Statistical analysis of network data—an application to diffusion of innovation[☆]

Jukka Nyblom [a,*], Steve Borgatti [b], Juha Roslakka [c], Mikko A. Salo [c]

[a] *Department of Statistics, University of Joensuu, P.O. Box 111, 80101 Joensuu, Finland*
[b] *Boston College, Boston, MA, USA*
[c] *Department of Social Policy, University of Joensuu, P.O. Box 111, 80101 Joensuu, Finland*

Abstract

General methodology is developed here to deal with the association between a a binary variable and network connections with or without confounding covariates. Also the case when the network is observed at several time periods is treated. As an application we consider the diffusion of organic farming in the province of North Karelia in Finland. It turns out that organic farms are more clustered than would be expected under pure random allocation. The neighborhood effect remains when adjusting for the production lines of the farms. The spatio-temporal analysis shows that new adopters are more often found within the neighborhoods of each others and of earlier adopters. © 2003 Elsevier Science B.V. All rights reserved.

JEL classification: C15; C52; O31; R11

Keywords: Logistic regression; Neighborhood effect; Organic farming; Permutation tests; Spatio-temporal association

1. Introduction

A common approach to the study of adoption of innovations within a social system is to regard adoption as the outcome of an information-seeking process that seeks to decrease the uncertainty relating to the adoption decision (Burt, 1987, p. 1288; Valente, 1995, p. 5). As Rogers (1995, p. 304) argues, a key method for individuals of obtaining information is communication with those around them, particularly those who have already faced the same decision. Hence, the adoption of innovations has come to be seen as the outcome of

a spatial and social diffusion process, what Hägestrand (1953), in a seminal paper, called the 'neighborhood effect'.

Starting in the 1940s, geographers have accumulated a set of statistical methods for examining diffusion. The general approach is as follows. If adoption is a function of diffusion among neighbors (whether defined socially, spatially or some combination), then pairs of adopters should be found more frequently among neighbors than we would expect by chance. In contrast, if adoption is principally a function of independent decisions based on, say, economic factors (e.g. government subsidies) or individual dispositions, then we would expect adopters to be uniformly randomly distributed with respect to being neighbors. To test for diffusion, then, we can distribute the 'is an adopter' property randomly among individuals and then count the number of adopter pairs that are neighbors. Repeating this many times, we can calculate how rare an event the observed number of connected adopter pairs is under the uniform distribution of adopters. This basic approach underlies the join-count statistic of Moran (1948), as well as the autocorrelation statistics developed by Moran (1950) and Geary (1954). In an epidemiological context, Mantel (1967) proposed a general quadratic statistic to measure the relation between two distance matrices and encouraged the use of simulation methods to assess its significance. Subsequently, the Mantel statistic was independently re-invented several times (see Whaley, 1983; Good, 1994) and used in a diversity of fields from epidemiology to archaeology and sociology (where it is known as QAP). An impressive set of references is found in Good (1994).

The objective of the present paper is to extend this basic body of work in two key directions. The first direction deals with the issue of confounding covariates. Potential adopters have varying attributes (e.g. level of education) which may affect their propensity to adopt an innovation. This becomes problematic when these same features affect who interacts with whom. Then, the observed association between being 'neighbors' and willingness to adopt innovations can be accounted for by these confounding factors. We examine two cases: one in which the covariates are entirely categorical, and the other in which continuous covariates are also present. We show that in the purely categorical case, we can eliminate the confounding effects by modifying the simulations so that the random allocations of the 'is an adopter' property are performed separately within categories. For the continuous covariate case, we use logistic regression to make the covariate adjustments.

The second direction deals with modeling the adoption process over time. Cross-sectional tests, even with the ability to control for covariates, provide only weak evidence for diffusion. While diffusion implies clustering, clustering in itself does not prove diffusion. Stronger evidence is provided if new adopters at any given time are neighbors of older adopters. We provide a relatively simple statistic for assessing this combined spatio-temporal effect.

We illustrate our methods by examining the adoption of organic farming from 1990 to 1999 among farmers in the province of North Karelia in Finland. The data were obtained from records kept by the Finnish Ministry of Agriculture and Forestry covering agricultural land use, livestock information, and most importantly, geographical location. Thus, we use physical distance among farm houses to define farmer's neighborhoods. Adoption of organic farming in Finland is occurring in a context in which clear economic incentives lead farmers to consider the adoption, creating a plausible case for the null model of adoption not based on social diffusion processes.

The article is organized as follows. First, in Section 2, we define a basic association measure. Its permutation distribution is derived in Section 3. Sections 4 and 5 discuss how to adjust the permutation distribution for categorical and continuous covariates. In Section 6, we extend the approach to longitudinal data. The article concludes with an empirical application of the methods to the diffusion of organic farming in Northern Karelia, Finland, in 1990–1999. The results have implications for institutional organizational theory (Meyer and Rowan, 1977; Scott, 1991).

2. Measuring network association

Let us assume that we have a population of N individuals labeled as $1, 2, \ldots, N$. Each individual is classified into one of two classes, the adopters which we are interested in and the non-adopters. This is denoted by the variables x_i taking the value 1 for an adopter and 0 for a non-adopter. Assume further that each pair (i, j) is either connected or disconnected. The connection is directional, i.e. i may be connected to j, $i \rightarrow j$, but j is not necessarily connected to i, $j \nrightarrow i$.

The connections are defined through the variables

$$\delta_{ij} = \begin{cases} 0, & \text{if } i \nrightarrow j, \\ 1, & \text{if } i \rightarrow j. \end{cases}$$

When put into a matrix $\mathbf{\Delta} = (\delta_{ij})$, they define a directed graph.

In many applications, as in ours, the connection means geographical nearness, and $\mathbf{\Delta}$ determines the neighborhoods for all individuals. The neighborhood of i consists of all individuals j with $\delta_{ij} = 1$. In other contexts the connection may be social rather than geographical, e.g. δ_{ij} can define whether i counts j as his/her friend. Also the geographical neighborhood is definable in various ways. If we have subregions instead of individuals, the neighbors of i are naturally defined to be those having linked by a join, i.e. having a boundary of positive length in common. If the individuals are points on a plane the natural definition employs circular neighborhoods consisting of all those living within a certain distance. The drawback here is that if the density of points varies greatly, the radius appropriate in the dense part may lead to a large number of individuals in having no neighbors in a sparse part. In order to avoid this, one could define the neighborhood as consisting of k nearest neighbors. A compromise between these two is to take the individuals that are within the radius which is the distance to the nearest neighbor plus some fixed constant. This is the definition that we have employed in our application. Nevertheless, we stress that the following discussion applies to whatever definition for the neighborhood is adopted.

Next we introduce the in- and out-degrees of the graph. The in-degree of j is $q_j = \sum_i \delta_{ij}$ (the number of edges directed to j) and the out-degree of i is $k_i = \sum_j \delta_{ij}$ (the number of edges directed from i). In our application k_i gives the number of the neighbors the farm i possess, while q_j is the number of the neighborhoods the farm j belongs to. If we assume that the individual i is influenced by its neighbors, then k_i is the measure how much i is influenced, while q_j measures how influential j is. The interpretation may also be vice versa depending on the application.

