Assessing the significance of global and local correlations under spatial autocorrelation; a nonparametric approach.

Júlia Viladomat, Rahul Mazumder, Alex McInturff,
Douglas J. McCauley and Trevor Hastie.

January 28, 2013

Abstract

In this paper we present a method to assess the significance of the correlation coefficient when at least one of the variables is spatially autocorrelated. The standard test assumes independence of the samples. If the data are smooth, the assumption does not hold and as a result we reject in many cases where there is no effect (the precision of the null distribution used by standard tests is over-estimated). We propose a method that recovers the null distribution taking into account the autocorrelation; it is based on Monte-Carlo methods, and focuses on permuting, and then smoothing and scaling one of the variables so that we destroy the correlation with the other variable while at the same time maintaining the initial autocorrelation. This research has been motivated by a project in biodiversity and conservation in the Biology Department at Stanford University.

Keywords: Geostatistics; Monte-Carlo methods; Resampling; Spatial autocorrelation; Spatial statistics; Variogram.

1 Motivation

Assessing whether the correlation coefficient is significant is not straightforward when the values of the variables involved vary smoothly with location. Under the presence of spatial autocorrelation, classical tests based on Student's t (Fisher (1915)) tend to produce incorrect and exaggerated results. Some work has been done, particularly in the field of geostatistics. For example, Clifford et al. (1989) propose a method that estimates an effective (much reduced) sample size. Spatial autocorrelation implies that two

close-by locations have similar values, one of them not giving much new information, and thus the variability of the sample is smaller than if the sample was independent of the same size. To take this into account, the correlation coefficient is compared to a Student's t distribution with larger variance (fewer degrees of freedom) which accounts for the loss of precision due to the (spatial) dependence of the observations. The method however is developed for Gaussian random fields, but not for general distributions, and in reality smoothed processes tend to be non-Gaussian.

In addition to that, existing methodology focuses on global correlation coefficients. With a good simulation model, it is possible to examine the null distribution of a larger variety of statistics. For instance, this project started because our coauthors were looking at local correlations produced by Geographically Weighted Regression (GWR) methods. GWR is a set of regression techniques that deal with spatially varying relationships. The book Fotheringham et al. (2002) has captured considerable attention in the geostatistics community. However, they do not provide tests for assessing significance of the regressors in the model, and focus on comparing coefficients for different spatial areas, identifying the relationships that are stronger, but with no assessment to whether they are significant or not.

In this paper we propose a method to obtain global, as well as local p-values for the correlation coefficient, that takes into account the spatial autocorrelation. In the previous example, it returns a map of p-values for the local correlations provided with GWR (or any other). Our approach uses Monte-Carlo methods to recover the null distribution. It permutes the values of X, one of the two variables, across space. This destroys the correlation with the other variable Y, as well as its spatial autocorrelation. The latter is recovered by smoothing and scaling the permuted variable in a way that approximately recovers the variogram of the original variable X. By repeating this process many times, we obtain approximate realizations of the null distribution of interest.

The rest of the paper is organized as follows. In section 2 we introduce the problem through a real example and analyze the limitations of the standard test. Section 3 describes the alternative method proposed by this paper, and Section 4 gives some evidence on the performance of the method and compares it with the approach in Clifford et al. (1989).

2 Introduction of the problem

Protecting remote ecosystems is the future of global diversity. Our collaborators in this project mapped the locations of sites over the world using two criteria: quantity of species richness (biodiversity) and travel time to reach the nearest city (remoteness), see McCauley et al. (2012) and Figure 1, where the smoothed nature of both variables is obvious. An important question that arises from the mapping is whether remoteness and biodiversity are correlated with one another; i.e. are there more species in remote areas that are better insulated from human disturbance? To succinctly communicate the strength of these correlations, the authors are interested in reporting a p-value map for the areas where overlap between remoteness and biodiversity occurs.

We will use this example to illustrate our methodology, but for simplicity will focus on the american region of the world. Biodiversity (X) is the number of different species in an area of size 100×100 km and centered at location s. The variable X is the result of estimating the number of species of plants, amphibians, birds and mammals in the area. Each of the 4 counts is normalized to a maximum score of 10, with X being the average of those 4 normalized counts. Remoteness (Y) takes values between $1-8$ and indicates the travel time in days needed to reach the nearest city larger than $50,000$ inhabitants from location s, where 8 represents any travel time larger than 7 days. Our sample is denoted by $(X_{s_1}, Y_{s_1}), \ldots, (X_{s_N}, Y_{s_N})$, $\mathbf{s} = (s_1 \ldots s_N)$, where $s_i \in \mathbb{R}^2$ are the longitude and latitude coordinates of observation i and $N = 19,926$.

Figure 2 plots the local correlations between $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$, using a gaussian kernel truncated at the bandwidth $\lambda = 5.281$. The local correlation at location s is calculated as follows:

$$\hat{r}^{\lambda}_{X_{\mathbf{s}},Y_{\mathbf{s}}}(s) = \frac{\sum_{\|s-s_j\|\leq\lambda} w_{s_j}(X_{s_j} - \bar{X}_{\mathbf{s}})(Y_{s_j} - \bar{Y}_{\mathbf{s}})}{\sqrt{\sum_{\|s-s_j\|\leq\lambda} w_{s_j}(X_{s_j} - \bar{X}_{\mathbf{s}})^2 \sum_{\|s-s_j\|\leq\lambda} w_{s_j}(Y_{s_j} - \bar{Y}_{\mathbf{s}})^2}} \qquad (1)$$

where $\bar{X}_{\mathbf{s}} = \frac{\sum w_{s_j} X_{s_j}}{\sum w_{s_j}}$ and $\bar{Y}_{\mathbf{s}} = \frac{\sum w_{s_j} Y_{s_j}}{\sum w_{s_j}}$. As we describe in detail in Section 3.1, the R package locfit fits a local constant regression at each location s using kernel weights (see expression (4), where in this case we use the fix bandwidth λ). We compute (1) by breaking it down and separately evaluating the quantities $\sum w_{s_j} X_{s_j}$, $\sum w_{s_j} Y_{s_j}$, $\sum w_{s_j} X_{s_j} Y_{s_j}$, $\sum w_{s_j} X_{s_j}^2$ and $\sum w_{s_j} Y_{s_j}^2$ using locfit. Note the we have not used GWR to calculate local correlations; the results are very similar, but locfit is much more efficient.

