geographical analysis

Geographical Analysis (2016) **48,** 115–131

Graphical Inference in Geographical Research

Holly M. Widen, James B. Elsner, Stephanie Pau, Christopher K. Uejio

Department of Geography, Florida State University, Tallahassee, FL, USA

Graphical inference, a process refined by Buja et al., can be a useful tool for geographers as it provides a visual and spatial method to test null hypotheses. The core idea is to generate sample datasets from a null hypothesis to visually compare with the actual dataset. The comparison is performed from a line-up of graphs where a single graph of the actual data is hidden among multiple graphs of sample data. If the real data is discernible, the null hypothesis can be rejected. Here, we illustrate the utility of graphical inference using examples from climatology, biogeography, and health geography. The examples include inferences about location of the mean, change across space and time, and clustering. We show that graphical inference is a useful technique to answer a broad range of common questions in geographical datasets. This approach is needed to avoid the common pitfalls of "straw man" hypotheses and "p-hacking" as datasets become increasingly larger and more complex.

Introduction

Geographic datasets tend to contain information from human and biophysical systems due to the integrative nature of the discipline (Gahegan and Brodaric 2002). Thus, geographic datasets are often products of multiple datasets linked together by a spatial or temporal aspect and are very extensive (Gahegan et al. 2001). As datasets become larger and more complex (big data), traditional statistical inference can become more difficult (Gahegan et al. 2001; Wickham et al. 2010). Thus, visualizations are crucial in processing complex datasets and developing accurate hypotheses for knowledge construction (Fox and Hendler 2011). Furthermore, Shneiderman (2014) asserts that humans will be critical to the conversion of massive datasets into usable information as they can recognize patterns more efficiently than machines. Graphical inference, a process refined by Buja et al. (2009) and introduced to the InfoVis community by Wickham et al. (2010), can be a useful tool for geographers as it assists in analyzing big data and can allow for human control in decision-making.

Graphical inference constructs empirical knowledge from data by fusing exploratory (descriptions) and confirmatory (conclusions) analysis. Tukey (1977) compares exploratory analysis to a "detective" who gathers evidence and confirmatory analysis as the judge/jury that

Correspondence: Holly M. Widen, 113 Collegiate Loop, Tallahassee, FL 32306
e-mail: hwiden@fsu.edu

Submitted: April 23, 2015. Revised version accepted: July 30, 2015.

doi: 10.1111/gean.12085

evaluates the strength of the evidence. Exploratory analysis may provide support for proceeding to confirmatory analysis (DiBiase 1990). Just as importantly, exploratory analysis may cast doubt on hypotheses or point toward unanticipated relationships. Popular exploratory analysis tools include counts, summary statistics, and geovisualization (e.g., Tukey 1977; Tufte 1983, 1990; Cleveland 1985, 1993; Cressie 1993; Keirn and Kriegel 1994; MacEachren and Taylor 1994; MacEachren et al. 2004, among others). More recently, exploratory data analysis (EDA) was expanded to accommodate spatial information and leverage geographic information system (GIS) data visualization and manipulation (Anselin 1999; Kwan 2000; Guo et al. 2005; Koua, MacEachren, and Kraak 2006; Chen et al. 2011, among others).

Similarly, visualization methods are mostly used for hypothesis generation through exploratory analysis, data mining, and geovisualization (Gahegan et al. 2001). Gahegan et al. (2001) call for integration of exploratory visual and computational methods throughout the entire knowledge construction process, including visualization of data or results, knowledge discovery in databases, and geocomputation. Gahegan and Brodaric (2002) discuss the gap between data exploration and explanation of geospatial data. The authors, as well as MacEachren et al. (2003), specifically use GeoVISTA *Studio* to integrate EDA methods; in addition, there are multiple CyberGIS platforms for spatial analysis and decision-making, such as GeoDa and PySAL (Wang et al. 2013). Although most of these platforms are used to discover relationships in data, they are not used to detect significance.

EDA and geovisualization may not evaluate the strength of the evidence, which traditionally requires numerical inference tools such as confidence intervals and *P*-values. Thus, confirmatory analysis or inference is performed following the exploratory analysis to draw conclusions from data. The interest lies in understanding the behavior of a population from a sample of observations. Traditional inference codifies the reasoning with quantitative procedures rooted in statistical theory. The procedure takes the form of a numerical calculation to evaluate the strength of the evidence and inference is made objectively. Graphical inference takes a different approach as it codifies the reasoning visually and spatially (Buja et al. 2009), making it interpretable to a broader audience.

Perhaps more importantly, graphical inference can help avoid two common issues that arise with formal statistical testing: "straw man" null hypotheses and "p-hacking." A straw man is an argument set up so that it is easily refuted (Cortina and Dunlap 1997). Straw man null hypotheses are simple to identify on plots making it less likely to draw wrong conclusions. The term p-hacking refers to the monitoring and manipulation of data and/or analyses to find the desired significant result (Nuzzo 2014). In this case, the researcher intentionally tries to minimize the *P*-value to reach the significance threshold ($P\text{-value} \leq 0.05$). This can be achieved through different practices, including, but not limited to, collecting more or fewer data based on preliminary analyses, stopping/repeating analysis to reach desired *P*-value, selectively reporting results or conditions, and falsifying data (John, Loewenstein, and Prelec 2012; Head et al. 2015). This approach makes the *P*-value the target of the analysis and alters the interpretation of the results. It also increases false positive results in scientific literature due to the pressure to publish only significant findings (Head et al. 2015). Statistically valid procedures (e.g., Bonferroni Correction and Tukey's Studentized Range) can account for multiple hypothesis testing by proportionately lowering the *P*-value threshold (Hochberg and Tamhane 1987). By requiring both thoughtful consideration of the null hypothesis as well as an unbiased jury, graphical inference makes manipulation difficult and thereby limits the opportunity for these problems to arise.

EDA visualizations can be merged with traditional statistical methods. For instance, Buja et al. (1988) and Buja, Cook, and Swayne (1999) create reference distributions to conduct visual comparisons by performing permutation tests from symmetries. In this example (and the following), the visualizations of the reference distributions become the test statistics in the confirmatory analysis. Gelman (2003, 2004) construct reference distributions for EDA graphics by generating simulations from statistical models within a Bayesian framework. Lastly, Buja et al. (2009) and Wickham et al. (2010) produce reference datasets based on null hypotheses for use in small-multiple displays to facilitate inference making. Here, we focus on the latter study where we use visualizations to conduct hypothesis testing through graphical inference. This method bridges the gap between exploratory and confirmatory analysis and allows for direct comparison with traditional statistical tests (Majumder, Hofmann, and Cook 2013).

