Progress in Human Geography
2017, Vol. 41(4) 512–523

DOI: 10.1177/0309132516648020
journals.sagepub.com/home/phg
SAGE

Quantitative methods II: Issues of inference in quantitative human geography

Christopher Brunsdon
Maynooth University, Ireland

Abstract
Although classical significance testing is the most commonly used inferential technique in quantitative geography, it is far from the only choice, and in some circumstances may not be the most appropriate. In the statistical literature and other disciplines, its utility has come under question in a number of contexts. This report overviews current progress in the development of quantitative inferential approaches, and considers their use and appropriateness in a number of human geographical contexts.

Keywords
Bayesian, inference, statistical testing, visualization

I Inference matters for quantitative geographers

In a time where terms such as *data science* are applied widely, the discipline of statistics is sometimes overlooked (Widen et al., 2015). However, one concern in 'traditional' statistics of importance in human geography is that of inference. Data analysis in human geography is carried out to understand the world – measurements, surveys and various forms of volunteered information are explored, analysed and modelled to generate insights into underlying processes and situations. Inference in this framework is the process of making deductions about the latter by observing the outcome of the former. It is therefore an essential part of quantitative geographical research.

It is timely to consider inference for a number of reasons. Firstly, within statistics – and by implication *spatial* statistics – standard statistical approaches to inference (classical inference) are being questioned. More generally, there is a longstanding debate in geography – perhaps initiated by Gould (1970), and re-visited by Brunsdon (2001), scrutinizing current practice as a tool for geographers – focusing on assumptions of *homogeneity* (characteristics of data being the same everywhere) and *spatial independence* (events at one location are not correlated to nearby events). Secondly, data scientists are also questioning these approaches, for different reasons (e.g. Gigerenzer, 2004). Here, criticism rounds more broadly on the fact that 'traditional' hypothesis tests do not necessarily provide answers to the questions that data analysts frequently need to ask – a 'right answer, wrong

Corresponding author:
Christopher Brunsdon, National Centre for Geocomputation, Maynooth University, National Centre for Geocomputation, NCG Iontas Building, NUI Maynooth, Ireland.
Email: christopher.brunsdon@nuim.ie

question' situation. Thirdly, a number of observers report that classical inference is widely misunderstood (Sterne and Davey Smith, 2001; Sterne, 2002). Relatedly, some studies stress the idea of correctly interpreting the testing procedures, rather than applying them mechanistically (see Spicer and Gangloff, 2016).

II Key ideas of classical statistical inference

1 The 'classical' approach and p-values

For many, the most familiar concept of statistical inference is the *p*-value, although arguably this is frequently explained as part of a procedure (plug numbers into a formula, check whether *p* is greater than 0.05 or not) rather than as an *idea*. The underlying idea is expressed by Wickham et al. (2010: 974), who state:

> Unlike the criminal justice system, in the statistical justice system (SJS) evidence is based on the similarity between the accused and known innocents, using a specific metric defined by the test statistic. . . . To determine the guilt of the accused we compute the proportion of innocents who look more guilty than the accused. This is the *p*-value, the probability that the accused would look this guilty if they were actually innocent.

The SJS analogy is helpful to outline some issues with the use of *p*-values as set out below (although it suggests scepticism if convicting on the basis of statistical evidence in an actual courtroom).

2 What does a p-value tell us?

A *p*-value measures the plausibility of the null hypothesis. Of itself, it doesn't imply anything *directly* about alternative hypotheses. Of course, for a given null hypothesis, there are many alternatives. Indirectly, potential alternative hypotheses do influence the testing procedure, mainly through the choice of the test statistic. But it is quite possible to test null hypotheses that spatial parameters are zero in both spatial lag and spatial error models, with both tests rejecting H_0. One may conclude there is some kind of spatial process occurring, but little has been learned about *which* spatial process model might apply. So in a sense, this takes only the very first footsteps toward understanding an underlying spatial process. In terms of the SJS analogy, this is fine if we know what charge is levelled at the accused, but offers no way of determining what the crime was, if it is not known at the outset. The above tests screen for the fact that there is no spatial process – an important task but one which fails to determine what kind of spatial process *is* occurring.

