*R. Randell et al. (Eds.)*

*doi:10.3233/978-1-61499-753-5-166*

# Application of Correspondence Analysis to Graphically Investigate Associations Between Foods and Eating Locations

Andrew N. CHAPMAN[a,] , Eric J. BEH[c], Luigi PALLA[a,b1]
[a] *Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine*
[b]*Farr Institute of Health Informatics,University College London*
[c]*School of Mathematical & Physical Sciences, University of Newcastle, Australia*

**Abstract**. This paper presents the application of correspondence analysis (CA) for investigating associations using confidence regions (CRs) with a focus on facilitating mining the data and hypothesis generation. We study the relationship between locations and “less-healthy” food consumption by UK teenagers. CA allows for a quick visual inspection of the various association structures that exist between the categories of cross-classified variables in large datasets derived with varying study designs. The hypotheses generated by the visual display can then be independently tested using suitable regression models. CA makes use of readily available software tools and of robust statistical tests amenable to interpretation.

**Keywords.** Correspondence Analysis, Confidence Regions, Hypothesis generation, Hypothesis testing, clustered data, hierarchic data, location, food-group, healthy.

## 1. Introduction

Effective public health policies are needed to discourage less-healthy eating habits. Identifying where such habits develop may facilitate any intervention to modify them. Recently, the location of eating has been related to change in social context [1] and also linked to diet quality in adolescents [2]. In order to investigate the latter issue further, we used the National Diet and Nutrition Survey (NDNS) database for the years 2008-2011 to analyse the association between eating locations and “less-healthy” food-groups consumed by UK teenagers aged 11-18 years. The published description of the NDNS programme [3] is used by the federal Government to monitor progress on diet and nutrition objectives. We focused on teenagers as habits taken up at that stage may have an effect throughout life and thus are potential intervention targets.

The response rate for the completion of a four-day food diary and a lifestyle interview was 56%. A total of 884 teenagers aged 11 to 18 years responded, providing a total of 62,523 diary records. The mean (standard deviation) of number of diary entries over four days for an individual teenager is 71 (22). Such entries are not independent of each other due to a two-level hierarchy of correlation in the diary records: individual and meal-time. Our analysis was performed purely on instances

[1] Corresponding author, luigi.palla@lshtm.ac.uk .

rather than quantities and addressed the open question whether there is a relationship between less-healthy eating and the location where the food is consumed.

## 2. Methods

We randomly selected half the NDNS data diary entries for hypothesis generation by cross-tabulating foods and eating locations, and the remaining half of the diary entries were used for performing the hypothesis tests that we generated.

### *2.1. Correspondence Analysis CA and Confidence Regions (CRs)*

The contingency table (frequency matrix) was analysed by Correspondence Analysis (CA) (cf [4] for practical guide and [5] for a historical and up-to-date theoretical treatment of CA). CA allows us to investigate the relationship between the categories of two variables, by projecting them on two dimensions (axes) that jointly represent a large percentage (displayed on the axes) of the $\chi^2$ deviation (called inertia) of the table from the condition of independence between the two variables (foods and locations).

There are two ways to plot CA results in the two most informative dimensions: symmetric plots and biplots. In symmetric CA plots, both row and column profile-points are normalised. This is useful for row-row and column-column associations but distorts [4] associations between rows and columns. Hence we preferred the biplot where row profiles are normalised (rescaled) but column profiles are not (or vice versa) and the direction of the column profiles are shown as arrows from the origin to facilitate the interpretation of row-column associations via the size of the angle between the categories. We can also sometimes find meaning for an axis by considering the opposition and placement of points along that axis. Importantly, this descriptive method does not require any assumptions about the data. Therefore, it can be used to explore correlated data. We also verified that the complex survey design should not affect the CA [7] and the application of survey weights in the analysis yielded negligible differences in the results/plots.

Two algorithms using the freely available "R" software routines are available to enhance CA plots with elliptical confidence regions (CRs) for each profile-point: CAvariants uses algebra and parametric-assumptions [8] and CABOOTCRS uses a bootstrap method [6]. We used CRs from CABOOTCRS with a 95% confidence level to eliminate those locations and food-groups with larger CRs that overlap the origin, since their deviation from the average profile could be explained by sampling variation. Associations amongst the remaining food-groups and locations suggested hypotheses which were tested using logistic regression/Generalized Estimating Equations (GEE).

Our CA plots for all food-groups were cluttered with too many overlapping profile points and CR ellipses. So, when constructing the CRs, we sub-divided the food-groups into categories: "healthy", "neutral" and "less-healthy" as defined in [9] using the UK Food Standards Agency (FSA) nutrient profiling system detailed in [10].

The CA method allows profiles to be sub-divided without introducing inconsistencies: this is one reason for using "inertia" as the metric in CA analysis [4].

### 2.2. GEE and Odds Ratios ORs

For hypothesis testing, we accounted for correlation by performing logistic regression by Generalised Estimating Equations (GEEs) yielding population mean ORs. GEEs provides unbiased estimates [11] of ORs despite our ignorance of the true correlation of diary entries within mealtimes, and within/between individuals. Since the correlation structure in the model is unreliable, the GEE model (via the SAS procedure GENMOD) provides empirical estimates of standard errors.

## 3. Results

Using a random process to split the diary records for the top 25 food-groups resulted in a hypothesis generating dataset of 20,567 diary records and a testing dataset of 20,455.

We plotted twelve of the top 25 food-groups which were classified as less-healthy: Biscuits, Crisps, Chocolate and Sweets, Buns & Cakes, Miscellaneous Foods, Cheese, Low-fibre Cereals, Sausages, Fried Chicken, Lower-fat spreads, Meat-pies & Pastries, Jams. Figure 1 and figure 2 are CA plots that summarise various associations. For both figures, the first axes reflects the greatest correlation (66.82%) between foods and location while the second axis accounts for 21.4% of this association. Therefore figures 1 and 2 reflect 21.4%+66.82% = 88.22% of the association between the two variables and so provides an excellent visual summary of how the two variables are related.

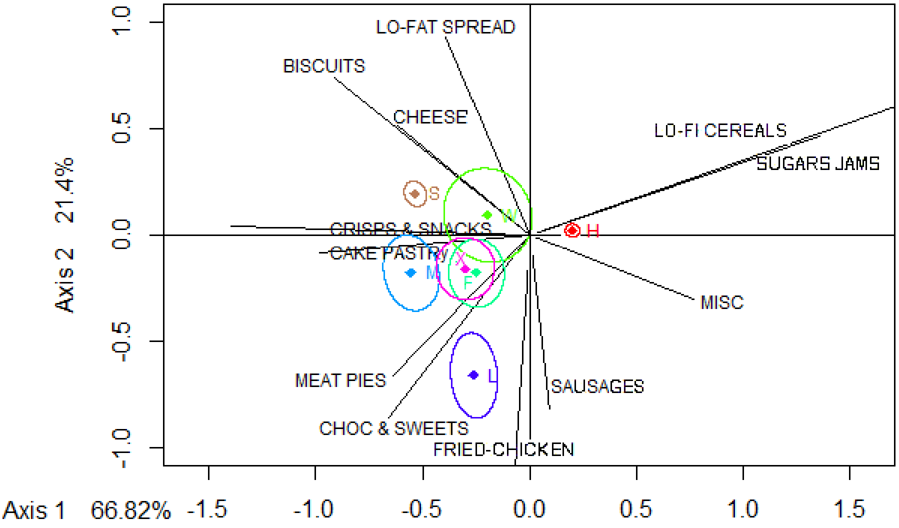

**Figure 1.** Biplots of Less-Healthy Foods with 95% bootstrap CRs for **Locations** H-Home S-School W-Work F-Friends/Carers L-Leisure M-Mobile X-Other

In the top-left quadrant of figure 1, Cheese, Less-fat spreads and Biscuits appear associated with School and Work. Moving anti-clockwise, we find Crisps, Cakes and Pastries are associated with non-Home locations. At the bottom of figure 2, we find Fried Chicken is associated with Leisure locations.

In both figures, Meat-Pies and Chocolate appear in the bottom-left quadrant associated with Leisure, Friend's & Carer's homes, Other and Mobile locations. In figure 2, the “Chocolate & Sweets” CR appears entirely inside the CR for Meat-Pies; this invited further investigation.

In figure 2, the locations have aligned themselves in three main directions which we have used to simplify our hypotheses: Home, then School and Work together, and then all Other locations (Leisure, Mobile, Other, Friend’s & Carer’s homes). As it is usually harder to collect information on eating behaviour at “Other” locations, we focus on those food-groups associated with eating locations away from Home and away from School among which Meat-Pies and “Choc and Sweets” are the main candidates, based on the biplots.

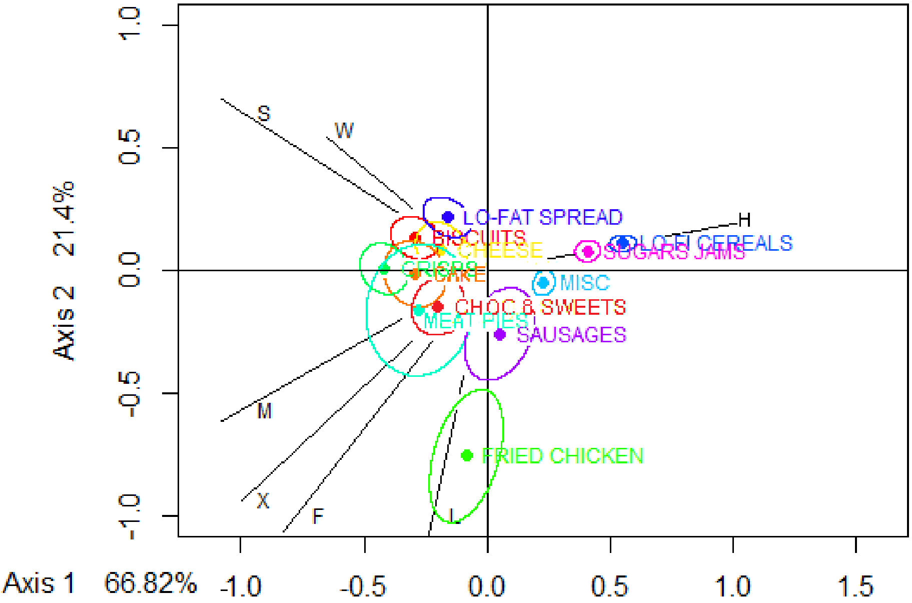

**Figure 2.** Biplots of Less-Healthy Foods with 95% bootstrap CRs for **Food_Groups** H-Home S-School W-Work F-Friends/Carers L-Leisure M-Mobile X-Other

