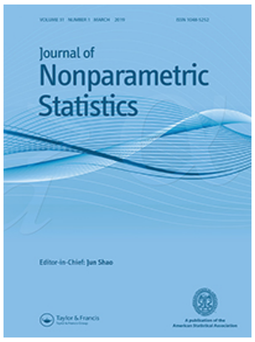


Taylor & Francis
Taylor & Francis Group

Journal of Nonparametric Statistics

ISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Moran's *I* statistic-based nonparametric test with spatio-temporal observations

Y. Xiong, D. Bingham, W. J. Braun & X. J. Hu

To cite this article: Y. Xiong, D. Bingham, W. J. Braun & X. J. Hu (2019) Moran's *I* statistic-based nonparametric test with spatio-temporal observations, Journal of Nonparametric Statistics, 31:1, 244-267, DOI: 10.1080/10485252.2018.1550197

To link to this article: https://doi.org/10.1080/10485252.2018.1550197

Published online: 25 Nov 2018.

JOURNAL OF NONPARAMETRIC STATISTICS
2019, VOL. 31, NO. 1, 244–267
https://doi.org/10.1080/10485252.2018.1550197

Moran's *I* statistic-based nonparametric test with spatio-temporal observations

Y. Xiong[a], D. Bingham[a], W. J. Braun[b] and X. J. Hu[a]

[a]Department of Statistics and Actuarial Science, Simon Fraser University, Canada; [b]Department of Computer Science, Mathematics, Physics, and Statistics, University of British Columbia, Okanagan, Canada

ABSTRACT

Moran's *I* statistic [Moran, (1950), 'Notes on Continuous Stochastic Phenomena', *Biometrika*, 37, 17–23] has been widely used to evaluate spatial autocorrelation. This paper is concerned with Moran's *I*-induced testing procedure in residual analysis. We begin with exploring the Moran's *I* statistic in both its original and extended forms analytically and numerically. We demonstrate that the magnitude of the statistic in general depends not only on the underlying correlation but also on certain heterogeneity in the individual observations. One should exercise caution when interpreting the outcome on correlation by the Moran's *I*-induced procedure. On the other hand, the effect on the Moran's *I* due to heterogeneity in the observations enables a regression model checking procedure with the residuals. This novel application of Moran's *I* is justified by simulation and illustrated by an analysis of wildfire records from Alberta, Canada.

ARTICLE HISTORY
Received 26 October 2017
Accepted 14 November 2018

KEYWORDS
Heterogeneity; model checking; regression residuals; spatio-temporal correlation; validity of assumption

AMS SUBJECT CLASSIFICATIONS
46N30; 67H11; 67K70

1. Introduction

The Moran's I statistic (Moran 1950) is commonly used to measure global spatial autocorrelation. With a collection of observations $\{Z_i : i = 1, \ldots, n\}$ from n individual units and a set of corresponding weights $\{w_{ij} : i = 1, \ldots, n; j = 1, \ldots, n\}$, the statistic is

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{W_0\sum_{i=1}^{n}(Z_i - \bar{Z})^2} = \frac{1}{W_0\hat{\sigma}^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z}),$$

where $\bar{Z}$ is the sample mean $\sum_{i=1}^{n} Z_i/n$, $\hat{\sigma}^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2/n = (n-1)S^2/n$ with S^2 the sample variance, $W_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ is the sum of the weights. With $w_{ii} = 0$, the original definition for the weights is $w_{ij} = 1$ for $i \neq j$ if units i and j are in the same 'neighbourhood'; otherwise, $w_{ij} = 0$. Geary's C statistic (Geary 1954), another popular

CONTACT X. J. Hu joanh@stat.sfu.ca

autocorrelation measure, is inversely related to the Moran's I as

$$C = \frac{1}{2W_0S^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - Z_j)^2 = \frac{\hat{\sigma}_w^2}{S^2} - \frac{n-1}{n}I,$$

where $\hat{\sigma}_w^2 = \sum_{i=1}^{n} w_{i.}(Z_i - \bar{Z})^2/W_0$ is the weighted sample variance with $w_{i.} = \sum_{j=1}^{n} w_{ij}$. Recent applications of the Moran's I range over a variety of fields, including real estate (Dubé and Legros 2012), public health (Jones et al. 2008; Helbich, Leitner, and Kapusta 2012), forest fires (Martell and Sun 2008), and disease mapping (Oden 1995).

The testing procedure based on the Moran's I is arguably the most widely used statistical method for testing spatial independence. Concerns have been raised about performing the Moran's I test with observations that depart from the conventional assumption of regional homogeneity. Alternative methods to accommodate heterogeneity in the data have been proposed (Oden 1995; Assunção and Reis 1999; Zhang and Lin 2016). In particular, Li, Calder, and Cressie (2007) emphasise the need to understand the limitations of the Moran's I when applying it to regression residuals to assess the model fit. Many practitioners, however, still take the Moran's I as an indicator of spatial autocorrelation regardless of whether the data satisfy the assumption of regional homogeneity. In addition, although several excellent exceptions exist (e.g. Dubé and Legros 2012), relatively little research has focused on the validity of applying the Moran's I test to spatial data collected over time.

Xiong (2015) presents a regression analysis of forest-fire records to explore how forest fires are related to ecological/environmental factors. The model checking with the regression residuals in Xiong (2015) adapts naturally the Moran's I-based procedure to test whether there is spatio-temporal correlation among the fires. The heterogeneity underlying the residuals, such as their different means or variances, appears to contribute a great deal to the Moran's I statistic. Misspecification of the regression model and the difference between the targeted and fitted models combine to contribute to the heterogeneity. It is thus unclear what causes the discrepancy between the resulting value of the Moran's I and the expected value under the null hypothesis.

This research explores the Moran's I statistic in several situations, both analytically and numerically. We pay particular attention to the performance of the Moran's I testing procedure in situations where the usual assumption of regional or temporal homogeneity is violated. We focus on two types of weights used to define the statistic: a natural adaptation of the weights originally proposed by Moran (1950) for spatio-temporal data, and an extended version of the weights considered in Dubé and Legros (2012). We observe that the magnitude of the Moran's I in general depends not only on the underlying correlation but also on certain heterogeneity in the individual observations. This suggests that one should exercise caution when interpreting the outcome of a test for correlation by the Moran's I. On the other hand, the effect on the Moran's I of heterogeneity in the individual observations makes it possible to conduct model diagnosis in regression analysis with the residuals via an inferential procedure based on the Moran's I.

The rest of this paper is organised as follows. Section 2 presents some theoretical preliminaries on the Moran's I. Section 3 examines the performance of the procedure for testing spatio-temporal correlation firstly with generated spatio-temporal data and then with simulated regression residuals. Section 4 discusses an application of the Moran's I based on the findings in Sections 2 and 3 using the records for the wildfires in Alberta, Canada in 2006. Section 5 provides concluding remarks.

2. Preliminaries on Moran's *I* statistic

2.1. Weights in Moran's I with spatio-temporal observations

Using appropriate weights in the Moran's I can be critical in several regards. For a collection of spatio-temporal observations $\{Z_i : i = 1, \ldots, n\}$, denote the time and location specified by its longitude and latitude associated with observation Z_i by t_i and $\mathbf{s}_i = (s_{1i}, s_{2i})'$, respectively. This paper focuses on symmetric weights. Particularly, we consider the following two types of weights in the numerical studies.

Adaptation of original Moran's I weights. We adapt the weights originally proposed in Moran (1950) and specify the neighbourhood as follows. Individuals i and j are viewed as ' neighbours' if the distance between their locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is within a prespecified limit d and their associated times t_i and t_j differ by at most a prespecified value τ. Thus, when $i \neq j$, weight w_{ij} indicates whether i and j are neighbours in space and time. We denote the resulting Moran's I by $I(d, \tau)$, a spatio-temporal extension of the original Moran's I; the latter can be viewed as $I(d, \infty)$ in our notation.

Dubé–Legros weights. The weights proposed by Dubé and Legros (2012) include the magnitudes of the location and time differences between observations. In our notation, $w_{ii} = 0$ and for $i \neq j$ the weight is

$$w_{ij} = \begin{cases} \|\mathbf{s}_i - \mathbf{s}_j\|^{-\gamma} |t_i - t_j|^{-\alpha}, & \text{if } \|\mathbf{s}_i - \mathbf{s}_j\| < d, \quad |t_i - t_j| < \tau \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $\|\mathbf{a} - \mathbf{b}\|$ is the Euclidean distance between locations $\mathbf{a}$ and $\mathbf{b}$, d, τ, γ, α are prespecified nonnegative constants, and the convention $0^{-\alpha} \equiv 1$ is taken. We denote the Moran's I with the Dubé–Legros weights by $I^*(d, \tau; \gamma, \alpha)$ or simply $I^*(d, \tau)$ with γ and α suppressed in the remainder of this paper, where the values of γ and α are clear from the context.

One may choose to use the values of d and τ that are practically meaningful in the application. We suggest to evaluate the Moran's I with multiple combinations of d and τ when conducting residual analysis for model diagnosis. This is illustrated via simulation and real data analysis in Sections 3 and 4.

In practical situations with a large difference between the magnitudes of the location distance and the time lag, we suggest to consider scaled location distance and/or time lag. This is exemplified in the simulation. One may choose to use different values of γ and α, to check whether the conclusion is robust.

2.2. Moran's I-based hypothesis testing

For a collection of observations $\{Z_i : i = 1, \ldots, n\}$, the Moran's I procedure for testing correlation may be presented as a test with the null hypothesis H_0: 'Z_i's are independent' and the test statistic as the standardised Moran's I under H_0 given the assumption of *homogeneity* $E(Z_i) = \mu$ and $\text{Var}(Z_i) = \sigma^2$ for $i = 1, \ldots, n$. For large sample size n, the rejection region of the testing procedure with an approximate type-I error rate of α is

$$\left\{a : a > E(I \mid H_0) + z_{1-\alpha/2}\sqrt{\text{Var}(I \mid H_0)} \text{ or } a < E(I \mid H_0) + z_{\alpha/2}\sqrt{\text{Var}(I \mid H_0)}\right\} \tag{2}$$

with observed Moran's I, where z_p is the $100p$ percentile of the standard normal distribution. Its theoretical justification is the asymptotic normality of the test statistic under H_0: $U = \{I - E(I \mid H_0)\}/\sqrt{\mathrm{Var}(I \mid H_0)} \to N(0,1)$ in distribution as $n \to \infty$, provided the population distribution has the first and second moments (Cliff and Ord 1981). Here the expectation $E(I \mid H_0) = -1/(n-1)$.

