2003

Monte Carlo assessments of goodness-of-fit for ecological simulation models

Lance A. Waller
Emory University

David Smith
University of Maryland

James E. Childs
Centers for Disease Control and Prevention, Atlanta

Leslie A. Real
Emory University

ELSEVIER

Ecological Modelling 164 (2003) 49–63

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Monte Carlo assessments of goodness-of-fit for ecological simulation models

Lance A. Waller [a,*], David Smith [b], James E. Childs [c], Leslie A. Real [d]

[a] *Department of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA*
[b] *Department of Epidemiology and Preventive Medicine, University of Maryland, Baltimore, MD, USA*
[c] *National Center for Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA*
[d] *Department of Biology, Emory University, Atlanta, GA, USA*

Received 22 April 2002; received in revised form 20 November 2002; accepted 2 December 2002

Abstract

One often develops stochastic ecologic simulation models based on local interactions between individuals or groups and bases systemic conclusions on trends summarized over multiple data sets generated from the model. In many cases, such models generate data sets ("realizations") each violating the usual assumptions associated with traditional statistical tests of goodness-of-fit, most notably that of independent observations. Monte Carlo hypothesis tests applied to multiple realizations from such models provide appropriate goodness-of-fit tests regardless of within-model peculiarities. The Monte Carlo tests address the question "Do the observed data appear consistent with the model?" in contrast to the usual question "Does the model appear consistent with the observed data?". In addition, such tests can make use of the same data sets used to draw systemic inference (i.e. the tests require no additional simulation runs). We illustrate the concept using Pearson's chi-square statistic with correlated data. We also consider the behavior of a similar statistic and of "modeling efficiency" in assessing the fit of a simulation model for the spatial spread of raccoon rabies in Connecticut.

Keywords: Model validation; Model assessment; Modeling efficiency; Goodness-of-fit; Simulation

1. Introduction

Stochastic ecological models often involve the specification of mathematical and probabilistic connections between experimental units or local collections of units, and analysis involves extension of such processes to system-wide outcomes for comparison to observed system level data. Extensions may be analytic or (increasingly) based on computer simulations that repeatedly apply models of local processes within a large system. In the latter case, one often bases system-wide inference on the average behavior observed across an ensemble of simulated data sets (realizations) from the same underlying model.

One may test the fit of a proposed ecological model by comparing observed values to those expected under that particular model. Very generally speaking, the analyst wishes to test the following conceptual null hypothesis:

$$H_0: \text{the data appear to be a typical realization of the model} \quad (1)$$

* Corresponding author. Tel.: +1-404-727-1057; fax: +1-404-727-1370.
E-mail address: lwaller@sph.emory.edu (L.A. Waller).

0304-3800/03/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0304-3800(03)00011-5

versus the alternative

H_A : the data do not appear to be a typical realization of the model.

We focus on assessments of fit for stochastic simulation models by statistical exploration of the observed variability between model realizations (simulated data sets) for a fixed set of model parameters. Deterministic models based on, e.g. natural laws and mass–balance relationships also play an important role in ecological modeling, however, variation between model output and the observed data in deterministic models often is attributed to uncertainty in model parameters, uncertainty in model structure, or measurement error rather than construed as an additional aspect of the system to be modeled in its own right. Our approach is a statistical consideration of the system-wide variability observed in stochastic models to assess model fit.

Statistical assessments of fit do not address all aspects of model validation (as discussed briefly in Section 2), however, the focus of this paper involves the specific component of model validation regarding the quantification of model fit through the conceptual statistical hypothesis testing approach above. Section 3 reviews the fundamental structure of Monte Carlo hypothesis tests and outlines their general application to the output of simulation-based ecological models. Section 4 illustrates the value of Monte Carlo tests via the impact of spatial correlation on the appropriateness of the asymptotic distribution of Pearson's chi-square statistic. Section 5 reviews a simulation model of the spread of rabies among raccoons in Connecticut beginning in 1991 and uses Monte Carlo testing to assess the fit of two proposed models. Finally, Section 6 provides general conclusions and directions for future developments.

2. Ecological model validation and goodness-of-fit

There is a broad literature on model validation (e.g. Hamilton, 1991 contains a list of 316 references on various aspects relating to the topic), and we attempt only the briefest outline here to motivate our approach and contrast it with existing statistical methods. The term "model validation" has very wide usage in the scientific literature and Rykiel (1996) notes some authors consider it an absolutely essential part of the modeling process while others consider it utterly impossible. Actual assessments of validity may vary in focus depending the model's purpose, e.g. contrast the purposes of prediction of the course of a new disease outbreak versus the identification of factors modifying a previously observed disease outbreak. Mayer and Butler (1993) stress that no single combination of validation tests or methods will be applicable across the diverse range of models and their possible uses.

Assessments of the goodness-of-fit (how well a model matches the observed data) are the focus of our discussion. While admittedly only a portion of the model validation process (Mayer and Butler, 1993; Power, 1993; Rykiel, 1996; Vanclay and Skovsgaard, 1997), goodness-of-fit provides some indication of the relationship between the model and the data from which it was derived, or the predictability of a model when applied to an independent set of data. Power (1993) and Mayer and Butler (1993) provide two oft-cited overviews of model validation containing reviews of statistical approaches to goodness-of-fit mostly comparing model traces (output realizations) to data sets. However, it appears the modeling literature contains very little regarding variability between independent realizations of a stochastic simulation model, a key component to our development below. In particular, to statistically assess the null hypothesis in Eq. (1) we need to know whether the observed data fall within the set of outcomes expected to result from the model.

Loehle (1997) stresses the notion of variability by pointing out that "too good" a fit reflects an overfit model (the model simply regenerates the data as observed). In particular, Loehle (1997, p. 155) notes: "We should not be asking how good a curve fit we can obtain with our model when evaluating how valid it is, but rather we should ask whether we can distinguish it from reality". Toward this end, Loehle (1997, p. 157) proposed assessing goodness-of-fit by focusing on the question: "Can one distinguish the model from the real system?" (paraphrase of Loehle, 1997, p. 157). In the absence of replicate observed data (either for the same study area at a different time or from a different study area) Loehle (1997) estimates the variability of the real system by confidence bands around the observed data values.

Many times, analysts are unlikely to have replicate data sets from the real system, but will have multiple replicate realizations from the model. That is, multiple model realizations provide analysts with ready assessment of the variability of *model* output allowing one to test the null hypothesis in Eq. (1) by testing whether the observed data (realization from the real system) are consistent with output from the model. In short, we propose reversing Loehle's (1997) question to read: "Can we distinguish the *data* from *the output of the model*?". The approach is the same as that proposed by Tsay (1992) in a time series setting, and adds a valuable tool to the model validation toolbox. We outline below how this reformulation of the question allows straightforward statistical assessment via Monte Carlo hypothesis testing, and provides probability inference contrasting with Loehle's (1997) confidence interval approach.

Reversing the question from whether the model falls within the observed variability of the data to whether the data falls within the observed variability of the model also distinguishes our approach from those of Whitmore (1991) and others based on replicate data sets. This is largely a matter of convenience from our perspective since, as mentioned above, replicate data sets from identical situations (e.g. identical field plots) are rare, while replicate output from the simulation model is plentiful. That is, we can estimate the variability between model realizations (with sample sizes limited only by computer time) much better than we can estimate the variability of the outcome measure, often even in the best designed experiment. Some may argue that the simulation models are designed to generate appropriate mean responses only and that the models are not necessarily attempting to accurately portray the between-realization variability of the real system. Even if this is the case, we argue that the between-realization variability of the model is of interest in defining the sorts of realizations possible from the model. For example, some portion of the data lying far outside the range of the corresponding values generated by the model suggests the model is unlikely to generate values consistent with these observations and identifies aspects of poor model fit.

