Transportation Research Part B 56 (2013) 217–233

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Development of an indicator to assess the spatial fit of discrete choice models

Antonio Páez [a,*], Fernando A. López [b], Manuel Ruiz [b], Catherine Morency [c]

[a] School of Geography and Earth Sciences, McMaster University, Canada
[b] Departamento de Métodos Cuantitativos e Informáticos, Universidad Politécnica de Cartagena, Spain
[c] Département des génies civil, géologique et des mines, École Polytechnique de Montréal, Canada

ARTICLE INFO

Article history:
Received 11 January 2013
Received in revised form 12 August 2013
Accepted 13 August 2013

Keywords:
Discrete choice models
Spatial analysis
Spatial fit
Q statistic
Auto ownership
Montreal

ABSTRACT

Discrete choice models are increasingly implemented using geographical data. When this is the case, it may not be sufficient to project market shares accurately, but also to correctly replicate the spatial pattern of choices. Analysts might then be interested in assessing the results of a model's fit relative to the spatial distribution of the observed responses. While canonical approaches exist for the exploratory spatial analysis of continuous variables, similar tools have not been widely implemented for discrete choice models, where the variable of interest is categorical. For this reason, despite recent progress with spatial models for discrete outcomes, there is still not a simple and intuitive tool to assess the quality of the spatial fit of a discrete choice model. The objective of this paper is to introduce a new indicator of spatial fit that can be applied to the results of discrete choice models. Utility of the indicator is explored by means of numerical experiments and then demonstrated by means of a case study of vehicle ownership in Montreal, Canada.

1. Introduction

Discrete choice models are the standard approach for the analysis of decision-making processes in transportation research, and many other disciplines as well. For decades, beginning with the work of McFadden (1974), models of this type have been used in the study of travel demand, travel behavior research, activity analysis, residential choice modeling, and land use analysis, to name just a few domains of application. More recently, as geo-referenced information has become more widely available, it can be seen that the number of discrete choice models that use spatial data has increased. Incorporating space more effectively into the behaviorally rich framework of discrete choice analysis has in fact been identified as an open area for research (Páez, 2007; Páez and Scott, 2004). Not surprisingly, with progress in data, techniques, and technology, there has been an attendant growth in the number of studies that emphasize the spatial dimension of discrete choices – in particular the possibility that outcomes display non-random spatial patterns, or *spatial association*. This includes, *inter alia*, the analysis of residential choices with spatially autocorrelated models (e.g. Bhat and Guo, 2004; Miyamoto et al., 2004), spatial analysis as a complement for destination choice analysis (e.g. Hammadou et al., 2008), the study of activity participation using spatial correlation structures (e.g. Bhat and Sener, 2009; Bhat et al., 2010), the effect on mode choice of socio-spatial dependencies (e.g. Dugundji and Walker, 2005; Goetzke, 2008; Sidharthan et al., 2011), the spatial association in sense of community (e.g. Whalen et al., 2012), and various studies of land use change with spatially explicit models (e.g. Chakir

* Corresponding author at: School of Geography and Earth Sciences, McMaster University, 1280 Main Street West, Hamilton, Ontario, L8S 3Z9, Canada. Tel.: +1 905 525 9140 x 26099; fax: +1 546 0463.
E-mail addresses: paezha@mcmaster.ca (A. Páez), fernando.lopez@upct.es (F.A. López), manuel.ruiz@upct.es (M. Ruiz), cmorency@polymtl.ca (C. Morency).

0191-2615/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trb.2013.08.009

and Parent, 2009; Frazier and Kockelman, 2005; Páez, 2006; Páez and Suzuki, 2001; Wang and Kockelman, 2009; Zhou and Kockelman, 2008).

As the number of applications with spatial data has grown so has the need to assess the spatial fit of the resulting models. In spatially-explicit applications, it may not be sufficient for the model to accurately replicate market segments. An analyst may well be interested in the ability of the model to replicate the spatial pattern of the observed responses. For instance, is the model able to predict the similarity in choices by individuals who are proximately located? Or in a model of land uses, are the correct combinations of land uses projected for neighboring tracts of land? While the spatial analysis literature provides canonical approaches to assess the spatial fit of models for continuous variables (Anselin, 1988; Bailey and Gatrell, 1995; Griffith, 1988; Haining, 1990; LeSage and Pace, 2009), the same cannot be said of models for discrete choices. Besides the join-count statistic (Dacey, 1968), until recently there were only limited efforts to develop tools for the exploratory spatial data analysis of categorical variables (e.g. Boots, 2006). As a consequence, the considerable progress that has been observed in the study of discrete choices using spatial data, has not been accompanied by diagnostics for assessing spatial fit. It seems clear, then, that there is a need for a simple and intuitive tool to assess the spatial fit of discrete choice models.

The objective of this paper is to develop an indicator of spatial fit for discrete choice models. We build on the $Q(m)$ statistic, introduced by Ruiz et al. (2010). $Q(m)$ can be used to assess the strength and significance of spatial association (i.e. non-random spatial pattern) of a geo-referenced categorical variable. In simple terms, the idea of an indicator of spatial fit is to use $Q(m)$ as a summary measure of spatial association, to compare whether the projected outcome has more than, less than, or a similar level of spatial association as the original variable. We propose that the indicator of spatial fit should be easy to interpret, and amenable to hypothesis testing. In this paper we present a new indicator of spatial fit. Further, to conduct a test of hypothesis regarding the differences in the level of spatial association between two categorical variables, we derive new results regarding the distribution of $Q(m)$ under spatial association. Numerical experiments demonstrate the utility of the approach under a broad range of circumstances. Finally, the approach is demonstrated by means of a case study of automobile ownership in Montreal, Canada. The simulations and empirical results suggest that the indicator can be a valuable addition to the set of diagnostics for discrete choice models when working with spatial data.

2. Methods

2.1. Spatial association of a categorical variable

$Q(m)$ was introduced by Ruiz et al. (2010) as a tool to assess the spatial association of categorical variables, with additional technical details provided by Páez et al. (2012). It is generally applicable to a variable $\boldsymbol{Y}$ that is the outcome of a spatial discrete process. In other words, each realization y of the variable can take one and only one of k different values, or categories, say $a_1, a_2, \ldots, a_k$. Each of these values is recorded a definite number of times, say $n_1, n_2, \ldots, n_k$, with $N = \sum_{j=1}^{k} n_j$. Further, the realizations are recorded at sites $i = 1,2,\ldots,N$ with coordinates $\boldsymbol{s}_i \in \Omega$.

One way to assess the spatial association of variable $\boldsymbol{Y}$ is to embed the realization y at a reference location, so that it becomes one element of a group of m observations that are spatially proximate. If we designate the reference location as $\boldsymbol{s}_0$, this embedding includes, in addition to the outcome at $\boldsymbol{s}_0$, the outcomes at $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{m-1}$. These coordinates are the $m-1$ nearest neighbors of $\boldsymbol{s}_0$, defined in terms of their distance from $\boldsymbol{s}_0$. In the case of ties, additional tie-breaking protocols can be invoked, such as direction. The spatial embedding so obtained can be denoted as a string in the following way:

$$\boldsymbol{Y}_{m}(\boldsymbol{s}_0) = (y_{s_0}, y_{s_1}, \ldots, y_{s_{m-1}}) \tag{1}$$

The string in (1) is called an m-surrounding. The m-surrounding is a summary of the spatial configuration of outcomes of $\boldsymbol{Y}$ in the vicinity of $\boldsymbol{s}_0$. Since there are k possible values that $\boldsymbol{Y}$ can take, and the number of elements in the surrounding is m, it is straightforward to see that there are exactly $\lambda_\sigma = k(k+1)\ldots(k+m-1)/m!$ unique configurations, in terms of the number of outcomes of each class that can be found in an m-surrounding.

Consider for example (see Fig. 1) the case where $k = 2$ (i.e. the outcomes are a_1 and a_2) and $m = 3$ (i.e. the surroundings are triads of observations). The possible combinations are: (1) three of a_1; (2) two of a_1 and one of a_2 (see panel (*i*) in Fig. 1); (3) one of a_1 and two of a_2 (see panel (*iii*) in Fig. 1); and (4) three of a_2 (see panel (*ii*) in Fig. 1). Each of these unique combinations of outcomes can be represented in a compact way by means of an abstract term, called a symbol, and denoted by σ.

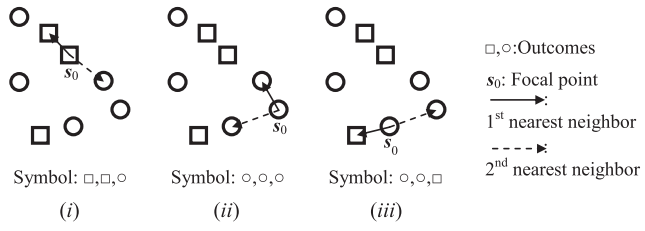


Fig. 1. Examples of spatial embedding with $k = 2$ outcomes (a_1 = □ and a_2 = ○) and surroundings of size $m = 3$.

Continuing with the example, there are four symbols, say σ_1 through σ_4, each representing one of the four unique combinations. The process of visiting an observation, identifying its embedding, and assigning a symbol to it, is called *symbolization*. Once this is done, the location is said to be of σ_i-type, the corresponding symbol. In principle, every observation can be symbolized, or a sub-set of observations S to reduce the degree of overlap between proximate m-surroundings (i.e. the number of common neighbors; more on this below). After symbolization, it is possible to count the number of locations $\boldsymbol{s}_i$ that are of σ_i-type, or X_{σ_i}, as follows:

$$X_{\sigma_i} = \sum_{j=1}^{S} I_j \tag{2}$$

where I_j is an indicator function taking the value of 1 if location j is of σ_i-type and 0 otherwise. Dividing by the number of sites that were symbolized (i.e. $S \leqslant N$) gives the relative frequency of the symbol:

$$p_{\sigma_i} = \frac{X_{\sigma_i}}{S} \tag{3}$$

The level of association in the spatial distribution of $\boldsymbol{Y}$ can be summarized by the *symbolic entropy* of the discrete spatial process for a fixed $m \geqslant 2$. Using Shannon's formula, the empirical entropy is:

$$h(m) = -\sum_j p_{\sigma_j} \ln(p_{\sigma_j}) \tag{4}$$

It can be appreciated that the empirical entropy tends to zero when the outcomes are strongly spatially associated. This is the case when one or a small number of symbols appear with high frequency, which could be similar outcomes in proximity or dissimilar outcomes in proximity. It follows that $p_{\sigma_i} \to 1$ and $p_{\sigma_j} \to 0$ for all $j \neq i$, which implies that $p_{\sigma_i} \ln(p_{\sigma_i}) \to 0$ and $p_{\sigma_j} \ln(p_{\sigma_j}) \to 0$. Intuitively, the entropy is low because the map is highly predictable. As the map becomes less organized, the entropy tends to increase, up to a maximum value corresponding to a random spatial sequence. The formula for the value of the symbolic entropy for a random spatial sequence, call it $\eta(m)$, is derived by Ruiz et al. (2010) as follows:

$$\eta(m) = -\sum_i \frac{X_{\sigma_i}}{S} \sum_{j=1}^{k} \alpha_{ij} \ln(q_j) \tag{5}$$

where α_{ij} is the number of times that class a_j appears in symbol σ_i and q_j is the probability that $y = a_j$.

