Kolmogorov–Smirnov Goodness–of–Fit Test for Spatial Data

N. Glenn Griesinger, Ph. D.,
Texas Southern University
Jose Guardiola, Ph. D.,
Texas A & M University Corpus Christi

Abstract

A key observation in the Kolmogorov–Smirnov Goodness–of–Fit test is that the largest difference between the cumulative distribution function $F(x)$ and the empirical distribution function $F_n(x)$ converges to 0 in probability. This observation allows one to test whether two underlying univariate probability distributions differ from each other or whether a given distribution differs from a hypothesized distribution, assuming the data are unbiased and independent. However, these assumptions render the test inadequate for spatial data. We extend the Kolmogorov–Smirnov goodness–of–fit test to spatial data, and assume the data are from a regularly or an irregularly spaced lattice.

1 Introduction

A spatial distribution depicts a trend or occurrence. Determining the distribution of a spatial dataset, or determining if two spatial datasets have the same distribution allows for more accurate data analysis of a trend. Tests such as goodness of fit tests are well established for independent, identically distributed data. However, goodness of fit tests for spatial data have received little attention (Olea, 2009). Spatial distribution applications include epidemiological, environmental, geographical, to name a few.

This paper is organized as follows. Section 2 discusses spatial data simulation methods. Since this research proposes a new Kolmogorov–Smirnov test for spatial data, Section 3 reviews the classical Kolmogorov–Smirnov test. The proposed Kolmogorov–Smirnov test uses a parametric spatial bootstrap, which is discussed in Section 4. Section 5 presents the new Kolmogorov–Smirnov test. A simulation study is provided in Section sec:ss. Final remarks are in Section 7.

2 Spatial Data Simulation

Spatial data are categorized according to a location j and a realized value

$$\{Z(j) : j \in d\}, \tag{1}$$

where j is in d–dimensional Euclidean space, $j \in \Re^d$. There are several categories of spatial data including geostatistical data, point patterns, objects, and lattice data. Our focus is regularly spaced lattice data.

A lattice, a countable collection of spatial sites, may be regular or irregular (Cressie, 1993). In the former, data are obtained over a regularly spaced set of points. In the latter, data are obtained over an irregularly spaced set of points. Lattice data are dependent observations from a spatial sampling region $\mathcal{R}_n \subset \Re^d$, $d \geq 1$ in d–dimensional Euclidean space on which a spatial process $\{Z_{\mathbf{s}} : \mathbf{s} \in \mathcal{Z}^d\}$ is observed on a grid $\mathcal{Z}^d$. A spatial process is the manner in which data values change from one spatial location to another. The variable d denotes the dimension of sampling, $\mathbf{s}$ represents the spatial sampling site, and n indicates sample size (Nordman, 2008).

Unlike random samples, statistical packages typically do not have functions that simulate spatial data. The following algorithm (Cressie, 1993) can be adapted to simulate spatial data:

The correlation structure is specified in the error term instead of in the mean as is conditional models.

Step 1: Determine a valid spatial covariance structure such as power, exponential, or circular, and determine its corresponding covariance matrix.

- Σ is an $n \times n$ covariance matrix.

Step 2: Obtain the Cholesky decomposition of the covariance matrix, Σ:

- $\Sigma = L \times L^T$, where L is the lower triangular matrix.
- Choose a constant mean μ for the spatial region.
- Generate a vector of independent, identically distributed random variables from the standard normal distribution.
- The error vector is $\epsilon = \epsilon_1, \ldots, \epsilon_n$, where ϵ_i $N(0, 1)$ for $i = 1, \ldots, n$.

Step 3: Generate a spatial data vector Z using the relationship

$$Z = \mu + L\epsilon \tag{2}$$

- The vector Z contains spatially correlated data in the region of interest.

Since lattice data are on a grid, the form of the data needs to be adjusted to accommodate the Kolmogorov–Smirnov test. This adjustment is accomplished through estimating equations.

3 Classical Kolmogorov–Smirnov Test

Although not originally intended as a goodness–of–fit test, the classical Kolmogorov–Smirnov test uses the empirical distribution $F_n(x)$ to test whether the cumulative distribution $F(x)$ is the distribution of x (Kolmogorov, 1933). The idea behind the Kolmogorov–Smirnov test is that the distribution of

$$\sup_{x \in \Re} |F_n(x) - F(x)| \tag{3}$$

does not depend on the unknown distribution of the sample when the data are continuous, independent, and identically distributed. The Kolmogorov–Smirnov test is motivated by Theorem 3.1.