Graphical inference utilizes maps or charts for analysis where the key idea is the display of static small-multiples in an array or line-up similar to those introduced by Tufte (1983, 1990). One map in the array shows the observed data and the others show—using the same scale and resolution—synthetic data under the null hypothesis. In this way, visual and spatial reasoning is applied in an inferential setting. Geographers are often interested in spatial relationships but it can be difficult to make inferences from a single static map without being able to make comparisons. Graphical inference allows for strict hypothesis testing through a visual inspection of many similar maps. The positioning of the maps is within the eye span so viewers can easily make comparisons and conduct visual reasoning (Tufte 1990).

Consider the array of maps in Fig. 1. One map displays recorded locations of the 2,035 tornado touchdowns in the Central Plains during May over the period 1950–2011. The other maps display events that are randomly assigned to locations in the east–west and north–south directions under the null hypothesis of complete spatial randomness (CSR) of tornado touchdowns. The map displaying the observed data is hidden in the array. If an independent and impartial observer can identify the real data as different from the others, we reject the null hypothesis and comfortably conclude there is a significant difference without the need to make a formal calculation.

This article explains and demonstrates graphical inference through examples from geography. It is based on the work of Buja et al. (2009) and Wickham et al. (2010) and provides a new example to test for spatial autocorrelation. Section *Is there a difference?* describes visual hypothesis testing using the example above (Fig. 1). Section *How does it work?* outlines the procedures for performing graphical inference. Section *Examples* provides examples of using graphical inference for the purpose of geographical research. The examples are from the subdisciplines of climatology, biogeography, and health geography. Finally, section *Summary and Discussion* summarizes the article and discusses methodological limitations. All the code used to generate the examples in this article is available from http://www.rpubs.com/hwiden/GraphicalInference.

Is there a difference?

Hypothesis testing is formalized with mathematics where a test statistic is calculated. The test statistic is described by a parametric distribution under the assumption that the null hypothesis is correct. The difference between the computed test statistic and more extreme values from the distribution provides a numerical answer (*P*-value) to the question “is there a difference?.”

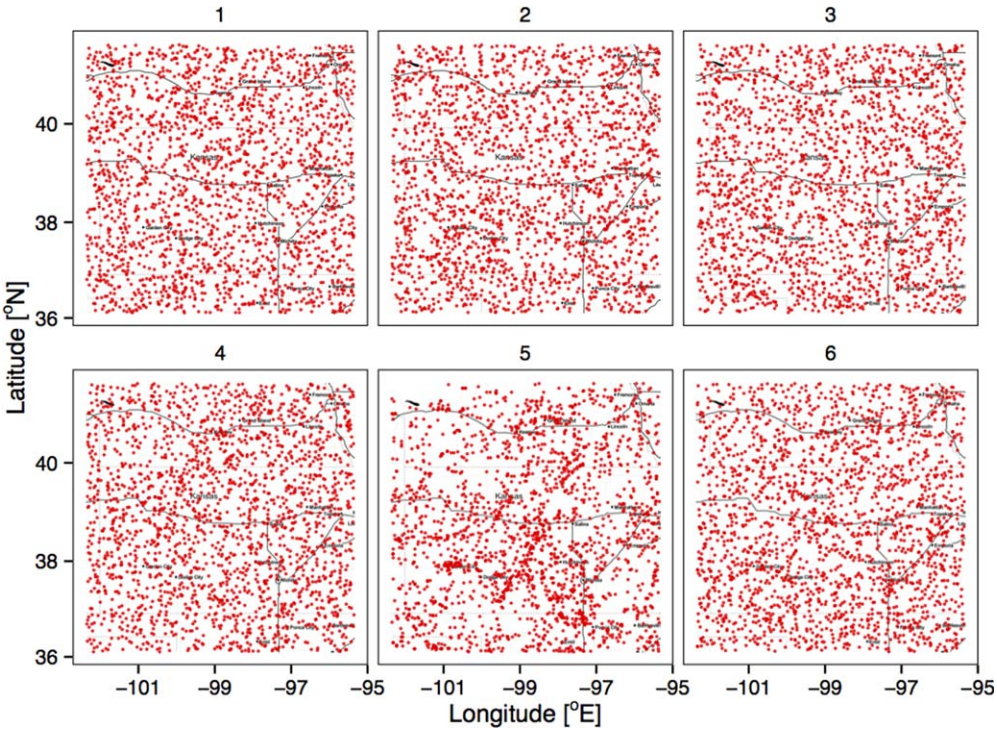

Figure 1. Maps showing location of recorded tornado touchdowns during May over the period 1950–2011. One map displays the observed locations and the others display samples of random locations. Can you identify the map containing the observed data?

This question can also be addressed visually. Consider the situation where we "hide" the observed data in a line-up of data generated from the null hypothesis. By "hide" we mean randomly assign a plot of the observed data to a location in an array of identically scaled plots using decoy data generated to be consistent with the null hypothesis. If an impartial jury can identify the location of the real plot in the array of decoys, then we reject the null hypothesis.

As an example, an interesting question concerns the clustering of tornadoes as the reports are historically affected by population bias (Verbout et al. 2006). Are tornado reports more clustered than one would expect under the null hypothesis of CSR? The observed data on tornado touchdowns are obtained from the U.S. Storm Prediction Center (SPC) [www.spc.noaa.gov/gis/svrgis/]. Here, we focus on a region centered on Russell, Kansas as defined in Elsner et al. (2013) and used in Elsner and Widen (2014). The region stretches across the central Plains from northern Texas to central Nebraska. It is an area with a high concentration of tornadoes and where there are no large spatial gradients in the occurrence rates of tornadoes.

The observed tornado data are placed on a map as points and the map is randomly assigned a location in an array of decoy maps (Fig. 1). The decoy maps are generated under the null hypothesis by randomly assigning spatial coordinates to the locations of the tornado touchdowns (i.e., Ho: tornado touchdowns are an example of CSR). In the array, there is a fairly obvious visual difference between map number 5 and the decoys. The arrangement of real tornado locations appears to be more clustered than decoy maps of events under the assumption of CSR so we confidently infer clustering in the tornado record. The clustering arises from a

historical bias whereby tornado reports tended to be more numerous near cities and roadways (Anderson et al. 2007).

How does it work?

Overview

Buja et al. (2009) and Wickham et al. (2010) describe two protocols in graphical inference: the "Rorschach" and the "line-up." The Rorschach protocol, named after the famous inkblot test, functions like a calibrator where the goal is to adjust our perception to the natural variability in data generated from a null distribution. The variability in the real dataset may also be compared visually with the variability of these decoys. The line-up protocol functions like a statistical test where the key idea is to try to visually identify the real data in an array of decoys, similar to a police line-up.