Another approach similar to ours is that of Reynolds et al. (1981) who provide detailed discussion of the role of statistical tests in assessing the goodness-of-fit of a simulation model. Reynolds et al. (1981) propose comparing the observed data separately to each of a number of realizations from the model, then combining these tests in a multiple comparisons framework to provide overall inference regarding goodness-of-fit. Our approach is somewhat different, exploring the distribution of a test statistic calculated for individual realizations of the model rather than combining comparisons between each individual realization and the observed data.

3. Monte Carlo hypothesis testing

Barnard (1963) (in a discussion of Bartlett, 1963) introduces the concept of Monte Carlo hypothesis testing. The basic idea is a very simple one and essentially operationalizes frequency-based statistical inference. Suppose one wishes to test the null hypothesis presented in Eq. (1). One selects some test statistic denoted S ("S" for "statistic") and calculates its value for the observed data, say s_{obs}. (We follow standard notation and denote random variables by capital letters and observed values by lower case letters.) Under the null hypothesis (i.e. the model is true), S will follow a probability distribution based on the randomness generated within the model. One determines the weight of statistical evidence against H_0 (the statistical significance of the observed value s_{obs}) by assessing how consistent the observed value s_{obs} appears to be with the distribution of the test statistic S given that the null hypothesis is true. Typically, one selects a statistic S where extreme values (very large or small values) are more likely under the alternative hypothesis than under the null. In the following development, we assume that larger values of S occur more often under the alternative than under the null hypothesis, but note extension to the opposite case is straightforward. Under our assumption of increased values under the alternative, the P-value represents the probability under the null hypothesis that the test statistic S (a random variable) exceeds the observed value s_{obs}, i.e.

$$P = \Pr[S > s_{\text{obs}}|H_0] \tag{2}$$

(note we make explicit the dependence on the null hypothesis through a conditional probability statement).

The frequentist interpretation of the P-value corresponds to the long-run frequency of the event $S > s_{\text{obs}}$

under the null hypothesis, i.e. if one were to observe values s of S from repeated independent data sets each consistent with the null hypothesis (here, repeated independent realizations from the model under consideration), the long-run proportion of s values (based on the model) exceeding s_{obs} would converge to P. A Monte Carlo test is simply a computational implementation of this concept. One generates a large number (say, *nsim*) of independent realizations from the model, calculates the observed value of S for each realization, denoted s_i, $i = 1, \ldots, nsim$. A histogram of the values associated with the simulated data sets $(s_1, \ldots, s_{nsim})$ provides an estimate of the probability density of the test statistic under the null hypothesis. The proportion of test statistic values based on simulated data exceeding the value of the test statistic observed for the actual data set (s_{obs}) provides a Monte Carlo estimate of the upper tail P-value for a one-sided hypothesis test. Specifically, suppose s_{obs} denotes the test statistic for the observed data and $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(nsim)}$ denote the test statistic values (ordered from largest to smallest) based on the simulated data sets. If $s_{(1)} \geq \cdots \geq s_{(\ell)} \geq s_{\text{obs}} > s_{(\ell+1)}$, i.e. only the ℓ largest test statistic values based on simulated data exceed s_{obs}, then the estimated P value is

$$\widehat{\Pr}[S \geq s_{\text{obs}}|\mathrm{H}_0] = \frac{\ell}{nsim + 1},$$

where we add one to the denominator since our estimate is based on the *nsim* values from the simulated data plus the value based on the observed data.

Monte Carlo testing has seen considerable application in spatial statistics where the distributions of tests statistics under a null hypotheses of complete spatial randomness are either intractable or based on questionable asymptotics. The ease of simulating data realizations of complete spatial randomness provides a ready mechanism for applying Monte Carlo tests (Besag and Diggle, 1977; Diggle, 1983, pp. 7–9; Ripley, 1987, pp. 16–18; Cressie, 1993, pp. 635–636; Stoyan et al., 1995, pp. 142–143).

Monte Carlo testing is similar in spirit to permutation tests (Fisher, 1935) and nonparametric bootstrap hypothesis tests (Manly, 1991, Chapter 2; Efron and Tibshirani, 1993, Chapter 16; Davison and Hinkley, 1997, Chapter 4). Rose and Smith (1998) consider permutation tests of model fit very similar to the Monte Carlo tests above. However, permutation tests, their randomized counterparts, and bootstrap tests typically involve resampling the observed data in some manner, while Monte Carlo tests involve the generation of "new" data under the null hypothesis. (Parametric bootstrap methods are based on the same concept as Monte Carlo tests, see Efron and Tibshirani, 1993, pp. 53–56.) At first glance the generation of additional data under the null hypothesis would seem to involve more computation than merely resampling the observed data. However, for assessing the fit of simulation-based ecological models, Monte Carlo tests merely use the realizations already generated for system-wide inference based on the model and only require the additional calculation of the test statistic for each data set, essentially the same computational effort required for resampling-based methods. In addition, Monte Carlo tests provide estimates of exact probabilities where, in some cases, bootstrap tests do not (even though the difference between Monte Carlo and bootstrapped P-values will often be small, see Efron and Tibshirani, 1993, p. 223). Finally, Hope (1968) shows that Monte Carlo tests approximate uniformly most powerful tests (i.e. those with the highest power over all alternative hypotheses, cf. Lehmann, 1993, Chapter 3) with the approximation improving with increased numbers of simulations, *nsim*.

We note the Monte Carlo hypothesis tests outlined above reflect only one application of Monte Carlo simulation techniques in ecological modeling. One common use of Monte Carlo methods involves sensitivity analysis where investigators randomly vary parameter values to determine the sensitivity of model output to individual and subsets of parameters. Such sensitivity analyses may be applied to either stochastic or deterministic models. Hornberger and Spear (1980), Spear and Hornberger (1980), and Humphries et al. (1984) provide detailed early examples of this approach, and Vanclay and Skovsgaard (1997) give a brief overview in the context of forest growth models. A related use of Monte Carlo methods involves assessments of variability in model output based on uncertainty or randomness in model parameters. Kremer (1983), Annan (1997, 1999, 2001), and Yool (1999) provide detailed discussion of such assessments. Finally, van Horssen et al. (2002) use Monte Carlo simulation to explore the simultaneous impact of parameter uncertainty and covariate measurement error in spatial predictions. In this paper, we maintain a narrow focus on Monte Carlo

goodness-of-fit tests as part of assessing model adequacy for stochastic simulation models.

4. Illustrative example: Pearson's chi-square statistic with correlated data

We generate a simple example similar to those considered by Besag and Diggle (1977) to illustrate the appropriateness of Monte Carlo tests in situations where traditional asymptotic distributions do not hold due to correlation among observations.

