SpringerPlus
a SpringerOpen Journal

METHODOLOGY Open Access

The association in a two-way contingency table through log odds ratio analysis: the case of Sarno river pollution

Ida Camminatiello[1*], Antonello D'Ambra[1] and Pasquale Sarnacchiaro[2]

Abstract

In this paper we are proposing a general framework for the analysis of the complete set of log Odds Ratios (ORs) generated by a two-way contingency table. Starting from the RC (M) association model and hypothesizing a Poisson distribution for the counts of the two-way contingency table we are obtaining the weighted Log Ratio Analysis that we are extending to the study of log ORs. Particularly we are obtaining an indirect representation of the log ORs and some synthesis measures. Then for studying the matrix of log ORs we are performing a generalized Singular Value Decomposition that allows us to obtain a direct representation of log ORs. We also expect to get summary measures of association too. We have considered the matrix of complete set of ORs, because, it is linked to the two-way contingency table in terms of variance and it allows us to represent all the ORs on a factorial plan. Finally, a two-way contingency table, which crosses pollution of the Sarno river and sampling points, is to be analyzed to illustrate the proposed framework.

Keywords: Log ratio analysis; Weighted log odds ratio analysis; Association measures; Two-way contingency table; Water pollution

Introduction

Polycyclic aromatic hydrocarbons (PAHs) are a group of lipophilic contaminants widespread in the environment. This class of compounds has been widely studied (Tolosa et al., 1995; Caricchia et al., 1993) because of its carcinogenic and mutagenic properties (Lehr and Jerima, 1997; Yan, 1985; White 1986).

PAHs are produced by both anthropogenic and natural processes and can be introduced into the environment through various routes. Anthropogenic inputs can originate from the incomplete combustion of organic matter (pyrolytic) and the discharge of crude oil-related material (petrogenic). PAHs can also originate from natural processes such as short-term diagenetic degradation of biogenic precursors (diagenesis). Each source (i.e. pyrolytic, petrogenic and diagenetic) gives rise to characteristic PAH patterns. Currently, the interest in multivariate statistical methodologies for identifying the main sources of PAH pollution and for quantifying the incidence of each source of pollution on total pollution levels, particularly in coastal environments, is increasing (Luo et al., 2006; Bihari et al., 2007; Xu et al., 2007; Sarnacchiaro et al., 2012). This study is part of a large project which has the objective of enhancing the knowledge of pollution in the Sarno River and its environmental impact on the gulf of Naples. This project has attempted to assess the pollution derived from local industries, agriculture and urban impact (Sarnacchiaro et al., 2012). In the present work, we have studied the association between the level of PAH pollution and the sampling points.

The analysis of the association for variables placed in a $I \times J$ two-way contingency table is a topic widely discussed. In this paper we focus our attention on the Odds Ratios (ORs) as measure of association. In a two-way contingency table the total number of ORs, that can be computed, may be too large, for their synthesis four main alternatives or complementary strategies have been performed. The first consists in the computation of statistical measures (Altham, 1970). The second is based on the construction of the model for frequencies and studies the ORs through the

* Correspondence: camminat@unina.it
[1]Department of Economics, Second University of Naples, Corso Gran Priorato di Malta, 81043 Capua, CE, Italy
Full list of author information is available at the end of the article

interaction between the row and column variables. The log-linear model for two-way contingency table belongs to this class. The third solution is the RC (M) association model (Goodman, 1985), which is more parsimonious than the usual log-linear model (Choulakian, 1988). The fourth strategy takes in consideration Singular Value Decomposition (SVD) of the matrix containing the basic set of log ORs (de Rooij and Anderson 2007).

In this paper we have proposed a general framework for the analysis of the complete set of log ORs generated by a two-way contingency table. The road map of the general framework is the following: we started from the RC (M) association model and we hypothesized a Poisson distribution for the counts of the two-way contingency table. Then for parameter estimations of RC (M) we use an alternative approach based on least squares. The matrix for the estimation of bilinear part of RC (M) has been linked to the matrix used in log-ratio analysis (Greenacre, 2009), moreover we have extended log-ratio analysis (LRA) to the study of log ORs. Then we have connected these methodologies with De Rooiij and Anderson's approach and we have introduced some important properties that have allowed us to have deeper knowledge on the association between variables troughs ORs. Differently from De Rooiij and Anderson, in our approach we have chosen to consider the matrix of a complete set of ORs, because, as we will show, this matrix is linked to the two-way contingency table and it allows us to represent all the ORs on a factorial plan. Moreover, the spanning cell odds ratios are useful when one of the categories defines a control or reference group. In that case, all other categories are described against this reference group. The local odds ratios are useful for ordinal variables when all local odds ratios are larger or equal to 1 (de Rooij and Anderson 2007).

Materials and methods

The research plan - study area, sampling points

Nicknamed "the most polluted river in Europe", the Sarno River originates in south-western Italy and has a watershed of about 715 km^2. An intensive sampling campaign was conducted in the spring of 2008. Surface sediment samples were collected at four locations along the Sarno river (near the source of the river, just before and after the junction with Alveo Comune and at the river mouth) and nine points in the continental shelf around the river mouth (three points, one for each direction North-West, West and South-West, were sampled 50 m from the Sarno River mouth, another three points 150 m away and, finally, another three points 500 m from the river mouth). The collected data were arranged in a two-way contingency table. The row variable is TPAHs (X) with three categories: Low (L), Medium (M), High (H) and column variable is the sampling points (Y) with five categories: Source (S), River (R), 50 m from the Sarno River mouth (50 m), 150 m from the Sarno River mouth (150 m), 500 m from the Sarno River mouth (500 m).