The $Q(m)$ statistic can be used to perform a test of spatial independence. In other words, it can be used to assess whether the outcomes of variable $\boldsymbol{Y}$ are spatially distributed in a random or non-random fashion. The test is designed as a likelihood ratio, and compares the empirical symbolic entropy of $\boldsymbol{Y}$ to the expected symbolic entropy under the null hypothesis of spatial independence:

$$Q(m) = 2S(\eta(m) - h(m)) \tag{6}$$

Clearly, the value of $Q(m)$ is bounded between zero when $\boldsymbol{Y}$ is spatially independent, and $2S\eta(m)$ when $\boldsymbol{Y}$ is very strongly spatially associated. Since $Q(m)$ is asymptotically χ^2 distributed with degrees of freedom equal to the number of symbols minus one (see Ruiz et al., 2010), the decision rule to reject the null hypothesis is to compare the value of $Q(m)$ to the critical value of the χ^2 distribution at a desired confidence level and using the appropriate number of degrees of freedom.

In practice, the analyst must decide the size of m to implement $Q(m)$.[1] This is done attending to the following considerations. First, for a given k, the number of symbols increases with larger values of m. This has three implications: (1) approximation to the χ^2 distribution can be poor if the ratio of the number of symbolized locations to the number of symbols (S/λ_σ) is less than five (Agresti, 1990; p. 49); (2) a large number of symbols can consume degrees of freedom available for inference quite rapidly; and (3) interpretation can be challenging with many symbols. Therefore m must be selected so that $S/\lambda_\sigma \geqslant 5$, inference is not hampered, and interpretability is deemed reasonable. Beyond this, the analyst enjoys flexibility to select m based on theoretical or conceptual considerations (analogous to selecting a pattern of contiguities in spatial econometrics; see Anselin, 1988).

Another practical consideration refers to the number of observations to symbolize (i.e. S). Visiting every observation implies a potentially large degree of overlap between proximate m-surroundings, a situation that increases the risk of false positives (Ruiz et al., 2010). To address this, the analyst can decide to limit the amount of overlap by symbolizing $S < N$ observations. A procedure to symbolize S observations without overlap is as follows:

1. Create a distance matrix $\mathbf{W}$ based on the coordinates of locations in set Ω.
2. Select a location $\boldsymbol{s}_0$ at random from set Ω.
3. Select the m locations in Ω that are the nearest neighbors of $\boldsymbol{s}_0$ according to distance matrix $\mathbf{W}$, namely $(\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{m-1}, \boldsymbol{s}_m)$.
4. The first m-surrounding is formed by $(y_{s_0}, y_{s_1}, \ldots, y_{s_{m-1}})$.

[1] Note that $Q(m)$ is currently defined only in terms of nearest neighbors, and does not allow for other contiguity schemes, such as inverse distance. This is a matter for future research.

5. Consider s_m and select the m locations in $\Omega \setminus (\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{m-1})$ that are the nearest neighbors of $\boldsymbol{s_m}$ according to distance matrix $\mathbf{W}$, namely $(\boldsymbol{s_{m+1}}, \ldots, \boldsymbol{s}_{2m-1}, \boldsymbol{s}_{2m})$.
6. The second m-surrounding is formed by $(y_{s_m}, y_{s_{m+1}}, \ldots, y_{s_{2m-1}})$.
7. Consider $\boldsymbol{s}_{2m}$ and select the m locations in $\Omega \setminus (\boldsymbol{s}_0, \ldots, \boldsymbol{s}_{2m-2}, \boldsymbol{s}_{2m-1})$ that are the nearest neighbors of $\boldsymbol{s}_{2m}$ according to distance matrix $\mathbf{W}$, namely $(\boldsymbol{s}_{2m+1}, \ldots, \boldsymbol{s}_{3m-1}, \boldsymbol{s}_{3m})$.
8. The following m-surrounding is given by $(y_{s_{2m}}, y_{s_{2m+1}}, \ldots, y_{s_{3m-1}})$.
9. Continue the process until no more locations are available to construct an m-surrounding.

Limiting the overlap can affect the power of the statistic, by reducing the number of observations S used in the analysis. There are a number of ways to work around this. For instance, additional rules may be introduced as part of the symbolization process that implicitly reduce the potential for overlap (for instance, by considering directionality). Alternatively, the classification scheme may be tweaked, for instance by considering one outcome versus the rest, and repeating for every outcome in turn. More simply, a non-significant result under a small degree of overlap may be checked by changing the degree of overlap, to assess whether loss of power is involved.

2.2. Indicator of spatial fit and hypothesis testing

Goodness-of-fit statistics for discrete choice models are typically based on the value of the log-likelihood of the model at convergence. These include summary measures such as McFadden's (adjusted) ρ^2, Nagelkerke R^2, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or tests of hypothesis that compare models, such as the likelihood ratio test. None of these statistics satisfies the sufficiency criterion, since they do not summarize the spatial information of the observed data (Griffith, 1988).

In this section we introduce a new indicator, namely $\Phi(m)$ (or Φ for short) that can be used to measure, for a given spatial embedding m, the degree of spatial fit of discrete choice models. This indicator is based on the $Q(m)$ statistic previously described. Suppose that a variable exists that has been observed as the outcome of a discrete choice process, say $\boldsymbol{Y}$, and that observations have been geo-referenced. Further, suppose that the outcome variable is used, in conjunction with a set of explanatory variables, to estimate a discrete choice model. The model then provides, for each observation, a vector of probabilities that the outcome belongs to class $a_1, a_2, \ldots, a_k$. These probabilities in turn can be used to obtain projections of the outcome variable, say $\widehat{\mathbf{Y}}$. Our interest is in assessing the spatial fit between the observed variable and the projected variable. This can be done by means of the following spatial fit indicator:

$$\Phi = \frac{\widehat{Q}(m) - Q(m)}{\widehat{Q}(m) + Q(m)} \tag{7}$$

where $\widehat{Q}(m)$ is the value of the statistic computed for the variable obtained from the discrete choice model ($\widehat{\mathbf{Y}}$), and $Q(m)$ is the value of the statistic for the observed variable ($\boldsymbol{Y}$).

The spatial fit indicator is bounded as: $-1 \leqslant \Phi \leqslant 1$. Negative values of Φ indicate that the projected variable $\widehat{\mathbf{Y}}$ is underfitted, i.e., it is more spatially random than the observed variable $\boldsymbol{Y}$. In contrast, positive values of Φ mean that the projected variable is over-fitted, i.e., it has a stronger pattern of association than the true variable. A value of zero indicates that the level of spatial association is identical for the two variables. Values of $|\Phi|$ close to 1 indicate a higher degree of under- or over-fit, either because the observed or the projected variable is closer to being spatially random.

The spatial fit indicator in Eq. (7) is a simple and intuitive summary measure to assess whether the degree of spatial association of a variable projected using a discrete choice model is similar or different (and if so, in which direction), relative to the observed variable. It would be useful, as well, to give statistical significance to the differences in the levels of spatial association. In other words, since discrete choice models incorporate an element of randomness, we wish to be able to assess how close to zero the difference must be before we are able to state, with a certain level of confidence, that the levels of spatial association in $\widehat{\mathbf{Y}}$ and $\boldsymbol{Y}$ are statistically equivalent.

To develop a test of hypothesis, the moments of the distribution are needed (i.e. the mean and variance). Ruiz et al. (2010) derived these terms for $Q(m)$ under the null hypothesis of spatial independence. In the present context, however, a variable $\boldsymbol{Y}$ (a set of discrete choices) is of interest *because* spatial association is present. For this reason, existing expressions for the mean and variance of $Q(m)$ are not of utility, and the asymptotic distribution of $Q(m)$ under spatial association must be derived.

We begin by noting that $Q(m)$ can be rewritten as:

$$Q(m) = f(X_{\sigma_1}, \ldots, X_{\sigma_{\lambda_\sigma}}) = 2\sum_{i=1}^{\lambda_\sigma} X_{\sigma_i} log\left(\frac{X_{\sigma_i}}{Sp^0_{\sigma_i}}\right) \tag{8}$$

In this equation, $p^0_{\sigma_i}$ is the probability of the symbol σ_i under the null hypothesis of independence, and X_{σ_i} is the random variable counting the number of locations that are of σ_i-type. If we denote $E(X_{\sigma_i}) = \mu_i = Sp_{\sigma_i}$, then the second order Taylor expansion of f at a neighborhood of $(\mu_1, \ldots, \mu_{\lambda_\sigma})$ is given by:

$$f(X_{\sigma_1}, X_{\sigma_2}, \ldots, X_{\sigma_{\lambda_\sigma}}) = f(\mu_1, \ldots, \mu_{\lambda_\sigma}) + \sum_{i=1}^{\lambda_\sigma} f_{X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma})(X_{\sigma_i} - \mu_i) + \frac{1}{2}\sum_{i=1}^{\lambda_\sigma} f_{X_{\sigma_i}X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma})(X_{\sigma_i} - \mu_i)^2 - \sum_{i \neq j} f_{X_{\sigma_i}X_{\sigma_j}}(\mu_1, \ldots, \mu_{\lambda_\sigma})(X_{\sigma_i} - \mu_i)(X_{\sigma_j} - \mu_j) + reminder \quad (9)$$

In the case in which $Q(m) = f(X_{\sigma_1}, \ldots, X_{\sigma_{\lambda_\sigma}})$ we have that:

$$f_{X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) = 2\left(1 + log\left(\frac{p_{\sigma_i}}{p^0_{\sigma_i}}\right)\right) \quad (10)$$

$$f_{X_{\sigma_i}X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) = \frac{2}{Sp_{\sigma_i}} \quad (11)$$

$$f_{X_{\sigma_i}X_{\sigma_j}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) = 0 \quad (12)$$

Therefore we obtain the following approximation for the estimation of the expected value of $Q(m) = f(X_{\sigma_1}, \ldots, X_{\sigma_{\lambda_\sigma}})$:

$$\mu_Q = E(Q(m)) \approx f(\mu_1, \ldots, \mu_{\lambda_\sigma}) + \frac{1}{2}\sum_{i=1}^{\lambda_\sigma} f_{X_{\sigma_i}X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Var(X_{\sigma_i}) \quad (13)$$

