Spatially simplified scatterplots for large raster datasets

Bin Li, Daniel A. Griffith and Brian Becker

1. Introduction

A scatterplot is a statistical graph that uses a point symbol to depict a corresponding value pair in the Cartesian plane. Typically with two graphic layers, a point cloud, and global/local fitted lines, a scatterplot is able to reveal various aspects of the relationship between two variables, based upon which researchers can specify more realistic statistical models. However, while this graphic tool works well for small size data-sets, it loses its effectiveness when the number of data points becomes so large that points overlap, forming an indiscernible clump (Figure 1). Because each point symbol needs to be plotted on a graphic device, the time to render a scatterplot can be so long that interactive exploration becomes impossible. For example, a well-equipped MacBook Pro needed about 170 s to complete plotting 9000000 data points, for a pair of 3000-by-3000 raster variables. Because fitting a LOESS (LOCal regrESSion) line on these data with a typical 70% span would take hours, we resorted to fitting a more restrictive GAM (Generalized Additive Model) line, losing some exploration powers and requiring an additional 45 s. Any typical interactive moves, such as zooming and panning, would cause the entire canvas to repaint the graphic elements, leading to another painful wait. The problems of overplotting and excessive delay prevent a scatterplot from being used in exploring large size data-sets, a common task in spatial data analysis and modeling. Cressie, Olsen, and Cook (1997) noted that statistical tools initially designed for small data-sets may be challenged by and ultimately fail with massive data-sets. To overcome the overplotting problem, researchers have developed numerous techniques, including alpha blending (Figure 2), kernel smoothing (Figure 3), contouring (Figure 4), and binning (Figures 5 and 6) (Unwin, Theus, and Hofmann 2007; Wand and Jones 1994; Carr et al. 1987). Among them, the nested lattice binning, by Carr et al., and the binned kernel smoothing, by Wand and Jones, are particularly effective in stratifying massively overlapped points, hence revealing the density structure of a scatterplot. Rendering speeds are also very fast. These benefits, however, are achieved at the cost of over-generalization, as a bin size needs to be sufficiently large. Equivalent to categorization, in most cases, binning continuous variables leads to a loss of information (Dinero 1996; MacCallum et al. 2002).

Limited options seem to remain – that is, if data are independent and each data point contributes a unique amount of information. Fortunately, this data characterization is not the case for most spatial data-sets for which data often are highly autocorrelated. The two variables used for generating Figure 1, for example, have Moran coefficients of 0.92 (X) and 0.90 (Y), respectively. Such high spatial autocorrelation indicates a large amount of...
calculating the effective sample size of a given spatial data. Exploratory applications of Griffith’s methods show a 99% reduction in sample size of a remotely sensed data-set.

Based on an approximation of effective sample size, we can extract a subset from an original data-set through spatial sampling while maintaining much of the data-set’s information. The resulting simplified data-set can be used to generate a scatterplot as well as fitted lines. To achieve such a goal, however, we need to overcome a number of methodological as well as computation challenges. This paper reports our initial experiments for developing a spatially simplified scatterplot for large raster data-sets. The remaining sections provide a review of related work, our research design, the outcomes, and a discussion of the remaining issues.

redundancy in the data, providing an opportunity to achieve a substantial reduction in data size. Although data generalization has long been a research domain in geography, the focus has been on cartographic representation (Monmonier 1983; Buttenfield and McMaster 1991; McMaster and Shea 1992; Mackaness, Rua, and Sarjakoski 2011). The theoretical foundation to exploit this unique characteristic of spatial data was not established until recently, when Griffith (2003, 2005) systematically introduced the mathematical formula for

Figure 1. Scatterplot and a GAM fitted line for Normalized Difference Vegetation Index (NDVI) (Y) and ETM + Band 4 (X). Data are standardized. The R package mgcv was used to fit the GAM (Wood 2006; Hastie and Tibshirani 1990).

Figure 2. Alpha blending scatterplot of a subset (1000 × 1000) of the same data in Figure 1, with alpha value set to 0.05. The alpha value specifies the level of transparency for the data points, ranging from 1 for opaque to 0 for black. Even at such a low alpha level, excessive overplotting remains a problem.

Figure 3. Two-dimensional binned kernel smoothing scatterplot. Notes: Black dots are bins with a single data point. Data are the same as Figure 2.

Figure 4. Scatterplot with contours. Due to the lack of sufficient memory to render the plot with the large data-set in Figure 2, 10,000 data points are randomly generated from a standard normal distribution.
2. Effective sample size

Spatial data-sets often exhibit strong spatial autocorrelations, which means a large amount of information contained in the data are redundant. Accordingly, a subset of the data exists containing independent information equivalent to that in the original data-set. Such a subset can be considered the effective sample, which was characterized by Clifford, Richardson, and Hémon (1989) as effective degrees of freedom. Recently, Griffith (2005) systematically developed theories and methods for estimating the effective sample size for georeferenced data, which can serve as an index of information content. This section describes the relevant concepts and methods furnishing the experiment summarized in this paper.