Assume that the weights w_{ij} are determined by $(\mathbf{s}_i, t_i)$, which are the locations and times associated with the individual observations; see, for example, the two sets of weights in Section 2.1. Since $E[I \mid (\mathbf{s}_i, t_i), i = 1, 2, \ldots, n] = E[I \mid w_{ij}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, n] \equiv -1/(n-1)$, the variance $\mathrm{Var}(I \mid H_0) = E\{\mathrm{Var}[I \mid H_0; (\mathbf{s}_i, t_i)'s]\}$. In general, one may calculate the variance using $E[I^2 \mid (\mathbf{s}_i, t_i), i = 1, 2, \ldots, n] = E\{E_R(I^2) \mid (\mathbf{s}_i, t_i), i = 1, 2, \ldots, n\}$ when n is large, where $E_R(I^2)$ is the second moment of I under the set of random permutations among the n location-time pairs, conditional on the observations $\{Z_i : i = 1, \ldots, n\}$. If the population distribution is normal,

$$E[I^2 \mid (\mathbf{s}_i, t_i), i = 1, 2, \ldots, n] = \left(n^2 W_1 - n W_2 + 3 W_0^2\right) / \left\{(n-1)(n+1) W_0^2\right\}$$

with $W_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$, $W_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (w_{ij} + w_{ji})^2$, and $W_2 = \sum_{i=1}^n (\sum_{j=1}^n w_{ij} + \sum_{j=1}^n w_{ji})^2 = \sum_{i=1}^n (w_{i.} + w_{.i})^2$. This gives

$$\mathrm{Var}[I \mid H_0; (\mathbf{s}_i, t_i)'s]\} = \frac{n^2 W_1 - n W_2}{(n^2 - 1) W_0^2} + \frac{2n - 4}{(n^2 - 1)(n - 1)}. \tag{3}$$

To explore the performance of the Moran's I test procedure, we derive the test statistic's expectations in the following settings, which differ from the one under H_0 combined with its assumed homogeneity of the observations. The difference between the expectation in each of the settings and the expectation under H_0 may reveal whether the testing procedure with data from the population can adequately detect the corresponding setting from the one under H_0. Let the Moran's I statistic be $I = I_N/I_D$ with $I_N = \sum_{i,j} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})/W_0$ and $I_D = \sum_{i=1}^n (Z_i - \bar{Z})^2/n = \hat{\sigma}^2$.

2.2.1. Expectation of Moran's I under H_a

When the Moran's I test is applied, the alternative hypothesis is often implicitly taken as H_a : 'H_0 is not true'. The following verifies the usefulness of the Moran's I procedure with the observations satisfying the *homogeneity* assumption for testing correlation specified as

H_a: 'Z_i's are correlated with $cov(Z_i, Z_j) = \rho_{ij} \not\equiv 0$ for $i \neq j$'.

With $\rho_{ii} = \sigma^2$, denote $\sum_{j=1}^n \rho_{ji} = \sum_{j=1}^n \rho_{ij}$ by $\rho_{i.}$. Provided the *homogeneity* assumption (the observations have the same mean and variance), which is often implicitly assumed in practice, the expectations of I_N and I_D are

$$E(I_N \mid H_a) = \frac{1}{W_0} \sum_{i,j} w_{ij}\rho_{ij} - \frac{1}{nW_0} \sum_{i} (w_{i.} + w_{.i})\rho_{i.} + \frac{\sigma^2}{n} + \frac{1}{n^2} \sum_{i \neq j} \rho_{ij}$$

and

$$E(I_D \mid H_a) = \left(1 - \frac{1}{n}\right)\sigma^2 - \frac{1}{n^2}\sum_{i \neq j} \rho_{ij}.$$

In the situations with symmetric weights w_{ij} and $\rho_{ij} = \rho w_{ij}$ for $i \neq j$, $E(I_N \mid H_a)$ reduces to

$$-\frac{1}{n}\sigma^2 + \frac{\rho}{W_0}\left\{\sum_{i,j} w_{ij}^2 - \frac{2}{n}\sum_i w_{i.}^2\right\} + \frac{W_0\rho}{n^2}$$

and $E(I_D \mid H_a)$ is approximate to $(1 - 1/n)\sigma^2 - (1/n^2)W_0\rho$.

Consider a simple set of weights:

$$w_{ij} = \begin{cases} 1, & \text{if } |i-j| = 1 \text{ or } n-1 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Here we view that units 1 and n are also ‘ neighbours’. Then $W_0 = 2n$, $W_1 = 4n$, $W_2 = 16n$. The expectation of the Moran's I under the alternative hypothesis $E(I \mid H_a)$ is now approximate to

$$-\frac{1}{n-1} + \frac{(n-4)\rho}{(n-1)\sigma^2 - 2\rho}.$$

With $E(I \mid H_0) = -1/(n-1)$ and $\mathrm{Var}(I \mid H_0) = n(n-3)/[(n+1)(n-1)^2]$ from Equation (3), the Moran's I test statistic $U \approx \sqrt{2n}\rho/\sigma^2$ differing from 0 significantly when n is sufficiently large. In particular, we see that the power to detect the correlation by the Moran's I test increases as $|\rho|/\sigma^2$ increases as well as the sample size n increases.

2.2.2. Expectations of Moran's I without assumed homogeneity

We now explore the robustness of the Moran's I test to the violation of the homogeneity assumption: $E(Z_i) = \mu$ and $\mathrm{Var}(Z_i) = \sigma^2$ for $i = 1, \ldots, n$. In particular, we examine the expectations of the Moran's I statistic under H_0 in the following two situations with observation heterogeneity:

Heterogeneity 1: $E(Z_i) = \mu$ and $\mathrm{Var}(Z_i) = \sigma_i^2 \not\equiv \sigma^2$.
Heterogeneity 2: $E(Z_i) = \mu_i \not\equiv \mu$ and $\mathrm{Var}(Z_i) = \sigma^2$.

Under H_0 with *Heterogeneity 1*: Denote $\sum_{i=1}^n \sigma_i^2/n$ by $\bar{\sigma}^2$. The expectations of I_N and I_D with *Heterogeneity 1* are

$$-\frac{1}{nW_0}\sum_{i=1}^n (w_{i.} + w_{.i})\sigma_i^2 + \frac{1}{n}\bar{\sigma}^2 \quad \text{and} \quad \left(1 - \frac{1}{n}\right)\bar{\sigma}^2,$$

respectively. When the weights are symmetric with $w_{ij} = w_{ji}$, the expectation of the Moran's I is close to

$$-\frac{1}{n-1} + \frac{2}{n-1}\left(1 - \frac{1}{W_0\bar{\sigma}^2}\sum_{i=1}^n w_{i.}\sigma_i^2\right).$$

The second term above is small when $w_{i.}$ do not vary much. That is, the expectation of the Moran's I is now close to the expectation of I under H_0 with the observation homogeneity assumption, $E(I \mid H_0) = -1/(n-1)$. The Moran's I test procedure for observation correlation can thus be rather robust to the heterogeneity in the variance. This analytical finding is verified by the simulation reported in Section 3.

Consider the special case with an even number of sample size n, and $\sigma_i^2 = \sigma_A^2$ if $i \leq n/2$ and $\sigma_i^2 = \sigma_B^2$ if $i > n/2$ with $\sigma_A^2 \neq \sigma_B^2$. Using the set of weights given in Equation (4), $I/E(I \mid H_0) \to 1$ almost surely as $n \to \infty$ and $V(I \mid H_0) = O(n^{-1})$. This indicates that the type I error of the test can be rather small in the situation.

Under H_0 with *Heterogeneity 2*: When there is heterogeneity in the mean, the test based on the Moran's I is less robust, compared to it in situations with *Heterogeneity 1*. To see this, denote $\sum_{i=1}^{n} \mu_i/n$ by $\bar{\mu}$. The expectations of I_N and I_D are now

$$-\frac{\sigma^2}{n} + \sum_{i,j} \frac{w_{i,j}}{W_0}(\mu_i - \bar{\mu})(\mu_j - \bar{\mu}) \quad \text{and} \quad \left(1 - \frac{1}{n}\right)\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \bar{\mu})^2,$$

respectively. We see that the expectation of the Moran's I is close to

$$-\frac{1}{n-1} + \frac{1}{1 + \tilde{\sigma}^2/\sigma^2}\left\{\frac{\tilde{\sigma}^2}{(n-1)\sigma^2} + \frac{n}{(n-1)W_0\sigma^2}\sum_{i,j} w_{ij}(\mu_i - \bar{\mu})(\mu_j - \bar{\mu})\right\}$$

with $\tilde{\sigma}^2 = \sum_{i=1}^{n}(\mu_i - \bar{\mu})^2/(n-1)$. While the first term in the curly brackets above can be small if n is large, the second is unlikely to be close to zero, especially if the products $(\mu_i - \bar{\mu})(\mu_j - \bar{\mu})$with i,j in the same neighbourhood have the same sign. Thus, even when the observations are truly independent, the procedure may reject H_0 with *Heterogeneity 2*.

Consider the special case with an even number of sample size n, and $\mu_i = \mu_A$ if $i \leq n/2$ and $\mu_i = \mu_B$ if $i > n/2$ with $\mu_A \neq \mu_B$. Using the set of weights given in Equation (4), the standardised Moran's I statistic U is now asymptotically equivalent to $\sqrt{2n}(\mu_A - \mu_B)^2/\{(\mu_A - \mu_B)^2 + 4\sigma^2\}$. Thus the test rejects H_0 with a high probability when n is large and/or $(\mu_A - \mu_B)^2$ is not small compared to σ^2. That is, the type I error can be rather high in those situations.

The finding that the Moran's I test is not robust to *Heterogeneity 2* may be used to diagnose a regression model. We elaborate this in the context of the application of the Moran's I statistic with regression residuals (Section 2.3), in the second simulation study (Section 3.2), and in the analysis of real data (Section 4).

2.3. Moran's I with regression residuals

Consider a response variable Y that follows $Y = f(\mathbf{x}) + \epsilon$, where $\mathbf{x}$ includes all the identified explanatory variables and the random error ϵ satisfies $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$. Let $\{(\mathbf{s}_i, t_i) : i = 1, \ldots, n\}$ be the set of locations and times associated with a collection of spatio-temporal observations $\{Y_i : i = 1, \ldots, n\}$. Suppose that $\mathbf{x}_i$ is a function of the location $\mathbf{s}_i$ and time t_i, and that $\mathbf{x}_i$ and $\epsilon_i = Y_i - f(\mathbf{x}_i)$ are independent.

The spatio-temporal correlation underlying the ϵ_i's can often be of interest. Since the ϵ_i's are unobservable in reality, model checking often uses the regression residuals, which are $e_i = y_i - \hat{g}(\mathbf{x}_i)$ with $\hat{g}(\cdot)$ the fitted model under the assumed regression model $Y = g(\mathbf{x}) + \epsilon$.

Note that

$$e_i = \epsilon_i + \big[f(\mathbf{x}_i) - g(\mathbf{x}_i)\big] + \big[g(\mathbf{x}_i) - \hat{g}(\mathbf{x}_i)\big],$$

which is a combination of the random fluctuation, the difference of the assumed model from the true model, and the difference of the fitted model from the assumed model. Thus, the Moran's I with the residuals e_i's may reveal the correlation combined with the heterogeneity in the residuals as a result of model misspecification together with the estimation precision. When the estimation procedure is unbiased $[E\{\hat{g}(\mathbf{x}) \mid \mathbf{x}\} = g(\mathbf{x})]$, the Moran's I test may be used for checking the model assumption that $f(\cdot) = g(\cdot)$ and the ϵ_i's are i.i.d. This use of the Moran's I is investigated by simulation in Section 3 and exemplified by real data in Section 4.

3. Simulation study

We conducted two simulation studies to examine the performance of the Moran's I test in various settings. The first study, which used spatio-temporal observations generated from normal distributions, is intended to verify our analytical findings presented in Section 2. The second study explores the performance of the test with regression residuals.

3.1. Simulation A: Moran's I hypothesis test

We simulated spatio-temporal observations $\{z_i : i = 1, \ldots\}$ as follows.

Step 1. Generate independently n locations $\mathbf{s}_i$ with each of the two location indices s_{1i} and s_{2i} from $Unif(0,1)$, the uniform distribution over $(0,1)$. Generate also the associated times $t_i \sim Unif(0,1)$.

