

Discovering Spatially Multiway Collocations

Didier Leibovici¹, Lucy Bastin² and Mike Jackson¹

¹Centre for Geospatial Sciences, University of Nottingham, UK

²School of Engineering & Applied Science, Aston University, UK

Tel. (+44(0)115 846 8408) Fax (+44(0)115 951 5249)

Didier.leibovici@nottingham.ac.uk

KEYWORDS: spatial statistics, co-occurrences, multiway data, marked point process

1. Introduction

Analysing geographical patterns by collocating objects, events or attributes, has a long history in applied science such as ecology or epidemiology. In spatial statistics, this is often associated with marked point processes, (Diggle, 2003). The problem of identifying patterns of co-occurrences, or at least establishing the existence of certain structures at some scales, is usually addressed by plotting the spatial dependence functions (cross-K functions) against distance, and by testing them against complete randomness or other generated stationary processes, (Ripley, 1977); Schlater and Diggle, 2004). This paper will investigate optimum spatial representations of collocations in terms of lack of spatial independence between two categorical variables or more. This can be complementary to the approach described above.

The framework, using a tensorial data representation developed in Leibovici and Sabatier (1998); Leibovici (2007) allows more detailed collocation by analysing co-occurrences of higher order than pairs.

2. Measuring Spatial cocurrences,

Let $i=1, \dots, I$ be categories of a variable νI , $j=1, \dots, J$ be categories of a variable νJ , $k=1, \dots, K$ be categories of a variable νK , and let $s=1, \dots, S$ be locations where one can record either νI , νJ or νK but also in some case studies all of them and sometimes more than one record, depending on the geometrical object and the semantic associated with the locations. The variables νI , νJ , νK describe a general “event”: e.g. **(i)** a person of age i with social class level j , diagnosed for a certain disease k and living at location s ; **(ii)** a plant species i , on a soil class j , at location s with annual rainfall k , or **(iii)** a crime of type i , at time slot j , in a zone of wealth class k of location s . So we are looking for associations of νI , νJ , νK variables in their spatial co-occurrences, either as multivariate observations on a spatial domain or as already collocated observations of different variables, or both. Clearly **(i)** is multivariate on the persons, **(ii)** is a collocation of different measurements **(iii)** is a mixture of both. One can argue that **(ii)** could be seen as multivariate characteristics of the location.

2.1. Counting pairs, triples

With, $n_{ii'}$, n_{ij} , and n_{ijk} the number of collocations of events or marks i and i' , or i and j , or i , j and k when the distance of collocating events is implicit, and n_{ij}^d when the distance d of collocating events is explicit, a general definition of collocations is, **Cg**: a collocation of marks or labels $\{i, j, k\}$ is recorded if, the distance between the locations $\{s, s', s''\}$ (which may be equal), all together expressing the labels $\{i, j, k\}$, is at most d . The standard way of computing the collocations (here without edge corrections) for two labels is:

$$n_{ij}^d = \sum_s \#_{C(s_i, d)}\{j\} = \sum_s \#_{C(s_j, d)}\{i\} = n_{ji}^d \quad (1)$$

where $\#C(s_i, d)\{j\}$ means the number of “events” or marked locations j at maximum distance d from a location labelled i : the number of s_j found in a circle of radius d and centre s_i . Adding flexibility about the searched geometry area ($G(s_i, d)$), with d being the buffer size of the geometry s_i and also about the way (O) the occurrence is recorded “within” the geometry (as depending also on the geometry of the j mark), gives a more general formula:

$$n_{ij}^d = \sum_s \mathcal{O}G_{(s_i, d)}(\{s_j\}) \quad (2)$$

Depending on the choices for O and $G(., d)$ the collocation value may not be symmetrical. When addressing 2nd-order collocations in multitype point patterns, Lotwick and Silverman (1982), generalise cross-Ripley’s K function and work with the labels two by two, and Baddeley and Turner (2005), use subsets of categories *e.g.* one from νI and one from νJ . These methods are used to test for clustering under assumptions for the spatial process such as stationarity and isotropy, which can be alleviated for particular studies (Diggle et al., 2007). Using (dis)similarities of order three, Heiser and Bannani Dosse (1997) demonstrated improved grouping or sorting descriptions and pattern recognition.