The column vector $\boldsymbol{k}$ with components k_i is obtained by the matrix multiplication $\boldsymbol{k} = \boldsymbol{\Delta 1}$, where $\boldsymbol{1}$ is the column vector of ones. Accordingly the vector of in-degrees is $\boldsymbol{q} = \boldsymbol{\Delta' 1}$, where the prime stands for transposing a matrix. We have, of course, the identity $\sum k_i = \sum q_j = \sum \sum \delta_{ij}$. Denote this sum, the total number of connections, by S. In matrix terms $S = \boldsymbol{k'1} = \boldsymbol{q'1} = \boldsymbol{1'\Delta 1}$. In the special case of an undirected connection, i.e. $i \rightarrow j$ implies $j \rightarrow i$, $\boldsymbol{\Delta}$ is symmetric and $\boldsymbol{k} = \boldsymbol{q}$. This applies to circular neighborhoods mentioned earlier. When the neighborhoods are defined as the k nearest neighbors we have $k_i = k$ for each i.

The interesting question we now wish to investigate is whether the adopters are clustered in a sense that they are more likely to be connected with one another than what might happen under purely random distribution. A simple measure for network association is

$$Q_{11} = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} x_i x_j = \boldsymbol{x'\Delta x}. \tag{1}$$

Clearly Q_{11} gives the number of pairs $i \rightarrow j$ in which both are adopters. Note that if $i \rightarrow j$ and $j \rightarrow i$, then this pair is counted twice.

The association between the non-adopter pairs is analogously

$$Q_{00} = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} (1 - x_i)(1 - x_j) = (\boldsymbol{1} - \boldsymbol{x})' \boldsymbol{\Delta} (\boldsymbol{1} - \boldsymbol{x}).$$

The remaining two associations are $Q_{01} = (\boldsymbol{1} - \boldsymbol{x})' \boldsymbol{\Delta x}$ and $Q_{10} = \boldsymbol{x'\Delta} (\boldsymbol{1} - \boldsymbol{x})$. A direct calculation shows that

$$\begin{aligned} Q_{10} &= \boldsymbol{k'x} - Q_{11}, \\ Q_{01} &= \boldsymbol{q'x} - Q_{11}, \\ Q_{00} &= \boldsymbol{1'\Delta 1} - \boldsymbol{k'x} - \boldsymbol{q'x} + Q_{11}. \end{aligned} \tag{2}$$

We can collect these into a 2×2 table as follows.

Association table for two types.

		total
Q_{00}	Q_{01}	$\boldsymbol{k'}(1-\boldsymbol{x})$
Q_{10}	Q_{11}	$\boldsymbol{k'x}$
total $\boldsymbol{q'}(1-\boldsymbol{x})$	$\boldsymbol{q'x}$	S

$$\tag{3}$$

If Q_{11} is high indicating association between adopters, it does not imply that there is also association between the non-adopters in the sense of Q_{00}. Eq. (2) shows that the increase in Q_{11} may be compensated by the increase in $\boldsymbol{k'x}$ and $\boldsymbol{q'x}$. These latter statistics are large when the adopters are more likely to be found among those having large number of neighbors (large $\boldsymbol{k'x}$) or belonging to many neighborhoods (large $\boldsymbol{q'x}$).

We conclude this section by relating our association measure Q_{11} in (1) to the previous literature. Firstly, it is a special case of Mantel's general quadratic statistic (Mantel, 1967)

$$\sum \sum \delta_{ij} y_{ij}. \tag{4}$$

The interpretation is that we measure closeness between two individuals in two different manner. In Mantel's application δ_{ij} measures spatial closeness and y_{ij} temporal closeness of

two individuals having the same disease. In our application δ_{ij} continues to measure spatial closeness but being zero-one valued. The temporal closeness y_{ij} is replaced by $x_i x_j$, i.e. with a binary variable expressing whether the individuals i and j are both adopters or not, a kind of closeness too.

Again, if we are dealing with subregions instead of individuals the statistic Q_{11} is seen to be a join count statistic proposed by Moran (1948) (see also Cliff and Ord, 1981, Chapter 1) with the exception that the connection matrix is not necessarily symmetric. Later, Moran (1950) proposed a closely related spatial autocorrelation

$$I = \frac{S^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij}(x_i - \bar{x})(x_j - \bar{x})}{N^{-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}.$$

After some algebra we find that

$$I = \frac{(M/N)Q_{00} + [(N-M)/N]Q_{11}}{S(M/N)(1 - M/N)} - \frac{1}{N},$$

which shows that Moran's I is a kind of weighted sum of Q_{00} and Q_{11}. It receives a high value also when non-adopters are clustered but adopters are not. Note also that $I_{11} = I_{00} = -I_{10} = -I_{01}$ with an obvious subscript notation. Therefore, we find some advantage to use the measure Q_{11} which is directly related to adopters. Another closely related statistic to I is proposed by Geary (1954).

3. Permutation test

From now on we concentrate on Q_{11} defined in (1). For simplicity let us denote $Q_{11} = Q$. It is plain that under clustering Q should take larger values than under no clustering. Assume now that the observations $x_1, \ldots, x_N$ are realizations of the Bernoulli random variables $X_1, \ldots, X_N$. Then the null hypothesis of no clustering can be stated formally as

$$H_0 : X_i \sim_{\text{ind}} \text{Ber}(\theta), \quad i = 1, \ldots, N,$$

i.e. the X_i are independent with $P(X_i = 1) = \theta$ (unknown). Under the null hypothesis $M = \sum X_i$ is a sufficient statistic. The conditional distribution of Q given $M = \sum X_i$ is free from θ making it possible to compute the exact significance probability (p-value). Start with any sequence $a_1, \ldots, a_N$ with $N - M$ zeros and M ones. Then

$$P\left(X_1 = a_1, \ldots, X_N = a_N \,\bigg|\, \sum X_i = M \right) = \frac{1}{\binom{N}{M}}.$$

Thus, the conditional distribution of Q given $M = \sum X_i$ is

$$P\left(Q \le c \,\bigg|\, \sum X_i = M \right) = \frac{\#\{\sum \sum \delta_{ij} a_i a_j \le c\}}{\binom{N}{M}},$$

where the symbol '#' stands for "the number of". Thus, the probability is the proportion of those sequences $a_1, \ldots, a_N$ with $N - M$ zeros and M ones yielding a value at most c for Q. Because a complete enumeration is impossible even for moderate M and N, a practical solution is found by taking a random sample from the sequences $a_1, \ldots, a_N$. This can be simply accomplished by a random permutation of the observed values $x_1, \ldots, x_N$. The (random) significance probability or p-value based on B random permutations is then

$$p = \frac{\#\{Q^\pi \geq Q\}}{B},$$

where $Q^\pi = \sum \sum \delta_{ij} x_i^\pi x_j^\pi$ with $(x_1^\pi, \ldots, x_N^\pi)$ being a random permutation of $(x_1, \ldots, x_N)$.