Log-ratio and log odds ratio analysis

Notations

Let $\mathbf{N} = (n_{ij})$ be a two-way contingency table that cross-classifies n units according to I row categories and J column categories of X and Y variables, respectively. Let X_i and Y_j be the i-th and j-th category of X and Y and let π_{ij} the probability that $X = X_i$ and $Y = Y_j$. The matrix of proportions is denoted by $\mathbf{P} = n^{-1}\mathbf{N}$ with general term p_{ij}. The marginal relative frequencies of the i-th row and j-th column of $\mathbf{P}$ are $p_{i\bullet}$ and $p_{\bullet j}$ and they may be represented in vector or matrix form. In this paper, the vector $\mathbf{r}$ (resp. $\mathbf{c}$) consists of $p_{i\bullet}$ (resp. $p_{\bullet j}$) as elements, while $\mathbf{D}_r$ (resp. $\mathbf{D}_c$) is the diagonal matrix of these quantities.

Let $OR_{ii'jj'} = \frac{n_{ij}n_{i'j'}}{n_{i'j}n_{ij'}}$ $(1 \le i < i' \le I;\ 1 \le j < j' \le J)$ be the OR, the complete set of ORs for table $\mathbf{N}$ is composed by $[I(I-1)]/2 \times [J(J-1)]/2$ ORs and it can be placed in a two-way table, called $\mathbf{S} = \left[s_{\tilde{i}\tilde{j}}\right]$, of dimension $\tilde{I} \times \tilde{J}$, where $\tilde{I} = I(I{-}1)/2$ and $\tilde{J} = J(J{-}1)/2$.

From association model to log Ratio Analysis

The association models (Goodman, 1985) are widely used to analyse two-way contingency tables. The first proposed version was the RC (1) association model (Goodman, 1979), then it was extended to the RC (M) association model to decompose the symmetric association into M components (Goodman, 1985). If $M = \min[(I-1), (J-1)]$, this model is called saturated. The RC (M) association model is given by

$$\pi_{ij} = \alpha_i\beta_j \exp\left(\sum\nolimits_{m=1}^{M} \varphi_m\mu_{im}\nu_{jm}\right)$$

where μ_{im} and ν_{jm} are X_i and Y_j scores on dimension m (standard coordinates), φ_m is a measure of the strength of the association between X and Y, α_i and β_j are the main effects of X and Y, respectively. With respect to the scores, the following constraints are assumed: $\sum\nolimits_{i=1}^{I}\pi_{i\bullet}\mu_{im} = \sum\nolimits_{j=1}^{J}\pi_{\bullet j}\nu_{jm} = \sum\nolimits_{i=1}^{I}\pi_{i\bullet}\mu_{im}\mu_{im'} = \sum\nolimits_{j=1}^{J}\pi_{\bullet j}\nu_{jm}\nu_{jm'} = 0$, and $\sum\nolimits_{i=1}^{I}\pi_{i\bullet}\mu_{im}^2 = \sum\nolimits_{j=1}^{J}\pi_{\bullet j}\nu_{jm}^2 = 1$where $\pi_{i\bullet} = \sum\nolimits_{j=1}^{J}\pi_{ij}$ and $\pi_{\bullet j} = \sum\nolimits_{i=1}^{I}\pi_{ij}$.

Assuming the previous constraints and that the distribution of counts within IJ categories is a multinomial distribution with parameters n and π_{ij}, the parameter estimation is computed by the maximum likelihood method.

An alternative estimation method is based on the least square procedure. Let $N_{ij} \sim \mathrm{Po}(n\pi_{ij} = \tau_{ij})$ be a random variable, if we perform the logarithm transformation we obtain the difference $\log(N_{ij}/n) - \log(\pi_{ij})$.

Replacing the random variable with its sample values and considering the RC (M) association model we have

$$log\left(p_{ij}\right) = log(\hat{\alpha}_i) + log\left(\hat{\beta}_j\right) + \sum_{m=1}^{M}\lambda_m u_{im} v_{jm}$$

Substituting the probabilities with observed frequency, taking into account the constraints and the condition $\sum_{i=1}^{I} log(\hat{\alpha}_i)p_{i\bullet} = 0$, we estimate the parameters $\log(\beta_j)$ and $\log(\alpha_i)$ as follows: $\log\left(\hat{\beta}_j\right) = \sum_{i=1}^{I} p_{i\bullet} \log\left(p_{ij}\right)$ and $\log(\hat{\alpha}_i) = \sum_{j=1}^{J} p_{\bullet j} \log\left(p_{ij}\right) - \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet} p_{\bullet j} \log\left(p_{ij}\right)$.

The estimation of the bilinear part is obtained through the least squares method (D’Ambra, 1988; Escoufier and Junga 1986), minimizing the quantity

$$min\left[\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet} p_{\bullet j}\left(a_{ij} - \sum_{m=1}^{M}\lambda_m u_{im} v_{jm}\right)^2\right]$$

where $a_{ij} = \log\left(p_{ij}\right) - \log\left(\hat{\beta}_j\right) - \log(\hat{\alpha}_i) = \log\left(p_{ij}\right) - \sum_i^I p_{i\bullet} \log\left(p_{ij}\right) - \sum_j^J p_{\bullet j} \log\left(p_{ij}\right) + \sum_i^I p_{i\bullet} \sum_j^J p_{\bullet j} \log\left(p_{ij}\right)$.

We have noted that a_{ij} is equivalent to the residual of the two-way analysis of variance. The same matrix $\mathbf{A} = (a_{ij})$, used in RC (M) association model, is analysed in Log-ratio analysis (Greenacre, 2009). Greenacre, starting from Correspondence Analysis (CA) and using Box-Cox transformation of p_{ij}^{α} (with $\alpha \to 0$) applied a SVD on the following matrix

$$\begin{aligned}\mathbf{Z} &= \mathbf{D}_r^{1/2}\left(\mathbf{I} - \mathbf{1}\mathbf{r}^T\right) L\left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{1/2}\right)\left(\mathbf{I} - \mathbf{1}\mathbf{c}^T\right)^T \mathbf{D}_c^{1/2} \\ &= \mathbf{D}_r^{1/2}\left(\mathbf{I} - \mathbf{1}\mathbf{r}^T\right) L(\mathbf{N})\left(\mathbf{I} - \mathbf{1}\mathbf{c}^T\right)^T \mathbf{D}_c^{1/2} \\ &= \mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c^{1/2}\end{aligned}$$

where L means logarithm transformation. Based on the different centring system, a comparison among CA, weighted LRA, and RC (M) has been done (Greenacre and Lewi 2009). For analogical criteria the weighted system of weighted LRA is the same of CA. This choice could be justified in a better way as follows.