Similarly the variance can be estimated by:

$$\begin{aligned}\sigma^2_Q &= Var(Q(m)) \\ &\approx \sum_{i=1}^{\lambda_\sigma} f^2_{X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Var(X_{\sigma_i} - \mu_i) + \frac{1}{4}\sum_{i=1}^{\lambda_\sigma} f^2_{X_{\sigma_i}X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Var((X_{\sigma_i} - \mu_i)^2) \\ &\quad + \sum_{i \neq j} f_{X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) f_{X_{\sigma_j}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Cov(X_{\sigma_i} - \mu_1, X_{\sigma_j} - \mu_j) \\ &\quad + \frac{1}{2}\sum_{i \neq j} f_{X_{\sigma_i}X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) f_{X_{\sigma_j}X_{\sigma_j}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Cov((X_{\sigma_i} - \mu_i)^2, (X_{\sigma_j} - \mu_j)^2) \\ &\quad + \frac{1}{2}\sum_{ij}^{2} f_{X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) f_{X_{\sigma_j}X_{\sigma_j}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Cov(X_{\sigma_i} - \mu_i, (X_{\sigma_j} - \mu_j)^2)\end{aligned} \quad (14)$$

Notice that since the joint distribution of $(X_{\sigma_1}, \ldots, X_{\sigma_n})$ is multinomial, it is possible to estimate the following variances and covariances, which we use in the computation of the mean and variance of $Q(m)$:

$$Var(X_{\sigma_i}) = Sp_{\sigma_i}(1 - p_{\sigma_i}) \quad (15)$$

$$Var((X_{\sigma_i} - Sp_{\sigma_i})^2) = Sp_{\sigma_i}(1 - p_{\sigma_i})(1 + S^2 - p_{\sigma_i}(4 + S) + p^2_{\sigma_i}(8 + S)) \quad (16)$$

$$Cov(X_{\sigma_i}, X_{\sigma_j}) = -Sp_{\sigma_i}p_{\sigma_j} \quad (17)$$

$$Cov(X_{\sigma_i} - Sp_{\sigma_i}, (X_{\sigma_j} - Sp_{\sigma_j})^2) = Sp_{\sigma_i}p_{\sigma_j}(2p_{\sigma_j} - 1) \quad (18)$$

$$Cov((X_{\sigma_i} - Sp_{\sigma_i})^2 (X_{\sigma_j} - Sp_{\sigma_j})^2) = 2S^2 p^2_{\sigma_i} p^2_{\sigma_j} - Sp_{\sigma_i}p_{\sigma_j}(6p_{\sigma_i}p_{\sigma_j} - 2p_{\sigma_i} - 2p_{\sigma_j} + 1) \quad (19)$$

By using (10), (11), (12), (15), (16), (17), (18), and (19) we get that $\frac{1}{2}\sum_{i=1}^{\lambda_\sigma} f_{X_{\sigma_i}X_{\sigma_i}}(\mu_1, \ldots, \mu_{\lambda_\sigma}) Var(X_{\sigma_i}) = n - 1$ and therefore:

$$\mu_Q = E(Q(m)) \approx f(\mu_1, \ldots, \mu_{\lambda_\sigma}) + n - 1 \quad (20)$$

Since each variable X_{σ_i} is sum of weakly dependent random variables, it follows that under the alternative hypothesis of association in the spatial process, by the Delta method and by the Central Limit Theorem for dependent variables (Hoeffding and Robbins, 1948) we can approximate the asymptotic distribution of $Q(m)$ to a normal distribution $N(\mu_Q, \sigma_Q)$ where μ_Q and σ_Q can be calculated using the preceding expressions.

Next, we are interested in testing the following null hypothesis:

H_0. The vector of projected outcomes $\hat{\mathbf{Y}}$ displays the same level of spatial association as Y.

To test for this null hypothesis, and therefore to test whether $\hat{\mathbf{Y}}$ is spatially associated to the same degree as $\mathbf{Y}$, it is possible to conduct a two-tailed test in the usual way as follows:

$$z = \frac{\widehat{Q} - \mu_Q}{\sigma_Q} \tag{21}$$

Alternatively, the confidence interval for μ_Q at a $100(1-\alpha)\%$ confidence level can be calculated as:

$$IC_Q = (\mu_Q - z_{\alpha/2}\sigma_Q, \mu_Q + z_{\alpha/2}\sigma_Q) \tag{22}$$

where $z_{\alpha/2}$ is defined such that $P(N(0,1) \geqslant z_{\alpha/2}) = \alpha/2$. The interval of confidence can be examined to see if it contains $\widehat{Q}$. An interval for Φ can be calculated based on the cut-off (interval of confidence), which has the same interpretation and can be examined to see if it contains the value of zero.

There are two reasons that could explain rejection of the null hypothesis. The first reason could be a failure of the model to project outcomes of each type that provide a reasonable approximation of actual market shares. In this case, the number of outcomes of each type may exceed or be insufficient to form the local configurations required to replicate the frequency of symbols observed in the sample. Alternatively, the model could project a reasonable approximation of market shares, but fail to capture the spatial association in the response. In other words, the number of outcomes of each type may be sufficient, but the local configurations are not replicated by the model. To what extent either of these situations is present in a specific application can be explored by means of more detailed analysis of the frequency of symbols.

Failure to reject the null hypothesis is possible only if the model captures the pattern of spatial association sufficiently well, which implies that the model projects the market shares reasonably well too. An attractive feature of the test is that the implications of not rejecting the null hypothesis are precise.

3. Montecarlo simulations

In this section we conduct a series of numerical experiments to assess the performance of the approach in a broad array of circumstances.

3.1. Experimental design

To design the numerical experiments, let us stipulate a decision-making situation with N individuals, each of whom, when faced with a choice, selects one and only one of k available alternatives (i.e. $a_1, a_2, \ldots, a_k$). Each individual evaluates the utility associated with each alternative and selects the one that gives the maximum utility. According to random utility theory, this rational decision making process is observed with some stochastic noise by the analyst.

Three potential sources of spatial pattern are identified in the literature.

First, the simplest, and most common reason of spatial association in the outcomes, is the presence of common environmental factors that affect the alternatives, decision-makers, or both. This could lead in turn to patterning of the responses. Inclusion of such factors in a model should help retrieve information relevant to the process, and contribute towards replicating the spatial pattern of responses. Models of this type can be estimated using standard methods (e.g. Alemu and Tsutsumi, 2011).

Secondly, omission of spatially patterned relevant factors gives rise to a structure commonly referred to in the spatial econometrics literature as spatial error autocorrelation (Anselin, 1988). Models that incorporate this type of structure include Miyamoto et al. (2004)'s study of location choices, Wang et al. (2012)'s model of land use change, and Sener and Bhat (2012)'s analysis of activity participation.

Third, spatial pattern could emerge when decision makers care about, or are influenced by, the decisions made by others. Models of this type are the rough analog of the spatial autoregressive model in spatial econometrics (Anselin, 1988). Examples include vehicle type selection (Adjemian et al., 2010; Paleti et al., 2013) and mode choice (Goetzke, 2008; Goetzke and Rave, 2011). Recent developments in estimation (e.g. Bhat, 2011) facilitate the implementation of spatial processes in discrete choice models.

For the experiment we assume that individuals derive utility as per the following set of utility functions with $j = 2,3,\ldots,k$ (taking alternative $j = 1$ as the reference):

$$\boldsymbol{u}_{(N\times k)\times 1} = \underbrace{\sum_{j=2}^{k} a_j \mathbf{1}_{N\times 1} \otimes \boldsymbol{H}^j_{k\times 1} + \sum_{j=1}^{k} b_j \boldsymbol{x}^j_{N\times 1} \otimes \boldsymbol{H}^j_{k\times 1}}_{\textit{non-spatial term}} + \underbrace{\sum_{j=2}^{k} c_j \boldsymbol{z}_{N\times 1} \otimes \boldsymbol{H}^j_{k\times 1}}_{\textit{spatial term}} + \boldsymbol{\varepsilon}_{(N\times k)\times 1} \tag{23}$$

In Eq. (23), the vector $\boldsymbol{u} = (u_{qj})$ contains the utility for qth individual and jth alternative. The term $\mathbf{1}_{N\times 1}$ is a vector of ones, $\boldsymbol{H}^j_k$ is a matrix of zeros of order $k \times 1$ with the jth element set to one, and $\boldsymbol{x}^j_N$ is a vector of values drawn from the distribution $U(0,1)$. These terms are not, by design, spatially associated. The third term on the right hand side of Eq. (23) incorporates a variable with spatial association, to simulate a common environmental factor. This term is generated based on a spatial random variable, using a purely auto-normal formulation as follows:

$$z_{N\times 1} = (\mathbf{I} - \delta\mathbf{W})^{-1}\boldsymbol{\mu}_{N\times 1} \text{ with } \mu_{N\times 1} \sim N(0,1) \tag{24}$$

Matrix **W** is a non-negative spatial weights matrix ($N \times N$) that captures the spatial topology of the system. In the present case, we randomly draw coordinates for each individual to locate them in the unit square. Based on these coordinates, a tessellation is generated which forms the basis for specifying **W**. In this matrix, a value of one in cell w_{ij} indicates that observations i and j are spatially proximate, whereas a value of zero indicates that they are not spatially proximate. Proximity is defined in terms of first order neighbors, to the fourth degree of separation. The spatial setup ensures that the topology of the system is reminiscent of a real-world spatial system (see Farber et al., 2009). As is usually the case, matrix **W** is row-standardized, that is, each element of the matrix is divided by the sum of the corresponding row. Parameter δ modulates the intensity of the spatial effect. Here we consider values between 0 and 1, to generate patterns of association whereby proximate outcomes tend to be similar. This is the most common form of spatial association in social and environmental processes.

Selection of parameters a_j, b_j, and c_j determine the utility that individual q derives from alternative j. Finally, the random part of the utility (ε) is drawn from the Type I Extreme Value distribution with mean zero and standard deviation 1, to give a multinomial logit model. These random terms are modulated by a factor of 0.2 so that their relative size with respect to the systematic utility is on average 0.08%. The parameters for the experiment are as follows:

- Four levels of N: 400, 900, 2500, and 4900.
- Two levels of k (number of alternatives): 3 and 4.
- Four levels of δ (see Eq. (24)): 0.2, 0.5, 0.7, and 0.9.
- For $k = 3$, parameters $a_j = [0.0, 1.0, 2.0]$, $b_j = [3.0, 1.0, 3.0]$, and $c_j = [0.0, 3.0, -3.0]$ (these values are selected to give overall shares of the outcomes of 12%, 26%, and 60%); for $k = 4$, parameters $a_j = [0.0; 1.0; 2.0; 2.0]$; $b_j = [4.0; 3.3; 0.7; 2.0]$, and $c_j = [0.0; 3.0; -3.0; 3.0]$ (these values are selected to give shares of the alternatives of 16%, 30%, 41% and 11%).
- Three sizes of m-surroundings: 3, 4, and 6, with zero overlap.
- Each combination of parameters is repeated in 1000 replications.