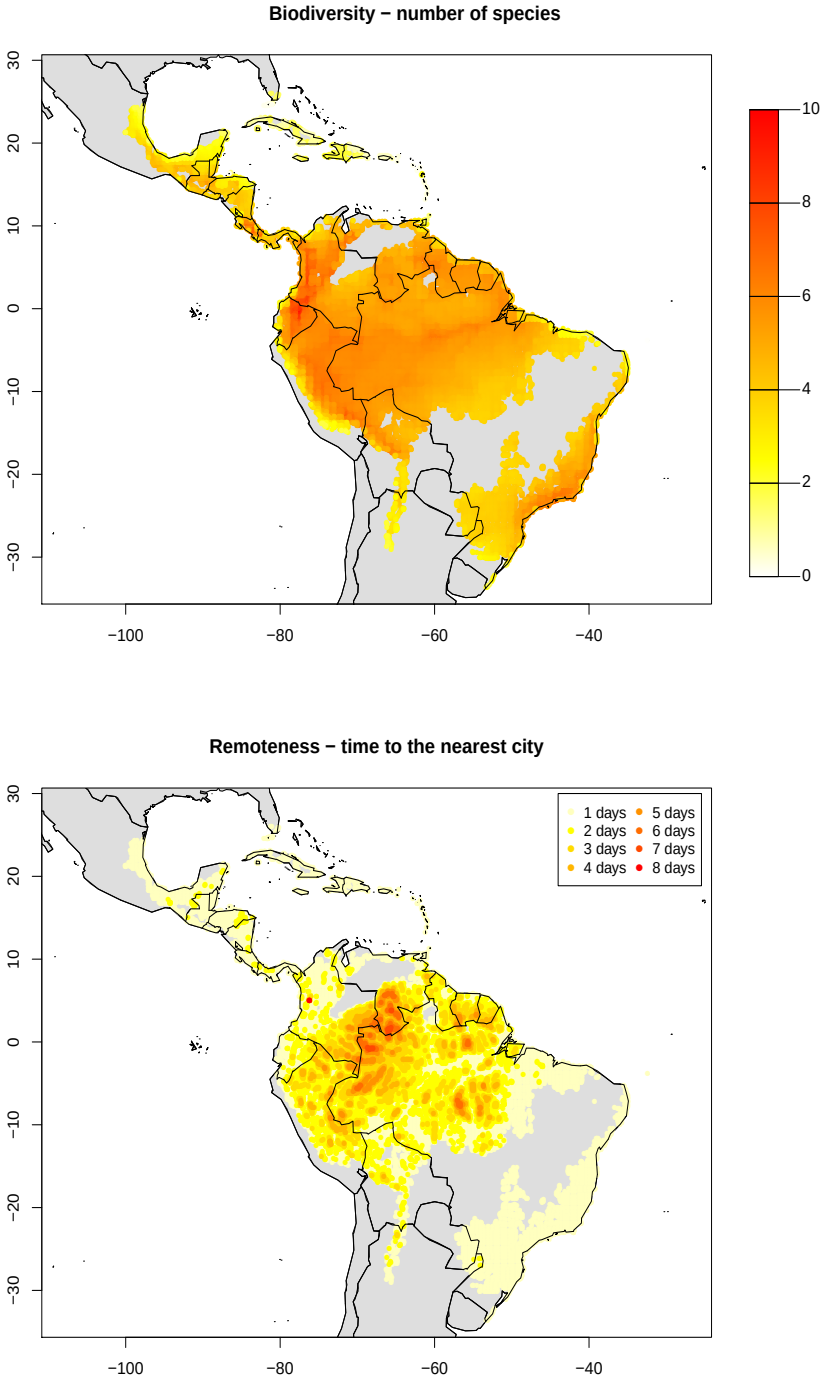


Figure 1: Variables biodiversity and remoteness (only areas where remoteness exceeds 1 day are considered, areas with no data are indicated in grey).

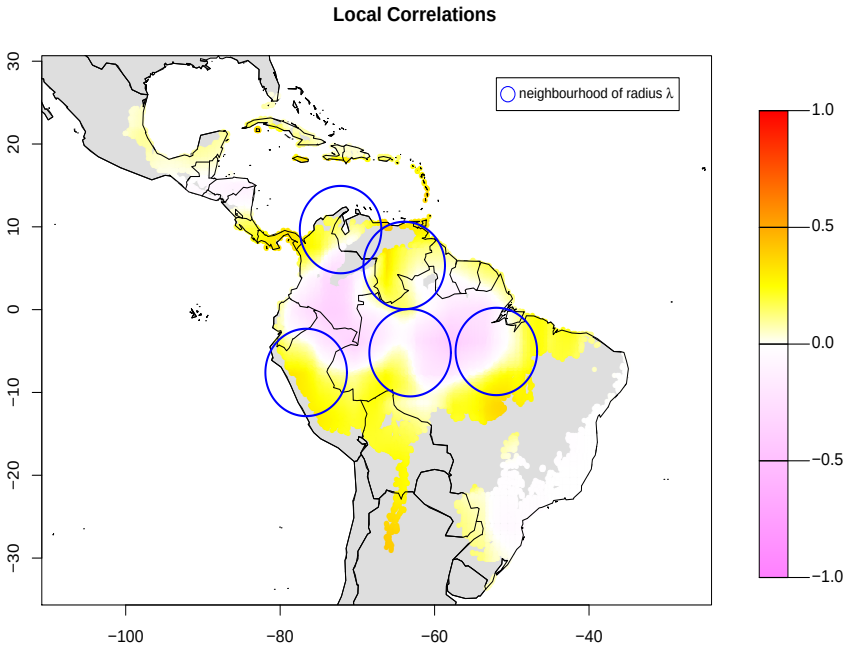


Figure 2: Local correlations between biodiversity and remoteness at locations $s_1, \ldots, s_N$ using a gaussian kernel with bandwidth $\lambda = 5.281$.

Assessing whether (and which) these correlations are significant, is the aim of this paper. The variogram is a useful tool to visualize the spatial autocorrelation of a process. It represents how the values of a variable X vary among different locations, and it is defined as the variance of the difference of X at two given locations s_i and s_j; $\gamma(u) = \frac{1}{2}\text{Var}[X_{s_i} - X_{s_j}]$, where $u = \|s_i - s_j\|$. In practice, we observe the empirical variogram, the collection of pairs of distances $u_{ij} = \|s_i - s_j\|$ between s_i and s_j, and their corresponding variogram ordinates $v_{ij} = \frac{1}{2}(X_{s_i} - X_{s_j})^2$.

The empirical variogram for biodiversity is plotted in Figure 3. The smoothed variogram $\hat{\gamma}$ is an estimate for γ and is plotted in red in Figure 3 [in Section 3.1 we define more formally γ and $\hat{\gamma}$]. At small distances, the variance among the values of X_{s_i} is small, indicating that the autocorrelation of locations close-by is high. As the distance increases, the correlation fades away (variance increases), which shows that locations sufficiently far apart are more independent.

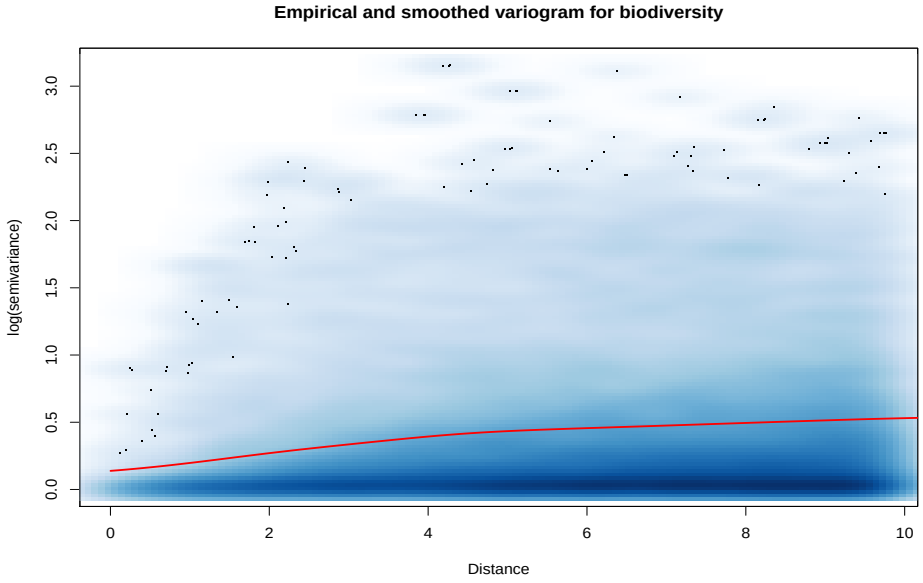


Figure 3: Empirical and smoothed variogram (in red) for biodiversity in a logarithmic scale.

2.1 The standard test and its limitations

If $(x_1, y_1), \ldots, (x_N, y_N)$ is an independent and normally distributed sample, the null distribution for the Pearson's correlation coefficient is

$$f_N(r) = \frac{(1-r^2)^{\frac{N-4}{2}}}{B[\frac{1}{2}, \frac{1}{2}(N-2)]}, \qquad \|r\| \leq 1 \tag{2}$$

A test for $\rho = 0$ is based on the statistic $t = \frac{(N-2)^{\frac{1}{2}} r}{(1-r^2)^{\frac{1}{2}}}$, which follows a Student's t distribution with $N-2$ degrees of freedom.

Variables biodiversity and remoteness are both spatially autocorrelated, and the pairs (X_{s_i}, Y_{s_i}) for $i = 1, \ldots, N$ are not independent. The classical assumption of independence does not hold and, as a consequence, the standard test produce incorrect and exaggerated results. Although we have a very large sample size, because of the strong autocorrelation, the effective dimension is much smaller (see Walther (1997)). The observed correlation will have more variance; behaving like a correlation with very small sample size. We illustrate this phenomenon in the following subsection.