Considering N_{ij}, when $\tau_{ij} \to +\infty$, then $\frac{\left(N_{ij} - \tau_{ij}\right)}{\sqrt{\tau_{ij}}}$ is a random variable with normal standard distribution[a]. If we consider the random variable $\sqrt{\tau_{ij}}\left[log\left(N_{ij}\right) - log\left(\tau_{ij}\right)\right] = \sqrt{\tau_{ij}}\, log\left[1 + \frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right]$, applying the Taylor series, we can say that $\left[\frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right]$ provides a useful approximation to $log\left[1 + \frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right]$ when $\left|\frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right| < 1$, thus $log\left[1 + \frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right] = \sum_t^{\infty} \frac{(-1)^{t+1}}{t}\left[\frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right]^t \approx \frac{N_{ij} - \tau_{ij}}{\tau_{ij}}$. Observing that $\sqrt{\tau_{ij}}\frac{N_{ij} - \tau_{ij}}{\tau_{ij}} = \frac{N_{ij} - \tau_{ij}}{\sqrt{\tau_{ij}}} \sim \mathrm{N}(0, 1)$, it follows that $\sqrt{\tau_{ij}}\left[log\left(N_{ij}\right) - log\left(\tau_{ij}\right)\right] = \sqrt{\tau_{ij}} log\left[1 + \frac{N_{ij} - \tau_{ij}}{\tau_{ij}}\right] \sim \mathrm{N}(0, 1)$, then $E[\log(N_{ij})] \cong \log(\tau_{ij})$ and $Var[\log(N_{ij})] \cong 1/\tau_{ij}$. Under the independence hypothesis τ_{ij} can be estimated by $np_{i\bullet}p_{\bullet j}$, justifying the weighting system based on row and column marginal totals of $\mathbf{P}$.

For the foregoing the association between the categories of X and Y variables could be studied, by performing a SVD of the double-centred matrix $\mathbf{Z}$ with respect to $p_{i\bullet}$ and $p_{\bullet j}$, $: \mathbf{Z} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \sum_{m=1}^{M}\mathbf{u}_m\lambda_m\mathbf{v}_m^T$ where $M = rank(\mathbf{Z}) = \min[(I-1), (J-1)]$ with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ where $\mathbf{u}_m$ is the m-th column of $\mathbf{U}$, $\mathbf{v}_m$ is the m-th column of $\mathbf{V}$ and the singular values down the diagonal of $\mathbf{\Lambda}$ are in descending order $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_M$. The total variance in weighted LRA can be written in terms of the complete set of the log ORs

$$tr\left(\mathbf{Z}^T\mathbf{Z}\right) = \sum_{i<i'}\sum\sum_{j<j'}\sum p_{i\bullet}p_{i'\bullet}p_{\bullet j}p_{\bullet j'}\left[logOR_{ii'jj'}\right]^2.$$

The principal and standard coordinates for rows and columns are computed as follows:
$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{\Lambda}$, $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{\Lambda}$, $\tilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{U}$, $\tilde{\mathbf{G}} = \mathbf{D}_c^{-1/2}\mathbf{V}$.

Setting $\mathbf{r} = (1/I)\mathbf{1}$ and $\mathbf{c} = (1/J)\mathbf{1}$ we obtain the unweighted Log Ratio Analysis (Aitchison 1990)[b]:

$$\mathbf{Z}^U = (IJ)^{-1/2}\left(\mathbf{I} - (1/I)\mathbf{1}\mathbf{1}^T\right)L(\mathbf{N})\left(I - (1/J)\mathbf{1}\mathbf{1}^T\right)$$

The LRA decomposes Altham’s measure for $Q = 2$ and the complete set of ORS, in fact:

$$tr\left(\left(\mathbf{Z}^U\right)^T\mathbf{Z}^U\right) = \left(\frac{1}{\tilde{I}\tilde{J}}\sum_{\tilde{i}}\sum_{\tilde{j}}\left[\log s_{\tilde{i}\tilde{j}}\right]^Q\right)^{1/2}$$

Althman’s index is an association measure based on the ORs. It can also be defined on the local odds ratio and spanning cell odds ratios. For a deep discussion of this measure, see Edwardes and Baltzan (2000).

Weighted LRA properties

The weighted LRA preserves the underlying properties the fundamental characteristics of classical CA: coordinates properties, distance measures, a reconstitution formula, rank of decomposed matrix. It is a powerful tool for analysing compositional data (Aitchison and Greenacre 2002). The weighted LRA to the specific case of log ORs has been introduced. As in RC (M) models (de Rooij and Heiser 2005) the row and column coordinates of the weighted LRA satisfy the following two important properties:

$$OR_{ii'jj'} = \exp\left(\sum_{m=1}^{M}\lambda_m\left(\tilde{f}_{im} - \tilde{f}_{i'm}\right)\left(\tilde{g}_{jm} - \tilde{g}_{j'm}\right)\right) \quad (1)$$

$$OR_{ii'jj'} = \exp\left(\frac{1}{2}d^2\left(\mathbf{f}_i, \mathbf{g}_j\right) + \frac{1}{2}d^2\left(\mathbf{f}_i, \mathbf{g}_j\right) - \frac{1}{2}d^2\left(\mathbf{f}_i, \mathbf{g}_j\right) - \frac{1}{2}d^2\left(\mathbf{f}_i, \mathbf{g}_j\right)\right) \quad (2)$$

Where $d^2(\mathbf{f}_i, \mathbf{g}_j)$ is the squared Euclidean distance between the points with coordinates $\mathbf{f}_i$ and $\mathbf{g}_j$ on the m dimensions. Thanks to these properties the factorial representation of the weighted LRA can be explained both in terms of inner product rule (type I), and distance rule (type II). For type I representation, at least one coordinate set should be drawn using vectors, and the points of the other set projected on these vectors to represent the relationship. For type II the categories for both sets can be represented by points in Euclidean space, with the distance between the points describing the relationship between categories of two sets. These properties permit to visualize in the same factorial plan the categories and the log-ORs. Unfortunately, these important properties do not work for unweighted LRA.

