doi.org/10.26398/IJAS.0029-012

TOTAL INERTIA DECOMPOSITION IN TABLES WITH A FACTORIAL STRUCTURE

Michael Greenacre[1]

Universitat Pompeu Fabra, Barcelona, Spain

Tor Korneliussen

Nord University, Bodø, Norway

Abstract. *In correspondence analysis the variance in a data set is measured by its total inertia. The decomposition of this inertia into parts per row or per column, per dimension, or per row or column on each dimension, is well known. In this paper we consider various other forms of inertia decomposition, inertia between demographic groups and inertia interactions, inertia of substantive versus non-substantive responses. We also consider the case where comparisons between pre-specified groups can be isolated on particular dimensions of the solution, using correspondence analysis of matched matrices. The idea of hierarchically breaking down the inertia into smaller parts of interest and into contrasts is illustrated using data on decision-making by tourists, collected in a Eurobarometer survey. This application shows patterns between the countries of the EU and gender and age effects and interaction within countries. These effects are also tested for statistical significance using permutation testing.*

Keywords*: Analysis of variance, Contrast, Correspondence analysis, Eurobarometer, matched matrices, Tourism.*

1. INTRODUCTION

In the correspondence analysis (CA) of a rectangular table of non-negative data, the total variance is measured by the *total inertia*, a measure highly related to the Pearson chi-square statistic computed on a contingency table. Specifically, the total inertia is equal to the chi-square statistic divided by the grand total of the table, the total inertia has substantive meaning even if the table is not a true contingency table and is decomposed into parts along respective principal dimensions of the CA, just like total variance is decomposed along principal axes in principal component analysis (PCA). As in PCA as well, these parts are expressed as percentages of the total to quantify how much the dimensions "explain" the data table.

[1] Corresponding author: Michael Greenacre, email: michael.greenacre@gmail.com

Each part of inertia along a principal dimension, which is an eigenvalue (or squared singular value), called a *principal inertia*, can also be decomposed into parts across the rows or across the columns. These parts lead to the well-known contributions to inertia that assist in the interpretation of the CA maps, giving diagnostics for deciding which rows or which columns are driving the dimensions of the solution, or for measuring how well the dimensions are explaining the inertias of the individual rows or individual columns.

A study can be centred around a two-way table, but can also include additional categorical variables that define a factorial structure. For example, a political survey might be aimed at comparing the voting patterns across the regions of a country, giving a matrix with regions as rows and political parties as columns, with the percentage of votes by region i for party j in the (i, j)-th cell of the data table. In addition, the survey could collect the gender of each respondent and the age group. So a similar table could be obtained for each gender, for each age group or even for each gender-age group combination. In this last case, there is a total inertia for all these tables and this can be broken down into parts that are reminiscent of an analysis of variance, so could be called an *analysis of inertia*, the only difference being that the data table serves as a response matrix rather than a single response variable. A part can be ascribed to gender difference, a type of "main effect" for gender, a part to age group differences, i.e. "main effect" for age group, and a part to the "interaction" between gender and age group. We put "main effect" and "interaction" between quotation marks because the original regions-by-political parties table is already summarizing a two-way interaction structure, so the "main effects" mentioned above are really a three-way interaction, while the "interaction" is a four-way interaction.

In this paper we will show how to decompose the total inertia into these parts, which in turn will decompose along dimensions and then for the set of rows or set of columns, just as in the simple case of a single table. The theory will be illustrated for the simple case of a two-way table on which a factorial structure is defined by two binary variables. The field of application is in tourism and concerns the information sources that European tourists use when deciding to take a holiday. The application of CA to such a table has already been considered by Korneliussen and Greenacre (2017). In the present paper we add the factorial structure to the data set in terms of gender and age group, and then quantify and visualize the parts of inertia explained by gender, age group and their interaction. Section 2 briefly introduces the data, Section 3 the methodology, Section 4 the results, and finally a discussion in Section 5.

2. DATA ON INFORMATION SOURCES USED BY EUROPEAN TOURISTS

This article uses data from the Eurobarometer survey, "Flash Eurobarometer 258"[2], conducted in 2009, to investigate associations between country and use of information sources. This survey about the attitudes of Europeans towards tourism provides data from representative samples of the 27 EU countries. The sample sizes vary between 501 and 2000 per country, with an average sample size of 1004. The data are mainly based on telephone interviews. Some eastern European countries have low fixed-line coverage. In these countries face-to-face interviews were also carried out. All in all, more than 27,000 randomly selected citizens aged 15 and over were interviewed. The average non-response rate was 6.3%.