2.1.1 Behaviour of the correlation coefficient under spatial autocorrelation

Let W_s be a stationary and isotropic gaussian random field in $\mathbb{R}^2$ ($s \in \mathbb{R}^2$) with autocorrelation function a member of the Matérn family:

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa} K_{\kappa}(u/\phi),$$

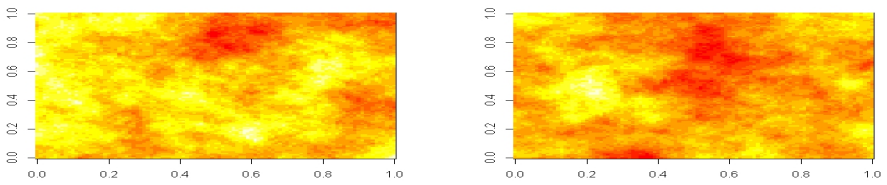
where $u = \|s_i - s_j\|$, $K_{\kappa}(\cdot)$ denotes a modified Bessel function of order κ, $\phi > 0$ is a scale parameter with the dimensions of distance, and $\kappa > 0$ is a shape parameter that determines the smoothness of the process. The variance of the process is $\sigma^2 = \text{var}(W_s)$.

Suppose X_{s_i} is generated by a stationary process

$$X_{s_i} = W_{s_i} + Z_i \tag{3}$$

where Z_i are mutually independent, identically distributed with zero mean and variance τ^2. The parameter τ^2 corresponds to the nugget variance, the measurement error variance.

Figures 4(a) and 4(b) are $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$, two independent realizations of this process observed at locations $\mathbf{s} = (s_1 \ldots s_N)$ with $s_i \in [0,1] \times [0,1]$, $\kappa = 0.5$ and $\phi = 0.30$ (simulated using R package RandomFields).

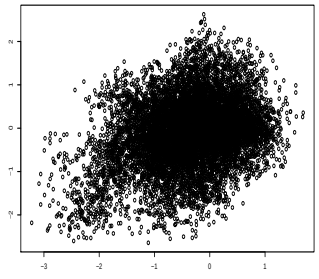


(a) gaussian random field $X_{\mathbf{s}}$ (b) random field $Y_{\mathbf{s}}$, independent of $X_{\mathbf{s}}$

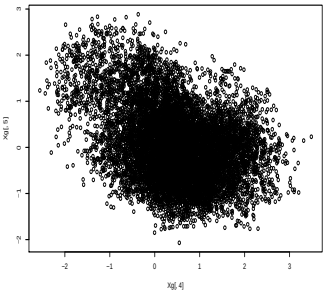
Figure 4: Two independent realizations of a gaussian random field with Matérn autocorrelation function and smoothing parameter $\kappa = 0.5$.

Figure 5(a) is the scatter plot of $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$, whereas Figure 5(d) is the scatter plot of two independent samples, each of them mutually independent (non-spatially correlated) and normally distributed. The correlation coefficient is much larger for $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$ ($r_{X_{\mathbf{s}},Y_{\mathbf{s}}} = 0.3$). However, Figure 5(b) is the scatter plot of two new independent realizations $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$, and now we observe a negative strong correlation. Figure 5(c) shows a third scatter plot, for another set of observed processes, with a correlation closer to zero. Due to the spatial component, the variance of the correlation coefficient is larger, in fact, the larger is κ, the larger the variance of the observed correlation. This is due to chance; because of the smoothness it is more likely that, just by chance, at a given region $\mathbf{s}^*$ of the support, $X_{\mathbf{s}^*}$ increases while $Y_{\mathbf{s}^*}$ increases as well (or decreases instead), contributing to a positive (or negative) linear correlation between $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$ (look at Figures 4(a) and 4(b) for an illustration of that).

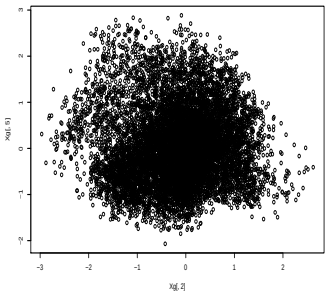
If we use the standard test to assess $r_{X_{\mathbf{s}},Y_{\mathbf{s}}} = 0.3$, the p-value is 0, although $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$ have been constructed to be independent of each other. A sample of the true null distribution of r_{X_s,Y_s} (obtained by simulation) is shown in Figure 6. Superimposed we plot the null distribution under the assumption of independence of the observations. The consequence of spatially autocorrelated data is a larger variance, which explains why it is more likely to reject when using the wrong null. If two close-by locations have similar values, one of the pairs in the sample is not giving new information; we know less about the distribution of r_{X_s,Y_s}, which is translated into less precision and an effective sample size smaller than N. Based on the true nulls, the probability of obtaining values of r as extreme as the observed ($r_{X_{\mathbf{s}},Y_{\mathbf{s}}} = 0.30$ and $r_{\text{ind}} = 0.01$) are 0.16 and 0.42 respectively, and there is no evidence to reject $\rho = 0$ in both cases.



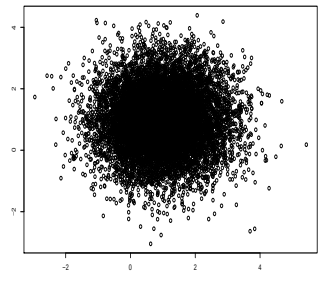
(a) Scatter plot of $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$ in Figure 4, the correlation coefficient is $r_{X_{\mathbf{s}},Y_{\mathbf{s}}} = 0.3$.



(b) Scatter plot of two new realizations of the same gaussian field, $r_{X_{\mathbf{s}},Y_{\mathbf{s}}} = -0.36$.



(c) Scatter plot of other two realizations, $r_{X_{\mathbf{s}},Y_{\mathbf{s}}} = 0.003$.



(d) Scatter plot of two independent samples, each mutually independent and normally distributed, $r_{\text{ind}} = 0.01$.

Figure 5: The observed correlation of two independent but spatially autocorrelated gaussian random fields $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$ has larger variance (5(a), 5(b) and 5(c)), due to chance, in comparison with two independent samples with no spatial autocorrelation (5(d)).

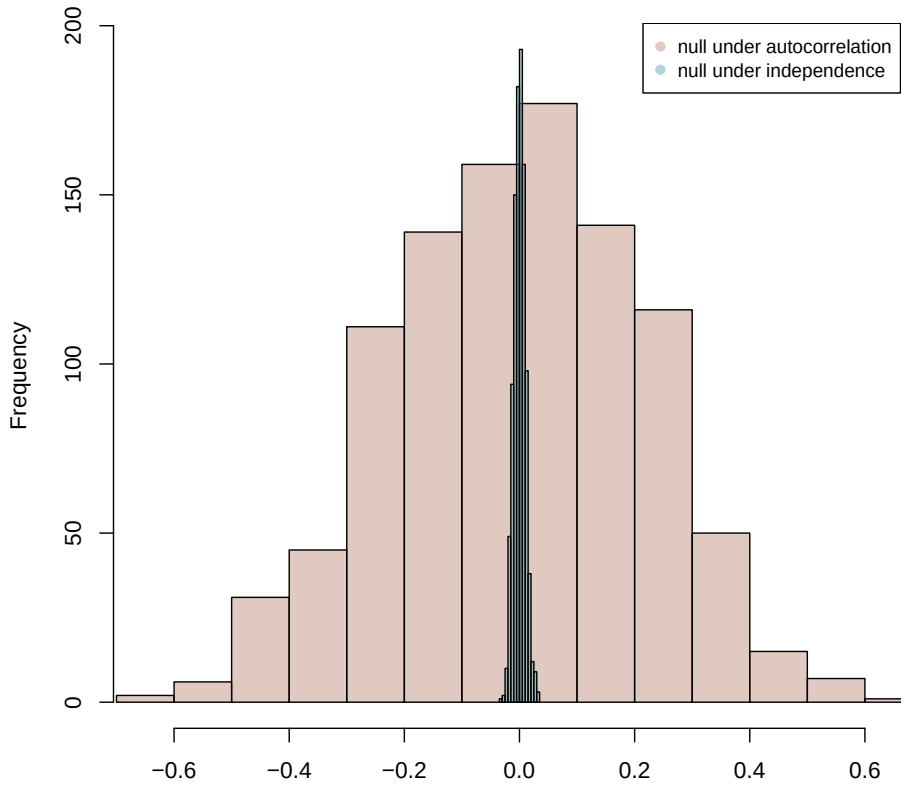


Figure 6: Empirical null distributions for the correlation coefficient between $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$, in contrast with the null distribution under the assumption of independence (no autocorrelation).

