariate region of the z_q^* variables $(q = 1,2,...,Q)$ determined by the observed vector of choices $(d_1, d_2, ..., d_Q)$. The dimensionality of the integration, in general, is equal to the number of observations Q (in the ordinary, spatially uncorrelated ordered response model, the density function collapses to that of the independence copula, and so we are left with only single dimension integrals). If the copula used is the Gaussian copula, and the marginals $F_q(.)$ are univariate normal, the result is the spatial ordered-response probit model. In this case, the probability expression for the observed vector of choices entails an analytically intractable Q-dimensional rectangular integral over the multivariate normal density. Equivalently, the expression involves the computation of the order of 2^Q terms, each term involving a Q-variate cumulative normal distribution.[11] Even with moderate-sized samples, it becomes numerically infeasible to evaluate such a high-dimensional integral. If the copula used is the FGM copula, one encounters the case of either numerically evaluating a Q-dimensional rectangular integral over the FGM density

[10] While the emphasis in this paper is on spatial dependence, the model formulation here is much more general, and can generate dependence between observational units q and k based on proximity on such non-spatial factors as income levels, education levels, and family structure. However, in the empirical analysis of this paper, we will confine attention to spatial measures of proximity.

[11] In the special case of just two categories – *i.e.*, the spatial binary probit model – the expression can be collapsed into a single Q-variate cumulative normal distribution.

function or computing in the order of 2^Q closed-form FGM cumulative distribution functions.[12] Thus, even for the FGM copula, the direct computation of the likelihood function becomes infeasible for even moderate-sized Q in the context of an ordered-response model (for instance, with even just 200 observations, the number of closed-form cumulative distribution functions to be computed is of the order of $1 \cdot 6 \times 10^{60}$). Hence, we propose the much simpler and robust composite marginal likelihood approach for spatial econometric models.

3.2. Estimation and Inference Approach

In the current paper, we use a pairwise marginal likelihood estimation approach, which corresponds to a composite marginal approach based on bivariate margins.[13] Let $\gamma = (\beta', \psi', \lambda', \zeta')'$. For the spatial copula-based ordered response (SCOR) model, the pairwise marginal likelihood function is given by:

$$L_{CML}(\gamma) = \prod_{q=1}^{Q-1} \prod_{k=q+1}^{Q} \left[P(z_q = d_q, z_k = d_k) \right]^{\omega_{qk}}, \tag{10}$$

where $\omega_{qk} (q = 1,...,Q-1, k = q+1,...,Q)$ are suitable non-negative weights, and

$$P(z_q = d_q, z_k = d_k) = C_\theta\left(u_{(d_q)}, u_{(d_k)}\right) - C_\theta\left(u_{(d_{q-1})}, u_{(d_k)}\right) - C_\theta\left(u_{(d_q)}, u_{(d_{k-1})}\right) + C_\theta\left(u_{(d_{q-1})}, u_{(d_{k-1})}\right),$$

$$u_{(d_q)} = \overline{F}_q\left(\frac{\psi_{(d_q)} - \beta' x_q}{\exp(\lambda' w_q)} \right).$$

$\overline{F}_q(.)$ in the above expression corresponds to the standard cumulative distribution form of $F_q(.)$. The bivariate-probability expressions in the likelihood function of Equation (10) are straightforward to compute, since they only entail four bivariate copula expressions (for example, in the case of

[12] In the special case of two categories and the FGM copula, the probability collapses to a single closed-form expression term, as exploited by Bhat and Sener (2009). In the special case of two categories and the generalized Gumbel (GG) copula, the probability still requires 2^Q closed-form GG cumulative distribution term evaluations because of the asymmetric nature of this copula. This is different from the case of the FGM copula with two categories.

[13] The analyst can also consider larger subsets of observations, such as triplets or quadruplets or even higher dimensional subsets (see Oman *et al.*, 2007; Engler *et al.*, 2006; Caragea and Smith, 2007). In general, the issue of how best to form a CML function corresponding to a full likelihood function remains an open, and under-researched, area of research because of the relatively limited results on the properties of CML inferential procedures (but see Zhao and Joe, 2005 and Cox and Reid, 2004). However, it is generally agreed that the CML construction should be based on balancing statistical and computational efficiency.

univariate normal margins and a Gaussian copula, the bivariate probability involves four bivariate cumulative normal expressions).[14]

The pairwise marginal likelihood function of Equation (10) comprises $Q(Q-1)/2$ pairs of bivariate probability computations, which can itself become quite time consuming. Fortunately, in a spatial case where dependency drops quickly with inter-observation distance, the pairs formed from the closest observations provide much more information than pairs that are very far away. In fact, as demonstrated by Varin and Vidoni (2009), Varin and Czado (2008), and Apanasovich *et al.* (2008), in different empirical contexts, retaining all $Q(Q-1)/2$ pairs not only increases computational costs, but may also reduce estimator efficiency. Typically, in a spatial context, there appears to be an optimal distance for inclusion of observation pairs. This distance threshold may be set based on knowledge about the spatial process or based on testing the efficiency of estimators with varying values of the distance threshold. Assume that this distance threshold is *m*, and let the set of observational units *k* within the threshold distance of unit *q* be M_q. Then, we propose dummy weights to include appropriate pairwise terms in the composite marginal likelihood function of Equation (10). In particular, $\omega_{qk} = 1$ if $k \in M_q$ and $\omega_{qk} = 0$ otherwise. This effectively reduces the number of pairwise terms in the CML function.

The properties of the CML estimator may be derived using the theory of estimating equations (see Cox and Reid, 2004). Specifically, under usual regularity assumptions (Molenberghs and Verbeke, 2005, page 191), the CML estimator is consistent and asymptotically normal distributed (this is because of the unbiasedness of the CML score function, which is a linear combination of proper score functions associated with the marginal event probabilities forming the composite likelihood). Of course, the maximum CML estimator loses some efficiency from a theoretical perspective relative to a full likelihood estimator (Lindsay, 1988; Zhao and Joe, 2005), but this efficiency loss has been shown to be minimal in most empirical cases (Lele and Taper, 2002; Henderson and Shimakura, 2003; Lele, 2006). Besides, in many situations, it is infeasible to use the full likelihood estimator anyway. Even if feasible, numerical simulation methods that are typically needed in such situations get imprecise as the number of dimensions increase, leading to convergence problems during estimation. Further, as indicated by Varin and Vidoni (2009), it is

[14] $C_\theta(u_1,u_2) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2),\theta)$ for the bivariate Gaussian copula
$C_\theta(u_1,u_2) = u_1u_2[1 + \theta(1-u_1)(1-u_2)]$ for the bivariate FGM copula

possible that the "maximum CML estimator can be consistent when the ordinary full likelihood estimator is not". This is because the CML procedures are typically more robust and can represent the underlying low-dimensional process of interest more accurately than the low dimensional process implied by an assumed high-dimensional multivariate model.

The CML estimator $\hat{\gamma}_{CML}$, obtained by maximizing the logarithm of the function in Equation (10) with respect to the vector γ, is asymptotically normal distributed with asymptotic mean γ and variance matrix given by the inverse of Godambe's (1960) sandwich information matrix[15]:

$$G^{-1}(\gamma) = \left[H(\gamma)\right]^{-1} J(\gamma) \left[H(\gamma)\right]^{-1}, \tag{11}$$

where

$$H(\gamma) = E\left[-\frac{\partial^2 \log L_{CML}(\gamma)}{\partial\gamma\partial\gamma'} \right] \text{ and}$$

$$J(\gamma) = E\left[\left(\frac{\partial \log L_{CML}(\gamma)}{\partial\gamma} \right) \left(\frac{\partial \log L_{CML}(\gamma)}{\partial\gamma'} \right) \right].$$ [16]

The "bread" matrix $H(\gamma)$ of Equation (11) can be estimated in a straightforward manner using the Hessian of the negative of $\log L_{CML}(\gamma)$, evaluated at the CML estimate $\hat{\gamma}$. This is because the information identity remains valid for each pairwise term forming the composite marginal likelihood. Thus, $H(\gamma)$ can be estimated as:

[15] The analyst can also use a two-stage approach to maximize the logarithm of the composite likelihood function in Equation (9). This entails assuming independence across each pair of observational units (*i.e.*, $\theta_{qk} = 0$ for all q and k) and estimating $\gamma_{-\zeta} = (\beta', \psi', \lambda')'$. Then, fixing these estimated parameters in Equation (9), the analyst can maximize the logarithm of the equation to estimate the ζ parameter vector embedded in the θ_{qk} terms. Zhao and Joe (2005) have compared this two stage approach with the pairwise approach adopted in the current paper for the case of a continuous dependent variable and a binary discrete variable, and have generally found that the pairwise approach outperforms the two stage approach. Besides, the computation of the covariance matrix of the parameters is more cumbersome in the two-stage approach than in the pairwise approach.