To begin, consider Pearson's chi-square goodness-of-fit test as generally defined by

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{3}$$

where O_i denotes the observed value for the ith observation, E_i its expected value under the model, and n is the total number of observations. Typically, X^2 is defined for count data (e.g. from a contingency table) where one presumes the O_i's follow independent Poisson(E_i) distributions where E_i is both the expected value and the variance of O_i. More generally, we could consider distributions with non-equal mean and variance by generalizing X^2 to

$$Y^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{V_i}, \tag{4}$$

where V_i denotes the variance of O_i under the model. Note that, under the model the square root of each summand in Y^2 has mean 0 and variance 1. Suppose the observed values are independent and each follow a Gaussian distribution such that

$$O_i \underset{\sim}{iid} N(E_i, V_i), \quad i = 1, \ldots, n.$$

The statistic Y^2 has an attractive interpretation as the sum of squared standardized residuals associated with the model. According to standard results regarding the definition of the χ_1^2 distribution as that of a squared standard $N(0,1)$ random variable, and the χ_n^2 distribution as that of a sum of n independent χ_1^2 random variables, we have

$$\frac{O_i - E_i}{(V_i)^{1/2}} \underset{\sim}{iid} N(0,1) \Rightarrow \frac{(O_i - E_i)^2}{V_i} \underset{\sim}{iid} \chi_1^2$$

$$\Rightarrow \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{V_i} \sim \chi_n^2.$$

The resulting chi-square distribution of Y^2 depends on two assumptions: (a) that the O_i are each distributed $N(E_i, V_i)$ and (b) that the O_i are mutually independent. Violating either (a) or (b) can result in values of Y^2 inconsistent with a χ_n^2 distribution, even if the data do originate from the model.

To explore the impact of violating (b) through the introduction of positive spatial correlation, consider the following simulation experiment. Consider a 12×12 grid of locations defined at locations $(x, y) = \{(0,0), (0,10), (0,20), \ldots, (120,120)\}$ (we use units of 10 for distance comparability with the raccoon rabies data in Section 5). We consider the distribution of the statistic Y^2 based on $\boldsymbol{O} = (O_1, O_2, \ldots, O_{144})$ following a multivariate normal distribution with mean zero, variance–covariance matrix $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{O} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, $i, j = 1, \ldots, 144$, $\sigma_{ii} = 1$, $\sigma_{ij} = \exp(-\gamma d_{ij})$, d_{ij} is the distance between the ith and jth location, and γ a parameter controlling the extent of positive spatial correlation. For independent observations (i.e. $\sigma_{ij} = 0$, for $i \neq j$), we expect Y^2 to follow a χ_{144}^2 distribution.

For each of three values of γ (0.5, 0.1, and 0.01), we define the variance–covariance matrix $\boldsymbol{\Sigma}$, and generate 500 data sets from a $MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution using the multivariate normal random number generator included in the freely available R statistical package. This generator uses a spectral decomposition of the variance–covariance matrix $\boldsymbol{\Sigma}$ to transform a sample of independent normal random variates into a sample of correlated multivariate (correlated) normal random variates (Ripley, 1987, p. 98). While elements within each data set are correlated, the data sets themselves are mutually independent. From each data set, we calculate the value of Y^2. The appendix provides a web link to software allowing replication of this example.

Fig. 1 compares the independence-based χ_{144}^2 probability density to histograms of 500 observations of Y^2 based on data generated with $\gamma = 0.5$ (top row), $\gamma = 0.1$ (middle row), and $\gamma = 0.01$ (bottom row). For $\gamma = 0.5$, our data are essentially independent (all spatial correlation occurs at distances shorter than the minimum distance between observations) and the histogram closely matches the χ_{144}^2 density, and the sample mean and variance are close to the theoretical values of 144 and 288, respectively. When we introduce appreciable spatial correlation in the data ($\gamma = 0.1$ or 0.05), we see the mean value of Y^2 is relatively

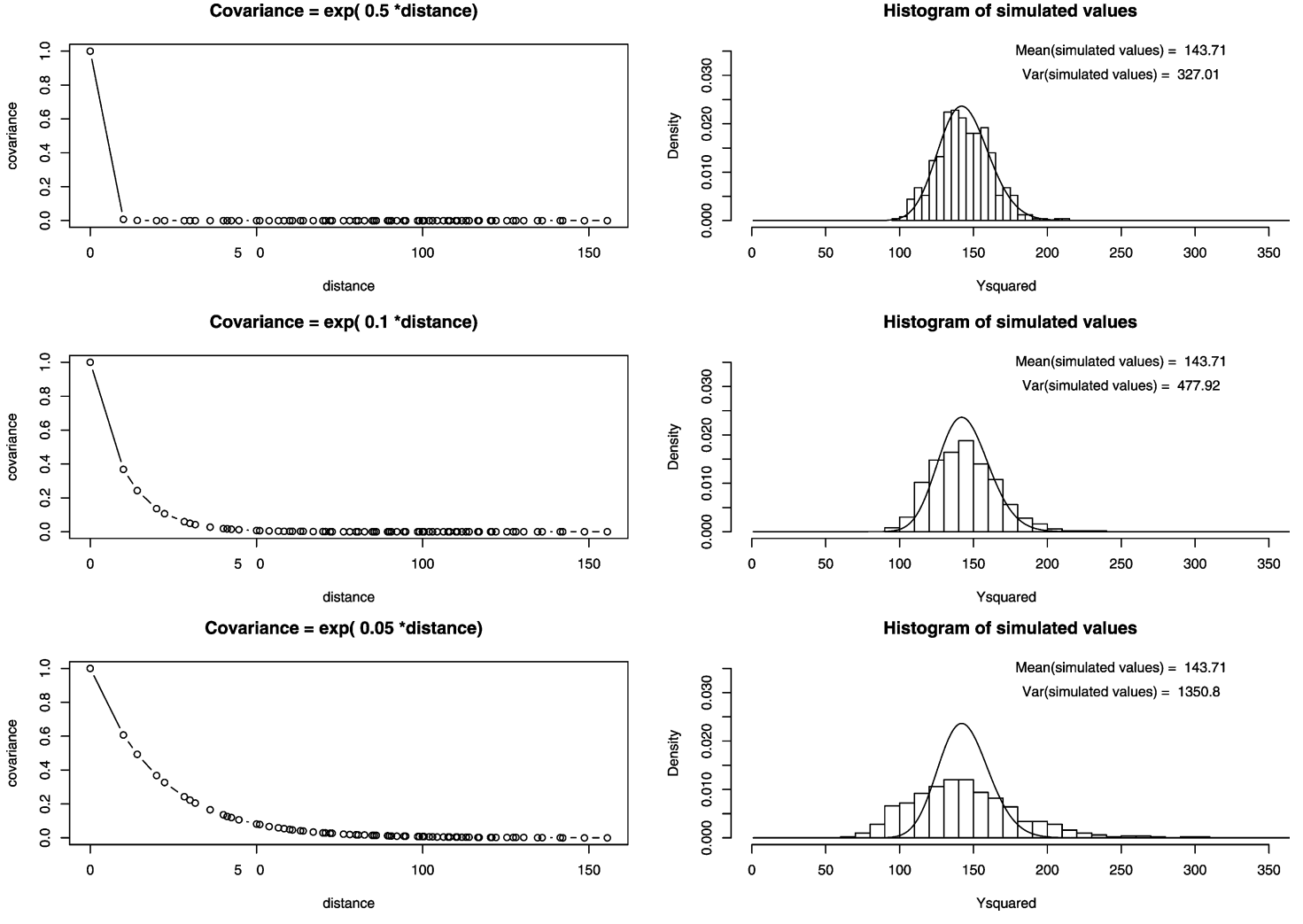

Fig. 1. The left column represents an exponential spatial covariance function. The right column represents the histogram of 500 modified chi-square statistics, Y^2, based on multivariate normal observations with mean zero, unit variance, and covariance defined by the function in the left column. The curve denotes the chi-square density based on an assumption of independence between observations within a single simulated data set.

unchanged but the variance is inflated over that of a χ^2_{144} random variable. The variance inflation arises since the amount of (statistical) information per observation is less in positively correlated observations than in a set of independent observations, for the same sample size. That is, one gains less precision in estimation from n correlated observations than from n independent observations.