In the next section we propose a methodology to assess the correlation at each location by extracting (eliminating) the component of the correlation due to spatial location. One of the results of this approach is that now it is easy to produce a p-value map indicating which areas have high values for both biodiversity and remoteness, areas known to be good refuges.

3 Proposed methodology

We propose a method that approximately recovers the null distribution of r_{X_s,Y_s}. The simulation model also allows us to examine the null distribution of a much bigger variety of statistics and thus we will be able to answer other distribution-related questions. The following scheme summarizes the ingredients of the method.

Let $X_{\mathbf{s}}$ and $Y_{\mathbf{s}}$ be a realization of two processes that have been observed. Repeat the following two steps B times:

1. Permute the indices of $X_{\mathbf{s}}$ over $\mathbf{s}$, which we denote by $X_{\pi(\mathbf{s})}$; this means $X_{\pi(\mathbf{s})}$ and $Y_{\mathbf{s}}$ are independent.

2. Smooth and scale $X_{\pi(\mathbf{s})}$ to produce $\hat{X}_{\mathbf{s}}$, such that its variogram approximately matches the variogram of $X_{\mathbf{s}}$; i.e. the transformed variable $\hat{X}_{\mathbf{s}}$ has the same autocorrelation structure as $X_{\mathbf{s}}$.

Hence the variables $\hat{X}^1_{\mathbf{s}}, \ldots, \hat{X}^B_{\mathbf{s}}$ are independent of $Y_{\mathbf{s}}$ but with autocorrelation similar to $X_{\mathbf{s}}$. A sample from a null that approximates the true null distribution of r_{X_s,Y_s} is $\hat{r}_1, \ldots, \hat{r}_B$, where $\hat{r}_j = \text{cor}(\hat{X}^j_{\mathbf{s}}, Y_{\mathbf{s}})$.

Finally, using this sample as the reference null, the p-value to assess whether the observed correlation $r^*_{X_{\mathbf{s}},Y_{\mathbf{s}}} = \text{cor}(X_{\mathbf{s}}, Y_{\mathbf{s}})$ is significant, is $P(|r_{X_s,Y_s}| > |r^*_{X_{\mathbf{s}},Y_{\mathbf{s}}}|) = \frac{1}{B}\sum_{j=1}^{B} I[\hat{r}_j > r^*_{X_{\mathbf{s}},Y_{\mathbf{s}}}]$.

By permuting the indices of one variable, while destroying the independence necessary to recover the null, we also destroy the smoothness (spatial autocorrelation). Step 2 restores it, the following section focuses on this step.

3.1 Matching variograms

We smooth $X_{\pi(\mathbf{s})}$ over the domain $\mathbb{R}^2$ by fitting a local constant regression at each location s. The smoothing is achieved via a kernel $K_{\lambda_s}(s, s_i)$ that assigns weights to observations based on their distance from s. We fit the

following function using the R package locfit (Loader (1999)):

$$\hat{f}_\delta(s) = \frac{\sum_{\|s-s_i\| \leq \lambda_s} w_i X_{\pi_i}}{\sum_{\|s-s_i\| \leq \lambda_s} w_i}. \tag{4}$$

The weights are $w_i = K_{\lambda_s}(\|s-s_i\|)$ where $K_{\lambda_s}(x) = \exp\left[\frac{-2.5x^2}{2\lambda_s^2}\right]$ is a gaussian kernel, and the bandwidth λ_s of the kernel controls the smoothness of the fit. For a fitting point s, the nearest-neighbour bandwidth λ_s is chosen so that the local neighbourhood contains the $k = \lfloor N\delta \rfloor$ closest points to s in euclidean distance, where δ is a smoothing parameter in $(0, 1)$. Using a non-constant bandwidth reduces data sparsity problems, because in areas with fewer points the radius of the neighbourhood is incremented to include more neighbours. Only observations belonging to the ball $B_{\lambda_s}(s)$ (centered at s and of radius λ_s) are used to estimate $\hat{f}_\delta(s)$, so the gaussian kernel truncates at one standard deviation, and the factor 2.5 in K_{λ_s} scales the kernel accordingly.

If we evaluate the function $\hat{f}_\delta(s)$ at the original locations $s_1, \ldots, s_N$ we obtain a new spatially autocorrelated variable $X^\delta_{\mathbf{s}}$. The smoothing parameter δ is chosen such that the variogram of $X^\delta_{\mathbf{s}}$ is close to the variogram of the original $X_{\mathbf{s}}$. Formally, the problem reduces to choosing a variogram of the family

$$\beta\ \gamma(X^\delta_{\mathbf{s}}) + \alpha \tag{5}$$

that best approximates $\gamma(X_{\mathbf{s}})$.

Before moving forward, we need to define γ. The theoretical variogram of a stationary process X_{s_i} in (3) is:

$$\gamma(u) = \sigma^2(1 - \rho(u)) + \tau^2. \tag{6}$$

The function $\rho(u)$ is the autocorrelation function of W_{s_i}, typically a monotone decreasing function with $\rho(0) = 1$ and $\rho(u) \to 0$ as $u \to \infty$. Its most important feature is its behaviour near $u = 0$, and how quickly it approaches zero when u increases, which reflects the physical extent of the spatial autocorrelation in the process. When $\rho(u) = 0$ for u greater than some finite value, this value is known as the range of the variogram. The intercept τ^2 corresponds to the nugget variance, the conditional variance of each measured value X_{s_i} given the underlying signal value W_{s_i}. The asymptote $\tau^2 + \sigma^2$ corresponds to the variance of the observation process X_{s_i} (the sill). Figure 7 gives a schematic illustration.

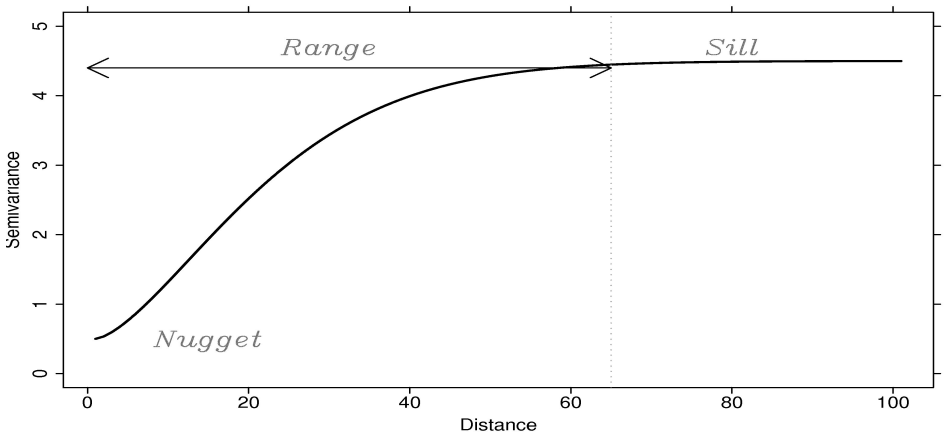


Figure 7: Typical semivariogram of a stationary spatial process: $\gamma(u) = \sigma^2(1-\rho(u)) + \tau^2$. The range is the distance u at which the autocorrelation function fades; $\rho(u) = 0$. The intercept τ^2 is the nugget variance, and $\tau^2+\sigma^2$ is the sill, the variance of the process.