3 Multiple significance tests

Another important issue is the multiple use of significance tests. This is especially relevant for geographers concerned with identifying spatial clusters such as 'hot spots' of disease or crime, since it arises in the use of scan statistics such as the Geographical Analysis Machine (GAM) of Openshaw et al. (1987) or the scan statistic of Kulldorf (1997). Rather than using data to assess a single hypothesis or identify a specific model, the aim is to look in a number of geographical locations to identify anomalies from the norm, to answer questions such as: 'Are there certain places where the incidence of some disease is notably higher than the population average?'

A null hypothesis that the local occurrence rate is the population average is applied. Places in which the *p*-value falls below some given threshold are mapped. However, the *p*-values refer to the chance of a false positive when there is just *one* test. When there are several tests, this must be allowed for. If there were 100 tests, and the threshold for significance is $p = 0.05$ then on average, in a world where H_0 is true *everywhere*, one would expect to see on average five significant results. Thus, for the question 'Are there *any* clusters of this disease?', a false positive answer is quite likely. Returning to the SJS analogy, if 100 innocent people were tried, then, given the 'trial' is based on the characteristics of the innocent population, around five of them would 'look guilty' despite not being so. If the results of the trials were used to answer the question 'Did anyone commit this crime?', it is quite likely that a completely innocent group of defendants could provide the wrong answer.

How this is addressed depends on the research question. If it is of the form above – 'Are there *any* exceptions to H_0? – then consider the probability of obtaining at least *one* significant result for a given p-threshold. This is referred to as the Family Wise Error Rate (FWER). One way of dealing with the multiple testing problem is to set the FWER at a desired level (say 0.05) and work backwards to identify an individual level p (see Šidàk, 1967). To get an FWER of 0.05, individual tests must have a much lower threshold for p. This is the approach adopted by Getis and Ord (1992).

However, this comes at a price. Evidence against H_0 for each test must be *very* strong to merit rejecting the hypothesis. However, this has implications for the *type II error*: p-values are defined in terms of the *type I error* – the probability of rejecting H_0 when it is actually true – whilst the type II error is the probability of not rejecting H_0 when it is false. Going back to the SJS analogy, type II error is the probability of convicting an innocent defendant, whilst type I error is the probability of acquitting a guilty party. Since very strong evidence against H_0 is required for multiple testing based on the FWER, smaller deviations from H_0 are less likely to be detected. In short, in reducing the risk that *any* of the tests return a false positive, the chances of false negatives (i.e. failure to identify a genuine anomaly) are notably increased. Such a procedure is termed *conservative,* as it will flag anomalies only with a great deal of caution, preferring to fail to identify them rather than falsely flag a non-anomalous situation. For example Wheeler (2007) notes in a study of childhood leukaemia in Ohio: 'the spatial scan method [. . .] does not find statistically significant local clusters, while the kernel intensity function method suggests statistically significant clusters in areas of central, southern, and eastern Ohio'.

This is reasonable if the key research question is to attempt to demonstrate the existence (or otherwise) of, say, clusters of some disease. However in many situations, this is not the case. Wheeler states:

> *Numerous studies have focused on childhood leukemia because of its relatively large incidence among children compared with other malignant diseases, its apparent tendency to cluster, and the substantial public concern over locally elevated leukemia incidence.* (emphasis added)

This suggests that, in this case, the key objective is to scan for potential clusters of a disease that has exhibited clustering in other studies. The issue is not the existence of clustering, but whether there are any clusters in Ohio. In this context, the FWER is perhaps an inappropriate metric. A more helpful measure of reliability may be the False Discovery Rate (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). This is the probability that, if a test flags an anomaly, it does so incorrectly. This differs from the FWER, which measures the chances that *any* false alarms occur. In the SJS analogy this is the probability of a mistaken conviction, given a conviction has occurred. A mathematical difference is that the denominator of this rate is the number of all significant results, regardless of whether H_0 is true. For FWERs, the denominator is the number of cases where H_0 is true. The adjustment in this case is not a simple formula to replace each unprocessed p-value, but generally the individual tests are less conservative. This approach is relatively rare in geographical studies, but see Brunsdon and Charlton (2011) and Caldas de Castro and Singer (2006) for examples.