Step 2. Conditional on the generated $\{(\mathbf{s}_i, t_i) : i = 1, \ldots, n\}$, generate the n-dimensional vector $\mathbf{z} = (z_1, \ldots, z_n)'$ from the multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$. To simulate different observations, we considered six combinations of the mean vector $\boldsymbol{\mu} = E(\mathbf{Z})$ and the covariance matrix $\Sigma = \mathrm{Var}(\mathbf{Z})$, denoted by Case A(a,b) with a = 0,1 for the two types of $\boldsymbol{\mu}$ and b = 0,1,2 for the three types of Σ.

(a) Two types of $\boldsymbol{\mu}$. Type 0: $\mu_i = E(Z_i) = 0$ for all $i = 1, \ldots, n$; Type 1: μ_i is defined according to the location $\mathbf{s}_i$ via $\mu_i = -3, -0.5, 1.0, 2.5$ for $\mathbf{s}_i \in (0, 0.5) \times (0, 0.5), (0, 0.5) \times [0.5, 1), [0.5, 1) \times (0, 0.5), [0.5, 1) \times [0.5, 1)$, respectively.

(b) Three types of Σ. Type 0: $\Sigma = \sigma^2\mathbf{I}$ with $\sigma^2 = 2^2$ and $\mathbf{I}$ the $n \times n$ identity matrix; Type 1: Σ the diagonal matrix with the elements on the diagonal σ_i^2 defined as $\sigma_i^2 = 0.1^2, 1.7^2, 2.6^2, 3.5^2$ for $\mathbf{s}_i \in (0, 0.5) \times (0, 0.5), (0, 0.5) \times [0.5, 1), [0.5, 1) \times (0, 0.5), [0.5, 1) \times [0.5, 1)$, respectively; Type 2: $\Sigma = (\sigma_{ij})_{n \times n}$ with $\sigma_{ij} = 2^2 \exp\{-0.1(|\mathbf{s}_i - \mathbf{s}_j| + |t_i - t_j|)\}$ for $i, j = 1, \ldots, n$.

Note that Cases A(0,0), A(0,1), A(1,0), A(1,1), A(0,2), and A(1,2) yield observations under H_0 with observation homogeneity, H_0 with observation *Heterogeneity 1*, H_0 with observation *Heterogeneity 2*, H_0 with both *Heterogeneity 1* and *2*, H_a with observation homogeneity, and H_a with observation *Heterogeneity 2*, respectively.

We evaluated the Moran's I in the form of $I(d,\tau)$, using the data generated in the above cases with the sample size $n = 50$, 200, or 400, neighbourhood distance limit $d = 0.3$, 0.6, or 1.0, and largest time lag $\tau = 0.1, 0.5, 1$, or ∞.

Table 1 presents the sample means and standard deviations based on 1000 evaluations in each of the simulation settings with $\tau = 0.5$. For comparison, we include (i) the conditional mean and variance of $I(d,\tau)$, where $E(I) = -1/(n-1)$ and $\mathrm{Var}\{I \mid (\mathbf{s}_i, t_i)'s\}$ was calculated following Equation (3) given a set of generated locations and times $\{(\mathbf{s}_i, t_i) : i = 1, \ldots, n\}$, and (ii) the marginal mean and variance of $I(d,\tau)$, where $E(I) = -1/(n-1)$ and $\mathrm{Var}(I)$ was approximated by the sample mean of the evaluations of $\mathrm{Var}\{I \mid (\mathbf{s}_i, t_i)'s\}$ with 1000 sets of generated locations and times.

The simulation results verify our findings from the analytical study presented in Section 2.2. The sample means and standard deviations of $I(d,\tau)$ in Case A(0,0), which generates i.i.d. normal observations, are rather close to the corresponding conditional or marginal means and standard deviations under H_0. Although the sample standard

Table 1. Summary statistics based on 1000 evaluations of Moran's $I(d, 0.5)$ in Simulation A.

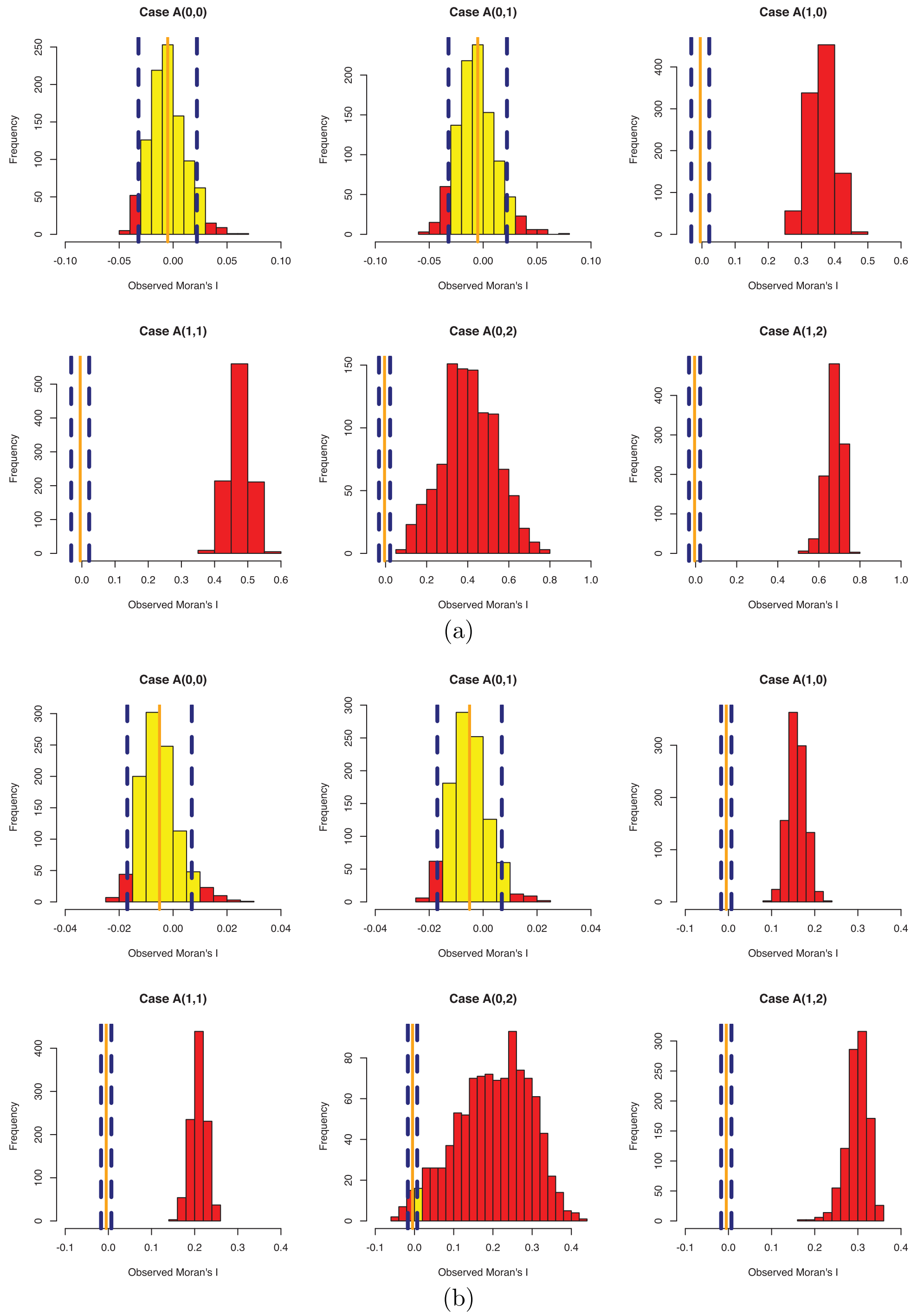
			[c]Simulation cases					
	[a]Marginal (under H_0)	[b]Conditional on **s**, *t* (under H_0)	A(0,0)	A(0,1)	A(1,0)	A(1,1)	A(0,2)	A(1,2)
$d = 0.3$								
			$n = 50$					
[d]Mean	−0.02041	−0.02041	−0.02197	−0.01955	0.29727	0.39553	0.42298	0.66373
[d]Std. dev	0.06305	0.05884	0.07055	0.06750	0.08820	0.15773	0.06826	0.04524
			$n = 200$					
Mean	−0.00503	−0.00503	−0.00499	−0.00603	0.36072	0.47351	0.40885	0.67432
Std. dev	0.01580	0.01383	0.01686	0.01806	0.03796	0.03088	0.13111	0.03778
			$n = 400$					
Mean	−0.00251	−0.00251	−0.00239	−0.00227	0.36950	0.49771	0.43506	0.70559
Std. dev	0.006266	0.00703	0.00871	0.01053	0.02738	0.02323	0.13392	0.03748
$d = 0.6$								
			$n = 50$					
Mean	−0.02041	−0.02041	−0.02098	−0.02074	0.09681	0.14235	0.18101	0.23494
Std. dev	0.02760	0.02407	0.03101	0.02695	0.04048	0.03317	0.10851	0.03335
$n = 200$								
Mean	−0.00503	−0.00503	−0.00509	−0.00519	0.15839	0.20919	0.19987	0.29862
Std. dev	0.00690	0.00610	0.00719	0.00700	0.02067	0.01752	0.09435	0.02558
			$n = 400$					
Mean	−0.00251	−0.00251	−0.00251	−0.00248	0.16854	0.22721	0.22819	0.32290
Std. dev	0.002742	0.00338	0.00374	0.00420	0.01499	0.01405	0.10231	0.02770
$d = 1.0$								
			$n = 50$					
Mean	−0.02041	−0.02041	−0.01998	−0.02132	−0.00710	−0.00151	0.06351	0.01585
Std. dev	0.01518	0.01370	0.01597	0.01317	0.01571	0.01205	0.08851	0.02019
			$n = 200$					
Mean	−0.00503	−0.00503	−0.00477	−0.00497	0.00330	0.00566	0.04929	0.01691
Std. dev	0.00381	0.00360	0.00441	0.00401	0.00337	0.00271	0.04258	0.01167
			$n = 400$					
Mean	−0.00251	−0.00251	−0.00259	−0.00243	0.01111	0.01558	0.06387	0.03032
Std. dev	0.001511	0.00192	0.00212	0.00219	0.00253	0.00228	0.04491	0.01244

[a] $E(I \mid H_0) = -1/(n-1)$ and $\mathrm{Var}\{I \mid (\mathbf{s}_i, t_i)'s\}$.

[b] $E(I \mid H_0) = -1/(n-1)$ and the sample mean of 1000 evaluations of $\mathrm{Var}\{I \mid (\mathbf{s}_i, t_i)'s\}$.

[c] Simulation settings: Case A(a,b) with $a = 0{,}1$ for Types 0,1 of $\boldsymbol{\mu}$ and $b = 0{,}1{,}2$ for Types 0,1,2 of Σ.