A georeferenced Gaussian variable $Y$ can be expressed as:

$$Y = \mu 1 + \varepsilon = \mu 1 + V^{-\frac{1}{2}}e^*$$  \hspace{1cm} (1)

where $Y$ is an $n$-by-$1$ vector for a normally distributed random variable; the population mean; $\varepsilon$ a vector of 1 with size $n$; $e$ the spatially autocorrelated error; $e^*$ the independently and identically distributed unautocorrelated error ($N(0, \sigma^2_e)$) and $V$ the $n$-by-$n$ matrix containing the covariation structure of the error, which takes different forms. For the simultaneous autoregressive model (SAR), which is used in this study, the matrix $V$ is expressed as $[(I - \rho W)^T(I - \rho W)]$, where $W$ is the row-standardized spatial weight matrix; $I$ is an $n$-by-$n$ identity matrix; and $\rho$ is the spatial autocorrelation parameter. Without considering spatial autocorrelation, the variance of the sample mean of $Y$ is formulated as:

$$E(\hat{\sigma}_y^2) = \frac{\sigma^2_e + V^{-1}1^T1}{n}$$

where superscript $T$ denotes matrix transpose, and which can be rearranged as:

$$E(\hat{\sigma}_y^2) = \frac{\text{TR}(V^{-1})\sigma^2_e}{\text{TR}(V^{-1}) + 1} \frac{1}{n}$$  \hspace{1cm} (2)

where $\text{TR}$ denotes matrix trace; the term $\text{TR}(V^{-1})/n$ is the variance inflation factor; and the denominator becomes the general expression of effective sample size,

$$n^* = \frac{\text{TR}(V^{-1})}{1^T V^{-1} 1} n$$  \hspace{1cm} (2A)

When no spatial autocorrelation is present, $V = I$, and $n^* = n$. When the observations are spatially autocorrelated, $n^* < n$. Depending on the objectives of geographic inquiries, $n^*$ takes different forms. If the interest is on estimating the mean of a single variable, the effective sample size can be approximated with the following empirical equation:

$$n^* \approx n \times \left[ 1 - \frac{1}{1 - e^{-1.92549 n}} \left( 1 - e^{-2.12373(\hat{\rho} + 0.20024 \sqrt{\hat{\rho}})} \right) \right]$$  \hspace{1cm} (3)

where $n$ is the size of the original data-set, and $\hat{\rho}$ is the spatial autocorrelation parameter estimate for a SAR model with the following form (Gelfand et al. 2010):

$$Y = \mu 1 + (I - \rho W)^{-1}\varepsilon$$  \hspace{1cm} (4)

which is the same specification as Equation (1). For a remotely sensed image of size 100-by-100, $n = 10,000,$
with an SAR spatial autocorrelation estimate $\hat{\rho} = 0.95$, the effective sample size is approximately 183. In other words, less than 2% of the original data points contain all of the attribute information in the original data set.

When the research interest is on estimating bivariate relationships, the effective sample size $n_{XY}^*$ is a function of the effective sample sizes of the respective variables $(n_X^*, n_Y^*)$, and needs to account for the spatial autocorrelation estimates of both variables ($\hat{\rho}_X, \hat{\rho}_Y$) as well as the correlation coefficient ($\hat{\rho}_{XY}$) between them:

$$n_{XY}^* \approx 1 + \frac{n - 1}{n} \times \left[ \frac{n_X^* + n_Y^*}{2} + \sqrt{2(\hat{\rho}_X + \hat{\rho}_Y)^2 \ln (1 - \hat{\rho}^2)} \right]^{0.04} \sqrt{n_X^* n_Y^*}$$

(5)

For the same hypothetical image in the previous example, denoted as variable $Y$, if we are interested in the correlation coefficient between it and another data set, denoted as variable $X$, of the same extent and resolution, denoted as $Y^*$, the effective sample size would be around 652, about 7% of the original data.

Griffith (2005) also derived the exact formula for the cases of two and more sample means as well as the equivalent specifications of effective sample sizes based on geostatistical models. Unlike the cases of a single mean, the calculations of effective sample sizes for estimating joint means involve the variances of the variables and do not have a simplified empirical form similar to Equations (3) and (5). Therefore, finding the exact effective sample size requires substantially more complex computations such as matrix multiplications and inversions, which becomes a barrier for large spatial data sets. Similarly, semivariogram-based models not only require the calculation of a full distance matrix, but also the fitting of a variogram; both processes need large amounts of RAM and CPU time that easily exceed the capacity of conventional research computers. Although all spatial statistical models more or less face similar computational challenges, methods recently were developed to alleviate some of the key bottlenecks, which substantially shortens the calculations of Equations (3) and (5) on conventional desktop computers.