III Broader inference issues

Here I review alternatives to classical inference; the various situations in which they may be used will be considered, and their strengths and weaknesses.

1 Exploratory data analysis

The points above relate to difficulties geographers may have using classical approaches in quite specific situations – and could be summarized by calling for attention to the nature of the hypotheses being tested, and a move away from 'button clicking' on statistical software to provide *p*-values with insufficient regard to the meaning of a significant result. However, there are broader debates questioning the nature of significance testing itself. Firstly, in the past century statistical theory relating to a number of approaches exists, but disproportionate attention has been focused on significance testing. Gigerenzer (2004) notes:

> Textbooks and curricula (p. 588) in psychology almost never teach the statistical toolbox, which contains tools such as descriptive statistics, Tukey's exploratory methods, Bayesian statistics, Neyman–Pearson decision theory and Wald's sequential analysis.

Similar could be said of geography, with the exception of descriptive statistics. However, Tukey (1977)'s approaches, Bayesian inference and decision theory all have potential uses in geographical data analysis. For example, Willmott, Robeson and Matsuura (2007) modify Tukey's idea of a box plot to explore geographical data. This is a box plot modified so that each observation is weighted by the physical land area associated with it. This report proposes a physical geography technique, but by substituting physical land area for population, the method could be used to explore human geography data. For example, in Figure 1 a population-weighted box plot is shown above a standard box plot for percentage of households adopting

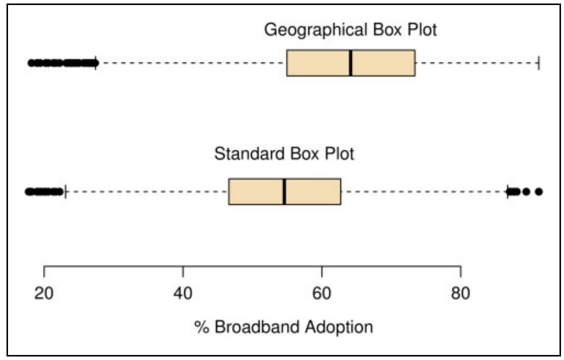

Figure 1. Standard vs. geographical boxplots.

broadband for Irish electoral districts in the 2011 census. Taking population weighting into account shows that on a 'per-household' basis, the uptake is somewhat higher than suggested with the standard boxplot, although a skew distribution with a left-hand tail having very low levels of adoption is seen.

The outlying EDs from the geographical box plot are shown as a map in Figure 2. It illustrates that many of these low uptake areas are in rural parts of Ireland, particularly in the midlands. Although no formal approach is employed here, this inferential technique is clearly useful – identifying noteworthy geographical patterns and possibly informing policy.

Another important idea for exploratory data analysis is the *cartogram*, in particular maps whose projections are designed to reflect the underlying population in each region, rather than their physical size. A recent algorithm is given by Sun (2013). Dorling, Barford and Newman (2006) make powerful arguments for the use of this approach. In Figure 3, Gastner and Newman's (2004) algorithm is used to show the low broadband uptake outliers. Here the inferential purpose of the visualization is a precautionary one. The EDs in more populous parts of Ireland are relatively small. On a standard map it may be possible that a shaded area (corresponding to an outlier) may not be seen. On the cartogram, no such outliers occur.