Feldman et al. (1984) provide an analytic procedure for analyzing goodness-of-fit in correlated data, based on the test statistic

$$U = [\boldsymbol{O} - E(\boldsymbol{O})]^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} [\boldsymbol{O} - E(\boldsymbol{O})] \sim \chi^2_n.$$

For independent data, Feldman et al.'s (1984) statistic U reduces to Y^2. In general applications, using U requires an estimate of the mean observations ($E(\boldsymbol{O})$), and estimates of the elements of $\boldsymbol{\Sigma}$ (with appropriate adjustments to the degrees of freedom for the chi-square distribution). In practice, estimation of covariances can be difficult, especially for values with few associated observed pairs of observations (e.g. for long distances in our spatial example), and inversion of the $n \times n$ matrix $\boldsymbol{\Sigma}$ can be problematic for sparse matrices. In contrast, Monte Carlo testing bases inference on the histograms in Fig. 1 and does not require estimation of means, variances, or covariances, requiring instead a mechanism for generating observations according to the model, and code for calculating the test statistic for each simulated data set (code identical to that for calculating the test statistic value from the observed data).

5. Example: raccoon rabies in Connecticut

We further illustrate the concept and benefit of a Monte Carlo assessment of goodness-of-fit using data and models from a recent analysis of factors influencing the spread of raccoon rabies in Connecticut (Smith et al., 2002). The data consist of the date of the first reported case of raccoon rabies for each of Connecticut's 169 townships beginning with the index case in 1991 in Ridgefield township on the western edge of the state.

Briefly, Smith et al. (2002) create an interaction network among townships where the rate of spread into a new township depends on the fraction of adjacent townships already reporting cases. At a given iteration, the simulation model randomly selects the next township to report cases (based on a multinomial probability depending on the current reporting status of all townships), and updates the reporting status of all townships. The simulation continues until all townships report cases. The appendix provides a web link giving additional detail on the simulation model and access to the software. For the purposes of this paper, we concentrate on two of the models considered by Smith et al. (2002), namely, a "homogeneous" model with constant rates of spread between any pair of townships, and a "river" model where transmission rates are lower between townships separated by a river than between townships not separated by a river. Both models also include a constant background rate of rabies reporting in all townships corresponding to the possibility of long distance translocation of rabid raccoons (e.g. intentionally through restocking of hunting areas or unintentionally as in transport in garbage trucks) (Wilson et al., 1997). Model parameters are selected to minimize the sum of weighted squared residuals (Pearson's chi-square statistic)

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

where O_i denotes the observed time to first appearance for township i, and E_i is the expected time to first appearance for township i for the model under consideration, calculated as the average time to first appearance across 5000 data sets generated by the model.

For drawing inference from the fitted model, Smith et al. (2002) use an additional 5000 data sets generated with parameter values fixed at their estimated values. Fig. 2 illustrates boxplots of the time to first report for each township based on 500 of these 5000 realizations, where we order townships left to right by increasing distance from the index township (Ridgefield). The boxplots indicate the mean, interquartile range, and extreme values of the time to first report generated by the homogeneous model. We note immediately the increasing variability with increasing mean for time to first report. The observed data appear as the black line and a single realization of the model appears as the gray line in Fig. 2.

The model proposed by Smith et al. (2002) combines two dispersal processes, one modeling local spread via transmission probabilities from one

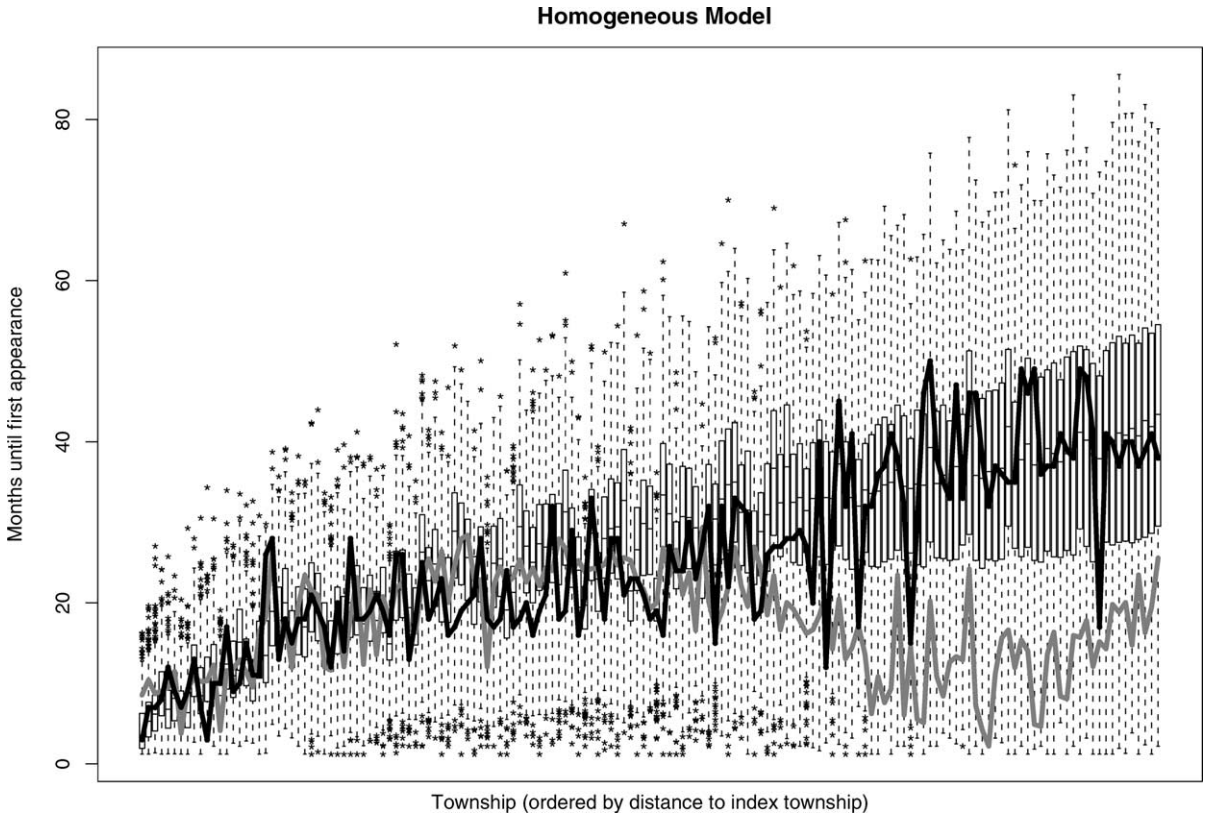

Fig. 2. Boxplots by township for homogeneous model in background, with townships ordered by distance to the index township (Ridgefield). The thick black line represents the observed data realization, the thick gray line represents a single realization from the model.

township to another, the other modeling random long distance translocations of infected animals. The first process provides a spatially predictable wavefront, while the second provides spatially unpredictable "shocks" to the system resulting in cases preceding the wavefront. As a result, a single realization from the model need not follow the mean behavior across multiple realizations, in particular, a single long distance translocation event may lower the time to first appearance for a number of relatively distant townships. The small but significant probability of a long distance translocation event results in the increasing variance at larger distances due to the range of multiple model pathways leading to initial disease incidence.