[16] This sandwich form for the asymptotic variance matrix for the CML estimator is similar to the asymptotic variance matrix for the ordinary maximum likelihood estimator under a mis-specified model. This is not surprising, since the CML setting is tantamount to a mis-specified model due to the consideration of only partial, though correctly specified, likelihood terms. Specifically, the use of the CML inference procedure implies the failure of the information identity (*i.e.*, $H(\gamma) \neq J(\gamma)$), implying a loss of efficiency with respect to the ordinary, but practically infeasible, maximum likelihood estimator.

$$\hat{H}(\hat{\gamma}) = -\left[\sum_{q=1}^{Q-1}\sum_{k=q+1}^{Q}\frac{\partial^2 \log L_{CML,qk}(\hat{\gamma})}{\partial\hat{\gamma}\partial\hat{\gamma}'}\right], \tag{12}$$

where

$$L_{CML,qk}(\hat{\gamma}) = [P(z_q = d_q, z_k = d_k)]^{\omega_{qk}} \mid \hat{\gamma}$$

However, the estimation of the "vegetable" matrix $J(\gamma)$ is more difficult. One cannot estimate $J(\gamma)$ as the sampling variance of individual contributions to the composite score function because of the underlying spatial dependence in observations. But, since the spatial dependence fades with distance, we can use the windows resampling procedure of Heagerty and Lumley (2000) to estimate $J(\hat{\gamma})$. This procedure entails the construction of suitable overlapping subgroups of the original data that may be viewed as independent replicated observations. Then, $J(\gamma)$ may be estimated empirically as the weighted average of the variance of composite score evaluations (computed at $\hat{\gamma}$) across the subgroups (the weights correspond to the size of each subgroup). In the current spatial context, we can consider all the observational units *k* in the data within a distance *m* of observation *q* as a subgroup or cluster (note that the dependence is weak beyond a distance *m*, and thus each subgroup as just defined would be only weakly dependent on other subgroups). Then, we propose the following as an estimate of the matrix $J(\gamma)$:

$$\hat{J}(\hat{\gamma}) = \sum_{q=1}^{Q}\frac{1}{N_q}[S_{CML,q}(\hat{\gamma})][S_{CML,q}(\hat{\gamma})]', \text{ where} \tag{13}$$

$$S_{CML,q}(\hat{\gamma}) = \sum_{q=k+1}^{Q}\left[\frac{\partial \log L_{CML,qk}(\hat{\gamma})}{\partial\hat{\gamma}}\right], \text{ and } N_q = \sum_{q=k+1}^{Q}\omega_{qk}.$$

As indicated earlier, for any copula model, one needs to determine the optimal threshold distance "*m*" that provides the most efficient parameter estimates. We establish this distance by estimating the variance matrix $G(\gamma)$ for different distance values and selecting the distance value that minimizes the total variance across all parameters as given by $tr[G(\gamma)]$, where $tr[A]$ denotes the trace of the matrix A.

3.3. Model Selection

Varin and Vidoni (2005) introduced a composite likelihood information criterion (CLIC) for model selection. In the current paper, this criterion can be used for selecting between different copula

models with the same threshold distance (*i.e.*, the same number of pairwise terms). We select the copula model that maximizes the following penalized log-composite likelihood[17]:

$$\log L^{*}_{CML}(\hat{\gamma}) = \log L_{CML}(\hat{\gamma}) - tr\left[J(\hat{\gamma}) H(\hat{\gamma})^{-1} \right] \tag{14}$$

An issue that is closely associated with model selection is testing null hypotheses. The composite likelihood ratio statistic may be used for this purpose. Consider the null hypothesis $H_0 : \tau = \tau_0$ against $H_1 : \tau \neq \tau_0$, where τ is a subvector of γ of dimension *d*; *i.e.*, $\gamma = (\tau', \alpha')'$. The statistic takes the familiar form shown below:

$$CLRT = 2\left[\log L_{CML}(\hat{\gamma}) - \log L_{CML}(\gamma_0) \right], \tag{15}$$

where γ_0 is the composite marginal likelihood estimate under the null hypothesis $(\tau_0', \alpha_{CML}'(\tau_0))$. The CLRT statistic does not have a standard chi-squared asymptotic distribution as in the case of the ordinary maximum likelihood ratio statistic. Molenberghs and Verbeke (2005; chapter 9) provides the appropriate asymptotic distribution, which is based on Kent's (1982) derivation of the distribution of the ordinary likelihood ratio statistic under a mis-specified likelihood function. To write this, first define $G_\tau(\gamma)$ and $H_\tau(\gamma)$ as the $d \times d$ submatrices of $G(\gamma)$ and $H(\gamma)$, respectively, which correspond to the vector τ. Then, CLRT has the following asymptotic distribution:

$$CLRT \sim \sum_{i=1}^{d} \lambda_i W_i^2, \tag{16}$$

where W_i^2 for $i = 1, 2, \ldots, d$ are independent χ_1^2 variates and $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ are the eigenvalues of the matrix $H_\tau(\gamma)^{-1} G_\tau(\gamma)$ evaluated under the null hypothesis. The problem with this approach though is that it cannot be used when the parameter value(s) under the null hypothesis is at the boundary of the parameter space. For instance, in the spatial dependence case of this paper, one may want to test if $\varsigma \to -\infty$ in Equation (7), which corresponds to the case of no spatial dependence. This represents a degenerate case for the application of Equation (14). Fortunately, one can resort to

[17] This penalized log-composite likelihood is nothing but the generalization of the usual Akaike's Information Criterion (AIC). In fact, when the candidate model includes the true model in the usual maximum likelihood inference procedure, the information identity holds (*i.e.*, $H(\gamma) = J(\gamma)$) and the CLIC in this case is exactly the AIC $[= \log L_{ML}(\hat{\gamma}) - (\#\text{ of model parameters})]$. Note also that the CLIC can be applied only for the case when the log composite likelihoods $[\log L_{CML}(\hat{\gamma})]$ are comparable across the models being tested or, equivalently, only when the same number of pairwise terms are used in the development of the log-composite likelihood function.

parametric bootstrapping to obtain the precise distribution of the CLRT statistic for any null hypothesis situation. Such a bootstrapping procedure is rendered very simple in the CML approach, and can be used to compute the p-value of the null hypothesis test. The procedure is as follows (see Varin and Czado, 2008):

1. Compute the observed CLRT value as in Equation (14) from the estimation sample. Let the estimation sample be denoted as y_{obs}, and the observed CLRT value as $CLRT(y_{obs})$.
2. Generate B sample data sets $y_1, y_2, y_3, ..., y_B$ using the CML convergent values under the null hypothesis
3. Compute the CLRT statistic of Equation (14) for each generated data set, and label it as $CLRT(y_b)$.
4. Calculate the p-value of the test using the following expression:

$$p = \frac{1 + \sum_{b=1}^{B} I\{CLRT(y_b) \geq CLRT(y_{obs})\}}{B+1}, \text{ where } I\{A\} = 1 \text{ if } A \text{ is true.}$$