Let $\tilde{\mathbf{f}}_{i*}$ and $\tilde{\mathbf{g}}_{j*}$ be the baseline of row and column score vectors, respectively. Substituting these baselines for the average with respect to i and j, in this case zero vectors ($\tilde{\mathbf{f}}_{i*} = \mathbf{0}$ and $\tilde{\mathbf{g}}_{j*} = \mathbf{0}$), and taking into account formula (1), the OR of the pair of categories ij-*th* respect the baseline (Eshima et al., 2001) can be defined

$$OR_{\tilde{\mathbf{f}}_i \mathbf{0} \tilde{\mathbf{g}}_j \mathbf{0}} = \exp\left(\sum\nolimits_{m=1}^{M} \lambda_m \tilde{f}_{im} \tilde{g}_{jm}\right)$$

This OR is theoretical and could be interpreted as the contribution of the pair of categories ij-*th* towards the association between X and Y variables. Considering log transformation and using vectors, the previous quantity can be written as

$$\log OR_{\tilde{\mathbf{f}}_i \mathbf{0} \tilde{\mathbf{g}}_j \mathbf{0}} = \tilde{\mathbf{f}}_i \boldsymbol{\Lambda} \tilde{\mathbf{g}}_j^T.$$

Denoting by $\bar{\mathbf{S}}$ the matrix of dimension $I \times J$, whose generic element is $\log OR_{\tilde{\mathbf{f}}_i \mathbf{0} \tilde{\mathbf{g}}_j \mathbf{0}}$, we have $\bar{\mathbf{S}} = \tilde{\mathbf{F}} \boldsymbol{\Lambda} \tilde{\mathbf{G}}^T$

In order to compute a synthesis measure of the complete set of ORs, the OR mean (Me) can be calculated by using formula (2)

$$\begin{aligned} Me\left(OR_{ii'jj'}\right) = {} & \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} \sum\nolimits_{m=1}^{M} \exp\left(\frac{1}{2}\left[\left(f_{i'm} - g_{jm}\right)^2 p_{i'j}\right.\right. \\ & \left.\left. + \left(f_{im} - g_{j'm}\right)^2 p_{ij'} - \left(f_{im} - g_{j'm}\right)^2 p_{ij} - \left(f_{im} - g_{j'm}\right)^2 p_{i'j'}\right]\right) \end{aligned}$$

Replacing $f_{i'm}$ and $g_{j'm}$ by means with respect to *i'* and *j',* in this case zero vectors, we have:

$$\begin{aligned} Me\left(OR_{i0j0}\right) = {} & \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} \sum\nolimits_{m=1}^{M} \exp\left(\frac{1}{2}\left[\left(0 - g_{jm}\right)^2 p_{i'j}\right.\right. \\ & \left.\left. + (f_{im} - 0)^2 p_{ij'} - (f_{im} - g_{im})^2 p_{ij} - (0 - 0)^2 p_{i'j'}\right]\right) \\ = {} & \sum\nolimits_{m=1}^{M} \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} f_{im} g_{jm} p_{ij} \end{aligned}$$

Using standard coordinates we obtain

$$Me\left(OR_{i0j0}\right) = \sum\nolimits_{m=1}^{M} \lambda_m \sum\nolimits_{i=1}^{I} \sum\nolimits_{j}^{J} \tilde{f}_{im} \tilde{g}_{jm} p_{ij} = \sum\nolimits_{m=1}^{M} \lambda_m \rho_m \quad (3)$$

In the RC (M) model, this quantity is expressed by the Kullback–Leibler information (Eshima et al., 2001). This property is also true for the weighted LRA:

$$\begin{aligned} Me\left(OR_{i0j0}\right) = {} & \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} p_{ij} \log\left(p_{ij} / p_{i\bullet} p_{\bullet j}\right) \\ & + \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} p_{i\bullet} p_{\bullet j} \log\left(p_{i\bullet} p_{\bullet j} / p_{ij}\right) \\ = {} & \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} p_{ij} \log p_{ij} - \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} p_{i\bullet} p_{\bullet j} \log p_{ij} \\ = {} & \sum\nolimits_{i=1}^{I} \sum\nolimits_{j=1}^{J} \left(p_{ij} - p_{i\bullet} p_{\bullet j}\right) \log p_{ij} \end{aligned}$$

This quantity shows the departure of the assumption of independence. As $Me(OR_{i0j0})$ is expressed by the Kullback–Leibler information, the larger this mean is, the stronger the association between X and Y is. Dividing this mean by the sum of singular values, an index for studying the relationship between the X and Y variables based on the log ORs is obtained:

$$I(X, Y) = \sum\nolimits_{m=1}^{M} \lambda_m \rho_m / \sum\nolimits_{m=1}^{M} \lambda_m$$

while the quantity $c_m = \lambda_m \rho_m / \sum\nolimits_{m=1}^{M} \lambda_m$ represents the contribution of the m-th pair of coordinate vectors to the relationship between the X and Y variables.

From LRA to log-odds Ratio Analysis

The weighted LRA permits the indirect representation of the log ORs. Starting from the de Rooij and Anderson approach (2007) we have proposed a methodology based on the singular value decomposition of log OR matrix for obtaining its direct representation. Summary measures of association were also obtained. Unlike de Rooij and Anderson, the method has been applied to the matrix with the complete set of log odds ratios because, as seen later, it is linked to LRA (see below).