Theorem 3.1 *If $H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}$ is the cumulative distribution function of the Kolmogorov–Smirnov distribution then*

$$P(\sqrt{n}\sup_{x\in\Re}|F_n(x) - F(x)| \leq t) \xrightarrow{D} H(t). \tag{4}$$

The statistic

$$D_n = \sqrt{n}\sup_{x\in\Re}|F_n(x) - F(x)| \tag{5}$$

is known as the Kolmogorov–Smirnov D statistic. It is not to be confused with $\xrightarrow{D}$, convergence in distribution. The D statistic is a measure of the discrepancy between the empirical and hypothesized distribution functions. Its values have been tabulated, which allows one to compute p–values.

Since the classical Kolmogorov–Smirnov Goodness–of–Fit test assumes that the data are independent, the authors propose a new Kolmogorov–Smirnov test for spatial data. The new test uses the parametric spatial bootstrap.

4 Parametric Spatial Bootstrap

Sampling bias and dependence, both inherent in spatiotemporal measurements, violate the assumptions of the classical Kolmogorov–Smirnov test. This necessitates a new Kolmogorov–Smirnov test for spatially correlated data. The proposed test obtains the distribution of the Kolmogorov–Smirnov D statistic by bootstrapping. However, the classical bootstrap (Efron and Tibshirani, 1993) cannot be used because it samples with replacement without regard to relative locations. Disregarding location destroys the correlation structure of the original spatial sample. The proposed method instead uses the parametric spatial bootstrap because its artificial samples maintain the correlation structure of the original data (Tang, 2005).

The parametric spatial bootstrap maintains the correlation structure by generating samples similar to the way in which spatial data are generated. The authors generate spatial data $\mathbf{Z}$ using the relationship

$$\mathbf{Z} = \mu + L\epsilon, \tag{6}$$

where μ is a constant mean value, L is the lower triangular Cholesky decomposition of the covariance matrix, and ϵ is the spatial error sequence. The purpose of L is to correlate the errors. The bootstrap resample

$$\mathbf{Z} = \hat{\mu} + \hat{L}\epsilon^*, \tag{7}$$

is obtained by bootstrapping the residuals ϵ^* which are generated from a Gaussian distribution.

The parametric spatial bootstrap algorithm is as follows (Tang, 2005). First estimate the spatial residuals $\hat{\delta} = \mathbf{Z} - \hat{\mu}$, and choose a semivariogram model based on empirical semivariogram estimates. Next, estimate the parameters in the semivariogram model using weighted least squares, then estimate the covariance matrix $\hat{C}$. Obtain the Cholesky decomposition matrix $\hat{L}$ of the estimated covariance matrix $\hat{C}$, then estimate the sampling distribution of $\hat{\theta}^*$ by repeating the following B times: Generate parametric bootstrap residuals ϵ^* from a Gaussian distribution. Transform the residuals and obtain bootstrap resamples Equation (7). Finally, calculate the statistic of interest $-2\log l(\theta) = \hat{\theta}^*$ from $\mathbf{Z}^*$.

5 Kolmogorov–Smirnov Test for Spatially Correlated Data

The proposed Kolmogorov–Smirnov test for spatially correlated data combines the spatial bootstrap with the spatial empirical likelihood method to calculate the distribution of the Kolmogorov–Smirnov D statistic at specified points for a finite sample of size n. It is based on an unbiased empirical sample of size n. The method is outlined as follows:

1. Determine spatial sampling region.
2. Vectorize spatial grid. The vectorization process breaks up the grid into several vectors.
3. Use the Parametric Spatial Bootstrap method to obtain artificial samples of vectorized grid. Each vectorized grid represents a sample of size n. The artificial sample has the same correlation structure as the empirical data. Spatial resamples are

$$Z^* = \hat{\mu} + \hat{L}\epsilon^*. \tag{8}$$

   Repeat to obtain at least 1000 artificial samples.
4. Determine empirical cumulative distribution function of the data, $F_n(x)$.
5. Determine D statistic:

$$D = \sup_x |F_n(x) - F(x)|, \tag{9}$$

   where $F(x)$ is the cumulative distribution function of the $\chi^2(r)$ distribution; r represent dimension of data.
6. Repeat previous steps at least 1000 times to obtain at least 1000 D statistic values.
7. Use cumulative probability plots to determine p–values:

$$p - value = 1 - quantile, \tag{10}$$

   where the null and alternative hypotheses are

   $H_0 : D = 0$

   $H_a : D > 0.$

6 Simulation Study

This simulation study uses the proposed Kolmogorov–Smirnov test to determine the D–statistic and p–value for spatially correlated data. The inference for variogram parameters are found in Table 1. Spatial data are generated using a regular grid of 50×50 with known Gaussian semi-variogram parameters, with values of: a partial sill σ^2 of 1.9, a nugget of 0.1 and range of 9 units. The corresponding covariance matrix was assembled using these values, and $2,500$ spatially correlated values were generated. There are $1,000$ samples of size 144 which were taken ignoring the spatial correlation structure, these samples are called uncorrelated samples due to the fact that we are ignoring the spatial correlation structure. Using these $1,000$ samples, the Kolmogorov–Smirnov test was performed for each sample and the corresponding D–statistics and p–values were saved. A second set of samples were