The first step of the Rorschach protocol is to identify a null hypothesis from which to generate null datasets. The null datasets can be created using imputation, resampling, or simulation techniques. Imputation involves replacing missing data with values from a given distribution, resampling involves permuting the observed data, and simulation involves sampling from a distribution. The second step is to create decoy plots from the null datasets and to display them in an array using identical scales. The third step is to ask an impartial judge what they see (e.g., patterns, trends, etc.) in the decoy plots. This analysis evaluates the tendency to make Type I errors (false positives) where one falsely identifies structure in the data due to random variability (Buja et al. 2009).

The line-up protocol extends the Rorschach protocol to make a graphical inference. It begins with the same two steps as the Rorschach protocol; however, the third step is to develop an array where one of the decoy plots is randomly replaced with a plot of the real data. Then the fourth step is to ask an impartial judge to identify which plot contains the real data. The ability of a judge to identify the real data from a line-up of decoys is related to a formal *P*-value. These third and fourth steps may also be used in the Rorschach protocol (replacing the original third step) to maintain the integrity of the analysis if the judge is aware that the array contains only decoy plots. Similarly, the real data should not be seen prior to either analysis by the impartial judge.

P-values

A *P*-value is an estimate of the "probability" that your data, or data more extreme than observed, could have occurred by chance if the null hypothesis were true. It is evidence in support of the null hypothesis and thus, the smaller the *P*-value, the less support there is for the null hypothesis. With graphical inference, the *P*-value is related to the likelihood that the judge (or judges) will identify the real data in the line-up. If there are K impartial judges and x of them are able to identify the real data from a line-up of m randomly placed plots, then the *P*-value is given by

$$P(X \geq x) = 1 - Binom_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right) \left(\frac{m-1}{m}\right)^{K-i} \quad (1)$$

where X has a binomial distribution with parameters K (number of judges) and P (probability of random selection = $1/m$) (Majumder, Hofmann, and Cook 2013). Therefore, if a single judge

is able to identify the real data in a line-up involving 19 decoys ($P = 1/20$), then the P-value is 0.05 (Buja et al. 2009).

Precision on the P-value can be increased with more decoys or more judges. However, there are limitations concerning the number of plots one person can absorb at a time (Buja et al. 2009). A reasonable amount is 20, which will generate a lower bound on the P-value of 0.05, the traditional threshold for significance in statistical inference (95% confidence level). However, the significance based on identifying the real data may be a continuum, not a specific threshold such as P-value of 0.01 or 0.05.

Power

The power of a statistical test is the likelihood of avoiding a Type II error (i.e., rejecting the null hypothesis when it is false). Power is often used to choose an appropriate plot design or test but it can also be used to compare the performance of visual and traditional tests (Hofmann et al. 2012; Majumder, Hofmann, and Cook 2013). With graphical inference, the power, V_θ, is influenced by the number of observers who identify the real data from the line-up and is determined by

$$\text{Power}_V(\theta) = \text{Power}_{V,K}(\theta) = 1 - \text{Bino}m_{K,1/m}(x_\alpha - 1) \tag{2}$$

where α is the significance level and x_α hinges on $P(X \geq x_\alpha) \leq \alpha$ (Majumder, Hofmann, and Cook 2013). Hofmann et al. (2012) estimate the power of a line-up simply as the ratio of correct identifications of the real data out of the total number of observations and evaluate competing plot designs using logistic regression and Amazon's Mechanical Turk service.

Using R

Graphical inference is made practical with modern graphing applications. The R project for statistical computing is particularly well suited because of the **ggplot2** package (Wickham 2009). The functions in **ggplot2** are based on the grammar of graphics theory of Wilkinson et al. (2006). The grammar specifies how a graphic maps data to attributes of geometric objects. Examples of attributes are color, shape, and size, whereas points, lines, bars, and polygons are examples of geometric objects. The plot is drawn on a specific coordinate system. Faceting is used to generate a line-up by replicating the plot with identical axes and scales but with different data.

The **nullabor** package (Wickham, Chowdhury, and Cook 2014) contains a **line-up()** function to generate the line-up of plots for visual analysis and two additional functions, **null_permute()** and **null_dist()**, to generate null datasets by means of resampling and simulation, respectively. The **null_permute()** function creates decoys by randomizing the data without replacement. This method can be used to assess independence as the arrangement of the data is varied but the distribution is preserved (Wickham et al. 2010). The **null_dist()** function creates decoys from a specified distribution based on the mean and standard deviation of the data. It can be used to uncover specific differences or relationships based on a hypothesized data distribution. More generally, there is a variety of functions from the base package for producing null datasets and additional code can be written to replace one of the decoy plots with the plot of the real data.

Table 1 Graphical Inference Techniques for Spatial and Spatiotemporal Research

Analysis	Visualization technique	Research question	Null hypothesis
Spatial analysis	Difference map	Does the mean/variance change over space?	No change/trend
Spatial dependence	Dot density/choropleth map/Moran's scatterplot	Are the data/residuals spatially dependent?	Complete spatial randomness
Temporal analysis	Time series graph	Does the mean/variance change over time?	No change/trend
Temporal dependence	Residual plot	Are the data/residuals temporally dependent?	Temporal independence

Common statistical inference graphs/maps used by geographers, the spatial and spatiotemporal questions they seek to answer, and the associated null hypotheses.

Examples

The line-up protocol is general enough to be used on a variety of inferential problems. Geographers tend to analyze spatial and spatiotemporal data for relationships, patterns, and trends. Table 1 lists a few common statistical inference plots and maps used by geographers and the research questions (regarding the null hypotheses) they seek to answer. In the following examples, we demonstrate the line-up protocol on four different inferential problems in order of increasing complexity. The first is inference about location of the mean, the second is inference about a change across space, the third is inference about a change across time, and the fourth is inference about clustering.

Inference about location

Given a sample of data from a population, a common inferential question concerns the location of the population mean. This example serves to illustrate the methodology by allowing us to vary the hypothesized population mean and by proceeding in steps to the line-up.

Testing for differences in means in health data is a common statistical problem for geographers studying public health. Here, we use a dataset consisting of statewide melanoma mortality rates for white males in the United States over the period 1950–1969 (Hothorn and Everitt 2009). We begin with the null hypothesis that the population mortality rate is 160 deaths per one million inhabitants. This value for the population mean mortality rate is chosen as the null hypothesis because it differs from the actual mean but not substantially. The data and hypothesized population mean are plotted in Fig. 2. The boxplot shows that the hypothesized population mean mortality rate is higher than the sample mean mortality rate of 153 deaths per one million inhabitants. The interquartile range, the range of the middle half of the data, is 50 and there are no outliers.

For comparison we perform a standard one-sample t-test with the t statistic given by

$$t = \frac{\underline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{3}$$

where $\underline{x}$ is the sample mean, μ is the hypothesized population mean, s is the sample standard deviation, and n is the sample size. The test results in a P-value = 0.14, which is suggestive, but inconclusive evidence against the null hypothesis.