The first two moments are obtained by a direct calculation. Plainly

$$E(x_i^\pi) = P(x_i^\pi = 1) = \frac{M}{N}.$$

Because

$$E(x_i^\pi x_j^\pi) = P(x_i^\pi = x_j^\pi = 1) = \frac{M(M - 1)}{N(N - 1)}, \quad i \neq j,$$

we obtain, recalling $\delta_{ii} = 0$, that

$$E(Q^\pi) = \sum \sum \delta_{ij} E(x_i^\pi x_j^\pi) = S \frac{M(M - 1)}{N(N - 1)}.$$

Note that the expected frequency is the product of the numbers of all directed connections S and of all ordered pairs $M(M - 1)$ of adopters divided by the number of all ordered pairs $N(N - 1)$.

A direct calculation yields

$$\begin{aligned} \mathrm{var}(Q^\pi) = {} & p_1(S + R) + p_2[(\boldsymbol{k} + \boldsymbol{q})'(\boldsymbol{k} + \boldsymbol{q}) - 2(S + R)] \\ & + p_3[S^2 + S + R - (\boldsymbol{k} + \boldsymbol{q})'(\boldsymbol{k} + \boldsymbol{q})] - p_1^2 S^2, \end{aligned}$$

where

$$p_k = \frac{M(M - 1) \cdots (M - k)}{N(N - 1) \cdots (N - k)}, \quad k = 1, 2, 3,$$
$$R = \sum \sum \delta_{ij} \delta_{ji}.$$

Note that R equals the number of pairs for which the relation is symmetric (each pair is counted twice!). The formula could befound also in Cliff and Ord (1981, p. 20), if we first symmetrize the δ_{ij}. Cuzick and Edwards (1990) consider the asymmetric case as we do but their neighborhoods are defined as the k nearest neighborhoods. Also the general formulas of Mantel (1967) are applicable.

We may expect that under some conditions on the connection matrix $\boldsymbol{\Delta}$ the permutation of Q^π tends to a normal distribution (for a discussion of this matter, see Cuzick and Edwards, 1990). Then we can compute $Z = [Q - E(Q^\pi)]/\sqrt{\mathrm{var}(Q^\pi)}$ and compare it to the standard normal distribution. However, this asymptotic result is doubtful, and our simulations confirm it, when the adopters are rare, i.e. for small M/N. A common statistical model for rare

events is the Poisson distribution. Therefore, we might proceed under the assumption that Q^π follows a Poisson distribution with mean $E(Q^\pi)$. But in Poisson model variance equals to mean, and in practice we have found that $\text{var}(Q^\pi)$ is often considerably larger than $E(Q^\pi)$. The negative binomial distribution is a simple alternative to Poisson model when variance exceeds mean.

The negative binomial probabilities are given by the formula

$$f(k; r, q) = \binom{r + k - 1}{k} q^k (1 - q)^r, \quad k = 0, 1, \dots; \ \ r > 0, \ 0 < q < 1.$$

The distribution has a natural interpretation for integer r, but any positive r defines a proper distribution (e.g. Feller, 1957, pp. 155–156). The computation of negative binomial probabilities is included in many statistical packages. The expectation and variance of a negative binomial variable are $rq/(1-q)$ and $rq/(1-q)^2$, respectively. Hence, if $\text{var}(Q^\pi) > E(Q^\pi)$, we equate the mean and variance of the negative binomial distribution and the permutation distribution as

$$E(Q^\pi) = \frac{rq}{1 - q}, \qquad \text{var}(Q^\pi) = \frac{rq}{(1 - q)^2}.$$

These yield

$$q = \frac{\text{var}(Q^\pi) - E(Q^\pi)}{\text{var}(Q^\pi)}, \qquad r = \frac{[E(Q^\pi)]^2}{\text{var}(Q^\pi) - E(Q^\pi)}.$$

Then the p-value is obtained with these r and q as the tail probability

$$\sum_{k=Q}^{\infty} f(k; r, q) = 1 - \sum_{k=0}^{Q-1} f(k; r, q).$$

We want to emphasize that the use of the negative binomial distribution is based on empirical findings rather than on theoretical arguments.

In some applications, those individuals that have many neighbors or belong to many neighborhoods are more exposed to innovations. A test for this can be based on the marginal counts $k'x$ and $q'x$ of the Table 3). Their first two moments are needed for normal approximation and given below. The expectations are equal and satisfy

$$E\left(\sum k_i x_i^\pi\right) = E\left(\sum q_j x_j^\pi\right) = S\frac{M}{N}.$$

The variances and the covariance are

$$\text{var}\left(\sum k_i x_i^\pi\right) = \frac{M(N - M)}{N(N - 1)} \sum (k_i - \bar{k})^2,$$

$$\text{var}\left(\sum q_j x_j^\pi\right) = \frac{M(N - M)}{N(N - 1)} \sum (q_i - \bar{q})^2,$$

$$\text{cov}\left(\sum k_i x_i^\pi, \sum q_j x_j^\pi\right) = \frac{M(N - M)}{N(N - 1)} \sum (k_i - \bar{k})(q_i - \bar{q}),$$

where $\bar{k} = \bar{q} = S/N$.

4. Adjusting for categorical covariates

It may happen that clustering occurs because of some confounding variable. In our application the adopters are organic farmers. Suppose that organic farming is more popular among crop producers, and that the crop producers are spatially clustered. Then the apparent clusters of organic farms may be explained by this association.

In general terms, assume now that the individuals divide into G disjoint groups or strata. Let the gth group have N_g individuals, $N_1 + \cdots + N_G = N$. Also, let us modify our notation by introducing double subscripts with x_{gi} referring to the ith individual in the gth group and further, write

$$Q = \sum_{g=1}^{G} \sum_{h=1}^{G} \sum_{i=1}^{N_g} \sum_{j=1}^{H_h} \delta_{ij}^{gh} x_{gi} x_{hj},$$

where δ_{ij}^{gh} defines the connection between the ith individual from the gth group to the jth individual from the hth group. Despite the complicated notation the Q-statistic is computed in exactly the same manner as in the case of no covariates. Adjustment for the covariates is made via a change of the distribution under the null hypothesis.