4. SIMULATION STUDY

To evaluate the performance of the CML estimation technique just discussed, we generate a sample of 500 observations with three independent variable and four ordered categories. The values for each of the independent variables are drawn from a standard univariate normal distribution. The coefficients applied to the independent variables are 1, 0.5, and 0.25. Next, values of the error terms ε_q (q = 1, 2, …, 500) in Equation (7) are generated with the dependence structure of Equation (8). We use a standard distribution for the ε_q terms (*i.e.*, we assume $\sigma_q = 1$ for all q) and focus on spatial dependence, which is what creates estimation complications in moderate to large sized samples. Further, for this simulation study, we consider normally distributed marginal error terms and consider a Gaussian copula to tie the error terms. To accommodate a global spatial dependence pattern among the error terms, we consider a single variable - the inverse of distance - in the s_{qk} vector that influences the level of spatial dependence between observational units q and k. We adopt this specification because it is simple and intuitive, and generates a global spatial dependence structure. The distance between members of each pair of the 500 observations is borrowed from the residential locations of 500 teenagers residing in the San Francisco Bay Area, based on the 2000 San

Francisco Bay Area Travel Survey (BATS) that is used in the empirical analysis of this paper (see Section 5). We then consider two values of ς for the coefficient on the inverse of distance [see Equation (8)] - $\varsigma = -0.5$ and $\varsigma = 0.5$. For a given distance between two observations, the former value for ς leads to lower spatial error dependence, while the latter value results in higher spatial error dependence. For instance, for two observations spaced 1 mile (10 miles) apart, the error correlation is 0.38 (0.06) when $\varsigma = -0.5$ and 0.62 (0.14) when $\varsigma = +0.5$. For each of the two values of ς, we generate a multivariate realization of the error term vector $\varepsilon = (\varepsilon_1, \varepsilon_2, ...\varepsilon_{500})$ with the desired correlation matrix. The error terms for each observation (ε_q) is then added to the systematic component $\beta' x_q$ as in Equation (7) and then translated to "observed" values of z_q (= 0, 1, 2,3) based on the pre-specified threshold values of ψ_0 = 0.25, ψ_1 = 0.75, and ψ_2 =1.5. The data generation process is undertaken 25 times with different realizations of the random vector ε to generate 25 different data sets. The CML estimation procedure is then applied to each dataset to obtain estimated values. In this estimation, we considered all possible pairings of the 500 observations.

The performance evaluation of the CML technique is based on the ability to recover the parameter vector $\gamma = (\beta', \psi', \varsigma')'$. The proximity of estimated and true values for each parameter is based on computing the following three metrics: (a) bias (or the difference between the mean of the relevant values across the 25 runs and the true values), (b) the relative bias (*i.e.*, the bias as a percentage of the true value), (c) the total error computed as the root mean-squared error (RMSE) between the estimated and true values across all 25 runs ($RMSE = 100 \times \sqrt{\left(\sum_{r=1}^{R} (\hat{\gamma}_r - \gamma)^2 \right) / R}$, where R is the number of data replications).

Table 1 presents the results, and shows that the CML approach recovers the parameters extremely well in terms of bias and relative bias (see the fourth and fifth columns of the table). The relative bias is, in general, higher for the case of higher dependence ($\varsigma = 0.5$), but still is very small. The RMSE values are also generally small, though it is interesting to note that the RMSE values for the thresholds are consistently higher than for other parameters in both the low and high correlation case. Overall, the simulation results clearly demonstrate the ability of the Composite Marginal Likelihood (CML) to recover the parameters in a spatially dependent discrete choice context.

Combined with its conceptual and implementation simplicity, the CML approach is an effective method in spatial and other high-dimensional multivariate distribution contexts.

5. THE EMPIRICAL CONTEXT

5.1. Background

The empirical context of the current study is the participation of teenagers in out-of-home, weekday, recreational episodes. Participation in out-of-home recreational activity is an important part of a balanced and healthy lifestyle, and contributes in important ways to the physical, emotional, and mental health of adolescents (see Campbell, 2007). Specifically, several earlier studies have shown that participation in out-of-home structured recreational activities helps adolescents develop social skills, self-esteem, teamwork abilities, fairness concepts, and tolerance. Such participation also provides increased learning opportunities, and reduces the incidence of drug and tobacco use, depression and level of illness (see, for example, Hofferth and Sandberg, 2001; Tiggemann, 2001; Marsh and Kleitman, 2003; Klentrou *et al.*, 2003).

In this paper, we distinguish between physically active and physically inactive activity episodes within the context of out-of-home recreation pursuits. Earlier studies in the literature strongly emphasize the importance of physical activity in improving health, since physical activity increases cardiovascular fitness and decreases obesity, heart disease, diabetes, high blood pressure, and several forms of cancer. Further, it also enhances agility and strength, reduces the need for medical attention, contributes to improved mental health, and decreases depression and anxiety (see, for instance, Nelson and Gordon-Larsen, 2006; Center for Disease Control (CDC), 2006; Ornelas *et al.*, 2007). But, in spite of its well-acknowledged benefits, adolescents' participation rates in physical activity have been very low. A report by the Center for Disease Control (2002) indicates that about a third of teenagers do not engage in adequate physical activity, and that the high school physical education class participation rate has been steadily declining over the past decade. Since physically inactive lifestyles may be transferred from adolescence to adulthood (Aaron *et al.*, 2002), the low physical activity participation among adolescents has become a particularly serious health concern in the U.S. and other countries.

Of course, the recreational activity participation of adolescents is not only relevant to the sociology or public health fields, but is also of considerable interest to transportation professionals in understanding children's activity engagement patterns (see, for example, Transportation Research

Board and Institute of Medicine, 2005; Goulias and Kim, 2005; Copperman and Bhat, 2007a,b; and Sener *et al.*, 2008). In particular, teenagers' activity engagement desires and needs impact the activity-travel patterns of adults through serve-passenger activities (such as the need to drop off a teenager at an out-of-home recreational activity site at a certain time of the day) and joint activity participations with parents (such as going to the park or taking a walk around the neighborhood). Also, the consideration of children's activity-travel patterns is important in its own right since these patterns contribute directly to travel demand.

In the current application, we model teenagers' participation in recreational activity (both physically active and physically inactive) in terms of the number of episodes of participation during the weekday, using a comprehensive set of socio-demographic and built environment variables as potential determinants, while also accommodating spatial effects based on residence patterns. The approach developed in this research allows spatial dependence in recreational pursuits across teenagers, as well as accommodates heteroscedasticity across teenagers based on spatial and individual-related characteristics.[18]

5.2. Data and Sample Formation

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 (see MORPACE International Inc., 2002). The sample used for the current analysis is confined to a single weekday of 1447 teenagers from 1447 different households residing in nine Counties (Alameda, Contra Costa, San Francisco, San Mateo, Santa Clara, Solano, Napa, Sonoma and Marin) of the San Francisco Bay Area. Each recreational activity episode of the teenagers is classified as being physically active (such as sports, games, walking around the neighborhood, and physically active play) or physically inactive (such as organized hobbies, attending sports events, going to the movies/concerts, and arts and crafts). The episodes are appropriately aggregated to obtain the number of physically active and physically inactive episodes

[18] It is possible that in addition to dependence due to unobserved factors in the propensities of participation in physically active recreation across teenagers, and similar dependence in the propensities of participation in physically inactive recreation across teenagers, there is also dependence in unobserved factors in the propensities of participation in physically active and physically inactive recreation within the same teenager. However, in this analysis, we consider each of the two recreation activity purposes in isolation to keep things relatively simple in this first application of a composite likelihood procedure to a spatial econometric context. The consideration of dependence across observational units as well as across multiple dependent variables of the same observational unit is left for future research.

undertaken by each teenager during the sampled weekday. These two variables constitute the dependent variables. The distributions of the number of active episodes and the number of inactive episodes in the sample are as follows: zero episodes (active: 89.0%, inactive: 79.1%), one episode (active: 9.9%, inactive: 17.3%), and two or more episodes (active: 1.1%, inactive: 3.6%).

In addition to the BATS survey, several other secondary Geographic Information System (GIS) data layers of highways, local roadways, bicycle facilities, businesses, and land-use/demographics were used to obtain spatial variables and built environment variables characterizing the residential traffic analysis zone (TAZ) of each teenager.[19] The residential neighborhood 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 a 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 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 (Fotheringham, 1983) 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

[19] Due to privacy considerations, the point coordinates of each teenager's residence is not available; only the TAZ of residence of each teenager is available.