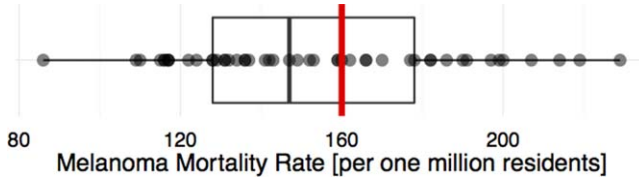

Figure 2. Melanoma mortality rates by U.S. state among white males from 1950 to 1969.

Another way to think about this is to generate random samples of mortality rates with the hypothesized mean value and see how these samples compare with the data. Continuing under the null hypothesis that the population mortality rate is 160 deaths per one million inhabitants, we use the **rnorm()** function (R Core Team 2013) to simulate a random sample of mortality rates ($n = 49$) based on the null hypothesis of a normal distribution with a mean of 160 (Fig. 3). It is visually clear that the observed data do not differ much from this particular decoy (sample generated under the null hypothesis). The range of values and sample means are quite similar.

But what about other decoys? How do they compare with the sample of observed data? Fig. 4 shows a line-up of 20 boxplots, 19 from decoys and 1 from the actual data. In panel (a) the decoys are based on random samples with a mean of 160. This corresponds to the formal t-test above.

As expected, it is difficult to find the real data (sample 2) against these decoys. This is consistent with the marginal *P*-value from the formal test. By comparison, in panel (b) where the decoys are based on samples with a mean of 185 (substantially higher than the actual mean), it is easy to see that sample 13 contains the real data. A formal t-test confirms this giving a *P*-value that is less than 0.001.

Inference about change across space

Next, consider the distributions of three tree species, *Pseudotsuga menziesii*, *Chamaecyparis lawsoniana*, and *Arbustus menziesii*, measured by Whittaker (1960) along a moisture gradient in the central Siskiyou Mountains. Here, we reinterpret this classic work, which was not originally analyzed statistically, to show how species are distributed across different habitats.

The data consist of numbers of tree stems (>1 cm dbh) per 0.5 hectare collected along 10 transect steps from 2,000 to 3,000 ft in elevation (Whittaker 1960). In this example, the null hypothesis is that the species sort randomly along the moisture gradient. The data are entered as a table and the columns and rows summed as in a chi-square analysis. Then, 19 random two-way tables are generated with the **r2dtable()** function (R Core Team 2013) using the marginals from the observed data. The marginals are the row and column sums (written in the "margins" of a table) and are held constant. This creates the samples for graphical inference based on the null hypothesis that the distribution of each species is random within each transect.

Fig. 5 is an array of heat maps displaying the distributions of stem densities for each species along the moisture-gradient transects using the observed data and the 19 decoys. The true data are easily picked out among the simulated samples (position 14) so we reject the null hypothesis that species occur randomly along the moisture gradient.

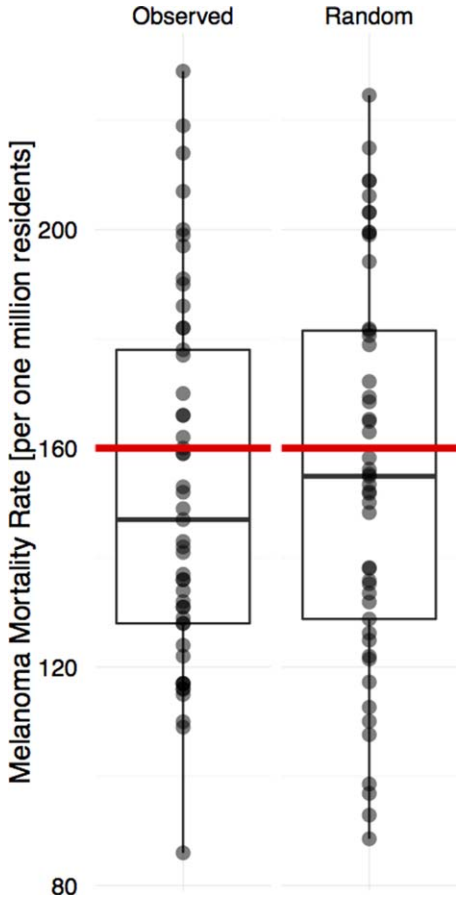

Figure 3. Melanoma mortality rates by state and a sample of rates assuming a normal distribution with mean of 160 and standard deviation equal to the standard deviation of the observed rates.

For comparison, a chi-square test is performed. The chi-square statistic is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{4}$$

where O is the observed frequency and E is the expected frequency. The resulting P-value is <0.001 confirming the graphical inference to reject the null hypothesis and conclude that the distribution of the three tree species is not random (i.e., the observed frequencies are significantly different than the expected frequencies) across the moisture-gradient transect.

Inference about change across time

Change is often examined by considering trends. In the case of time-series data, the inference problem takes the form of determining whether the slope of the trend line is significantly different from zero. Trend analysis is a common statistical approach for geographers that utilize historical data for prediction purposes.

Physical theory predicts a seven percent increase in atmospheric moisture per degree of warming assuming a fixed relative humidity (Held and Soden 2006). So, one might argue that as the

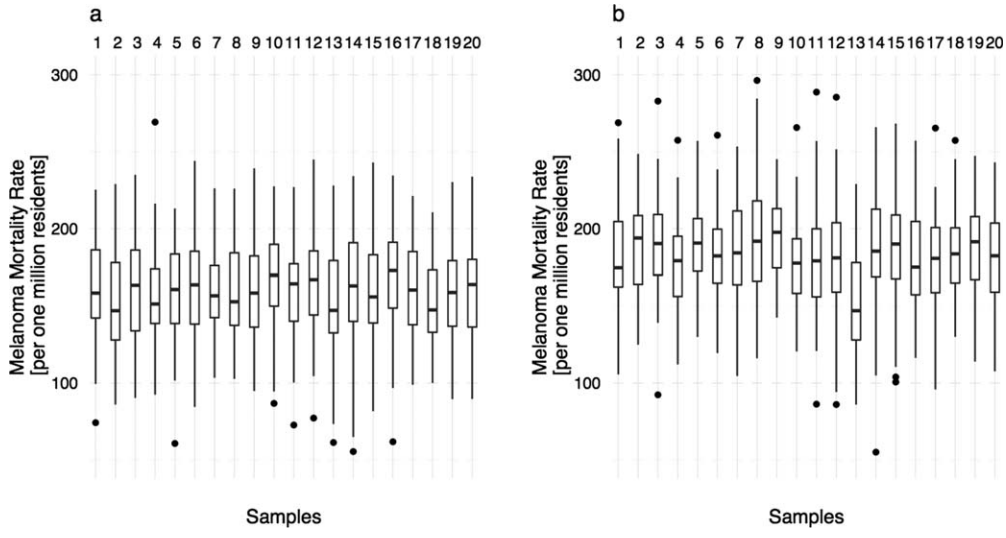

Figure 4. Melanoma mortality rates by state and 19 random samples of mortality rates. The random samples are generated assuming a normal distribution with mean of 160 (a) and mean of 185 (b). Can you pick out the box plot of the observed data in each panel?