The null hypothesis of no clustering but allowing for varying levels over the groups formalizes to

$$H_0 : X_{gi} \sim_{\text{ind}} \text{Ber}(\theta_g), \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G. \tag{5}$$

The variables x_{gi} are the observed values of X_{gi}. The clustering of the adopters is continued to measure with Q, but when calculating its significance we wish to eliminate the confounding grouping effect by conditioning on the group totals

$$M_g = \sum_{i=1}^{N_g} X_{gi}, \quad g = 1, \ldots, G.$$

They are sufficient for the model (5). This leads, by the similar reasoning as in the previous section, to the conclusion that all the permutations within the groups are equally likely. Hence, the p-value based on B random permutations is

$$p = \frac{\#\{Q^\pi \geq Q\}}{B},$$

where

$$Q^\pi = \sum_{g=1}^{G} \sum_{h=1}^{G} \sum_{i=1}^{N_g} \sum_{j=1}^{H_h} \delta_{ij}^{gh} x_{gi}^\pi x_{hj}^\pi$$

with $x_{g1}^\pi, \ldots, x_{g,N_g}^\pi$ being a permutation of $x_{g1}, \ldots, x_{g,N_g}$, $g = 1, \ldots, G$. Furthermore, the permutations are independent in different groups. Previously Pike and Smith (1974) have suggested similar procedure in case–control studies in epidemiology, where a number of cases and controls matched by adjusting for age, sex, etc. The method of permuting within

categories is well known in the linear model context (e.g. Noreen, 1989, Chapter 2.7) and Good (1994, Chapter 4.3).

The expected value of Q^π is

$$E(Q^\pi) = \sum_{g \neq h} \sum S_{gh} \frac{M_g M_h}{N_g N_h} + \sum S_{gg} \frac{M_g(M_g - 1)}{N_g(N_g - 1)}, \tag{6}$$

because

$$E(x_{gi}^\pi x_{hj}^\pi) = \begin{cases} \dfrac{M_g(M_g - 1)}{N_g(N_g - 1)}, & \text{for } g = h, \\[3mm] \dfrac{M_g M_h}{N_g N_h}, & \text{for } g \neq h. \end{cases}$$

The exact formula for the variance is complicated and thus omitted (Pike and Smith, 1974).

5. Logistic regression in adjusting for covariates

The case with two categorical covariates with groups $C_1, \ldots, C_{G_1}$ and $D_1, \ldots, D_{G_2}$ can be treated as in the previous section by introducing the intersections $C_i \cap D_j$ as a new grouping scheme. Clearly, this can be generalized to arbitrary number of categorical covariates. The case of continuous covariates may sometimes be handled by discretizing it into disjoint groups. But these devices are not always practical. With several categorical covariates the number of intersections increase rapidly while the number of individuals in them become small. The discretizing device is, in any case, arbitrary.

In order to avoid the above drawbacks let us assume that under the null hypothesis

$$X_i \sim_{\text{ind}} \text{Ber}(\theta_i)$$

with

$$\log\left(\frac{\theta_i}{1 - \theta_i}\right) = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip}, \quad i = 1, \ldots, N,$$

i.e. we assume the logistic regression. A good introduction to these models is Hosmer and Lemeshow (1989). The covariates $z_{i1}, \ldots, z_{ip}, i = 1, \ldots, N$, can be continuous or categorical dummies. In case of categorical dummies (including all interactions if necessary), we are back in the situation of the previous section. Denote $z_{10} = \cdots = z_{N0} = 1$. Then the likelihood function at the observed data $X_i = x_i, i = 1, \ldots, N$, is

$$\prod_{i=1}^N \theta_i^{x_i}(1 - \theta_i)^{1-x_i} = \exp\left[\sum_{j=0}^p \beta_j \sum_{i=1}^N x_i z_{ij} - \sum \log(1 + e^{\beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip}})\right].$$

The factorization criterion (Lehmann 1994, p. 54) yields sufficient statistics

$$M_j = \sum_{i=1}^N X_i z_{ij}, \quad j = 0, \ldots, p.$$

Again, the conditional distribution of $X_1 \ldots, X_N$ given $M_0, \ldots, M_p$, is free from the nuisance parameters $\beta_0, \ldots, \beta_p$. The test statistic Q remains as such, but the significance is obtained by enumerating all permutations $x_1^\pi, \ldots, x_N^\pi$ with the restrictions $M_j = \sum_{i=1}^N x_i^\pi z_{ij}$. Then we count the relative number of those values not below the observed one. Mehta et al. (2000) give references to fast numerical algorithms. Needless to say that only small data sets can be analyzed by exact methods. Not only the exact solution but also simulation leads to difficulties. The straightforward rejection algorithm is highly inefficient. A practical solution was recently given by Mehta et al. (2000). Yet, their method seems to be unsuitable for large data sets as ours.

For large data sets we propose the following approximate, unconditional method. If we knew the parameters $\beta_0, \ldots, \beta_p$ we could simulate the unconditional distribution of $\sum \sum \delta_{ij} X_i X_j$. But because $\beta_0, \ldots, \beta_p$ are not known we generate random variables from the estimated model, i.e. take $X_i^* \sim_{\text{ind}} \text{Ber}(\hat{\theta}_i)$ with

$$\log \left(\frac{\hat{\theta}_i}{1 - \hat{\theta}_i} \right) = \sum_j \hat{\beta}_j z_{ij}, \quad i = 1, \ldots, N,$$

where the $\hat{\beta}_j$ are the maximum likelihood estimates of the β_j under the null hypothesis. Then compute $Q^* = \sum \sum \delta_{ij} X_i^* X_j^*$. This method can be termed the parametric bootstrap (Efron and Tibshirani, 1993). With B bootstrap samples the estimated p-value is

$$p = \frac{\#\{Q^* \geq Q\}}{B},$$

i.e. the number of bootstrap samples leading to the excess of the observed value Q.

The expected value of Q^* (over bootstrap distribution) is

$$E(Q^*) = \sum \sum \delta_{ij} E(X_i^*) E(X_j^*) = \sum \sum \delta_{ij} \hat{\theta}_i \hat{\theta}_j. \tag{7}$$

By a direct calculation the variance is seen to be

$$\text{var}(Q^*) = \sum (\delta_{ij} + \delta_{ij}\delta_{ji}) \hat{\theta}_i (1 - \hat{\theta}_i) \hat{\theta}_j (1 - \hat{\theta}_j) + \sum_i \left[\sum_j (\delta_{ij} \hat{\theta}_j + \delta_{ji} \hat{\theta}_j) \right]^2 \hat{\theta}_i (1 - \hat{\theta}_i).$$

Because the moments depend basically on the estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)'$, we denote shortly $E(Q^*) = \mu(\hat{\boldsymbol{\beta}})$ and $\text{var}(Q^*) = \sigma^2(\hat{\boldsymbol{\beta}})$. Under mild restrictions on the values δ_{ij} the bootstrap distribution of Q^* tends to a normal distribution, and thus the standardized variable $(Q^* - \mu(\hat{\boldsymbol{\beta}}))/\sigma(\hat{\boldsymbol{\beta}})$ is approximately normal with zero expectation and unit variance. Again, with $\mu(\hat{\boldsymbol{\beta}}) < \sigma^2(\hat{\boldsymbol{\beta}})$ and $\mu(\hat{\boldsymbol{\beta}})$ small, the negative binomial approximation can be tried.

Inference based on logistic regression with bootstrap samples is, of course, possible for all confounders whether categorical or not. But the solution remains approximate irrespective how large B, the number of samples, is chosen. On the other hand the procedure of Section 4 becomes exact when $B \to \infty$. Therefore, we prefer the latter method whenever possible, especially if some of the categories contain a small number of individuals.