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 and 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 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.
8) *Spatial dependence variables*, characterize the spatial dependence of the residences of each pair of teenagers (these are the elements of the s_{qk} vector in Section 3.1). These include (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 and square of inverse of distance.[20]

6. EMPIRICAL ANALYSIS

6.1 Model Specification

Several different variable specifications, functional forms, and variable interactions were considered to identify the final model specification. The variables included (1) individual characteristics (age, sex, race, driver's license holding, physical disability status, *etc.*), (2) household characteristics (number of adults, number of children, household composition and family structure, household income, dwelling type, whether the house is owned or rented, parents' activity participation characteristics, *etc.*), (3) activity-day variables (season of the year, day of week, *etc.*), and (4) residential neighborhood variables (as discussed in Section 5.2).

[20] 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.

We estimated separate ordered response models for teenagers' physically active recreational activity (or active recreation) and physically inactive recreational activity (or inactive recreation). The models were estimated with two different univariate distribution assumptions (normal and logistic) for the random error term ε_q and two different copula structures (FGM and Gaussian). Thus, for each of the active and inactive recreation ordered response structures, we estimated four distribution-copula models: (1) Normal-FGM, (2) Normal-Gaussian, (3) Logistic-FGM, and (4) Logistic-Gaussian.

The final model specification was based on a systematic process of eliminating variables found to be statistically insignificant, intuitive considerations, insights from previous literature, and parsimony in specification. 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. Further, we retained the same set of variables across all the four distribution-copula models for consistency and comparison purposes, so some variables that may turn out to be statistically significant in one model may be marginally significant in another.

6.2. Model Selection

As discussed before, the optimal distance for selecting pairwise terms for inclusion in the composite likelihood was set based on minimizing the trace of the variance-covariance matrix, tr($G(\gamma)$). To achieve this, tr($G(\gamma)$) was computed for five distance thresholds (5 miles, 10 miles, 20 miles, 40 miles and 151.46 miles, the last one representing the case of including all the $Q(Q-1)/2$ possible pairs in the CML function). Our results showed that the trace values did not change substantially based on the distance threshold used, particularly for the inactive recreation category. But, in general and across the four distribution-copula models, the best estimator efficiency was obtained at about 40 miles for the active recreation category and at 151.46 miles for the inactive recreation category.

The next step in the selection process was to identify the best model from the distribution-copula models estimated using the 40 miles (151.46 miles) distance threshold for active (inactive) recreation using the composite likelihood information criterion (CLIC). Table 2 provides the values of the log-composite likelihood at convergence $\log L_{CML}(\hat{\gamma})$, the trace value in the CLIC statistic ($tr\left[J(\hat{\gamma})H(\hat{\gamma})^{-1}\right]$), and the CLIC statistic value [see Equation (12)]. As can be observed from the CLIC statistic column, the Logistic-Gaussian model turns out to be the best specification for both the

active and inactive recreation categories. The usual multivariate normal specification for the error terms (as captured by the Normal-Gaussian model) has a poorer fit. This finding highlights the value of the copula approach that is able to separate out the univariate marginal distribution form from the multivariate dependence structure.

Finally, one can compare the Logistic-Gaussian model for each of the two recreation activity categories with a standard ordered logistic (SOL) model with no spatial dependence and no heteroscedasticity. This SOL model can be easily estimated using the classical maximum likelihood procedure. However, we estimate it using the CML approach so that we can compare the data fit of these models with those that incorporate spatial dependence.[21] The CLIC statistic value at convergence for the SOL model is -447,759 for the active recreation category and -1,271,030 for the inactive recreation category. Comparing these numbers with the corresponding ones from Table 2, we observe that all the copula models perform better than the SOL model in terms of the CLIC statistic, rejecting the null hypothesis of no spatial dependence and no heterogeneity. One can also use the more powerful CLRT statistic to compare the SOL model with the best Logistic-Gaussian model from Table 2. Using the parametric bootstrap procedure discussed in Section 3.3, we can compute the p-value corresponding to the null hypothesis of $\lambda = 0$ and $\varsigma \to -\infty$. The estimated p-value based on 25 bootstrap samples is 0.115 for the active recreation category and 0.038 for the inactive recreation category. The low p-values reject the null hypotheses of absence of heterogeneity and spatial dependence, and highlight the value of the Logistic-Gaussian models estimated in the current paper.[22]

The empirical results are presented in the following section, in which we focus our attention on the results of the best Logistic-Gaussian copula models.

---

[21] The CML and ML estimates are, as expected, almost identical for the SOL model. For the active recreation category, we observed some very small differences between the CML and ML estimates because of the use of the 40 mile threshold in the CML approach, which effectively has the result of weighting some observations more than others. For the inactive category, the CML and ML estimates are pretty much identical because of the use of all pairs in the CML estimation, which weights all observations equally as in the ML case.

[22] We also compared the Logistic-Gaussian models for each of the active and inactive recreation categories with corresponding pure heteroscedastic ordered logistic (HOL) models with no spatial dependence. This tests the exclusive null hypothesis of the absence of spatial dependence. Again, the CLIC statistics for the Logistic-Gaussian models were higher than those for the HOL models (the HOL CLIC statistics were -437,171 for the active recreation category and -1,255,749 for the inactive recreation category, both of which are lower than the corresponding Logistic-Gaussian CLIC values). As we will see later in Section 6.4.2, we can also reject the null hypothesis of no spatial dependence based on the statistically significant t-statistics on the spatial dependence effects.

6.3. Estimation Results

Table 3 presents the estimation results for the best ordered response models as identified in the previous section. The coefficients provide the effects of variables on the latent propensity of teenagers to participate in active recreation (second main column) and inactive recreation (third main column). The first two rows provide estimates of the threshold values that do not have any substantive interpretation. These thresholds simply serve to translate the latent propensity into the observed ordered categories of the number of recreational activity participations.

6.3.1 Individual Characteristics

The effects of individual characteristics indicate that, among teenagers, males are more likely than females to participate in active recreation (see Mhuircheartaigh, 1999 and Bhat, 2008 for similar results). The age variable suggests a significantly higher propensity for inactive recreation among teenagers aged 16 to 19 years (relative to their younger teenage counterparts), perhaps because these older teenagers can drive themselves to recreation activity locations and not be dependent on others to chauffeur them. The race variable effects reveal that Hispanic teenagers have a lower propensity to partake in recreational activity (active and inactive) relative to Caucasians, African Americans, Asians, and other ethnic groups, while Asian teenagers are less likely to participate in inactive recreation relative to their non-Hispanic teenager peers.

The part-time student status (employment status) variable effect reflects a higher (lower) prevalence of inactive recreation among part-time students (employed individuals) compared to full-time students (non-employed individuals), possibly due to time constraints of full-time students (employed teenagers). Interestingly, we did not find any statistically significant effects of these variables on active recreation propensity.

6.3.2 Household Characteristics

The household-related variable effects show that teenagers in households with more children (age less than 18 years) are associated with a higher propensity to participate in active recreation, possibly due to increased opportunities for joint physical recreational activities with siblings. The household structure effects indicate that teenagers living in nuclear family households (*i.e.*, households with both parents living with the teenager) are less likely to partake in active recreation compared to those in other household structure types (single parent families, roommate families, and

joint families with several adults), a result that needs further exploration. Finally, among the household demographic variables, teenagers in low income households (less than an annual income of $35,000) have a lower propensity for active recreation, presumably due to financial constraints (see Bhat *et al.*, 2006 for a similar income-related effect). At the same time, the results indicate that teenagers living in high income households (more than an annual income of $90,000) are more likely to participate in inactive recreation.

The final variables in the category of household characteristics indicate, as expected, a higher level of active recreation among teenagers whose parents participate in physically active recreation (for the purpose of this research, we designate a parent as participating in physically active recreation if the parent pursues one or more active recreation episodes on the survey day). This is presumably because parents serve as role models to children. Further, the joint activity participation of parents and children can significantly motivate and increase physical activity participation among teenagers (such as bicycling or walking together as a family around the neighborhood). Perhaps, an appropriate policy strategy to encourage physical activity participation among teenagers would be to develop family oriented programs focusing on encouraging physical activity levels among parents as well as other household members.

