Signed chi-squares revisited

Martin Charlton[*1], Chris Brunsdon[†1], Paul Harris[‡2] and Lex Comber[§3]

[1]National Centre for Geocomputation, Maynooth University, Maynooth, Co Kildare, Ireland
[2]Rothamsted Research, North Wyke, Okehampton, Devon EX20 2SB
[3]School of Geography, University of Leeds, Leeds LS2 9JT

December 3, 2018

Summary

In considering maps for contingency tables, such as those which represent an age-sex distribution, a question arises as to a means to displaying their spatial variation and identifying areas for closer scrutiny. Visvalingam's (1976) signed χ^2 maps form a starting point which leads us to examine the utility of local Cramer's V statistics, and various modifications thereof. We show their use in mapping the spatial variation in the age-sex distribution in Ireland.

Deleted: f

Deleted: y

Deleted: ing

KEYWORDS: contingency table, χ^2, Cramer's V, spatial variation

1. Introduction

The appearance of a census of population is often followed by an intensive period of activity in mapping the spatial variations of numerous indicator variables to visualise the patterns of socio-economic activity.

Such variables are usually normalised by some suitable denominator (for example: percent of total resident population aged between 15 and 19). This pre-supposes some underlying statistical model (usually the binomial), but this tends to be overlooked. For smaller spatial units problems with the stability of the estimates can arise, which raises the question whether there are other ways of examining such data.

Choynowski (1959) appears to the one of the first to use Poisson probabilities in constructing a map of the spatial variation of brain tumour incidence among the poviats in Rzeszow, Poland. The observed number of tumors in each poviat was small (apart from Rszeszow city, none had more than 6), and thus the incidence rates were small. This technique is discussed by Cliff and Haggett in their atlas of disease distributions (Cliff and Haggett, 1988, 99-101).

Deleted: was examined

Another choice for disease incidence mapping, the SMR, presents problems of stability when the populations of the spatial units are small. To deal with this Clayton and Kaldor (1987) describe a method based on an empirical Bayes estimates of the relative risks.

Work at the Census Research Unit at Durham University in the 1970s led to the development of the signed χ^2 statistic (Visvalingam, 1976; 1978; 1983). This statistic was used in the production of an atlas of indicators from the 1971 Census of Population for the county of Durham (Dewdney and Rhind, 1975) and a national Census Atlas (CRU/OPCS, 1980); the production of the latter is outlined in Rhind et al, 1980). The expected values for Visvalingam's statistic were obtained from national proportions – the values being mapped are the signed components of a one sample χ^2 statistic. Jones

[*] martin.charlton@mu.ie
[†] christopher.brunsdon@mu.ie
[‡] paul.harris@rothamsted.ac.uk
[§] a.comber@leeds.ac.uk

and Kirby (1980) caution that this measure should not be regarded as a statistical test, but rather as a data transformation procedure that allows for small numbers.

Deleted: e

This naturally leads to the question of what happens with the case of a two sample contingency table. Comber at al (2017) considered the spatial variation of geographically weighted correspondence matrices in to visualise the accuracy of remotely sensed image classifications. The matrices are square, and a single statistic is reported for each matrix (for example: the portmanteau accuracy).

More challenging is the case of spatially varying contingency tables, such as those that arise from considering variables with more than two categories, for example: age/sex distributions for a geographically heterogenous population, where the statistic may also vary over space. This forms the basis for this paper.

Deleted:

2. Two sample contingency tables

In Visvalingham’s (1976) signed χ^2 the expected frequencies for the cells in a binary case (males/females, unemployed/employed) were generated from the national probabilities. A two sample χ^2 test for independence tests the null hypothesis that the observed frequencies in the table do not differ from those expected under independence.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where E_{ij} is the joint expectation of being a member of row i and column j under the hypothesis of independence. This is:

$$E_{ij} = p_i p_j \sum_{**} O_{ij} = \frac{\sum_{i*} O_{ij} \sum_{*j} O_{ij}}{\sum_{**} Oij}$$

and a related statistic, Cramer's V, is

$$V = \sqrt{\frac{\chi^2}{N\min(r-1,c-1)}}$$

V (Cramer, 1946) can be thought of as a χ^2 which is normalised by the sample size. It has an interpretation as the ratio of the obtained departure from independence to the maximum departure from independence, and runs from 0 to 1 (Acock and Stavig, 1979). The N*min(r-1,c-1) term provides the maximum value of the chi-square under the hypothesis of independence for a rectangular table, which constrains V to the range 0-1.

Note that V does not indicate how the departure from independence occurs, that is up to the inspection and modelling of the requisite contingency table. What is does provide is a mappable statistic that is comparable between zones; this can suggest where anomalous cases may lie. The sampling distribution of a spatial version of this statistic is for future research; however, we may use the boxplot (and a corresponding boxmap) as a means of identifying zones for closer scrutiny.