### 3. Computational issues

The SAR spatial autocorrelation coefficient $\hat{\rho}$, which appears in the calculations of effective sample size, can be estimated by solving the following maximum likelihood normal equation (Griffith 2015):

$$\frac{\partial \ln(L)}{\partial \rho} = \frac{\sum_{j=1}^{n} n_j}{n} \left[ \frac{1}{1 - \rho^2} \right] - \frac{(Y - \hat{\mu})^T (W^T + W - \rho^2 W^T W) (Y - \hat{\mu})}{2n\hat{\sigma}^2} = 0$$

(6)

with the constant mean and variance specified as:

$$\hat{\mu} = 1^T (I - \rho W) Y / [n(1 - \rho)]$$

$$\hat{\sigma}^2 = [Y - \rho W Y - (1 - \rho) \mu]^T [Y - \rho W Y - (1 - \rho) \mu] / n$$

where $\lambda_j$ is the $j$th eigenvalue of matrix $W$. Finding the eigenvalues is computationally intensive, which for a long time had been a major barrier to applying spatial autoregressive modeling, before a series of algorithms were developed to exploit the special characteristics of different types of matrices (Pan and Chen 1999; Bivand, Hauke, and Kossowski 2013). Topologically based spatial weights matrices for raster data, in particular, often are symmetrical and sparse, which provides an opportunity for substantial optimization. With a sufficient amount of RAM, a desktop computer can find $\hat{\rho}$ for a raster data set as large as 3000-by-3000, with a 30,000$^2 \times 30,000^2$ corresponding $W$ matrix, although it would still take hours. If the data structure is a regular square tessellation and the data cover an entire rectangular region, such as the case for many remotely sensed data and raster data sets, the eigenvalues of the spatial weights matrix can be approximated by (Griffith 2000):

$$-1 \leq \lambda_{hk} = \text{sign} \left\{ \frac{\cos \left( \frac{hk}{P-1} \right) + \cos \left( \frac{k Q}{Q+1} \right)}{2Q} \right\} \leq 1$$

(7)

where $P$ and $Q$ denote the number of rows and the number of columns; $h = 0, 1, 2, \ldots, P-1$ and $k = 0, 1, 2, \ldots, Q-1$; and sign represents the sign of an eigenvalue, either positive or negative. $\gamma$ is the inflation exponent which approaches 1 as $P$ and $Q$ increase to infinity. Based on simulations with a range of $P$ and $Q$ values, Griffith (2015) put forth the following approximation of $\gamma$:

$$\hat{\gamma} = 1 - 0.1466 \left( \frac{1}{P+0.04} + \frac{1}{Q+0.04} \right) - 0.7784 \frac{(P-0.38)(Q-0.38)}{PQ}$$

(8)

Griffith (2004a) subsequently showed that, for a regular square tessellation, the Jacobian term $\sum_{j=1}^{n} \ln \left( 1 - \rho \lambda_j \right)$ in Equation (6) can be approximated by the following equation:

$$- \frac{\sum_{j=1}^{n} \ln \left( 1 - \rho \lambda_j \right)}{PQ} \approx 2\omega \ln(\delta) - \omega \ln(\delta + \rho) - \omega \ln(\delta - \rho)$$

(9)

Another simplification is possible (Griffith 2004b):

$$\frac{- \sum_{j=1}^{n} \ln \left( 1 - \rho \lambda_j \right)}{PQ} \approx \ln \left( 1 + q_2 \rho^2 + q_4 \rho^4 + q_6 \rho^6 \right)$$

(10)
The coefficients ω, δ, q_2, q_4, and q_∞ in (9) and (10) can be calibrated using Equation (7) for raster data of different sizes and different contiguity specifications. With these empirical coefficients, no eigenvalues need to be numerically calculated. The polynomials in both (9) and (10) are relatively simple to solve, making the calculation of the Jacobian term very fast and highly scalable.

4. Spatial resampling

Selecting a subset of the original data while maintaining the geographic extent is an act of resampling, although it may be different from the common use of the term that refers to such procedures as bootstrap and jackknife, which are performed to assess the precision of statistical estimates (Chun and Griffith 2011). Our primary goal of resampling, meanwhile, is to extract a judiciously selected subset to represent the key statistical and graphical characteristics of the original data as shown by a scatterplot. Once the sample size is determined through an effective sample size calculation, methods of spatial sampling can be adopted to resample the original data-set.