6.3.3 Household Location, Season, and Activity Day Variables

Teenagers residing in San Francisco County have a higher (lower) tendency to pursue active (inactive) recreation 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). Further, the tendency for inactive recreation among teenagers is lower if they reside in Alameda County.[23]

The seasonal variables reflect the higher propensity to participate in active recreation during the temperate summer months and the higher inclination for inactive recreation during cold winter months. This suggests that public health policies aimed at encouraging year-round teenager physical

[23] These location dummy variables are perhaps capturing micro-scale urban form characteristics and crime-related characteristics. Further exploration of the effects of such attributes is an important avenue for future research. In the current paper, a number of built environment measures were considered, but these are at the relatively coarse spatial scale of the traffic analysis zone (because we could not obtain the point coordinates of each teenager's residence, and only have the teenager's residence tagged to a traffic analysis zone). Further, we were able to obtain crime statistics only at the county level, and this aggregate crime variable did not have a statistically significant impact on recreation activity participation.

activity participation should focus on providing more indoor recreational activity opportunities at affordable cost during the non-summer months in general, and the winter season in particular.

The results also indicate that teenagers have a higher propensity to participate in inactive recreation on Fridays compared to other days of the week.

6.3.4 Residential Neighborhood Variables

The next set of variables in Table 3 corresponds to the impacts of the residential neighborhood measures identified in Section 5.2. Many of these variables did not turn out to be statistically significant even at the 15% level in both of the models, and hence do not appear in Table 3. Of course, as indicated earlier, this result may be the consequence of using a relatively coarse spatial resolution for computing these variables.

The effect of the "Fraction of African-American population" variable in Table 3 shows that there is a lower propensity to participate in active recreation among teenagers living in zones with a high percentage of African-American population relative to teenagers in other areas. A similar result is obtained by Gordon-Larsen *et al.* (2005, 2006), who suggest that this may be because of poor neighborhood quality and lack of good recreational facilities in areas with a high fraction of African-American population. As expected, the presence of natural recreation sites (such as county/state/national parks, gardens, nature centers) in a zone has a positive influence on active recreation among teenagers residing in the zone, suggesting that providing more opportunities for natural recreation and improving accessibility to natural recreation sites may be an effective urban and transportation policy to improve public health. Finally, teenagers in households that own several bicycles and that are in residential areas with a high bicycle facility density (as measured by miles of bicycle lanes per square mile in the residential TAZ) are more likely to participate in physically active recreational pursuits than their peers in other households.

6.4. Heteroscedasticity and Spatial Dependency

This section presents the parameter estimates characterizing heteroscedasticity and spatial dependence in the models.

6.4.1 Heteroscedasticity

Several variables were considered in the w_q vector that generates scale heteroscedasticity among individuals (note that $\sigma_q = \exp(\lambda' w_q)$), though only a handful turned out to be statistically significant. The estimates provided in Table 3 under "(Spatial) heteroscedasticity variables" correspond to the λ vector. The results indicate a higher variation (*i.e.*, more spread) in the propensity to participate in active recreation during the winter season (relative to other seasons) and among teenagers living in Contra Costa County (compared to other counties). Further, the results reveal a much tighter variation (*i.e.*, less spread) in the propensity to participate in active recreation among teenagers residing in San Francisco and Solano County. In combination with the direct positive effect of the San Francisco location variable on active recreation propensity (see Section 6.3.3), the net implication is that teenagers residing in San Francisco have a uniformly higher propensity to participate in active recreation relative to teenagers living elsewhere.

The season and "Alameda County" residence variables influence the scale of the error term for the inactive recreation category. Specifically, there is a much tighter variation (*i.e.*, less spread) in the propensity to participate in inactive recreation among teenagers during the winter and fall seasons (relative to other seasons), and a higher variation (*i.e.*, more spread) among teenagers residing in Alameda County (relative to other counties).

6.4.2 Spatial Dependence Effects

In addition to heteroscedasticity, the estimated ordered choice models also incorporate spatial dependency across observational units through the s_{qk} vector and the corresponding ς coefficient vector. In this regard, the best specification for the models included a single "inverse of distance" variable in the s_{qk} vector of Equation (7). The corresponding ς coefficient is reported in Table 2, and has a value of -1.690 (with a standard error estimate of 0.074) for active recreation, and -1.116 (with a standard error estimate of 0.047) for inactive recreation. The implied value of $\mu = e^{\varsigma}$ is 0.185 for active recreation and 0.327 for inactive recreation, with corresponding t-statistics values of 13.53 and 21.18 with respect to the null hypothesis that $\mu = 0$ (the standard error for μ may be computed from that for ς using the familiar *delta* method). These t-statistics clearly reject the hypothesis of no spatial dependence (note that $\theta_{qk} = 0$ for all q and k pairs in Equation (7) if $\mu = 0$,

and so rejection of the hypothesis that $\mu = 0$ is a clear rejection of spatial independence). In fact, it is easy to show that the t-statistic corresponding to the spatial correlation coefficient θ_{qk} (θ_{qk} for the Gaussian copula corresponds to the traditional Spearman's correlation) may be written as follows:

$$tstat(\theta_{qk}) = \left(\frac{1}{dist_{qk}} + \frac{1}{\mu} \right) \times \frac{\mu^2}{std(\mu)},$$

which then implies that as $dist_{qk} \to \infty$, $\theta_{qk} \to 0$ and the t-statistic for θ_{qk} tends toward the t-statistic of μ. As the distance between two teenagers *q* and *k* decreases, θ_{qk} increases and the t-statistic of θ_{qk} also increases. Essentially then, one can reject the null hypothesis that θ_{qk}=0 even for the most distant pair of teenagers, since the t-statistic for testing this hypothesis will be at least 13.53 for active recreation and 21.18 for inactive recreation for any pair of teenagers. Given the range of the distance between teenagers' residences in the sample, the Spearman's correlation ranges, for active recreation, from 0.001 (for two teenagers located 151.460 miles apart) to 0.575 (for two teenagers located 0.135 miles apart). The correlation for two teenagers spaced 1 mile apart in the active recreation case is 0.156 and that for two teenagers spaced 2 miles apart is 0.085. The correlation values for inactive recreation range from 0.002 (for two teenagers located 151.460 miles apart) to 0.708 (for two teenagers located 0.135 miles apart). The correlation for two teenagers spaced 1 mile apart in the inactive recreation case is 0.247 and for two teenagers spaced 2 miles apart is 0.141. These results indicate that the spatial extent of dependence in unobserved factors for active recreation is smaller than the spatial extent of dependence in unobserved factors for inactive recreation. Alternatively, the spatial dependence effect is more localized for active recreation relative to inactive recreation. Thus, including distant teenager pairs (in the construction of the CML function) provides more useful information (and better estimator efficiency) for inactive recreation compared to active recreation, as our empirical results indicated in Section 6.2.

It is clear from the discussion above that the spatial dependence effect is very highly statistically significant, and needs to be accommodated. The standard ordered logistic (SOL) model ignores these spatial dependencies, while the Logistic-Gaussian (LG) copula models of this paper consider these dependencies and accommodates (spatial and individual) heteroscedasticity. The result is that the SOL model provides less efficient estimates. In particular, the average of the trace of the covariance matrix of parameter estimates for the active recreation (inactive recreation) model

is 0.062 (0.011) for the LG model and 0.074 (0.024) for the SOL model, indicating the higher standard errors of the SOL model. Further, as we discuss in the next section, the SOL model also provides inconsistent elasticity effects.

6.5. Aggregate-Level Elasticity Effects

The parameters on the exogenous variables in Table 3 do not directly provide the magnitude of the effects of the variables on the probability of each number of weekday recreation episodes. To do so, 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 expected aggregate shares of each number of activity episodes 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 each number of activity episodes 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 number of children), we increase the value of the variable by 1 and compute a percentage change in the expected aggregate share of teenagers participating in each number of activity episodes. Finally, the aggregate-level "arc" elasticity effect of a continuous exogenous variable (such as fraction of African-American population) 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 each number of activity episodes. While the aggregate level elasticity effects are not strictly comparable across the three different types of independent variables (dummy, ordinal, and continuous), they do provide order of magnitude effects.