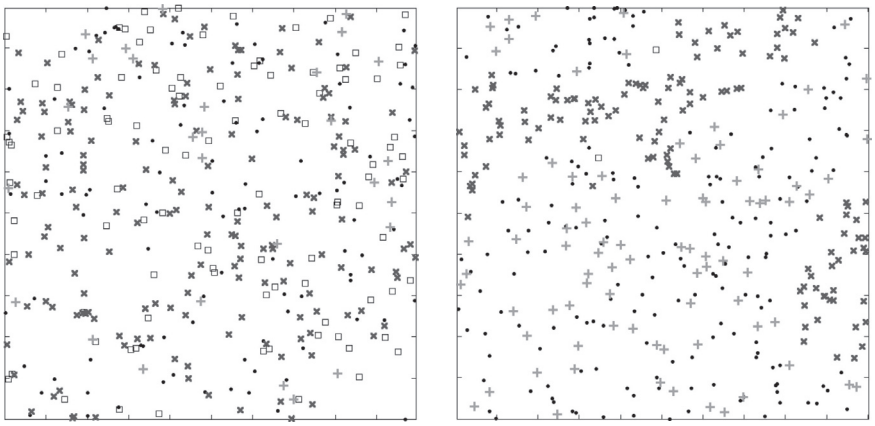
Furthermore, actual choices are compared to the estimated values based on the models. The relation between individual utilities and chosen alternatives y is:

$$y_q = a_j \quad \text{if } u_{qj} = \max(u_{q1}, \ldots, u_{qk}) \tag{25}$$

Fig. 2 provides two examples of variables obtained using this scheme. In panel (i) there is a variable with moderately weak spatial association, whereas panel (ii) shows a variable with a strong pattern of association.

3.2. Size and power of test

To explore the performance of the test, and the ability of Φ to assess the spatial fit of models, the data generation process is retrieved by estimation of multinomial logit models. A number of model specifications are tested. First, an exhaustive model M_F^* is estimated that contains all variables used in the data generation process. This model represents the best case scenario, whereby all relevant variables are correctly included in the specification. Secondly, we consider the case where the (spatial) variable $\boldsymbol{y}$ is not available to the researcher or is otherwise missing. We denote this situation M_0^*. Third, we explore the case where $\boldsymbol{y}$ is available, but the modeler does not correctly specify the model, by failing to incorporate it in all relevant utility functions. In this case, spatial information is available but only partially used. This is entered in the estimation by



(i) δ=0.2 (ii) δ=0.9

Fig. 2. Examples of spatial association in the outcomes of the data generation process (moderate-weak spatial association: $\delta = 0.2$; strong spatial association: $\delta = 0.9$).

using only the term $z_{N\times 1} \otimes \boldsymbol{H}^{j}_{k\times 1}$ corresponding to utility j, and the models are denominated M^{*}_{j} with $j = 2,3,4$, as appropriate (since $j = 1$ is the reference).

Inferential values are reported in terms of the proportion of times that the test rejects the null hypothesis at $p < 0.05$. As well, the values corresponding to the 5th, 50th, and 95th percentile of Φ for each combination of parameters are reported.

Tables 1 and 2 show the test's power and size for $k = 3$ and $k = 4$ with different embedding dimensions (m) and sample sizes (N). For each value of m, the first column (grey) shows information about test size, that is, the proportion of times that the analyst rejects the null hypothesis when the model is supplied with all relevant information (M^{*}_{F}). This is the risk of false positives. Test size is greater than the nominal value of 0.05 in the presence of strong spatial association and small sample sizes. Size values tend towards nominal values as N increases. This result is expected, since when data are strongly grouped together, the multinomial distribution under the alternative hypothesis of association (according to which $Q(m)$ is constructed) presents overdispersion (Agresti, 1990, p. 42). The test size also tends to decrease when the length of the m-surrounding increases. There are two reasons for this. First, when large m-surroundings are used, increasingly distant observations are symbolized, thus reducing the overdispersion effect. Secondly, when very long m-surroundings are used, it is possible to find highly unlikely symbols, leading to undersized tests (Agresti, 1990). In any event, in the most common situations, where spatial association is low or moderate, the test size is correct.

Table 1
Power and size of Φ with $k = 3$ alternatives.

	ρ	$m = 3$ ($\lambda_\sigma = 10$)				$m = 4$ ($\lambda_\sigma = 15$)				$m = 6$ ($\lambda_\sigma = 28$)			
		M^{*}_{F}	M^{*}_{2}	M^{*}_{3}	M^{*}_{0}	M^{*}_{F}	M^{*}_{2}	M^{*}_{3}	M^{*}_{0}	M^{*}_{F}	M^{*}_{2}	M^{*}_{3}	M^{*}_{0}
$N = 400$	0.2	0.017	0.015	0.016	0.330	0.015	0.026	0.012	0.588	–	–	–	–
	0.5	0.032	0.109	0.039	0.805	0.026	0.031	0.039	0.571	–	–	–	–
	0.7	0.062	0.187	0.109	0.946	0.074	0.198	0.132	0.999	–	–	–	–
	0.9	0.175	0.421	0.456	0.999	0.192	0.457	0.430	1.000	–	–	–	–
$N = 900$	0.2	0.022	0.080	0.024	1.000	0.017	0.047	0.015	0.999	0.000	0.001	0.000	0.960
	0.5	0.053	0.170	0.100	1.000	0.038	0.148	0.070	1.000	0.001	0.009	0.002	0.997
	0.7	0.089	0.303	0.194	1.000	0.067	0.247	0.159	1.000	0.025	0.151	0.072	0.894
	0.9	0.169	0.704	0.590	1.000	0.100	0.580	0.520	1.000	0.110	0.610	0.580	1.000
$N = 2500$	0.2	0.024	0.588	0.063	1.000	0.033	0.471	0.043	1.000	0.001	0.417	0.006	1.000
	0.5	0.057	0.421	0.124	1.000	0.050	0.370	0.108	1.000	0.021	0.388	0.060	1.000
	0.7	0.031	0.550	0.072	1.000	0.040	0.556	0.078	1.000	0.051	0.581	0.082	1.000
	0.9	0.117	0.907	0.787	1.000	0.118	0.882	0.731	1.000	0.097	0.830	0.738	1.000
$N = 4900$	0.2	0.035	0.885	0.112	1.000	0.024	0.840	0.059	1.000	0.000	0.069	0.000	1.000
	0.5	0.058	0.721	0.182	1.000	0.044	0.679	0.143	1.000	0.013	0.561	0.064	1.000
	0.7	0.056	0.864	0.124	1.000	0.047	0.852	0.171	1.000	0.039	0.599	0.237	1.000
	0.9	0.101	0.991	0.967	1.000	0.087	0.983	0.950	1.000	0.085	0.955	0.897	1.000

Notes: Approximate percentages of classes 5% ($k = 1$); 40% ($k = 2$); 55% ($k = 3$). Model M^{*}_{F}: Full model; M^{*}_{j} = Model estimated with $\boldsymbol{z}$ in alternative j only; Model M^{*}_{0}: Model estimated without $\boldsymbol{z}$.

Table 2
Power and size of Φ with $k = 4$ alternatives.

	ρ	$m = 3$ ($\lambda_\sigma = 20$)					$m = 4$ ($\lambda_\sigma = 35$)					$m = 6$ ($\lambda_\sigma = 84$)				
		M^{*}_{F}	M^{*}_{2}	M^{*}_{3}	M^{*}_{4}	M^{*}_{0}	M^{*}_{F}	M^{*}_{2}	M^{*}_{3}	M^{*}_{4}	M^{*}_{0}	M^{*}_{F}	M^{*}_{2}	M^{*}_{3}	M^{*}_{4}	M^{*}_{0}
$N = 400$	0.2	0.048	0.259	0.062	0.291	0.262	0.019	0.100	0.039	0.086	0.023	–	–	–	–	–
	0.5	0.070	0.466	0.149	0.537	0.426	0.025	0.371	0.052	0.278	0.192	–	–	–	–	–
	0.7	0.106	0.953	0.252	0.932	1.000	0.078	0.777	0.158	0.801	1.000	–	–	–	–	–
	0.9	0.218	1.000	0.854	0.999	1.000	0.189	0.995	0.779	0.997	1.000	–	–	–	–	–
$N = 900$	0.2	0.058	0.356	0.109	0.343	0.509	0.032	0.155	0.074	0.227	0.046	0.000	0.022	0.004	0.019	0.002
	0.5	0.072	0.929	0.234	0.885	1.000	0.043	0.627	0.120	0.800	0.985	0.001	0.343	0.005	0.376	0.559
	0.7	0.099	0.999	0.387	0.990	1.000	0.080	0.995	0.286	0.982	1.000	0.006	0.800	0.291	0.792	1.000
	0.9	0.156	1.000	0.967	1.000	1.000	0.129	1.000	0.937	1.000	1.000	0.116	1.000	0.900	1.000	1.000
$N = 2500$	0.2	0.068	0.837	0.198	0.742	0.742	0.055	0.566	0.104	0.746	0.746	0.009	0.272	0.016	0.214	0.214
	0.5	0.064	1.000	0.422	0.999	0.999	0.072	0.996	0.346	0.996	0.996	0.030	0.968	0.169	0.938	0.938
	0.7	0.085	1.000	0.573	1.000	1.000	0.063	1.000	0.532	1.000	1.000	0.035	1.000	0.984	1.000	1.000
	0.9	0.131	1.000	1.000	1.000	1.000	0.101	1.000	1.000	1.000	1.000	0.098	1.000	0.996	1.000	1.000
$N = 4900$	0.2	0.058	0.981	0.436	0.989	0.989	0.057	0.988	0.320	0.978	0.978	0.028	0.762	0.150	0.846	0.846
	0.5	0.060	1.000	0.492	1.000	1.000	0.063	1.000	0.467	1.000	1.000	0.036	1.000	0.347	1.000	1.000
	0.7	0.073	1.000	0.768	1.000	1.000	0.084	1.000	0.752	1.000	1.000	0.040	1.000	1.000	1.000	1.000
	;0.9	0.125	1.000	1.000	1.000	1.000	0.101	1.000	1.000	1.000	1.000	0.071	1.000	1.000	1.000	1.000

Notes: Approximate percentages of classes 16% ($k = 1$), 30% ($k = 2$), 41% ($k = 3$), 11% ($k = 4$). Model M^{*}_{F}: Full model; M^{*}_{j} = Model estimated with $\boldsymbol{z}$ in alternative j only; Model M^{*}_{0}: Model estimated without $\boldsymbol{z}$.