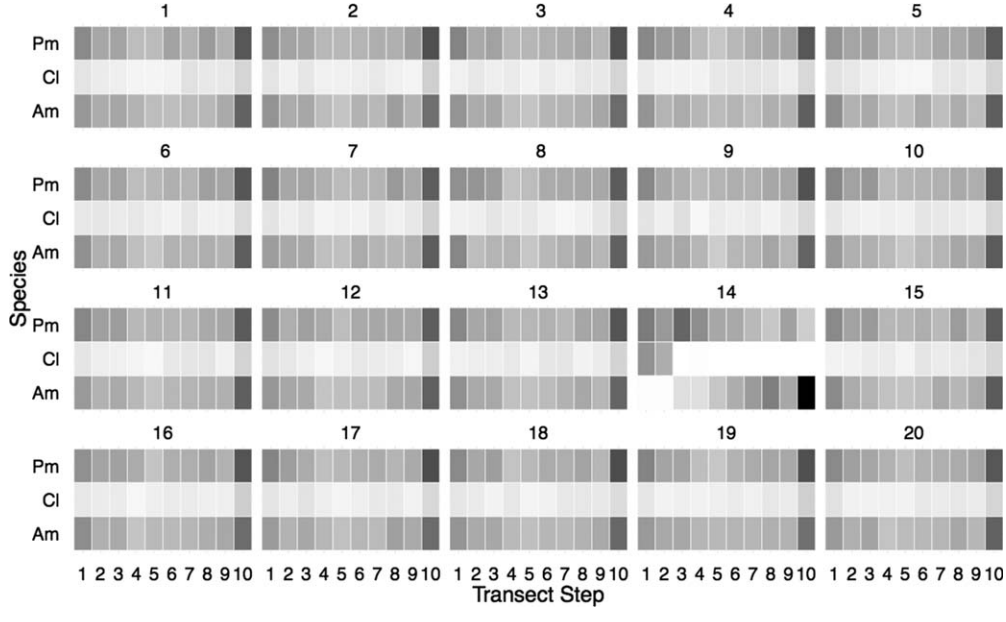

Figure 5. Stem densities per .5 hectares by species along a moisture-gradient transect on quartz diorite and 19 random samples of stem densities. The random samples are generated assuming constant marginals. The darker (lighter) shading represents higher (lower) stem densities.

earth warms, increased atmospheric moisture will lead to more rainfall. Consider the time series of average March precipitation in Florida over the period 1985–2012 [www.myweb.fsu.edu/jelsner/data/FLprecip.txt]. The null hypothesis is that there is no trend in the data. Nineteen random

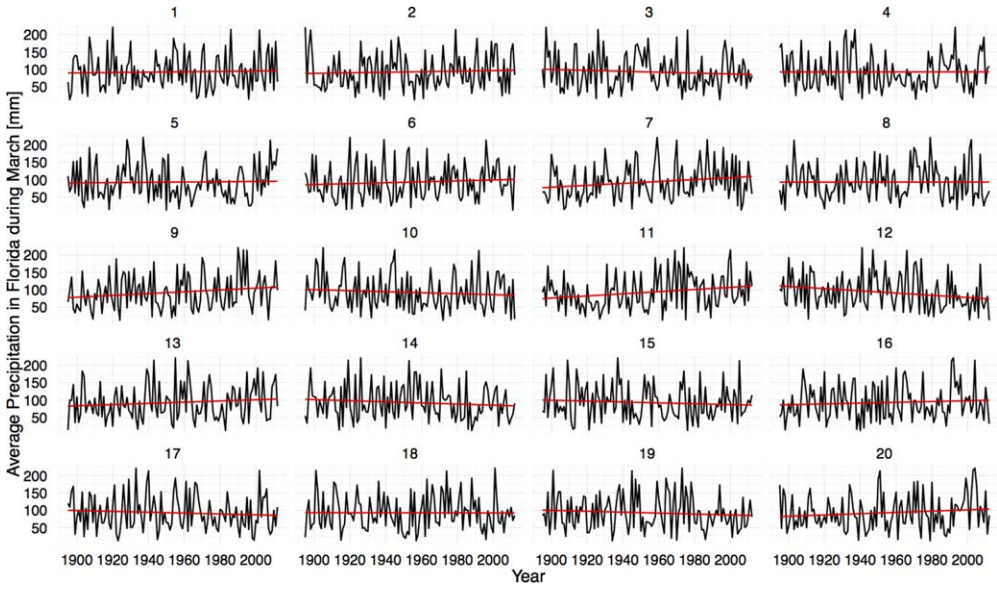

Figure 6. March rainfall in Florida (1895–2012) and 19 random samples of rainfall. The random samples are generated by permuting the rainfall amount across the years under the assumption of independence. Can you find the observed time series?

samples of March precipitation are generated by permuting the values across the years under the assumption that there is no autocorrelation between years. If there is temporal dependency, one could generate null distributions with autoregressive or moving average (ARMA) components to produce normally distributed comparison time series with specified correlation structures.

Fig. 6 is an array of time-series with the 19 decoys and one actual time-series randomly positioned in the array. Best-fit linear slope lines are drawn on each plot. Some of the decoy series show shallow upward and downward trend lines and the upward trend in the real data (sample 7) is not large enough to stand out against trends computed on randomized data under the null hypothesis. Thus, we fail to reject the null hypothesis of no trend.

For comparison we test the regression slope with the t statistic given by

$$t = \frac{b}{se(b)} \tag{5}$$

where b is the slope coefficient and se is its standard error. The test results in a P-value = 0.04 and a slope coefficient = 0.28. Although marginally significant, the slope has a very small positive magnitude. This corresponds to the results of the line-up analysis as the shallow upward trend hides among the decoys displaying the natural variability in both directions. Thus, although traditional statistical tests can uncover small effects in data, the results may not be very meaningful and may lead to a Type I error if the dataset is large.

Inference about clustering

In geographical research, spatial autocorrelation or clustering may be an interesting spatial signal or it may be an error that needs to be accounted for to meet assumptions of other statistical tests. Regardless of the reason, it is common for geographers to test for clustering as many utilize spatial data.