6. Testing association over time

When dealing with the diffusion of innovation, as in our application, the non-adopters become adopters over time. If the change-points are recorded, we can study their dynamics. Unfortunately, exact description of such dynamics entails somewhat complicated notation. First, define

$$x_{ti} = \begin{cases} 1, & \text{if the individual } i \text{ becomes adopter at time } t \\ \text{otherwise}, & t = 1, \ldots, T. \end{cases} \tag{8}$$

We assume that there are no backward changes, i.e. no adopter changes back to a non-adopter. Then, with fixed i, the sequence $x_{1i}, \ldots x_{Ti}$ can be regarded as an indicator sequence giving the timing for a change-point to an adopter. If the change occurs at t, then $x_{ti} = 1$ and $x_{si} = 0$ for $s \neq t$. If $x_{1i} = \cdots = x_{Ti} = 0$, the individual i stays as a non-adopter for the whole period $1, \ldots, T$. Thus, the sum

$$y_{ti} = x_{1i} + \cdots + x_{ti}$$

describes the status of the individual at t. It is one for an adopter and zero for a non-adopter. The set

$$G_t = \{i \mid y_{ti} = 0\}$$

defines those individuals that are not adopted the innovation till t. The number of non-adopters at t is

$$N_t = \sum_i (1 - y_{ti}).$$

For completeness set $N_0 = N$, i.e. at the beginning all individuals are non-adopters. The number of new adopters at t is $M_t = \sum_i x_{ti}$. When considering the transition from t to $t+1$, the non-adopters at t, N_t in number, make up the group from which the new adopters, M_{t+1} in number, arise. Clearly the recursion $N_{t+1} = N_t - M_{t+1}$ holds.

With the above notation the association at t can be tested with $\sum \sum \delta_{ij} y_{ti} y_{tj}$ which is simply Q_{11} from (1) written at time t. But a more interesting question is whether there is association between successive time periods. First, consider two successive time periods t and $t+1$. The old adopters at $t+1$ are those with $y_{ti} = 1$ and the new ones those with $x_{t+1,i} = 1$. The association between the new and old adopters is measured by

$$\sum \sum \delta_{ij} x_{t+1,i} y_{tj}. \tag{9}$$

This reflects our assumption that the individual i is influenced by its neighbors but can influence only those whose neighbor it is. That is, if $\delta_{ij} = 1$ then farm j influences farm i. The choice $\sum \sum \delta_{ij} y_{ti} x_{t+1,j}$ would lead to the reverse interpretation.

Although we cannot speak of the influence of the new adopters to potential new adopters at the same time point $t+1$, it is, nevertheless, meaningful to assume that new adopters may cluster around themselves provided that the transition time from t to $t+1$ is long enough.

In our application, the interval is one year, offering plentiful chances for mutual contacts. Then, naturally, the measure between the new adopters is

$$\sum \sum \delta_{ij} x_{t+1,i} x_{t+1,j} \tag{10}$$

again Q_{11} from (1) applied to new adopters at t. The measures (9) and (10) could be used separately but we prefer their sum

$$Q_t = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} x_{t+1,i} y_{tj} + \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} x_{t+1,i} x_{t+1,j}, \tag{11}$$

$$Q_t = \sum_{i \in G_t} \sum_{j=1}^{N} \delta_{ij} x_{t+1,i} y_{tj} + \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij} x_{t+1,i} x_{t+1,j}, \tag{12}$$

which combines the association measures between old and new adopters and that within new adopters. Eq. (12) follows from the fact that $x_{t+1,i} = 0$, if i has adopted the innovation before $t + 1$, or with our notation $i \notin G_t$. This formula is needed when permutations are performed. We remind that Q_t is related to the step from t to $t + 1$. Note also that Q_t is not a special case of Mantel's statistic (4).

The hypothesis of no space-time association between t and $t + 1$ may be stated in terms of the random variables $X_{t+1,i}$ conditioned on the observed values y_{ti}, as follows

$$X_{t+1,i} = \begin{cases} 0, & i \notin G_t, \\ X_{t+1,i} \sim_{\text{ind}} \text{Ber}(\theta_t), & i \in G_t. \end{cases}$$

Again, arguing conditionally on the sum

$$M_{t+1} = \sum_{i \in G_t} X_{t+1,i}$$

the significance of Q_t is found from the permutation distribution of

$$Q_t^{\pi} = \sum_{i \in G_t} \sum_{j} \delta_{ij} x_{t+1,i}^{\pi} y_{tj} + \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij} x_{t+1,i}^{\pi} x_{t+1,j}^{\pi},$$

where the values $x_{t+1,i}^{\pi}$ are obtained by randomly permuting $x_{t+1,i}$, $i \in G_t$. The values y_{ti} are treated fixed. The expectation is

$$E(Q_t^{\pi}) = \frac{M_{t+1}}{N_t} \sum_{i \in G_t} \sum_{j} \delta_{ij} y_{tj} + \frac{M_{t+1}(M_{t+1} - 1)}{N_t(N_t - 1)} \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij},$$

The variance is found by direct calculation

$$\text{var}(Q_t^{\pi}) = (p_{t0} - p_{t1}) \sum_{i \in G_t} (z_{ti} - \bar{z}_t)^2$$

$$+ 2 \left[(p_{t1} - p_{t2}) \sum_{i \in G_t} (k_{ti} + q_{ti}) z_{it} + p_{t2} S_t S_{tz} - p_{t0} p_{t1} S_t S_{zt} \right]$$

$$+ (p_{t1} - 2p_{t1} + p_{t3})(S_t + R_t) + (p_{t2} - p_{t3}) \sum_{i \in G_t} (k_{it} + q_{it})^2 + (p_{t3} - p_{t1}^2) S_t^2$$

with

$$p_{tk} = \frac{M_{t+1}(M_{t+1} - 1) \cdots (M_{t+1} - k)}{N_t (N_t - 1) \cdots (N_t - k)},$$

$$S_t = \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij},$$

$$R_t = \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij} \delta_{ji},$$

$$k_{ti} = \sum_{j \in G_t} \delta_{ij}, \quad i \in G_t,$$

$$q_{tj} = \sum_{i \in G_t} \delta_{ij}, \quad j \in G_t,$$

$$z_{ti} = \sum_j \delta_{ij} y_{tj}, \quad i \in G_t.$$

For easy reference, we give the algorithm as follows.