Research about spatial sampling is abundant, as reviewed by Wang et al. (2012). In the classic scheme, four broad categories of sampling exist—random, systematic, stratified, and clustering; all four designs seek to represent some aspects of the statistical characteristics of a population. For raster data, simple random sampling uses a pseudo-random number generator to identify random coordinate pairs (row and column) within a given geographic extent. Systematic sampling, also referred to as regular sampling, selects data points at equal geographic intervals determined by the sample size. Spatially stratified sampling extracts data within geographic regions. In a simple stratified random design, stratification is with a regular tessellation whose number of cells equals the sample size, and a single data point is selected within each stratification unit. In a simulation study, Chun and Griffith (2013) report that while all three sampling schemes are able to produce subsets containing the mean and variance of the original data, the hexagon stratified random sampling design produces the best results. More sophisticated methods are needed as spatial heterogeneity increases.

GIS and remote sensing software programs often provide a set of functions for resampling raster data. The primary objective is to enable integrated analysis of data at different spatial resolutions. Common methods include nearest neighbor, bilinear interpolation, and cubic convolution (ERDAS, Inc. 1999). In nearest neighbor resampling, the value of the source cell closest to the center of a target cell is selected. This sampling often is used for nominal data. Bilinear interpolation generates a new value for a target cell through a weighted distance average of its four nearest source cells. Cubic convolution finds a new value by fitting a spline through 16 nearest source cells. Both bilinear interpolation and cubic convolution result in a certain level of smoothing, resulting in loss of information, and undesirable alteration of the original data.

5. Research design

Our objective is to find a solution to the problems of overplotting and excessive delays in generating scatterplots for large spatial data-sets, particularly for remotely sensed data and their derivatives. We hypothesize that a small subset of the original georeferenced data-set contains sufficient information for producing a scatterplot with the capacity for visual exploration equivalent to the original plot, because raster data in general and remotely sensed data in particular often exhibit substantial spatial autocorrelation. We contend that the effective sample size provides the critical guideline for generating such a subset. Classic spatial sampling designs (random, systematic, and stratified) will be evaluated individually and in combination for their effectiveness in uncovering the effective samples.

A simulated data-set and a remote sensing data-set were used in an experiment. The simulated data-set has two randomly generated raster variables in a 100-by-100 rectangular tessellation. The two variables have a nonlinear relationship (cubic polynomial), and each has a significant level of spatial autocorrelation (Table 1, Figures 7 and 8). The simulated data-set was used to obtain a more controlled result, which furnishes a guide to real data applications. The real-world data-set includes the NDVI (Normalized Difference Vegetation Index) and Band 4 image from the Landsat Enhanced Thematic Mapper, acquired on 16 July 2002, near Palisades, Idaho. The NDVI is a popular measure of plant “greenness,” and ETM Band 4 captures the spectral range (wavelength) from 0.77 to 0.90 mm and is characterized as the “near infrared band (NIR)” that is sensitive to biomass content. Therefore, the two variables should be highly positively correlated. The size of the image is 3000-by-3000 pixels, with a spatial resolution of 30 m (Figure 9).

Table 1 summarizes the statistical characteristics of these two variables, in which ρ is the SAR spatial autocorrelation coefficient and ρ_{xy} is the correlation coefficient.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Moran coefficient</th>
<th>ρ</th>
<th>ρ_{xy}</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>-0.0193</td>
<td>8.2652</td>
<td>0.033</td>
<td>3.000</td>
<td>0.7705</td>
<td>0.9658</td>
<td>0.4688</td>
</tr>
<tr>
<td>Y</td>
<td>176.1985</td>
<td>3689.0297</td>
<td>0.443</td>
<td>7.432</td>
<td>0.2221</td>
<td>0.5136</td>
<td>0.4688</td>
</tr>
</tbody>
</table>
The experiments were conducted with the following steps. Using the simulated data, we first calculated the SAR spatial autocorrelation coefficients for both variables, and then computed the effective sample size for each variable as well as the effective sample size for the bivariate correlation case. The latter is used to perform a spatial sampling and generate a simplified scatterplot, overlaid with a local regression line and a linear regression line, along with the same visual features for the original data. Both data-sets may be described by a range of statistics, including the mean, standard deviation, skewness, kurtosis, Moran coefficient, and SAR spatial autocorrelation coefficient. A Kolmogorov-Smirnov test evaluates if the resampled set and the original set represent the same population. Resampling, comparison between the spatially simplified scatterplot with the original plot, as well as statistical summary and testing are repeated for three spatial sampling designs: random, systematic, and hexagon stratified random. Results from all three schemes are evaluated, followed by subsequent experiments with the remote sensing data-set using the same procedures.

