Correlation as a LISA

Martin Charlton*[1] and Chris Brunsdon[1]

[1]National Centre for Geocomputation, Maynooth University, Ireland

*Email:martin.charlton@nuim.ie

Abstract

This paper describes a decomposition of three commonly used measures of correlation, Pearson's r, Spearman's ρ and Kendall's τ. As the individual components sum to the global statistics these decompositions could the thought of as a LISA. We illustrate their use with an example using the Georgia educational attainment data. These coefficients are also special cases of a generalised correlation coefficient.

Keywords: correlation, local indicators of spatial association, Pearson's r, Spearman's ρ, Kendall's τ.

1. Correlation

`Correlations are difficult to assess and interpret' warns Chatfield (1995, p.167). Nevertheless they are frequently used in the exploratory stage of an analysis as they measure linear association. A number of popular choices are available include Pearson's (1896) product-moment correlation coefficient, Spearman's (1904) rank correlation coefficient (ρ) and Kendall's (1938) τ, another rank correlation coefficient. The three measures are related as will be discussed later.

As exploratory statistics correlations should perhaps not be examined in isolation, but with a visualisation in a scatterplot or scatterplot matrix. Anscombe (1973) demonstrates that very different configurations in a scatterplot can have identical correlation coefficients and regression lines. Once the number of variables involved increases, so the number of possible pairings, ($(n^2 - n)/2$), increases rapidly and the analyst may be encouraged towards techniques for dealing with multivariate correlation structures such as principal component analysis (PCA).

A formula for the Pearson (1896) product moment coefficient between two variables x and y is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \qquad \text{Equation 1}$$

which is the ratio of the covariance to the product of the standard deviations of the two variables in question.

A frequently quoted formula for Spearman's ρ is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n^3 - n} \qquad \text{Equation 2}$$

*

where d_i is the difference between the the rank positions on each variable. This may be derived by noting that the definition of ρ is equivalent to r where the values of the variables replaced by their rank positions. Spearman (1904, p87) in discussing the ``method of rank differences'' also presents the alternative:

$$\rho = 1 - \frac{3\sum |d_i|}{n^2 - 1} \qquad \text{Equation 3}$$

which he suggests differs from r by a factor of $\sqrt{r^3}$.

Kendall's (1938) τ_b examines all possible pairings of the observations and sums the concordant and discordant pairings with a correction for tied observations. Kendall's coefficient can be computed from:

$$\tau = \frac{P - Q}{\sqrt{(P + Q + x_t - xy_t)(P + Q + y_t - xy_t)}} \qquad \text{Equation 4}$$

where P is the count of concordant pairs, Q the count of discordant pairs, x_t the count of observations where the xs are tied, similarly for y_t, and xy_t the count of pairs for which both variables are tied.

2. Decomposing the coefficients

The numerator in each case is a summation of measures of the covariation of the variables in question. For r, ρ and τ the individual elements can be considered separately, and when divided by the numerator they become the components of a decomposition of each coefficient. They can then be used to show:

- the contribution of each observation to the global statistic
- the influence of each observation on the global statistic

The latter allows us to determine whether any of the individual locations can be considered as outlying in some sense.

Anselin (1995) suggests that a LISA satisfies the requirements of:

1. yielding an indication of the extent of significant spatial clustering of similar values around each observation
2. the sum of the individual values is proportional to the global indicator of spatial association.

Anselin demonstrates local decompositions of Moran's I and Geary's C, as well as Getis and Ord's (1992) G and G^* statistics. The first of Anselin's requirements implies the existence of a inferential framework. We see the local versions of r, ρ and τ as very much part of an exploratory framework, although the existence of outlying values points to the locations of unusual data in either each or both variables. Unlike I, C, or G the correlation decompositions provide a measure of spatial association between a *pair* of variables and not just a single variable. The decompositions also show us *where* the covariation is strongest.

Hawkins (1980) characterises an outlier as *an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*. In this respect observations in which the individual product of the deviances is outlying in the boxplot would be worth scrutiny as either potential *hot spots* or *cold spots*. These would be hot or cold spots of *association*. A permutation test could be used to determine whether the values were significant. We do not need a weight matrix with