In order to measure information source use, this study uses the responses to the question: "From the following information sources, which one do you consider to be the most important when you make a decision about your travel/holiday plans?" Then followed seven information sources: a) personal experience, b) recommendation of friends and colleagues, c) guidebooks and magazines (commercial), d) catalogues, brochures (non-commercial), e) the internet, f) travel / tourist agencies, g) media (newspaper, radio, TV).

Korneliussen and Greenacre (2017) publish the 27×8 table of percentage responses in each country, with the seven information sources and a missing response category as columns. We extend this study to include the gender and age group (in two groups, approximately of equal sizes in the total data set: "younger", i.e. less than 50 years, and "older", i.e. 50 years or older) of the respondents (in the discussion we will comment on the effect of including more age groups). Percentages were calculated relative to the country totals after reweighting the data according to the respondent weights that accompanied the Eurobarometer data file – "this removes sample bias from the estimates of the percentages for each country. Overall estimates for the EU show that "recommendation of friends and colleagues" was, with 29.3% of the cases, the information source most often reported to be the most important. This was followed by the internet (21.9%) and personal experience (18.8%), travel/tourist agencies (11.4%), catalogues and brochures (5.5%), guidebooks and magazines (4.8%) and media (3.2%).

---

[2] See http://ec.europa.eu/public_opinion/flash/fl_258_en.pdf for more information about the survey

3. METHODOLOGY

Because two dichotomous variables, gender and age, are now taken into account, the basic data set now consists of four data tables: **A**, the 27×8 table of percentages for younger males; **B**, the 27×8 table of percentages for older males; **C**, the 27×8 table of percentages for younger females; and **D**, the 27×8 table of percentages for older females. If the missing data column is maintained in the analysis, it usually turns out to be an important feature of the result, which was indeed true for this data set – see, for example, the large percentage of missing data for Bulgaria. Hence, we use a variation of CA called *subset CA* (Greenacre and Pardo 2006; Greenacre 2007, chap. 21), which allows all data to be analysed, including the missing data percentages, but uses just the percentages for the subset of seven information sources to determine the solution. The computations were done in the ca package (Nenadić and Greenacre, 2007) of the R software platform (R Core Team, 2016), as well as with R code written specially for the novel aspects of this analysis. Furthermore, we can test whether the studied effects are statistically significant or not, conveniently using the permutation test for canonical correspondence analysis (CCA) in the **vegan** package (Oksanen et al., 2015) in R so that our analysis extends beyond data description and visualization. CCA is a constrained form of correspondence analysis, where the constraining variables in this application are the gender and age groups – see Greenacre (2010a) for the definition of CCA and Greenacre (2010b) for an explanation in the social science context.

The four tables **A**, **B**, **C** and **D**, can be analysed in a "stacked" 108×8 format (Greenacre, 2016, Chap. 17), in which case the gender and age effects as well as their interaction are all competing for being displayed in the principal plane of the first two principal dimensions. Alternatively, a special block circulant format, called affectionately an "ABBA" format, can be set up to split the effects into separate visualizations (Greenacre, 2003, 2016, Chap. 23). The ABBA format for a single dichotomous variable, for example the analysis of males only, distinguishing between younger and older males, would be as follows:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \tag{1}$$

In the present example, the CA of this 54×16 matrix would lead to dimensions corresponding to all the males, and separately dimensions corresponding to the difference between younger and older males. The dimensions corresponding to these two respective parts can easily be distinguished from the signs of the sub-vectors of the coordinate matrices. This result can be extended to two dichotomous

variables, by nesting the ABBA format as follows:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{D} \\ \mathbf{B} & \mathbf{A} & \mathbf{D} & \mathbf{C} \\ \mathbf{C} & \mathbf{D} & \mathbf{A} & \mathbf{B} \\ \mathbf{D} & \mathbf{C} & \mathbf{B} & \mathbf{A} \end{bmatrix} \tag{2}$$

In the CA of this 108×32 matrix each coordinate vector has four sub-vectors with repeated values but different patterns of signs, which identify the overall "constant", "main effects" and "interaction". This result is illustrated when applied to the tourism information source data in the next section.

The total inertia in this example can be apportioned to the three categorical variables (country, gender and age group) that define the factorial structure of the rows of the four matrices **A**, **B**, **C** and **D** that can be stacked one on top of another. This is conveniently done using the **vegan** package again, where the factors are defined as additional columns of the data set. In the results that follow we start with this decomposition of inertia.

4. RESULTS

The parts of inertia due to country, gender and age as main effects are obtained as follows, using two simple R commands in the **vegan** package, assuming the stacked table is stored as ABCD, apply canonical correspondence analysis (CCA) with country, gender and age coded as categorical explanatory variables (called factors in R):

```
> ABCD.cca <- cca(ABCD ~ country + gender + age)
> anova(ABCD.cca)
```

The results are as follows:

factor	inertia	percentage
country	0.1432	52.7%
gender	0.0058	2.1%
age	0.0688	25.3%
residual	0.0541	19.9%
total	0.2718	100.0%

This shows that the country differences are higher than the age and gender differences, but this is also due to the fact that countries define 27 categories

whereas age and gender only two each. A by-product of the anova function above is a permutation test for each variable, and they are all highly significant at $p<0.001$. Note that this test is conducted at the country level, not at the individual level.

In the above analysis the missing value category is included, but this can be omitted in the subset CAs that follow, and could also have been omitted by a slight rewriting of the cca function. The inertia due to this missing value can be quantified in the subset CA: it contributes 0.0815 to the total in the above table, or almost 30% of the total. This inertia is absent in the subset CAs.

Figure 1 shows the subset CA of the original countries-by-information sources table, where the male-female and younger-older splits for each country are plotted as supplementary points, using plotting symbols m, f, y and o respectively. This map contains all the country, gender and age effects and is clearly too crowded. Our idea is to split this analysis into parts that specifically look at the different effects. Nevertheless, Figure 1 does illustrate the large variance of the countries, then the larger differences between age groups within countries compared to the gender differences (y-o segments generally longer than the m-f segments). For example, for both Finland (FI) and Denmark (DK) at top left of Figure 1, the segments for the gender male-female contrasts are seen to be much shorter than the age younger-older ones.

The analysis of the nested ABBA matrix of (2) provides four CA maps of the separate parts of inertia. The different dimensions can be determined by the pattern of the signs of the four subvectors that form the CA solution (Greenacre, 2016, Chapter 23):

dimension	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
pattern	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	+	+	+	+	+	-	+	+	-	+	+	-	+	-	+	-	-	-	+	+	-	-	-	+	-	-	-	-
	-	+	+	+	+	+	+	-	+	-	+	-	-	+	-	-	+	-	-	+	-	+	-	-	-	+	-	+
	-	+	+	+	+	-	+	-	-	-	+	+	-	-	-	+	-	+	-	+	+	-	+	-	+	-	+	-
part	a	c	c	c	c	g	c	a	g	a	c	i	a	g	a	i	g	i	a	c	i	g	i	a	i	g	i	g

In the above the column pattern ++++ for a particular dimension corresponds to the four tables merged, i.e. the country effect (c), then ++-- focuses specifically on age differences (a), because **A** and **B** correspond to younger respondents, **C** and **D** to older, +-+- on gender differences (g), because **A** and **C** correspond to male respondents, **C** and **D** to females, and +--+ on gender-age interaction (i) (this interaction effect was not measured in the CCA performed above, but could also be included).

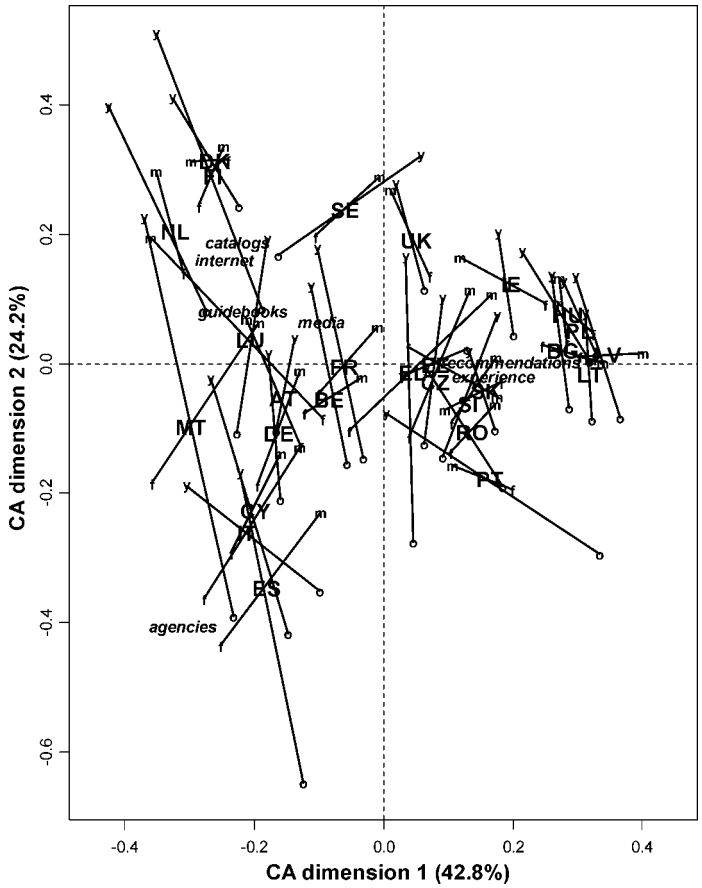

Figure 1: Subset CA of countries-by-information sources data set, with the male (m) and female (f) subgroups within a country, as well as younger (y) and older (o) subgroups shown as supplementary points and connected by line segments that pass through the country points.

The parts of inertia corresponding to the above separation of effects are as follows, obtained by summing the inertias of the corresponding dimensions (remember that this is based on the subset analysis, that contains less total inertia than the previous CCA analysis, which included the missing response category):

part	inertia	percentage
c	0.1015	53.3%
g	0.0160	8.4%
a	0.0650	34.2%
i	0.0079	4.1%
total	0.1903	100.0%

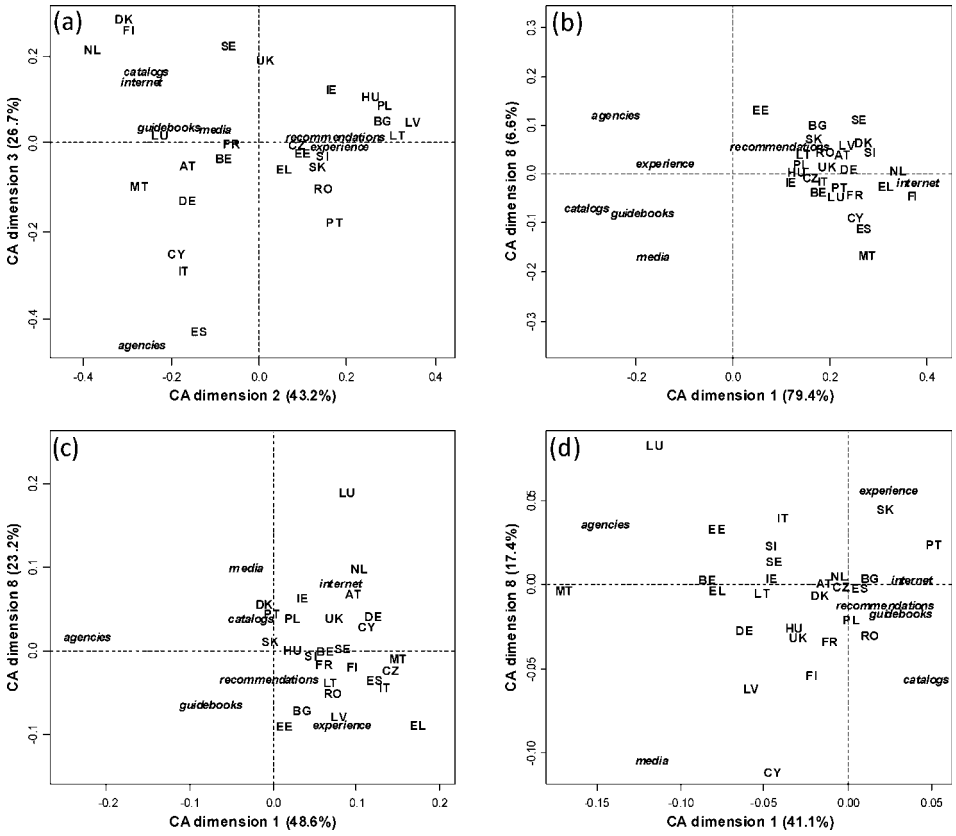

Figure 2: Four pairs of dimensions from the subset CA of the nested block-circulant matrix that visualize four separate effects: (a) the basic country-by-information source analysis; (b) the younger-older differences within countries; (c) the male-female differences between countries; and (d) the interaction of gender and age across countries. Percentages are with respect to the total inertias of the respective parts.

Thanks to the pattern of signs we can identify the best two dimensions for visualizing each effect: dimensions 2 and 3 for c, dimensions 1 and 8 for a, dimensions 6 and 9 for g and dimensions 12 and 16 for i. The four CA maps are shown in Figure 2(a)-(d).

Figure 2(a) essentially shows the same country differences as Figure 1, with eastern European countries on the right generally favouring recommendations and experience when choosing between tourism options, the Scandinavian countries and Holland upper left preferring internet and catalogues, while Spain, Italy and Cyprus, tending to use agencies more than average.

Figure 2(b) shows the differences in the countries between younger and older respondents. These are the line segments y-o of Figure 1, but dimension-reduced from their respective full space of differences onto this map, with the o anchored at the origin, so the country points indicate younger minus older differences. The interpretation clearly shows that in all countries younger respondents use the internet more than older ones, with the biggest differences found in Finland, Holland and Greece.

Figure 2(c) shows the gender differences, smaller than the age ones, where this time female (f) is anchored at the origin, so it is the male minus female differences that are shown. Clearly, males are not using agencies as much as females, in all countries, and in some countries, e.g. Greece, Latvia, Estonia and Bulgaria, males rely on experience more than females, while in Holland and Luxembourg, for example, internet and the media are preferred more by males.

Finally, in Figure 2(d), the more complex interaction effect is visualized. This effect has the least inertia, only 4% of the total as seen previously, of which almost 60% is visualized here. Here we understand questions such as the following: does the male-female difference itself differ when looking at younger versus older participants? Malta seems to be a country lying towards agencies, but from Figure 2(b) younger respondents in Malta using agencies much less than older ones. The implication is that this younger versus older difference is greater for Maltese men than women, which is a component of the interaction effect.

5. DISCUSSION AND CONCLUSION

In this paper we have exposed a methodology that allows the total inertia of a table of data on which a factorial structure is overlaid to be decomposed into parts corresponding to effects defined by the factors (see also Le Roux and Rouanet (2010) for a different treatment of this topic). In particular, for factors with two categories, the separation of the parts can be achieved efficiently by setting up the matrix in a nested block-circulant format, which is then analysed by CA in a single analysis that generates the CAs of all the parts. This separation of the effects makes the interpretation of the results much easier, thanks to both the quantification of the inertia contributions of each effect as well as their visualization in separate maps. In the case of three or more categories defining the factorial structure, the situation is a bit more complex. The block-circulant format with three matrices circulating, say **A**, **B** and **C**, leads to contrasts being quantified and displayed, for example, the contrast $\mathrm{A} - {}^{1}/_{2}(\mathbf{B}+\mathbf{C})$. Of course, any such contrasts or other combinations of the factorial parts can be analysed *per se*, but a regular CA software program can then not be employed because the data should not be recentred, so that the origin of the

display corresponds to the contrast value of zero. In addition, proper attention to the masses and chi-square distances needs to be given. The block-circulant format avoids these issues and assures that the same chi-square distance is used in all analyses and that the centre of the displays for the difference parts (or contrasts, as the case may be) are identified with zero values.

REFERENCES

Greenacre, M. (2003). The singular value decomposition of matched matrices. In *Journal of Applied Statistics.* 30: 1-13.

Greenacre, M. (2010a). *Biplots in Practice*. Bilbao: BBVA Foundation. URL: http://www.multivariatestatistics.org (free download).

Greenacre, M. (2010b). Canonical correspondence analysis in social science research. In H. Locarek-Junge and C.Weighs, editors, *Classification as a Tool for Research*. Heidelberg: Springer-Verlag: 279-286.

Greenacre, M. (2016). *Correspondence Analysis in Practice*. *3rd edition.* Boca Raton, FL: Chapman and Hall/CRC Press.

Greenacre, M. and Pardo, R. (2006). Subset correspondence analysis: visualization of selected response categories in a questionnaire survey. In *Sociological Methods and Research.* 35: 193-218.

Greenacre, M. and Pardo, R. (2007). Multiple correspondence analysis of subsets of response categories. In Greenacre, M.J. and Blasius, J., editors, *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC Press, London: 197–217.

Korneliussen, T. and Greenacre, M. (2017). Information sources used by European tourists: a cross-national study. In *Journal of Travel Research,* DOI: https://doi.org/10.1177/0047287516686426.

Le Roux, B. and Rouanet, H. (2010). *Multiple Correspondence Analysis*. Thousand Oaks, CA: Sage.

Nenadić O. and Greenacre, M. (2007). Correspondence analysis in R, with two – and three – dimensional graphics: the ca package. In *Journal of Statistical Computing*. 20(3). URL: http://www.jstatsoft.org/v20/i03/ (free download).

Oksanen, J., Guillaume Blanchet, F., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H. and Wagner, H. (2015). vegan: *Community Ecology Package. R package version 2.3-2*. URL: http://CRAN.R-project.org/package=vegan.

R Core Team (2016). R: a language and environment for statistical computing. Vienna, Austria. URL: https://www.R-project.org.