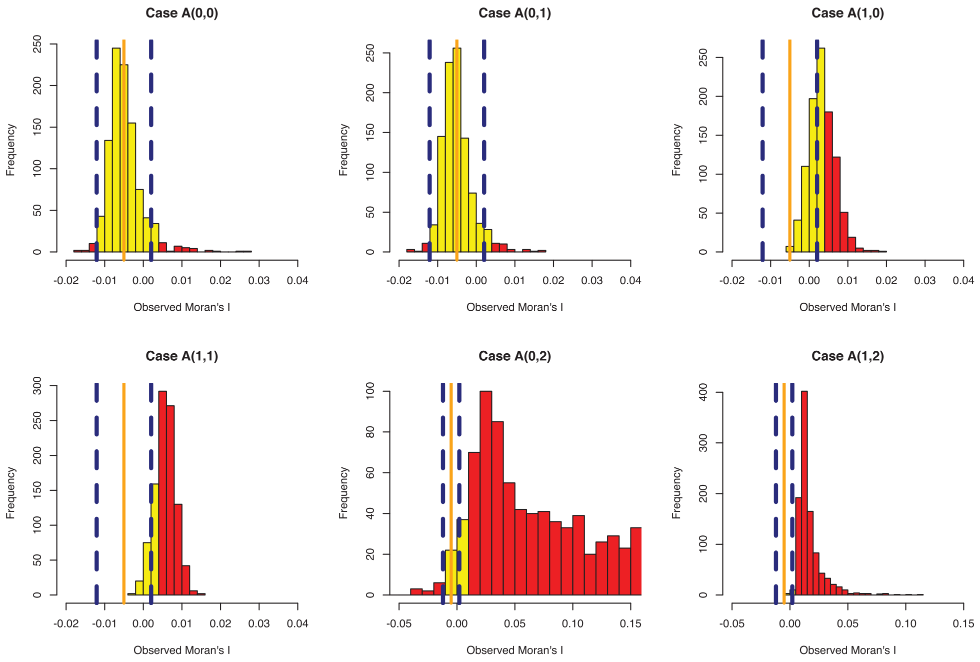
[d] Sample mean and sample standard deviation.



(a)

(b)

Figure 1. Histogram of Moran's $I(d, \tau)$ with $d = 0.3,0.6,1.0$ and $\tau = 0.5$ when $n = 200$. The solid line represents the mean under the null hypothesis. The dashed lines define 95% acceptance region. The bars in the acceptance region are shaded lightly; in rejection region, darkly. (a) Histogram of $I(0.3, 0.5)$, (b)Histogram of $I(0.6, 0.5)$ and (c) Histogram of $I(1.0, 0.5)$.



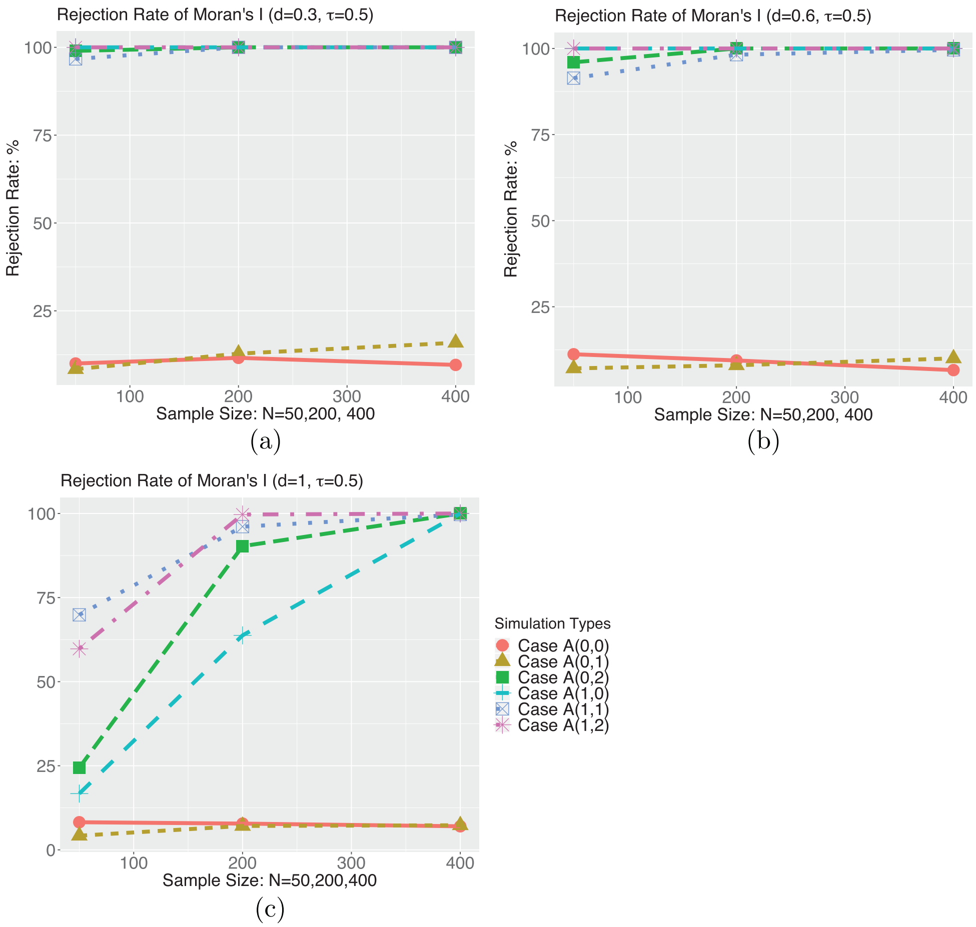
(c)

Figure 1. Continued.

deviations of $I(d,\tau)$ in Case A(0,1) are larger, the sample means are similar to those from Case A(0,0). This confirms that the Moran's I test for correlation can be rather robust to varying variances of the individual observations. The results from Cases A(1,0) and A(1,1), however, verify that the Moran's I test for correlation may be violated as a result of heterogeneity in the expectations of the observations. Regardless of the independent observations, the test likely rejects H_0 with a commonly used type-I error rate in each of the settings of Cases A(1,0) and A(1,1). On the other hand, this suggests a new application of the Moran's I test: model checking with the residuals in regression analysis.

The simulation results for Cases A(0,2) and A(1,2) are also as expected. They are illustrated by the histograms of the Moran's I evaluations in the settings with sample size $n = 200$ in Figure 1. Each histogram includes the evaluations of the rejection region with the type-I error rate 0.05 given by Equation (2) for its lower and upper limits.

Figure 2 presents the empirical rejection rates of the Moran's I test for all simulation settings. The rejection rates with $d = 1.0$ in Cases A(1,0), A(1,1), A(0,2), and A(1,2) are not as high as those with $d = 0.3$ or 0.6 when the sample sizes are not large (i.e. $n = 50$ and 200). This is because the Moran's I is a weighted average over the global surface. With larger neighbourhoods, individual contributions can be averaged out, making the test less efficient. For comparison, we also evaluated the Moran's I with the Dubé–Legros weights, $I^*(d,\tau;\gamma,\alpha)$ with $\gamma = 0.5, \alpha = 0.5$. Based on Equation (1), we used the standardised form $\|\mathbf{s}_i - \mathbf{s}_j\|/\sqrt{2}$ instead of $\|\mathbf{s}_i - \mathbf{s}_j\|$ to limit the distance between any pair of individuals to the range $[0,1]$ so that it is comparable to the time difference.



(a)

(b)

(c)

Figure 2. Rejection rate of Moran's $I(d, \tau)$ with $d = 0.3, 0.6, 1.0$, $\tau = 0.5$, sample size $n = 50, 200, 400$. (a) Rejection rate of $I(0.3, 0.5)$, (b)Rejection rate of $I(0.6, 0.5)$ and (c) Rejection rate of $I(1.0, 0.5)$.

We also examined the histograms of the evaluations of the two Moran's I statistics $I(d, \tau)$ and $I^*(d, \tau)$ in Case A(1,2). $I^*(d, \tau)$ appears to have a larger variation than $I(d, \tau)$. The two types of weights yield results that are consistent with each other.

In summary, this simulation study confirms that the magnitude of the Moran's I depends on the heterogeneity in the expectations of the observations and on the underlying spatio-temporal correlation.

3.2. Simulation B: Moran's I in analysis of residuals

To investigate the test procedure applied to regression residuals, we simulated data in the following way. The outcome Y was generated as $Y = f(\mathbf{x}) + \epsilon$, where ϵ is the random error, $\mathbf{x} = (\mathbf{s}, t, w)$, and

$$f(\mathbf{x}) = \begin{cases} sin(2s_1^2) - \sin(2s_2) - 0.5w, & \text{if } 0 < t < 0.5 \\ sin(2s_1^2) + \sin(2s_2)t - 0.5w, & \text{if } 0.5 \leq t < 1, \end{cases} \tag{5}$$

where $\mathbf{s}$ is the location index with the two components s_1 and s_2, and t and w are the time and additional predictor, respectively. Observations on Y together with $\mathbf{x}$ were generated as follows:

> Step 1. Generate the n locations $\mathbf{s}_i$ and times t_i as for Simulation A. In addition, independently generate observations on the additional predictor $w_i \sim N(0,1)$.
> Step 2. Conditional on the generated $\{(\mathbf{s}_i, t_i) : i = 1, \ldots, n\}$, generate the n-dimensional vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ from the multivariate normal distribution $N(\mathbf{0}, \Sigma)$. Here $\mathbf{0}$ is the zero vector and Σ is either Type 0: $\Sigma = 0.5^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix; or Type 1: $\Sigma = (\sigma_{ij})_{n\times n}$, where $\sigma_{ij} = 0.5^2 \exp\{-0.1(\|\mathbf{s}_i - \mathbf{s}_j\| + |t_i - t_j|)\}$ for $i, j = 1, \ldots, n$.
> Step 3. Obtain the response y_i's as $f(\mathbf{x}_i) + \epsilon_i$, where $f(\mathbf{x}_i)$ is the function in Equation (5) evaluated at $(\mathbf{s}_i, t_i, w_i)$ generated by Step 1.

We conducted regression analyses with the observations generated above under various models in the form $Y_i = g(\mathbf{x}_i) + \zeta_i$, where $\zeta_i \sim N(0, \sigma^2)$ independently. Specifically, we took $g(\cdot)$ to be one of the following functions:

Type 1. Ordinary Linear Regression Model

$$g(\mathbf{x}) = \beta_0 + \beta_1 w + \beta_2 t + \beta_3 s_1 + \beta_4 s_2; \tag{6}$$