Smoothed variogram $\hat{\gamma}$. Since $\gamma(u)$ is expected to be a smooth function of u, we smooth the empirical variogram (defined in Section 2) to improve its properties as an estimator of $\gamma(u)$, using the following kernel smoother:

$$\hat{\gamma}(u_0) = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_{ij} v_{ij}}{\sum_{i=1}^{N}\sum_{j=i+1}^{N} w_{ij}}. \tag{7}$$

It assigns weights that die off smoothly as distance to u_0 increases, with $w_{ij} = K_h(\|u_0 - u_{ij}\|)$ and $K_h(x) = \exp[\frac{-(2.68x)^2}{2h^2}]$, the gaussian kernel is scaled so that their quartiles are at $\pm 0.25h$, with h being the bandwidth (R function ksmooth). The variogram $\hat{\gamma}$ is obtained evaluating (7) at distances $\mathbf{u} = [u_1, \ldots, u_{100}]$, uniformly chosen within the range of distances u_{ij}. As an example, Figure 3 shows the empirical as well as the smoothed variogram (in red) for biodiversity, with bandwidth $h = 0.746$. Variograms in Figure 3 are truncated at distance $u = 10$, corresponding to the 25% percentile of all pairs of distances, because the precision of the estimate is expected to decrease as the distance increases, since a decreasing number of pairs are involved in the estimate.

How do we choose δ, α and β in (5)?

1. Given a permuted variable $X_{\pi(\mathbf{s})}$, for each $\delta \in \Delta$ we do the following:
   (a) Construct the smoothed variable $X_{\mathbf{s}}^{\delta}$ as indicated above.
   (b) Fit a simple linear regression between $\hat{\gamma}(X_{\mathbf{s}}^{\delta})$ and $\hat{\gamma}(X_{\mathbf{s}})$, where $(\hat{\alpha}_{\delta}, \hat{\beta}_{\delta})$ are the least-squares estimates.
2. The optimal δ^* is such that the sum of squares of the residuals of the fit is minimized, and so the estimates for (α, β) are $(\hat{\alpha}_{\delta^*}, \hat{\beta}_{\delta^*})$.

By varying the tuning parameter δ we obtain a family of variograms $\hat{\gamma}(X_{\mathbf{s}}^{\delta})$ with different shapes. The shape of the optimal variogram $\hat{\gamma}(X_{\mathbf{s}}^{\delta^*})$ is the closest to $\hat{\gamma}(X_{\mathbf{s}})$.

The smoothing has changed the scale of $X_{\mathbf{s}}$ (the smoother $X_{\mathbf{s}}^{\delta^*}$ is, the smaller the variance), in addition to the intercept (nugget variance) of $\hat{\gamma}(X_{\mathbf{s}})$, that is why we we need to transform $X_{\mathbf{s}}^{\delta^*}$ in the following way:

$$\hat{X}_{\mathbf{s}}^{\delta^*} = |\hat{\beta}_{\delta^*}|^{\frac{1}{2}} X_{\mathbf{s}}^{\delta^*} + |\hat{\alpha}_{\delta^*}|^{\frac{1}{2}} Z,$$

and ensure that the scale and intercept of $\hat{\gamma}(\hat{X}_{\mathbf{s}}^{\delta^*})$ match those of the target variogram $\hat{\gamma}(X_{\mathbf{s}})$, where Z is a vector of mutually independent and identically distributed Z_i's with zero mean and unit variance. Note that $\hat{\gamma}(\hat{X}_{\mathbf{s}}^{\delta^*})$ is a member of the family in (5).

From models (3) and (6), $|\hat{\beta}_{\delta^*}|\text{var}(X_{\mathbf{s}}^{\delta^*})$ is an estimate of σ^2, $|\hat{\alpha}_{\delta^*}|$ is an estimate of τ^2, and correspondingly $\text{var}(\hat{X}_{\mathbf{s}}^{\delta^*}) = |\hat{\beta}_{\delta^*}|\text{var}(X_{\mathbf{s}}^{\delta^*}) + |\hat{\alpha}_{\delta^*}|$ is an estimate of $\sigma^2 + \tau^2$.

To conclude, $\hat{X}_{\mathbf{s}}^{\delta^*}$ has been constructed to match the target variogram $\hat{\gamma}(X_{\mathbf{s}})$ in shape, scale, and intercept.

Note that the notation for $\hat{X}_{\mathbf{s}}^{\delta^*}$ has been used previously at the beginning of Section 3 simplified as $\hat{X}_{\mathbf{s}}$ $(\hat{X}_{\mathbf{s}}^1, \ldots, \hat{X}_{\mathbf{s}}^B)$.

3.2 Illustration of the method

The global correlation between biodiversity and remoteness is $\hat{r}_{X_{\mathbf{s}}, Y_{\mathbf{s}}} = 0.224$. The local correlations between both variables at locations $s_1, \ldots, s_N$ are plotted in Figure 2. If we apply our methodology, we can test whether the global correlation is significant, and provide a map of p-values for the local correlations. The algorithm, described in Section 3, returns $\hat{X}_{\mathbf{s}}^1, \ldots, \hat{X}_{\mathbf{s}}^B$ ($B = 1000$ proxies for $X_{\mathbf{s}}$) and the null distribution, which is plotted in Figure 8; the red line indicates the observed value $\hat{r}_{X_{\mathbf{s}}, Y_{\mathbf{s}}} = 0.224$. The p-value for the global correlation is 0.057. If we had used the classical test, the p-value would have been 0, rejecting the null hypothesis of the global correlation being equal to zero.

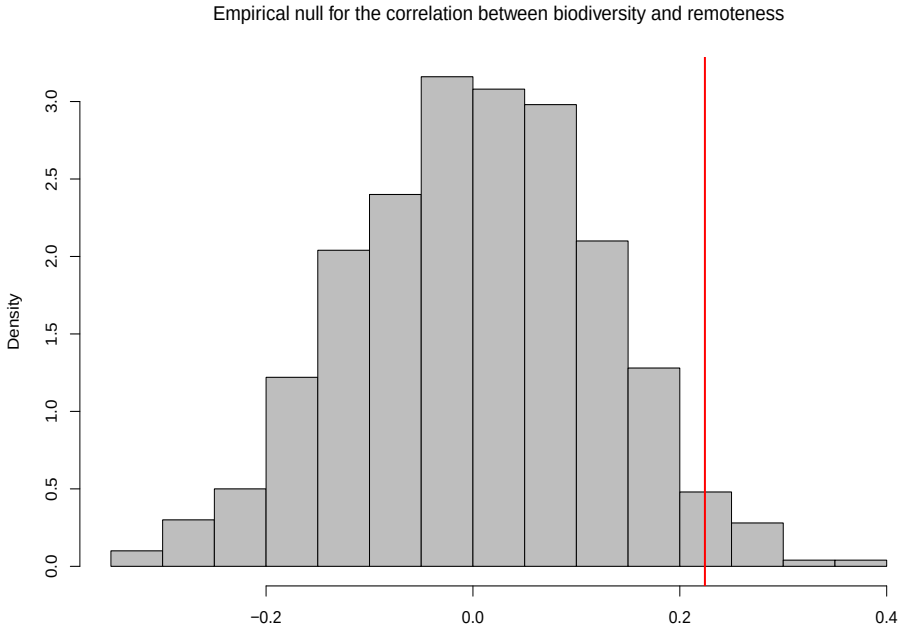


Figure 8: Empirical null distribution of the correlation between biodiversity and remoteness obtained with the proposed methodology, the red line corresponds to the observed correlation $\hat{r}_{X_{\mathbf{s}},Y_{\mathbf{s}}} = 0.224$.

To assess the local correlations consider the following. Each pair of variables $(\hat{X}^i_{\mathbf{s}}, Y_{\mathbf{s}})$ can be used to calculate a map of local correlations under the null hypothesis of independence, since $\hat{X}^i_{\mathbf{s}}$ is constructed to be independent of $Y_{\mathbf{s}}$, $i = 1, \ldots, B$ (the local correlations are calculated as described in Section 2).