6. Results

Given $\rho_x = 0.9658$, $\rho_y = 0.5136$, $\rho_{xy} = 0.4688$, and $n = 10,000$, the effective sample sizes for the simulated data are $n^*_x = 124$, $n^*_y = 2832$, and $n^*_{xy} = 2785$. Clearly, effective sample size is very sensitive to the SAR coefficient – at a lower level of spatial autocorrelation, variable $Y$ has far fewer redundant data points than $X$. Also worth noting is that the sample sizes need to be slightly adjusted for regular sampling and hexagon stratified random sampling in order for the resampling to cover the entire geographic region. Figure 10 depicts the sample sites for all three schemes, with $X$ in the background for reference purpose (except for the hexagon case).

Table 3 summarizes the resampling results with the simulated data. The large $p$-values and small $D$ for the K–S tests for all three resampled sets are good indications of similarities with the original data. In addition, the confidence intervals of the means, standard deviations, and correlation coefficients, although varying in width, all contain the corresponding statistics of the original data. Skewness and kurtosis, though not affecting the visual characteristics of the scatterplot, are included for completeness, and show some variation among different sampling schemes. For skewness, random sampling and hexagon stratified random sampling result in relatively large deviations from the original data. For kurtosis, the stratified sample set stands out as an outlier. Overall, however, statistical measures indicate the resampled subsets from all three spatial sampling schemes are close representations of the original data, with hexagon stratified random sampling showing a better performance.

Figures 11 and 12 confirm the statistical summaries. The point clouds of the resampled data capture the general distribution characteristics of the original data, but with a substantial reduction in overplotting. The three sampling schemes produce similar patterns, but regular sampling seems to be slightly more representative, being
Factors. First is the normality assumption for the variables when specifying the effective sample sizes. For the two simulated variables, $X$ was generated from a normal density function but $Y$ was derived from a polynomial of $X$ that results in a departure from normality. Furthermore, we used relatively simple sampling schemes that are unable to remove redundancy completely. Both, however, tend to increase the effective sample size without negatively affecting the representativeness of resulting simplified scatterplots.

The existence of substantial levels of spatial autocorrelation in the resampled set does lead us to speculate the possibility of performing a second round reduction. For example, given $\rho_x = 0.8539$, $\rho_y = 0.3490$, $\rho_{xy} = 0.4533$, and $n = 2438$, a new effective sample sizes becomes $n_{xy}^* = 1326$. A simple random selection of 1326 data
points (Figure 13) from the 2438 samples extracted in the previous step (through hexagon stratified random sampling) shows some interesting results, as summarized in row Hex2 of Table 3 and Figure 14. The K–S statistics show the resulting subset, though only half of the size of the initial resampled set, still constitutes a good representation of the original data. The means, standard deviations, and correlation coefficient start to deviate from those of the population but the amount of departure is very small (Figure 14). We repeated the experiments for the resampled sets from the other two schemes, and the results were similar.

For the remote sensing data set, we adopted a slightly different two-stage resampling design where hexagon stratified random sampling was first performed with the correlation-based joint effective sample size, followed by a second round of resampling using the random, systematic, and stratified schemes. Initial experiments with hexagon stratified random sampling showed advantages over the other methods, similar to the published research that suggested it to be a superior design (Chun and Griffith 2013). Rectangular stratified sampling was included for comparison.

Table 4 summarizes the statistical characteristics of the two-stage resampling experiment. At stage one, the effective sample size was calculated based on the SAR spatial autocorrelation coefficients, $\rho_{\text{NIR}} = 0.9839$, $\rho_{\text{NDVI}} = 0.9852$, $n = 9000,000$, $\rho_{\text{NIR-NDVI}} = 0.2281$, and $n_{\text{NIR-NDVI}} = 200,800$. More digits for $\rho_{\text{NIR}}$ and $\rho_{\text{NDVI}}$ are included in the calculation because for such a large $n$, a small increase in precision in these coefficients would result in a substantial change in the effective sample size. The number of corresponding hexagons constructed was 174,212, within each of which one data point was randomly chosen for each variable. As the two rows “Hex stratified” in Table 4 show, the summary statistics (mean, standard deviation, skewness, and kurtosis) from this much smaller subset (2% of the original data) are almost identical to those of the original data, with a strong confirmation by the K–S tests. The correlation coefficient slightly deviates. This outcome is expected because there are clearly non-linear components in the relationship that may become more dominating in the resampled set.