Let $L(\mathbf{S})$ be a two-way table of dimension $\tilde{I} \times \tilde{J}$ containing the complete set of log ORs, in this table the rows (resp. columns) are formed by all pairs of categories of X (resp. Y). Let **B** and **D** be two square diagonal matrices of dimensions $\tilde{I}$ and $\tilde{J}$ respectively with general terms $1/\tilde{I}$ and $1/\tilde{J}$ Performing a SVD of L(**S**) with the matrices **B** and **D,** we have:

$$\mathbf{B}^{1/2} L(\mathbf{S}) \mathbf{D}^{1/2} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}^T$$

We have called this analysis Unweighted Log Odd Ratio analysis (ULORA). ULORA is linked to unweighted LRA, and, consequently, is joined with the Altman measure:

$$\begin{aligned} tr\left(\mathbf{B}^{1/2} L(\mathbf{S}) \mathbf{D} L(\mathbf{S}^T) \mathbf{B}^{1/2}\right) &= tr\left(\left(\mathbf{Z}^U\right)^T \mathbf{Z}^U\right) \\ &= \left(\frac{1}{\tilde{I}\tilde{J}} \sum\nolimits_{\tilde{i}} \sum\nolimits_{\tilde{j}} \left[\log s_{\tilde{i}\tilde{j}}\right]\right)^2. \end{aligned}$$

Table 1 Complete set of log ORs

	SR	S50m	S150m	S500m	R50m	R150m	R500m	50m150m	50m500m	150m500m
LM	3.296	4.171	4.394	1.399	0.875	1.099	−1.897	0.223	−2.773	−2.996
LH	2.187	2.854	3.045	−1.268	0.667	0.858	−3.455	0.191	−4.123	−4.313
MH	−1.109	−1.317	−1.350	−2.668	−0.208	−0.241	−1.558	−0.033	−1.350	−1.317

This method does not take into account the weight structures of the rows and columns. Let $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{D}}$ be two square diagonal matrices of dimensions $\tilde{I}$ and $\tilde{J}$ respectively with general terms $p_{i\bullet}p_{i'\bullet}$ and $p_{\bullet j}p_{\bullet j'}$. Performing a SVD of $L(\mathbf{S})$ with the matrices $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{D}}$ we get a weighted analysis of the log OR matrix (WLORA):

$$\tilde{\mathbf{B}}^{1/2}L(\mathbf{S})\tilde{\mathbf{D}}^{1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$$

The coordinates are:

$$\breve{\mathbf{F}} = \mathbf{U}\boldsymbol{\Lambda}\,,\ \breve{\mathbf{G}} = \mathbf{V}\boldsymbol{\Lambda}$$

We obtain a factorial representation in which pairs of categories of X and Y are drawn. Following this approach, at least one coordinate set should be drawn using vectors, and the points of the other set projected on these vectors to represent the weighted ORs. In this case we show that:

$$tr\left(\mathbf{Z}^T\mathbf{Z}\right) = tr\left(\tilde{\mathbf{B}}^{1/2}L(\mathbf{S})\tilde{\mathbf{D}}L\left(\mathbf{S}^T\right)\tilde{\mathbf{B}}^{1/2}\right) \\ = \sum\nolimits_{i<i'}\sum\sum\nolimits_{j<j'}\sum p_{i\bullet}p_{i'\bullet}p_{\bullet j}p_{\bullet j'}\left[\log OR_{ii'jj'}\right]^2.$$

Therefore both the weighted LRA and ULORA decompose a synthetic measure of the log ORs. This last one is a weighted version of Altham's measure.

Results and discussion

The association between TPAHs (X) and sampling points (Y) is significative, with Pearson's chi-squared equal to 687.017. The complete set of log ORs is computed (Table 1).

The log ORs are very different from 0, therefore the association is confirmed. The synthesis of the complete set of log ORs can be performed computing the Altham index and formula (3), the results are 0.72158 and 0.62545, respectively.

Afterwards the unweighted and weighted LRA of the two-way contingency table have been performed. The number of dimensions to be retained is two and the factorial representations have been presented (Figure 1, Figure 2). In these representations we have the categories of X and Y. In Figure 1 we can observe that on the first axis there is a juxtaposition between a low level of pollution and a medium-high level of pollution, with "Source" and "500 meters" associated with a low level of pollution and the other categories of X ("River", "50 meters" and "150 meters") linked with a medium-high level of pollution. Figure 2, instead, appears more readable and interesting thanks to the effect of the weighting system. In fact "Source" and "500 meters" remain at a low level of pollution, but the other group is further divided in two more homogeneous groups: "River" associated with a

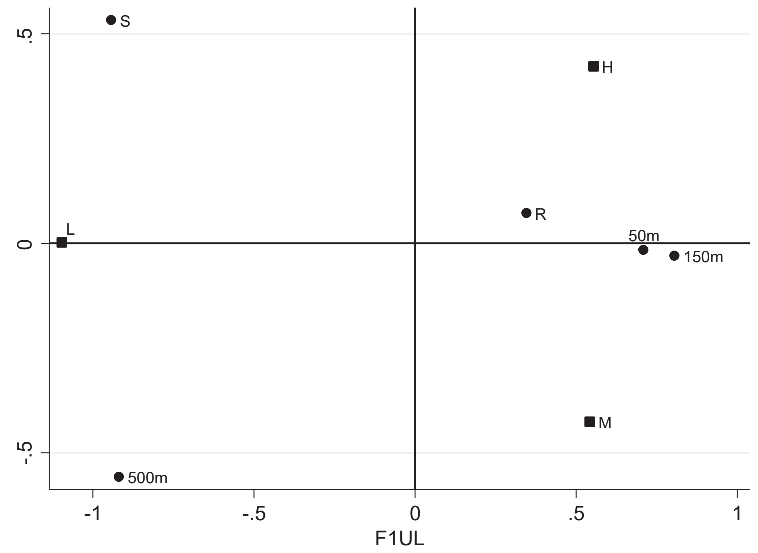

Figure 1 Factorial representation for unweighted LRA.