As a result, we have a sample (of size B) distribution of local correlations for each s_j, $j = 1, \ldots, N_s$, and we can calculate a p-value for each location. Figure 9(a) is the map of p-values for the local correlations in Figure 2, which identify the areas with strong correlation. For comparison, Figure 9(b) are the p-values using the classical test, which contrasts with Figure 9(a), since most of them are significant.

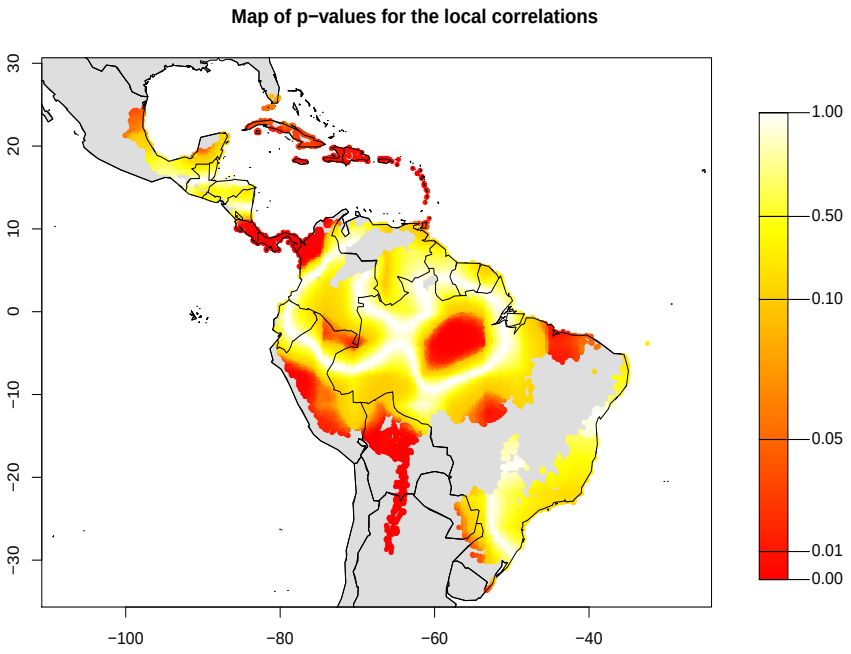
We illustrate in Figure 10 the variogram matching that takes place when smoothing and transforming a permuted $X^i_{\pi(\mathbf{s})}$ in a a way that it resembles the target variogram of $X_{\mathbf{s}}$ in Figure 3. Four different values for δ are used to smooth and scale $X^i_{\pi(\mathbf{s})}$. In this case, the best match between the target $\hat{\gamma}(X_{\mathbf{s}})$ (in black) and $\hat{\gamma}(\hat{X}^{i,\delta}_{\mathbf{s}})$ (in red) is reached when $\delta^* = 0.085$. The estimates $(\hat{\alpha}^\delta, \hat{\beta}^\delta)$ are obtained by linearly regressing $\hat{\gamma}(X^{i,\delta}_{\mathbf{s}})$ on $\hat{\gamma}(X_{\mathbf{s}})$. Figure 11 plots the residual sum of squares of this fit for different values of δ. We choose δ such that the sum of squares is minimized: $\delta^* = 0.085$.

4 Some evidence on the performance

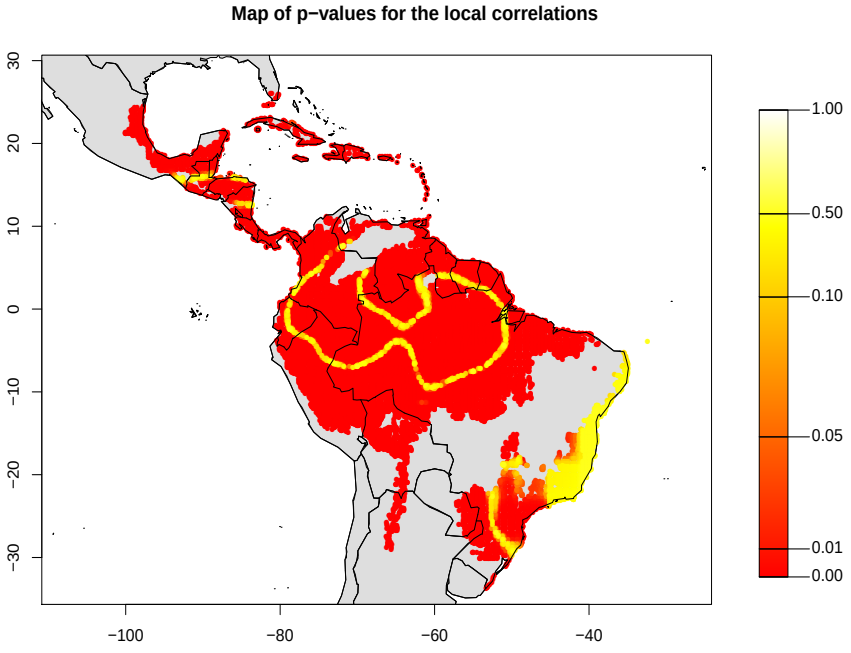
In this section we use simulations to demonstrate the effectiveness of our approach. By simulating random fields with a known theoretical model, we can recover the true empirical null distribution and compare it with the one obtained with our method, and therefore give some evidence of its performance.

Let X_s and Y_s be two independent gaussian random fields that follow model (3) with gaussian autocorrelation function (a particular case of the Matérn model when $\kappa \to \infty$), scale parameter $\phi = 0.3$, no nugget variance, variance $\sigma^2 = 1$ and mean $\mu = 0$.

We simulate the processes at locations $\mathbf{s} = (s_1 \ldots s_N)$ with s_i belonging to a grid $[0, 1] \times [0, 1]$, with 101 equally spaced points per interval, and $N = 10201$. A sample of the null for r_{X_s, Y_s} is plotted in Figure 12(a), and obtained by simulating several times the pairs $(X^i_{\mathbf{s}}, Y^i_{\mathbf{s}})$, $i = 1, \ldots, 1000$. To compare this null to the one given by our method, we consider one of the pairs $(X^i_{\mathbf{s}}, Y^i_{\mathbf{s}})$, and apply our method with bandwidths $\Delta = (0.1, 0.2, \ldots, 0.9)$ (in 75% of cases the optimal bandwidth is either $\delta = 0.2$ or $\delta = 0.3$). The resulting null is also plotted in Figure 12(a). It does recover fairly well



(a) Map of p-values obtained using the proposed methodology.



(b) Map of p-values using the classical test.

Figure 9: For each local correlation at location s_i in Figure 2, we associate a p-value assessing whether it is different from zero.

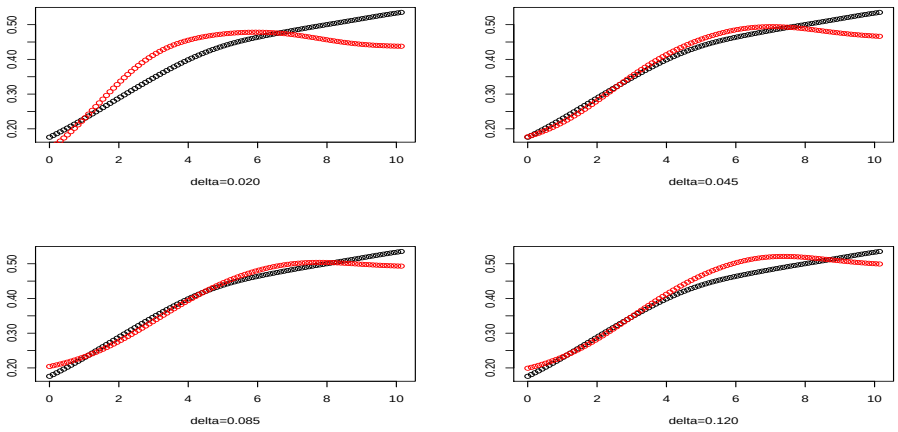


Figure 10: Matching that takes place between the target variogram of $X_{\mathbf{s}}$ (in black) and the variogram of $\hat{X}_{\mathbf{s}}^{i,\delta}$ (in red), variable result of permuting, smoothing and scaling $X_{\mathbf{s}}$, for 4 different values of the bandwidth δ.