As shown in the introduction, the distribution of tornado touchdown points illustrates patterns of spatial clustering particularly well. The data are more clustered than samples generated under the null hypothesis of CSR. Another way these data might be analyzed is with quadrats. Here, we place a 12 by 12 quadrat over the domain and record the number of tornado touchdowns in each cell. This is accomplished using functions in the **raster** package (Hijmans and van Etten 2011).

We assess clustering by calculating local Moran's I values from the quadrats of tornado counts. Local Moran's I is a local spatial autocorrelation statistic that measures the extent of spatial clustering of similar values neighboring a particular observation (Anselin 1995). The local Moran statistic is given by

$$I_i = z_i \sum w_{ij} z_j \tag{6}$$

where z_i and z_j are observations in deviation form, w_{ij} is the spatial weight in row-standard form, and the summation over j includes only neighboring values.

Fig. 7 displays the local Moran's I values computed from the quadrats. Areas across the south central part of the region have cells with a relatively high number of tornadoes and are adjacent to cells that also have a high number of tornadoes. These cells illustrate positive local spatial autocorrelation and clustering within the study region.

For comparison, the touchdown locations are randomized and counts are again tabulated for each cell. Local Moran's I maps are shown for the observed and one sample of the randomized data in Fig. 8 using the same color scale range. The muted colors on the map indicate lower values of local spatial autocorrelation.

But what about other decoys? Fig. 9 shows a line-up consisting of 19 decoy boxplots of local Moran's I values and one boxplot from the actual data randomly positioned in the line-up. Here, the true data are not difficult to identify (sample 4) as the distribution of values is higher on average and much higher in a few cells. Without a formal test of the null hypothesis, it can be stated confidently that the null hypothesis of no spatial autocorrelation on this scale is untenable.

Summary and discussion

Geographers have a rich history of using large, observational datasets and a long interest in graphical and visual analysis. Graphical inference can be a useful tool for geographers as it assists in analyzing large, complex data and is a visual and spatial method to test null hypotheses. This article demonstrates the utility of using graphical inference for a set of common statistical tests used in geographical research, particularly spatial autocorrelation. Examples are illustrated from climatology, biogeography, and health geography. The examples include inferences about location of the mean, change across space and time, and clustering. Although we only offer examples with null datasets generated from simulation and permutation techniques, one could use imputation to produce null datasets for spatial or spatiotemporal data with missing values.

Like geovisualization, which constructs knowledge through interactive maps and solves the problem of limited exploratory capability, graphical inference constructs knowledge through small-multiple maps and solves the problem of limited inferential capability. Graphical inference is "a tool for skepticism that can be applied in a curiosity-driven context" (Wickham

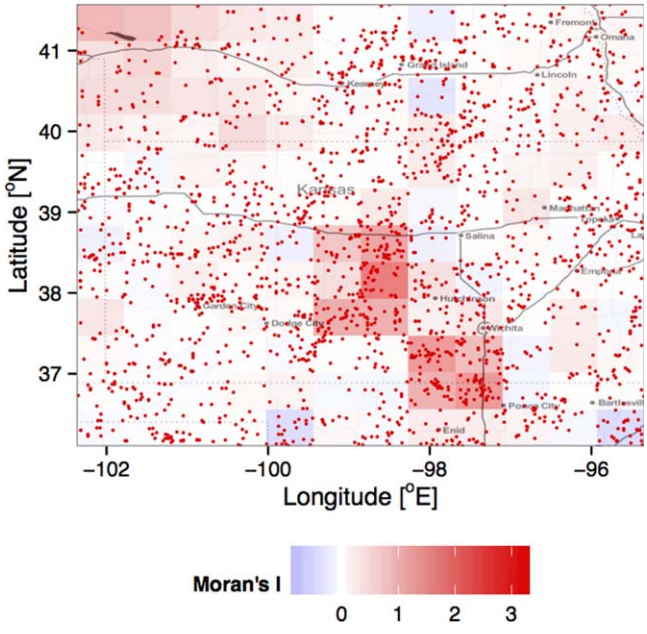

Figure 7. Local Moran's I computed from a raster of tornado touchdown frequencies.

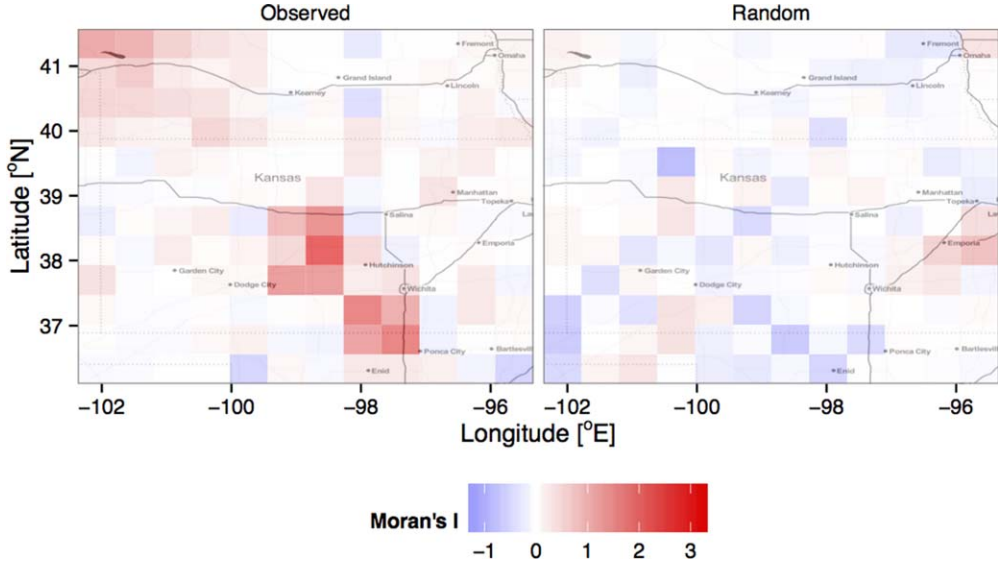

Figure 8. Local Moran's I computed from rasters of observed and one sample of random tornado frequencies. Random frequencies are based on a uniform distribution of touchdown latitude and longitude.

et al. 2010). It provides a way to make new discoveries with data while controlling for our keen ability to see a pattern in noise. In practice, graphical inference involves visual comparison of a line-up of graphs, where a single graph of the actual data is randomly positioned among multiple graphs of sample data under the null hypothesis. The significance test is then conducted by determining if an impartial judge (or jury) can identify the real data. P-values