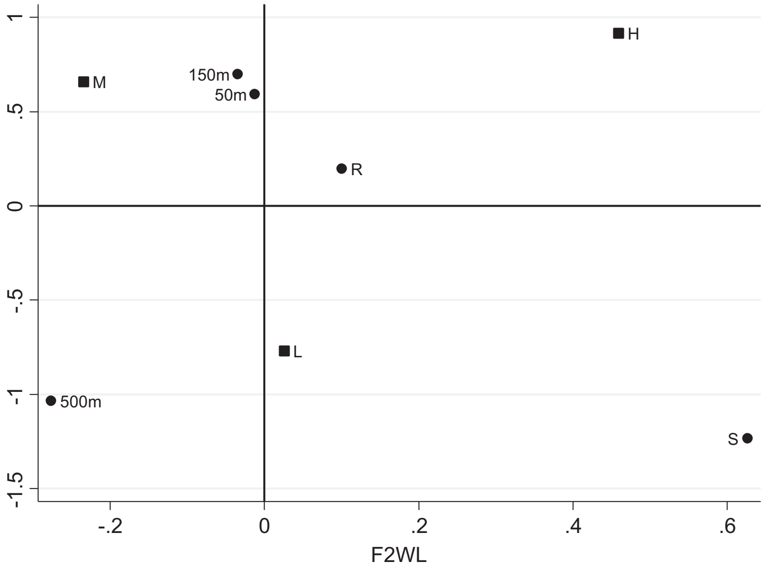

Figure 2 Factorial representation for weighted LRA.

high level of pollution and "50-150 meters" with a medium level of pollution.

In order to improve the data analysis we have divided the data table into three sub-tables, in which the last three categories of X have been further subdivided into three sub-levels concerning the direction of detection (North-West, West, South-West). This division was made because it was found that the level of pollution of the sea is influenced by the direction of detection. In this paper we have showed only the direction South-West (for the other analysis the authors can be contacted). In factorial representation of unweighted LRA (Figure 3) three associations are clear: "500 meters" with a low level of pollution, "50 and 150 meters" with a medium level of pollution and "River" with a high level of pollution. The position of category "Source" is ambiguous, in fact it is in the middle between low and high level of pollution. In our opinion, this ambiguity depends on the different values of the margins of the data table. In order to take into account this feature of the data, the weighted LRA, that allows us to include a system of weights, has been performed. The factorial representation (Figure 4) is better, in fact the classification of the ambiguous category "Source" has been resolved and is correctly associated with a low

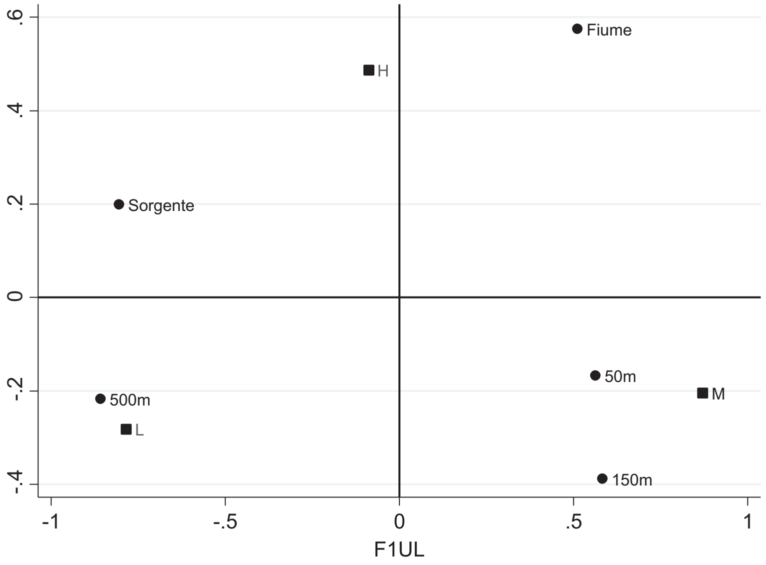

Figure 3 Factorial representation for unweighted LRA (South-West).

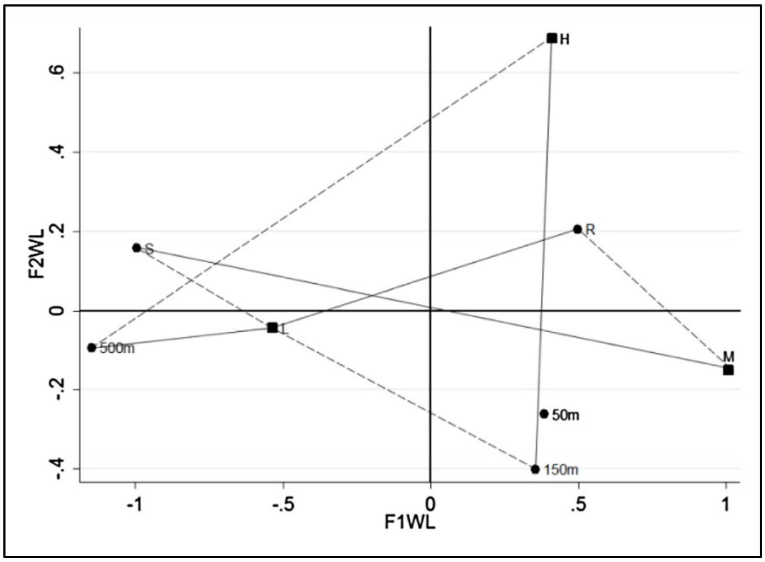

Figure 4 Factorial representation for weighted LRA (South-West).

level of pollution. As the marginal relative frequencies of data table (rows: 0.614, 0.284, 0.102; columns: 0.149, 0.419, 0.145, 0.143, 0,144) are different, then weighted LRA is preferred. Moreover, for weighted LRA, the indirect representation of log-ORs can be appreciated. For example considering $\log OR_{L,M;S,R}$ and $\log OR_{L,H;R,500}$, according to formula (2) the log-OR depends on the length of the distances between categories. In our case $\log OR_{L,M;S,R}$ is clearly greater than 1, since the solid lines are much longer than the dotted ones, for the second $\log OR_{L,H;R,500}$ the contrary happens therefore it is smaller than 1. We fitted the RC (2) association model using the marginal proportions as weights. The results are very similar to those obtained by the weighted LRA.