1. For every individual $i = 1, \ldots, N$ find the time t when i changes to an adopter. Then set $x_{ti} = 1$ and $x_{si} = 0$ for all $s \neq t$. If i stays non-adopter set $x_{ti} = 0$ for all $t = 1, \ldots, T$.
2. Compute the cumulative sums $y_{ti} = x_{1i} + \cdots + x_{ti}$, $t = 1, \ldots, T$, $i = 1, \ldots, N$.
3. Find the index sets $G_t = \{i \mid y_{ti} = 0\}$, i.e. the non-adopters at t, $t = 1, \ldots, T$.
4. Repeat the steps 5–7 for $t = 2, \ldots, T$.
5. Compute the observed test statistic.

$$Q_t = \sum_{i \in G_t} \sum_j \delta_{ij} x_{t+1,i} y_{tj} + \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij} x_{t+1,i} x_{t+1,j}.$$

6. Randomly permute the indices i on G_t and compute

$$Q_t^{\pi} = \sum_{i \in G_t} \sum_j \delta_{ij} x_{t+1,i}^{\pi} y_{tj} + \sum_{i \in G_t} \sum_{j \in G_t} \delta_{ij} x_{t+1,i}^{\pi} x_{t+1,j}^{\pi}.$$

7. The p-value for the spatio-temporal association between t and $t + 1$ is

$$\frac{\#\{Q_t^{\pi} \geq Q_t\}}{B},$$

where B is the number of random permutations.

A natural overall test statistic is obtained by the sum $\sum_{t=1}^{T-1} Q_t$. Repeating the argument for simultaneous association between new adopters discussed above, and applying it to the

first adopters at $t = 1$ leads to the association measure

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} x_{1i} x_{1j} + \sum_{t=1}^{T-1} Q_t.$$

Let us develop this formula further. Using (11) and the definition of y_{ti} together we have

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} x_{1i} x_{1j} + \sum_{t=1}^{T-1} \sum_{i} \sum_{j} \delta_{ij} x_{t+1,i} (y_{t,j} + x_{t+1,j}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{ij} \sum_{t=1}^{T} x_{ti} y_{tj}.$$

Now we are ready to give Q a clear interpretation. Suppose that both i and j have adopted the innovation at some time point $1, \ldots, T$. Further, suppose that i adopts the innovation at s. Then $x_{si} = 1$ and $x_{ti} = 0$ for $t \neq s$ yielding $\sum_{t=1}^{T} x_{ti} y_{tj} = y_{sj}$. Because y_{sj} is 1 if j adopts the innovation not later than at s, we can rewrite

$$Q = \sum \sum \delta_{ij} v_{ij}, \tag{13}$$

with

$$v_{ij} = \begin{cases} 1, & \text{if } j \text{ has adopted the innovation not later than } i, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Note that Q in (13) is of Mantel's form (4).

The appropriate null hypothesis of no temporal association can be defined by first specifying the initial distribution for the X_{1i}

$$X_{1i} \sim_{\text{ind}} \text{Ber}(\theta_1), \quad i = 1, \ldots, N,$$

and then the conditional distributions for the X_{ti} given the past values $X_{1i}, \ldots, X_{t-1,i}$ as follows:

$$X_{ti} = 0 \qquad \text{if some } X_{si} = 1, \ s = 1, \ldots, t-1,$$
$$X_{ti} \sim_{\text{ind}} \text{Ber}(\theta_t) \quad \text{if } X_{1i} = \cdots = X_{t-1,i} = 0 \ t = 2, \ldots, T.$$

In addition, let us define $X_{T+1,i} = 1 - X_{1i} - \cdots - X_{Ti}$. Then the vectors $(X_{1i}, \ldots, X_{T+1,i})$ are independent draws from the multinomial distribution with class probabilities

$$P(X_{1i} = 1) = \theta_1,$$
$$P(X_{ti} = 1) = (1 - \theta_1) \cdots (1 - \theta_{t-1}) \theta_t, \quad t = 2, \ldots, T,$$
$$P(X_{T+1,i} = 1) = (1 - \theta_1) \cdots (1 - \theta_T).$$

Note that the class $T + 1$ consists of those that remain non-adopters during the whole time span $t = 1, \ldots, T$. The class frequencies $M_t = \sum X_{ti}$ (the number of new adopters at t) are the sufficient statistics. This implies that the permutations of the observed vectors $(x_{1i}, \ldots, x_{T+1,i})$ are all equally probable under the null hypothesis of no temporal association. Therefore, the permutational distribution could be given in terms of these variables. But we prefer another formula that is more amenable to simulation.

Define t_i to be the time when i adopts the innovation, $i = 1, \ldots, N$. If i remains non-adopter, then we set $t_i = T + 1$. Let us introduce the unit step function $u(x) = 1$ for $x \geq 0$ and 0 otherwise. Then by (14)

$$v_{ij} = u(t_i - t_j)u(T - t_i)u(T - t_j).$$

With this notation

$$Q = \sum_i \sum_j \delta_{ij} u(t_i - t_j)u(T - t_i)u(T - t_j). \tag{15}$$

The permutation distribution is obtained by permuting the timings t_i

$$Q^\pi = \sum_i \sum_j \delta_{ij} u(t_i^\pi - t_j^\pi)u(T - t_i^\pi)u(T - t_j^\pi).$$

Its first two moments can be obtained from the tables of Mantel (1967). We give here only the expectation

$$E(Q^\pi) = \frac{S}{N(N-1)} \left[\sum_{t=1}^{T} M_t \left(\sum_{s=1}^{t} M_s \right) - \sum_{t=1}^{T} M_t \right].$$

7. Application

7.1. Cross-section data in 1998

Let us start our analysis with the data in 1998. In that year we have full information on all 3685 farms (of which 250 or 6.8% are organic), including the exact coordinates of their ranch houses. The neighborhood is defined as those farms having the ranch house within the distance to the nearest neighbor plus 1 km. Then 156 (= Q_{11} in the notation of Section 2) organic farm pairs are observed. The expected number of organic farm pairs under the randomness hypothesis is 79.1 with standard deviation 12.2. Since the expectation is fairly large the normal approximation can be tried. This leads to $Z = (156 - 79.1)/12.2 = 6.29$ that is significant at all reasonable levels. The negative binomial approximation does not contradict this conclusion. The maximum of Q^π in 1000 random permutations was 131 well below the observed value 156. This is in line with the previous approximations.

Because the variance of Q^π is complicated, it might be easier to compute the first two moments by simulation and then use normal or negative binomial approximation with these estimates. The computation of moments usually needs less replications than the actual p-value. Also the accuracy of approximations may be checked against the results from the simulation. In the present case 1000 simulations yield a mean of 79.45 and a standard deviation of 12.36. To get a better understanding the accuracy of the approximation suppose for a moment that we would have observed 110 instead of 156. Then the normal deviate is 2.47 yielding $p = 0.007$. The negative binomial approximation gives $p = 0.011$ that is in a close agreement to the simulated $p = 0.012$. With $Q = 120$ we would have p-values 0.0005 (normal), 0.0016 (negative binomial) and 0.003 (simulated).

Table 1
Spatial associations for various variables in 1998

	Q	Mean	S.D.	Normal	Negative binomial
Organic farms	156	79.1	12.2	0.0000	0.0000
GRAIN	11420	10262	104	0.0000	–
PLANT	3614	3404	80.1	0.0044	–
ANIM1	7577	7094	102	0.0000	–
ANIM2	296	271	23.0	0.14	0.14

Table 2
Associations between organic farming and types of products

	χ^2	d.f.	p	Type of association	Production (%)[a]	No production (%)[b]
GRAIN	18.0	1	0.0000	Positive	7.7	3.6
PLANT	71.1	1	0.0000	Positive	10.7	3.7
ANIM1	61.3	1	0.0000	Negative	4.4	11.1
ANIM2	29.3	1	0.0000	Positive	12.8	5.9

[a] Percentage of organic farms among those farms having the type production.