Fig. 2 also illustrates that assessments of model fit via comparison of the data to a single realization of the model (i.e. comparing the black and gray lines in Fig. 2 and ignoring model variability) could result in a conclusion of poor fit when in fact the data are quite consistent with the entire distribution of model realizations. A corresponding plot for the river model appears in Fig. 3. In both cases, the observed data tend to fall within the interquartile range (the white box) of the simulated values.

While parameters were selected to minimize the Pearson's chi-square criteria over the possible values of model parameters, the chosen parameters still may not provide good fit (i.e. the "best" model still may fit the data poorly). To this end we seek an assessment of model fit for the set of selected parameters. For the remainder of the paper, we assume model parameters fixed at their estimated values and wish to test the null hypothesis expressed in Eq. (1). (Allowing parameter values to vary in a manner similar to the sensitivity analyses mentioned briefly above adds additional complexity to the process which we ignore here for simplicity's sake.) We assess model fit for the homogeneous and river models for 159 of Connecticut's 169

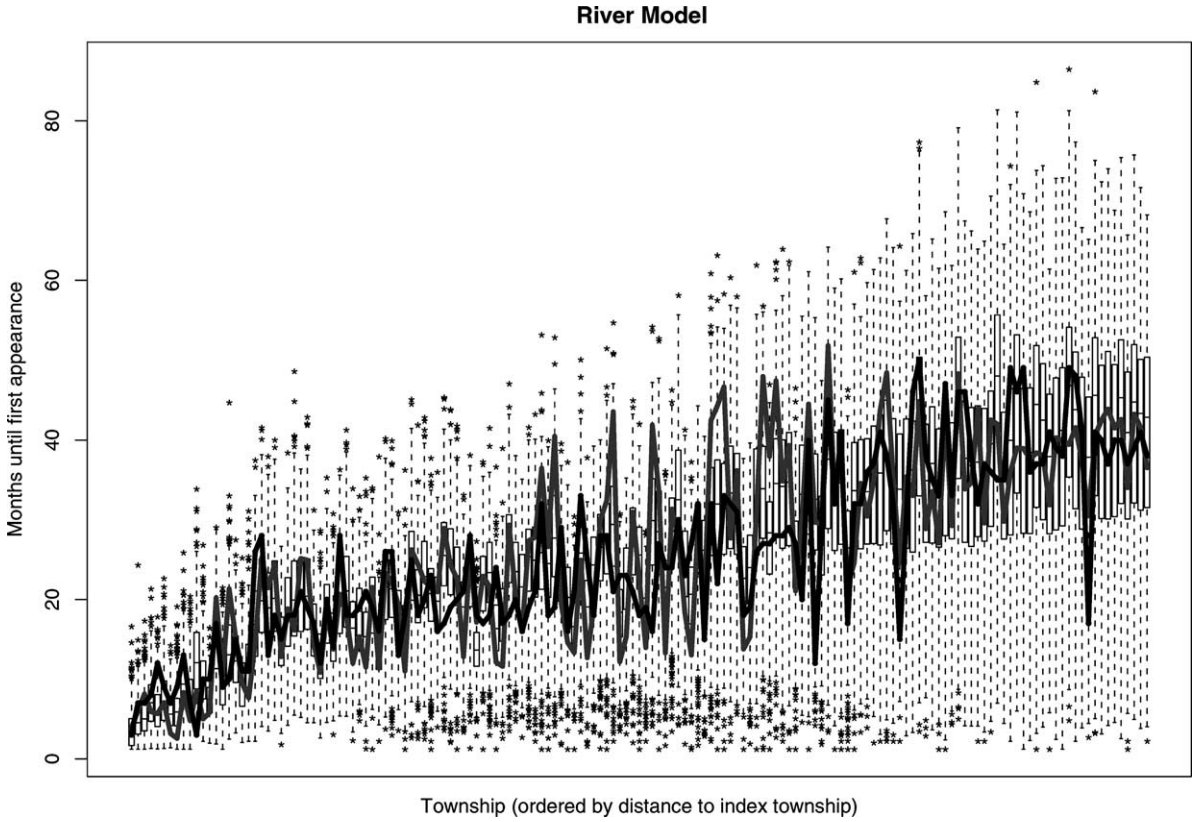

Fig. 3. Boxplots by township for river model in background, with townships ordered by distance to the index township (Ridgefield). The thick black line represents the observed data realization, the thick gray line represents a single realization from the model.

townships, ignoring the index township and townships along the western border of the state that may have acquired infection from New York rather than other townships in Connecticut.

First, consider goodness-of-fit as assessed by Y^2 as defined in Eq. (4). The observed values are 79.024 for the homogeneous model and 81.209 for the river model. If we assume independence between the times to first appearance for each township, we should compare these values to chi-square distributions with (159 − the number of estimated parameters) degrees of freedom. However, due to the spatial nature of the spread from township to neighboring township we might expect spatial correlation within each simulated data set (model realization). Such correlation between observations would not be a problem per se unless it also results in spatial correlation among the summands of Y^2, namely the squared standardized residuals

$$\frac{(O_i - E_i)^2}{V_i}.$$

Based on the example in Section 4, we would expect any such spatial correlation to inflate the variance of Y^2. Fig. 4 indicates the observed values of Y^2 for each model, the associated reference chi-square distribution, and a histogram of Y^2 values based on 500 realizations of each model. The histograms indicate variance inflation (very much similar to that observed in Section 4) and indicate that P-values based on the chi-square distributions are inaccurate. The P-values based on the (inaccurate) chi-square distributions are both >0.99 indicating a model fit that is far "too good to be true" (i.e. the data would appear to fit the model much better than would be expected even by chance), while the Monte Carlo P values of 0.914 for the

