Transactions in GIS, 2001, 5(1): 1–3

Guest Editorial

Is 'Statistix Inferens' Still the Geographical Name for a Wild Goose?

After the recent death of Peter Gould, I decided to look once again at his insightful paper 'Is Statistix Inferens the Geographical Name for a Wild Goose?' (Gould 1970) – hence the title for this guest editorial. For those readers who have not seen this article, Gould outlines a number of shortcomings of the common statistical practices of geographers of the day. Many of these relate to the assumptions made in the statistical models used:

- The functional form used in regression models (typically linear)
- The random nature of the data sample
- The probability distributions of random variates and error terms (typically normal)
- The assumption of independence of random variates and error terms.

He argues that often these assumptions are inappropriate in the geographical context, and goes on to question the validity of statistical inferences made using techniques based on these assumptions. In particular he finds the assumption of independence problematic in a spatial context. I found the paper to be as thought provoking now as it must have been when it was first published 30 years ago.

One obvious question that arises is 'to what extent have these issues been addressed since 1970?'. The answer is not a particularly simple one. The years between 1970 and now have seen the creation of a large number of new statistical techniques. Other techniques in their infancy in 1970 have now reached a stage of maturity. Many of these relax one or more of the above assumptions – generalized additive modelling, nonparametric regression, kernel density estimation, randomisation tests and regression models with autocorrelated errors readily come to mind, and these by no means provide an exhaustive list.

So why do I claim the answer to my earlier question is not simple? Firstly, one must ask to what extent these methods have diffused from the statistical community into the GIS literature. Although this has happened to some extent, there are still plenty of studies published which rely on non-spatial data analysis techniques such as least squares regression modelling. Why does the use of such methods continue? In some situations these methods may be genuinely appropriate – this could only be the case if the relative or absolute geographical locations associated with the data played no part

in the process under investigation. A geographer's instinct would lead one to suspect that this is rarely the case. To quote another paper from the Statistix Inferens era:

> '... everything is related to everything else, but near things are more related than distant things.' (Tobler 1970)

It seems reasonable that one should at least test for the presence of spatial effects in order to assess the suitability of an OLS approach. Unfortunately, in a number of studies no attempt has been made to do this.

Another reason for the continuing popularity of non-spatial methods is that they are readily available in mainstream statistical software packages, unlike their spatial counterparts. It seems that none of the larger GIS or statistical software manufacturers feel the need to include techniques such as spatially autocorrelated regression in their main packages. Some advocate data sharing between the packages, but this often leads to a process where data is transferred from a GIS to a statistics package, analysed using standard (non-spatial) techniques and returned to the GIS. This approach seems to fly in the face of Gould's analysis. At this stage, I should point out that a number of specialist packages for spatial statistics do exist, many of them excellent. However, I suspect that to gain widespread use we will need to see these techniques appear in the larger general purpose packages. Perhaps this issue is best resolved by lobbying the software manufacturers.

A second issue that arises from the paper is the very notion of the significance test. When based on flawed assumptions it is clear that significance tests are not helpful, but one can go on to consider their role even when they are based on reasonably sound assumptions. The problem here is that tests of whether some quantity is equal to zero – which most commonly applied tests amount to – are of less practical use than an attempt to estimate that quantity. There is a difference between a result being of statistical significance at the somewhat arbitrary 1% or 5% levels, and it being of practical significance. Although Gould considered this from a geographical viewpoint, there has been little debate about this in the GIS and quantitative geography literature. Conversely, it has become quite a 'hot topic' in the statistical community. Although it would be misleading to state that statisticians had abandoned the technique, there have been various levels of protest ranging from an admission that significance testing is 'over-rated' – as documented by Nester (1996) – to explicit assaults:

> 'The tyranny of the Neyman-Pearson theory in many branches of empirical science is detrimental, not advantageous, to the course of science.' (Wang 1993)

It might be interesting to consider the level of awareness of this debate in the GIS and spatial analysis communities. Clearly, these issues are of importance even to those of us using the 'best' spatial models. Perhaps it is time to re-open the debate that Gould began.

Next, it is worth noting some of the issues that have arisen since 1970 that also call into question some of the assumptions made in statistical models. One particular case in point is Stan Openshaw's modifiable areal unit problem (MAUP). As demonstrated, for example, in Openshaw and Taylor (1979) results of statistical analysis for spatially aggregated data can alter dramatically when the areal units change. Whereas this has little consequence for point based spatial statistical techniques such as kriging, the implications for area-based approaches require attention. For example, many spatial regression models use area based data, and model the spatial dependance of the

residuals on the connectivity matrix of the areal units. Changing these units will change the connectivity matrix and essentially change the model of the spatial process. In addition to this, for any given areal units there are a number of ways to define connectivity – as a binary indicator based on edge contiguity, or vertex contiguity, perhaps as a continuous variable based on distance decay between zonal centroids – the list goes on. Here we have a compounded MAUP – even if we fix the aggregation zones we still have an infinite selection of connectivity matrices! The typical approach seems to be to take the connectivity matrix as fixed a priori, but then considering inferences related to the other parameters in the model seems like only half solving the problem. In the spirit of Statistix Inferens we are prompted to question the utility of such inferences. The aim here is not to assert that these techniques are of no use – in general they present a great advance on the techniques that concerned Gould, providing statistical assumptions that are much closer to reality. However by scrutinising the validity of even these assumptions we can hopefully improve the situation even further.

Hopefully this discussion shows that Gould's paper still proves to be a stimulating read. The issues I have outlined here are only a small selection of some of the ideas that re-reading the paper has prompted. Other issues include the implications of spatial non-stationarity for some statistical techniques, and the somewhat patchy nature of the uptake of spatial statistical methods – why, for example, are there so few instances of spatial statistical methods for survey data? The issues addressed in 1970 are still important today. Although the emphasis in much GIS research is less explicitly placed on statistical inference, it still plays an important role. If we in the GIS community wish to behave scientifically, we need inferential tools. But, as Gould argues, we need appropriate inferential tools. These are tools which take geographical processes and geographical data collection issues into account. We've come a long way since 1970 but we still have some distance to go.

Chris Brunsdon
Department of Geography
University of Newcastle-upon-Tyne

References

Gould P R 1970 Is statistix inferens the geographical name for a wild goose? *Economic Geography* 46: 439–48

Nester M 1996 An applied statistician's creed. *Applied Statistics* 45: 401–10

Openshaw S and Taylor P 1979 A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In Wrigley N (ed) *Statistical Applications in the Spatial Sciences*. London: Pion: 127–44

Tobler W R 1970 A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–40

Wang C 1993 *Sense and Nonsense of Statistical Inference*. New York: Dekker