[b] Percentage of organic farms among those farms not having the type production.

Having found a significant spatial association between organic farmers we might ask, if it is due to the types of products they produce which might be spatially correlated. We made four dichotomous variables GRAIN (grain production), PLANT (other plants than grain), ANIM1 (pork, beef and/or milk production) and ANIM2 (other animal products). Apart from ANIM2 they are seen to be spatially correlated (see Table 1). Treating these variables separately they all are significantly associated with organic farming (Table 2). Among those who have grain or other plant production or other animal products, the proportion of organic farmers is more than twice as high as among the complementary groups. Among milk and meat producers the association is negative but equally strong.

We wish now to eliminate the possible confounding effects of GRAIN, PLANT, ANIM1 and ANIM2. Because one farm can produce several products or none of the chosen ones, we have 16 disjoint classes within which we make the random permutations. Thus, the adjustment is done not only to main effect but also for all interactions. The observed value remains the same 156, but its expectation from (6) increases to 85.7. In 10000 simulation all values of the test statistic remain below 156 (the row "categorical" in Table 3). Thus, adjusting for the chosen covariates does not remove all spatial autocorrelation.

Table 3
Spatial association of organic farms adjusted for covariates

Type of covariates	Q	Mean	S.D.	Normal	Negative binomial	Simulated
Purely categorical	156	85.7	13.0[a]	0.0000	0.0000	0.0000
Mixed	156	100	18.8	0.0014	0.0042	0.0033 ± 0.0015[b]

Simulations are based on 10000 replications.

[a] Based on simulations.

[b] 99% confidence interval.

It turns out that organic farming is associated with the size of the neighborhoods. The observed values are $k'x = 1284$ (out-degrees) and $q'x = 1263$ (in-degrees). The expected value is 1171 for both. The observed values are significantly higher ($p = 0.004, 0.015$, respectively, by normal approximation). Thus, there seems to be more organic farms in those parts of the province which are more densely populated (in the sense that farms there have more neighbors).

The next step is to adjust for both the neighborhood size and production sector effects. This is done via logistic regression. We use out- and in-degrees together with GRAIN, PLANT, ANIM1 and ANIM2 as predictors. For simplicity, here we use only the main effects of the production sector variables. The coefficients of degree predictors are not significant. This is mainly due their correlation with each other. If we used only one of them, it would be significant. Because we are not interested in distinguishing their actual effects but regard their role purely confounding, we decided to adjust for them both. The bootstrap test of Section 5 is significant (the row "mixed" in Table 3) showing that the spatial autocorrelation cannot be attributed to neighborhood size and production types. Note again that the test statistic stays at the same value 156 but the expectation from (7) further increases to value 100 and the negative binomial approximation is clearly inside the confidence interval.

7.2. Time series 1990–1999

Table 4 gives the yearly tests for the randomness hypothesis. Before 1995 the number of organic farmers was too small to indicate any spatial clustering. But from 1995 on, the randomness hypothesis is rejected in each year. The Monte Carlo *p*-value is computed for the year 1995 only, because in the other years the normal and negative binomial approximations show that we would need perhaps millions of replications to get a value exceeding the observed one.

The theory in Section 6 assumes that farmers do not stop farming and that once they have switched to organic farming, they do not switch back. However, in real life, 22 farmers quit farming and 5 switched back from organic to traditional methods. The yearly tests in Table 4 are applied to the farms that are present in the data base in the corresponding year. In the study of the spatio-temporal dynamics we have to simplify things slightly, and we proceed as follows. We take all farms present in the data base at least once in 1995–1999. In each year, a farm is organic, inorganic or absent. If a farm is in some years absent but

Table 4
Yearly spatial associations

Year	Q	Mean	S.D.	Normal	Negative binomial	Simulated
1995	35	16.5	5.53	0.0004	0.0031	0.0034 ± 0.0015[a]
1996	83	36.7	8.29	1.2×10^{-8}	4.1×10^{-6}	–
1997	85	36.7	8.29	2.9×10^{-9}	1.9×10^{-6}	–
1998	147	70.9	11.62	4.7×10^{-10}	2.0×10^{-7}	–
1999	173	92.2	13.29	6.1×10^{-10}	1.5×10^{-7}	–

Simulations are based on 10000 replications.
[a] 99% confidence interval.

Table 5
Spatial associations between new organic farms and the old ones in two successive years

Years	Q	Mean	S.D.	Normal	Negative binomial	Simulated
1995–1996[a]	28	13.7	4.37	0.0005	0.0032	0.0028 ± 0.0014[b]
1997–1998[a]	37	24.4	5.87	0.016	0.027	0.0234 ± 0.0039[b]
1998–1999[a]	23	16.3	4.68	0.076	0.098	0.1021 ± 0.0078[b]
Overall[c]	114	63.9	9.16	2.3×10^{-8}	8.9×10^{-7}	0.0000

Simulations are based on 10000 replications.

[a] From (12).

[b] 99% confidence interval.

[c] From (15).

otherwise inorganic, we keep its status as inorganic also in the years of absence. If a farm is organic in some year, we keep it organic also in the later years, even if it becomes absent or switches back to inorganic farming. Before the first appearance as organic the farm is regarded always as inorganic.

Table 5 gives results for testing time dependence between two successive years t and $t + 1$ based on Q_t defined in (12). In 1997 there are only 2 new organic farms, which is too few to allow any inference. The tests, though significant in 1996 and 1998, show increasing tendency towards nonsignificance. Also the ratios of the observed counts over the expected ones decrease. They are 2.0, 1.5 and 1.4 in succession. This may be an indication that the neighborhood effect is diminishing among new adopters. This overall measure, based on (15), clearly significant, indicates a definite tendency of new organic farms to follow neighbors who adopted earlier.

8. Discussion

This paper is intended to make both methodological and substantive contributions to study of the effect of spatial proximity on the adoption of a behavioral practice. The specific case we treat is the adoption by farmers of organic farming techniques. Some of the models we present are well known individually; in such cases the contribution of our paper is in bringing them all together in a single framework with comparable notation. Other models, such as the combined space-time model, are original.

We began our analysis with a traditional permutation test of spatial autocorrelation, of the kind well-known to geographers (e.g. Cliff and Ord, 1981). The results showed unequivocally that organic farms are much more likely to have organic farms as neighbors than we would expect by chance if adoption of organic methods were independent of proximity. These results are very much in line with a long string of similar results reported in the literature, and establish that there is something worth studying here.