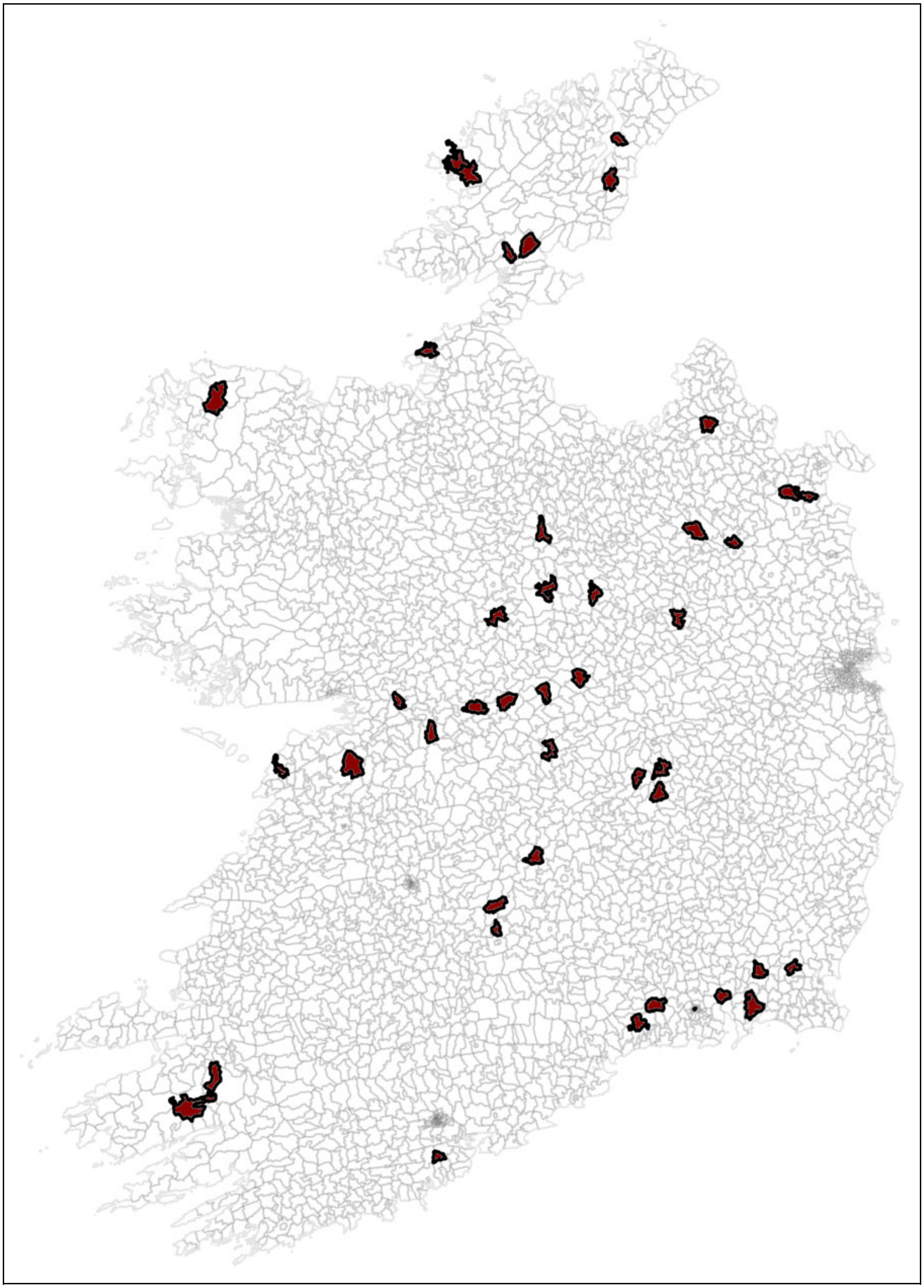

Figure 2. Conventional map of outliers.