The test's power depends on the quantity of spatial information supplied to the model. The test rapidly reaches maximum power when the model is estimated without the spatial variable (M_0^* column), which indicates the ability to identify missing spatial information. In the case of $k = 3$, with sample sizes of 2500 and greater, power is always at its greatest. The same occurs for $k = 4$ with sample sizes of 4900.

If partial spatial information is provided, the test's power depends on the category's importance in the model. If the spatial information is provided for an option with a high percentage of occurrence relative to the reference category, the power of the test increases. For example, for $k = 3$, $N = 2500$, $m = 3$ and $\delta = 0.7$, the proportions of each alternative are 12.9% ($k = 1$), 26.7% ($k = 2$), and 60.3% ($k = 3$). If $z_{N\times 1} \otimes \boldsymbol{H}_{k\times 1}^2$ is included in the model as category $k = 2$ it has little effect on the utility function when spatial information is supplied and the test's power increases, with 55.0% rejection of the null hypothesis. On the other hand, if the information corresponding to $z_{N\times 1} \otimes \boldsymbol{H}_{k\times 1}^3$ is included in the model, it represents most of the spatial pattern in the utility function, as the third category ($k = 3$) has great weight in the model, so the test's power is reduced by 7.2%. This pattern of behavior is maintained irrespective of sample sizes, embedding dimensions, spatial association, and the number of categories.

The embedding dimension is also a key factor that affects the test's behavior. Firstly, this dimension should not exceed certain limits in small samples, as high values of m involve a very large number of symbols. For example, for $N = 400$ it is not sensible to use $m = 6$ as we would obtain only 400/6 = 64 symbolized observations for a total of 84 equivalent symbols. The relationship between the number of symbolized observations and the number of symbols must be greater than 5 and our experience indicates that more robust results are obtained when the ratio between the number of symbolized observations and the number of symbols is greater than 10. There is often a slight decrease in power as the size of the m-surrounding increases, due to the reduction in the number of locations that are symbolized.

As discussed above, the sign and size of Φ also provide relevant information about the spatial association that is missing/exceeds that in the observed variable. Fig. 3 shows the distribution of Φ after 1000 iterations for each model, together with the table showing the percentiles for $N = 4900$, $k = 4$, $m = 4$, and two values of δ: 0.2 and 0.5. Similar results are obtained in the other cases.

In situations where the model is supplied with all information (M_F^*), the expected value[2] of the indicator is zero $E[\Phi] = 0$. Consistent with this result, the boxplots in Fig. 3 show how the values of Φ are symmetrically distributed around zero. The variability of the indicator is also smaller as spatial association increases, due to the increase in the mean value of the $Q(m)$ statistic in the model. This situation is also expected when considering the variance of the indicator.

For the other cases, the median values of Φ are negative, reflecting the fact that the spatial association of the projected outcomes is inferior to that of the true model. This is more marked as spatial association increases. For example, for $\delta = 0.5$ when information pertaining to the fourth alternative M_4^* is included in the utility function, the median value of the indicator is -0.542 and the graph contains no value close to zero (in this case the test's power was 1, see Table 2). On the other hand, if information pertaining to the third alternative M_3^* corresponding to the most common category (41%) is included, the median value is -0.073 and the Φ indicator presents a high percentage of values around zero (test power in this case was only 0.467).

As a result of this Monte Carlo, it would seem appropriate to provide some guidelines for use. First, a large enough ratio should be established between the number of symbolized observations and the number of symbols. This ratio should be at least 5, and higher if possible. Secondly, use the $Q(m)$ statistic (Ruiz et al., 2010) to assess the degree of spatial association. If the statistic's value is very high relative to its expected value, very strong spatial association can be suspected, so large values of m would prevent false negative (oversize) problems. On the contrary, if medium or low levels of spatial association are suspected, lower values of m tend to increase the power of the test. In any event, it makes sense to test different values of m. Special attention should also be paid to the least frequent categories, as they will induce even less frequent symbols and can lead to an undersized test.

3.3. Relationship to McFadden's ρ^2

An intriguing question is the potential relationship between the new spatial fit indicator, and other established indicators of fit. Conventionally, indicators of fit are derived based on the value of the log-likelihood function, and include McFadden's ρ^2, Akaike Information Criterion, Bayesian Information Criterion, and Nagelkerke's R^2. These indicators differ

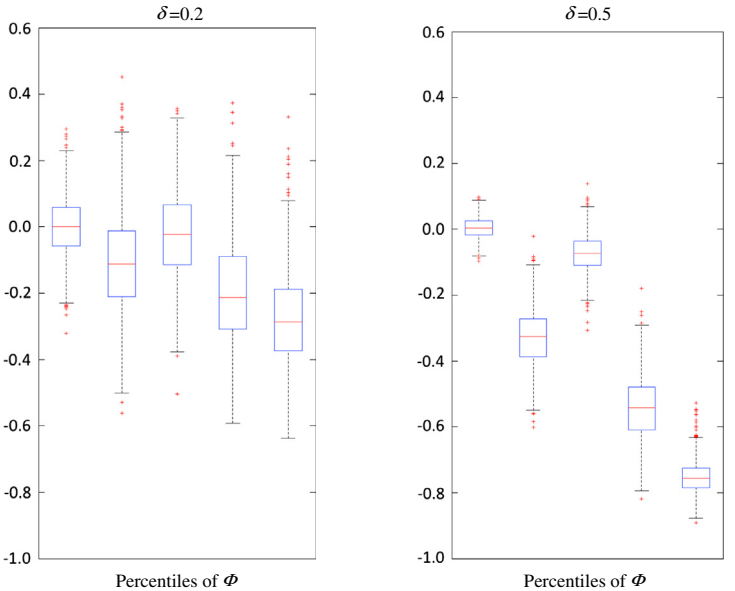
[2] The moments of this indicator may be obtained in function of the moments of Q and $\widehat{Q}$

$$E[\Phi] = \frac{\mu_{\widehat{Q}} - \mu_Q}{\mu_{\widehat{Q}} + \mu_Q} + \frac{\sigma_{\widehat{Q}}^2 - \sigma_Q^2}{(\mu_{\widehat{Q}} + \mu_Q)^2} + \frac{(\sigma_{\widehat{Q}}^2 + \sigma_Q^2 + 2\sigma_{\widehat{Q}Q}^2)(\mu_{\widehat{Q}} - \mu_Q)}{(\mu_{\widehat{Q}} + \mu_Q)^3}$$

$$Var[\Phi] = \frac{\sigma_{\widehat{Q}}^2 + \sigma_Q^2 - 2\sigma_{\widehat{Q}Q}^2}{(\mu_{\widehat{Q}} + \mu_Q)^2} - \frac{2(\sigma_{\widehat{Q}}^2 - \sigma_Q^2)(\mu_{\widehat{Q}} - \mu_Q)}{(\mu_{\widehat{Q}} + \mu_Q)^3} + \frac{(\sigma_{\widehat{Q}}^2 + \sigma_Q^2 + 2\sigma_{\widehat{Q}Q}^2)(\mu_{\widehat{Q}} - \mu_Q)^2}{(\mu_{\widehat{Q}} + \mu_Q)^4}$$

in case $\mu_{\widehat{Q}} = \mu_Q$ and $\sigma_{\widehat{Q}}^2 = \sigma_Q^2$ we get $E[\Phi] = 0$ and $Var[\Phi] = (\sigma_Q^2 - \sigma_{QQ}^2)/2\mu_Q^2$

.



	M_F^*	M_2^*	M_3^*	M_4^*	M_0^*
P_5	-0.147	-0.369	-0.227	-0.449	-0.476
P_{50}	0.001	-0.112	-0.022	-0.212	-0.286
P_{95}	0.140	0.144	0.199	0.066	-0.035

	M_F^*	M_2^*	M_3^*	M_4^*	M_0^*
P_5	-0.049	-0.469	-0.168	-0.690	-0.825
P_{50}	0.004	-0.326	-0.073	-0.542	-0.756
P_{95}	0.058	-0.186	0.025	-0.381	-0.657

Fig. 3. Bloxplots and percentiles of spatial fit indicator. Case $N = 4900$, $k = 4$, $m = 4$.

in the way the likelihood function is used, and how strongly they penalize the number of parameters. Intuitively, it is plausible that the log-likelihood of the model will increase with the inclusion of relevant variables, however these may not necessarily improve the spatial fit if their spatial pattern is irrelevant to the process. Conversely, it is possible that variables can increase the spatial fit, with only relatively small gains in the value of the log-likelihood function.

In what follows, we compare McFadden's ρ^2 to Φ. In a numerical experiment, we consider a sample size of 900, $k = 3$ categories, and $m = 3$. The level of spatial association is moderate ($\delta = 0.5$). Further, we consider 10 different values of parameters of the non-spatial term in model (23), namely b_j ($j = 1,2,3$). The rest of the parameters in model (23) remain as before. For each parameter configuration we have computed 100 pairs (ρ^2, Φ).

Fig. 4 shows a scatterplot of the resulting values of McFadden's ρ^2 (horizontal axis) and Φ (vertical axis). As seen in the figure, there is no a clear relationship between these two indicators of fit. Since in this example we estimate the full model,

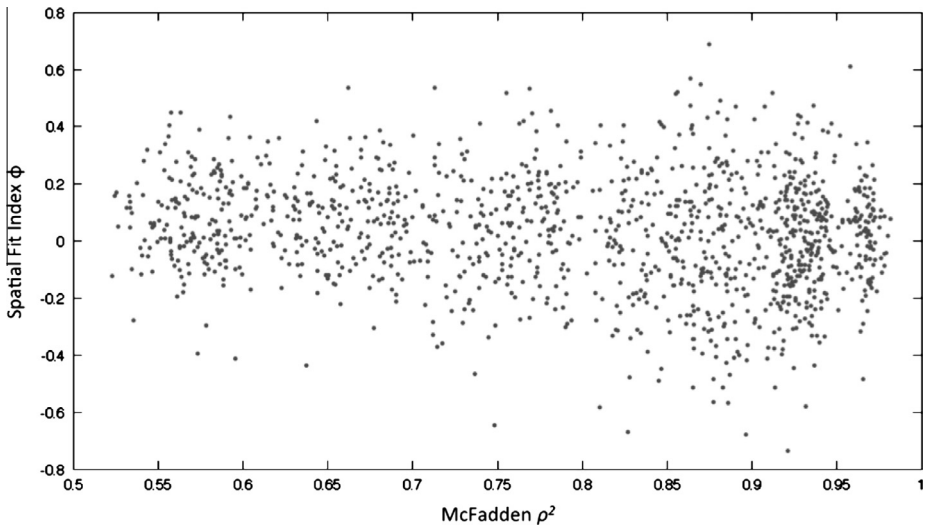


Fig. 4. Scatterplot ρ^2 versus Φ.

the spatial fit indicator should have an absolute value close to zero (in the rejection zone). Clearly, even for large values of ρ^2, the absolute value of Φ remains close to zero. This suggests that McFadden's ρ^2 and the spatial fit indicator give complementary information about model fit.