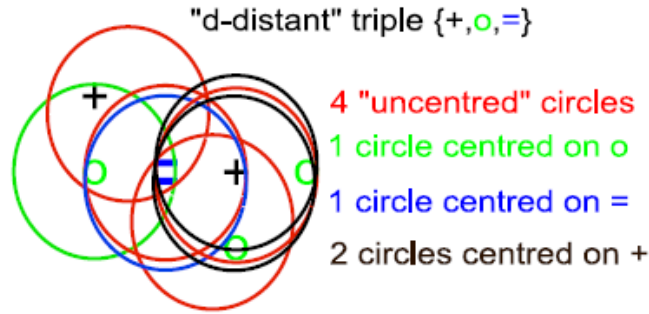


Figure 1. Symmetrical and asymmetrical ways of counting collocations of a triple with circles of radius d (not necessarily verifying definition Cg)

For three or more marked events the standard formula has to be rewritten if symmetry is to be maintained:

$$\begin{aligned} n_{ijk}^d &= \sum_s \#C(d)\{i, j, k\} & (3) \\ &\neq \sum_s \#C(s_i, d)\{j, k\} \neq \sum_s \#C(s_j, d)\{i, k\} \neq \sum_s \#C(s_k, d)\{i, j\} \\ n_{ijk}^d &=_{Cg} \sum_s \#C(s_i, d) \cap C(s_j, d) \cap C(s_k, d)\{i, j, k\} \end{aligned}$$

As seen in Figure 1, the non-location-centred circle method is symmetrical but collocation counts from asymmetrical methods based on marked-location-centred circles are different. In fact a triple, collocation will be recorded if the intersection of the three circles contains the three labels. Different counting methods and interpretations can introduce a range of methods suitable for discovering spatial collocations. Considering S itself as a variable instead of marginalising through it, will allow spatial interaction of associations between categorical variables.

These counts may be very low, depending on the study as well as the approach taken: multivariate observations or collocated observations. Can a collocation approach (case (ii)) be informative for multivariate observations? Sample sizes needed for a multivariate approach (case (i)) are larger! The approach of this paper being multitype, multi-variable, without assumptions on the process distribution, the focus is on a very simple and well known

statistic: the χ^2 measure for lack of independence, with the aim of demonstrating its potential use with appropriate tables of counts for multiway collocation.

3. Co-occurrences and tensorial framework

The tensorial framework developed previously is used here to analyse and represent data “summaries” in a similar approach to multidimensional analysis. This framework allows some flexibility about representing or modelling data information within an array of any dimensions, affording new ways to discover spatial pattern. From Figure 2 some two way tables can be analysed by correspondence analysis, (method **CA200**). The extension of correspondence analysis to a k-way table provided in Leibovici (2007, 2000) can be used to analyse triple (or more) collocations (method **CAk00**). This allows us to analyse multiway dissimilarities as in Benanni-Dosse’s thesis, (published partially in Heiser and Bennani Dosse, 1997), with generalised unfolding metric multidimensional scaling.

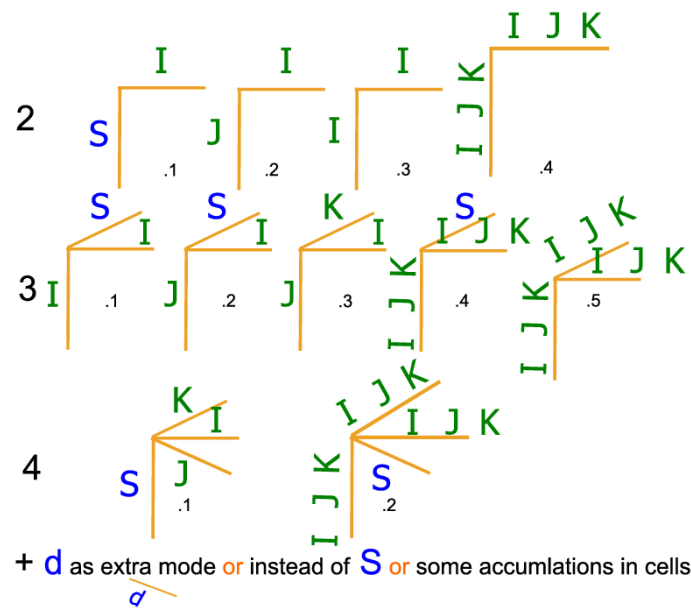


Figure 2. k-way Tables for collocation discovery analysis: all 2-modes, 3-modes, 4-modes possibilities with cell values as counted collocations as explained in the text: *I*, *J*, *K* categorical variables, *S* spatial locations, *d* distance for collocation event.