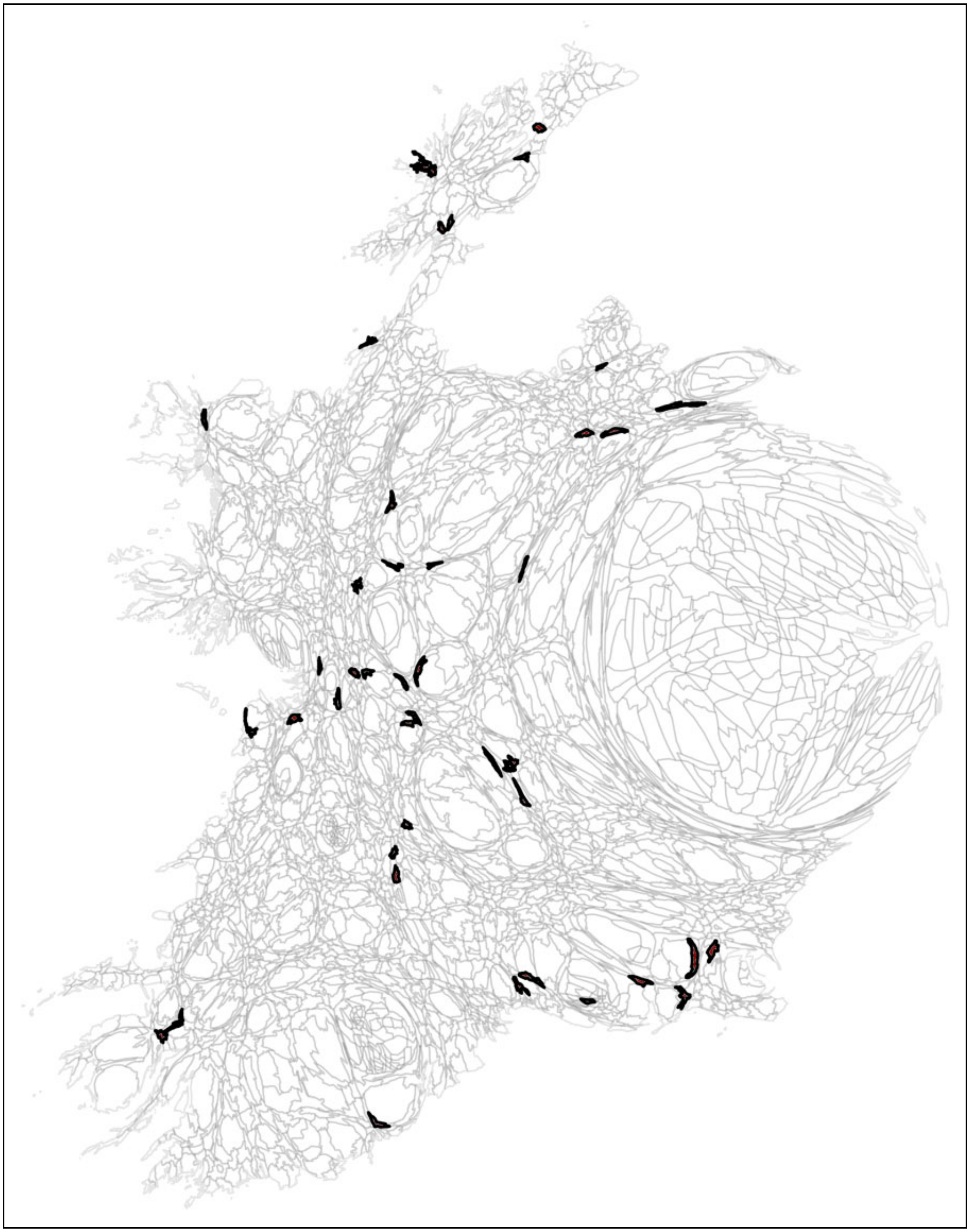

Figure 3. Cartogram of outliers.

2 A Bayesian approach

The EDA approach considered above provides a less formal approach to inference and is particularly useful if no prior hypotheses have been formulated. But the significance testing procedures considered earlier are not the only formal approach. One alternative being widely adopted is the Bayesian approach (see Withers, 2002, for a useful discussion on its adoption in human geography). In recent years, Rohde, Corcoran and Chhetri (2010) demonstrate the use of Bayesian inference to analyse patterns in the occurrence of urban fires, and Jonker et al. (2013) apply this approach to analyse geographical variation in life expectancy.

The most contentious aspect of this approach is the inclusion of prior beliefs in the analysis – which can be seen as adding a subjective element to an approach generally intended to be objective. Although there are many mathematical parallels between Bayesian and classical approaches, philosophically, the meaning of probability differs fundamentally between them (Spiegelhalter, 2004). Due to this distinction and some other factors, the Bayesian approach can be seen as extending the portfolio of analytical techniques rather than replacing existing 'classical' approaches. Others also argue that the subjective element of Bayesian inference is necessary: 'Bayesian statistics treats subjectivity with respect by placing it in the open and under the control of the consumer of data' (Berger and Berry, 1988: 163).