The elasticity effects by variable category for each of the active and inactive recreation categories, and for both the (aspatial) standard ordered logistic (SOL) and the best spatial model, are presented in Table 4. To reduce clutter, we have simplified the presentation by translating the elasticity effects of variables from the ordered models to a simple binary elasticity effect of variables on the share of teenagers not participating, and participating, in each recreation activity category.[24]

[24] The more detailed elasticity effects for each number of activity episodes (0,1,2) are available from the authors.

Further, we present only the elasticity effect on the "1 or more activity episodes" category in the table (rather then presenting the effect on the "no activity episodes" category too). Thus, the numbers in the table may be interpreted as the percentage change in the share of teenagers participating in recreational activity. For instance, the first number "24.94" corresponding to the "male" variable in the SOL model indicates that the share of male teenagers participating in active recreation is about 25% higher than the share of female teenagers participating in physical activity. Similarly, the number "22.56" corresponding to the "Number of children" variable in the SOL model reflects that an increase in number of children by 1 leads to about a 23% increase in teenager participation in active recreation, while the number "-0.86" for the effect of the "Fraction of African-American population" implies that teenager participation in active recreation decreases by 0.9% due to a 10% increase in the zonal fraction of African-American population.

The elasticity results provide several insights. First, for active recreation, parents' physical activity participation constitutes the most important factor influencing teenagers' physical activity participation levels. This suggests that an effective way to increase active recreation among teenagers would be to direct informational and education campaigns (that raise awareness of the health benefits of active recreation) toward parents, perhaps at special physical education sessions at schools for parents of teenagers studying there. Interestingly, regardless of the sex of the teenager, it is the teenager's mother's physical activity participation that appears to have a higher influence (than the teenager's father's physical activity participation) on the teenager's active recreation participation. Second, another variable with a strong influence on recreation activity participation is the "San Francisco" location dummy variable. Specifically, living in San Francisco County substantially increases active recreation among teenagers, perhaps because of better active recreation opportunities and access to opportunities. On the other hand, living in San Francisco reduces participation in inactive recreation. As indicated earlier, while we considered several built environment and other measures of the residential environment, these are at a coarse geographic level and may not be capturing the micro-urban form attributes that get manifested in the effects of the location dummy variables in the current estimation. Third, the share of teenagers in nuclear family and low income households participating in active recreation is about 50-75% lower than the share of teenagers in non-nuclear family and high income households, respectively, that participate in active recreation. The effect of the "nuclear family" variable may be because teenagers in nuclear family households perceive less independence and feel more "controlled" by parents in their activity

schedules (contributing to less opportunity for physically active free play), while teenagers in non-nuclear families are more independent and participate in more active free-play. Fourth, the presence of natural recreation sites in and around a teenager's residence has a clear and strong impact on active recreation participation. Fifth, there is a higher likelihood of active recreation during the summer season compared to other seasons, and a higher propensity for inactive recreation during the winter season, suggesting that public health policies need to aim at providing more indoor active recreation opportunities at affordable cost to promote year-round teenager physical activity participation. Sixth, individual characteristics (age, race, and student and employment status) appear to play a much more important role in determining inactive recreation participation, while household characteristics play a more dominant role in influencing active recreation participation. This is an interesting result for activity scheduling models, pointing to a more individual-orientated decision process for participation in inactive recreation and a more household interactive influence mechanism for participation in active recreation. Finally, there are differences in the elasticity effects between the SOL and spatial models. This, combined with the better data fit of the spatial model, points to the inconsistent elasticity effects from the SOL model. For instance, for the active recreation category, the SOL model underestimates the influence of gender and family structure, and overestimates the impact of the teenager's father's physical activity participation. Further, the SOL model also underestimates the effect of the presence of natural recreation sites on active recreation participation. There are similar differences in the elasticity effects for the inactive recreation category. Further, note that some of the location variables (Contra Costa and Solano dummy variables) have an impact on active recreation participation in the spatial model, but not in the aspatial SOL, because these location variables appear in the heteroscedasticity specification in the spatial model. The same is the case for the Fall season dummy variable effect for inactive recreation. Overall, ignoring spatial effects, when present, can lead to inconsistent estimation of variable effects that, in turn, can lead to misinformed policy actions.

7. SUMMARY AND CONCLUSIONS

This paper proposes a copula-based ordered-response spatial dependence formulation across decision agents that can incorporate a variety of different kinds of marginal distribution forms for the random terms of each decision agent as well as dependence forms that characterize the multivariate relationship among the decision agents. Regardless of the dependence form used, extant methods in

spatial econometrics become practically infeasible to implement with a moderate-to-large sized sample of decision agents. To address this situation, we propose a simple pseudo-likelihood estimation technique based on a composite marginal likelihood (CML) inference approach to estimate spatial ordered-response discrete choice models. The approach is applicable to data sets of any size, provides standard error estimates for all parameters, and does not require any simulation machinery. It also represents a conceptually and pedagogically simpler procedure relative to current simulation techniques, and has the advantage of reproducibility of the results. The estimation of the asymptotic standard errors and model selection/hypothesis testing procedures are a little more tedious than in the case of the traditional maximum likelihood method, but the appropriate expressions/statistics are presented for spatial econometric models in the current paper. These expressions/statistics are easy to code and implement.

The ability of the CML approach to recover the parameters of a spatially ordered process is evaluated using a simulation study, which clearly points to the effectiveness of the approach. In addition, the combined copula-CML approach is applied to study the daily episode frequency of teenagers' recreational activity participation (both physically active and physically passive), a subject of considerable interest in the transportation, sociology, and adolescence development fields. The data for the analysis is drawn from the 2000 San Francisco Bay Area Survey. Several model forms were tested during the empirical specification, from which the Logistic-Gaussian model emerged as the best specification for both the active and inactive recreation categories. The usual multivariate normal specification for the error terms, as captured in the Normal-Gaussian model, has a poorer fit. This finding highlights the value of the copula approach that is able to separate out the univariate marginal distribution form from the multivariate dependence structure. A further comparison of the aspatial standard ordered logit (SOL) model with the Logistic-Gaussian spatial model indicates the significant presence of heteroscedasticity across observations and spatial dependence between teenager pairs. This underscores the need to consider spatial effects in recreational activity participation to obtain consistent and efficient parameter estimates and elasticity effects.

The variable effects indicate that parents' physical activity participation constitutes the most important factor influencing teenagers' physical activity participation levels, suggesting that one of the most effective ways to increase active recreation among teenagers would be to direct physical activity benefit-related information and education campaigns toward parents, perhaps at special

physical education sessions at schools for parents of teenagers studying there. Another important general result is that individual characteristics (age, race, and student and employment status) are more important than household characteristics in determining teenagers' inactive recreation participation, while household characteristics (number of children, household structure, household income, and parents' recreation participation) are more important determinants of teenagers' active recreation participation.

To summarize, we have proposed a combined copula-CML approach to accommodate, estimate, and test different forms of multivariate dependence in the spatial process underlying observed ordinal discrete choices of decision agents. However, the approach should be very appealing for application to several other multivariate modeling contexts too because it is simple and flexible, and is easy to implement.

ACKNOWLEDGEMENTS

The authors acknowledge the helpful comments of two anonymous reviewers on an earlier version of the paper. Thanks to Lisa Macias for her help in typesetting and formatting this document. This research was partially funded by a Southwest Region University Transportation Center grant.

REFERENCES

Aaron, D.J., Storti, K.L., Robertson, R.J., Kriska, A.M., LaPorte, R.E., 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(11), 1075-1080.

Aas, K., Berg, D., 2009. Models for construction of multivariate dependence – a comparison study. *The European Journal of Finance* 15(7-8), 639-659.

Aas, K., Czado, C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* 44(2), 182-198.