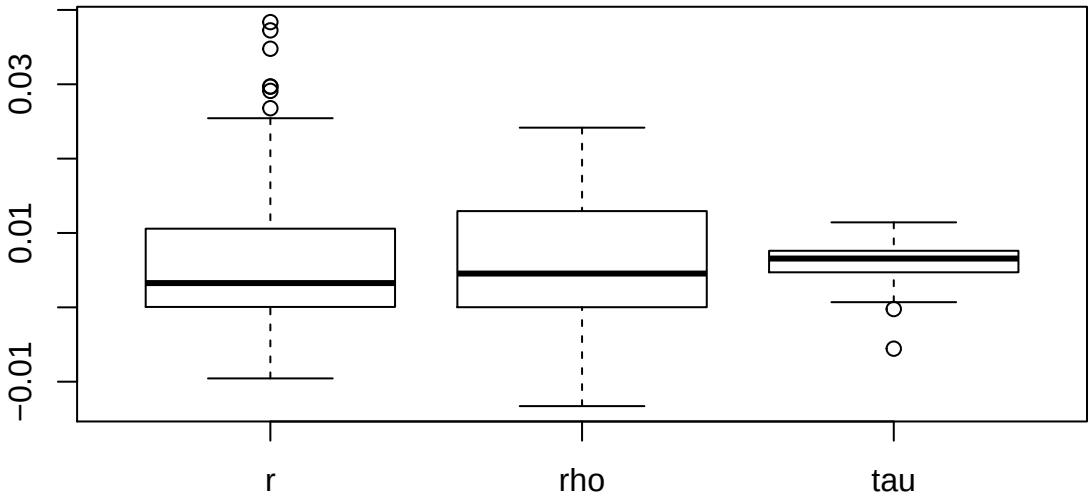

Figure 1: Boxplots of Distributions of Local Correlation Decompositions

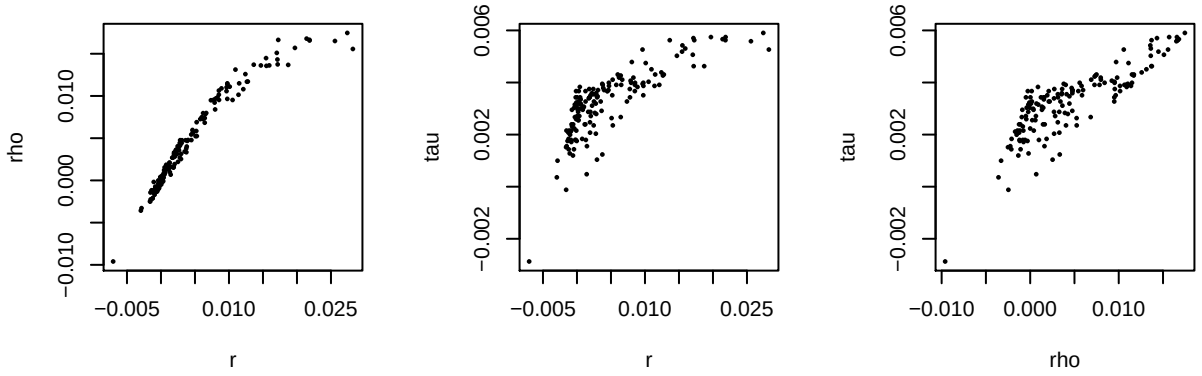

Figure 2: Pairwise Scatterplots of Local Components of Correlation ($r \leftrightarrow \rho$, $r \leftrightarrow \tau$ and $\rho \leftrightarrow \tau$)

the decomposed correlation coefficient. Wall (2004) has also observed that weights matrices can induce undesirable dependencies into CAR and SAR models.

3. Example - Educational Attainment in Georgia

We illustrate the local decompositions of the three coefficients by considering the local relationships between the proportion of residents educated to bachelor's level or higher to the proportion of residents born outside the USA.

It is instructive to examine the relationship between the individual components of the three measures.

The boxplot (Figure 1) shows the distributions of the components when divided by their corresponding global values. The Pearson and Spearman components have similar distributions, with some larger outlying values among the Pearson components. The median of the Kendall values is slightly larger, and the variability of the distribution is lower than the other two. The Kendall values have two low-valued outliers. We can compare the distributions in scatterplots (Figure 2).

The r and ρ distributions are reasonably closely related, which is hardly surprising given their derivation.

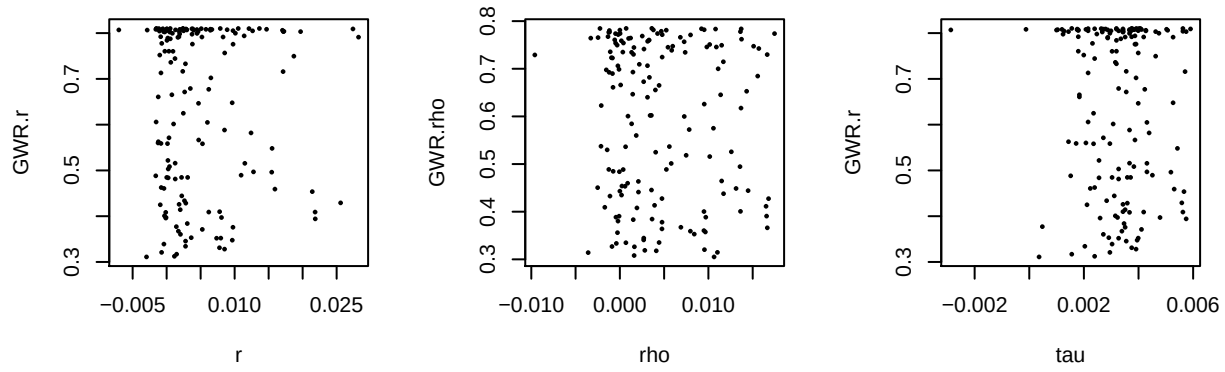

Figure 3: Local Components of Correlation vs. Geographically Weighted Equivalents

There is more variation in the τ values. To what extent does the Geographically Weighted Correlation coefficient mirror any of these coefficients? This is investigated in Figure 3.