Classical inference *does* have some subjective elements, such as the choice of variables to include in a regression model, or the choice of sampling methods in a survey. By explicitly including a subjective prior distribution, a more open treatment of this occurs. Others often employ a *noninformative prior* – typically a uniform distribution – to represent a state of no prior knowledge. In such cases, the posterior distribution is often similar to the *likelihood function* of classical theory, although the meaning of the two expressions differs notably between the two approaches.

A key practical advantage of the Bayesian approach is the ability to use simulation-based methods such as Markov Chain Monte Carlo (MCMC) to generate random numbers drawn from the posterior for relatively complex models that may not be evaluated analytically. For example, Wheeler et al. (2014) use this approach to model house prices so that complex and more realistic geographical processes may be incorporated, in which the connections between drivers of house price and the price itself vary geographically.

3 The Akaike Information Criterion

A quite different approach to inference is evident in Akiake's (1973) Information Criterion (AIC). Rather than focusing on parameter estimates or hypotheses tests, a key goal of this tool is model selection (Burnham and Anderson, 2002, 2004). Fitting a number of models using maximum likelihood, the AIC of each model is computed as

$$2k - 2\log(L) \tag{1}$$

where L is the likelihood of the model, and k is the number of parameters. Very generally, lower AICs indicate better models. The criterion is derived in terms of information theory, and tries to estimate the information loss in approximating the process that generated the data by each of the models under consideration. None of the models is considered to be 'true', and the aim is to mitigate information loss by choosing appropriate models. Generally, advice is not always to choose the model having the lowest AIC without question, but to consider relative differences in AICs. If the 'winner' is clearly ahead of the others then this model is favoured. But in less clear circumstances it may be that the set of appropriate models is narrowed down to a smaller 'shortlist' rather than one individual model. Hirschfield et al. (2014) offer an example of the use of the AIC (alongside other inferential tools), applying a number of spatial models to patterns in urban crime rates. As well as simply detecting that some form of spatial process is at work, this allows some degree of comparison as to which spatial process is most likely, and to use this as an approach towards gaining better understanding of underlying social processes and human behaviour patterns driving crime patterns.

4 Graphical inference

A very recent development is the concept of *graphical inference* (Wickham et al., 2010). It is based on the idea that if a statistical model encapsulates a real world process faithfully, synthetic data generated from it should 'look like' real world data. Thus, maps of house prices simulated from a hedonic model should look like maps of real house prices. Although precise definitions of 'looks like' may be elusive, one approach is to use several human subjects to assess similarity. Subjects view a number of maps (or other graphic displays), one of which was created using real world data, whilst the others were simulated. Each is asked to identify which map is the 'odd one out'. If the model is a poor reflection of reality, a high proportion of subjects should easily identify the real world. For example, in Figure 4 six maps show patterns in reported household burglary – where either forced entry was used (grey) or no force was used to enter the property (black). One map is based on hypothetical data where unforced entry was more common closer to the centre of the study area. The other five aspatial models were created by randomly assigning forced or unforced to each burglary location in the same proportion. (The top left map differs from the rest – although of course in a real example this would have to be put to the test).

The approach, although novel in some ways, has a very similar framework to classical hypothesis testing. Returning to the statistical justice system (SJS) analogy, the graphical test works like an identity parade, and consensus is sought as to whether any member appears to be the guilty party. Following the original phrase, consensus is actually sought as to whether any member 'looks' more guilty. Also, although this appears to be an informal approach, it can be re-cast in a more formal way. If the survey in which observers are asked to identify the distinct item is a randomized trial, and a formal significance test of the hypothesis that no map is more likely to be chosen is used, the result is no less valid than any other well-founded trial.

However, there are still methodological questions. When choosing a particular symbology on maps, are some kinds of pattern more visually striking than others? In which case, although the trial may be well founded, it may not be very powerful, with a high type II error. Also, are there biases associated with the location of the 'true' map – are people more likely to choose the top left map regardless of pattern, for example? Widen et al. (2015) consider a number of these issues in a geographical context, together with a number of applications (see also Brunsdon and Comber, 2015). Despite the need to consider some methodological questions, this approach holds promise. It allows complex models to be assessed, since simulation of the process being modelled is required, rather than a full maximum likelihood approach in order to perform a classical hypothesis test.