4. Example

In this section we use an empirical example to demonstrate the use of the indicator of spatial fit introduced in the preceding section. The case study is concerned with household vehicle ownership in Montreal, Canada.

Vehicle ownership is an important variable in travel behavior. In the specific case of Montreal, it is known to influence trip generation (Roorda et al., 2010) and trip length (Morency et al., 2011). More generally, it has been found that vehicle ownership can depress the use of alternative modes of transportation and local consumption of goods and services (Giuliano and Dargay, 2006).

The preferred approach to model vehicle ownership at the level of individual households is by means of discrete choice models. Bhat and Pulugurta (1998) compared two plausible behavioral mechanisms for household decisions regarding vehicle ownership, namely the ordinal and multinomial structures. These authors observe that the ordinal model is parsimonious, but may oversimplify by assuming a single propensity variable. The multinomial model, on the other hand, invokes trade-offs between alternatives and is consistent with global utility maximization. Furthermore, after extensive experiments, Bhat and Pulugurta conclude that the multinomial logit model appears to be more flexible in capturing elasticity patterns between alternatives. On the downside, a loss of efficiency may be incurred if the process is indeed ordinal. In the present case we weigh the risk of losing some efficiency against the large sample size, and prefer to use the multinomial logit model for the appealing features noted by Bhat and Pulugurta.

It is important to note that, while the analysis below uses the multinomial logit, the indicator of spatial fit is generally applicable to any choice structure, as long as the responses are categorical.

4.1. Data

Data used for the analysis are drawn from Montreal's Travel Survey of 2003. This survey is part of one of the largest ongoing travel data collection efforts in the world. Conducted approximately every five years since 1970, in 2003 the survey was deployed using Computer Assisted Telephone Interviews to reach a 4.7% sample of a population consisting of 3.61 million residents in the Greater Montreal Area. In addition to travel information, and socio-economic and demographic attributes of households and individual household members, a key aspect of this survey is that origins and destinations are geocoded at a high level of resolution, using various types of identifiers such as trip generators, addresses, or nearest intersections.

For the purpose of the analysis, a sub-sample is extracted from the larger dataset, corresponding to all residents in Montreal Island. This yields a total of $N = 34{,}027$ households. The dependent variable is number of vehicles per household, coded as follows (with proportion of sample in parentheses): 0 vehicles (29.33%), 1 vehicle (46.89%), 2 vehicles (20.17%), and 3 or more vehicles (3.61%). The number of households with 4 or more vehicles is extremely small (only 0.76%) and is therefore collapsed into the same class as 3 vehicles. The information provided by the travel survey is complemented with information from the Census at the level of Dissemination Areas (the smallest publicly available Census geography), and neighborhood attributes obtained from detailed geographic files for the region. For the analysis below, we consider three broad classes of independent variables: Socio-Economic and Demographic, Urban Structure and Built Environment, and Locational variables. These are defined in Table 3.

4.2. Analysis and results

The first step in the analysis is to apply $Q(m)$ to the dependent variable $\boldsymbol{Y}$. For the purpose of the example, this is done using surroundings of size $m = 3$ (experiments have been conducted using $m = 4$ and $m = 5$, and the results are in general agreement). The number of symbols with $k = 4$ (the number of outcomes) and $m = 3$ is 20. Further, the analysis is conducted without overlap between proximate m-surroundings, which yields $S = 11{,}342$. As noted above, limiting the overlap reduces the risk of false positives.

The ratio of the number of symbolized locations to number of symbols is 567.10 (since $S = 11{,}342$ and $\lambda_\sigma = 20$), well in excess of the recommended minimum value of 5. The value of $Q(m)$ for $\boldsymbol{Y}$ is calculated as 1956.11. This value can be used to reject the null hypothesis of a random spatial sequence with a very high level of confidence ($p < 0.0001$).

The pattern of spatial association of the variable can be examined in more detail by means of the frequency of symbols along with intervals of confidence for the frequency of individual symbols (see Páez et al., 2012). In Fig. 5, the vertical axis is the frequency, and the horizontal axis is the list of symbols, which are read as follows: the first digit of each symbol is the number of neighbors (in a surrounding of size 3) with zero vehicles; the second digit is the number of neighboring households with 1 vehicle; and so on. For instance, 1020 indicates that in a group of three, one household owns zero vehicles and two households own two vehicles each, whereas 0021 means that two households own two vehicles and one household owns three or more vehicles. The whiskers are the 95% confidence intervals of the frequencies. Accordingly, white bars

Table 3
Variable definitions.

Socio-economic and demographic variables		
Household structure	These variables take the value of 1 for household of the corresponding type and 0 otherwise	Couple without children; couple with children; single parent with children; other types of multi-person households
Income	These variables take the value of 1 if the household income is of the corresponding level and 0 otherwise	Income less than $20,000; income range $20,000–$40,000; Income range $40,000–$60,000; income range $60,000–$80,000; income range $80,000–$100,000; income greater than $100,000; respondent refused to answer, or did not know income
Occupation	Number of household members by occupation type	Full-time work; part-time work; student; retired; stay-at-home;
Driver License	Number of household members who hold a driver license	Driver License
Urban structure and built environment variables		
Activity density	Density class of Dissemination Area at place of residence	Low population density; medium population density; high population density; low employment density; medium employment density; high employment density
Land use mix	Land use mix in Dissemination Area at place of residence	Land use mix: calculated using the entropy formulation, where P_i is the proportion of land use of type i in zone and n is the number of different land uses considered: $-\sum p_i \ln(P_i)/\ln(n)$
Transit	Presence of a transit facility within 500 m of place of residence	Transit
Feature density	Feature density in dissemination area at place of residence	Street density in km/km^2; Intersection density in units/km^2; total built up area divided by DA area
Locational variables		
Distance to CBD	Distance to CBD in 100's of km	Distance to Central Business District (CBD)
Coordinates	Coordinates of place of residence in 100's of km using a false origin	XC: easting; YC: northing

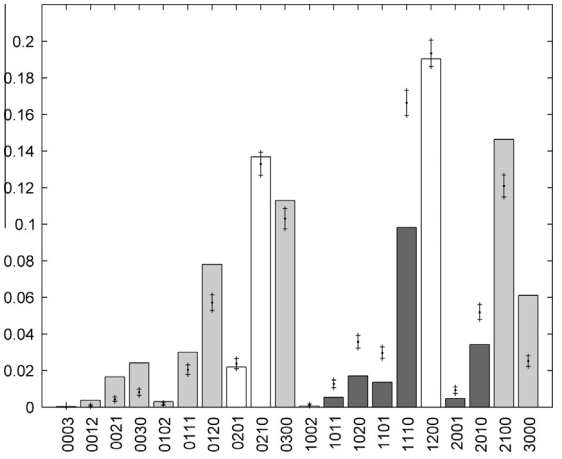


Fig. 5. Analysis of symbolic frequency, dependent variable *Y* (dark gray bars are symbols that are observed significantly less frequently than expected under the null; light gray bars are symbols that are observed significantly more frequently than expected under the null).

correspond to symbols whose frequency is not significantly different from the expectation, whereas dark gray and light gray bars are symbols with frequencies below or above the expectation, respectively.

Examination of the figure makes it clear that there is a high degree of affinity in vehicle ownership levels among neighboring households. The symbols that appear with significantly higher frequency than expected include those where all households in the *m*-surrounding are in the same vehicle ownership category, namely three households with no vehicles (3000), with one vehicle (0300), and with two vehicles (0030). As well, it is frequent to observe little variation in vehicle ownership levels among neighboring households. In this way, we find that a common occurrence is to observe two households that are in the same category of vehicle holdings, next to a neighbor in an adjacent category (see 2100, 0120, 0021, and 0012). Relatively rare are instances where neighboring households greatly differ in their levels of vehicle ownership. For instance, it is uncommon (significantly less frequent than expected) to see three households, each in a different ownership class (1110, 1101, 1011). The only case where the opposite is true is when no households without car are present in the *m*-surrounding (0111).

Various factors could potentially contribute to explain the high degree of spatial association observed in vehicle ownership by households in Montreal. This includes spatial affinity by level of income, or similarity in terms of family structure due to broader demographic processes (i.e. birds of a feather flocking; McPherson et al., 2001). Concurrently, there could be shared environmental factors that exert similar influence over the decision to own vehicles (e.g. if households respond similarly to attributes such as density and land uses). Yet another possibility is that social referencing processes, such as status-seeking or herd behavior (cf. Páez and Scott, 2007; Smirnov, 2010), exert some influence on auto ownership decisions (Axsen and Kurani, 2012; Goetzke and Weinberger, 2012). One purpose of the modeling effort is to identify which factors are significant covariates of the behavior, as well as their relative effect on decision making.

Four models are estimated, to illustrate the effect of adding (alternatively removing) variables, as part of a specification search. Once a model has been estimated, there are a several ways to convert the probabilities into a vector of projected outcomes $\widehat{\mathbf{Y}}$. A crude way of doing this is to identify, for each individual, the alternative with the highest probability, and then assigning the outcome to that alternative. In what follows, instead, we simulate the outcomes by using the estimated probabilities and a multinomial random number generator.

The first specification that we assess is the constants-only model (Model 0). This type of naïve model is never used in practice, except as a benchmark to assess the goodness of fit of fully specified models. The model has a complete set of alternative specific constants, using as reference the category corresponding to zero vehicles. The value of ρ^2 is 0.1650. A vector of projected outcomes, call it $\widehat{\mathbf{Y}}_0$, is generated and analyzed for spatial association using $Q(m)$ under identical terms as previously defined for $\boldsymbol{Y}$ (i.e. $m = 3$ and zero overlap). The spatial fit indicator is $\Phi = -0.9856$, indicating that the projected outcomes are almost perfectly random. As seen in Fig. 6, while the model replicates the general distribution of market shares (as the constants-only model is wont to do), when mapped, the projected responses lack spatial association (compare to Fig. 5).

Three models using groups of independent variables are reported in Table 4. Note that only variables that are significant at the 0.05 level for at least one utility function are retained in these models. The results of the analysis are in general agreement with our understanding of vehicle ownership decisions, including the importance of considering household structure, occupation and licensing of household members, as well as the positive effect of income (the probability of owning more cars increases with increasing income). In terms of urban structure and built environment, higher density, both of population and employment, tends to facilitate the decision to own fewer vehicles. One feature of the built environment, land use mix, is significant, and with the anticipated sign: greater diversity of land uses decreases the probability of owning more vehicles. Finally, we calculate distance to the Central Business District (CBD), and use the coordinates as independent variables to create a quadratic trend surface (Bailey and Gatrell, 1995). The shape of the trend surface follows the east–west direction, which is the general orientation of Montreal Island.