Subsequently to study the association between X and Y we have considered the matrix of the complete set of log ORs (Table 2).

The ULORA and WLORA have been performed. The factorial representations, on retained axes, are in Figure 5 and Figure 6. At first glance, the factorial representations look very similar but differences exist, as a matter of fact in Figure 6 the category "SR" and "MH" have coordinates greater than in Figure 5.

This is a consequence of the system of weights; in fact in WLORA we have taken into consideration the marginal relative frequencies of the rows and the columns as a weighting system. When there are large differences among the marginal relative frequencies of the rows or of the columns, it could be relevant to take this information into account, therefore the WLORA is preferred. ULORA and WLORA permits the visualization of the log-ORs through the projection of the column points onto the row vectors (or viceversa). For example, we have focalized our attention on three log-ORs computed using the same row vector: $\log OR_{L,M;S,R}$, $\log OR_{L,M;S,50}$ and $\log OR_{L,M;S,150}$. The projections are different in the two analysis depending on the system of weights. In Figure 6 we can see that there is a strong association between the categories "Source-River" and Low-High level of pollution, "Source-50 meter" and Low-Medium level of pollution and "Source-150 meter" and Low-Medium level of pollution. This association can be justified both by the proximity of the categories and through the log-ORs (represented by projections of the first category on the second). Moreover, in this case it is also possible to make an interpretation in terms of variations. In fact, the $\log OR_{L,H;S,R}$, tells us that when you switch from "Source" to "River" is very likely that the level of pollution jumps from Low to High, the same happens when we switch from category "Source" to "50 meters", where it is very likely that the pollution goes from low to medium. Therefore, given the sequence of the measured

Table 2 Complete set of log ORs (South-West)

	SR	S50m	S150m	S500m	R50m	R150m	R500m	50m150m	50m500m	150m500m
LM	3.296	3.255	3.255	−0.223	−0.041	−0.041	−3.520	0.000	−3.477	−3.477
LH	2.187	0.588	0.118	−0.981	−1.599	−2.069	−3.168	−0.470	−1.569	−1.099
MH	−1.109	−2.668	−3.137	−0.758	−1.558	−2.028	0.351	−0.470	1.910	2.380

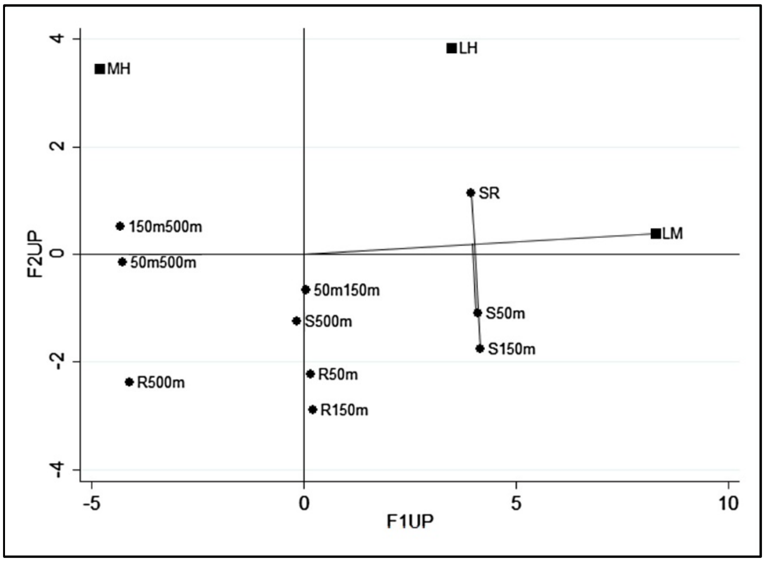

Figure 5 Factorial representation for ULORA (South-West).

points, it is possible to argue that passing from "River" to see the pollution decreases from high to medium.

Conclusions

In this paper we discussed the RC (M) association model and weighted LRA providing a justification for the logarithm transformation and weighting system. RC (M) association models and LRA are based on Newton–Raphson (NR) algorithm and the SVD, respectively. The convergence of the NR depends on the starting point. On the contrary the SVD is extremely stable and computationally simpler. Regarding the number of dimensions, the LRA allowed to choose the number of dimensions to be retained later. A criterion could be the variance explained by the first components. Another criterion could be the application of a bootstrap or jackknife procedure for verifying the stability of singular values. In the RC (M) association models, for a given dimensionality, main effects and interaction terms were estimated. Then we extended LRA to the study of log-ORs, obtaining the indirect representation of the log ORs and its synthesis measures. Finally we applied the SVD to the unweighted and weighted Log-ORs matrix (ULORA and WLORA) obtaining a direct representation of log-ORs. We also

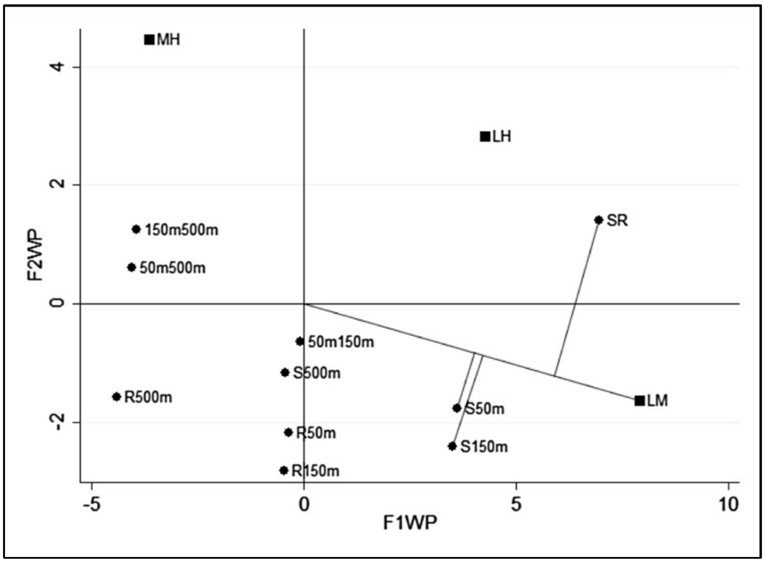

Figure 6 Factorial representation for WLORA (South-West).

got summary measures of association. The ULORA and WLORA were applied to the complete set of log-ORs for linking these methods to LRA.