Table 1: Inference for Variogram Parameters

	Parameters	Estimates
Nugget	0.1	0.1
Partial Sill	1.9	1.91
Range	9.0	9.53

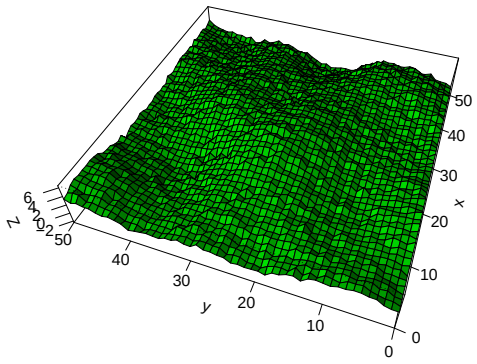

Figure 1: Graph depicts spatially correlated data using Equation (6). Inference for variogram parameters are in Table 1.

Table 2: Results for Simulation Study

Samples	Mean p–values	Mean D–satistics
Ignore Spatial Correlation	0.42	0.16
Consider Spatial Correlation	0.61	0.06

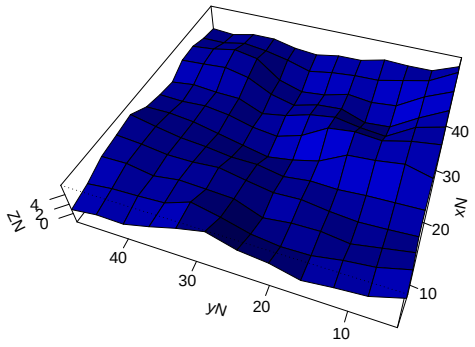

Figure 2: Graph depicts 1000 simulations from obtaining samples with the same correlation structure using Equation (8).

prepared, this time taking into account the spatial correlation structure. The semi–variogram parameters were estimated using maximum likelihoods estimation methods in R, as these values in a real world problem are unknown. The inference for the semi–variogram parameters are shown in Table 2. The estimated parameters are very close to the true values used to generate the original data set. Using the estimated parameters we generated a second set of $1,000$ samples of size 144 with each one of these samples followed the same semi-variogram structure imposed by the estimated semi–variogram. The Kolmogorov–Smirnov test was performed for each one of the 1 spatially correlated samples and the corresponding D–statistics and p–values were computed and saved.

7 Summary

The simulations results described in the previous section indicate that ignoring the spatial correlation produces p-values that are smaller than the p–values that were obtained when the samples were generated with the same correlation structure as the original data set. These results are in agreement with previous research (Olea, 2009). Therefore, given that obtaining the correlated samples involved more computational effort and time, as we need to infer the semi-variogram parameters and $1,000$ samples were generated using the same spatial correlation structure, we can conclude that the additional effort for preparing the correlated

samples is not necessary in most cases. Specifically, when the p-value ignoring the spatial correlation structure allows you to reject the null hypothesis for the Kolmogorov-Smirnov test, preparing the correlated samples that follow the same correlation structure as the original data set is not necessary. In another words, preparing the spatially correlated samples that follow the same spatial correlation as the original data set is not going to change the conclusion for the Kolmogorov-Smirnov when the null hypothesis was not rejected for the uncorrelated samples. The null hypothesis is not going to be rejected again by using the correlated samples for the Kolmogorov-Smirnov test.

References

Cressie, Noel A. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Inc.

Efron, Brad and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Glenn, Nancy L. and Zhao, Yichuan (2007). Weighted Empirical Likelihood estimates and their robustness properties. *Computational Statistics & Data Analysis* . **51**, 5130 – 5141.

Kolmogorov, A. N. (1933). Sulla Determinazione Empirica Di Una Legge Di Disribuzione. *Giornale dell'Istituto Italiano delgli Attuari* **4**: 1 – 11.

Nordman, Daniel J. (2008). A Blockwise Empirical Likelihood for Spatial Lattice Data. *Satistica Sinica* **18**: 1111 – 1129.

Olea, Ricardo A. and Pawlowsky–Glahn, Vera. (2009). Kolmogorov–Smirnov test for spatially correlated data. *Stochastic Environmental Research and Risk Assessment* **23**: 749–757.

Tang, Liansheng. (2005). Undercoverage of Wavelet–Based Resampling Confidence Intervals and a Parametric Spatial Bootstrap, Ph. D. thesis. Southern Methodist University. Dallas, Texas.