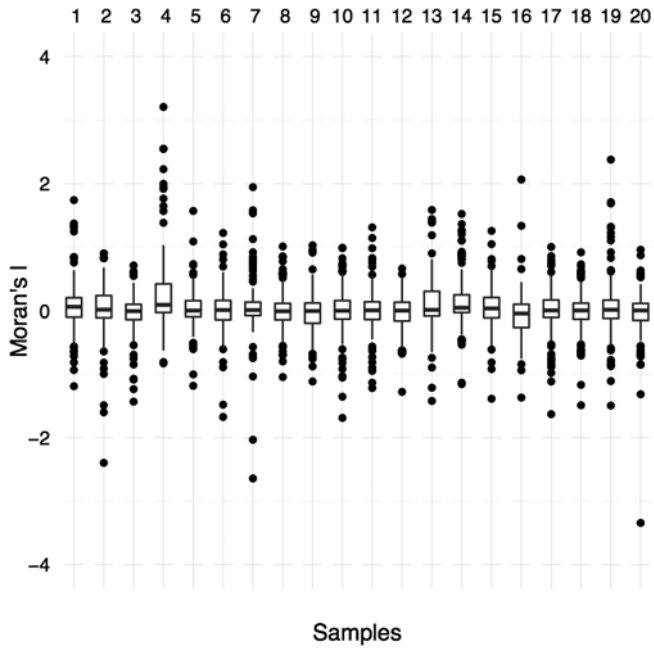

Figure 9. Local Moran's I of observed tornado touchdown frequencies and 19 samples of random frequencies. Can you pick out the box plot showing the autocorrelations from the observed tornado frequencies?

and power may also be calculated when implementing this methodology with independent observers evaluating the line-ups. If human subjects are used, removing the axis labels may provide greater neutrality.

Graphical inference does not replace formal statistical tests but can be used in conjunction with the tests as a "check" against a weak null hypothesis (straw man). It forces the researcher to think carefully about the null hypothesis through visualization. Using graphical inference may reduce the chance of making Type I (false positive) and Type II (false negative) errors, which tend to arise with conventional hypothesis testing (Buja et al. 2009; Wickham et al. 2010). In fact, Majumder, Hofmann, and Cook (2013) find the line-up protocol to outperform formal statistical tests when data do not meet the underlying test assumptions.

However, even if the data meet the assumptions, standard statistical tests only generate numerical results and can cause researchers to search for significant outcomes. This searching is commonly referred to as "p-hacking" or "significance-chasing" and is not the intended purpose of statistical testing (Nuzzo 2014). Graphical inference can help to deter p-hacking. A line-up comparison of the actual statistic with a statistic generated from the null hypothesis provides information on the size or magnitude of the effect that is not available with a *P*-value (although it may be provided by another test statistic). A small *P*-value from a statistical test may be associated with a very small actual effect and thus not particularly relevant. With graphical inference, data exhibiting a marginally significant small effect may be difficult to pick out among the decoys, as in the time-series example above. Thus, it reduces the opportunity to p-hack or chase small effects hidden in large datasets and make Type I errors. This is important today as pursuing small effects in noisy data is common (Simonsohn, Nelson, and Simmons 2013).

Bayesian methods also dispense with *P*-values by allowing researchers to incorporate their previous knowledge into their results as in Gelman (2004). Hierarchical Bayesian analysis allows multiple datasets to be assigned separate weights through partial pooling (Gelman 2010). Using multiple analysis methods may also assist in avoiding errors or erroneous *P*-values (Nuzzo 2014). Regardless of the method, moving beyond *P*-values allows us opportunities to publish negative findings, which may be important although not statistically significant.

Correspondingly, researchers need to be honest and detailed when reporting results, such as magnitude of effect, confidence intervals, and so forth, to efficiently communicate the importance of their findings. This promotes transparency and reproducibility, which may assist in avoiding errors and misinterpretations (Nuzzo 2014). Gelman (2012) illustrates an example where a researcher purposely deemphasizes nonsignificant results (by not disclosing them) causing misinterpretation of the presented *P*-values. The *P*-value is the main focus of that study and reproducibility would be difficult. Using code to conduct research guards against these issues and yields transparency and reproducibility (Wand et al. 2014). R code is relatively easy to understand without having previous programming skills (as it has many semi-automated functions) and the software is open-source. This is critical as the future calls for open science, where data, software, and publications are open and collaboration is essential (Rey 2014).

In addition, the visual setting of statistical inference makes it easier to communicate in a presentation and a poster as visual representation is an easily understood form of communication (Wand et al. 2014). Thus, graphical inference can be a helpful pedagogical tool for self-teaching and for students. For instance, this method can help determine arbitrary features that we may falsely identify as signal. Hence, it assists in developing an understanding of natural randomness in data structure, which is a key scientific concept (Buja et al. 2009 and Wickham et al. 2010).

A practical limitation of graphical inference is the finite number of automated R functions for performing it. In most situations that go beyond a normal density or a simple reordering, the researcher will need to create the null hypothesis and null datasets themselves. Another limitation is not all tests can be used (graphed or mapped) for visual analysis or inference. Wickham et al. (2010) list many possibilities including creating treemaps for analyzing distributions, choropleth maps for spatial trends, scatterplots for associations, and so forth. However, this method is not well suited for multiple variable analyses, hierarchical relationships, and dealing with confounding factors or random effects. In these cases, Bayesian inference techniques (Gelman et al. 2003, 2004) or other visual analysis methods, such as quilt plots (Wand et al. 2014), may be useful.

Acknowledgements

We would like to thank the anonymous reviewers for their assistance with improving this manuscript.

References

Anderson, C. J., C. K. Wikle, Q. Zhou, and J. A. Royle. (2007). "Population Influences on Tornado Reports in the United States." *Weather & Forecasting* 22(3), 571–79.

Anselin, L. (1995). "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27(2), 93–115.

Anselin, L. (1999). "The Future of Spatial Analysis in the Social Sciences." *Geographic Information Sciences* 5(2), 67–76.

Buja, A., D. Asimov, C. Hurley, and J. A. McDonald. (1988). "Elements of a Viewing Pipeline for Data Analysis." *Dynamic Graphics for Statistics, 277–308*, edited by W. S. Cleveland and M. E. McGill. Belmont, CA: Wadsworth.

Buja, A., D. Cook, and D. Swayne. (1999). "Inference for Data Visualization." In *Talk given at Joint Statistical Meetings*. Baltimore, Maryland. Accessed 12 March 2014 from http://www-stat.wharton.upenn.edu/~buja/PAPERS/visual-inference.pdf.

Buja, A., D. Cook, H. Hofmann, M. Lawrence, E. K. Lee, D. F. Swayne, and H. Wickham. (2009). "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367(1906), 4361–4383.

Chen, J., S. L. Shaw, H. Yu, F. Lu, Y. Chai, and Q. Jia. (2011). "Exploratory Data Analysis of Activity Diary Data: A Space–Time GIS Approach." *Journal of Transport Geography* 19(3), 394–404.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Monterey, CA: Wadsworth Advanced Books and Software.