Figure 15 shows an overlay of the point clouds and the fitted lines (linear, GAM/LOESS, and quartic
A new sample size was calculated as approximately 25% of the initial resampled set and 0.5% of the original data. Actual sampling sizes for regular and stratified schemes vary slightly, due to the geometric restrictions to cover the geographic region.

As Table 4 shows, the stage-two resampled set from all three spatial sampling methods produced summary statistics virtually identical to not only its parent data-set (initial resampling), but also the original set; the K–S tests again confirm the data-sets are from the same population. Although the SAR autocorrelation coefficients were only reduced by a small amount, the Moran coefficients have dropped considerably. The simplified scatterplots from all three stage-two resampled data-sets were almost identical, so we selected the one generated from regular sampling as representative. Figure 16 depicts an excellent portrayal of the original point cloud from the parent data-set and all fitted lines are overlapped with little deviation from those of the original data-set.

Despite a substantial reduction in sample size and a general increase in spacing among data points, spatial autocorrelations in the resampled set are still quite strong, which offers the opportunity for further reduction. Based on the SAR coefficients and the correlation coefficient ($\rho_{\text{NIR}} = 0.8084$, $\rho_{\text{NDVI}} = 0.8579$, $n = 174,212$, $\rho_{\text{NIR-NDVI}} = 0.2571$), a new sample size was calculated as 42,128, about 25% of the initial resampled set and 0.5% of the original data. Actual sampling sizes for regular and stratified schemes vary slightly, due to the geometric restrictions to cover the geographic region.

As Table 4 shows, the stage-two resampled set from all three spatial sampling methods produced summary statistics virtually identical to not only its parent data-set (initial resampling), but also the original set; the K–S tests again confirm the data-sets are from the same population. Although the SAR autocorrelation coefficients were only reduced by a small amount, the Moran coefficients have dropped considerably. The simplified scatterplots from all three stage-two resampled data-sets were almost identical, so we selected the one generated from regular sampling as representative. Figure 16 depicts an excellent portrayal of the original point cloud from the parent data-set and all fitted lines are overlapped with little deviation from those of the original data-set.

With less than 45,000 data points, it took a conventional polynomial). The point cloud from the resampled set captures the general visual structure of the original data, with much less overlap. The GAM/LOESS lines are virtually identical, except for the upper region of the NIR ($X$ axis) variable. The OLS fits show similar patterns. The quartic polynomial fits are a complete overlap, while the cubic polynomial from the resampled set provides a good generalization of the relationship, and may alleviate the possible problem of overfitting in the lower and upper regions. These results clearly show the spatially simplified scatterplot, using a fraction of the original data, can sufficiently represent the key characteristics of the original scatterplot, making it possible to perform interactive exploration.

Despite a substantial reduction in sample size and a general increase in spacing among data points, spatial autocorrelations in the resampled set are still quite strong, which offers the opportunity for further reduction. Based on the SAR coefficients and the correlation coefficient ($\rho_{\text{NIR}} = 0.8084$, $\rho_{\text{NDVI}} = 0.8579$, $n = 174,212$, $\rho_{\text{NIR-NDVI}} = 0.2571$), a new sample size was calculated as 42,128, about 25% of the initial resampled set and 0.5% of the original data. Actual sampling sizes for regular and stratified schemes vary slightly, due to the geometric restrictions to cover the geographic region.

As Table 4 shows, the stage-two resampled set from all three spatial sampling methods produced summary statistics virtually identical to not only its parent data-set (initial resampling), but also the original set; the K–S tests again confirm the data-sets are from the same population. Although the SAR autocorrelation coefficients were only reduced by a small amount, the Moran coefficients have dropped considerably. The simplified scatterplots from all three stage-two resampled data-sets were almost identical, so we selected the one generated from regular sampling as representative. Figure 16 depicts an excellent portrayal of the original point cloud from the parent data-set and all fitted lines are overlapped with little deviation from those of the original data-set. With less than 45,000 data points, it took a conventional polynomial). The point cloud from the resampled set captures the general visual structure of the original data, with much less overlap. The GAM/LOESS lines are virtually identical, except for the upper region of the NIR ($X$ axis) variable. The OLS fits show similar patterns. The quartic polynomial fits are a complete overlap, while the cubic polynomial from the resampled set provides a good generalization of the relationship, and may alleviate the possible problem of overfitting in the lower and upper regions. These results clearly show the spatially simplified scatterplot, using a fraction of the original data, can sufficiently represent the key characteristics of the original scatterplot, making it possible to perform interactive exploration.

Figure 12. Spatially simplified scatterplots (point clouds with local/global fits) from three resampling schemes (left: random sampling; center: regular sampling; right: hexagon stratified sampling).