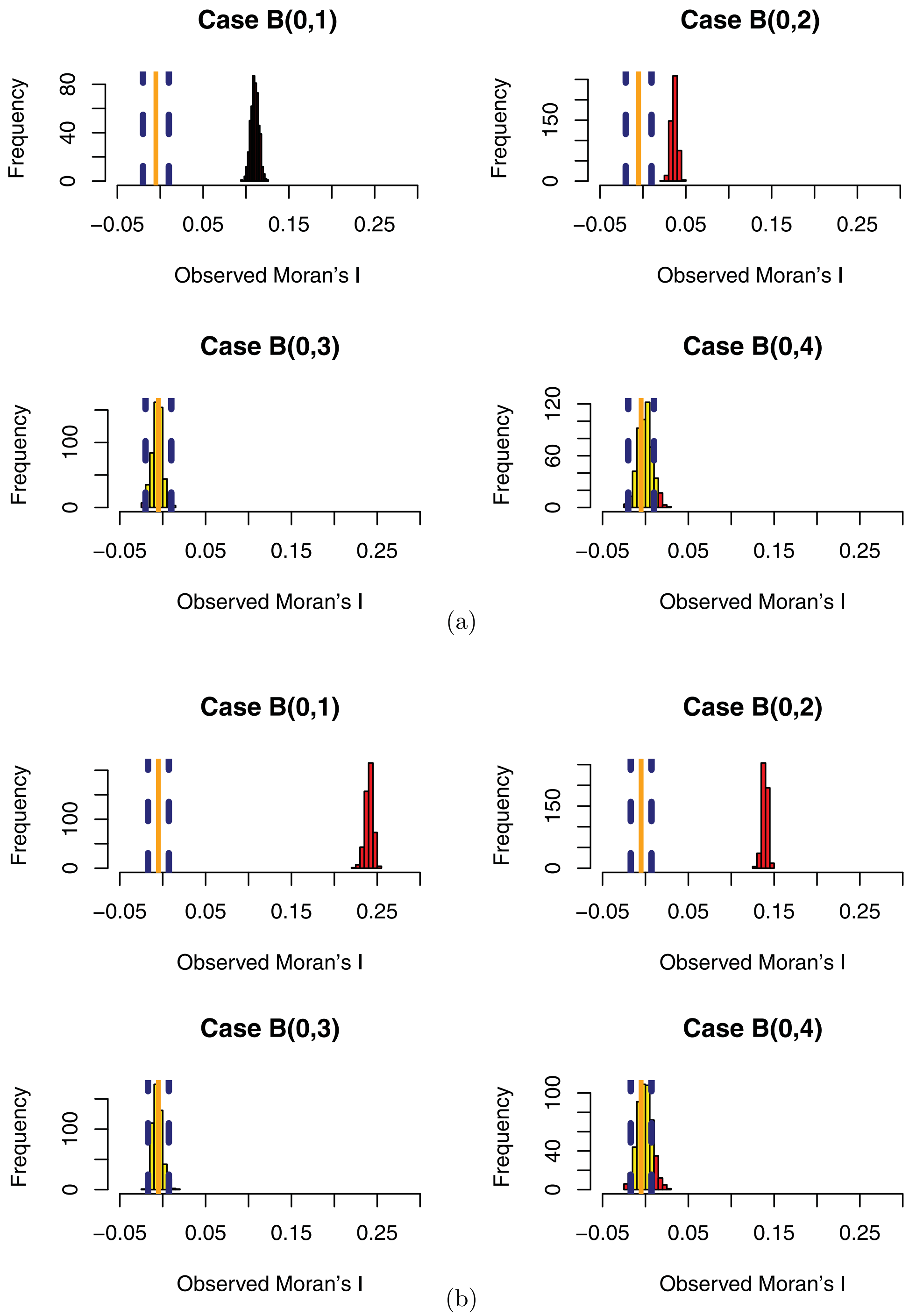
Type 2. General Linear Regression Model

$$g(\mathbf{x}) = \beta_0 + \beta_1 w + \beta_2 t + \beta_3 s_1 + \beta_4 s_2 + \beta_5 s_1^2 + \beta_6 s_2^2; \tag{7}$$

Table 2. Summary statistics based on 200 evaluations of Moran's $I(0.6, 0.5)$ and $I^*(0.6, 0.5)$ in Simulation B.

			[c]Simulation cases							
			(independent random errors)				(correlated random errors)			
	[a]Marginal (under H_0)	[b]Conditional on $\mathbf{s}, t$ (under H_0)	B(0,1)	B(0,2)	B(0,3)	B(0,4)	B(1,1)	B(1,2)	B(1,3)	B(1,4)
$n = 200$										
					$I(0.6, 0.5)$					
[d]Mean	−0.00503	−0.00503	0.24041	0.13889	−0.00638	−0.00111	0.28117	0.19205	0.03125	0.05350
[d]Std. dev	0.00690	0.00619	0.00455	0.00327	0.00506	0.00631	0.09210	0.05516	0.00592	0.01030
					$I^*(0.6, 0.5)$					
Mean	−0.00503	−0.00503	0.11008	0.03632	−0.00656	0.00639	0.28389	0.19274	0.09381	0.11113
Std. dev	0.00832	0.00767	0.00496	0.00405	0.00580	0.00644	0.09025	0.05087	0.00700	0.00960
$n = 400$										
					$I(0.6, 0.5)$					
Mean	−0.00251	−0.00251	0.29364	0.13808	−0.00107	−0.00297	0.26938	0.13360	0.01401	0.03994
Std. dev	0.00344	0.00343	0.00305	0.00200	0.00211	0.00454	0.09835	0.03732	0.00248	0.00624
					$I^*(0.6, 0.5)$					
Mean	−0.00251	−0.00251	0.14449	0.02997	−0.00056	−0.00539	0.27355	0.13952	0.04240	0.05155
Std. dev	0.00416	0.00414	0.00333	0.00264	0.00324	0.00411	0.09775	0.03879	0.00370	0.00651

[a] $E(I \mid H_0) = -1/(n-1)$ and $V\{I \mid (\mathbf{s}_i, t_i)'s\}$.
[b] $E(I \mid H_0) = -1/(n-1)$ and the sample mean of 1000 evaluations of $V\{I \mid (\mathbf{s}_i, t_i)'s\}$.
[c] Simulation settings: Case B(a,b) with $a = 0{,}1$ for Σ Types 0, 1 and $b = 1{,}2{,}3{,}4$ for Types 1,2,3,4 of the specified $g(\cdot)$.
[d] Sample mean and sample standard deviation.



(a)

(b)

Figure 3. Histograms of $I(d,\tau)$ and $I^*(d,\tau)$ with independent data, $n = 200$. (a) $N = 200$, $I(d,\tau)$ with independent data and (b) $N = 200$, $I^*(d,\tau)$ with independent data.

Type 3. Partial Linear Regression Model

$$g(\mathbf{x}) = h(\mathbf{s}, t) + \beta_1 w \text{ with unspecified function } h(\cdot); \tag{8}$$

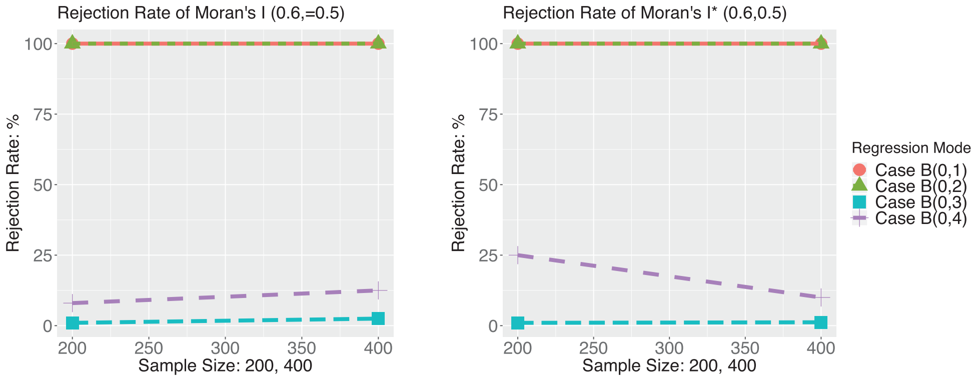
Type 4. Nonparametric Regression Model

$$g(\mathbf{x}) = h(\mathbf{s}, t, w) \text{ with unspecified function } h(\cdot). \tag{9}$$

Under each of the above models, residuals were calculated via $e_i = y_i - \hat{g}(\mathbf{x}_i)$ for $i = 1, \ldots, n$, where $\hat{g}(\cdot)$ is the fitted model in the regression analysis with the generated data $(y_i, \mathbf{x}_i)$. Here, the parameters in the first two models (linear models) were estimated by least-squares estimation (LSE) procedures. Estimates of the partial linear regression models were obtained by LSE integrated with local linear estimation. For nonparametric regression, the unspecified functions were estimated by kernel smoothing methods using the Gaussian kernels, as implemented in the `R` package `np` (Hayfield and Racine 2008). We used the GCV (generalised cross validation) approach to choose the bandwidths in the estimation. We evaluated the Moran's I with two sets of weights, $I(d, \tau)$ with the original weights and $I^*(d, \tau; \gamma, \alpha)$ with the Dubé–Legros weights given in Equation (1), and the simulated residuals. We set $\gamma = 0.5, \alpha = 0.5$ to control the Dubé–Legros weights and $\|\mathbf{s}_i - \mathbf{s}_j\|$ is standardised as $\|\mathbf{s}_i - \mathbf{s}_j\|/\sqrt{2}$.

Table 2 presents the sample means and standard deviations of $I(0.6, 0.5)$ and $I^*(0.6, 0.5)$ based on 500 repetitions in the simulation settings Cases B(a,b), where a = 0, 1 for $\text{Var}(\boldsymbol{\epsilon}) = \Sigma$ specified by types 0, 1 and b = 1, 2, 3, 4 for $g(\cdot)$ given in Equations (6)–(9) specified by types 1–4, respectively. The table also presents the conditional and marginal means and standard deviations of the two Moran's I statistics, calculated in the same manner as those presented in Table 2.

The histograms of the evaluations of $I(d, \tau)$ and $I^*(d, \tau)$ with the regression residuals when $n = 200$ for the cases with correlated random errors (i.e. Cases B(1,b) for $b = 1,2,3,4$) indicate that all the tests reject H_0 with a type-I error rate of 0.05, regardless of the regression model. Figure 3 presents histograms of the evaluations of $I(d, \tau)$ and $I^*(d, \tau)$ with



(a) (b)

Figure 4. Empirical rejection rates of $I(0.6, 0.5)$ and $I^*(0.6, 0.5)$ based on regression residuals under models (6), (7), (8), and (9) with independent random errors, sample size $n = 200{,}400$. (a) Rejection rate of $I(0.6, 0.5)$ and (b) Rejection rate of $I^*(0.6, 0.5)$.