Cleveland, W. S. (1993). "A Model for Studying Display Methods of Statistical Graphics." *Journal of Computational and Graphical Statistics* 2(4), 323–343.

Cortina, J. M., and W. P. Dunlap. (1997). On the Logic and Purpose of Significance Testing. *Psychological Methods* 2(2), 161–172.

Cressie, N. (1993). *Spatial Statistics*. Oxford, UK: Wiley.

DiBiase, D. (1990). "Visualization in the Earth Sciences." *Earth and Mineral Sciences* 59(2), 13–18.

Elsner, J. B., L. E. Michaels, K. N. Scheitlin, and I. J. Elsner. (2013). "The Decreasing Population Bias in Tornado Reports across the Central Plains." *Weather, Climate & Society* 5(3), 221–232.

Elsner, J. B., and H. M. Widen. (2014). "Predicting Spring Tornado Activity in the Central Great Plains by 1 March." *Monthly Weather Review* 142(1), 259–267.

Fox, P., and J. Hendler. (2011). "Changing the Equation on Scientific Data Visualization." *Science (Washington)* 331(6018), 705–708.

Gahegan, M., M. Wachowicz, M. Harrower, and T. M. Rhyne. (2001). "The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation." *Cartography and Geographic Information Science* 28(1), 29–44.

Gahegan, M., and B. Brodaric. (2002). "Computational and Visual Support for Geographical Knowledge Construction: Filling in the Gaps between Exploration and Explanation." In *Proceedings of the 10th International Symposium on Spatial Data Handling*, Ottawa, Canada.

Gelman, A. (2003). "A Bayesian Formulation of Exploratory Data Analysis and Goodness–of–Fit Testing*." *International Statistical Review* 71(2), 369–382.

Gelman, A. (2004). "Exploratory Data Analysis for Complex Models." *Journal of Computational and Graphical Statistics* 13(4), 755–779.

Gelman, A. (2010). "Bayesian Statistics then and Now." *Statistical Science* 25(2), 162.

Gelman, A. (2012). "P Values and Statistical Practice." *Epidemiology* 24(1), 69–72.

Guo, D., M. Gahegan, A. M. MacEachren, and B. Zhou. (2005). "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach." *Cartography and Geographic Information Science* 32(2), 113–132.

Head, M. L., L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions. (2015). "The Extent and Consequences of P-Hacking in Science." *PLoS Biology* 13(3), e1002106.

Held, I. M., and B. J. Soden. (2006). "Robust Responses of the Hydrological Cycle to Global Warming." *Journal of Climate* 19, 5686–5699.

Hijmans, R. J., and J. van Etten. (2011). "Raster: Geographic Analysis and Modeling with Raster Data." *R Package Version 1.8-39*, http://CRAN.R-project.org/package=raster.

Hochberg, Y., and A. C. Tamhane. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Hofmann, H., L. Follett, M. Majumder, and D. Cook, (2012). "Graphical Tests for Power Comparison of Competing Designs." *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2441–2448.

Hothorn, T., and B. S. Everitt. (2009). *A Handbook of Statistical Analyses Using R*, 2nd ed. Boca Raton, FL: CRC Press.

John, L. K., G. Loewenstein, and D. Prelec, (2012). "Measuring the prevalence of questionable research practices with incentives for truth telling." *Psychological Science* 23(5) 524–32.

Keirn, D., and H. P. Kriegel. (1994). "VisDB: Database Exploration Using Multidimensional Visualization." *IEEE Computer Graphics and Applications*, 14(5), 40–49.

Koua, E. L., A. MacEachren, and M. J. Kraak (2006). "Evaluating the Usability of Visualization Methods in an Exploratory Geovisualization Environment." *International Journal of Geographical Information Science* 20(4), 425–48.

Kwan, M. P. (2000). "Interactive Geovisualization of Activity-Travel Patterns using Three-Dimensional Geographical Information Systems: A Methodological Exploration with a Large Data Set." *Transportation Research Part C: Emerging Technologies* 8(1), 185–203.

MacEachren, A. M., and D. R. F. Taylor (edited by). (1994). *Visualization in Modern Cartography*, Vol. 2. Oxford, U.K.: Pergamon Press.

MacEachren, A., D. Xiping, F. Hardisty, D. Guo, and G. Lengerich. (2003). "Exploring high-D spaces with multiform matrices and small multiples." In *Proceedings of the International Symposium on Information Visualization*, Seattle, Washington.

MacEachren, A. M., M. Gahegan, W. Pike, I. Brewer, G. Cai, E. Lengerich, and F. Hardistry. (2004). "Geovisualization for knowledge construction and decision support." *IEEE Computer Graphics and Applications*, 24(1), 13–17.

Majumder, M., H. Hofmann, and D. Cook. (2013). "Validation of visual statistical inference, applied to linear models." *Journal of the American Statistical Association* 108(503), 942–956.

Nuzzo, R. (2014). "Scientific method: Statistical errors." *Nature* 506(7487), 150–152.

R Core Team. (2013). *"R: A language and environment for statistical computing." R Foundation for Statistical Computing*, Vienna, Austria. Available online: www.r-project.org (accessed on December 2013).

Rey, S. J. (2014). "Open regional science." *The Annals of Regional Science* 52(3), 825–837.

Shneiderman, B. (2014). "The big picture for big data: Visualization." *Science (New York, NY)* 343(6172), 730.

Simonsohn, U., L. D. Nelson, and J. P. Simmons. (2013). "P-Curve: A key to the file-drawer." *Journal of Experimental Psychology: General* 143(2), 534–547.

Tukey J. W. (1977). *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, Connecticut: Graphics Press.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz. (2006). "Evolution of the US Tornado Database: 1954–2003." *Weather and Forecasting* 21(1), 86–93.

Wand, H., J. Iversen, M. Law, and L. Maher, (2014). "Quilt Plots: A Simple Tool for the Visualisation of Large Epidemiological Data." *PloS One* 9(1), e85047.

Wang, S., L. Anselin, B. Bhaduri, C. Crosby, M. F. Goodchild, Y. Liu, and T. L. Nyerges. (2013). "CyberGIS Software: A Synthetic Review and Integration Roadmap." *International Journal of Geographical Information Science*, 27(11), 2122–2145.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer Science & Business Media.

Wickham, H., D. Cook, H. Hofmann, and A. Buja. (2010). "Graphical inference for infovis." *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 973–979.

Wickham, H., N. R. Chowdhury, and D. Cook. (2014). "nullabor: Tools for Graphical Inference." *R Package Version 0.3.1*.

Wilkinson, L., D. Wills, D. Rope, A. Norton, and R. Dubbs. (2006). *The Grammar of Graphics*. Secaucus, New Jersey: Springer-Verlag.

Whittaker, R. H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California." *Ecological Monographs* 30(3), 279–338.