Figure 13. A random sample ($n = 1326$, right) of the hexagon stratified random sample set ($n = 2438$, left). Note: Sample sites are shown as yellow dots.
desktop computer less than a second to generate the scatterplot constituting a distinguishable point cloud and the local and global fitting lines.

7. Discussion

Experiments with simulated data and a remote sensing data-set support the hypothesis that overplotting and excessive delay in generating scatterplots for large spatial data-sets can be resolved through exploiting the information redundancy in the data, as measured by spatial autocorrelation. Though the original intended use of effective sample size was to determine the correct sample size for estimating different statistical properties of a population (Griffith 2005), the concept can serve as an approximation of the true information content contained in the data-set, and as a guide for extracting a small subset closely representing the original data on the corresponding scatterplot. As the experiment shows, when the Moran coefficients are as high as 0.98, less than 1% of the original data can render scatterplots identical to the original. As such levels of spatial autocorrelation are common among raster data, especially for remotely sensed data and their derivatives, the prospects of a spatially simplified scatterplot are promising.

Calculating the effective sample size involves computational tasks that can be highly complex. Effective sample size is a function of the SAR spatial autocorrelation coefficient whose estimation involves the calculation of the eigenvalues of the spatial weights matrix that is the square of the data size. The amount of RAM required often exceeds even a well-equipped desktop computer, not to mention the hours of processing time it would take. Other methods, such as Monte Carlo approximation and LU decomposition, can accommodate large sizes but would still be unable to handle a $W$ matrix as large as $900,000 \times 900,000$ within a reasonable range of response time. The computation issues with finding the SAR coefficient, fortunately, were largely resolved for raster data, thanks to the approximation methods developed by Griffith, as described in the previous section. Because these estimations do not require numerically intensive calculations of the eigenvalues, we can obtain the SAR coefficient within minutes instead of hours. For other geospatial data types, fast approximation of eigenvalues has yet to be developed.

Throughout the course of our experiment, we adopted a two-stage resampling design because the resampled data-set, although substantially smaller than the original data-set (2%), still exhibited a high level of spatial autocorrelation. The information redundancy is likely due to the relatively simplistic sampling design where hexagons or rectangles are regularly laid out, and where sample sites are either systematically or randomly selected, without considering the spatial associations among data elements. For example, a large and elongated water body, which should likely be sampled just once because of its homogeneity, may end up being sampled many times. Meanwhile, an area with a high level of heterogeneity, which should be sampled more densely, would have only one sample to represent its rich variation. Adopting more sophisticated sampling methods, such as those reviewed by Wang et al. (2012), will help reduce spatial autocorrelation and hence improve the quality of the resampled set.

However, because the purpose of a spatially simplified scatterplot is to generate a close match to the original plot by extracting a small subset, the level of spatial autocorrelation serves as an indicator of potential reduction, instead of a criterion. Among various statistical measures we monitored, the K–S test serves as a good threshold of substantial deviation from the original data set. Confidence intervals of such fitting functions as GAM and LOESS are also potential candidates for formulating
Table 4. Summary of resampling statistics for the remote sensing data-set.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (CI)</th>
<th>Standard deviation (CI)</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>p (value)</th>
<th>Dp-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI</td>
<td>0.852 (0.843, 0.853)</td>
<td>0.052 (0.051, 0.053)</td>
<td>-0.500 ( -0.505, -0.505)</td>
<td>2.497 (2.495, 2.499)</td>
<td>0.001 (0.000, 0.002)</td>
<td>0.984 (0.983, 0.985)</td>
</tr>
<tr>
<td>Hex-stratified</td>
<td>0.852 (0.843, 0.853)</td>
<td>0.052 (0.051, 0.053)</td>
<td>-0.500 ( -0.505, -0.505)</td>
<td>2.497 (2.495, 2.499)</td>
<td>0.001 (0.000, 0.002)</td>
<td>0.984 (0.983, 0.985)</td>
</tr>
<tr>
<td>Rec-stratified</td>
<td>0.852 (0.843, 0.853)</td>
<td>0.052 (0.051, 0.053)</td>
<td>-0.500 ( -0.505, -0.505)</td>
<td>2.497 (2.495, 2.499)</td>
<td>0.001 (0.000, 0.002)</td>
<td>0.984 (0.983, 0.985)</td>
</tr>
<tr>
<td>NIR</td>
<td>0.852 (0.843, 0.853)</td>
<td>0.052 (0.051, 0.053)</td>
<td>-0.500 ( -0.505, -0.505)</td>
<td>2.497 (2.495, 2.499)</td>
<td>0.001 (0.000, 0.002)</td>
<td>0.984 (0.983, 0.985)</td>
</tr>
<tr>
<td>Rec-stratified</td>
<td>0.852 (0.843, 0.853)</td>
<td>0.052 (0.051, 0.053)</td>
<td>-0.500 ( -0.505, -0.505)</td>
<td>2.497 (2.495, 2.499)</td>
<td>0.001 (0.000, 0.002)</td>
<td>0.984 (0.983, 0.985)</td>
</tr>
</tbody>
</table>

Notes: Mean and standard deviation based on 4,000,000 samples. p-value is the probability of observing a particular data set under the null hypothesis of no difference.