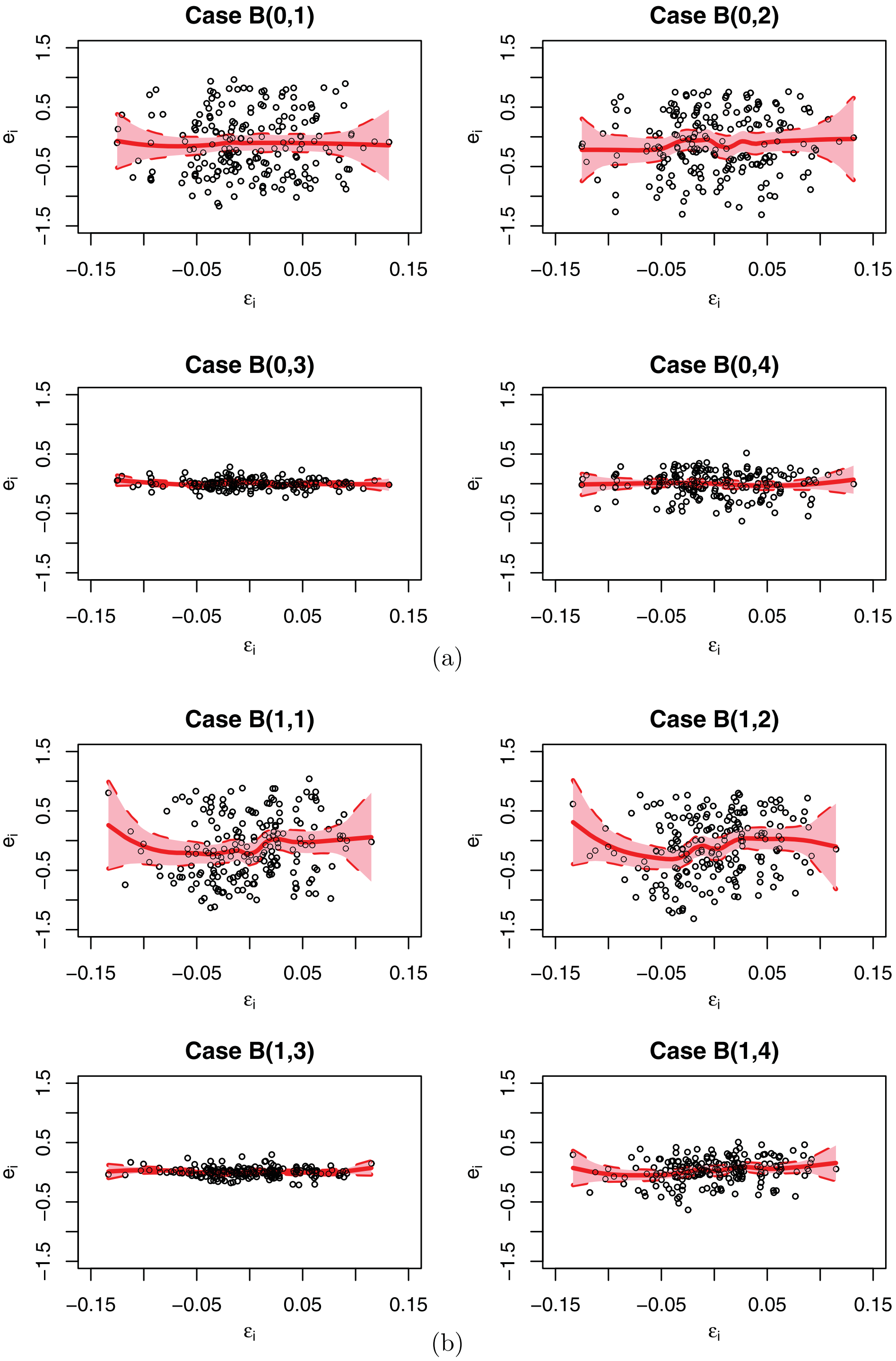
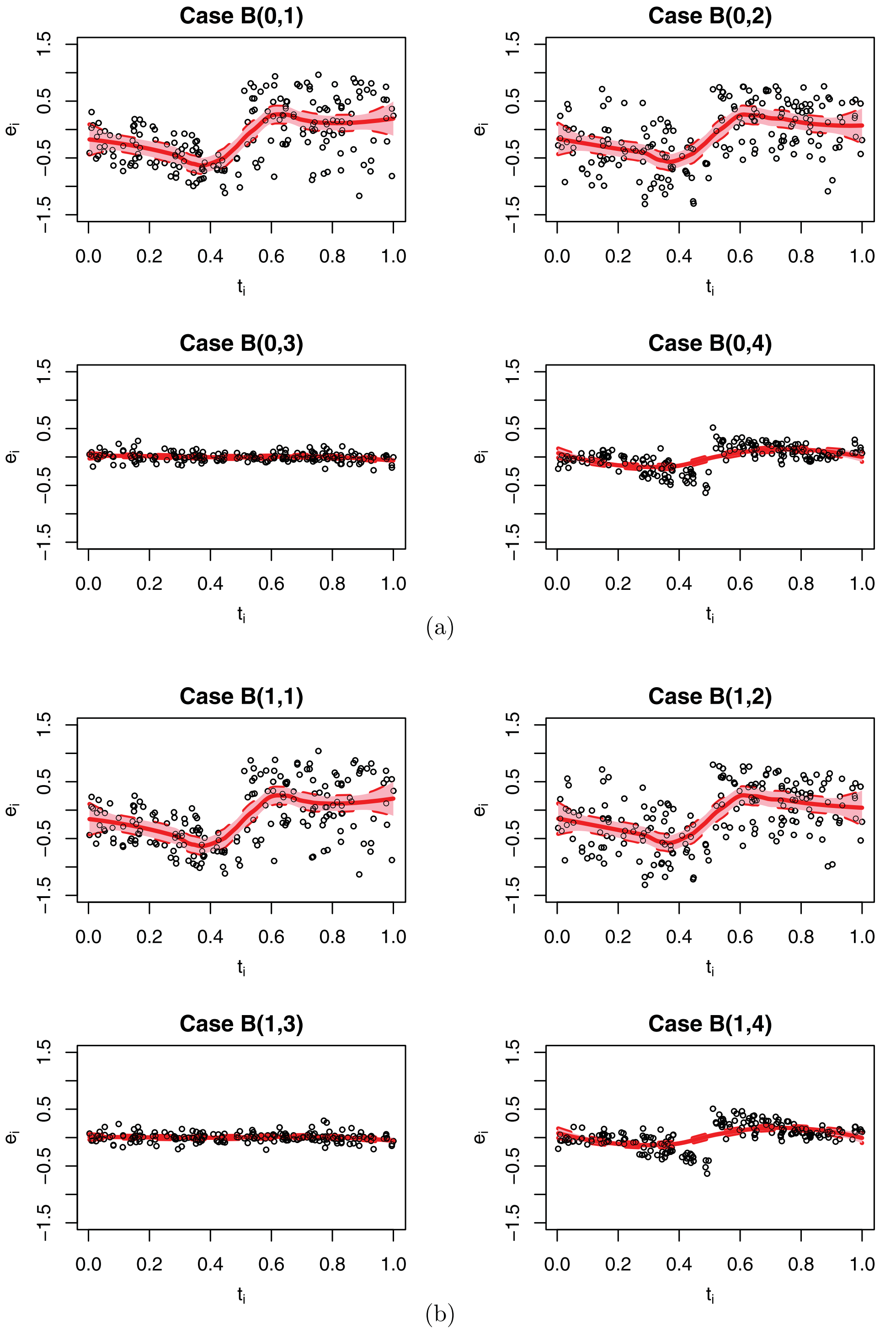


Figure 5. Scatterplots of residuals vs. generated random errors, sample size $n = 200$. (a) with independent data and (b) with correlated data.



(a)

(b)

Figure 6. Scatterplots of residuals vs. times in the generated observations, sample size $n = 200$. (a) with independent data and (b) with correlated data.

the regression residuals when $n = 200$ for the cases with independent random errors. We see that, as expected, the variation associated with $I(d, \tau)$ is generally smaller than that associated with $I^*(d, \tau)$. However, the tests based on the two Moran's I statistics reject H_0 for Cases B(0,1) and B(0,2), where the random errors are i.i.d. and the residuals are under misspecified regression models.

Figure 4 shows the empirical rejection rates with the two testing procedures for Cases B(0,1) and B(0,2). This confirms the suggestion in Section 2.3 that the testing procedure based on the Moran's I can be used as a tool for checking regression models. Case B(0,3) simulates the situation of the null hypothesis H_0. The empirical rates of H_0 rejection for both testing procedures are around the nominal type-I error rate of 0.05. However, the empirical rates of rejection for Case B(0,4) associated with the two procedures based on $I(d, \tau)$ and $I^*(d, \tau)$ indicate uncertain outcomes. This is likely because a satisfactory fitting by nonparametric regression usually requires a large collection of observations.

The commonly used residual plots can show in detail the performance of the testing procedures in the simulated regression analyses. As an example, Figure 5 presents the scatter plots of the residuals e_i vs. the random errors ϵ_i for one generated data set ($n = 200$) together with the local regression curves produced by the R function LOESS and confidence intervals of approximately 95%. We see that the residuals from each regression are near zero but apparently differ from the corresponding random errors ϵ_i. The differences are especially large in the cases with misspecified regression models for both independent and correlated observations. The scatter plots of the residuals e_i vs. the generated times t_i displayed in Figure 6 show that the regression analyses with the two misspecified models can not adequately capture the generated temporal correlation. Those for nonparametric regression reveal an unsatisfactory model fit near $t = 0.5$, where the underlying regression function changes unsmoothly. This may explain the behaviour of the two tests in the associated cases.

The Moran's I test with the regression residuals generally performs well in detecting model misspecification with the functional form of the mean function and with the spatio-temporal pattern underlying the random errors. In particular, we conclude that the regression model is appropriate for the current data if no significant evidence exists that counters the null hypothesis based on the Moran's I test with the regression residuals.

4. Example with real data

Wildfire records constitute typical spatio-temporal data. Many researchers have studied the relationship between wildfires and associated ecological factors together with the spatial and temporal characteristics of the fires (Martell and Sun 2008; Podur 2001). The sizes of fires adjacent to each other in time and/or location are likely correlated. Xiong (2015) explores the spatio-temporal correlation of wildfires via regression analysis using the Moran's I test with regression residuals as a tool to diagnose models. The difficulty she encountered when interpreting the test outcomes was a motivation for our research. To illustrate our findings about the usefulness of the Moran's I, we analyse the records from 10 wildfire-management areas of Alberta, Canada of the 746 lightning-caused wildfires in 2006.

4.1. Data description and regression modelling

Figure 7(a) summarises the fire sizes based on the information associated with the variable '*ex_hectares*' in the data, which records the area (in hectares) that each fire has burnt when it is extinguished. Because its distribution appears quite skewed, we took a base-10-log transformation of the variable, which is referred to as the fire size in the rest of this paper. Figure 7(b) shows the wildfires across Alberta in the months of the fire season: May to September. Most large fires occurred in the High Level and Slave Lake areas, which are in the northeastern part of the province. The size distribution varies; most large fires occurred in June and July.

We considered a general regression model with the log-transformed fire size as the response variable:

$$Y_i = \mu(\mathbf{s}_i, t_i; \mathbf{z}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{10}$$

where Y_i is the log-transformed *burn area* (i.e. wildfire size) of wildfire i, $\mathbf{s}_i$ is the *location* (i.e. the vector of latitude and longitude) of wildfire i, t_i is the *starting time* of wildfire i, $\mathbf{z}_i$ is the additional vector of environmental exposures, and ϵ_i is the random error. Here $\mu(\mathbf{s}_i, t_i; \mathbf{z}_i) = E[Y_i \mid \mathbf{s}_i, t_i, \mathbf{z}_i]$, and the random errors are assumed to be i.i.d. with $E[\epsilon_i \mid \mathbf{s}_i, t_i, \mathbf{z}_i] = 0$ and $\mathrm{Var}[\epsilon_i \mid \mathbf{s}_i, t_i, \mathbf{z}_i] = \sigma^2$. Following the Canadian Forest Fire Weather Index System, the variables *daily temperature*, *relative humidity*, and *wind speed* were used as explanatory variables in addition to the fire location and time.