The results of estimating the models indicate that, as expected, the goodness of fit tends to increase with the use of additional variables. Given the large sample size, there are only negligible differences when adjusting ρ^2 to account for number of parameters. There is a large gain in goodness of fit (relative to the constants-only model) when introducing all socio-economic, demographic, and urban structure and built environment variables (Models 1 and 2). The improvement in the goodness of fit when incorporating the trend surface, on the other hand, is more modest ($\rho^2 = 0.450$ increases to $\rho^2 = 0.459$). Another measure of goodness of fit, the Bayesian Information Criterion, confirms the gains in fit with the introduction of each additional set of variables.

After obtaining projected outcomes for Models 1, 2, and 3, call these $\widehat{\mathbf{Y}}_1$, $\widehat{\mathbf{Y}}_2$, and $\widehat{\mathbf{Y}}_3$, Φ can be calculated. In contrast to ρ^2, the spatial fit of the models shows a more marked change as variables are added. Models 1 and 2 lead to projected outcomes that are under-fitted, and the interval of Φ does not include the value of zero. Model 3, in contrast, is slightly over-fitted with the addition of a trend surface, but the interval in this case includes the value of zero.

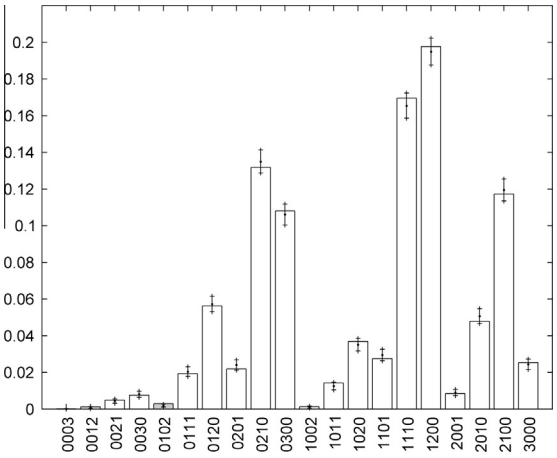


Fig. 6. Analysis of symbolic frequency, projected outcomes $\widehat{\mathbf{Y}}_0$ (dark gray bars are symbols that are observed significantly less frequently than expected under the null; light gray bars are symbols that are observed significantly more frequently than expected under the null).

Table 4
Multinomial logit model: estimation results and diagnostics. Dependent variable is number of cars.

	Model 1			Model 2			Model 3		
	1 Car	2 Car	3+ Car	1 Car	2 Car	3+ Car	1 Car	2 Car	3+ Car
Constant	−2.210	−6.760	−9.950	−1.360	−4.980	−7.800	−4.510	−9.560	−13.000
Socio-economic and demographic									
Household: couple	0.151	1.130	−0.992	0.151	1.150	−0.969	0.094	1.090	−1.030
Household: Couple + Children	1.580	3.230	1.510	1.490	3.080	1.340	1.340	2.900	1.140
Household: single parent	1.340	2.010	2.260	1.220	1.800	2.050	1.100	1.680	1.910
Household: other type	*−0.031*	0.983	*−0.037*	*−0.053*	1.030	*0.039*	−0.133	0.943	*−0.040*
Income 20–0 K	0.698	0.870	1.070	0.714	0.859	1.050	0.717	0.827	1.010
Income 40–60 K	1.180	1.710	2.050	1.200	1.700	2.000	1.240	1.690	1.960
Income 60–80 K	1.360	2.250	2.620	1.370	2.190	2.510	1.430	2.210	2.500
Income 80–100 K	1.510	2.570	3.170	1.520	2.480	3.000	1.680	2.620	3.100
Income >100 K	2.040	3.510	4.240	2.040	3.410	4.080	2.250	3.640	4.300
Income: Refuse/don't know	0.768	1.540	1.890	0.739	1.390	1.680	0.726	1.330	1.590
Part time work	−0.214	−0.161	*−0.133*	−0.190	−0.136	*−0.108*	−0.158	*−0.100*	*−0.073*
Student	−0.684	−0.913	−1.090	−0.646	−0.863	−1.050	−0.608	−0.827	−1.000
Driver license	2.130	3.370	4.450	2.130	3.340	4.410	2.180	3.380	4.430
Urban structure and built environment									
Pop. den. med.				−0.189	−0.690	−1.010	*0.010*	−0.216	−0.419
Pop. den. high				−0.838	−1.870	−2.410	−0.287	−0.801	−1.120
Job den. med.				−0.196	−0.313	−0.325	−0.101	*−0.102*	*−0.057*
Job den. high				−0.386	−0.671	−0.618	−0.168	−0.301	*−0.184*
Land use mix				−0.472	−0.860	−1.010	−0.196	−0.455	−0.551
Locational									
Distance to CBD							1.080	1.600	1.850
XC^2							−0.314	−0.477	−0.494
XC							1.530	2.170	2.280
Null LL	−47170.052	−39393.166		−25959.944	00.450052,514.67		−25513.789	00.459051,716.280.0162	
Constants LL	−26729.871	00.433053,808.00		−0.3126[−0.3558 … −0.2732]			[−0.0324 … 0.0592]		
Final LL	−0.6434[−0.6710 … −0.6175]								
ρ^2									
BIC									
Spatial fit Φ									
Interval of Φ (based on the 95% C.I. of Q)									

Note: All coefficients are significant at $p < 0.05$, with the exception of italicized cells for which $p \geqslant 0.05$.

The approach to assess spatial fit is useful to investigate the elements of the response that may contribute to under- or over-fitting. For instance, assuming this to be the objective, inspection of the symbolic frequency of outcomes projected using Model 3 could potentially lead to some useful insights. As seen in Fig. 7, Model 3 closely resembles the spatial pattern of the original variable. At a frequency of approximately 0.10, it still over-estimates the number of sites where all three

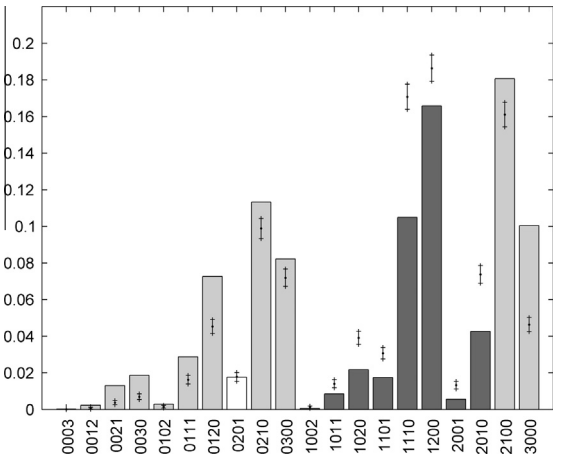


Fig. 7. Analysis of symbolic frequency, projected outcomes $\hat{\mathbf{Y}}_3$ (dark gray bars are symbols that are observed significantly less frequently than expected under the null; light gray bars are symbols that are observed significantly more frequently than expected under the null).

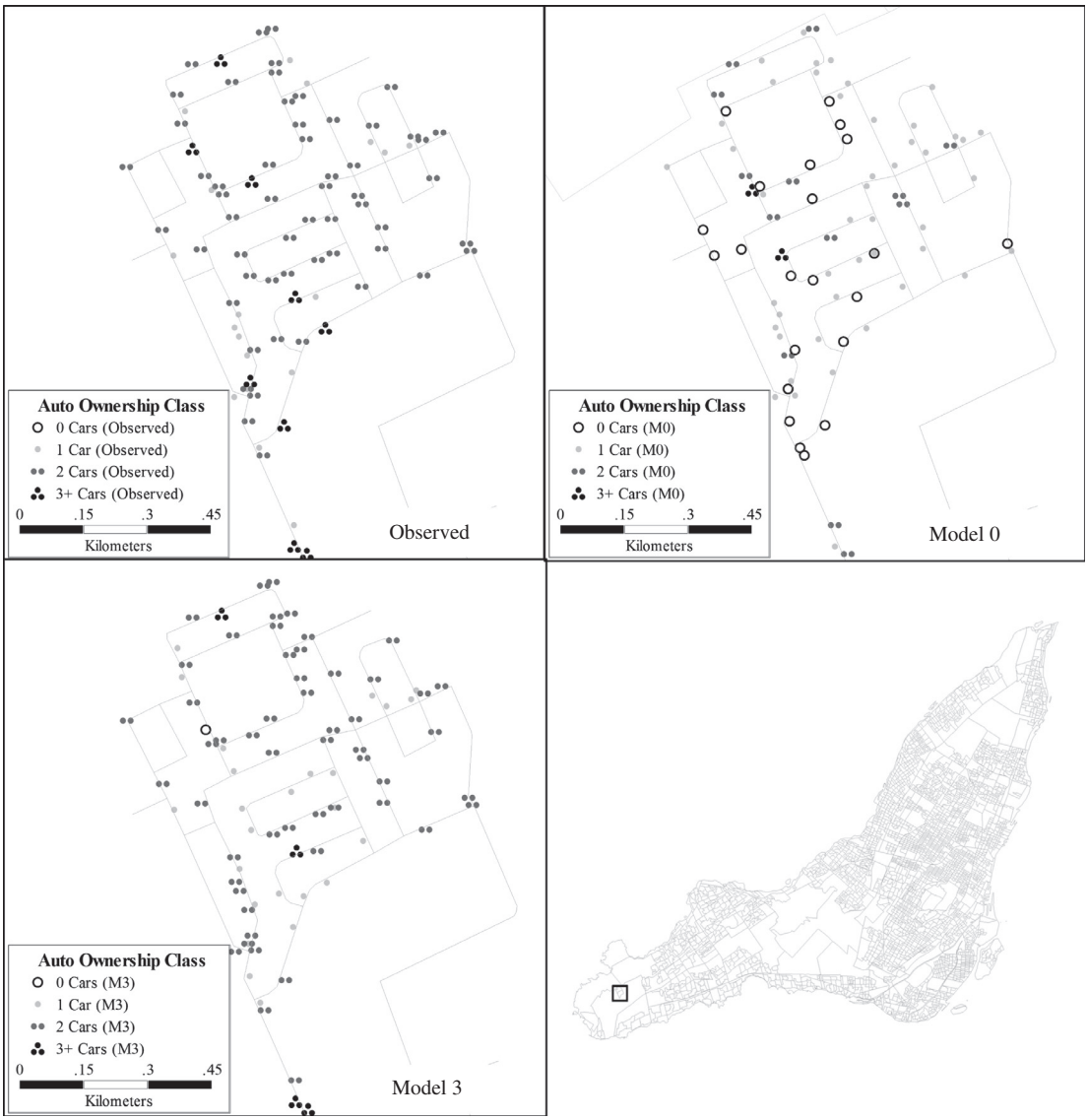


Fig. 8. Micro-geography of observed and projected responses according to Model 0 (M0) and Model 3 (M3). Location of inset shown in bottom-right panel.

neighbors have no cars (3000). The respective frequency for the original variable was approximately 0.06. Similarly, the model over-estimates the case of two neighbors without cars, in the neighborhood of a household with one car (2100), which is estimated at approximately 0.18 compared to 0.15 in the original variable. Contrariwise, the model under-estimates the situation where two neighbors have one car and another does not own a vehicle (1200), at approximately 0.12–0.14. Inspection of the pattern of the original and projected outcomes can generate questions about the specification of the model. For instance, why does the model predict more carless households in close proximity of each other than actually observed? This, in turn, could lead to ideas useful to refine the model as desired.