3.1. Two-way Correspondence Analysis

Correspondence analysis of a two-way contingency table (**FCA**) $p_{ij} = n_{ij}/N$ is achieved by performing the Principal Component Analysis, **PCA**, (or generalised **PCA**) of the following triplet, (Escoufier, 1987); Dray and Dufour, 2007):

$$(D_I^{-1}PD_J^{-1}, D_I, D_J) \quad (4)$$

with D_I and D_J diagonal metrics containing vector margins $p_{.i}$ and $p_{.j}$; where a **PCA** of a triplet

(X, Q, D) with X a data matrix $n \times p$, Q a $p \times p$ metric on the rows (or in the column-space) and similarly D a $n \times n$ metric on the columns (or in the row-space), is generalising a standard **PCA** by diagonalising with Q -normed vectors the matrix $'XDXQ$ equivalent, to the covariance matrix if X is column-centred, $D = 1/nIdn$ and $Q = Idp$, or to the correlation matrix if instead of the identity metric $Q = \text{diag}(1/\text{var}_1 \cdots 1/\text{var}_p)$.

The measure of lack of independence is linked to the analysis by:

$$1 + \frac{\chi^2}{N} = \sum_{ij} \frac{(p_{ij} - p_{i..}p_{.j.})^2}{p_{i..}p_{.j.}} = \sum_{ij} p_{i..}p_{.j.} \left(\frac{p_{ij}}{p_{i..}p_{.j.}} \right)^2 = \sum_s \sigma_s^2 \quad (5)$$

where the σ_s^2 are the eigenvalues. $\sigma_0 = I$ with components equal to unit vectors in their respective spaces.

3.2. Correspondence Analysis of k-way coOccurrences

For a **CAkOO** analysis we propose to use a generalisation of correspondence analysis to k-way tables, **FCA-kmodes**, which is a specific **PTA-kmodes** just as **FCA** is a specific **PCA**. The **PTA-kmodes** decomposition of a k-way table was already applied in a spatial context in Leibovici et al. (2007). With similar notations for a three-way table $I \times J \times K$, one performs the **PTA-3modes** of the quadruplet:

$$((D_I^{-1} \otimes D_J^{-1} \otimes D_K^{-1})P, D_I, D_J, D_K) \quad (6)$$

Equation 7 shows that three-way independence can be orthogonally decomposed into deviations from independence for the two-way margins of the three-way table, and a three-way interaction term. Each two-way margin's deviation from independence is reminiscent of (simple) correspondence analysis:

$$\begin{aligned} \frac{\chi^2}{N} &= \sum_{ijk} p_{i..}p_{.j.}p_{..k} \left(\frac{p_{ijk} - p_{i..}p_{.j.}p_{..k}}{p_{i..}p_{.j.}p_{..k}} \right)^2 \\ &= \sum_{jk} p_{.j.}p_{..k} \left(\frac{p_{.jk} - p_{.j.}p_{..k}}{p_{.j.}p_{..k}} \right)^2 + \sum_{ik} p_{i..}p_{..k} \left(\frac{p_{i.k} - p_{i..}p_{..k}}{p_{i..}p_{..k}} \right)^2 + \sum_{ij} p_{i..}p_{.j.} \left(\frac{p_{ij.} - p_{i..}p_{.j.}}{p_{i..}p_{.j.}} \right)^2 \\ &+ \sum_{ijk} p_{i..}p_{.j.}p_{..k} \left(\frac{p_{ijk} - \delta_{ijk}}{p_{i..}p_{.j.}p_{..k}} \right)^2 \end{aligned} \quad (7)$$