5 Machine learning

The machine learning approach, although very computational, differs from the classic, Bayesian, AIC and graphical inference approaches in that it does not require a *generative model* – that is, a statistical model for the data, outlining a stochastic process that could have generated it. Machine learning focuses more on using algorithms to detect patterns in data, rather than on calibrating models or testing formal hypotheses. This emphasis on pattern detection tends to lead to a stronger focus on either prediction (where patterns relating to a response variable and a number of predictor variables are sought) or exploration (where patterns are sought more generally). A distinction between this method and the others listed is the generality of what constitutes a pattern. As outlined in Gahegan (2003: 70):

> Inferential statistics uses observations to condition (shape) the form of a distribution model that is usually provided by the analyst.... By

Figure 4. Example of graphical inference.

contrast, many machine learning techniques construct a distribution model using evidence gleaned from the data alone, i.e. they are data-driven.

This data-drivenness is a strength and a weakness. The danger lies in the idea that this leads to an 'end of theory' (Anderson, 2008) – inference in this world-view consists merely of identifying regularities in the data, rather than offering any explanation of how they came into being. This viewpoint is not without criticism (Kitchin, 2014; Brunsdon, 2015). However, even when this is considered, the application of machine learning techniques can still play some inferential role – perhaps in identifying patterns, whose explanation can then be sought. Inference here is perhaps taking on the role of hypothesis generator rather than hypothesis evaluator.

IV Conclusions

This report has provided an overview of inferential tools for quantitative geography, ranging from the very traditional (such as significance testing), to the emergent at the time of writing (graphical inference). Many textbooks in quantitative geography place emphasis on the traditional approaches. However, new methods are being derived, and in the statistical literature and beyond, the universal application of well-established procedures such as significance testing are being questioned. It is timely for geographers working with quantitative data to take stock of this, and consider which – if any – of the inferential tools currently being developed best suit their needs. Some of the newer approaches offer promise. A major inferential task that could be usefully addressed is that of distinguishing which kind of spatial process is the most appropriate model. Tests of a null hypothesis of spatial independence are essentially screening for situations in which there actually *is* a geographical process. But to many geographers, once this basic situation is established a more interesting question is: 'what kind of geographical process is occurring?' – or 'is a particular spatial model appropriate?'

Some of the inferential approaches outlined here may be more appropriate than others. For example, a graphical inference approach could be used to compare simulated data from a given spatial process with real data – rather than simulated data without spatial pattern. Classical inferential processes could be used to test compound models, with components from several kinds of process, for example, a model with both a spatial error and a spatial lag term, both of whose coefficients could be tested – although this increase in model complexity has consequences for the sampling distribution of parameters and the type II error of the tests involved. Some of the ideas could be adapted to answer different kinds of question. AIC seems to be best suited to the model selection question – 'which of a set of candidate models are most appropriate?' – but graphical inference could also address this problem. If the real data was in map 1, and maps 2 to n were of data generated from competing candidate models, observers could be asked to state which of maps 2 to n most resembled map 1. Inference could then be drawn from the proportions prefering each map about its associated data generating model's plausibility as a generator of the actual data. In summary, this is a call for geographers to help shape development in inferential tools for spatial data, by looking more critically at existing approaches, and considering how they may be adapted (or entirely new approaches developed) to best answer questions arising in geography.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Akaike H (1973) Information theory and the maximum likelihood principle. In: Petrov BN and Csaki F (eds) *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267–281.

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Available at: http://www.wired.com/2008/06/pb-theory/ (accessed 20 April 2016).

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57(1): 289–300.

Benjamini Y and Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29(4): 1165–1188.

Berger JO and Berry DA (1988) Statistical analysis and the illusion of objectivity. *American Scientist* 76: 159–165.

Brunsdon C (2001) Is 'statistix inferens' still the geographical name for a wild goose? *Transactions in GIS* 5(1): 1–3.