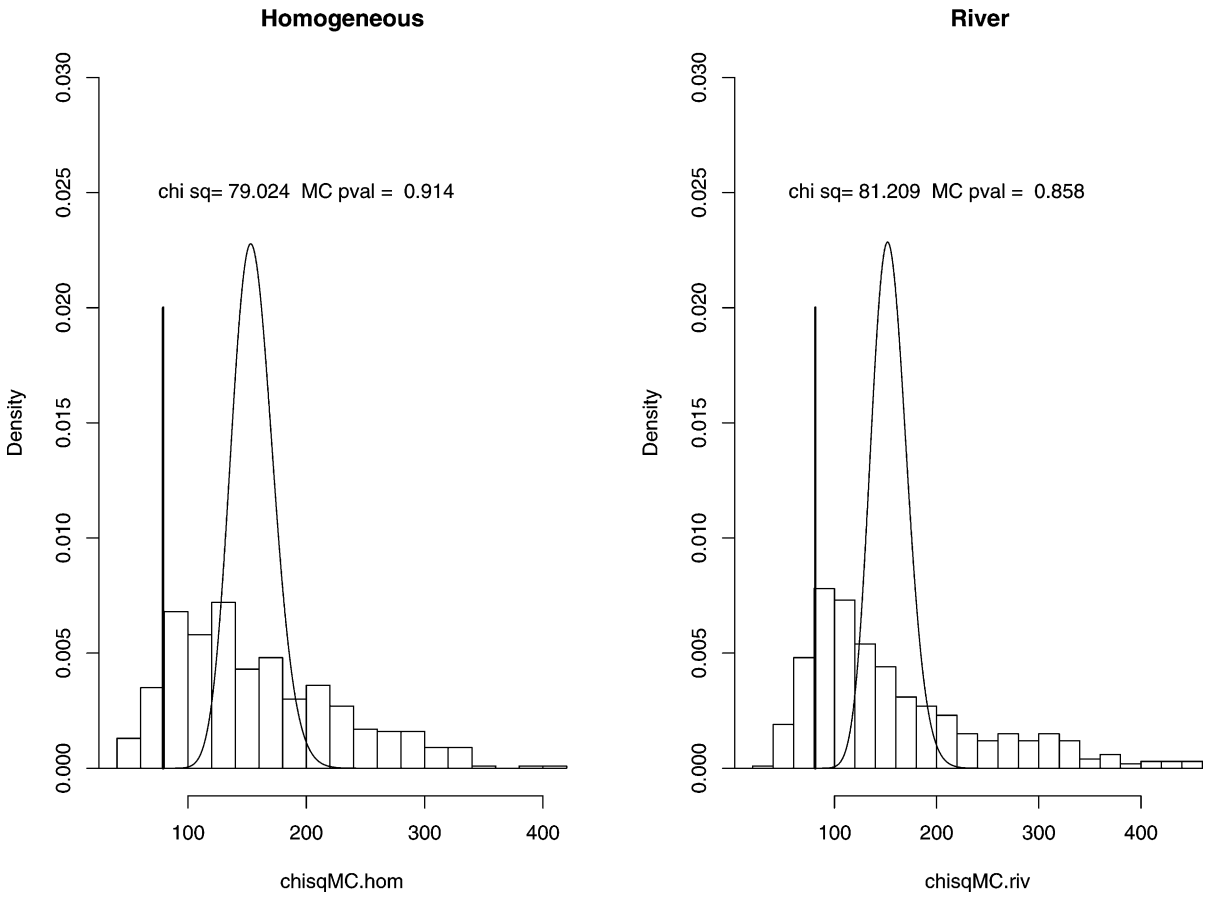

Fig. 4. Histograms of the goodness-of-fit statistic Y^2 (sum of squared standardized residuals, see text) based on 500 Monte Carlo simulations treating each data realization as a data set and calculating the expected values and variances based on the remaining 499 simulations. The vertical segments indicate the observed statistics based on the homogeneous model (left plot) and the river model (right plot). The solid curves represent chi-square densities assuming independence between townships, demonstrating the variance inflation in Y^2 due to spatial correlation induced by the models.

homogeneous model, and 0.856 for the river model are much more reasonable (even while indicating very good model fit).

In addition, we may also use the simulated realizations to indicate whether spatial autocorrelation among the squared standardized residuals appears to drive the variance inflation observed in Fig. 4. For each model, we calculate the spatial correlogram of the squared standardized residuals (elements of Y^2) and plot these as squares in Fig. 5. Note that the correlograms are slightly different for each model since they are based on residual values which depend on the model through E_i and V_i. For each realization, we also calculate the spatial correlogram of squared standardized residuals comparing that model realization to the other 499 model realizations, and display boxplots of these correlogram estimates in Fig. 5. The boxplots indicate the variability associated with the sample correlogram of the squared standardized residuals under the null hypothesis expressed in Eq. (1). Both models result in appreciable spatial correlation among the squared standardized residuals, and the correlogram of the observed standardized residuals (the squares) fall into the range expected under each model.

Note the range of values in the boxplots in Fig. 5 reflects the probability distribution associated with the correlogram estimates under H_0, i.e. the boxplots indicate Monte Carlo estimates of the probability of the

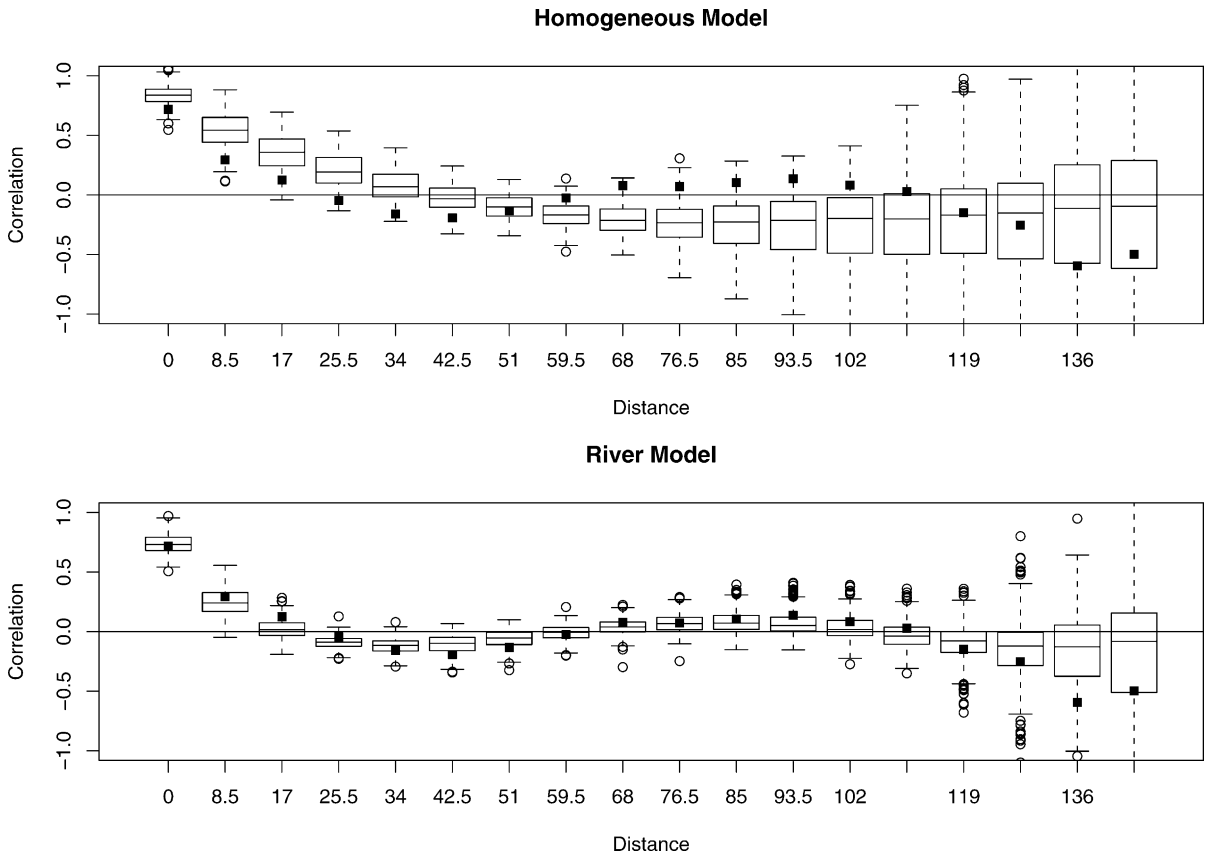

Fig. 5. Boxplots of correlogram values combined over 500 realizations each of the homogeneous and river models. The correlograms are based on the standardized residuals based on comparing each of 500 simulated data sets to the township-specific means and variances (estimated from the remaining 499 simulated data sets). The top correlogram corresponds to the homogeneous model and exhibits significant residual positive spatial correlation among the standardized residuals. The estimated correlograms based on the squared standardized residuals for the observed data appear as solid squares.

correlogram estimate falling in a given range of values, under H_0. Contrast these "probability intervals" with the notion of confidence intervals centered around data estimates (e.g. the squares in Fig. 5). Recall that (on average) 95% of the 95% confidence intervals constructed on independent realizations under the null hypothesis will contain the true (unknown) value of the estimand, but one cannot say that there is a 95% chance that a *given* 95% confidence interval contains the true value of the estimand. In this example, the "true" correlogram is an unknown function related to the underlying stochastic process defined by the model under consideration. However, one *can* say there is a 95% chance that the probability intervals in Fig. 5 contain the sample correlogram estimated from any single realization of the underlying model.