3. Independence and comparison

A question arises as to the underlying hypothesis. The χ^2 measures the divergence of the observed values from those that would arise if the variables in the rows and columns were independent. We might wish for a different hypothesis: perhaps that the local pattern differs from that which we would observe if it followed the national proportions: in this case the E_{ij} would be generated from the joint marginal probabilities for the national table. The resulting statistic no longer follows a χ^2 distribution, and the maximum value compensation introduced by Cramer is not the maximum value of the statistic we have created,

A simple workaround is to divide the local V by the national V. The local V measures the extent to which the pattern differs from that which would be given by the maximum value possible. The ratio of local to global measures the extent that the local divergence differs from the national divergence, with 1 being identical.

If the local sample population was confined to the cell where the joint expectation (based on the national probabilities) was lowest, we obtain the maximum value of the χ^2 for that population. This means we can create a statistic, V*, which runs from zero to unity for baseline models other than that of independence.

Deleted: hypotheses

Percentage points for V* distribution might be found by simulation. At the moment we follow Jones and Kirby’s (1980) advice that such statistics should be regarded as providing exploratory information.

4. Experiment: age-sex variation in Ireland

We consider the age sex distribution among the 3409 Electoral Divisions (EDs) in the Republic of Ireland, using data from the 2011 Census of Population. Five broad age groups are considered: 0-14, 15-24, 25-44, 45-64, and 65 and over. The resulting contingency tables, which are easily formed for each ED, have 4 degrees of freedom. The median ED population is 620, so care must be taken to ensure that the incidence of expected values under 5 is kept to a minimum.

The distributions of the V* values between the counties are shown in Figure 1.

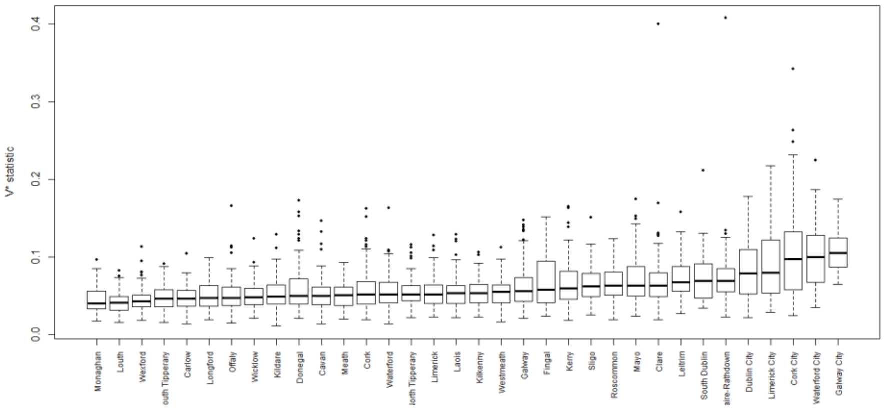

Figure 1 Boxplots of the V* statistic by county – ordered by median V*

Of note is that the counties with the highest median values are those for the cities and main urban areas around Dublin. These have areas with notable age imbalances, particular towards the younger population. The lowest imbalances are to be found in the more rural counties. However, the county breakdown is not always helpful.

The spatial variation at ED level is shown in Figure 2 below.

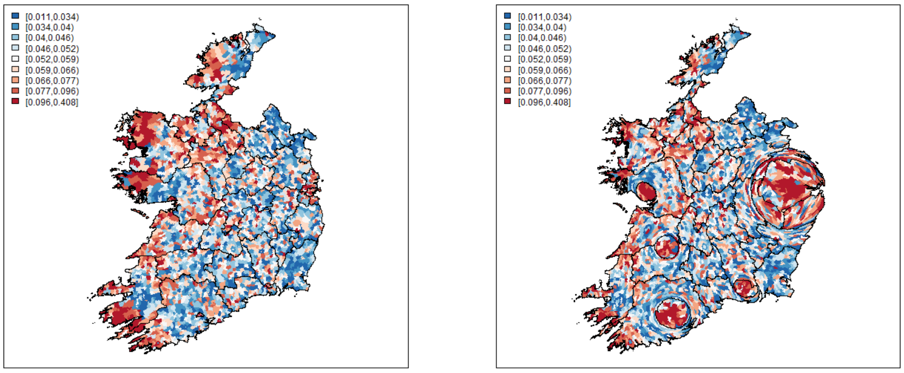

Figure 2 V* by ED – county boundaries shown as a guide

The two map bases are (i) in the Irish National Grid system and (ii) a mild cartogram transform based on the log of the population. The conventional projection indicates imbalances in the western parts of Cork, Kerry, Galway, Mayo and Donegal and hints at imbalances in the main urban area. The cartogram reveals imbalances in Dublin, Galway City, Limerick City, Cork City, and Waterford City.