We conducted analyses under the following two specifications of Equation (10) (one is the ordinary linear regression model and the other is a partial linear regression model) for $i = 1, \ldots, 746$:

$$\mu_i(\mathbf{s}_i, t_i; \mathbf{z}_i) = \beta_0 + \boldsymbol{\alpha}'\mathbf{s}_i + \gamma_{h(t_i)} + \boldsymbol{\beta}'\mathbf{z}_i, \tag{11a}$$

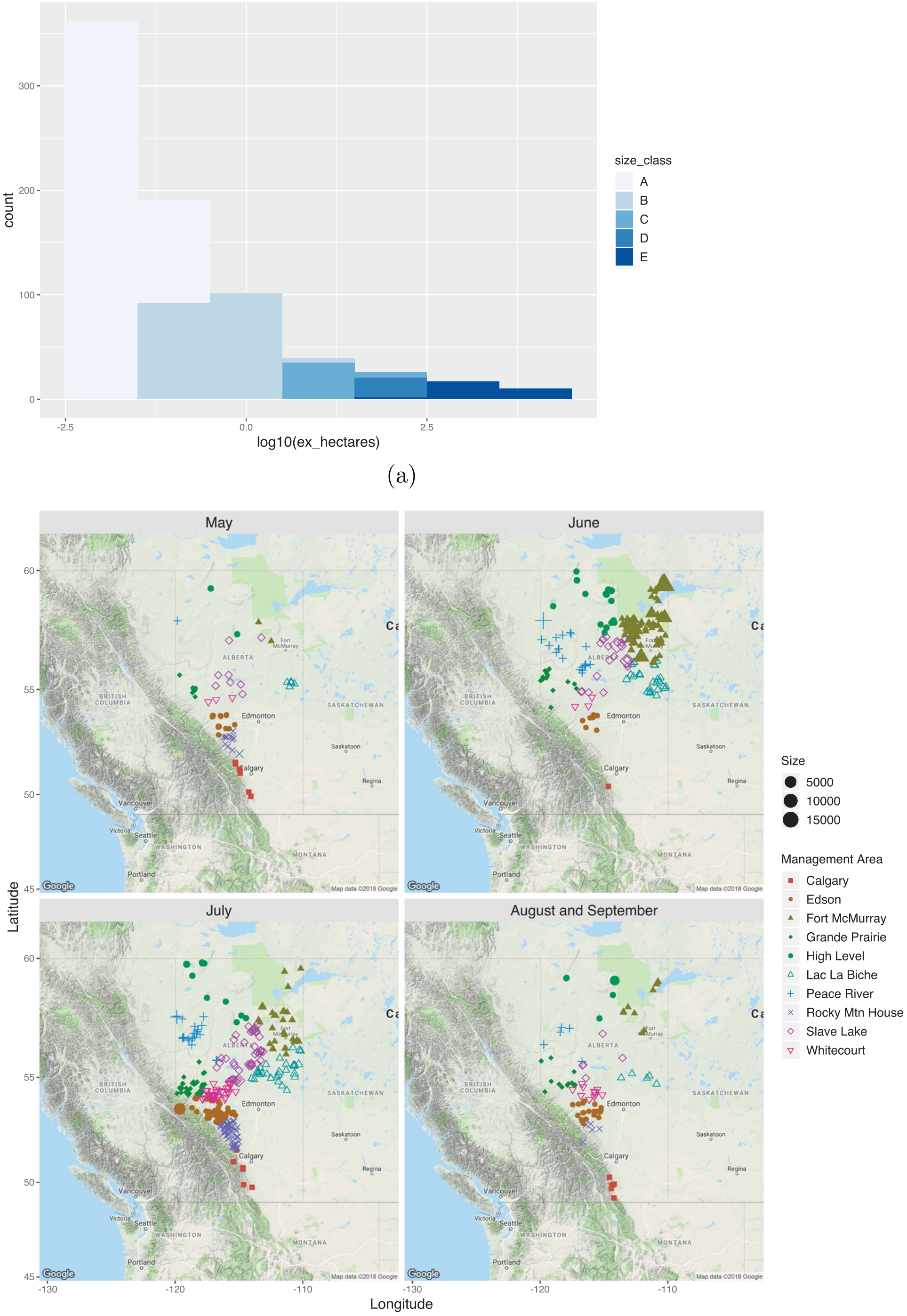
$$\mu_i(\mathbf{s}_i, t_i; \mathbf{z}_i) = g_{h(t_i)}(\mathbf{s}_i) + \boldsymbol{\beta}'\mathbf{z}_i, \tag{11b}$$

where $h(t_i)$ is a factor with four levels for May, June, July, and August/September, and $g_j(\cdot)$ is an unspecified function corresponding to fire month j.

4.2. Analysis outcome and model diagnosis

The parameters in model (11a) were estimated by LSE, and the unspecified functions and regression parameters in model (11b) were estimated by kernel smoothing integrated with LSE. The parameter estimates under both models (11a) and (11b) are listed in Table 3. The variables *relative humidity* and *wind speed* were identified as significant predictors for fire size by both analyses, but the effect of the third environmental variable *temperature* did not appear significant in either analysis.

Figure 8 presents the contours of the estimated spatio-temporal surfaces $\hat{g}_j(\mathbf{s})$ under model (11b). We also checked the corresponding perspective plots of $\hat{g}_j(\mathbf{s})$ together with the corresponding planes under the ordinary linear regression model (11a) $\hat{\beta}_0 + \hat{\boldsymbol{\alpha}}'\mathbf{s}_i + \hat{\gamma}_{h(t_i)}$. The summarised spatio-temporal patterns under the two models are similar in the central area of Alberta but differ in the northeastern and southwestern areas. The plots indicate graphically that the partial linear regression model is more appropriate for the wildfire data than the ordinary linear model.



(a)

(b)

Figure 7. Distribution of sizes and occurrences: Alberta wildfires in 2006. (a) Distribution of wildfire sizes and (b) distribution of wildfire occurrences.

Table 3. Regression parameter estimates for the records of Alberta wildfires in 2006.

	Estimate	[a]St. err	[b]Test statistic	[b]p-value
Under Model 10(a)				
(Intercept)	−1.02618	0.31416	−3.26643	.00114
relative_humidity	−0.00872	0.00268	−3.35450	.00084
wind_speed	0.04711	0.00596	8.24128	$<.0001$
temperature	0.01173	0.00970	1.24153	.21480
longitude (standardised)	0.11433	0.04660	2.45328	.01439
latitude (standardised)	0.31688	0.04988	6.35291	$<.0001$
June vs. May	0.01513	0.17853	0.08475	.93249
July vs. May	−0.01145	0.15954	−0.07174	.94283
AugSep vs. May	−0.09221	0.18941	−0.48683	.62652
Under Model 10(b)				
relative_humidity	−0.00795	0.00260	−2.96443	.00152
wind_speed	0.05153	0.00572	8.63991	$<.0001$
temperature	0.00775	0.00945	0.79879	.21221

[a] Estimated standard errors.
[b] Ratio of the estimate of β to the estimated standard error, and p-value of the t-test for H_0: $\beta = 0$.

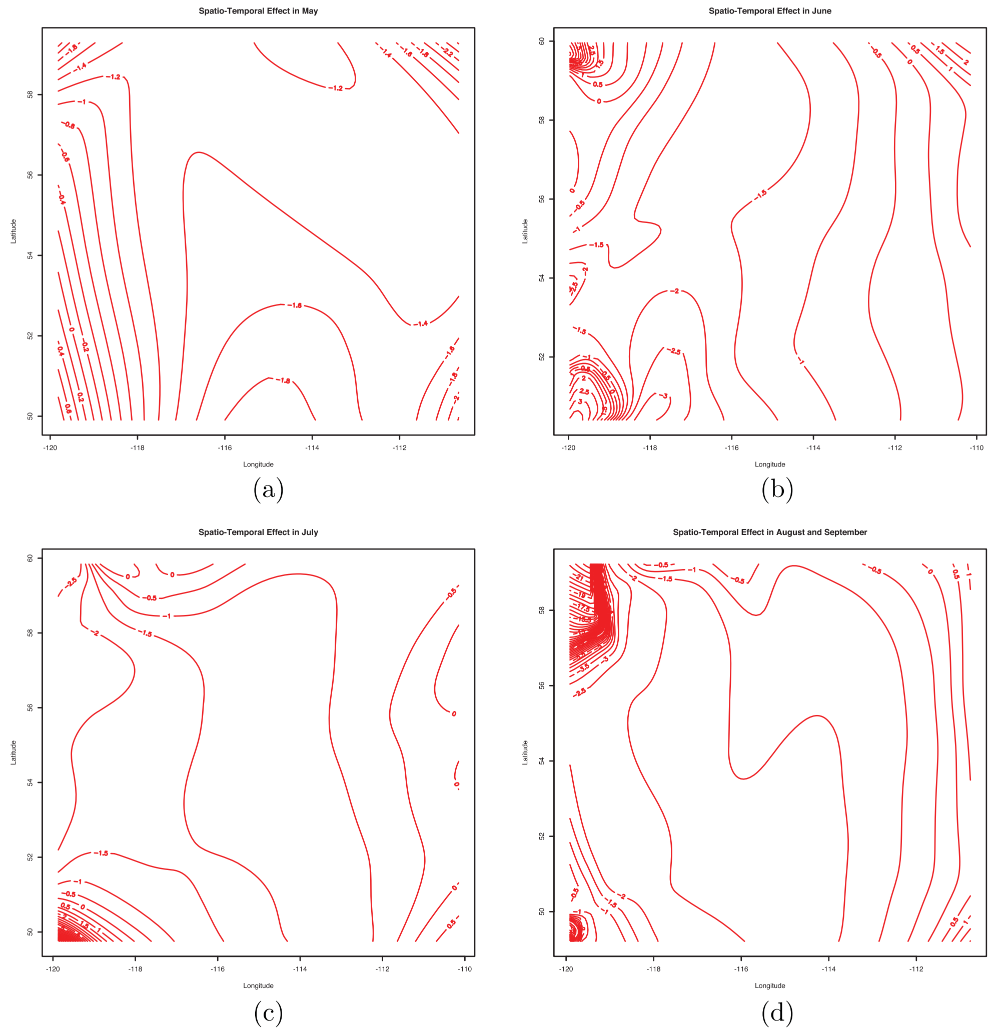
We evaluated the Moran's I statistic with four different sets of weights with the regression residuals under models (11a) and (11b). The first three sets are the original weights corresponding to the three definitions of neighbourhood: two fires are neighbours if the time lag between them is bounded by a predetermined τ and (i) they are from the same fire management area, (ii) they are detected by the same fire station, or (iii) the distance between their locations is bounded by a predetermined d.

Figures 9 and 10 show the three Moran's I statistics with the regression residuals under models (11b) and (11a). We used different combinations of d and τ in Figure 9(a) with type (i) weights; in Figure 9(b), type (ii) weights; in Figure 10, type (iii) weights. In each plot, we also include the nominal value of the Moran's I under H_0 of i.i.d. random errors with mean zero and the rejection region with a type-I error rate of .05. We also considered a fourth type of weights (type (iv)), the Dubé–Legros weights given in Equation (1). The plots of the evaluations of the Moran's I with the weights are similar to those in Figure 10.

The conclusions based on the Moran's I tests with all four sets of weights for model checking are consistent with each other. The partial linear regression model (11b) appears quite appropriate for the wildfire data whereas the ordinary linear regression model (11a) does not. This indicates that a nonlinear spatio-temporal pattern exists in the wildfire sizes.