2-way margins analysis from decomposition of 3-way co-occurrences are not equivalent to 2-way co-occurrence analysis as in correspondence analysis of contingency tables. Nonetheless, considering co-occurrence counts and contingency tables one can symbolically write for a multiway table like 3.3 or 4.1, on Figure 2:

$$\lim_{d \rightarrow 0} CAkOO = FCAk \quad (8)$$

4. Examples

Different examples for CA2OO and CAkOO methods will be shown for the presentation using datasets from the literature but also with datasets from recent epidemiological studies:

- *langsing* from **spatstat**
- an epidemiological dataset of persons (5 age groups) located to home postcode, having contracted a bacterium resistant (R) or sensitive (S) to an antibiotic at one of three successive time periods,
- a similar epidemiological dataset where the resistance to antibiotics has been classified into 3 or 20 categories and where patient gender is known.

On each dataset, we compared a range of possible analyses from Figure 2, and will demonstrate the added value of the approach for classical 2nd order analysis.

References

Baddeley, A. and Turner, R. (2005). spatstat: An r package for analyzing spatial point patterns. *Journal of Statistical Software*, **12(6)**, 1–42.

- Diggle, P., Gomez-Rubio, V. Brown, P., Chetwynd, A., and Gooding, S. (2007). Second order analysis of inhomogeneous spatial point processes using case-control data. *Biometrics*, **63**, 550–557.
- Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Hodder Arnold, London.
- Dray, S. and Dufour, A.-B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22(4)**, 1–20
- Escoufier, Y. (1987). The duality diagram : a means of better practical applications. In P. Legendre and L. Legendre, editors, *Development in numerical ecology*, pages 139–156. NATO advanced Institute, Springer Verlag, Berlin.
- Heiser, W. and Bennani Dosse, M. (1997). Triadic distance models : axiomatization and least squares representation. *Journal of Mathematical Psychology*, **41**, 189–206.
- Leibovici, D. (2000). Multiway Multidimensional Analysis for Pharmacology-EEG Studies. Report initiated at SANOFI-RECHERCHE , TR00DL2, FMRIB Centre, University of Oxford, UK.
- Leibovici, D. (2007). PTak: Principal Tensor Analysis on k modes. *Contributing R-package*, version 1.1-16.
- Leibovici, D. and Sabatier, R. (1998). A Singular Value Decomposition of k-Way Array for a Principal Component Analysis of Multiway Data, PTA-k. *Linear Algebra and Its Applications*, **269**, 307–329.
- Leibovici, D., Quillevere, G., and Desconnets, J.-C. (2007). A Method to Classify Ecoclimatic Arid and Semi-Arid Zones in Circum-Saharan Africa Using Monthly Dynamics of Multiple Indicators. *IEEE Transactions on Geoscience and Remote Sensing*, **45(12)**, 4000–4007.
- Lotwick, H. and Silverman, B. W. (1982). Methods for analysing spatial processes of several types of points. *Journal of Royal Statistical Society*, **B(44)**, 406–413.
- Ripley, B. (1977). Modelling spatial patterns. *Journal of Royal Statistical Society*, **B(39)**, 172–212.
- Schlater, M. Riberio, P. and Diggle, P. (2004). Detecting dependence between marks and locations of marked point process. *Journal of Royal Statistical Society*, **B(66)**, 79–93.

Biography

Dr Didier Leibovici is a Research Fellow in geospatial modelling and analysis, with previous posts as statistician in epidemiological/medical imaging research and as geomatician for landscape changes in agro- ecology. Interests refer to interoperability and conflation models for cross-scales for integrated modelling applications within an interoperable framework chaining web services.

Lucy Bastin is a Lecturer in GIS at Aston University. After a PhD on urban plant metapopulations, and research into fuzzy classification / uncertainty visualisation at Leicester University, she spent 3 years as a GIS software developer. Her current research interests include Web Processing Services for automatic interpolation, and spatial epidemiology.

Pr. Mike Jackson is head of department. He worked for the geospatial industry (QinetiQ, Hutchison 3G, Laser_Scan) and in research for NERC and as investigator for NASA. Mike is non-executive director of the Open Geospatial Consortium Europe, with interests in combining new technologies such as location based services with geospatial intelligence.