Figure 15. An overlay of scatterplots from the original data and the resampled set.
Notes: Light grey point cloud: original data (3000 x 3000 points). Darker grey point cloud: hexagon stratified random sampled data (174,212 points). Black wavy curve: GAM fitted line from the original data. Violet wavy curve: GAM fitted line from sample data. Blue curve: a cubic polynomial fit to the sample data. Green curve: a quartic polynomial fit to the original data. Black dash: linear fit to the original data. Violet dash: linear fit to the sample data.

Figure 16. Scatterplot of data from the stage-two regular resampling.
Notes: Light grey point cloud: stage-one hexagon random sampled data (174,212 data points). Darker grey point cloud: stage-two regular sampled data (42,025 data points). Black wavy curve: GAM fitted line from stage-one data. Violet wavy curve: GAM fitted line from stage-two resampled data. Blue curve: a cubic polynomial fit to the stage-two resampled data. Red curve: a quartic polynomial fit to the stage-two resampled data. Green curve: a quartic polynomial fit to the stage-one data. Black dash: linear fit to the stage-one data. Violet dash: linear fit to the stage-two resampled data. X axis = normalized Band 4 radiance, Y axis = normalized NDVI. NDVI = [(Band 4 - Band 3) / (Band 4 + Band 3)]. NDVI is a simple algebraic combination of remotely sensed spectral information (e.g., Bands 4 and 3 from Landsat Enhanced Thematic Mapper) that provides meaningful information about vegetative structure (leaf/cellular structure) and condition (chlorophyll content). Generally, high NDVI values indicate that photosynthetically active plant biomass is present.
the simplification tolerance parameters. After the second stage resampling, these measures actually showed there was more room for further reduction, because there was minimum departure from the original data-set and the level of spatial autocorrelation was still relatively high. A sequential resampling, or multi-stage resampling, will result in a smaller subset and a lower level of spatial autocorrelation, but greater deviation from the original data, a tradeoff that is to be determined by researchers.

The fact that sequential resampling can produce a data subset smaller than the effective sample size while maintaining a close match with the original data, in the attribute domain, suggests we may not need all the independent samples to represent the key characteristics of a scatterplot. Further research is needed to characterize these controlling data points and to develop methods to identify them.

8. Conclusions

The spatially simplified scatterplot presented in this paper serves as an addition to the continuous efforts in enhancing the usability of the scatterplot for large data-sets. While most of the current methods tend to focus on generalization in the attribute domain, the spatially simplified scatterplot is originated in the simplification in the spatial domain. Although both approaches exploit redundancies in data, attribute generalization methods seem to emphasize visual manipulations, while the spatially simplified method seeks a balance between both domains with a sound theoretical basis. We are optimistic that it will gain acceptance as a tool for exploring large spatial data-sets.

The methods presented in this paper are tailored to rectangular tessellated interval/ratio scale spatial data and with an inclination toward normally distributed variables. For vector data with interval/ratio scale attributes, different sampling methods are needed. Calculating the SAR spatial autocorrelation coefficient will be a major bottleneck if the data size jumps to millions, because currently there is no approximation method similar to those for raster data. Converting data from vector to raster offers a possible solution.

Notes on contributors

Bin Li is a professor at Central Michigan University. He has a PhD in Geography from Syracuse University, an MS from University of Nebraska, and a BS from South China Normal University. He conducts research in GIS with a current focus on applying spatial statistical methods to solve theoretical as well as application problems in Geography.

Daniel A. Griffith is an Ashbel Smith professor at University of Texas-Dallas. He has a PhD in Geography from University of Toronto, an MS in Statistics from Pennsylvanian University, an MA in Geography and a BS from Indiana University of Pennsylvania. His main research interests are in spatial statistics, quantitative urban and economic geography, and spatial epidemiology. He publishes extensively and his scholarly contributions are recognized world-wide.

Brian Becker is a professor at Central Michigan University. He has a PhD in Geography from Michigan State University, an MS from University of Illinois, and a BS from East Illinois University. His research interests are in remote sensing and its applications in wetland ecology.

References


Spatially Simplified Scatterplots for Large Raster Datasets

CC BY 4.0 (Attribution) License
©2016 Wuhan University

Citation: