# General marginal-free association indices for contingency tables: From the Altham index to the intrinsic association coefficient

Milan Bouchet-Valat

# General Marginal-Free Association Indices for Contingency Tables: From the Altham Index to the Intrinsic Association Coefficient

Milan Bouchet-Valat*

**Abstract:**
Notwithstanding a large body of literature on log-linear models and odds ratios, no general marginal-free index of the association in a contingency table has gained a wide acceptance. Building on a framework developed by L. A. Goodman, we put into light the direct links between odds ratios, the Altham index, the intrinsic association coefficient and coefficients in log-multiplicative models including Unidiff and RC(M) association models. We devise a normalized version of the latter coefficient varying between 0 and 1, which offers a simpler interpretation than existing indices similar to the correlation coefficient. We illustrate with the case of educational and socioeconomic homogamy among 149 European regions how this index can be used either alone in a non- or semi-parametric approach or combined with models, and how it can protect against incorrect conclusions based on models which rely on strong assumptions to summarize the strength of association as a single parameter.

**Keywords:** association index; Altham index; intrinsic association coefficient; odds ratio; log-linear model; homogamy; intergenerational mobility.

Despite the existence of a substantial and long-standing literature on odds ratios and log-linear models, it is surprising that no general marginal-free index of the association between categorical variables has become standard. While a number of indices based on Pearson contingencies (such as the mean square contingency coefficient $\varphi^2$ or Cramér's V) are frequently used in various fields, no equivalent index exists for researchers in need for the odds ratio's essential property of margin-insensitivity (Bishop, Fienberg, and Holland [1975] 2007:11; Liebetrau 1983).

This lack is particularly striking in studies of intergenerational social mobility or homogamy, which rely heavily on contingency table analysis. Indeed, in the recent years the economic literature on intergenerational mobility has much developed, focusing on income rather than categorical measures like social class. One of the strengths of this literature rests in its methodological unification, with the use of intergenerational income elasticities or correlations as standard tools. We suggest that the lack of such a standard index is a comparative disadvantage for sociology in this field. As Jo Blanden (2013:44) notes in her comparison of the two approaches: "It is one of the disadvantages of the social class literature that there is not a more intuitive summary measure of mobility; for the purpose of this summary we would benefit greatly from a single mobility parameter for each nation and point in time, which could be easily compared with the measures for income and education mobility."

Yet, several marginal-free association indices have been proposed in the literature over the last fifty years. The best known of them, the Altham index (Altham 1970) has not benefited from the attention it deserved, although it has recently gained some popularity in historical studies (Ferrie 2005; Altham and Ferrie 2007; Bourdieu, Ferrie, and Kesztenbaum 2009; Long and Ferrie 2013). As we will show below, a closely related index has been developed under different names by different authors, both in theoretical (Goodman 1996) and empirical (Hout, Brooks, and Manza 1995; Breen et al. 2009) works. Their lack of success may be attributed to difficulty to give them an intuitive interpretation (Bishop et al. [1975] 2007:393).

The objective of the present article is twofold. First, we would like to highlight the value of several related association indices based on the odds ratio. To this end, we will mobilize the framework established in a major article by Leo Goodman (1996) that sought to reconcile two opposed traditions: on the one hand, Pearson contingencies ($\chi^2$) and correspondence analysis; on the other hand, odds ratios and log-linear, log-multiplicative and association models. Using this framework, we will show that the Altham index is very directly related to the intrinsic association coefficient found in log-multiplicative association models. This coefficient is in turn the equivalent in the odds ratio tradition of the mean square contingency coefficient $\varphi^2$ or of Cramér's V in the Pearsonian tradition. Moreover, it has strong connections with classical log-multiplicative models like the layer effect model also known as Unidiff (Xie 1992; Erikson and Goldthorpe 1992) and association models (Wong 2010): analyses carried out using either approach can therefore easily be made comparable, and we will argue that combining them can equip researchers with the best of each method.

As a second goal, we will try to help the adoption of odds ratio-based indices by making them easier to use. Indeed, one possible reason for the lack of success of the Altham index may be difficulty in interpreting it. In an attempt to alleviate this issue, we will first show that the intrinsic association coefficient and the Altham index are equal (up to a multiplicative factor) to the logarithm of the (geometric) standard deviation of all the odds ratios that can be computed from a table. Then, we will propose a new normalization of

*Institut national d'études démographiques (INED), F-75020, Paris, France, and Laboratoire de sociologie quantitative (CREST-LSQ), Palaiseau, France. E-mail: milan.bouchet-valat@ined.fr. A previous version of this work has been presented at the 2015 Spring Meeting of the Research Committee on Social Stratification and Mobility (ISA RC28) in Tilburg (Netherlands). The author would like to thank Louis-André Vallet and Richard Breen for their comments.

the intrinsic association coefficient varying between 0 and 1, similar to the well-known Pearson correlation coefficient, mean square contingency coefficient $\varphi^2$ and Cramér's V. This normalized index has a direct relation with the correlation coefficient in the case of a bivariate normal distribution.

We begin by presenting the general framework as established by Goodman, unifying Pearson's approach and the odds ratio approach, so as to derive the intrinsic association coefficient, an index of the general intensity of association insensitive to the table's margins and dimensions. Then we show that this index is directly related to the standard deviation of all log-odds ratios in a table, and therefore to the Altham index. We then develop the strong relation between these indices derived from the odds ratio and standard log-multiplicative models: Unidiff; RC(M) and RC(M)-L association models. Finally, we illustrate the interest of these association indices for the analysis of the determinants of socioeconomic homogamy in 149 regions of the European Union, using the R package logmult (Bouchet-Valat, 2018b) for the estimations.

# 1 General framework unifying the Pearson and odds ratios traditions

We begin by presenting the general framework established by Goodman (1996) for the analysis of the association in a contingency table, using that article's terminology and notation. The article systematizes results the author gradually developed in a series of papers (in particular Goodman 1986, 1991). We then briefly show how this framework can be used to revisit the Pearson approach from a new perspective, and then we develop the elaboration of the intrinsic association coefficient as a general measure of the association in a contingency table according to the odds ratio tradition.

## 1.1 Preliminary definitions

Let $P_{ij}$ be the proportions observed in the cell belonging to row $i$ and column $j$ in a table of dimensions $I \times J$. In this case the row and column marginal proportions are respectively

$$P_{i+} = \sum_{i=1}^{I} P_{ij} \quad \text{and} \quad P_{+j} = \sum_{j=1}^{J} P_{ij} \tag{1}$$

Using this notation, let us define the Pearson ratio, also known "independence ratio", "mobility ratio" or "homogamy ratio" in the sociological literature:

$$\psi_{ij} = \frac{P_{ij}}{P_{i+}P_{+j}} \quad \text{and a derived quantity:} \quad R_{ij} = \mathrm{R}[\psi_{ij}] \tag{2}$$

with $\mathrm{R}[x]$ a monotonically increasing function called "interaction link".

We may then define the unweighted interaction corresponding to a given cell as

$$\lambda_{ij} = R_{ij} - \frac{1}{I}\sum_{i=1}^{I} R_{ij} - \frac{1}{J}\sum_{j=1}^{J} R_{ij} + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} R_{ij} \tag{3}$$

Similarly we may define the weighted interaction as

$$\tilde{\lambda}_{ij} = R_{ij} - \sum_{i=1}^{I} R_{ij}P_{+j} - \sum_{j=1}^{J} R_{ij}P_{i+} + \sum_{i=1}^{I}\sum_{j=1}^{J} R_{ij}P_{i+}P_{+j} \tag{4}$$

Although for simplicity's sake we use weighting equal to the table's marginal proportions, any strictly positive set of weights summing to unity may be used. This applies in particular to uniform weights equal respectively to $1/I$ for rows and $1/J$ for columns, as we shall see below.

In order to obtain a general measure of the intensity of association within a table, Goodman (1996:7) proposed a generalized index of nonindependence. In its unweighted version, denoted $\lambda$, it is equal to the Euclidean norm of the $\lambda_{ij}$:

$$\lambda = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{ij}^2} \tag{5}$$

And in its weighted version, denoted $\tilde{\lambda}$, equal to the weighted standard deviation of the $\tilde{\lambda}_{ij}$[1]:

$$\tilde{\lambda} = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J} \tilde{\lambda}_{ij}^2 P_{i+}P_{+j}} \tag{6}$$

Instead of marginal weighting, the uniform weighting already mentioned above provides a third often-used version of the coefficient, denoted $\lambda^\dagger$, which is equal to the standard deviation of the $\lambda_{ij}$:

$$\lambda^\dagger = \sqrt{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{ij}^2} = \frac{\lambda}{\sqrt{IJ}} \tag{7}$$

[1] Since the weights sum to unity, the weighted version corresponds to a mean, and the unweighted version to a sum.

An interesting property of the various versions of the nonindependence index is that the square of the index, defined as a sum of contributions per cell, can also be broken down into contributions per row and column.

We first show that the weighted version of the index, denoted $\tilde{\lambda}$, is a generalization of Pearson's mean square contingency coefficient $\varphi^2$, and then explain at greater length how a version of this index derived from odds ratios (called the intrinsic association coefficient) can be devised within this framework[2].

## 1.2 Pearson's mean square contingency coefficient

If we define R as the identity function, that is $\mathrm{R}[x] = x$, then Equation (4) implies that

$$\tilde{\lambda}_{ij} = R_{ij} - 1 - 1 + 1 = \psi_{ij} - 1 = \frac{P_{ij} - P_{i+}P_{+j}}{P_{i+}P_{+j}} \tag{8}$$

In this case, we find that the  values, using Goodman's terminology, are the Pearson contingencies, or the square roots of the Pearson residuals. The general index of nonindependence defined in Equation (6) as $\tilde{\lambda}$ is thus equal to the square root of Pearson's mean square contingency coefficient $\varphi^2$:

$$\tilde{\lambda}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(P_{ij} - P_{i+}P_{+j})^2}{P_{i+}P_{+j}} = \varphi^2 = \frac{\chi^2}{N} \tag{9}$$

where N is the sum of counts in the table.

## 1.3 Intrinsic association coefficient

If we instead define R as the natural logarithm, that is $\mathrm{R}[x] = \log x$, we find:

$$\lambda_{ij} = \log P_{ij} - \frac{1}{J}\sum_{j=1}^{J} \log P_{ij} - \frac{1}{I}\sum_{i=1}^{I} \log P_{ij} + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \log P_{ij} \tag{10}$$

It can be shown that the $\lambda_{ij}$ are in that case equal to the interaction coefficients of the saturated log-linear model (justifying the notation chosen), which are more often presented (Bishop et al. [1975] 2007:2; Agresti 2002:5) in the form $\log P_{ij} = \lambda^0 + \lambda_i^I + \lambda_j^J + \lambda_{ij}$.

Therefore, the odds ratio contrasting rows $i$ and $i'$ and columns $j$ and $j'$ is equal to:

$$\theta_{ij,i'j'} = \exp\left[\lambda_{ij} + \lambda_{i'j'} - \lambda_{i'j} - \lambda_{ij'}\right] \tag{11}$$

Like the odds ratio, the $\lambda_{ij}$ possess the property of margin-insensitivity that is central to log-linear modeling: they are not affected by the rows and columns being multiplied by arbitrary values. Therefore, the index of intensity of association $\lambda$ defined in Equation (5) is also marginal-free.

We obtain a similar equation for the weighted interaction:

$$\tilde{\lambda}_{ij} = \log P_{ij} - \sum_{j=1}^{J} P_{+j} \log P_{ij} - \sum_{i=1}^{I} P_{i+} \log P_{ij} + \sum_{i=1}^{I}\sum_{j=1}^{J} P_{i+}P_{+j} \log P_{ij} \tag{12}$$

However, in that case, the $\tilde{\lambda}_{ij}$ coefficients are not equivalent to the log-linear interaction coefficients. These coefficients, and therefore the weighted index $\tilde{\lambda}$ defined in Equation (6), are marginal-free only if the weights themselves do not depend on the margins. A case in point is when uniform weights are used, as in Equation (7), for the version of the index denoted $\lambda^\dagger$. The relationship with the odds ratio contrasting rows $i$ and $i'$ and columns $j$ and $j'$ is still direct:

$$\theta_{ij,i'j'} = \exp\left[\tilde{\lambda}_{ij} + \tilde{\lambda}_{i'j'} - \tilde{\lambda}_{i'j} - \tilde{\lambda}_{ij'}\right] \tag{13}$$

It is also interesting to note that the marginal-weighted version of the index does not change if two rows (respectively, columns) with the same conditional distributions are combined (Goodman 1996:425). This property is also possessed by Pearson's mean square contingency coefficient $\varphi^2$.

We call the index derived in this section the intrinsic association coefficient[3], following the terminology introduced by Goodman (1981a, 1985, 1986, 1991) for association models. The link between the definition of the index presented above and log-linear and log-multiplicative models, justifying this terminology, will be developed below. Let us note, however, that an application of this index to logistic regression has previously been called $\kappa$ index (Hout et al. 1995; Breen et al. 2009).

## 1.4 Normalized intrinsic association coefficient

The intrinsic association coefficient is expressed on the scale of the logarithm of odds ratios: it equals zero when independence holds, and has no upper limit. Therefore, although it possesses the desired marginal-independence property and is thus a useful tool, it does not make it easy to assess the strength of the association, given that the log-odds ratio scale is not familiar to a wide audience. We propose normalizing the coefficient so that it follows the well-known scale from 0 to 1 used by the Pearson correlation coefficient, the mean square contingency coefficient $\varphi^2$ and Cramér's V.

This transformation of the intrinsic association coefficient from a scale from 0 to infinity to a scale from 0 to 1 is not just an artificial device to make the index appear more familiar. Indeed, it has been shown

[2] We shall not demonstrate here the links with Yule's Y and Q coefficients, less often used, which display properties similar to those of odds ratios.

[3] Despite the notation, this index has no relationship with Goodman and Kruskal's $\lambda$ coefficient.

that there exists a direct relationship between the marginal-weighted intrinsic association coefficient and the correlation coefficient in the special case when frequencies in a contingency table are distributed according to a discretized bivariate normal distribution (Goodman 1981b, 1985; Becker and Clogg 1988; Becker 1989). We can therefore define a normalized version of the intrinsic association coefficient with uniform weighting $\tau^\dagger$ as:

$$\tau^\dagger = \sqrt{1 + 1/(2\lambda^\dagger)^2} - 1/(2\lambda^\dagger) \quad \text{or equivalently} \quad \lambda^\dagger = \tau^\dagger/(1 - \tau^{\dagger 2}) \tag{14}$$

and with arbitrary weighting $\tilde{\tau}$ as:

$$\tilde{\tau} = \sqrt{1 + 1/(2\tilde{\lambda})^2} - 1/(2\tilde{\lambda}) \quad \text{or equivalently} \quad \tilde{\lambda} = \tilde{\tau}/(1 - \tilde{\tau}^2) \tag{15}$$

The normalized and non-normalized versions of the intrinsic association coefficient are very close for values below 0.3: a normalized intrinsic association coefficient of 0.3 corresponds to a (non-normalized) intrinsic association coefficient of 0.33. The difference then increases quickly beyond that limit: 0.5 corresponds to 0.67, 0.7 to 1.37, and 0.9 to 4.74. For the special case of the complete absence of association, we define in accordance with the limit that $\tau^\dagger = \lambda^\dagger = 0$ and $\tilde{\tau} = \tilde{\lambda} = 0$.

In practical use, the normalized index can be preferred when reporting results to ease interpretation. However, the non-normalized index is more appropriate in contexts where the absence of an upper bound is an advantage, notably when the strength of the association is used as the dependent variable in a linear model, as we will illustrate below.

# 2 A derivation of the intrinsic association coefficient from odds ratios

We have shown in the previous section that the intrinsic association coefficient could be considered as the odds ratio-based equivalent of the Pearson mean square contingency coefficient $\varphi^2$ or of Cramér's V. In this section, we would like to give a more straightforward interpretation of this index by showing that it can be derived in a quite direct way from all the odds ratios in a contingency table. This derivation will also show that the intrinsic association coefficient and the Altham index are actually very close quantities which can easily be translated from one another.

For the purposes of the demonstration, let us define the standard odds ratio (SOR) as the geometric standard deviation of all the odds ratios that can be calculated for a table. Just as the intrinsic association coefficient defined at Equations (6) and (7) equals the standard deviation of interaction coefficients $\lambda_{ij}$ or $\tilde{\lambda}_{ij}$, so the aim here is to measure the distance between the odds ratio and its reference value (in this case 1), but on a multiplicative rather than linear scale. Indeed, the geometric standard deviation is the equivalent on a multiplicative scale of the arithmetic standard deviation, i.e. the exponential of the standard deviation of the log-odds ratios.

One way of enumerating all the odds ratios that can be constructed from an $I \times J$ table is to take as reference each cell in the table in turn and calculate for that cell all the $IJ$ odds ratios involving that cell and all the cells in the table including those on the same row or column, for which the odds ratio is equal to unity. This amounts to constructing $(IJ)^2$ odds ratios, most of which are redundant, since the "basic set" (Goodman 1969; Rudas 1998) of $(I-1)(J-1)$ spanning cell odds ratios corresponding to any cell is sufficient to recalculate all the others. Furthermore, this series of odds ratios contains exactly 4 times each of the $(IJ)^2/4$ distinct square log-odds ratios that can be constructed from the table. These redundancies are not a problem, because the standard deviation is not affected by the repetition of all the values the same number of times.

## 2.1 Using uniform weighting

We can now define the SOR with uniform weighting[4], and then in the next section generalize it for arbitrary weights:

$$\begin{aligned} SOR &= \exp \sqrt{\frac{1}{(IJ)^2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{i'=1}^{I} \sum_{j'=1}^{J} (\log \theta_{ij,i'j'})^2} \\ &= \exp \sqrt{\frac{1}{(IJ)^2} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{i'=1}^{I} \sum_{j'=1}^{J} (\lambda_{ij} + \lambda_{i'j'} - \lambda_{i'j} - \lambda_{ij'})^2} \end{aligned} \tag{16}$$

It appears that the quadruple sum in Equation (16) is actually equal to the square of the Altham index measuring the distance of a two-way table from independence (Altham 1970; Altham and Ferrie 2007). We will return to this below.

The uniform-weighted SOR can also be expressed in terms of the intrinsic association coefficient. Indeed, since the $\lambda_{ij}$ have null sums in rows and columns, Equation (16) simplifies as follows:

[4] An unweighted version may also be defined, equivalent to the unweighted intrinsic association coefficient presented above. However, it is not a standard deviation strictly speaking.

$$
\begin{aligned}
&(IJ)^2 \log^2 SOR \\
&\quad = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} \left(\lambda_{ij} + \lambda_{i'j'} - \lambda_{i'j} - \lambda_{ij'}\right)^2 \\
&\quad = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} (\lambda_{ij}^2 + \lambda_{i'j'}^2 + \lambda_{i'j}^2 + \lambda_{ij'}^2) \\
&\qquad + 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} (\lambda_{ij}\lambda_{i'j'} + \lambda_{i'j}\lambda_{ij'} - \lambda_{ij}\lambda_{i'j} - \lambda_{ij}\lambda_{ij'} - \lambda_{i'j'}\lambda_{i'j} - \lambda_{i'j'}\lambda_{ij'}) \\
&\quad = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} (\lambda_{ij}^2 + \lambda_{i'j'}^2 + \lambda_{i'j}^2 + \lambda_{ij'}^2) \\
&\quad = 4IJ \sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{ij}^2
\end{aligned} \tag{17}
$$

So we find that the intrinsic association coefficient with uniform weighting, defined in Equation (7), is equal to half[5] the logarithm of the geometric standard deviation of all the table's odds ratios (here expressed as the standard deviation of the log-odds ratios):

$$
\lambda^\dagger = \sqrt{\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{ij}^2} = \frac{1}{2}\sqrt{\frac{1}{(IJ)^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} (\log \theta_{ij,i'j'})^2} \tag{18}
$$

This result also allows deriving the very direct relation between the Altham index $d(P, Q)$, with $P$ the analyzed table and $Q$ the table expected under independence, and the intrinsic association coefficient, in both its unweighted and uniform-weighted versions:

$$
d(P, Q) = 2\sqrt{IJ}\lambda = 2IJ\lambda^\dagger \tag{19}
$$

While being very close to the Altham index, the intrinsic association coefficient offers a significant advantage over its competitor: it is insensitive to the dimension of the table, i.e. using a larger number of categories does not mechanically increase the value of the index. In that regard, the intrinsic association coefficient has a similar relationship to the Altham index as Cramér's V to the mean square contingency coefficient $\varphi^2$. Apart from this, the two indices are equivalent: in particular, they will give the same conclusions when comparing tables of the same dimension. However, the intrinsic association coefficient is superior in that it allows comparing tables of different dimensions, as we will illustrate below.

## 2.2 Using arbitrary weighting

Following the approach used above, we now define the standard odds ratio with arbitrary weighting, which generalizes the results of the previous section. For simplicity's sake, as before, we present the specific case of marginal weighting but the demonstrations hold, unless otherwise indicated, for any set of strictly positive weights that sum to unity (on condition that the $\tilde{\lambda}$ interaction coefficients have been calculated with the same weights). The value of this approach is not so much in making it possible to use marginal weightings in a two-dimensional table – which would lose the property of margin insensitivity – but rather, as we shall see, to use average-marginal weighting in a three-dimensional table.

This second version of the SOR, denoted $\widetilde{SOR}$, is defined as the geometric weighted standard deviation of all the odds ratios of an $I \times J$ table:

$$
\begin{aligned}
\widetilde{SOR} &= \exp\sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} (\log \theta_{ij,i'j'})^2 P_{i+}P_{+j}P_{i'+}P_{+j'}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} P_{i+}P_{+j}P_{i'+}P_{+j'}}} \\
&= \exp\sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J} (\tilde{\lambda}_{ij} + \tilde{\lambda}_{i'j'} - \tilde{\lambda}_{i'j} - \tilde{\lambda}_{ij'})^2 P_{i+}P_{+j}P_{i'+}P_{+j'}}
\end{aligned} \tag{20}
$$

Similar to the previous section, we observe that the quadruple sum is a weighted generalization of the square of the Altham index. By replacing in Equation (20) all the $P_{i+}$ and $P_{+j}$ by unit weights, we obtain the standard (uniform-weighted) version of the index presented above.

Using the same procedure as for the uniform-weighted index in Equation (17), we can establish the link between the weighted geometric standard deviation of odds ratios and the weighted intrinsic association coefficient. Indeed, since the weighted row and column sums of the $\tilde{\lambda}_{ij}$ are zero, Equation (20) simplifies to:

[5] We can remark that the constant 2 corresponds to $\sqrt{4}$, and reflects the fact that each odds ratio is calculated from four $\lambda_{ij}$ interaction coefficients, i.e. from 4 cells.

$$
\begin{aligned}
&\log^2 \widetilde{SOR} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J}(\tilde{\lambda}_{ij}+\tilde{\lambda}_{i'j'}-\tilde{\lambda}_{i'j}-\tilde{\lambda}_{ij'})^2 P_{i+}P_{+j}P_{i'+}P_{+j'} \\
&= 4\sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{\lambda}_{ij}^2 P_{i+}P_{+j} \\
&+ 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J}(\tilde{\lambda}_{ij}\tilde{\lambda}_{i'j'}+\tilde{\lambda}_{i'j}\tilde{\lambda}_{ij'}+\tilde{\lambda}_{ij}\tilde{\lambda}_{i'j}+\tilde{\lambda}_{ij}\tilde{\lambda}_{ij'}+\tilde{\lambda}_{i'j'}\tilde{\lambda}_{i'j}+\tilde{\lambda}_{i'j'}\tilde{\lambda}_{ij'})P_{i+}P_{+j}P_{i'+}P_{+j'} \\
&= 4\tilde{\lambda}^2
\end{aligned}
\tag{21}
$$

Again, we find that the weighted intrinsic association coefficient, defined at Equation (6), is equal to half the logarithm of the weighted geometric standard deviation of all odds ratios that can be constructed from a table:

$$
\tilde{\lambda} = \sqrt{\sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{\lambda}_{ij}^2 P_{i+}P_{+j}} = \frac{1}{2}\sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J}(\log\theta_{ij,i'j'})^2 P_{i+}P_{+j}P_{i'+}P_{+j'}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{i'=1}^{I}\sum_{j'=1}^{J}P_{i+}P_{+j}P_{i'+}P_{+j'}}} \tag{22}
$$

As already indicated, we can see that by replacing, in Equations (21) and (22), $P_{i+}$ by $1/I$ and $P_{+j}$ by $1/J$, we return to the formula for the uniform-weighted intrinsic association coefficient.

# 3 Relation to the Unidiff model

Replacing the Altham index and the intrinsic association coefficient in a common framework is particularly useful as it allows unifying the descriptive approach of the Altham index and the parametric approach of log-linear and log-multiplicative modeling. One area where the similarity is striking is the analysis of variations in the overall strength of the association over the last dimension (layer) of a three-way table.

Indeed, the indices presented above are directly related to the association represented by the log-multiplicative layer effect model (Xie 1992), better known as the Unidiff model (Erikson and Goldthorpe 1992)[6]. The proportions expected under this model follow the equation, with $F_{ijk}$ the number predicted by the model for the cell at the intersection of row $i$, column $j$ and layer $k$:

$$
\log F_{ijk} = \lambda^0 + \lambda_i^I + \lambda_j^J + \lambda_k^K + \lambda_{ik}^{IK} + \lambda_{jk}^{JK} + \phi_k\psi_{ij} \tag{23}
$$

For a given layer $k$, if the model applies, the interaction coefficients between rows and columns can be written as $\lambda_{ijk} = \phi_k\psi_{ij}$, with $\psi_{ij}$ the interaction coefficients common to all layers and $\phi_k$ the layer coefficient measuring the intensity of the association on layer $k$. According to Equation (5), the unweighted intrinsic association coefficient for layer $k$ thus equals:

$$
\lambda_k^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(\phi_k\psi_{ij})^2 = \phi_k^2\sum_{i=1}^{I}\sum_{j=1}^{J}\psi_{ij}^2 \tag{24}
$$

It follows that the ratio between the intensities of the associations relating to layers $k$ and $k'$ equals the ratio between the respective intrinsic association coefficients of these layers:

$$
\frac{\lambda_{k'}}{\lambda_k} = \frac{\phi_{k'}}{\phi_k} \tag{25}
$$

Therefore, if we denote by index $k = 0$ the reference layer for which the layer effect coefficient is fixed by convention at unity, that is $\phi_0 = 1$, we obtain:

$$
\lambda_k = \phi_k\lambda_0 \tag{26}
$$

The same properties are verified for the Altham index (Zhou 2015), because of its direct relation with the intrinsic association coefficient evidenced in Equation (19) above.

It is easy to verify with the same procedure that this property is verified when arbitrary weighting is used, as long as the weights are independent of the layer under consideration. This holds in particular for the weighting by margins of the whole table (average-marginal weighting, see Becker and Clogg 1989), which is an interesting alternative to uniform weighting when one seeks to examine the variations between layers in the intensity of the association independent of the table margins. Let us note that, extending a result highlighted in the first section regarding two-dimensional tables, the values of the index computed with average-marginal weights do not change when combining rows (respectively, columns) with identical conditional distributions. This makes this weighting system particularly appealing.

Using intrinsic association coefficients or Altham indices corresponding to layers therefore allows comparing them in the same way as by using the layer effect coefficients $\phi_k$, and if necessary recomputing the latter

[6] This relation with the intrinsic association coefficient has already been partly presented by Leo Goodman and Michael Hout (2001), but with respect to models that, while more general in some ways, are in other ways much more restrictive than Unidiff.

coefficients. These indices have the further advantage that they provide a measure of the absolute extent of the association, whereas layer effect coefficients can only be interpreted with respect to the reference layer. This property is particularly useful to compare results across studies, or to compare the intensity of the association between a variable and a series of other variables.

Finally, when the Unidiff model does not accurately fit the data, these three indices can be used to measure the intensity of the association relating to the various layers without assuming that the structure of this association is homogeneous between layers. In this sense, they are generalizations of the measure provided by the layer effect coefficient of the Unidiff model. This approach can be carried out either by calculating the value of these indices directly from the observed data, or by combining them with models more complex than Unidiff, such as the regression-type model (Goodman and Hout 1998) or the row-column association model with layer effect RC(M)-L that we describe in the next section.

# 4 Relation to row-column association models

## 4.1 RC(M) association model

The intrinsic association coefficient was devised by Goodman for association models (Goodman 1981a, 1985, 1986; Becker and Clogg 1989; Clogg and Shihadeh 1994; Wong 2010): it is thus directly related to these models, and the Altham index inherits this close relation. With the log-multiplicative row-column association model (also known as RC(M) or Goodman's RC type II model) the expected proportions follow the equation, with $F_{ij}$ the number predicted by the model for the cell at the intersection of row $i$ and column $j$:

$$\log F_{ij} = \lambda^0 + \lambda_i^I + \lambda_j^J + \sum_{m=1}^{M} \phi_m \mu_{im} \nu_{jm} \tag{27}$$

In this equation, $\phi_m$ is the intrinsic association coefficient for dimension $m$, and $\mu_{im}$ and $\nu_{jm}$ are the scores on dimension $m$ for row $i$ and column $j$. By convention, without loss of generality, $\phi_m$ is always chosen to be positive (incurring if necessary a change of sign for the $\mu_{im}$ or the $\nu_{jm}$ scores). These coefficients are made identifiable using the following constraints (of position, scale and orthogonality across dimensions):

$$\begin{aligned}
&\sum_{i=1}^{I} \mu_{im} = \sum_{j=1}^{J} \nu_{jm} = 0, \\
&\sum_{i=1}^{I} \mu_{im}^2 = \sum_{j=1}^{J} \nu_{jm}^2 = 1, \\
&\sum_{i=1}^{I} \mu_{im} \mu_{im'} = \sum_{j=1}^{J} \nu_{jm} \nu_{jm'} = 0 \text{ for all } m \neq m'
\end{aligned} \tag{28}$$

In the weighted version, the equation of the model is:

$$\log F_{ij} = \tilde{\lambda}^0 + \tilde{\lambda}_i^I + \tilde{\lambda}_j^J + \sum_{m=1}^{M} \tilde{\phi}_m \tilde{\mu}_{im} \tilde{\nu}_{jm} \tag{29}$$

$\tilde{\phi}_m$, $\mu_{im}$ and $\tilde{\nu}_{jm}$ are defined similarly, but with the following weighted identification constraints:

$$\begin{aligned}
&\sum_{i=1}^{I} \tilde{\mu}_{im} P_{i+} = \sum_{j=1}^{J} \tilde{\nu}_{jm} P_{+j} = 0, \\
&\sum_{i=1}^{I} \tilde{\mu}_{im}^2 P_{i+} = \sum_{j=1}^{J} \tilde{\nu}_{jm}^2 P_{+j} = 1, \\
&\sum_{i=1}^{I} \tilde{\mu}_{im} \tilde{\mu}_{im'} P_{i+} = \sum_{j=1}^{J} \tilde{\nu}_{jm} \tilde{\nu}_{jm'} P_{+j} = 0 \text{ for all } m \neq m'
\end{aligned} \tag{30}$$

In an association model, the significance of a dimension is measured by the corresponding intrinsic association coefficient, generally denoted $\phi_m$ for dimension $m$. This coefficient is the direct equivalent of the coefficient of the same name denoted $\lambda$ above, but calculated from the component of total interaction between rows and columns which can be attributed to the dimension under consideration. From Equation (27) it can be seen that with the RC(M) model, the row-column interaction coefficient equals:

$$\lambda_{ij} = \sum_{m=1}^{M} \phi_m \mu_{im} \nu_{jm} \tag{31}$$

So the contribution of each dimension to the interaction is $\phi_m \mu_{im} \nu_{jm}$ (a value that may be either negative or positive for a given cell and dimension). Taking Equation (5) defining the overall intrinsic association coefficient $\lambda$, but replacing the term $\lambda_{ij}$ by $\phi_m \mu_{im} \nu_{jm}$ so as only to account for the contribution from dimension $m$, we obtain, in line with Equations (27) and (28):

$$\lambda_m^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (\phi_m \mu_{im} \nu_{jm})^2 = \phi_m^2 \sum_{i=1}^{I} \mu_{im}^2 \sum_{j=1}^{J} \nu_{jm}^2 = \phi_m^2 \tag{32}$$

And in the weighted version, using Equations (6), (29) and (30) this time:

$$\tilde{\lambda}_m^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(\tilde{\phi}_m\tilde{\mu}_{im}\tilde{\nu}_{jm})^2 P_{i+}P_{+j} = \tilde{\phi}_m^2 \sum_{i=1}^{I}\tilde{\mu}_{im}^2 P_{i+} \sum_{j=1}^{J}\tilde{\nu}_{jm}^2 P_{+j} = \tilde{\phi}_m^2 \tag{33}$$

It can be seen that the intrinsic association coefficient $\phi_m$ (respectively $\tilde{\phi}_m$) is indeed the equivalent, applied to a particular dimension, of the coefficient $\lambda$ (respectively $\tilde{\lambda}$) of the same name defined in the first section regarding total association. This relation goes beyond a mere analogy: the overall intrinsic association coefficient equals the Euclidean norm of the coefficients corresponding to each dimension. This can be seen from Equations (5) and (28):

$$\begin{aligned}
\lambda^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{m=1}^{M}\phi_m\mu_{im}\nu_{jm}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M}(\phi_m\mu_{im}\nu_{jm})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{m'\neq m}\phi_m\phi_{m'}\mu_{im}\mu_{im'}\nu_{jm}\nu_{jm'} \\
&= \sum_{m=1}^{M}\lambda_m^2 + \sum_{m=1}^{M}\sum_{m'\neq m}\phi_m\phi_{m'}\sum_{i=1}^{I}\mu_{im}\mu_{im'}\sum_{j=1}^{J}\nu_{jm}\nu_{jm'} \\
&= \sum_{m=1}^{M}\lambda_m^2
\end{aligned} \tag{34}$$

Similarly, in the weighted version, by (6) and (30):

$$\begin{aligned}
\tilde{\lambda}^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\sum_{m=1}^{M}\tilde{\phi}_m\tilde{\mu}_{im}\tilde{\nu}_{jm}\right)^2 P_{i+}P_{+j} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M}(\tilde{\phi}_m\tilde{\mu}_{im}\tilde{\nu}_{jm})^2 P_{i+}P_{+j} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{m'\neq m}\tilde{\phi}_m\tilde{\phi}_{m'}\tilde{\mu}_{im}\tilde{\mu}_{im'}\tilde{\nu}_{jm}\tilde{\nu}_{jm'}P_{i+}P_{+j} \\
&= \sum_{m=1}^{M}\tilde{\lambda}_m^2 + \sum_{m=1}^{M}\sum_{m'\neq m}\tilde{\phi}_m\tilde{\phi}_{m'}\sum_{i=1}^{I}\tilde{\mu}_{im}\tilde{\mu}_{im'}P_{i+}\sum_{j=1}^{J}\tilde{\nu}_{jm}\tilde{\nu}_{jm'}P_{+j} \\
&= \sum_{m=1}^{M}\tilde{\lambda}_m^2
\end{aligned} \tag{35}$$

So the overall intrinsic association coefficient (weighted or otherwise) equals the Euclidean norm of the intrinsic association coefficients corresponding to each dimension of the model. An association model is thus a way of decomposing the total association in the table into a series of dimensions of diminishing significance. This decomposition is valid whether the model is saturated or not, as long as it fits the data properly. Association models therefore stand in the same relation to the odds ratio tradition as correlation or correspondence analysis do to the Pearson tradition (Goodman 1985, 1986, 1991, 1996; Gilula and Haberman 1986). Once again, the Altham index is also tightly linked to this approach, though this relation is less direct than for the intrinsic association coefficient.

## 4.2 Extension to RC(M)-L association model

RC(M)-L models (Clogg 1982; Wong 2010) are an extension of RC(M) models to three-dimensional tables: the intrinsic association coefficient and/or scores can vary from one layer to another. One version of this model postulates that the association is identical for all layers (homogeneous scores and intrinsic association coefficients); another, that it differs entirely between layers (heterogeneous scores and intrinsic association coefficients): these two versions can be reduced either in the first case to the scores of a single RC(M) model (but with layer-specific marginal parameters) or to those of as many RC(M) models are there are layers.

Only the third version of the RC(M)-L model requires an extension of the approach presented so far. This version of the model assumes that the scores are homogeneous between layers but that the intrinsic association coefficients are heterogeneous. Its equation, with $F_{ijk}$ the frequency predicted by the model in the cell at the intersection of row $i$, column $j$ and layer $k$, is as follows:

$$\log F_{ijk} = \lambda^0 + \lambda_i^I + \lambda_j^J + \lambda_k^K + \lambda_{ik}^{IK} + \lambda_{jk}^{JK} + \sum_{m=1}^{M}\phi_{mk}\mu_{im}\nu_{jm} \tag{36}$$

or, in its weighted version,

$$\log F_{ijk} = \lambda^0 + \tilde{\lambda}_i^I + \tilde{\lambda}_j^J + \tilde{\lambda}_k^K + \tilde{\lambda}_{ik}^{IK} + \tilde{\lambda}_{jk}^{JK} + \sum_{m=1}^{M}\tilde{\phi}_{mk}\tilde{\mu}_{im}\tilde{\nu}_{jm} \tag{37}$$

Only the first two constraints of Equations (28) and (30), applying to the scores, are required: cross-dimensional constraints can no longer be applied and there is generally a non-zero correlation between the scores in different dimensions. Consequently, the reasoning followed in Equations (34) and (35) cannot

be used. The relation between the intrinsic association coefficients for each dimension and the overall intrinsic association coefficient is not a simple summation: it must take into account the correlation between dimensions. In Equations (34) and (35), the respective terms:

$$\sum_{m=1}^{M} \sum_{m' \neq m} \phi_m \phi_{m'} \sum_{i=1}^{I} \mu_{im} \mu_{im'} \sum_{j=1}^{J} \nu_{jm} \nu_{jm'} \tag{38}$$

and

$$\sum_{m=1}^{M} \sum_{m' \neq m} \tilde{\phi}_m \tilde{\phi}_{m'} \sum_{i=1}^{I} \tilde{\mu}_{im} \tilde{\mu}_{im'} P_{i+} \sum_{j=1}^{J} \tilde{\nu}_{jm} \tilde{\nu}_{jm'} P_{+j} \tag{39}$$

corresponding to the sum of the products of the intrinsic association coefficients and the correlations between (respectively) the row and column scores of the dimensions taken two at a time do not generally equal zero. For example, the intensity of association on a given layer depends on the positive, negative or zero correlation between the scores and the intensities of the different dimensions. Intuitively, one dimension may offset the association represented by another if it is strong enough and the two dimensions have sufficiently different scores. Note too that the intrinsic association coefficients here may be negative, which amounts in practice to inverting the sign of the row or column scores and thus inverting the direction of the link compared to layers where the coefficient was positive.

Despite this greater complexity, which is due to the richness of the RC(M)-L model, both the intrinsic association coefficient and the Altham index can always be calculated separately, for the overall association and for each dimension. The analysis of the correlation between dimensions can also be of interest to better understand the variations of the overall association.

In conclusion, note that, as with the Unidiff model, analysis of the differences in association independently of marginal variations between layers can be achieved by adopting either uniform weighting or weighting by the average row and column margins of the table (rather than by the margins of each layer).

# 5 Application: Educational and Socioeconomic Homogamy Among European Regions

This section illustrates the interest of the association indices presented above to analyze the spatial variations of educational and socioeconomic homogamy among European regions (see Bouchet-Valat 2018a for a more complete analysis). Like intergenerational social mobility, homogamy has typically been studied in the literature using marginal-free methods such as log-linear models and other odds-ratio-based techniques (for international comparisons see Smits, Ultee, and Lammers 1998, 1999; Raymo and Xie 2000; Smits, Ultee, and Lammers 2000; Smits 2003; Park and Smits 2005; Katrňák, Martin Kreidl, and Fónadová 2006; Domański and Przybysz 2007; Katrňák, Fučík, and Luijkx 2012). Multiple families of log-linear and log-multiplicative models have been used by different authors, so that no straightforward comparison of the results is possible. The association indices presented in this article would allow summarizing the model results in a single figure given the overall strength of homogamy in each studied society, despite the variety of the chosen modeling strategies.

The example presented here will also highlight a risk which researchers may run when trying to use models to obtain a single measure of the strength of the association. Often, only relatively restrictive models will provide such a summary parameter: the log-multiplicative layer effect (Unidiff) estimates a layer coefficient; the log-multiplicative row-column association model (RC-L) estimates an intrinsic association coefficient; the distance log-linear model estimates a step parameter. When these simple models do not fit the data adequately, more complex models may be more appropriate, like the regression-type log-multiplicative model (Goodman and Hout 1998), or multidimensional association models (like RC(M)-L models). Even more frequently, cell-specific parameters will have to be introduced for the main diagonal of the homogamy table in order to account for the varying intensity of homogamy between groups; these parameters may also be country-specific. In these cases, no single measure of the strength of homogamy in a given country can be obtained. Researchers may then be tempted either to analyze the determinants of one of the components of the association, and ignore the others (as did Domański and Przybysz 2007 by regressing step parameters and leaving aside diagonal parameters); or to use simpler models which may not give a completely accurate description of the data. In what follows, we illustrate this risk using the Unidiff model.

The illustration is based on the analysis of educational and socioeconomic homogamy tables for 149 infra-national regions of the European Union (NUTS1 and NUTS2 levels, regrouping between 800,000 an 7 million people) for years 2014 to 2016. These tables have been computed from the corresponding waves of the European Union Labour Force Survey, covering 26 European Union member States[7]: Austria (AT), Belgium (BE), Bulgaria (BG), Croatia (HR), Czech Republic (CH), Cyprus (CY), Estonia (EE), France (FR), Germany (DE), Greece (GR), Hungary (HU), Ireland (IE), Italy (IT), Latvia (LV), Lithuania (LT), Luxembourg (LU), the Netherlands (NL), Norway (NO), Poland (PO), Portugal (PT), Romania (RO), Slovakia (SK), Slovenia (SI), Spain (ES), Sweden (SE) and the United Kingdom (UK). Cohabiting couples (both married and unmarried) have been identified within each household using partner identifiers. To ensure the reliability of the information on occupations and stability of the rate of individuals in a relationship, only couples in which both partners are aged 30 to 59 years are considered. The sample is made of 1,400,000

[7] Norway is included as an associated country. On the other hand, the data for Denmark, Finland and Malta could not be used.

couples for educational homogamy and of 1,100,000 couples for socioeconomic homogamy (with regional samples ranging from 1,000 to 55,000)[8].

The educational levels of the partners is measured in the International Standard Classification of Education (ISCED) 2011, in four categories: lower secondary or less (ISCED 0-2, including short vocational education); upper secondary (ISCED 3); lower tertiary (ISCED 4-6: up to and including Bachelor's); upper tertiary (ISCED 7-8: Master's and beyond). The socioeconomic groups of the partners are measured using the European Socio-economic Groups classification (ESeG, see Meron and Amar 2014) in seven categories: Managers; Professionals; Technicians and associated professionals; Small entrepreneurs; Clerks and skilled service employees; Industrial skilled employees; Less skilled employees. For individuals not employed at the time of the survey, we use information on the last occupation (available only for those who worked within eight years before the survey).

One major question in the comparative literature on homogamy concerns its relationship with the level of economic development. Studies have sought to test the empirical validity of the inverted U curve hypothesis (Smits et al. 1998), according to which educational homogamy would increase in the first stage of development, but then decrease in a later stage. A variation of this hypothesis posits that a stabilization will be observed at the highest levels of development (saturation hypothesis). We will only deal with the part of the curve concerning advanced economies, to which European Union countries belong. To this end, two independent variables will be used. First, the average disposable income per inhabitant (in purchasing power parity) as computed by Eurostat for NUTS1 and NUTS2 regions in 2006 is used to measure economic development at the regional level. Second, we classify regions according to whether they contain the capital city of their countries, or a large metropolis[9]. This second variable will allow us to distinguish the role of economic development per se and that of the peculiarities of very dense regions which are generally richer, but also present higher inequality levels and are large enough so that inter-group contacts may not be as developed as in less populated regions (meeting opportunity effect).

## 5.1 Measuring Homogamy

Multiple approaches can be used to measure the strength of relative homogamy (i.e. controlling for the population structure in each region). We may opt for a fully non parametric approach by computing the intrinsic association coefficient (or equivalently the Altham index) directly on the observed data. We may also retain a semi-parametric estimator of the intrinsic association coefficient, like the Bayesian shrinkage method proposed by Zhou (2015) for the Altham index, whose principle is to estimate log-odds ratios for a given region more accurately by "borrowing strength" from the tables for other regions. Finally, we may also fit several models to the data and choose the one which provides the most accurate description according to classical criteria; the indices can then be computed on the fitted tables. We illustrate all three approaches in order to compare their results below. In all cases, we use average-marginal weighting.

As usual, we start with the conditional independence model, which only controls for the marginal distribution of men and women among the seven socioeconomic groups or the four educational categories (respectively) in each region, but does not allow for any tendency to relative homogamy. The equation of this model is:

$$\log F_{ijk} = \lambda^0 + \lambda^I_i + \lambda^J_j + \lambda^K_k + \lambda^{IK}_{ik} + \lambda^{JK}_{jk} \tag{40}$$

Fit statistics confirm that this model does not describe accurately the data (Table 1), with respectively 23% and 21% of misclassified couples (dissimilarity index) for education and socioeconomic group. The second model, called stability model, extends the first one by allowing for a common association to all regions once margins have been controlled. Its equation is:

$$\log F_{ijk} = \lambda^0 + \lambda^I_i + \lambda^J_j + \lambda^K_k + \lambda^{IK}_{ik} + \lambda^{JK}_{jk} + \lambda^{IJ}_{ij} \tag{41}$$

This model improves the fit significantly, with only 5.3% and 6.2% of misclassified couples and a clear reduction in both the BIC and AIC.