Deleted: a

What these show is *where* the imbalances occur but not necessarily their direction. As such the V* statistic is a useful tool for exploring spatial variation in imbalance, and leads to questions such as 'why'.

The Moran I for this V* pattern is 0.37 which is significant at the 0.05 level – there is mild positive autocorrelation. There is evidence of spatial dependency effects.

A complementary approach is to use the bias correction described in Bergsma (2013) which is based on the phi coefficient. This controls for varying population by converting the values in the contingency table to probabilities. Mapping the resulting spatial variation yields Figure 3:

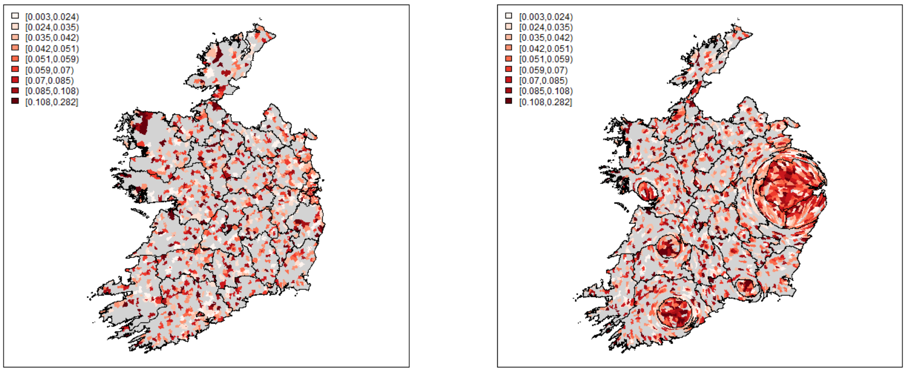

Figure 3 Bergsma's bias corrected V

The areas where there is no divergence from the maximum value are shown light grey. Once again, the imbalance in the urban areas is notable. Note, however, that the hypothesis being tested is one of independence rather than divergence from the national age-sex pattern. The correlation between the non-zero values of this bias corrected V and the corresponding V* is 0.42.

5. Conclusions

This is very much work in progress. A geographically weighted version of the statistic is suggested by the Moran I.

6. Acknowledgements

We gratefully acknowledge support from Science Foundation Ireland's Strategic Research Initiative (07/SRC/I1168) which allowed some of the development of geographically weighted discriminant methods. We also acknowledge support from Science Foundation Ireland's Investigator Award Programme (15/IA/3090).

7. Biography

Martin Charlton is Senior Lecturer in Geocomputation at Maynooth University.

Chris Brunsdon is Professor of Geocomputation at Maynooth University.

Paul Harris is Project Leader in Sustainable Agriculture Sciences at Rothamsted Research's North Wyke Farm Platform in Devon. He used to work at Maynooth University.

Lex Comber is Professor of Geography at the University of Leeds. He wrote the abstract of a paper for GIScience 2016 on a flight from Rhodes to London, aided by Martin and Chris.

References

Acock AC and Stavig GR, 1979, A measure of association for nonparametric statistics, *Social Forces*, 57(4), 1381-1386

Bergsma, W, 2013, A bias-correction for Cramer's V and Tschuprow's T, *Journal of the Korean Statistical Society*, 42, 323-328

Choynowski M, 1959, Maps based on probabilities, *Journal of the American Statistical Association*, 54, 385-388

Clayton D and Kaldor J, 1987, Empirical Bayes estimates of age-standardised relative risks for use in disease mapping, *Biometrics*, 43(3), 671-681

Comber A, Brunsdon CF, Charlton M, and Harris P, 2017, Geographically weighted correspondence matrices for local change analyses and error reporting: Mapping the spatial distribution of errors and change. *Remote Sensing Letters,* 8: 234–243

Cramer, H, 1946, *Mathematical Methods of Statistics*, Princeton: Princeton University Press

Dewdney JC and Rhind DW, 1975, *People in Durham: a Census Atlas*, Census Research Unit, University of Durham

Jones K and Kirby A, 1980, The use of chi-square maps in the analysis of census data, *Geoforum*, 11, 409-417

Rhind DW, Evans Is, and Visvalingam M, 1980, Making a national atlas of population by computer, *The Cartographic Journal*, 17(1), 3-10

Visvalingam M, 1976, *Chi-square as an alternative to ratios for statistical mapping and analysis*, Census Research Unit Working Paper No.8, University of Durham

Visvalingam M, 1978, The signed chi-square measure for mapping, *The Cartographic Journal*, 15(2), 93-98

Visvalingam M, 1983, Area-based social indicators: signed chi-square as an alternative to ratios, *Social Indicators Research*, 13(3), 311-329