Finally, Fig. 8 provides an example of the geography of observed and projected responses, according to Model 0 (naïve model, constants only), and Model 3. It can be seen from the figure that the two models differ in their degree of similitude to the observed spatial pattern. According to Model 0, of course, the distribution of outcomes across space is random. This is evident from the figure, and although this model closely yields the observed shares of choices, these provide a poor spatial fit. Model 3, in contrast, provides the best spatial fit.

5. Conclusions

In this paper we proposed a simple and intuitive indicator of spatial fit to assess the performance of discrete choice models estimated using geo-referenced data. The indicator builds on the $Q(m)$ statistic, whereby the degree of spatial association

between the predicted variable obtained from a discrete choice model of interest is contrasted against the original variable used for estimation of the model. The indicator is bounded and provides a valuable guide regarding the degree to which the projected outcomes $\hat{\mathbf{Y}}$ approximate the level of spatial association of the observed outcomes $\boldsymbol{Y}$. In addition, we derived the results needed to conduct hypothesis testing, whereby the null hypothesis is that $Q(m)$ calculated for $\hat{\mathbf{Y}}$ and for $\boldsymbol{Y}$ are identical, to a predetermined level of confidence. Failure to reject the null hypothesis indicates that the degree of spatial association of these two variables is not comparable.

A set of numerical experiments was conducted that provide information useful to suggest guidelines regarding the application of the spatial fit indicator and testing. Furthermore, the approach was demonstrated by means of an empirical example of auto ownership in Montreal. Several models were estimated to illustrate the effect of adding variables (as in a forward search using blocks of variables as in the present paper). The approach could also be used in a similar way if the search strategy involved removing variables (as in a general-to-specific backward search). The various models indicate that, as expected, the goodness of fit measured by the adjusted ρ^2 indicator and the Bayesian Information Criterion, improves with the addition of significant variables. The spatial fit also tends to improve. The final model including a relatively simple trend surface achieves the objective of replicating the level of spatial association of the original discrete variable at the 0.05 level of confidence.

Together, the numerical experiments and case study suggest the utility of employing the proposed indicator of spatial fit as a diagnostic and exploratory tool in the estimation of discrete choice models when using geo-referenced data.

Acknowledgments

The authors received funding from several organizations. Antonio Páez is supported by a grant from Canada's Natural Sciences and Engineering Research Council. Fernando López was supported by projects 11897/PHCS/09 (Fundación Seneca, Comunidad Autónoma de Murcia) and ECEO2009-10534 (Ministerio de Ciencia e Innovación del Reino de España). Manuel Ruiz Marín was supported by projects ECO2012-36032-C03-03 and MTM2012-35240 (Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional) and COST Action IS1104. Catherine Morency is supported by Chaire Mobilité (co-sponsored by La Ville de Montréal, l'Agence métropolitaine de transport, le Ministère des transports du Québec, and la Société de transport de Montréal). This paper also benefitted from the expert opinion of eight anonymous reviewers. The views expressed here are those of the authors alone, and do not represent the official position of any of these organizations.

References

Adjemian, M.K., Lin, C.Y.C., Williams, J., 2010. Estimating spatial interdependence in automobile type choice with survey data. Transportation Research Part A 44 (9), 661–675.
Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.
Alemu, D.D., Tsutsumi, J.G., 2011. Determinants and spatial variability of after-school travel by teenagers. Journal of Transport Geography 19 (4), 876–881.
Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.
Axsen, J., Kurani, K., 2012. Interpersonal influence within car buyers' social networks: applying five perspectives to plug-in hybrid vehicle drivers. Environment and Planning A 44 (5), 1047–1065.
Bailey, T.C., Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Addison Wesley Longman, Essex.
Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (Macml) estimation of multinomial probit-based unordered response choice models. Transportation Research Part B 45 (7), 923–939.
Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transportation Research Part B 38 (2), 147–168.
Bhat, C.R., Pulugurta, V., 1998. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. Transportation Research Part B 32 (1), 61–75.
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., Sener, I.N., Eluru, N., 2010. A flexible spatially dependent discrete choice model: formulation and application to teenagers' weekday recreational activity participation. Transportation Research Part B 44 (8–9), 903–921.
Boots, B., 2006. Local configuration measures for categorical spatial data: binary regular lattices. Journal of Geographical Systems 8 (1), 1–24.
Chakir, R., Parent, O., 2009. Determinants of land use changes: a spatial multinomial probit approach. Papers in Regional Science 88 (2), 327–344.
Dacey, M.F., 1968. A Review on measures of contiguity for two and *k*-color maps. In: Berry, B.J.L., Marble, D.F. (Eds.), Spatial Analysis: A Reader in Statistical Geography. Prentice Hall, Englewood Cliffs, NJ, pp. 479–495.
Dugundji, E.R., Walker, J.L., 2005. Discrete choice with social and spatial network interdependencies – an empirical example using mixed generalized extreme value models with field and panel effects. Transportation Research Record 1921, 70–78.
Farber, S., Páez, A., Volz, E., 2009. Topology and dependency tests in spatial and network autoregressive models. Geographical Analysis 41 (2), 158–180.
Frazier, C., Kockelman, K.M., 2005. Spatial econometric models for panel data – incorporating spatial and temporal data. Transportation Research Record 1902, 80–90.
Giuliano, G., Dargay, J., 2006. Car ownership, travel and land use: a comparison of the us and Great Britain. Transportation Research Part A 40 (2), 106–124.
Goetzke, F., 2008. Network effects in public transit use: evidence from a spatially autoregressive mode choice model. Urban Studies 45 (2), 407–417.
Goetzke, F., Rave, T., 2011. Bicycle use in Germany: explaining differences between municipalities with social network effects. Urban Studies 48 (2), 427–437.
Goetzke, F., Weinberger, R., 2012. Separating contextual from endogenous effects in automobile ownership models. Environment and Planning A 44 (5), 1032–1046.
Griffith, D.A., 1988. Advanced Spatial Statistics: Special Topics in the Exploration of Quantitative Spatial Data Series. Kluwer, Dordrecht.
Haining, R., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University Press, Cambridge.
Hammadou, H., Thomas, I., Verhetsel, A., Witlox, F., 2008. How to incorporate the spatial dimension in destination choice models: the case of Antwerp. Transportation Planning and Technology 31 (2), 153–181.
Hoeffding, W., Robbins, H., 1948. The central limit theorem for dependent random variables. Duke Mathematical Journal 15 (3), 773–780.
LeSage, J., Pace, R.K., 2009. Introduction to Spatial Econometrics. CRC Press, Boca Raton.

McFadden, D., 1974. The measurement of urban travel demand. Journal of Public Economics 3 (4), 303–328.
McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annual Review of Sociology 27, 415–444.
Miyamoto, K., Vichiensan, V., Shimomura, N., Páez, A., 2004. Discrete choice model with structuralized spatial effects for location analysis. Transportation Research Record 1898, 183–190.
Morency, C., Páez, A., Roorda, M.J., Mercado, R.G., Farber, S., 2011. Distance traveled in three Canadian cities: spatial analysis from the perspective of vulnerable population segments. Journal of Transport Geography 19 (1), 39–50.
Páez, A., 2006. Exploring contextual variations in land use and transport analysis using a probit model with geographical weights. Journal of Transport Geography 14 (3), 167–176.
Páez, A., 2007. Spatial perspectives in urban systems: developments and directions. Journal of Geographical Systems 9 (1), 1–6.
Páez, A., Ruiz, M., López, F., Logan, J., 2012. Measuring ethnic clustering and exposure with the *q* statistic: an exploratory analysis of Irish, Germans, and Yankees in 1880 Newark. Annals of the Association of American Geographers 102 (1), 84–102.
Páez, A., Scott, D.M., 2004. Spatial statistics for urban analysis: a review of techniques with examples. GeoJournal 61 (1), 53–67.
Páez, A., Scott, D.M., 2007. Social influence on travel behavior: a simulation example of the decision to telecommute. Environment and Planning A 39 (3), 647–665.
Páez, A., Suzuki, J., 2001. Transportation impacts on land use change: an assessment considering neighborhood effects. Journal of the Eastern Asia Society for Transportation Studies 4 (6), 47–59.
Paleti, R., Bhat, C., Pendyala, R.M., Goulias, K., 2013. The modeling of household vehicle type choice accommodating spatial dependence effects. In: 92nd Annual Meeting of the Transportation Research Board, Washington, DC.
Roorda, M.J., Páez, A., Morency, C., Mercado, R.G., Farber, S., 2010. Trip generation of vulnerable populations in three Canadian cities: a spatial ordered probit approach. Transportation 37 (3), 525–548.
Ruiz, M., López, F., Páez, A., 2010. Testing for spatial association of qualitative data using symbolic dynamics. Journal of Geographical Systems 12 (3), 281–309.
Sener, I.N., Bhat, C.R., 2012. Flexible spatial dependence structures for unordered multinomial choice models: formulation and application to teenagers' activity participation. Transportation 39 (3), 657–683.
Sidharthan, R., Bhat, C.R., Pendyala, R.M., Goulias, K.G., 2011. Model for children's school travel mode choice accounting for effects of spatial and social interaction. Transportation Research Record 2213, 78–86.
Smirnov, O.A., 2010. Spatial econometrics approach to integration of behavioral biases in travel demand analysis. Transportation Research Record 2157, 1–10.
Wang, X., Kockelman, K.M., Lemp, J.D., 2012. The dynamic spatial multinomial probit model: analysis of land use change using parcel-level data. Journal of Transport Geography 24, 77–88.
Wang, X.K., Kockelman, K.M., 2009. Application of the dynamic spatial ordered probit model: patterns of land development change in Austin. Texas. Papers in Regional Science 88 (2), 345–365.
Whalen, K., Páez, A., Bhat, C.R., Moniruzzaman, M., Paleti, R., 2012. T-communities and sense of community in a university town: evidence from a student sample using a spatial ordered response model. Urban Studies 49 (6), 1357–1376.
Zhou, B., Kockelman, K.M., 2008. Neighborhood impacts on land use change: a multinomial logit model of spatial relationships. Annals of Regional Science 42 (2), 321–340.