To measure geographic variations in the strength of relative homogamy, we have to find models which fit the data better than the stability baseline. Since the socioeconomic groups cannot be unequivocally ordered, the log-multiplicative layer effect model or Unidiff (Erikson and Goldthorpe 1992; Xie 1992) is a natural choice[10]. As mentioned above, this model is a good candidate for our purpose since it provides a single coefficient for each region, measuring the intensity of relative homogamy assuming that the structure of the row-column interaction is the same in all regions. This model follows the equation:

$$\log F_{ijk} = \lambda^0 + \lambda^I_i + \lambda^J_j + \lambda^K_k + \lambda^{IK}_{ik} + \lambda^{JK}_{jk} + \phi_k \psi_{ij} \tag{42}$$

The Unidiff model reveals significant variations of relative homogamy between regions, as both the AIC and the BIC decrease very clearly. However, the improvement to the description of the data is modest: the proportion of misclassified couples goes down by only two percentage points for education, and by less than one percentage point for socioeconomic group.

Does the assumption that the pattern of the association is the same in all regions on which rests the Unidiff model hold? Clearly not, as a fourth model including one additional parameter for each cell on the

[8]Survey weights are used in all analyses. For log-linear models, weighted tables are normalized for each region to sum to the actual sample size (Skinner and Vallet 2010). For linear regression models, regions are weighted using the population size in the 30-59 age range. This allows taking into account the large variations in population size across regions (Ebbinghaus 2005:136), which frequently result from arbitrary data availability issues.

[9]The list of metropolitan regions corresponds to Dijkstra's (2009) "second-tier metropolitan areas". See Bouchet-Valat (2018a) for details.

[10]For ordered variables like education, the distance model would also be a common choice (Smits, Ultee, and Lammers 1998; Domański and Przybysz 2007), but we do not explore it here for simplicity.

Table 1: Fit statistics for log-linear and log-multiplicative models

| | D. F. | Deviance | Δ (%) | BIC | AIC |
|---|---|---|---|---|---|
| **Education** | | | | | |
| Independence | 1,341 | 553,924 | 23.2 | 534,930 | 551,242 |
| Stability | 1,332 | 42,983 | 5.3 | 24,116 | 40,319 |
| Unidiff | 1,184 | 19,221 | 3.2 | 2,451 | 16,853 |
| Unidiff + diagonal parameters | 592 | 3,012 | 0.8 | **-5,373** | **1,828** |
| **Socioeconomic group** | | | | | |
| Independence | 5,364 | 342,606 | 21.1 | 267,857 | 331,878 |
| Stability | 5,328 | 38,566 | 6.2 | -35,681 | 27,910 |
| Unidiff | 5,180 | 31,228 | 5.6 | **-40,957** | 20,868 |
| Unidiff + diagonal parameters | 4,144 | 20,390 | 3.8 | -37,358 | **12,102** |

D. F.: Degrees of Freedom. Δ: Dissimilarity Index.

Table 2: Correlations Between the Indices Estimated via Four Different Methods

| | Non-parametric | Shrinkage | Unidiff | Unidiff + diagonal |
|---|---|---|---|---|
| **Education** | | | | |
| Non-parametric | 1.00 | 0.97 | 0.82 | 0.97 |
| Shrinkage | 0.97 | 1.00 | 0.89 | 0.97 |
| Unidiff | 0.82 | 0.89 | 1.00 | 0.85 |
| Unidiff + diagonal | 0.97 | 0.97 | 0.85 | 1.00 |
| **Socioeconomic group** | | | | |
| Non-parametric | 1.00 | 0.97 | 0.87 | 0.93 |
| Shrinkage | 0.97 | 1.00 | 0.92 | 0.95 |
| Unidiff | 0.87 | 0.92 | 1.00 | 0.91 |
| Unidiff + diagonal | 0.93 | 0.95 | 0.91 | 1.00 |

Correlations are weighted using the population size of each region.

main diagonal of the table for each region allows reducing the share of misclassified couples by about two percentage points, and improves the fit both according to the BIC and the AIC for education, and according to the AIC for socioeconomic homogamy. The equation of this model is:

$$\log F_{ijk} = \lambda^0 + \lambda_i^I + \lambda_j^J + \lambda_k^K + \lambda_{ik}^{IK} + \lambda_{jk}^{JK} + \phi_k \psi_{ij} + \delta_{ik} \mathbb{1}_{i=j} \tag{43}$$

Even this more complex model fails to fit the data according to the AIC (as the strongly positive value shows that the saturated model should be preferred), which indicates that more statistically significant deviations remain. We will not try to find better models here since our goal is precisely to evaluate how close are relatively classic models to the non-parametric and semi-parametric estimators.

The association indices presented in this article make the comparison of the estimates of the strength of the association provided by the above models very straightforward. Indeed, despite the different specifications, we can simply compute the values of the indices based on the fitted counts of the models (i.e. the $F_{ijk}$). We illustrate this using the intrinsic association coefficient, with weights equal to the average marginal counts for each row and column across all regions. For the standard Unidiff, this is strictly equivalent to applying E quation (24) above. Since our sample size is sufficiently large for the number of cells in each region, we can also compute the non-parametric estimator of the index[11], and compare it with the semi-parametric Bayesian shrinkage estimator (Zhou 2015) and the two Unidiff models. The model fitting and the computation of the two estimators of the intrinsic association coefficient can be achieved easily using the R package logmult (Bouchet-Valat 2018b)[12].

The comparison shows that the results obtained using the different methods are very similar overall (Table 2). For both types of homogamy, the weakest correlation across regions is observed between the non-parametric estimator and the Unidiff model: 0.82 for education and 0.87 for socioeconomic group. This is expected since these are respectively the least and the most restrictive estimation methods. This correlation is already quite high, indicating that using the Unidiff model as an approximation would globally yield correct results. The strongest correlations are observed between the non-parametric and the shrinkage estimators, at 0.97 for both education and socioeconomic group. The Unidiff model with diagonal parameters also agrees very closely with the non-parametric and shrinkage estimators (from 0.93 to 0.97).

[11]Since 26 cells (out of 7,301, i.e. 0.4%) for socioeconomic group and 36 (out of 2,384, i.e. 1.5%) for education contain zero counts, we added 1/2 to all cells for the computation of the non-parametric and semi-parametric estimators (Agresti 2002:9.8.7). The same operation was applied for education to two Slovakian regions (SK01 and SK03) to stabilize the Unidiff with diagonal parameters model due to the presence of one and sometimes two empty cells in the corners of the tables.

[12]Tables and code to reproduce the analyses are available in the data supplement, on the author's personal webpage at http://bouchet-valat.site.ined.fr and upon request.

Table 3: Average, Dispersion and Range of the Indices Estimated via Four Different Methods

| | Non-parametric | Shrinkage | Unidiff | Unidiff + diagonal |
|---|---|---|---|---|
| **Education** | | | | |
| Mean | 0.72 | 0.70 | 0.66 | 0.72 |
| Standard deviation | 0.21 | 0.16 | 0.16 | 0.19 |
| Minimum | 0.43 | 0.46 | 0.40 | 0.43 |
| Maximum | 1.50 | 1.17 | 1.24 | 1.46 |
| **Socioeconomic group** | | | | |
| Mean | 0.57 | 0.55 | 0.51 | 0.53 |
| Standard deviation | 0.09 | 0.07 | 0.09 | 0.08 |
| Minimum | 0.37 | 0.39 | 0.33 | 0.32 |
| Maximum | 0.86 | 0.77 | 0.87 | 0.78 |

Averages are weighted using the population size of each region.

For both types of homogamy, the average of the coefficients across all regions (Table 3) is the highest with the non-parametric approach (0.72 for education and 0.57 for socioeconomic group), and the lowest with the standard Unidiff (respectively 0.66 and 0.51). This is consistent with the correlation between these two methods being the lowest. The Bayes shrinkage estimator is mid-way between the two extremes in both cases, and the Unidiff with diagonal parameters is closer to one or the other indicator depending on the type of homogamy considered.

One advantage of the intrinsic association coefficient is that the number of categories used to measure each type of homogamy does not mechanically affect the level of the index, allowing for (cautious) comparisons between different types of homogamy. Here, educational homogamy consistently appears as stronger than socioeconomic homogamy, by 25% to 35% depending on the estimator[13]. This is also the case considering each region separately, in 120 to 129 regions out of 149.

The standard deviation of association across regions and the difference between maximum and minimum associations are the lowest for the shrinkage estimator, which is expected due to the definition of the estimator, which brings log-odds ratios closer to their European average. They are quite higher for the non-parametric estimator, which is again expected. It is more surprising to remark that the standard Unidiff model estimator has the largest standard deviation and range for socioeconomic group. Contrary to the Bayesian shrinkage estimator, it appears it is not always the case that the Unidiff model provides conservative estimates of deviations from the average.

The normalized variant of the index varying between 0 and 1 presented at Equation (15) can be used to ease the interpretation of these results. For the non-parametric approach, the average association of 0.72 for education gives a normalized coefficient of 0.52; the average association of 0.57 for socioeconomic group gives a normalized coefficient of 0.45. Regional normalized coefficients range from 0.37 to 0.72 for education, and from 0.33 to 0.58 for socioeconomic group. According to standard effect strength conventions for the Pearson correlation and contingency coefficients, these associations would range from medium to very large (Cohen 1988, ch. 7).