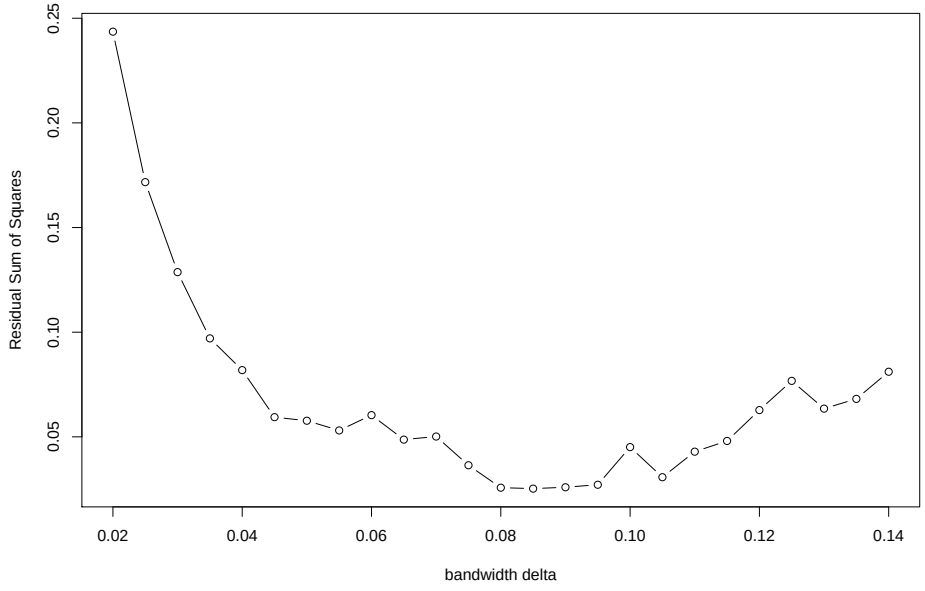


Figure 11: Residual sum of squares of linearly regressing $\hat{\gamma}(X_{\mathbf{s}}^{i,\delta})$ on $\hat{\gamma}(X_{\mathbf{s}})$, for different values of δ.

the true null, and the upper and lower limits of the corresponding 95% confidence intervals are very close.

Since the smoothing of the permuted variables is done with a gaussian kernel, we also simulate random fields with non-gaussian (Matérn with $\kappa = 0.5$) autocorrelation function to not favour our method when we match variograms. The results are very similar, specially in the tails, and are plotted in Figure 12(b).

4.1 Comparison with Clifford's method

In this section we compare our method to the one proposed in Clifford et al. (1989), where they suggest to estimate an effective sample size that takes into account the loss of precision due to spatial autocorrelation. The distribution of reference is $f_M(r)$ in (2) with $M-2$ degrees of freedom instead, where M is the effective sample size. Their approach is to equate σ_r^2, the variance of the sample correlation, to $\frac{1}{M-1}$, the variance of $f_M(r)$. An estimate for M is thus $\hat{M} = \lfloor 1 + \frac{1}{\hat{\sigma}_r^2} \rfloor$. They prove that,

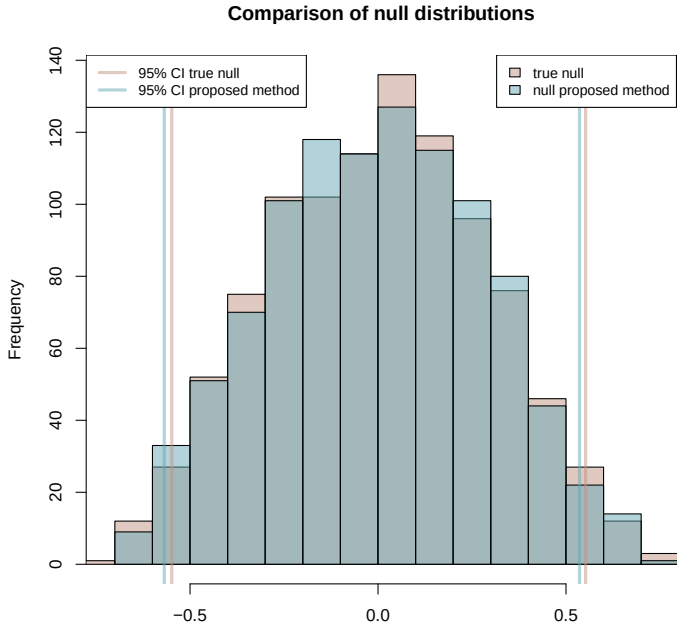
$$\sigma_r^2 = \frac{\mathrm{var}(S_{X_\mathbf{s}Y_\mathbf{s}})}{E(S_{X_\mathbf{s}}^2)E(S_{Y_\mathbf{s}}^2)}$$

to the first order, and under the assumption of normality (see Appendix in Clifford et al. (1989)), where $S_{X_\mathbf{s}Y_\mathbf{s}}$ is the sample covariance, and $S_{X_\mathbf{s}}^2$, $S_{Y_\mathbf{s}}^2$ are the sample variances of $X_\mathbf{s}$ and $Y_\mathbf{s}$. The term in the numerator is $\mathrm{var}(S_{X_\mathbf{s}Y_\mathbf{s}}) = \mathrm{trace}(\Sigma_{\xi_\mathbf{s}}\Sigma_{\eta_\mathbf{s}})$, where $\Sigma_{\xi_\mathbf{s}} = P\Sigma_{X_\mathbf{s}}P$, $\Sigma_{\eta_\mathbf{s}} = P\Sigma_{Y_\mathbf{s}}P$, $\Sigma_{X_\mathbf{s}}$ and $\Sigma_{Y_\mathbf{s}}$ are the covariance matrices of the processes $X_\mathbf{s}$ and $Y_\mathbf{s}$ respectively, $P = I - \frac{1}{N}\mathbf{1}\mathbf{1}'$ and $\mathbf{1}$ is a vector of 1's of dimension N.

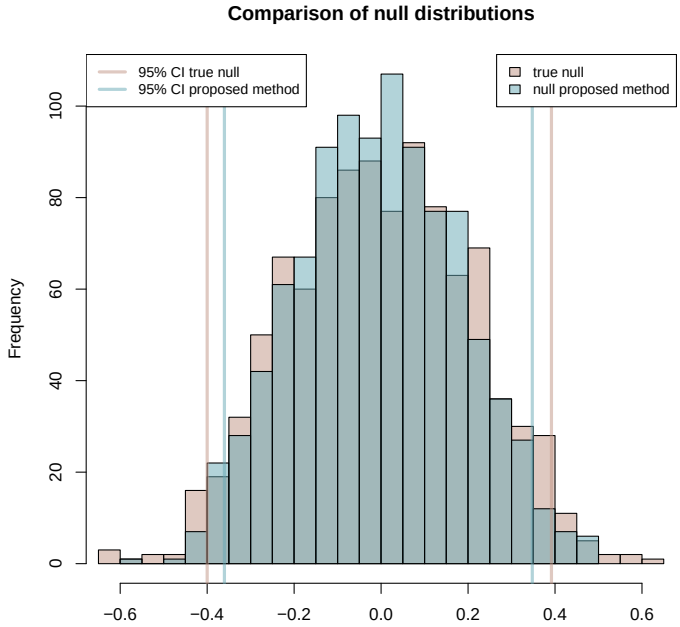
They impose a stratified structure on $\Sigma_{X_\mathbf{s}}$ and $\Sigma_{Y_\mathbf{s}}$ to estimate $\mathrm{var}(S_{X_\mathbf{s}Y_\mathbf{s}})$. More precisely, they assume that the set of all ordered pairs of elements of $\mathbf{s}$ can be divided into strata $S_0, S_1, S_2, \ldots$ so that the covariances within strata are constant. Then, the estimate for σ_r^2 is

$$\hat{\sigma}_r^2 = \frac{\sum_k N_k \hat{C}_{X_\mathbf{s}}(k)\hat{C}_{Y_\mathbf{s}}(k)}{N^2 S_{X_\mathbf{s}}^2 S_{Y_\mathbf{s}}^2}$$

where N_k is the number of pairs in stratum S_k and $\hat{C}_{X_\mathbf{s}}(k) = \frac{1}{N_k}\sum_{(i,j)\in S_k}(X_{s_i} - \bar{X}_k)(X_{s_j} - \bar{X}_k)$ is an auto-covariance estimate for stratum S_k. The number of strata is chosen as the number of bins used for the sample variogram of $X_\mathbf{s}$.