5. Concluding remarks

This paper explores analytically and numerically the Moran's I statistic with the weights originally proposed by Moran (1950) and the Dubé–Legros weights. In particular, we find that the test procedure based on the Moran's I with regression residuals can be a nonparametric tool for diagnosing regression models, including both the functional form of the mean function and the assumption regarding random errors. When the goal is to detect the underlying spatio-temporal correlation, the test is rather robust to varying variance among the individual units but can be misled by heterogeneity in the means. On the other hand, when checking for a regression model, the test with the regression residuals can be rather sensitive to the bias caused by either the procedure for estimating the regression parameter



(a)

(b)

(c)

(d)

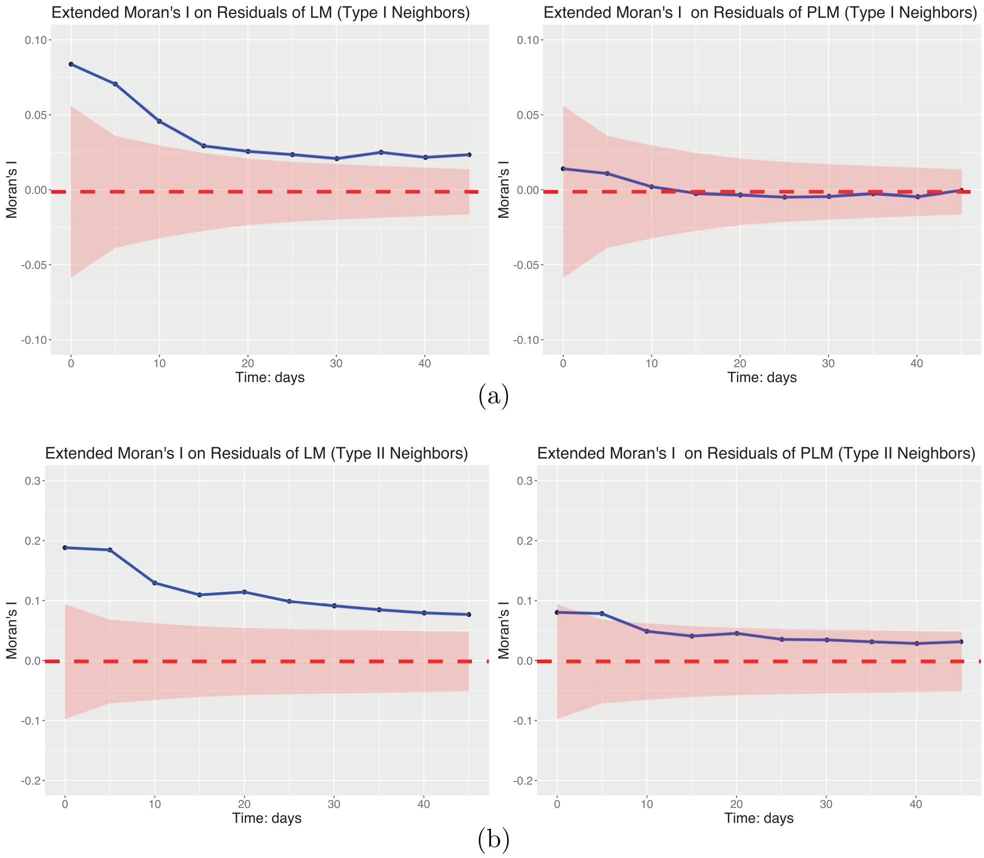
Figure 8. Contour plots for spatio-temporal characteristics of fire size.

or the underlying correlation in the random errors. The usefulness of these general findings is illustrated by a regression analysis of real wildfire data. This novel application of the Moran's *I* provides an inferential procedure for model checking with regression residuals. We could potentially extend this to checking for regression models with respect to general explanatory variables if the neighbourhood used to determine the weights is defined accordingly.

The simulation results show that the Moran's *I* with different weights may cause the test to draw different conclusions. For example, the original weights with a larger upper bound for the location distance for the neighbourhood may have a lower power against the null hypothesis. Other types of weights are worth exploring. One choice is the type of weights account for population density (Oden 1995). Another consideration is to choose a different measure of distance between two locations in the Dubé–Legros weights, not necessarily the Euclidean distance, to accommodate certain geographic environment. The development of a systematic procedure for choosing an appropriate set of weights is an interesting but challenging problem.

We plan to explore different ways to implement the Moran's I based testing procedure. One may consider, for example, to implement it via randomly permute the observations Z_i's to the associated times and locations. As suggested by one of the referees, we can also conduct the test via a bootstrap procedure. By either of the two approaches, one may easily obtain estimates for the mean and variance of the Moran's I under the current population. Some preliminary analysis of the real data by the bootstrap approach agrees to what is presented on the model checking in this paper.

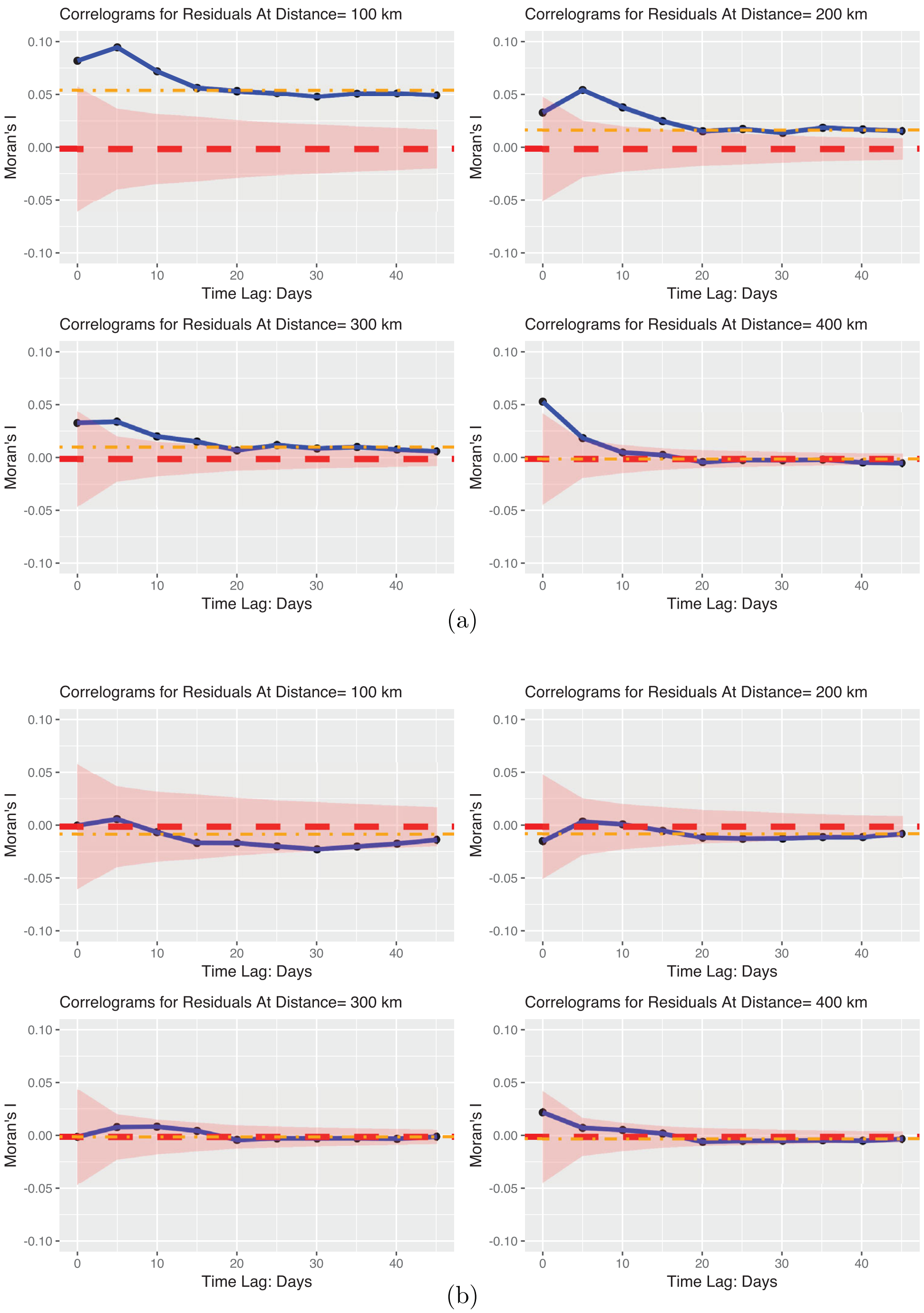
Several other investigations would be worthwhile. We may follow Li et al. (2007) and explore the Moran's I with a collection of observations that are generated from a particular stochastic process or random field. Another avenue of research is to follow the work reported in Zhang and Lin (2016) to examine the asymptotic distributions of the Moran's I



(a)

(b)

Figure 9. Moran's I with residuals from the ordinary/partial linear regression analysis. The solid lines are the observed Moran's I values; the dashed lines are $E[I \mid H_0]$; and the shaded areas are the 95% acceptance regions. (a) Moran's I with type (i) weights: Neighbors are from the same wildfire management areas with time lags $\leq \tau$ and (b) Moran's I with type (ii) weights: Neighbors are detected by the same wildfire watch stations with time lags $\leq \tau$.



(a)

(b)

Figure 10. Moran's $I(d,\tau)$ with type (iii) weights. The solid lines are the observed Moran's I values; the dashed lines are $E[I \mid H_0]$; and the shaded areas are the 95% acceptance regions. (a) Moran's I with residuals from the ordinary linear regression analysis (b) Moran's I with residuals from the partially linear regression analysis.

under various types of heterogeneity in the situations with the normally distributed observations or conditional on the current set of observations. This may help us further assess performance of the Moran's I test procedure. A third possibility is to extend the Moran's I to accommodate data with a complex structure, such as censored data. Last but not least, we plan to develop a software package to make this research accessible to practitioners.

Acknowledgements

The authors acknowledge the funding support from the Canadian Statistical Sciences Institute (CANSSI). We are grateful to Neal McLoughlin of Alberta Wildfire Management Branch for providing the wildfire data. We thank the two referees and an AE for their constructive comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is partly supported by Natural Sciences and Engineering Research Council of Canada [NSERC 038317] (for Bingham), [NSERC 116860] (for NSERC), and [NSERC DG and DAS 177430] (for Hu).

References

Assunção, R.M., and Reis, E.A. (1999), 'A New Proposal to Adjust Moran's *I* for Population Density', *Statistics in Medicine*, 18, 2147–2162.

Cliff, A.D., and Ord, J.K. (1981), *Spatial Processes: Models and Applications* (Vol. 44), London: Pion.

Dubé, J., and Legros, D. (2012), 'A Spatio-Temporal Measure of Spatial Dependence: An Example using Real Estate Data', *Papers in Regional Science*, 92, 19–30.

Geary, R.C. (1954), 'The Contiguity Ratio and Statistical Mapping', *The Incorporated Statistician*, 5, 115–145.

Hayfield, T., and Racine, J.S. (2008), 'Nonparametric Econometrics: The np Package', *Journal of Statistical Software*, 27(5), 1–32.

Helbich, M., Leitner, M., and Kapusta, N.D. (2012), 'Geospatial Examination of Lithium in Drinking Water and Suicide Mortality', *International Journal of Health Geographics*, 11, 19.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., and Daszak, P. (2008), 'Global Trends in Emerging Infectious Diseases', *Nature*, 451, 990–993.

Li, H., Calder, C.A., and Cressie, N. (2007), 'Beyond Moran's *I*: Testing for Spatial Dependence Based on the Spatial Autoregressive Model', *Geographical Analysis*, 39, 357–375.

Martell, D.L., and Sun, H. (2008), 'The Impact of Fire Suppression, Vegetation, and Weather on the Area Burned by Lightning-Caused Forest Fires in Ontario', *Canadian Journal of Forest Research*, 38, 1547–1563.

Moran, P.A.P. (1950), 'Notes on Continuous Stochastic Phenomena', *Biometrika*, 37, 17–23.

Oden, N. (1995), 'Adjusting Moran's *I* for Population Density', *Statistics in Medicine*, 14, 17–26.

Podur, J.J. (2001), 'Spatial and Temporal Patterns of Forest Fire Activity in Canada', Dissertation, University of Toronto.

Xiong, Y. (2015), 'Analysis of Spatio-Temporal Data for Forest Fire Control', Master's Thesis, Simon Fraser University.

Zhang, T., and Lin, G. (2016), 'On Moran's *I* Coefficient under Heterogeneity', *Computational Statistics and Data Analysis*, 95, 83–94.