Apanasovich, T.V., Ruppert, D., Lupton, J.R., Popovic, N., Turner, N.D., Chapkin, R.S., Carroll, R.J., 2008. Aberrant crypt foci and semiparametric modeling of correlated binary data. *Biometrics* 64(2), 490-500.

Beck, N., Gleditsch, K.S., Beardsley, K., 2006. Space is more than geography: using spatial econometrics in the study of political economy. *International Studies Quarterly* 50(1), 27-44.

Bellio, R.,Varin, C., 2005. A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling* 5(3), 217-227.

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

Beron, K.J., Murdoch, J.C., Vijverberg, W.P.M., 2003. Why cooperate? Public goods, economic power, and the Montreal protocol. *Review of Economics and Statistics* 85(2), 286-97.

Besag, J.E., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B* 36(2), 192-236.

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., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B* 37(9), 837-855.

Bhat, C.R., 2008. The multiple discrete-continuous extreme value (MDCEV) model: role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B*, 42(3), 274-303.

Bhat, C.R., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B* 43(7), 749-765.

Bhat, C.R., Guo, J.Y., 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., Sener, I.N., 2009. A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. *Journal of Geographical Systems* 11(3), 243-272.

Bhat, C.R., Srinivasan, S., Sen, S., 2006. A joint model for the perfect and imperfect substitute goods case: application to activity time-use decisions. *Transportation Research Part B* 40(10), 827-850.

Campbell, J., 2007. Adolescent identity development: the relationship with leisure lifestyle and motivation. Master of Arts Thesis, Department of Recreation and Leisure Studies, University of Waterloo, Waterloo, Ontario, Canada.

Caragea, P.C., Smith, R.L., 2006. Approximate likelihoods for spatial processes. Technical Report. Department of Statistics. Iowa State University. Available at: http://www.stat.iastate.edu/preprint/articles/2006-18.pdf

Caragea, P.C., Smith, R.L., 2007. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis* 98(7), 1417-1440.

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).

Chaganty, N., Joe, H., 2004. Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society Series B*, 66(4), 851-860.

Copperman, R.B., Bhat, C.R., 2007a. An analysis of the determinants of children's weekend physical activity participation. *Transportation* 34(1), 67-87.

Copperman, R.B., Bhat, C.R., 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.

Cox, D., Reid N., 2004. A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91(3), 729-737.

de Leon, A.R., 2005. Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters* 75(1), 49-57.

Engler, D.A., Mohapatra, M., Louis, D.N., Betensky, R.A., 2006. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* 7(3), 399-421.

Fleming, M., 2004. Techniques for estimating spatially dependent discrete choice models. In: Florax, R., Anselin, L. (Eds.), *Advances in Spatial Econometrics*, 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.

Franzese, R.J., Hays, J.C., 2008. Empirical models of spatial interdependence. In: Box-Steffensmeier, J.M., Brady, H.E., Collier, D. (Eds.), *The Oxford Handbook of Political Methodology*, Oxford University Press Inc., New York.

Godambe, V., 1960. An optimum property of regular maximum likelihood equation. *Annals of Mathematical Statistics* 31, 1208-1211.

Gordon-Larsen, P., McMurray, R.G., Popkin, B.M., 2005. Determinants of adolescent physical activity and inactivity patterns. *Pediatrics*, 105(6), E83.

Gordon-Larsen, P., Nelson, M., Page, P., Popkin, B.M., 2006. Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics*, 117(2), 417-424.

Goulias, K.G., Kim, T., 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.

Guan, Y., 2006. A composite likelihood approach in fitting spatial point process models. *Journal of the American Statistical Association* 101(476), 1502-1512.

Hanfelt, J.J., 2004. Composite conditional likelihood for sparse clustered data. *Journal of the Royal Statistical Society Series B* 66(1), 259-273.

Heagerty, P.J., Lumley, T., 2000. Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association* 95(449), 197-211.

Henderson, R. Shimakura, S., 2003. A serially correlated gamma frailty model for longitudinal count data. *Biometrika* 90(2), 355-366.

Heyde, C.C., 1997. *Quasi-likelihood and its application*. Springer, New York.

Hjort, N.L., Omre, H., 1994. Topics in spatial statistics (with discussion). *Scandinavian Journal of Statistics* 21(4), 289-357.

Hjort, N.L., Varin, C., 2008. ML, PL and QL for markov chain models. *Scandinavian Journal of Statistics* 35(1), 64-82.

Hofferth, S.L., Sandberg, J., 2001. How American children spend their time. *Journal of Marriage and Family* 63(2), 295-308.

Jones, K., Bullen, N., 1994. Contextual models of urban home prices: a comparison of fixed and random coefficient models developed by expansion. *Economic Geography* 70(3), 252-272.

Kent, J.T., 1982. Robust properties of likelihood ratio tests. *Biometrika* 69(1), 19-27.

Klentrou, P., Hay, J., Plyley, M., 2003. Habitual physical activity levels and health outcomes of Ontario youth. *European Journal of Applied Physiology* 89(5), 460-465.

Kuk, A.Y.C., 2007. A hybrid pairwise likelihood method. *Biometrika* 94(4), 939-952.

Lele, S.R., 2006. Sampling variability and estimates of density dependence: a composite-likelihood approach. *Ecology* 87(1), 189-202.

Lele, S.R., Taper, M.L., 2002. A composite likelihood approach to (co)variance components estimation. *Journal of Statistical Planning and Inference* 103(1-2), 117-135.

LeSage, J.P., 2000. Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis* 32(1), 19-35.

Lindsay, B.G., 1988. Composite likelihood methods. *Contemporary Mathematics* 80, 221-239.

Mardia, K.V., Hughes, G., Taylor, C.C., 2007. Efficiency of the pseudolikelihood for multivariate normal and von Mises distributions. University of Leeds, UK. Available at: http://www.amsta.leeds.ac.uk/Statistics/research/reports/2007/STAT07-02.pdf

Marsh, H.W., Kleitman, S., 2003. School athletic participation: mostly gain with little pain. *Journal of Sport and Exercise Psychology* 25(2), 205-228.

McMillen, D.P., 1995. Spatial effects in probit models: a monte carlo investigation. In: Anselin, L., Florax, R. (Eds.), *New Directions in Spatial Econometrics*, pp. 189-228, 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.

Molenberghs, G., Verbeke, G., 2005. *Models for discrete longitudinal data*. Springer Science + Business Media, Inc., New York.

MORPACE International, Inc., 2002. Bay area travel survey final report, March, ftp://ftp.abag.ca.gov/pub/mtc/planning/BATS/BATS2000/BATS%20Final%20Report/Vol%20I/Volume%20I_Tab1.pdf, accessed January 20, 2010.

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

Nelson, M.C., Gordon-Larsen, P., 2006. Physical activity and sedentary behavior patterns are associated with selected adolescent health risk behaviors. *Pediatrics* 117(4), 1281-1290.

Oman, S.D., Landsman, V., Carmel, Y., Kadmon, R., 2007. Analyzing spatially distributed binary data using independent-block estimating equations. *Biometrics* 63(3), 892-900.

Ornelas, I.J., Perreira, K.M., Ayala, G.X., 2007. Parental influences on adolescent activity: a longitudinal study. *The International Journal of Behavioral Nutrition and Physical Activity* 4:3.

Paelinck, J.H.P., 2005. Spatial econometrics: history, state-of-the-art and challenges ahead. Keynote Paper for the Workshop on Spatial Econometrics. Kiel Institute for World Economics, April.

Páez, A., 2007. Spatial perspectives on urban systems: developments and directions. *Journal of Geographic Systems* 9(1), 1-6.

Páez, A., Scott, D., 2004. Spatial statistics for urban analysis: a review of techniques with examples. *GeoJournal* 61(1), 53-67.

Pinkse, J., Slade, M.E., 1998. Contracting in space: an application of spatial statistics to discrete-choice models. *Journal of Econometrics* 85(1), 125-154.

Sener, I.N., Copperman, R.B., Pendyala, R.M., Bhat, C.R., 2008. An analysis of children's leisure activity engagement: examining the day of week, location, physical activity level, and fixity dimensions. *Transportation* 35(5), 673-696.