It is interesting to note that even though differences between estimation methods are limited, the standard Unidiff model, which is the most common method method used in the literature to estimate overall the level of association, tends to slightly underestimate the mean association, even if maximum values are in some cases higher than with other methods. We can therefore conclude that while Unidiff remains a useful tool, non- and semi-parametric estimators or more complex models should be preferred.

## 5.2 Accounting for Variations in Relative Socioeconomic Homogamy

In order to illustrate the fact that the relatively limited differences between estimators of the association observed above can lead to substantively different conclusions, we now turn to the analysis of the macro determinants of the intensity of relative socioeconomic homogamy. For simplicity we will not cover the case of educational homogamy, since for this dimension differences between estimators are less marked. The full analysis is available in a separate article (Bouchet-Valat 2018a).

We take the logarithm of the intrinsic association coefficient obtained by the four estimation methods as the dependent variable in an ordinary least squares regression model. This is appropriate since the index cannot take negative values. Variables therefore have a multiplicative effect on the association level. As developed above, the model includes as independent variables the disposable income per inhabitant and its square (variable was standardized so that its mean is zero and its standard deviation is one) and whether the region includes a capital city or a second-tier metropolitan area. Finally, we introduce country fixed effects (i.e. one dummy variable for each country) so that the coefficient estimates reflect the deviation in the strength of relative homogamy with reference to the country average.

The comparison of the $R^2$ for the four models (Table 4) shows that proportions of explained variance are similar, from 0.82 to 0.88. This very high figure is due to the inclusion of country fixed effects. The within-country $R^2$ (that is, the share of the variance not explained by country fixed effects which is explained by the full model) varies more significantly, from 0.15 for the non-parametric estimator down to 0.10–0.11 for the other three estimators.

[13]This difference persists when using a more aggregated socioeconomic classification.

Table 4: Linear Regression Results for Relative Socioeconomic Homogamy

| | **Non-parametric** | **Shrinkage** | **Unidiff** | **Unidiff + diagonal** |
|---|---|---|---|---|
| Disposable income* | 0.94 | 0.97 | 0.98 | 0.96 |
| | (0.89–0.99) | (0.93–1.00) | (0.94–1.03) | (0.92–0.99) |
| Disposable income$^{2*}$ | 1.05 | 1.04 | 1.04 | 1.05 |
| | (1.02–1.09) | (1.02–1.05) | (1.01–1.07) | (1.02–1.09) |
| Other Region (Ref.) | 1.00 | 1.00 | 1.00 | 1.00 |
| Capital Region | 1.10 | 1.07 | 1.08 | 1.13 |
| | (1.04–1.16) | (1.03–1.11) | (1.03–1.13) | (1.08–1.18) |
| Second-Tier Metro Region | 1.04 | 1.04 | 1.04 | 1.05 |
| | (1.01–1.07) | (1.02–1.06) | (1.01–1.07) | (1.03–1.08) |
| Intercept | 0.53 | 0.53 | 0.49 | 0.50 |
| | (0.50–0.56) | (0.51–0.55) | (0.47–0.51) | (0.47–0.52) |
| Country fixed effects | x | x | x | x |
| Observations (regions) | 146 | 146 | 146 | 146 |
| $R^2$ | 0.85 | 0.90 | 0.89 | 0.90 |
| Adjusted $R^2$ | 0.82 | 0.87 | 0.87 | 0.88 |
| Within-country $R^2$ | 0.15 | 0.10 | 0.11 | 0.10 |

95% normal bootstrap confidence intervals in parentheses.
* Standardized variable (zero mean and unit standard deviation).
The model is fitted on 146 regions out of 149 due to the non-availability of independent variables.

This result is consistent with the fact that estimated coefficients are generally farther away from 1 (indicating no effect) for the non-parametric estimator. The effect of level of metropolization varies in a non-negligible way across estimating methods. Capital regions are characterized by a higher socioeconomic homogamy than regions with no metropolis by 7% according to the shrinkage and standard Unidiff estimators, by 10% according to the non-parametric estimator, and by 13% according to the Unidiff with diagonal parameters estimator. This effect is statistically significant at the 5% level for all estimators. Regions with a second-tier metropolitan area also show a higher homogamy by 4% to 5%.

Differences between estimators are even more visible regarding the effect of the level of development. Using the non-parametric estimator, the coefficients for disposable income (0.94, significant at the 5% level) and its square (1.05, also significant) imply that moving from a disposable income two standard deviations below the average to a value equal to the average decreases relative homogamy by 27%, and that moving from the average to two standard deviations above average slightly increases relative homogamy (by 7%)[14]. Socioeconomic homogamy therefore tends to decline with economic development, but stabilizes above the average level of development. A similar, though weaker effect (and borderline significant at the 5% level) is observed using the shrinkage (respectively -20% and +10%) and Unidiff with diagonal parameter (-24% and +12%) estimators. On the contrary, when the association is measured using the standard Unidiff model, the effect of disposable income is so small (0.98) that it is no longer statistically significant at the 10% level[15]. Only its square has a significant positive effect, implying that socioeconomic homogamy is the highest for both the least and the most developed regions within a country, but that no decreasing trend is detected.

These results illustrate the impact small inaccuracies in the model-based measurement of the association can have on the results of subsequent analyses. Even if the overall correlation between the association measures obtained using the three different methods are quite high (over 0.8), the assumption of a common pattern of the association across all European regions does not really hold, which becomes more visible when considering within-country differences. This problem is likely to be more severe in cases where the standard Unidiff model does not fit the data as accurately as it does in the present study (dissimilarity index of 5.6%). Using the standard Unidiff model, we would have (incorrectly) concluded that a U-shaped relationship exists between level of development and relative socioeconomic homogamy, while the comparison with other estimators of the association shows that homogamy actually stabilizes rather than increases at higher levels of development. This result is consistent with that obtained for educational homogamy with all four estimating methods (Bouchet-Valat 2018a).

We have shown how a general-purpose marginal-free index of the association like the intrinsic association coefficient could be used to compare the results obtained using various model specifications, as well as using a semi- or non-parametric approach. This can be particularly useful to carry out a sensitivity analysis testing multiple modeling assumptions. Depending on the researcher's needs, this index can be used as a way of summarizing the strength of the association described by a chosen model whose coefficients are commented in detail, or as a way to measure the strength of the association without fitting any model to the data. Association indices offer a summary measure of the level of the association, while models are most useful to describe its patterns in a more fine-grained way or to test specific hypotheses.

[14] These figures are computed respectively via $(1 - 0.94^{-2} \times 1.05^4)/(0.94^{-2} \times 1.05^4)$ and $0.94^2 \times 1.05^4 - 1$.

[15] We should note that the Unidiff and the shrinkage estimators may not be completely appropriate methods for a within-country analysis, since they tend to make the estimates closer to the European average, rather than to national averages which are the reference in the fixed-effects model. The disturbance thus introduced could explain at least in part why the estimated effects are smaller than with the non-parametric method.

# 6 Conclusion

We have presented three closely related marginal-free indices of the association in a contingency table: the Altham index, the intrinsic association coefficient, and a normalized variant of the latter index. The Altham index has been used several times in recent empirical works, but its relationship to odds ratios, log-linear and log-multiplicative models (in particular association models) had not been developed systematically until now. On the contrary, the intrinsic association coefficient was originally proposed by Goodman in the context of row-column (RC) association models, and later identified by him as a fundamental quantity, equivalent in the odds ratio framework to the Pearson mean square contingency coefficient $\varphi^2$ or to Cramér's V. Yet, it has not been used in empirical applications. We have shown that this index is actually equal to the standard deviation of all the log-odds ratios that can be constructed from a table, and that it is directly related to the layer coefficient estimated by the Unidiff model. Finally, we have proposed a normalized variant of the intrinsic association coefficient varying between 0 and 1 which is equivalent to the correlation coefficient under a bivariate normal distribution, in order to make the interpretation and the presentation of results more intuitive.

Despite the very strong links between the three indices, it seems that the intrinsic association coefficient should be preferred to the Altham index. Indeed, the Altham index mechanically increases with the number of rows and columns of the table, making its scale somewhat arbitrary and its interpretation difficult. This is not the case of the two variants of the intrinsic association coefficient, which on the contrary allow for (careful) comparisons across different classifications, for example between different socioeconomic classifications, or even between socioeconomic and educational dimensions of homogamy or social mobility. Other advantages can be highlighted: the intrinsic association coefficient fits very well in the framework of association models, since it appears directly in their equations; the normalized version varying between 0 and 1 is measured on a more easily interpretable scale.

We hope that these indices can be useful for empirical research regarding at least three aspects. First, they allow comparing the overall strength of the association as predicted by several, possibly very different log-linear or log-multiplicative models. This is particularly useful for models which do not provide a single parameter summarizing the strength of the association. As we have shown above regarding educational and socioeconomic homogamy among European regions, these indices therefore make it easy to test multiple specifications and check whether results are robust, which can in some cases prevent drawing incorrect conclusions.

Second, using one of the indices proposed in the present article will make it possible to compare results of several studies after the fact (as in a meta-analysis). This is currently hindered by the diversity of models used in the literature, even when the research questions and methods are very similar (as in the case of homogamy and intergenerational mobility). To this end, the insensitivity of the intrinsic association coefficient to the dimensions of the table is essential.

Third, the standardization of the measurement of the association on a single quantity should help establishing the credibility of the sociological approach to phenomena such as intergenerational mobility, notably in comparison with economic approaches based on the intergenerational elasticity or correlation coefficient. Any of the indices described here can be used for this purpose, since one index can easily be translated into the other just from published tables.