(a) Gaussian autocorrelation function.



(b) Matérn autocorrelation function.

Figure 12: Comparison of the true empirical null for $r_{X_{\mathbf{s}},Y_{\mathbf{s}}}$ (obtained by generating samples from model (3)) with the empirical null obtained by applying our methodology to one realization of the same model. The corresponding 95% Confidence Intervals are added to the plot.

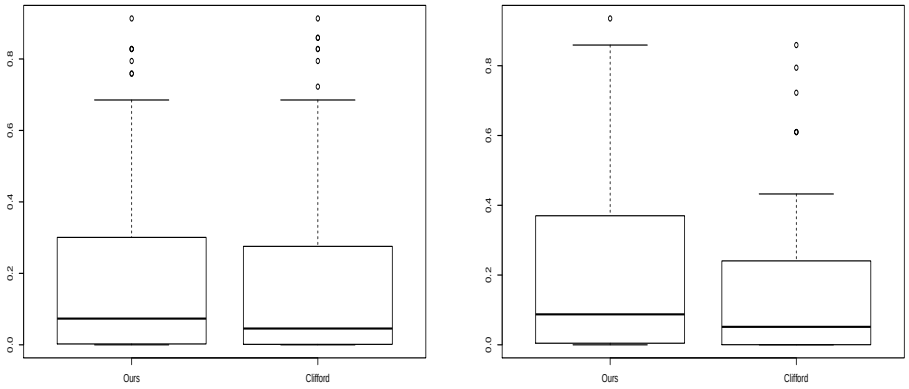
Hence, an approximation of the null distribution of r_{X_s,Y_s} is $f_{\hat{M}}(r)$. The statistic $t = \frac{(\hat{M}-2)^{\frac{1}{2}}\hat{r}}{(1-\hat{r}^2)^{\frac{1}{2}}}$ follows a Student's t with $\hat{M}-2$ degrees of freedom and is used to assess significance of a given correlation $\hat{r}$. We can also assess significance using a sample of $f_{\hat{M}}(r)$ as the reference null, that we can obtain generating independent and normally distributed random samples of size $\hat{M}$. The elements of the null are $r_i^{\text{Cl}} = \text{cor}(X_i, Y_i)$, where X_i and Y_i are independent random vectors of dimension $\hat{M}$, $i = 1, \ldots, 1000$.

We have now all the ingredients to carry out the following simulation experiment to compare both methods: (1) generate pairs $(X_{\mathbf{s}}^j, Y_{\mathbf{s}}^j)$ following model (3) with gaussian autocorrelation function, for $j = 1, \ldots, 100$, (2) apply both methods to each pair. As a result, we have two empirical null distributions for pair j: Clifford's null $\mathbf{r}_j^{\text{Cl}} = (r_{1j}^{\text{Cl}}, \ldots, r_{1000j}^{\text{Cl}})$ and our null $\mathbf{r}_j = (r_{1j}, \ldots, r_{1000j})$. We compare each null to the empirical true null of Figure 12(a) using a Kolmogorov-Smirnov test of comparison between distributions. The p-values of these tests are summarized in Figure 13(a) for each method. Both methods behave quite similar when the data is normal, although our method does slightly better.

Clifford's method is based on the assumption of normality. To see to which extent it is robust to deviations of normality, we generate data from the same gaussian random field and transform the marginal distribution. We generate gamma random numbers $(T_{\mathbf{s}})$ with scale and shape parameters equal to 2, and use its CDF $F_{T_{\mathbf{s}}}$ to transform the original observations as $Z_{s_i} = F_{T_{\mathbf{s}}}^{-1}(F_{X_{\mathbf{s}}}(X_{s_i}))$, $i = 1, \ldots, N$. The marginal distribution is now non-gaussian, the results of applying the same simulation experiment to these data are summarized in Figure 13(b). In this context, our method gives better results.

4.2 Type I error estimates

The type I error of the test should be equal to the significance level α. We use the nulls $(\mathbf{r}_1, \ldots, \mathbf{r}_{100})$ and $(\mathbf{r}_1^{\text{Cl}}, \ldots, \mathbf{r}_{100}^{\text{Cl}})$ to estimate the type I error rates associated to both methods. We generate 100 samples $(X_{\mathbf{s}}^i, Y_{\mathbf{s}}^i)$ under the null hypothesis and use respectively $\mathbf{r}_j$ and $\mathbf{r}_j^{\text{Cl}}$ to assess significance of $\hat{r}_i = \text{cor}(X_{\mathbf{s}}^i, Y_{\mathbf{s}}^i)$, for $i = 1, \ldots, 100$. Out of the 100 samples, the proportion of times the p-values are smaller than $\alpha = 0.05$ is an estimate of the type I error. We repeat the process for all nulls, $j = 1, \ldots, 100$, and average the results, which are found in Table 1. We see that our method provides better estimates in both cases, for gaussian and non-gaussian data.



(a) Gaussian random field.

(b) Non-gaussian random field.

Figure 13: Comparison, using a Kolmogorov-Smirnov test, of the true empirical null with the empirical nulls obtained by applying our methodology $(\mathbf{r}_1, \ldots, \mathbf{r}_{100})$ and Clifford's method $(\mathbf{r}_1^{\mathrm{Cl}}, \ldots, \mathbf{r}_{100}^{\mathrm{Cl}})$. The boxplots are the p-values of the tests for normal and non-normal samples.

Table 1: Estimated Type I error for ours and Clifford's method for gaussian and non-gaussian samples (%).

	our method	Clifford
gaussian	5.62	7.59
non-gaussian	5.8	7.92

Discussion

This paper aims to bring attention to the consequences of spatial autocorrelation when analyzing correlations, and propose a method that minimizes its effect. It provides a p-value for the global correlation of a spatial region, as well as a map of p-values that indicate the areas of high correlation, given a map of local correlations. It is of interest to explore correlation at both scales since association, as stated in Clifford et al. (1989), 'can exist simultaneously at a number of different geographical scales, and it is possible that negative association at small scales is swamped by positive association at large scales'.

The corresponding null distributions are recovered using Monte-Carlo methods. The procedure behaves well in practice, both for isotropic gaussian and non-gaussian random fields. The results are more precise than when the problem is approached by estimating effective sample sizes, as in Clifford et al. (1989)), and our method does not rely on the assumption of normality.

One of the consequences of autocorrelation is that increasing the resolution (getting more data) does not necessarily increase the power to find significance. Even if we have tons of fine resolution points, at some point we get no or little more information, since it is limited by the spatial autocorrelation of the variables. Consequently we would estimate the same significance if we used 20,000 fine resolution points, or a sample of 2,000 of them, for instance. In practice, it may be more important to focus on using methods that adjust for autocorrelation, than to focus on collecting a lot more data.

Acknowledgements

We thank Paul Switzer for some suggestions early on in this project.

References

Clifford, P., Richardson, S., and Hemon, D. (1989). Assessing the Significance of the Correlation Between Two Spatial Processes. *Biometrics* **45,** 123–134.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation

coefficient in samples from an indefinitely large population. *Biometrika* **10,** 507–521.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley.

Loader, C. (1999). *Local Regression and Likelihood*. Springer.

McCauley, D. J., McInturff, A., Nuñez, T. A., Young, H. S., Viladomat, J., Mazumder, R., Hastie, T., Dunbar, R. B., Dirzo, R., Ceballos, G., Power, E. A., Durham, W. H., Bird, D. W., and Micheli, F. (2012). In review. Nature's last stand: Identifying the world's most remote and biodiverse ecosystems. *Nature* .

Walther, G. (1997). Absence of Correlation between the Solar Neutrino Flux and the Sunspot Number. *Physical Review Letters* **79,** 4522–4524.