We note the probability intervals for correlogram estimates based on the river model are considerably tighter than those for the homogeneous model, probably due to the influence of rivers on the allowable range of model realizations. The correlogram of the observed standardized residuals follows both models, including the tighter pattern expected under the river model.

In addition to Y^2, we also consider a second summary measure of fit, namely the modeling efficiency EF recommended by Mayer and Butler (1993) and investigated by Alewell and Manderscheid (1998).

Modeling efficiency is defined by

$$\mathrm{EF} = 1 - \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}, \quad (6)$$

where $\bar{O}$ is the sample mean observed value. The modeling efficiency EF provides a dimensionless summary statistic very similar in structure to the coefficient of determination R^2 from linear regression, and we similarly interpret EF as the proportional reduction in variation of observed values around the model expectation to variation around the observed mean value. Note $\bar{O}$ represents the "worst case" regression line (slope = 0) indicating a lower bound of 0 for R^2, but Loehle (1997) point out that no such lower bound exists for EF.

While modeling efficiency is generally presented as a single summary with no associated variability, we could easily assess the between-realization variation in EF for a particular model by calculating EF for the rth simulated data realization through

$$\mathrm{EF}_r = 1 - \frac{\sum_{i=1}^{n}(O_{r,i} - E_{-r,i})^2}{\sum_{i=1}^{n}(O_{r,i} - \bar{O}_r)^2}, \quad (7)$$

where $O_{r,i}$ represents the (simulated) time to first appearance in township i for the rth simulated data set, $\bar{O}_r$ the sample mean time to first appearance across townships within simulated data set r, and $E_{-r,i}$ the sample mean time to first appearance across all simulated data sets except the rth set within township i and $r = 1, \ldots, 500$. The calculation of EF_r ignores the observed data, treats each realization as the observed data set, and calculates ER_r using Eq. (7).

The observed values of EF are 67.9% for the homogeneous model and 75.9% for the river model. These observed values suggest a better fit for the river model than for the homogeneous model, but one wonders how variable the EF value is under the respective null hypotheses of each model being true in turn. Again, our simulated data sets provide a sample of EF_r values under each null hypothesis, indicating the variability of EF in each situation. Histograms of EF_r for each model appear in Fig. 6. We notice immediately the long lower tail in each situation, and the associated range of values of (−4.543, 0.870) for the homogeneous model and (−7.447, 0.915) for the river model. Table 1 provides Monte Carlo probability estimates, in particular, one observes negative EF values

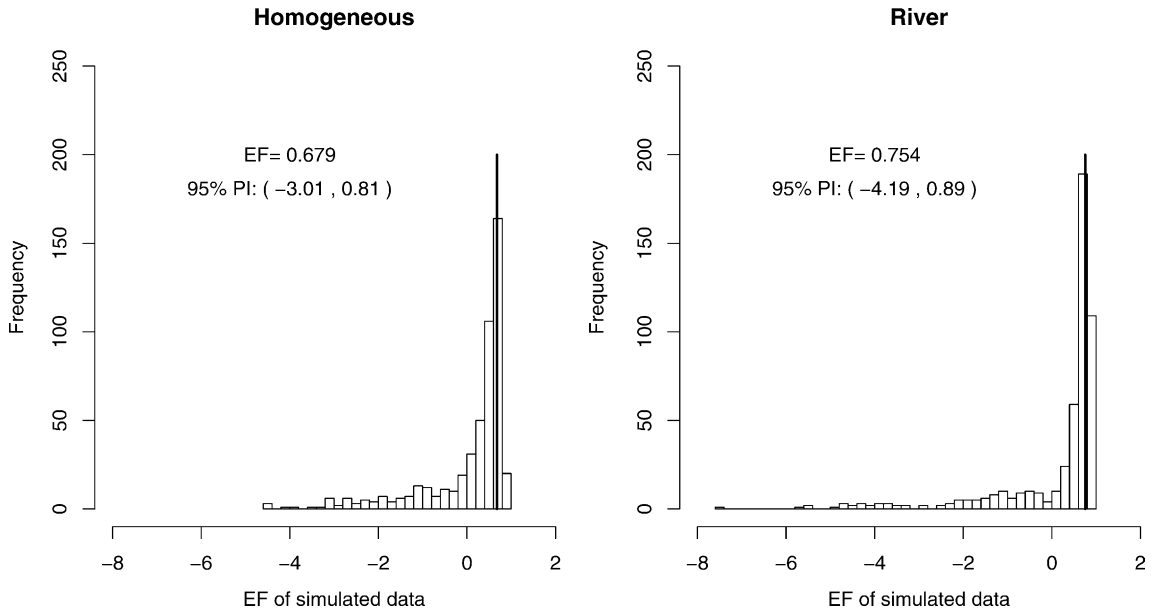

Fig. 6. Histograms of the modeling efficiency (EF, see text) based on Monte Carlo simulations treating each data realization as a data set and calculating the expected values based on the remaining 499 simulations. The vertical segments indicate the observed modeling efficiencies for both the homogeneous and the river models.

Table 1
Monte Carlo estimates of probabilities for the modeling efficiency EF based on 500 realizations of each model

Model	Event	Probability of event
Homogeneous	EF > 0.5	0.48
River	EF > 0.5	0.68
Homogeneous	EF > 0.75	0.10
River	EF > 0.75	0.37
Homogeneous	EF > 0.85	0.002
River	EF > 0.85	0.08
Homogeneous	EF $\in (0, 1)$	0.74
River	EF $\in (0, 1)$	0.78

with probability 0.26 and 0.22 for the homogeneous and river models, respectively, even if the associated null hypothesis is true. That is, a data realization generated under either of the two models will result in a negative value of EF approximately one quarter of the time, even in comparison to the "correct" model. Such a wide variability in the index EF, and such a high proportion of negative values *under the appropriate associated null hypothesis* casts some doubt on the utility of EF as a measure of goodness-of-fit, at least in this particular application.

6. Conclusions

The examples above illustrate the application and utility of Monte Carlo testing in assessing the goodness-of-fit of ecological simulation models. The approach operationalizes a basic question in goodness-of-fit expressed by the null hypothesis in Eq. (1), and provides probabilistic inference regarding model fit using the same model realizations used to provide inference about the modeled ecologic process. As detailed above, the approach assesses model fit via comparison to the variability generated by the model, rather than the assumed distribution of the data.

More generally, the methods described above easily assess the predictability of models derived and parameterized on one set of data and applied to another, and are perhaps most appropriate in this situation. For instance, we could replicate the results in Section 5 to assess the performance of the homogeneous and river models using parameters derived from the Connecticut data to predict the spread of raccoon rabies in Pennsylvania.

We note that statistical hypothesis tests do not prove the validity of a model, i.e. failing to reject a null hypothesis of "the model fits" does not prove the model fits, but the rejection of such a null hypothesis can point to particular problems by identifying particular portions of the data resulting in poor model fit.

In addition to hypothesis testing, Monte Carlo simulations also provide readily interpretable probability intervals. Using these, we show the modeling efficiency EF, at least for our raccoon rabies application, is too variable under the null hypothesis to allow useful assessment or comparison of models.