The results are illuminating. The GW coefficients are on the y-axis of each plot. They represent the *locally weighted* correlation between the two variables, and include the neighbouring observations under each kernel. In the case of the plots shown here, the kernel size was determined to obtain the best fit from a GW regression model - there are 109 neighbours. Decreasing the size of kernel decreases the number of observations in both the numerator and denominator, whereas with the decomposed coefficient, *all* the observations contribute to the denominator. Not surprisingly, the plots suggest that the GW and decomposed coefficients measure different phenomena.

4. Generalised Correlation Coefficients

Kendall (1948) points to Daniels (1944) as the originator of a generalised correlation coefficient Γ of which the product moment correlation coefficient, Spearman's (1904) rank correlation coefficient and Kendall's (1928) τ are special cases.

The generalised coefficient is given by:

$$\Gamma = \frac{\sum a_{ij} b_{ij}}{\sqrt{\sum a_{ij}^2 \sum b_{ij}^2}} \qquad \text{Equation 5}$$

The *ij* subscripts indicate that the summation is over all values of *i* and *j* from $1...N$.

Statistic	a_{ij}	b_{ij}
Pearson's r	$x_j - x_i$	$y_j - y_i$
Spearman's ρ	$j - i$	$j - i$
Kendall's τ	$\pm 1\ j \gtrless i$	$\pm 1\ j \gtrless i$

Table 1: Generalised Correlation Forms for r, ρ and τ

This gives great flexibility in arriving at a set of complementary measures of correlation. The numerator is a measure of covariation, and the denominator ensures that the resulting dimensionless statistic scales between -1 and +1. There may well be other measures of association which fit into this framework.

5. Do we have another LISA?

The decompositions are candidates for consideration as a LISA in that their values add up to the value of the parent coefficient. Anselin's chosen measures provide a local version of a global coefficient, but for a single variable at a time. The decomposed coefficients also provide a local version of a global coefficient, which shows *where* a pair of variables are most strongly associated. In this sense, they are a LISA.

While the Pearson and Spearman coefficients are widely encountered in the social sciences as measures of correlation the relative infrequency of Kendall's coefficient is puzzling. Moran (1948) comments on the `superiority of $\tau$ as a measure of rank correlation', and a few months later Daniels (1948) remarks that it `is seen to measure in a rather arbitrary sense the degree of agreement between ranks'. Daniel also notes Moran's observation that τ is related to the `least number of interchanges required to bring the two rankings into perfect agreement'. Perhaps greater use of τ might be made in preference to ρ as a non-parametric measure of correlation.

6. References

Anscombe, FJ. 1973. Graphs in statistical analysis. *The American Statistician*. 27(1). 17-21

Anselin, L. 1995. Local indicators of spatial association-LISA. *Geographical Analysis*. 27(2). 93-115

Chatfield, C. 1995. *Problem Solving: a statistician's guide*. 2nd edition. London: Chapman & Hall

Daniels, HE. 1944. The relation between measures of correlation in the universe of sample correlations. *Biometrika*. 33(2). 129-135

Daniels, HE. 1948. A property of rank correlations. *Biometrika*. 34(3/4). 416-417

Hawkins, D. 1980. *Indentification of Outliers*. London: Chapman and Hall

Kendall, MG. 1938. A new measure of rank correlation. *Biometrika*. 30(1/2). 81-93

Kendall, MG. 1948. *Rank Correlation Methods*. London: Griffin (p.24)

Moran, PAP. 1948. Rank correlation and product-moment correlation. *Biometrika*. 35(1/2). 203-206

Pearson, K. 1896. Mathematical contributions to the theory of evolution. III. Regression. heredity and panmixia. *Philosophical Transactions of the Royal Society of London*. 187. 253-318

Pearson, K. 1920. Notes on the history of correlation. *Biometrika*. 13(1). 25-45

Spearman, C. 1904. The proof and measurement of association between things. *The American Journal of Psychology*. 15(1). 72-101

Wall, MM. 2004. A close look at the spatial structure implied by CAR and SAR models. *Journal of Statistical Planning and Inferemnce*. 121. 311-324

7. Acknowledgements

We gratefully acknowledge funding from Science Foundation Ireland under the Investigators' Award Programme award number 15/IA/3090.