Brunsdon C (2015) Quantitative methods I: Reproducible research and human geography. *Progress in Human Geography*. DOI: 10.1177/0309132515599625.

Brunsdon C and Charlton M (2011) An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection. *Environment and Planning (B)* 38: 216–230. DOI: 10.1068/b36093.

Brunsdon C and Comber L (2015) *An Introduction to R for Spatial Analysis and Mapping*. London: SAGE.

Burnham KP and Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed*. New York: Springer-Verlag.

Burnham KP and Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research* 33(2): 261–304.

Caldas de Castro M and Singer B (2006) Controlling the false discovery rate: A new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis* 38(2): 180–208. DOI: 10.1111/j.0016-7363.2006.00682.x.

Dorling D, Barford A and Newman M (2006) Worldmapper: The world as you've never seen it before. *IEEE Transactions on Visualization and Computer Graphics* 12(5): 757–764.

Gahegan M (2003) Is inductive machine learning just another wild goose (or might it lay the golden egg)? *International Journal of Geographical Information Science* 17(1): 69–92.

Gastner MT and Newman MEJ (2004) Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America* 101(20): 7499–7504.

Getis A and Ord K (1992) The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24(3): 189–206.

Gigerenzer G (2004) Mindless statistics. *The Journal of Socio-Economics* 33: 587–606.

Gould P (1970) Is *statistix inferens* the geographical name for a wild goose chase? *Economic Geography* 46: 439–448.

Hirschfield A, Birkin M, Brunsdon C, Malleson N and Newton A (2014) How places influence crime: The impact of surrounding areas on neighbourhood burglary rates in a British city. *Urban Studies* 51(5): 1057–1072.

Jonker MF, Congdon P, Van Lenthe F, Donkers B, Burdorf A and Mackenbach J (2013) Small-area health comparisons using health-adjusted life expectancy: A Bayesian random-effects approach. *Health and Place* 23: 70–78.

Kitchin R (2014) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 3(3): 262–267.

Kulldorf M (1997) A spatial scan statistic. *Communications in Statistics: Theory and Methods* 26: 1481–1496.

Openshaw S, Charlton M, Wymer C and Craft A (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1(4): 335–358.

Rohde D, Corcoran J and Chhetri P (2010) Spatial forecasting of residential urban fires: A Bayesian approach. *Computers, Environment and Urban Systems* 34(1): 58–69.

Šidàk Z (1967) Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62: 626–633.

Spicer RC and Gangloff HJ (2016) Permutation/randomization-based inference for environmental data. *Environmental Monitoring and Assessment* 188(3): 1–12. DOI: 10.1007/s10661-016-5090-0.

Spiegelhalter DJ (2004) Incorporating Bayesian ideas into health-care evaluation. *Statistical Science* 19(1): 156–174.

Sterne JAC (2002) Teaching hypothesis tests – Time for significant change? *Statistics in Medicine* 21(7): 985–994.

Sterne JAC and Davey Smith G (2001) Sifting the evidence: What's wrong with significance tests? *British Medical Journal* 322: 226–231.

Sun S (2013) An optimized rubber-sheet algorithm for continuous area cartograms. *The Professional Geographer* 65(1): 16–30.

Tukey JW (1977) *Exploratory Data Analysis*. London: Addison-Wesley.

Wagenmakers E-J and Farrell S (2004) AIC model selection using Akaike weights. *Psychonomic Bulletin and Review* 11(1): 192–196.

Wheeler DC (2007) A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996–2003. *International Journal of Health Geographics* 13(6). DOI: 10.1186/1476–072X–6–13.

Wheeler DC, Páez A, Spinney J and Waller LA (2014) A Bayesian approach to hedonic price analysis. *Papers in Regional Science* 93(3): 663–683.

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

Widen HM, Elsner JB, Pau S and Uejio C (2015) Graphical inference in geographical research. *Geographical Analysis*. DOI: 10.1111/gean.12085.

Willmott CJ, Robeson SJ and Matsuura K (2007) Geographic box plots. *Physical Geography* 28(4): 331–344.

Withers SD (2002) Quantitative methods: Bayesian inference, Bayesian thinking. *Progress in Human Geography* 26(4): 553–566.