The examples also illustrate that the Monte Carlo approach is no panacea for model validation. In particular, while the approach provides valid estimated P-values for the modified Pearson's chi-square statistic Y^2, our examples did not consider direct comparisons between the homogeneous and river models in the raccoon rabies example. Model comparison is another component of model validation, and is an important topic for future research. Of particular interest is the comparison of *nested* models such as the homogeneous and river models above. Simulation-based statistical inference provides some promise for model comparison in general with particular examples appearing in Efron and Tibshirani (1993, pp. 190–198, comparing linear models) and Davison and Hinkley (1997, pp. 393–396, comparing time series models).

The examples above compare observed data to that generated by models assuming parameter values set at their estimated values. Trying to simultaneously estimate parameters and assess model fit complicates the approach considerably, raising the possibility of identifiability issues or difficulty in model convergence. In particular, the conditional probability statement given in Eq. (2) is not valid if the model and its associated parameters are not fixed. Statistically speaking, if we allow more than one model or more than one set of parameter values, we have a *composite null hypothesis* rather than a *simple null hypothesis* (cf. Hall and Titterington, 1989; Efron and Tibshirani, 1993, p. 210; Davison and Hinkley, 1997, Chapter 4). Monte Carlo methods still apply for composite null hypotheses, but the simulation approaches are more involved (Theiler and Prichard, 1996; Bølviken and Skovlund, 1996; Engen and Lillegård, 1997). In short, model fitting and

assessment of model fit are two separate but not necessarily independent operations, and the development of accurate and appropriate methodologies combining the two goals remains an active research area.

In conclusion, Monte Carlo methods offer an approach to draw statistical inference beyond just mean (average) behavior from ecological simulation models, particularly when realizations of such models violate many traditional statistical assumptions (e.g. independence). The observed variability in the output of such models provides valuable summary information regarding model fit and performance, and Monte Carlo methods offer a ready means to extract this information.

Acknowledgements

This research was supported in part by the National Center for Ecological Analysis and Synthesis (a center funded by NSF grant DEB-94-21535, University of California, Santa Barbara, the California Resources Agency, and the California Environmental Protection Agency) and NIH Grant R01 AI47498-03 (LAR). The authors also thank the state epidemiologists and state laboratory personnel in the United States who acquired and distributed the rabies incidence data over many years.

Appendix A

The web site http://www.sph.emory.edu/~lwaller contains links to the R statistical package, R code for the multivariate normal Pearson's chi-square example described in Section 4, and the raccoon rabies simulator described in Section 5.

References

Alewell, C., Manderscheid, B., 1998. Use of objective criteria for the assessment of biogeochemical ecosystem models. Ecol. Model. 107, 213–224.

Annan, J.D., 1997. On repeated parameter sampling in Monte Carlo simulations. Ecol. Model. 97, 111–115.

Annan, J.D., 1999. Reply to: comments on the paper: on repeated parameter sampling in Monte Carlo simulations. Ecol. Model. 124, 255–257.

Annan, J.D., 2001. Modelling under uncertainty: Monte Carlo methods for temporally varying parameters. Ecol. Model. 136, 297–302.

Barnard, G.A., 1963. Discussion of; The spectral analysis of point patterns. J. Roy. Stat. Soc., Ser. B 25, 294.

Bartlett, M.S., 1963. The spectral analysis of point patterns (with discussion). J. Roy. Stat. Soc., Ser. B 25, 264–296.

Besag, J., Diggle, P.J., 1977. Simple Monte Carlo tests for spatial patterns. Appl. Stat. 26, 327–333.

Bølviken, E., Skovlund, E., 1996. Confidence intervals from Monte Carlo tests. J. Am. Stat. Assoc. 91, 1071–1078.

Cressie, N.A.C., 1993. Statistics for Spatial Data, Revised Edition. Wiley, New York.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge, UK.

Diggle, P.J., 1983. Statistical Analysis of Spatial Point Patterns. Academic Press, London.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall/CRC Press, Boca Raton, FL.

Engen, S., Lillegård, M., 1997. Stochastic simulations conditioned on sufficient statistics. Biometrika 84, 235–240.

Feldman, R.M., Curry, G.L., Wehrly, T.E., 1984. Statistical procedure for validating a simple population model. Environ. Entomol. 13, 1446–1451.

Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edingburgh, UK.

Hall, P., Titterington, D.M., 1989. The effect of simulation order on level accuracy and power of Monte Carlo tests. J. Roy. Stat. Soc., Ser. B 51, 459–467.

Hamilton, M.A., 1991. Model validation: an annotated bibliography. Commun. Stat.—Theory Methods 20, 2207–2266.

Hope, A.C.A., 1968. A simplified Monte Carlo significance test procedure. J. Roy. Stat. Soc., Ser. B 30, 582–598.

Hornberger, G.M., Spear, R.C., 1980. Eutrophication in Peel Inlet—I. The problem-defining behavior and a mathematical model for the phosphorus scenario. Water Res. 14, 29–42.

Humphries, R.B., Hornberger, G.M., Spear, R.C., McComb, A.J., 1984. Eutrophication in Peel Inlet—III. A model for the nitrogen scenario and a retrospective look at the preliminary analysis. Water Res. 18, 389–395.

Kremer, J.N., 1983. Ecological implications of parameter uncertainty in stochastic simulation. Ecol. Model. 18, 187–207.

Lehmann, E.L., 1993. Testing Statistical Hypotheses. Chapman and Hall, London.

Loehle, C., 1997. A hypothesis testing framework for evaluating ecosystem model performance. Ecol. Model. 97, 153–165.

Manly, B.F.J., 1991. Randomization and Monte Carlo Methods in Biology. Chapman and Hall, London.

Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecol. Model. 68, 21–32.

Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model. 68, 33–50.

Reynolds, M.R., Burkhart, H.E., Daniels, R.F., 1981. Procedures for statistical validation of stochastic simulation models. For. Sci. 27, 349–364.

Ripley, B.D., 1987. Stochastic Simulation. Wiley, New York.

Rose, K.A., Smith, E.P., 1998. Statistical assessment of model goodness-of-fit using permutation tests. Ecol. Model. 106, 129–139.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90, 229–244.

Smith, D., Lucey, B., Childs, J.E., Real, L.A., Waller, L.A., 2002. Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. Proc. Natl. Acad. Sci. U.S.A. 99, 3668–3672.

Spear, R.C., Hornberger, G.M., 1980. Eutrophication in Peel Inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res. 14, 43–49.

Stoyan, D., Kendall, W.S., Mecke, J., 1995. Stochastic Geometry and its Applications. Wiley, Chichester, UK.

Theiler, J., Prichard, D., 1996. Constrained-realization Monte-Carlo method for hypothesis testing. Physica D 94, 221–235.

Tsay, R.S., 1992. Model checking via parametric bootstraps in time series analysis. Appl. Stat. 41, 1–15.

van Horssen, P.W., Pebesma, E.J., Schot, P.P., 2002. Uncertainties in spatially aggregated predictions from a logistic regression model. Ecol. Model. 154, 93–101.

Vanclay, J.K., Skovsgaard, J.P., 1997. Evaluating forest growth models. Ecol. Model. 98, 1–12.

Whitmore, A.P., 1991. A method for assessing the goodness of computer simulation of soil processes. J. Soil Sci. 42, 289–299.

Wilson, M.L., Bretsky, P.M., Cooper, G.H., Egbertson, S.H., Van Kruinigen, H.J., Cartter, M.L., 1997. Emergence of raccoon rabies in Connecticut, 1991–1994: spatial and temporal characteristics of animal infection and human contact. Am. J. Trop. Med. Hygiene 57, 457–463.

Yool, A., 1999. Comments on the paper: on repeated sampling in Monte Carlo simulations. Ecol. Model. 115, 95–98.