Finally, let us note that extensions of the intrinsic association coefficient can be devised to decompose the overall association into a symmetric component and a skew-symmetric component. The index can very naturally be combined with various quasi-symmetric specifications and with the skew-symmetric log-multiplicative association model proposed by Peter van der Heijden and Ab Mooijaart (1995). Further work would also be in order to derive confidence intervals for the non-parametric and semi-parametric estimators of the indices.

# 7 References


Agresti, Alan. 2002. *Categorical Data Analysis*. 2nd ed. New York: Wiley.

Altham, Patricia M. E. 1970. "The Measurement of Association of Rows and Columns for an R x S Contingency Table." *Journal of the Royal Statistical Society. Series B (Methodological)* 32(1):63–73.

Altham, Patricia. M. E. and Joseph P. Ferrie. 2007. "Comparing Contingency Tables Tools for Analyzing Data from Two Groups Cross-Classified by Two Characteristics." *Historical Methods* 40(1):3–16.

Becker, Mark P. 1989. "On the Bivariate Normal Distribution and Association Models for Ordinal Categorical Data." *Statistics & Probability Letters* 8(5):435–40.

Becker, Mark P. and Clifford C. Clogg. 1988. "A Note on Approximating Correlations from Odds Ratios." *Sociological Methods & Research* 16(3):407–424.

Becker, Mark P. and Clifford C. Clogg. 1989. "Analysis of Sets of Two-Way Contingency Tables Using Association Models." *Journal of the American Statistical Association* 84(405):142–51.

Bishop, Yvonne M., Stephen E. Fienberg, and Paul W. Holland. [1975] 2007. *Discrete Multivariate Analysis: Theory and Practice*. New York: Springer.

Blanden, Jo. 2013. "Cross-Country Rankings in Intergenerational Mobility: A Comparison of Approaches from Economics and Sociology." *Journal of Economic Surveys* 27(1):38–73.

Bouchet-Valat, Milan. 2018a. "Educational and socioeconomic homogamy, development level, and metropolisation across 149 European regions." *Revue européenne des sciences sociales/European Journal of Social Sciences* 56(1): 53–84.

Bouchet-Valat, Milan. 2018b. "logmult: Log-multiplicative models, including association models." R package. Version 0.7.0. http://CRAN.R-project.org/package=logmult

Bourdieu, Jérôme, Joseph P. Ferrie, and Lionel Kesztenbaum. 2009. "Vive La Différence? Intergenerational Mobility in France and the United States during the Nineteenth and Twentieth Centuries." *Journal of Interdisciplinary History* 39(4):523–57.

Breen, Richard, Ruud Luijkx, Walter Müller, and Reinhard Pollak. 2009. "Nonpersistent Inequality in Educational Attainment: Evidence from Eight European Countries." *American Journal of Sociology* 114(5):1475–1521.

Clogg, Clifford C. 1982. "Some Models for the Analysis of Association in Multiway Cross-Classifications Having Ordered Categories." *Journal of the American Statistical Association* 77(380):803–15.

Clogg, Clifford C. and Edward S. Shihadeh. 1994. *Statistical Models for Ordinal Variables*. Thousand Oaks, CA: Sage.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, MI: Lawrence Erlbaum Associates.

Dijkstra, Lewis. 2009. *Metropolitan Regions in the EU*. European Union Regional Policy. http://ec.europa.eu/regional_policy/

Domański, Henryk and Dariusz Przybysz. 2007. "Educational Homogamy in 22 European Countries." *European Societies* 9(4):495–526.

Ebbinghaus, Bernhard. 2005. "When Less Is More. Selection Problems in Large-N and Small-N Cross-National Comparisons." *International Sociology* 20(2):133–52.

Erikson, Robert and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford, United Kingdom: Clarendon Press.

Ferrie, Joseph P. 2005. "History Lessons: The End of American Exceptionalism? Mobility in the United States since 1850." *The Journal of Economic Perspectives* 19(3):199–215.

Gilula, Zvi and Shelby J. Haberman. 1986. "Canonical Analysis of Contingency Tables by Maximum Likelihood." *Journal of the American Statistical Association* 81(395):780–88.

Goodman, Leo A. 1969. "How to Ransack Social Mobility Tables and Other Kinds of Cross-Classification Tables." *American Journal of Sociology* 75(1):1–40.

Goodman, Leo A. 1981a. "Association Models and Canonical Correlation in the Analysis of Cross-Classifications Having Ordered Categories." *Journal of the American Statistical Association* 76(374):320–334.

Goodman, Leo A. 1981b. "Association Models and the Bivariate Normal for Contingency Tables with Ordered Categories." *Biometrika* 68(2):347–55.

Goodman, Leo A. 1985. "The Analysis of Cross-Classified Data Having Ordered And/or Unordered Categories: Association Models, Correlation Models, and Asymmetry Models for Contingency Tables With or Without Missing Entries." *The Annals of Statistics* 13(1):10–69.

Goodman, Leo A. 1986. "Some Useful Extensions of the Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables." *International Statistical Review* 54(3):243–70.

Goodman, Leo A. 1991. "Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data." *Journal of the American Statistical Association* 86(416):1085–1111.

Goodman, Leo A. 1996. "A Single General Method for the Analysis of Cross-Classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule, and Fisher, and Also Some Methods of Correspondence Analysis and Association Analysis." *Journal of the American Statistical Association* 91(433):408–28.

Goodman, Leo A. and Michael Hout. 1998. "Statistical Methods and Graphical Displays for Analyzing How the Association Between Two Qualitative Variables Differs Among Countries, Among Groups, Or Over Time: A Modified Regression-Type Approach." *Sociological Methodology* 28(1):175–230.

Goodman, Leo A. and Michael Hout. 2001. "Statistical Methods and Graphical Displays for Analyzing How the Association Between Two Qualitative Variables Differs Among Countries, Among Groups, Or Over Time. Part II: Some Exploratory Techniques, Simple Models, and Simple Examples." *Sociological Methodology* 31(1):189–221.

van der Heijden, Peter G. M. and Ab Mooijaart. 1995. "Some New Log-Bilinear Models for the Analysis of Asymmetry in a Square Contingency Table." *Sociological Methods & Research* 24:7–29.

Hout, Michael, Clem Brooks, and Jeff Manza. 1995. "The Democratic Class Struggle in the United States, 1948-1992." *American Sociological Review* 60(6):805–28.

Katrňák, Tomáš, Petr Fučík, and Ruud Luijkx. 2012. "The Relationship between Educational Homogamy and Educational Mobility in 29 European Countries." *International Sociology* 27(4):551–73.

Katrňák, Tomáš, Martin Kreidl, and Laura Fónadová. 2006. "Trends in Educational Assortative Mating in Central Europe: The Czech Republic, Slovakia, Poland, and Hungary, 1988–2000." *European Sociological Review* 22(3):309–22.

Liebetrau, Albert M. 1983. *Measures of Association*. Thousand Oaks: Sage.

Long, Jason and Joseph Ferrie. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States Since 1850." *The American Economic Review* 103(4):1109–1137.

Meron, Monique and Michel Amar, eds. 2014. *Final Report of the European Statistical System Network on the Harmonisation and Implementation of a European Socio-Economic Classification: European Socio-Economic Groups (ESeG)*. Paris: Insee. http://ec.europa.eu/eurostat/cros/sites/crosportal/files/ESEG_finalReport_V10j

Park, Hyunjoon and Jeroen Smits. 2005. "Educational Assortative Mating in South Korea: Trends 1930-1998." *Research in Social Stratification and Mobility* 23:103–27.

Raymo, James M. and Yu Xie. 2000. "Temporal and Regional Variation in the Strength of Educational Homogamy." *American Sociological Review* 65(5):773–81.

Rudas, Tamás. 1998. *Odds Ratios in the Analysis of Contingency Tables*. Thousand Oaks: Sage.

Skinner, Chris and Louis-André Vallet. 2010. "Fitting Log-Linear Models to Contingency Tables From Surveys With Complex Sampling Designs: An Investigation of the Clogg-Eliason Approach." *Sociological Methods & Research* 39(1):83–108.

Smits, Jeroen. 2003. "Social Closure Among the Higher Educated: Trends in Educational Homogamy in 55 Countries." *Social Science Research* 32(2):251–77.

Smits, Jeroen, Wout Ultee, and Jan Lammers. 1998. "Educational Homogamy in 65 Countries: An Explanation of Differences in Openness Using Country-Level Explanatory Variables." *American Sociological Review* 63(2):264–85.

Smits, Jeroen, Wout Ultee, and Jan Lammers. 1999. "Occupational Homogamy in Eight Countries of the European Union, 1975-89." *Acta Sociologica* 42(1):55–68.

Smits, Jeroen, Wout Ultee, and Jan Lammers. 2000. "More or Less Educational Homogamy? A Test of Different Versions of Modernization Theory Using Cross-Temporal Evidence for 60 Countries." *American Sociological Review* 65(5):781–88.

Wong, Raymond Sin-Kwok. 2010. *Association Models*. Thousand Oaks: Sage.

Xie, Yu. 1992. "The Log-Multiplicative Layer Effect Model for Comparing Mobility Tables." *American Sociological Review* 57(3):380–95.

Zhou, Xiang. 2015. "Shrinkage Estimation of Log-Odds Ratios for Comparing Mobility Tables." *Sociological Methodology* 45(1):320–56.