What is missing from this general test—and from most previous research of this type—is a sense of process and time. This approach essentially observes the cumulated outcomes to date of an unknown process and establishes the unlikeliness of independence from spatial proximity. But, having done so, we do not know whether we are looking at an influence-type process (farmer i is influenced by neighbor j), or whether there is an unobserved underlying

variable such as soil conditions that explains the autocorrelation. In addition, assuming that organic farming is an innovation that could eventually spread to all farms, we were fortunate that we observed the system at an intermediate point in time when a good number of farms are organic and a good number were still inorganic, since the traditional cross-sectional autocorrelation test would have difficulty detecting the spatial effect at the endpoint of the process, when nearly all farms are organic.

Hence, to go beyond simply establishing spatial autocorrelation, we need the ability to (a) control for other variables, and (b) examine the interplay of proximity and adoption over time. In this paper, we have sought to provide both. Our first step was to suggest a method of controlling for categorical covariates. The basic technique we proposed (restricting per-mutations within categories of the covariates) is known in the statistical literature, but has not previously been applied to the study of diffusion. Using this technique, we established that the spatial autocorrelation among organic farmers could not be explained in terms of the kinds of products produced by the farms.

We then consider the case of non-categorical covariates. Here, we propose an approximate logistic regression method. Using this method, we controlled for not only type of production, but the number of farms that each farm is adjacent to. The result was that the observed spatial autocorrelation could not be attributed to production type nor the number of neighbors.

The other contribution we make to the study of adoption of innovation is the development of an approach to examining adoption over time. Consider a hypothetical study in which we observe that organic farmers at the present time tend to be clustered spatially. Suppose we now examine the pattern of adoption over time and find that adoptions at time t have no particular spatial relation to the adopters of previous times. Rather than an influence process, such a pattern would be more consistent with a "geographic suitability" hypothesis that says that certain geographical areas have a higher predisposition to acquire organic farms, and these will appear in random order. In contrast, if the pattern over time is that new adoptions tend to appear adjacent to farms that recently adopted, this is consistent with a contagion or influence process from neighbor to neighbor. To test this, we developed an easily calculated model that tests the null model that adoptions do not tend to occur in farms adjacent to farms that have already adopted. The test can be calculated year-by-year, or for all years together. The results showed that, overall, there is a clear tendency for organic farms to develop adjacent to existing organic farms, supporting the hypothesis of a social influence process.

Interestingly, the year-by-year test appears to show that this effect is declining over time. This is consistent with institutional theory (Meyer and Rowan, 1977; Scott, 1991), which argues that legitimacy is the key factor in adoption of innovation. The argument says that, in the beginning, conventional norms favor traditional farming techniques, and adoption of organic techniques is done by a few independent thinkers, along with farmers that come into close contact with the rebels and are influenced by them. This creates strong spatio-temporal autocorrelation at the early stages. As more and more farmers adopt, however, organic farming becomes legitimized and enters the mainstream as a "normal" option. Once this occurs, the decision process becomes more of a cost/benefit analysis conducted individually by farmers, weakening the spatio-temporal diffusive effect (but not necessarily the simple spatial autocorrelation if there are some regions that are more suitable than others). In addition to institutional and cultural reasons for the weakening spatio-temporal effect, we

　J. Nyblom et al. / Social Networks 25 (2003) 175–195

also have to take into consideration the importance of the agricultural policy in European Union (Finland has been a member state since 1995). The growing political interest in organic farming after mid-1990s has led to the development of subsidy programs which have clearly made organic farming more attractive among farmers with increasing economic problems (Michelsen, 2001).

It is important to realize that all of the data we have studied in this paper are archival data obtained from government records. Such data are much less expensive to obtain than survey data, and the analytical methods we propose allow the researcher to make tests which could obviate the need for surveys or, on the other hand, provide justification for taking the next step. In our case, the spatial autocorrelation effect survived controlling for covariates, and showed a temporal pattern consistent with a social influence process. These results provide justification for a survey of farmers to understand how they may be influencing each other to adopt—for example, is it simply that adoption provides a model that neighbors can copy? Or are organic farmers actively persuading neighbors to join? In order to answer these questions two of us (Roslakka and Salo, 2001) have conducted a survey among the organic and inorganic farmers in which we investigate the ways the farmers interact with and possibly influence and are influenced by their neighbors.

Of course, the methods we propose have limitations. One important limitation of the spatio-temporal models is that they assume that once an adoption is made, it remains. That is, no farmer could switch back to inorganic methods. In our application this is not a problem, since the number of farmers who switched back is insignificant. In other applications, however, such as consumer purchasing of products, this could constitute a serious limitation. Another important limitation is that the models assume that the underlying structure of adjacency does not change over time. Again, this causes us no problems in the case of farms, but could be a significant factor in applications where adjacency refers to social relations such as friendship. There, it would not be surprising if the pattern of friendships changed over time, possibly even as a result of the adoption/non-adoption decision (i.e. organic farmers become friends as a result of both being organic farmers).

References

Burt, R.S., 1987. Social contagion and innovation: cohesion versus structural equivalence. American Journal of Sociology 92, 1287–1335.
Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion Ltd., London.
Cuzick, J., Edwards, R., 1990. Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society B 52, 73–104.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.
Feller, W., 1957. An Introduction to Probability Theory and Its Applications, vol. I, 2nd ed. Wiley, New York.
Geary, R.C., 1954. The contiguity ratio and statistical mapping. The Incorporated Statistician 5, 15–145.
Good, P., 1994. Permutation Tests. Springer, New York.
Hägestrand, T., 1953. Innovationsförloppet ur korologisk synpunkt (Innovation Diffusion as a Spatial Process). Lunds Universitet, Lund.
Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.
Lehmann, E.L., 1994. Testing Statistical Hypotheses, 2nd ed. Chapman & Hall, New York.
Meyer, J.W., Rowan, B., 1977. Institutionalized organizations: formal structure as myth and ceremony. American Journal of Sociology 83, 340–363.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27, 209–220.

Mehta, C.R, Patel, N.R., Senchaudhuri, P., 2000. Efficient Monte Carlo methods for conditional logistic regression. Journal of the American Statistical Association 95, 99–108.

Michelsen, J., 2001. Recent development and political acceptance of organic farming in Europe. Sociologia Ruralis 41, 3–20.

Moran, P.A.P., 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society B 10, 243–251.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17–23.

Noreen, E.W., 1989. Computer Intensive Methods for Testing Hypotheses: An Introduction. Wiley, New York.

Pike, M.C., Smith, P.G., 1974. A case–control approach to examine diseases for evidence of contagion, including diseases with long latent period. Biometrics 30, 263–279.

Rogers, E.M., 1995. Diffusion of Innovations, 4th ed. Free Press, New York.

Roslakka, J., Salo, M.A., 2001. The effect of the farmers' personal networks on the adoption of organic farming, in press.

Scott, W.R., 1991. Unpacking institutional arguments. In: DiMaggio, P.J., Powell, W.W. (Eds.), The New Institutionalism in Organizational Analysis. University of Chicago Press, Chicago, pp. 164–182.

Valente, T.W., 1995. Network Models of the Diffusion of Innovations. Hampton Press, Cresskill.

Whaley, F.S., 1983. The equivalence of three individually derived permutation procedures for testing the homogeneity of multidimensional samples. Biometrics 39, 741–745.