In the further study we intend to extend the introduced methodologies to three-way contingency table (Gallo and Simonacci 2013), to the other types of ORs (i.e. cumulation, continuation and global) and to the ratio of two contingency tables. Considering that the two contingency tables are of same dimensions: one representing the target population, X_{ij} the second a subset of this population with a specific character, Y_{ij}. Let $X_{ij} \sim \mathrm{Po}(\tau_{ij})$ and $Y_{ij}| X_{ij} = k_{ij} \sim \mathrm{Bin}(\mathrm{k_{ij}}, p_{ij})$ be. One demonstrates $Y_{ij} \sim \mathrm{Po}(\tau_{ij}p_{\mathrm{ij}})$. Then the methodologies presented could be extended to the analysis of ratios $r_{ij} = y_{ij}/x_{ij}$.

Endnotes

[a] A continuity and symmetry correction can also be applied when one discrete distribution is approximated by the normal distribution.

[b] Other weight systems can be applied, for example $p_{i\bullet} + p_{\bullet j}$ for squared tables.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors read and approved the final manuscript.

Author details

[1]Department of Economics, Second University of Naples, Corso Gran Priorato di Malta, 81043 Capua, CE, Italy. [2]University of Rome Unitelma Sapienza, Viale Regina Elena 295, 00161 Rome, Italy.

Received: 27 March 2014 Accepted: 30 June 2014
Published: 28 July 2014

References

Aitchison JS (1990) Relative variation diagrams for describing patterns of variability in compositional data. Math Geol 22:487–512

Aitchison J, Greenacre MJ (2002) Biplots of compositional data. Appl Stat 51:375–392

Altham PME (1970) The measurement of association of rows and columns for an r x s contingency table. J R Stat Soc B 32:63–73

Bihari N, Fafandel M, Hamer B, Kralj-Bilen B (2007) PAH content, toxicity and genotoxicity of coastal marine sediments from the Rovinj area, Northern Adriatic, Croatia. Sci Total Environ 366:602–611

Caricchia AM, Chiavarini S, Cremisini C, Marrini F, Morabito R (1993) PAHs, PCBs and DDE in the Northern Adriatic Sea. Mar Pollut Bull 26:581–583

Choulakian V (1988) Exploratory analysis of contingency tables by loglinear formulation and generalization o Correspondence analysis. Psychometrika 53(2):235–250

D'Ambra L (1988) Least squares criterion for asymmetric dependence models in three-way contingency table. Unité de biometrie, Montpellier technical report n° 8802

de Rooij M, Anderson CJ (2007) Visualizing, Summarizing, and Comparing Odds Ratio Structures. Methodology: Eur J Res Methods Behav Soc Sci 3(4):139–148

de Rooij M, Heiser WJ (2005) Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. Psychometrika 70:99–123

Edwardes MD, Baltzan M (2000) The generalization of the odds ratio, risk ratio and risk difference to r × k tables. Stat Med 19:1901–1914

Escoufier Y, Junga S (1986) Least squares approximation of frequencies or their logarithms. Int Stat Rev 54(3):279–283

Eshima N, Tubata M, Tsujitani M (2001) Property of the RC (M) association model and a summury measure of association in the contingency table. J Japan Stat Soc 31(1):15–26

Gallo M, Simonacci V (2013) A procedure for three analysis of compositions. Electron J Appl Stat Anal 06(02):202–210

Goodman LA (1979) Simple models for the analysis of association in cross classifications having ordered categories. J Am Stat Assoc 74:537–552

Goodman LA (1985) The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetric models for contingency tables with or without missing entries. Ann Stat 13:10–69

Greenacre MJ (2009) Power transformations in correspondence analysis. Comput Stat Data Anal 53(8):3107–3116

Greenacre MJ, Lewi PJ (2009) Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency table and ratio scale measurement. J Classif 26(1):29–54

Lehr RE, Jerima DM (1997) Metabolic activations of polycyclic hydrocarbons. Arch Environ Contam Toxicol 39:1–6

Luo XJ, Chen SJ, Mai BX, Yang QS, Sheng GY, Fu JM (2006) Polycyclic aromatic hydrocarbons in suspended particulate matter and sediments from the Pearl River Estuary and adjacent coastal areas, China. Environ Pollut 139:9–20

Sarnacchiaro P, Diez S, Montuori P (2012) Polycyclic Aromatic Hydrocarbons Pollution in a Coastal Environment: the Statistical Analysis of Dependence to Estimate the Source of Pollution. Curr Anal Chem 8:300–309

Tolosa I, Bayona JM, Albaigés J (1995) Aliphatic and polycyclic aromatic hydrocabons and sulfur/oxygen derivatives in northwestern Mediterranean sediments: spatial and temporal vatriability, fluxes and budgets. Environ Sci Technol 29:2519–2527

White KL (1986) An overview of immunotoxicology and polycyclic aromatic hydrocarbons. Environ Carcinogenesis Rev 2:163–202

Xu J, Yu Y, Wang P, Guo W, Dai S, Sun H (2007) Polycyclic aromatic hydrocarbons in the surface sediments from Yellow River, China. Chemosphere 67:1408–1414

Yan LS (1985) Study of carcinogenic mechanisms for aromatic hydrocarbons – extended bay region theory and its quantitative model. Carcinogenesis 6:1–6

doi:10.1186/2193-1801-3-384
Cite this article as: Camminatiello *et al.*: The association in a two-way contingency table through log odds ratio analysis: the case of Sarno river pollution. *SpringerPlus* 2014 3:384.