Sklar, A., 1973. Random variables, joint distribution functions, and copulas. *Kybernetika*, 9, 449-460.

Tiggemann, M., 2001. The impact of adolescent girls' life concerns and leisure activities on body dissatisfaction, disordered eating and self-esteem. *The Journal of Genetic Psychology* 162(2), 133-142.

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., Zimmer, D.M., 2007. Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1), Now Publishers.

Vandekerkhove, P., 2005. Consistent and asymptotically normal parameter estimates for hidden markov mixtures of markov models. *Bernoulli* 11(1), 103-129.

Varin, C., 2008. On composite marginal likelihoods. *Advances in Statistical Analysis* 92(1), 1-28.

Varin, C., Czado, C., 2008. Modeling pain severity diaries with mixed autoregressive ordinal probit models. Available at: http://www-m4.ma.tum.de/Papers/Czado/varin_czado_pain_diaries.pdf

Varin, C., Vidoni, P., 2005. A note on composite likelihood inference and model selection. *Biometrika* 92(3), 519-528.

Varin, C., Vidoni, P., 2006. Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics & Data Analysis* 51(4), 2365-2373.

Varin, C., Vidoni, P., 2009. Pairwise likelihood inference for general state space models. *Econometric Reviews* 28(1-3), 170-185.

Varin, C., Høst, G., Skare, Ø., 2005. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics & Data Analysis* 49(4), 1173-1191.

Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models and the gauss-newton method. *Biometrika* 61(3), 439-447.

Zhao, Y., Joe, H., 2005. Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics* 33(2), 335-356.

LIST OF TABLES

Table 1. CML Estimation Results with 25 samples of 500 Observations

Case 1: ς = -0.50

Parameter	True Value	Mean CML estimate	Bias	Relative Bias (%)	RMSE
β_1	1.0000	1.0052	-0.0052	0.5228	0.0793
β_2	0.5000	0.4981	0.0019	0.3832	0.0696
β_3	0.2500	0.2490	0.0010	0.4080	0.0561
ψ_0	-0.7500	-0.7504	0.0004	0.0544	0.1678
ψ_1	0.2500	0.2563	-0.0063	2.5328	0.1899
ψ_2	1.2500	1.2494	0.0006	0.0480	0.1959
ς	-0.5000	-0.5011	0.0011	0.2224	0.0020

Case 2: ς = 0.50

Parameter	True Value	Mean CML estimate	Bias	Relative Bias (%)	RMSE
β_1	1.0000	1.0203	-0.0203	2.0300	0.0829
β_2	0.5000	0.5203	-0.0203	4.0512	0.0752
β_3	0.2500	0.2570	-0.0070	2.7952	0.0554
ψ_0	-0.7500	-0.7768	0.0268	3.5771	0.2185
ψ_1	0.2500	0.2516	-0.0016	0.6544	0.2233
ψ_2	1.2500	1.2714	-0.0214	1.7155	0.2577
ς	0.5000	0.4918	0.0082	1.6368	0.0098

Table 2. Model Selection based on the CLIC Statistic

Recreation Category	Distribution	Copula	Log-composite likelihood	Trace of the matrix in the CLIC statistic	CLIC statistic
Active recreation (40 miles)	Normal	FGM	-432,318.74	10,629.02	-442,947.77
		Gaussian	-432,370.84	10,303.54	-442,674.37
	Logistic	FGM	-433,256.40	11,441.04	-444,697.44
		Gaussian	-433,233.25	3,839.25	-437,072.49
Inactive recreation (151.46 miles)	Normal	FGM	-1,249,341.25	8,035.03	-1,257,376.28
		Gaussian	-1,251,218.01	7,807.39	-1,259,025.39
	Logistic	FGM	-1,247,454.36	7,754.06	-1,255,208.42
		Gaussian	-1,247,475.00	4,120.94	-1,251,595.94

Table 3. Estimation Results for the Number of Recreational Activity Episodes

	Active Recreation		Inactive Recreation	
Variable	**Parameter**	**t-stat**	**Parameter**	**t-stat**
Threshold parameters				
Threshold 1	3.114	23.31	1.357	23.01
Threshold 2	5.781	31.03	3.131	29.14
Individual characteristics				
Male	0.386	4.19	-	-
Age greater than 15	-	-	0.285	5.73
Hispanic	-0.438	-1.86	-0.523	-4.56
Asian	-	-	-0.244	-3.18
Part time student	-	-	0.397	2.59
Employed	-	-	-0.272	-4.38
Household characteristics				
Number of children	0.299	6.15	-	-
Nuclear family	-0.631	-5.98	-	-
Household income less than 35k	-1.423	-4.65	-	-
Household income greater than 90k	-	-	0.131	2.77
Teenager's father physically active	0.933	5.57	-	-
Teenager's mother physically active	1.471	12.32	-	-
Household location, season and activity-day variables				
San Francisco County	1.732	7.55	-0.651	-2.34
Alameda County	-	-	-0.228	-1.98
Summer	0.502	4.81	-	-
Winter	-0.420	-1.74	0.799	13.98
Friday	-	-	0.280	5.82

Table 3 (Continued) Estimation Results for the Number of Recreational Activity Episodes

	Active Recreation		Inactive Recreation	
Variable	**Parameter**	**t-stat**	**Parameter**	**t-stat**
Residential neighborhood variables				
Fraction of African-American population	-2.395	-2.69	-	-
Presence of natural recreation sites such as county/state/national parks, gardens, nature centers	0.740	6.12	-	-
Bicycling facility density multiplied with number of bikes in the household (miles of bike lanes per square mile)	0.011	2.48	-	-
(Spatial) heteroscedasticity variables				
Winter	0.125	1.96	-0.915	-11.12
Fall	-	-	-0.283	-7.44
Alameda County	-	-	0.144	1.91
Contra Costa	0.268	5.49	-	-
San Francisco	-0.485	-2.81	-	-
Solano	-0.187	-2.42	-	-
Spatial dependence variable				
ζ in the θ parameter "*Inverse of distance between zonal centroids*"	-1.690	-22.88	-1.116	-23.63
Number of Observations	1447		1447	
Trace of G	1.2452		0.1783	
Log-composite likelihood at convergence	-433,233.25		-1,247,475.00	
Trace of the matrix in the CLIC statistic	3,839.25		4,120.94	
Penalized log-composite likelihood (PLCL)	-437,072.49		-1,251,595.94	

Table 4. Aggregate-level Elasticity Effects of the Aspatial SOL and Spatial Models

Variable	Active Recreation		Inactive Recreation	
	SOL	Spatial	SOL	Spatial
	1 or more episodes	1 or more episodes	1 or more episodes	1 or more episodes
Individual characteristics				
Male	24.94	30.70	-	-
Age greater than 15	-	-	21.76	26.44
Hispanic	-31.33	-30.54	-33.71	-40.28
Asian	-	-	-28.83	-20.69
Part time student	-	-	35.15	40.95
Employed	-	-	-24.16	-23.49
Household characteristics				
Number of children	22.56	26.08	-	-
Nuclear family	-46.38	-51.07	-	-
Household income less than 35k	-61.45	-72.65	-	-
Household income greater than 90k	-	-	12.96	12.08
Teenager's father physically active	126.20	94.37	-	-
Teenager's mother physically active	166.49	166.76	-	-
Household location, season, and activity-day variables				
San Francisco	136.40	165.94	-27.23	-46.74
Alameda County	-	-	-6.63	-5.18
Contra Costa	0.00	43.07	-	-
Solano	0.00	-26.29	-	-
Summer	42.54	42.33	-	-
Winter	1.36	-11.44	19.92	15.50
Fall	-	-	0.00	-28.66
Friday	-	-	16.33	26.90
Residential neighborhood variables				
Fraction of African-American population	-0.86	-0.77	-	-
Presence of natural recreation sites such as county/state/national parks, gardens, nature centers	41.99	68.26	-	-
Bicycling facility density multiplied with # of bikes in the household (miles of bike lanes per square mile)	0.53	0.66	-	-