Based on the findings from the two CA plots, we tested two hypotheses using logistic regression (adjusted by sex, age, weekend and socio-economic status): that “Choc and Sweets” and Meat-Pies were more likely to be consumed when a teenager finds himself/herself at “Other” locations away from School/Work and away from Home. Our results are summarised in Table 1:

**Table 1.** Adjusted Odds Ratio estimates for food-groups as outcomes and location-types as exposures.

| Food-Group | OR vs<br>Home | 99% CI<br>p-value | OR vs<br>School-Work | 99% CI<br>p-value |
|---|---|---|---|---|
| Choc and<br>Sweets | **2.5** | (1.8, 3.4)<br>p<0.0001 | **1.8** | (1.2, 2.8)<br>p=0.0002 |
| Meat Pies | **2.8** | (1.5, 5.0)<br>p<0.0001 | 1.3 | (0.6, 3.0)<br>p=0.44 |

We also detected strong evidence (p=0.002) of a linear trend for age so that each additional year of a teenager’s age increases the odds of his/her eating a Meat-Pie by a factor of 1.15 with 99% CI (1.02, 1.30).

## 4. Discussion and Conclusion

In this paper, we showed how correspondence analysis with confidence regions facilitated the systematic investigation of the relationship between categorical variables from a large dataset. This analysis generated plausible hypotheses concerning associations which were then tested using regression methods that take into account the correlation between observations.

Our specific aim was to explore and test for associations between foods consumed and eating locations, recorded in NDNS diet diaries. We focussed on less-healthy foods and found evidence of higher odds of consumption of those at locations away from Home and School. Therefore, public health interventions are warranted to reduce the consumption of foods like meat-pies and chocolate and sweet snacks in other locations.

Some methodological challenges still arise from our work. For example, the CRs in CA pertain to statistical significance of each of the categories of the variables and not to any measure of the association between the categories. Moreover, while the CRs were the key data-mining tool, it is plausible that the correlation in the data warrants further investigation to refine CA and the CRs. Another fundamental issue that may affect the results here is the possible bias due to under-reporting should the probability to report eating a food vary according to eating location. However, evidence of under-reporting for diet diaries is likely related to amount rather than to reporting vs non reporting, which we have overcome by counting food instances (0/1) rather than summing food quantity (grams) per location.

Despite these challenges, we stress the usefulness of CA and its CRs as a data mining tool for hypothesis generation for the analysis of cross-classified categorical variables.

## References

[1] Holm L., et al, 2016, Changes in the social context and conduct of eating in four Nordic countries between 1997 and 2012: *Appetite* 103, p. 358-368.

[2] Williams J.L., 2016, Spaces between home and school: The effect of eating location on adolescent nutrition: *Ecology of Food and Nutrition*, v.55-1, p.65-86.

[3] www.gov.uk/government/uploads/system/uploads/attachment_data/file/310995/NDNS_Y1_to_4_UK_report.pdf , 2013, NDNS Report Y1 to Y4.

[4] Greenacre M. J., 1993, *Correspondence Analysis in Practice*: London, Academic Press, 195 p.

[5] Beh E. J., Lombardo R., 2014, *Correspondence Analysis: Theory, Practice and New Strategies*, Wiley.

[6] Ringrose T. J., 2012, Bootstrap confidence regions for correspondence analysis: *Journal of Statistical Computation and Simulation*, v. 82, p. 1397-1413.

[7] Nyfjall M., 2002, Aspects on Correspondence Analysis Plots under Complex Survey Sampling Designs, Research Report, Department of Information Science, Division of Statistics, Uppsala University.

[8] Beh E. J., Lombardo R., 2015, Confidence Regions and Approximate p-values for Classical and Non Symmetric Correspondence Analysis: *Commun.. in Statistics-Theory and Methods*, v. 44, p. 95-114.

[9] Pechey R., et al. 2013, Socioeconomic differences in purchases of more vs. less healthy foods and beverages: Analysis of over 25,000 British households in 2010: *Social Science & Medicine*, v. 92, p. 22-26.

[10] www.food.gov.uk, 2011, Nutrient Profiling Technical Guidance, in F. F. S. Agency

[11] Liang K. Y., Zeger S.L., 1986, Longitudinal data-analysis using generalized linear-models: *Biometrika*, v. 73, p. 13-22.